## AN ABSTRACT OF THE DISSERTATION OF

Adam J. Chouinard for the degree of Doctor of Philosophy in Zoology presented on June 10, 2016.

Title: Pheromone Variation in Plethodontid Salamanders.

Abstract approved:

Virginia M. Weis

Plethodontid salamanders have served as an informative vertebrate system for studying the role of chemical signals in facilitating social and reproductive behaviors. Individuals produce complex mixtures of chemicals from multiple glandular regions. In total, these secretions convey a wide variety of information, and are important for numerous inter- and intraspecific interactions. In order for a signal to convey complex information, it must vary in its composition. This research explores additional aspects of pheromone variation, within and between species, and discusses how this variation may affect the many functions of chemical signals in these species. Chapter One provides an overview of communication, chemical signaling, and essential background for the system in study. Chapter Two documents the degree of intraspecific variation in a major courtship pheromone protein, Plethodontid Receptivity Factor (PRF). Chapter Three explores the evolution of an additional level of signal complexity (post-translational modification via glycosylation) in the same pheromone, and documents substantial interand intraspecific variation in this trait. Lastly, Chapter Four uses high-throughput RNA sequencing to provide the first detailed description of the genes expressed by three other signaling glands: the cloacal glands, postcloacal gland, and the dorsal tail base. These tissues are involved in different aspects of reproductive, territorial, and other social behavior. Finally, Chapter Five summarizes the findings of this dissertation, incorporates these findings into an integrated model of chemical communication in plethodontids, and discusses key hypotheses to consider and directions for future research.
©Copyright by Adam J. Chouinard June 10, 2016
All Rights Reserved

## by

Adam J. Chouinard

# A DISSERTATION 

submitted to

Oregon State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Presented June 10, 2016
Commencement June 2017

## APPROVED:

Major Professor, representing Zoology

Chair of the Department of Integrative Biology

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Adam J. Chouinard, Author

## ACKNOWLEDGEMENTS

I would like first and foremost to thank my advisor, Dr. Virginia Weis, for her support and mentorship throughout my doctoral studies. Dr. Weis represents the very best type of professor, and this dissertation would literally not have been possible without her. Secondly, I extend my gratitude to the other four members of my committee. Dr. Mason has been an incredible mentor and friend throughout my time here, and I thank him sincerely for the faith he has placed in me, the opportunities he has provided me, and the support and guidance he has given me over the years. Likewise, Dr. Denver and Dr. Karplus have both provided excellent insights and counsel as I opened the door to the molecular world. They have pushed me to think more critically about biological phenomena and the process of science. Lastly, I thank Dr. Magnusson for her support and assistance as my Graduate Council Representative. Each of these faculty have their own interesting research projects and little time to spare between their many committments. Yet, they have dedicated a significant amount of that time over the years to reading and talking about my own research and professional development as a scientist. For that I am truly grateful.

This dissertation would also not have been what it is without the help of many brilliant peers. Sarah Eddy has been an incredible role model and resource for both research and teaching, and inspired me to think more formally about my commitment to quality instruction. Josef Uyeda was an invaluable resource for gaining practical skills and theoretical knowledge about evolutionary biology. Chris Friesen was an indispensable ally in discussing physiology, behavior, and, of course, reproduction. There are many more graduate students that helped in other ways, and I thank all of them for the great memories in talking about science and life in general.

Lastly, Damien Wilburn was an instrumental collaborator for this research, and his technical expertise helped several of my ideas become reality. This thesis greatly benefitted from his enthusiasm and generosity. Likewise, I thank Drs. Houck and Feldhoff for their support in the early stages of this work. All of these individuals helped reinforce the type of scientist, teacher, and person I want to be.

## CONTRIBUTIONS OF AUTHORS

## Chapter Two

Samples from previous years $(2007,2008)$ were provided by RCF and LDH. Animals from 2010 were collected by AJC and DBW with help from RCF and LDH. AJC and LDH performed gland excisions as RCF and DBW carried out mental gland extractions. HPLC analyses were performed in the lab of RCF. AJC performed all data analyses and figure preparation. AJC wrote the paper with help from DBW, LDH, and RCF.

## Chapter Three

Samples from NH were collected by AJC, including animal collection, mental gland excisions, and protein extractions. DBW ran these samples in the SDS-PAGE gel used for a different publication; the gel was adapted and modified from this reference with the NH samples included. DBW carried out cDNA synthesis and sequencing of NH samples. Additional PRF sequences from previous publications were obtained from the NCBI database by AJC. AJC prepared all sequences, performed all data analyses, and prepared all figures. AJC wrote the paper with help from VW.

## CHAPTER FOUR

Animals were collected by AJC, SLE, and DBW. AJC performed gland excisions and DBW carried out RNA exrtraction and cDNA synthesis. SLE provided the motivation to include cloacal glands. Sequencing was performed by a third-party (Otogenetics). AJC performed all bioinformatic analyses and prepared all figures. AJC wrote the paper with help from VW.

## Author Abbreviations

Adam J. Chouinard (AJC)
Damien B. Wilburn (DBW)
Richard C. Feldhoff (RCF)
Lynne D. Houck (LDH)
Sarah L. Eddy (SLE)
Virginia Weis (VW)

## TABLE OF CONTENTS

Chapter ..... Page

1. Introduction .....  1
General Introduction ..... 1
Communication and Chemical Signaling ..... 2
Communication Defined ..... 2
Communication via Chemical Signals ..... 4
Variation and Specificity in Chemical Signals ..... 5
Species Background: Plethodontidae. ..... 6
General Description and Phylogenetics ..... 6
Range, Ecology, and Territoriality ..... 8
Reproduction and Courtship. ..... 9
The Role and Composition of Chemical Signals ..... 12
Research Focus and Chapter Outline. ..... 14
2. Individual Variation in Pheromone Isoform Ratios of the Red-legged Salamander, Plethodon shermani ..... 23
Introduction. ..... 24
Methods ..... 26
Animal Collection and Protein Extraction ..... 26
Biochemical Analysis. ..... 26
Data Analysis ..... 27
Results ..... 27
Pheromone Composition ..... 27
Minor Components ..... 28
Relative Isoform Ratios ..... 29
Discussion. ..... 30
Proportional Abundance of Major Components ..... 30
Minor Components in the Range of PRF ..... 32
Intraspecific Diversity of Relative Isoform Ratios ..... 33
Conclusions and Future Directions. ..... 36

## TABLE OF CONTENTS (continued)

Chapter Page
3. Inter- and Intraspecific Variation in Pheromone Glycosylation in Plethodontid Salamanders. ..... 43
Introduction ..... 44
Plethodontids as a Model for Reproductive Isolation ..... 44
Courtship Pheromones in Plethodontid Lineages ..... 45
Observations of Glycosylation and Goals of the Present Study. ..... 47
Methods ..... 48
Animal Collection, Gland Removal and Extraction ..... 48
SDS-PAGE Glycoprotein Analysis ..... 49
Sequencing of NH P. cinereus PRFs ..... 49
Inter- and Intraspecific Analyses of PRF Glycosylation. ..... 50
Results ..... 53
SDS-PAGE Analyses of Inter- and Intraspecific Differences in Pheromone Glycosylation ..... 53
Sequence Analyses for NH P. cinereus PRFs ..... 54
Glycosylation Predictions and Sequence Annotations ..... 55
Phylogenetic Analyses of PRF Glycosylation Motifs ..... 56
Structural Implications of PRF Glycosylation. ..... 57
Discussion. ..... 59
Inter- and Intraspecific Differences in Pheromone
Sequence and Glycosylation Potential. ..... 60
Phylogenetic Inferences ..... 61
Structure-Function Relationships ..... 62
Functional Hypotheses. ..... 63
Putative Roles in Speciation. ..... 64
Glycosylated Pheromones in Vertebrates and Others ..... 65
Conclusions and Future Directions. ..... 66

## TABLE OF CONTENTS (continued)

Chapter Page
4. De Novo Transcriptome Assembly of Chemical Signaling Glands in a Plethodontid Salamander. ..... 85
Introduction ..... 86
Chemical Communication in Plethodontid Salamanders. ..... 86
Regions of Chemical Signal Production. ..... 87
Known Plethodontid Pheromones ..... 89
Goals of the Present Study ..... 91
Methods ..... 91
Animal Collection and Gland Removal ..... 91
RNA Extraction and Sequencing ..... 92
Bioinformatic Analyses ..... 92
Results ..... 94
De Novo Transcriptome Assembly ..... 94
Transcriptome and Secretome Annotation. ..... 95
Differential Expression of Transcripts ..... 96
Known Pheromone Identification and Expression ..... 96
Discussion. ..... 97
Identification and Expression of Known Plethodontid Pheromones ..... 98
Identification and Expression of Putative Chemical
Signaling Proteins. ..... 100
Carbohydrate-binding Proteins and Mucins ..... 100
Protease Inhibitors ..... 102
Cysteine Rich Secretory Proteins ..... 102
Reproductive Proteins ..... 103
Hormone-like Proteins. ..... 104
Conclusions and Future Directions. ..... 105

## TABLE OF CONTENTS (continued)

Chapter ..... Page
5. Conclusion ..... 121
The Findings of the Present Research ..... 121
An Integrated Model of Chemical Communication in Plethodontids ..... 123
Future Directions ..... 126
The Function of Mental Gland Pheromones ..... 127
Target Tissues and Receptors in Different Plethodon Lineages ..... 131
The Role of Other Signaling Glands. ..... 134
Summary ..... 136
Bibliography ..... 137
Appendices. ..... 160

## LIST OF FIGURES

Figure Page
1.1. Family-level phylogeny of salamanders (Order Caudata). ..... 19
1.2. Time-scaled phylogeny of the family Plethodontidae ..... 20
1.3. Proteinaceous pheromones in plethodontids. ..... 22
2.1. RP-HPLC chromatogram of mental gland extract. ..... 38
2.2. Proportions of major proteinaceous components of the mental gland ..... 39
2.3. Example chromatograms of individual PRF profiles. ..... 40
2.4. Distribution of individual PRF isoform proportions ..... 41
2.5. Variation in male PRF profiles ..... 42
3.1. SDS-PAGE analysis of mental gland extract from three lineages ..... 68
3.2. Bayesian consensus phylogeny of PRF sequences ..... 70
3.3. Bayesian consensus phylogeny for all known PRF sequences. ..... 71
3.4. Similarity matrices for primary and tertiary structure of P. cinereus PRFs ..... 72
3.5. Annotated sequence alignment for all known PRFs. ..... 74
3.6. Annotated sequence alignment for two populations of $P$. cinereus ..... 75
3.7. Potential for O-linked glycosylation between P. cinereus PRF isoforms ..... 76
3.8. Phylogeny of PRF glycosylation for three main lineages of $P$. cinereus ..... 77
3.9. Representative structural model of PRF from a dermal delivery species. ..... 78
3.10. Representative structural model of N -linked glycan attachment ..... 79
3.11. Structural model of a hypothetical ancestral PRF receptor interaction. ..... 81
3.12. Comparison of glycosylation sites of interleukin-6 and PRF ..... 83
4.1. Full-length transcript reconstruction for all BLAST results. ..... 117
4.2. Differential expression of transcripts between gland tissues ..... 118
4.3. Transcript expression clusters. ..... 119
4.4. Cumulative transcript production per gland for known pheromones ..... 120
A1.1. Individual male mental gland profiles ..... 162
A3.1. Read quality trimming and adapter clipping. ..... 252
A3.2. Distribution of TransRate contig scores. ..... 253

## LIST OF TABLES

Table Page
1.1. Information encoded by plethodontid semiochemicals. ..... 17
4.1. Read count and length metrics before and after trimming and clipping. ..... 106
4.2. Transcriptome assembly quality metrics. ..... 107
4.3. Common BLAST homologies grouped by categories ..... 109
4.4. Relevant secretory proteins in each expression clusters. ..... 112
4.5. Significantly enriched Gene Ontology Terms in each expression cluster ..... 114
A3.1. The frequency and total expression of each unique BLAST homology. ..... 254

## DEDICATION

# "He that breaks a thing to find out what it is has left the path of wisdom." <br> - Gandalf 


#### Abstract

Alas, I am but a fool.

I would like to dedicate this research, as ridiculous as it sounds, to the many creatures of the world, big and small, and salamanders chief among them.


Not long ago we crawled out of the sea, and lived much as they do now.

Humans often take the living world for granted, and in this selfish tradition, I sacrificed animals for this work, to satisfy my own curiosity.

## CHAPTER ONE: INTRODUCTION

## General Introduction

Communication is a fundamental part of being alive. Practically every aspect of an organism's existence depends in some way upon communicating with members of its own species (intraspecific communication) and the many others with which it shares the world (interspecific communication). These interactions define which individuals and species will survive, thrive, and reproduce. In this way, communication between organisms ultimately shapes the very evolution of life on Earth. Studying the many ways in which organisms communicate, including means foreign to our own anthropocentric perspective, is an important endeavor in our efforts to understand biology. This dissertation aims to contribute a few small pieces to this very large puzzle.

The goal of this research is to investigate uncharted aspects of the chemical communication system of an early-diverging tetrapod vertebrate, represented herein by members of the salamander family Plethodontidae (Order Caudata). Specifically, this work will document new aspects of signal complexity in the pheromones used to facilitate various sexual and social behaviors in these species. Previous research has identified numerous behavioral functions for these chemical signals, and parallel efforts have documented several proteins involved in these phenomena. These findings have revealed a complex, multicomponent mixture of rapidly-evolving proteins, but in-depth analyses have largely been restricted to a single gland and function. There are multiple ways in which these composite signals may differ between individuals of the same species, or between lineages, but more work is needed to understand the full extent of signal variation. This will allow us to examine its consequences for the function of these pheromones, and provide a more intimate view the evolution of the species that produce them.

The working hypothesis that unifies all of the following chapters is that variation in a signal used in communication affects the ultimate function of that communication. In the same way that changing the words I write affects the meaning of this text, so too should changing the form of a salamander's chemical signals affect the way its peers
respond to the pheromones involved in so many aspects of their life. In employing this hypothesis, there are several underlying questions that provide the motivation for this research. How do individuals communicate complex information through a chemical medium? How might changes in signal composition affect the physiological and/or behavioral functions of these pheromones? What are the ways in which these chemical signals actually vary between individuals, or between species? How do other, unstudied, pheromone glands within an individual differ in the chemicals they produce, and how might these other signals relate to the behavioral phenomena in which they are implicated? This research alone cannot answer all of these questions, but seeks only to continue exploring the fascinating communication system of one of our important relatives.

To introduce this research, I will first define my use of two key terms: communication and information. I will then provide some background on the role of chemical signals in animal behavior, followed by a brief overview of the species to be studied. This will include a discussion of the relevant biology of plethodontids, the behavioral functions of plethodontid pheromones, and the composition of the pheromone signals studied to date. Lastly, I will summarize the contribution of each individual chapter to the overall endeavor.

This introduction is not intended as a comprehensive review, but is instead meant to provide essential context for the species and communication system to be studied throughout. For more thorough information about salamander and amphibian biology, the reader is referred to Petranka (1998) and Duellman and Trueb (1994) respectively. For a recent review of amphibian (and plethodontid) chemical signaling, see Woodley (2014; 2015). I have also provided a more detailed account of the biology, territoriality, and chemical signaling of plethodontids (particularly $P$. cinereus) in a previous work (Chouinard, 2010).

## Communication and Chemical Signaling

Communication Defined
Most people have a colloquial appreciation for what it means to communicate, and just how important it is for everyday life. And yet, this is not a term for which we can easily ascribe a strict definition, and little consensus exists as to what that definition should be (Bradbury and Vehrencamp, 1998). Perhaps one of the best to date was coined by Wilson (1970), who defined communication as "an action on the part of one organism (or cell) that alters the probability pattern of behavior in another organism (or cell) in an adaptive fashion [though this advantage is not always for both parties]." In this sense, the first organism would be considered the sender and the second a receiver. Thus, in its broadest sense, to communicate is to generate a stimulus (passively or actively) that was selected for its propensity to affect the probability of another entity's actions.

Due to the particular actions and responses selected for in this way, the stimuli involved in communication may appear to encode some aspect of information (or meaning) as a function of its adaptive role. While the use of these words has caused several justifiable concerns from a philosophical standpoint (Bradbury and Vehrencamp, 1998), they are nevertheless a useful framework for discussing communication in a reasonable, efficient manner, as long as authors are explicit in their meaning. As such, I will herein use the word information as shorthand for the trait(s) of the sender that are associated with the use of a specific stimulus (or combination of stimuli) in an act of communication. For example, a sex-specific stimulus selected for its use as a mate attractant may be thought of as encoding the information of the sender's sex.

Other philosophical difficulties in discussing communication have stemmed from the intentionality of the sender (and of course none of the following means to suggest conscious intent). Not all stimuli created by a sender need be produced "for" that purpose (i.e. selected for in senders due to its role in communication). Rather, some stimuli may be produced by a sender for other functions, but still used in communication through selection on the receiver. For example, a prey item does not communicate its presence for a predator to find them, but stimuli they generate may have this effect nonetheless. These concerns have led some authors to distinguish between signals and cues in
communication (Maynard-Smith and Harper, 2003): signals are those acts of communication that have evolved through selective advantage for both parties (sender and receiver), while cues are other stimuli that are adaptive for receivers alone. In the case of all the above terminology, other variations of their definitions exist due to additional considerations, but this is how I will employ these terms throughout this work.

## Communication via Chemical Signals

Chemical signals are undoubtedly the oldest form of communication to evolve on Earth (Bradbury and Vehrencamp, 1998). Even our earliest prokaryotic ancestors would have used these stimuli as cues to navigate and survive in their chemical world. It is thought that many of these stimuli were then exapted (co-opted) for use as inter- and intraspecific communication systems (Wyatt, 2014a). After more than three billion years of evolution, chemical signaling systems are found in literally every branch on the tree of life and are involved in a great diversity of biological functions.

In general, chemicals that act in any capacity as signaling molecules are referred to as semiochemicals (sensu Law and Regnier, 1971). By definition this includes both inter- and intraspecific communication. Within this category, semiochemicals that expressly function in intraspecific communication are called pheromones. Karlson and Lüscher (1959) originally defined pheromones as "substances secreted to the outside by an individual of the same species, in which they release a specific reaction [in another individual of that species]," and it is to this definition I will adhere throughout. Much as hormones are chemicals that communicate between cells of the body, pheromones are chemicals that communicate between individuals of the same species. Indeed, this analogy led to their original definition as ectohormones, the term Karlson and Lüscher replaced with pheromones for greater clarity.

Despite many years of research, there is still a great deal to learn about the production, reception, function, and evolution of chemical signals. This is largely due to the many challenges associated with their study (Wyatt, 2014a). Many chemical signals exist as complex mixtures with large numbers of compounds. Not all of these necessarily act as signals, but identifying a semiochemical from the background noise is difficult.

Many chemical signals are also produced in very small quantities and are difficult to synthesize in the lab, further complicating experimental study. Even having identified a putative compound, bioassays are needed to determine whether said molecule is necessary for the responses observed. Furthermore, individual components are often not as effective as the whole mixture, so demonstrating that a given molecule is sufficient for a given response is even more dubious.

Yet, diligent studies over the years have provided many insights into the roles and mechanisms of chemical signals in several prominent systems. In the realm of animal behavior, particularly influential work has been performed in insects (e.g. moths, fruit flies, hymenopterans), mammals (e.g. rodents), reptiles (e.g. snakes), and amphibians (among others). These studies have revealed that semiochemicals are involved in a wide variety of behavioral functions in diverse animal taxa. For example, chemical signals are used for: aggregation, inter- and intraspecific spacing through scent marking, alarm signaling, homing/orientation, myriad social behaviors via individual, group, and/or kin recognition, and reproduction via roles in mate attraction, mate assessment, modifying a receiver's reproductive state, facilitating courtship, and even affecting sperm competition. The scope of these many functions and the diversity of taxa relying heavily on chemical signals underscore their importance in many facets of animal life. For more information about chemical signals and animal behavior, including detailed examples for each of the behaviors mentioned above, the reader is referred to the invaluable textbook on the subject by Wyatt (2014a).

## Variation and Specificity in Chemical Signals

As indicated earlier, the hypothesis employed throughout this dissertation is that variation in the composition of a signal should (in many cases) affect the information encoded by the signal, as inferred by its corresponding effect on receivers. In other words, in order to convey variable information, the composition of the signal itself needs to differ in a reliable manner according to the trait in question. The capability of varied signals to convey different information is referred to as the specificity of the signal.

There are several ways in which the composition of a multicomponent chemical signal can vary. First, two signals may differ in terms of the presence/absence of individual components. Secondly, even if the same components are present, they may exist in different structural variations. Depending on the nature of the chemical involved, there are often several ways in which the structure might vary. For example, organic molecules can exist as different enantiomers, and protein sequences can differ in the sequence of their amino acids. Third, the signals may also differ in the absolute and relative levels of expression of individual components.

Several illustrative examples of signal specificity resulting from each of these separate mechanisms are present in the literature. For example, females of the red-sided garter snake, Thamnophis sirtalis, produce several methyl ketones from the skin that are not present in males. In this case, the inclusion of additional components allows for a sexspecific attractant pheromone (Mason et al., 1989). An example of specificity from structural variation comes from newts of the genus Cynops, in which males produce a decapeptide mate attractant named Sodefrin. The peptides from two congeneric species differ by only two amino acids, but even this subtle structural difference allows for a reliable species-specific signal (Kikuyama and Toyoda, 1999). The relative ratios of semiochemicals (also called the blend of components in some systems) can likewise be important for signal specificity. For example, unique blends of the same pheromone components are capable of conveying population- and species-specific signals in some moths (Linn and Roelofs, 1989; Wanner et al., 2010). These mechanisms of signal variability are not mutually-exclusive; on the contrary, they often work in concert. In the context of complex multicomponent mixtures, this variation affords a tremendous potential for specificity in the information encoded by chemical signals.

## Species Background: Plethodontidae

## General Description and Phylogenetics

The family Plethodontidae is a widespread and speciose group of salamanders (Order Caudata), comprising $\sim 66 \%$ of all caudate species under current taxonomic nomenclature (AmphibiaWeb, 2016). These species occupy a wide variety of habitats;
they are found throughout North and South America, in addition to a few isolated species in parts of Europe and Asia (Shen et al., 2016). Previous hypotheses posited an "Out of Appalachia" biogeographic history, in which species were thought to have radiated from southeastern North America in the Cretaceous after losing their lungs as a streamdwelling adaptation (Ruben and Boucot, 1989). Recent advances, however, indicate that a radiation out of western North America in the early Paleocene epoch is more likely (Vieites et al., 2007; Shen et al., 2016). This radiation is considered non-adaptive, but nevertheless occurred within a relatively short time frame. Estimates of lineage accumulation place this event as comparable to some of the most rapid adaptive radiations known to date (Kozak et al., 2006). Even these ancestral species probably relied heavily on chemical communication, and indeed some pheromone components still present in extant plethodontids are plesiomorphic (Van Bocxlaer et al., 2015).

The inference of phylogenetic relationships resulting from this rapid radiation is unsurprisingly challenging, but efforts using larger data sets and new methods have made significant progress (Pyron and Wiens, 2011; Shen et al., 2016). Current hypotheses place the Plethodontidae as the sister family of the Amphiumidae (Figure 1.1), both of which diverged relatively late compared to most of the other families. Within the Plethodontidae are several major lineages (Figure 1.2), although the taxonomic levels used for these clades are not always consistent. This research will primarily investigate two species: the red-legged salamander ( $P$. shermani; of the $P$. glutinosus group) and the eastern redbacked salamander ( $P$. cinereus). This is due to the abundance of prior literature on the behavior and biochemistry of chemical signaling in these species. Plethodon cinereus in particular has been the cornerstone of behavioral research concerning the role of chemical signals in facilitating many aspects of plethodontid social behavior (Jaeger, 1986). In contrast, P. shermani has been the most extensively-studied system in terms of the biochemistry of pheromone communication (Woodley, 2010). As such, a complementary and comparative approach to studying this clade is necessary in order to fully understand the relationship between biochemistry and behavior.

## Range, Ecology, and Territoriality

As noted, plethodontids inhabit all of North America, but this work will focus mostly on two eastern species. These focal species have different geographical distributions in terms of their location and size. Plethodon cinereus has a large range, found throughout the entirety of the northeastern United States and southeast Canada. In contrast, $P$. shermani was formerly a population of $P$. jordani, described by Highton and Peabody (2000) as the lineage with a range limited to specific mountains in western North Carolina. This distinction, based upon geographic protein (allozyme) variation, illustrates a trend in Plethodon systematics in which species designations are troublesome. Most species with large ranges exist in reality as distinct, isolated populations with limited dispersal. Some researchers have a tendency to divide these groups into separate species, and indeed this same phenomenon has also occurred for certain populations of $P$. cinereus (Highton, 1999).

Plethodontids as a whole occupy a great diversity of habitats, including streamdwelling, fossorial, and even arboreal species (Wake, 1966). While other families rely on water for reproduction and have an aquatic larval stage, many plethodontids are highly terrestrial. Most species spend a significant amount of their life on land, and several have lost the requirement of aquatic habitats altogether. Both of the species in question inhabit the leaf litter on the forest floor of eastern deciduous forests, establishing territories beneath cover objects such as rocks or downed logs (Petranka, 1998). While territoriality has been observed in $P$. jordani/shermani (Selby et al., 1996), much of what we know about plethodontid social and territorial behavior comes from studies of $P$. cinereus (Jaeger, 1986; Jaeger and Forester, 1993). In this species, the size of territories have been estimated as $0.16-0.33 \mathrm{~m}^{2}$ (Mathis, 1991a), and extend at least 0.3 m underground (Taub, 1961). Territories are defined as the space aggressively defended (Jaeger, 1984) against con- and heterospecifics as a means to protect limited access to suitable foraging areas (Jaeger, 1971; 1972; 1974).

Individuals of both species are often highly abundant where present. Population density estimates for $P$. cinereus range from 0.3 salamanders $/ \mathrm{m}^{2}$ in a northern population (New Hampshire; Burton and Likens, 1975) to an order of magnitude higher (2.8
salamanders $/ \mathrm{m}^{2}$ ) in a southern population (Virginia; Mathis, 1991a). While comparable density estimates do not exist specifically for $P$. shermani to my knowledge, on wet evenings it was common to find hundreds of individuals (in 10 person-hours) on the surface at the study site for this research (Eddy, 2012; pers. obs.). This high population density, coupled with fluctuating prey availability and poor sub-surface feeding opportunities (Jaeger, 1972; 1978; 1980; Fraser, 1976), undoubtedly contribute to the high degree of territoriality in these species. Rainfall is a major limiting factor for access to suitable prey; thus, bouts of surface activity during wet weather are precious foraging opportunities (Jaeger, 1972; 1978; 1980). As these periods also provide the best opportunity to find, assess, and reproduce with potential mates, there is an inherent conflict between the need to forage and the chance to mate.

## Reproduction and Courtship

In many plethodontids, males mate annually while females mate biennially (Petranka, 1998; Marvin, 1996). In particular, this is true for most populations of $P$. cinereus (Sayler, 1966), and for P. shermani (Hairston, 1983). In cases where females mate every other year, this results in a skewed operational sex ratio of 2:1 (males to gravid females; Thomas et al., 1989). Reproduction is thought to be limited mainly by access to sufficient nutritional resources for egg development and defense (Sayler, 1966; Ng and Wilbur, 1995), and female reproductive fitness is dependent upon body size (Tilley, 1968; Peacock and Nussbaum, 1973; Nagel, 1977; Lotter, 1978; Fraser, 1980; Verrell, 1995). The timing of breeding varies between species, and is restricted to one or more specific mating seasons per year (Petranka, 1998).

Courtship in plethodontids consists of several stages surrounding a ritualized behavior called the "tail-straddling walk" (TSW; Arnold, 1976; 1977). While the form and timing of behaviors differ slightly between species, overall courtships within the major lineages (see below) are highly conserved (Arnold 1976; 1977; Dyal, 2006; Eddy, 2012). A generalized example of courtship is as follows. The first stages involve male orientation toward the female, after which he performs tactile behaviors such as nudging the female and tapping his snout to her body (presumably receiving olfactory cues).

Visual signals in the form of "foot dancing" are also performed at this stage in some species. This stage is considered important for "persuasion" of the female to continue courtship and progress to subsequent stages (Eddy et al., 2012). The next phase (TSW) commences once the male nudges his way under the chin of the female, and walks forward until her chin is at the dorsal base of the male's tail. At this point, the male periodically walks forward while the female follows behind, with her front legs straddling the male's tail (hence the name). Throughout this process the male undulates his tail in a circular manner, while keeping its base arched in the air.

During TSW (and sometimes surrounding it), the male applies proteinaceous pheromone secretions from the mental gland (MG), a male-specific submandibular gland (on the underside of the chin). The sole effect of MG secretions documented to date is a reduction in courtship duration, an effect interpreted as increasing female receptivity (Rollmann et al., 1999; Houck, 2009). The way in which the male applies the pheromone is one of the major differences in courtship between Plethodon lineages. There are two major behaviors for administering pheromone to females: dermal and olfactory delivery (Figure 1.2). The ancestral method is dermal delivery. Males of these species (including P. cinereus) possess modified, protruding premaxillary teeth (PPT) which they use to scratch the female's dorsum; they then rub their MG over the wound, presumably administering the pheromone directly to the bloodstream (Organ, 1961). As such, this method is sometimes referred to as the "scratching" or "vaccination" mode of delivery (Palmer et al., 2005). The second method of pheromone delivery is through the olfactory pathway, a derived trait that is characteristic of the $P$. glutinosus complex (including $P$. shermani). Males of these species have lost their PPT, but most have large, distended mental glands situated posteriorly relative to the ancestral state (Palmer et al., 2007a; Sever, 1976). During TSW, males in this group periodically turn around and apply MG secretions directly to the nares of the female, a behavior termed "slapping" (Arnold, 1976). The transition between these clades is referred to as the intermediate group, although the phylogenetic relationships of this transition and the exact method of pheromone delivery remain poorly understood. These species also lack PPT and have enlarged posterior MGs, but the pheromone delivery mode is less clear; a degree of both
dermal and olfactory delivery has been implicated by previous observations (Palmer et al., 2007a).

After TSW, males remain in one place and commence the deposition of a spermatophore, consisting of a gelatinous base with a sperm cap (Organ and Lowenthal, 1963). Once the spermatophore is placed on the substrate, the male again walks forward and the female follows, traversing over top of the spermatophore. Once the spermatophore has reached her cloaca, if she is appropriately receptive, she picks up the spermatophore and courtship will have finished. There is substantial interspecific variation in the length of the individual stages of courtship and the overall process (measured only in laboratory settings). In $P$. shermani the entire sequence takes more than 90 minutes on average, while in $P$. cinereus it lasts for more than 60 minutes. Even within a species, there is a great deal of variation in the timing of courtship between individual male-female pairs (Eddy, 2012).

After mating, females store sperm for prolonged periods in a specialized organ (the spermatheca) to be used for internal fertilization during oviposition (Sayler, 1966; Sever et al., 1978; Eddy et al., 2015). For example, in P. shermani viable sperm can be stored for well over nine months, suggesting that sperm competition is probably a powerful sexual selective pressure (Eddy et al., 2015). After oviposition, females provide parental care by guarding their clutch of eggs (brooding), aggressively protecting them from predation and cannibalism, and aerating them to avoid desiccation and inhibit the growth of fungal pathogens (Jaeger and Forester, 1993). Unguarded clutches have very low survivability, but females do not actively forage throughout the duration of brooding. In $P$. cinereus, this brooding period can last up to twelve weeks, during which females deplete their fat reserves (Ng and Wilbur, 1995). This situation again demonstrates the importance of sufficient energy stores for female fitness. Some plethodontids have a larval stage while others undergo direct development (Mueller et al., 2004). After hatching, the young often stay within the parent's territory, sometimes for long periods of time, before eventually dispersing (Jaeger and Forester, 1993).

## The Role and Composition of Chemical Signals

Semiochemicals are used for a variety of both inter- and intraspecific behavioral functions in plethodontids. Between species, they are used for: prey-detection (Dickens, 1999; Placyk and Graves, 2002), predator avoidance (Maerz et al., 2001; Sullivan et al., 2005), allocation of spatial resources (Thurow, 1975; Jaeger, 1986; Quinn and Graves, 1999), and species recognition (Jaeger and Gergits, 1979; Dawley, 1986ab). Within species, these molecules are used for: scent marking and territoriality (Jaeger, 1986; Jaeger and Forester, 1993), homing and orientation (Madison, 1969; Madison and Shoop, 1970; Kleeberger and Werner, 1982), sex recognition (Jaeger and Gergits, 1979; Dawley, 1984a; Dantzer and Jaeger, 2007ab), mate assessment (Walls et al., 1989; Mathis, 1991b; Marco et al., 1998; Chouinard, 2012), and facilitating courtship (Rollmann et al., 1999). Controlled behavioral experiments have used bioassays to determine that receivers can infer a great deal of information from these scents (Table 1.1).

The site of production for the chemical signals involved in each of these behavioral phenomena has been a longstanding question (Jaeger, 1986). Several regions of the skin have specialized glands that differ in anatomical location, morphology, and chemical composition (Simons and Felgenhauer, 1992; Simons et al., 1994; 1997; 1999; Hecker et al., 2003; Sever, 2003). Among these regions of chemical signaling importance are the aforementioned mental gland (MG), cloacal glands (CG), the postcloacal gland (PCG; ventral surface of the tail posterior to the cloaca), and the dorsal tail base (DTB; top of the tail above the cloaca). (These glands and their behavioral roles are discussed in detail in Chapter 4.) In general, these tissues produce a mixture of mostly proteins and carbohydrates, but histochemical analyses demonstrate different chemical compositions between glands (Hecker et al., 2003; Sever, 2003). To date, specific semiochemicals have only been documented in the mental gland. In this case, several protein pheromones involved in courtship have been identified (Figure 1.3).

The first plethodontid pheromone identified (from P. shermani) is Plethodontid Receptivity Factor (PRF), a 22 kDa protein related to interleukin-6-type (IL-6) cytokines (Rollmann et al., 1999). It has four alpha-helices that alternate in a characteristic IL-6 fold. Gene duplications have led to individual males expressing several isoforms, and
these sequences demonstrate a signature of strong positive selection. Sequencing efforts in other Plethodon spp. have revealed numerous additional isoforms and substantial intraand interspecific variation, but this protein has only been documented in eastern Plethodon (Watts et al., 2004; Palmer et al., 2005; 2007a). Recent observations have confirmed that PRF in P. cinereus is glycosylated, but is not glycosylated in P. shermani (Wilburn et al., 2014a). This finding adds yet another level of structural variability for this MG component, and suggests differences in PRF function among the major Plethodon lineages.

The second major component of MG secretions, Plethodontid Modulating Factor (PMF), is a 7 kDa member of the three-finger protein (TFP) superfamily (Palmer et al., 2007b). PMF is present throughout the Plethodontidae, and exhibits even greater interand intraspecific variation than PRF. Individual males produce dozens of PMF isoforms; it has been described as a "hypervariable" pheromone component, and its sequences also indicate strong positive selection (Palmer et al., 2007b; 2010; Wilburn et al., 2012). Typical TFPs share a distinct fold consisting of sets of antiparallel beta-sheets that form the eponymous three "fingers" of the structure; they also possess a characteristic disulfide bonding pattern and a net positive charge. In contrast, PMF has a net negative charge, and recent structural analyses revealed a novel topology and disulfide bonding pattern (Wilburn et al., 2014b).

The final MG pheromone documented prior to this research is highly expressed in Desmognathus, a genus that diverged from Plethodon between 35 and 110 mya (with estimates varying between studies; Figure 1.2). This protein, Sodefrin Precursor-like Factor (SPF), is related to the larger protein from which the decapeptide mate attractant Sodefrin is cleaved in Cynops newts (Kikuyama and Toyoda, 1999). Subsequent research has shown that this pheromone is an ancestral component of cloacal glands, dating back to the late Paleozoic (>300 mya; Janssenswillen et al., 2014; Van Bocxlaer et al., 2015; Maex et al., 2016). Like PMF, it is also thought to have originally evolved from TFPs; unlike PMF, however, it contains two tandem TFP domains and different disulfide bonding patterns (Janssenswillen et al., 2014; Doty et al., 2015). In plethodontids SPF is a 20 kDa protein, also present in numerous isoforms within a given male. Palmer et al.
(2007a) hypothesized that SPF is undergoing evolutionary replacement by PRF because it is not present as protein in substantive levels in the MG of the Plethodon spp. investigated to date. Several isoforms are glycosylated in salamandrids (Van Bocxlaer et al., 2015), which may also be true in plethodontids (Doty et al., 2015).

Application of the whole MG secretion to females reduces the duration of courtship (Houck et al., 1998; Houck and Reagan, 1990). This effect has also been demonstrated for PRF (Rollmann et al., 1999), PMF (Wilburn et al., 2015), and SPF alone (Houck et al., 2008a). In the case of PMF, Houck et al. (2007) initially reported that when applied singly it actually increased the duration of courtship. This has since been attributed to the use of an incomplete fraction of isoforms in this initial experiment, as a more comprehensive PMF mixture elicits the expected reduction in courtship duration (Wilburn et al., 2015). Thus, signal complexity appears to be important for appropriate function.

## Research Focus and Chapter Outline

Despite these intriguing advances in the behavioral roles and biochemical composition of plethodontid semiochemicals, many questions remain and large aspects of this communication system have gone unstudied. For example, chemical signals facilitate a wide array of behavioral phenomena, but essentially all of the research investigating the identity of gland secretions has been restricted to the mental gland and its "courtship pheromones." Several other specialized glands exist; these have been implicated in different behaviors, and possess distinct anatomies and chemical compositions. Various experimental designs have demonstrated that different aspects of information can be inferred from volatile and/or non-volatile components (Table 1.1), again suggesting a wide variety of chemicals used in specific aspects of communication. Thus, more biochemical research is needed to document the full extent of proteins, carbohydrates, and other chemicals produced by these many glandular regions. Complementary behavioral and physiological studies are also needed to elucidate the specific functions of these chemicals.

Even in regard to the courtship pheromones receiving the most attention to date, we have only begun to appreciate the true complexity of these multicomponent signals. Previous research has found that MG pheromone composition differs between the major lineages of the Plethodontidae in terms of the types of proteins present (Palmer et al., 2007a; Figure 1.2). Each protein component is also produced in numerous sequence variants, which differ in primary protein structure both within and between lineages (Watts et al., 2004; Palmer et al., 2010). Because the overall MG secretion consists of multiple isoforms, the relative ratios of these components may likewise serve as an emergent aspect of signal complexity. For example, different populations are known to possess different MG pheromone profiles (Rollmann et al., 2000). Unfortunately, little is known about what this signal complexity means to female receivers, how it affects their neurobiology and reproductive physiology, or why and how it has evolved. For now, the focus of this research will be to explore further aspects of signal complexity within an individual, within a population, and between species. In doing so, several new and existing hypotheses are discussed concerning the function and evolution of these signals.

Chapter Two: Individual variation in pheromone isoform ratios of the red-legged salamander, Plethodon shermani. This research will examine the degree of individual variation in MG pheromone composition, with specific reference to variation in PRF profiles between individuals. Previous research has documented MG pheromone profiles at the level of entire populations (as pooled samples), but this work will investigate these profiles on a finer scale. In this way, these analyses explore additional signal complexity within and between individual males of a single population.

Chapter Three: Inter- and intraspecific variation in pheromone glycosylation in plethodontid salamanders. Prior biochemical analyses have also revealed the presence of post-translational modifications of PRF in the form of glycosylation. These modifications were documented in PRF from P. cinereus, a member of the ancestral clade (dermal pheromone delivery), but are not present in $P$. shermani, a member of the derived clade (olfactory pheromone delivery). This work will use bioinformatic
approaches to investigate the extent that these observations apply to the rest of the Plethodon phylogeny. In addition, the functional implications of glycosylation are investigated through evolutionary analyses and protein structural modeling. This research examines the variation in glycosylation potential within and between lineages, and addresses a previously overlooked aspect of signal complexity within populations and between species.

Chapter Four: De novo transcriptome assembly of three chemical signaling glands in a plethodontid salamander. This final chapter provides the first in-depth analysis of the genes expressed by three chemical signaling glands: the cloacal glands, postcloacal glands, and the dorsal tail base. High-throughput RNA sequencing technology is used to construct de novo transcriptome assemblies. The identities of many putative secretory proteins that may be relevant for chemical communication are documented, as are differences in their expression between glands. Thus, this research describes additional aspects of chemical signal complexity within an individual, by documenting the composition of several uncharted signaling glands.

Table 1.1. Information encoded by plethodontid semiochemicals. Listed are the types of information inferred from chemical stimuli, the species in which the inference is reported (i.e. the receiver), the source of the odors, whether this source allowed access to only volatile odors ("Volatile") or non-volatile as well ("Both"), and the reference. The experimental scent sources typically consist of substrates marked by the sender ("Substrate"), including (active) scent marks and other (passive) skin and urogenital secretions. Other possible stimuli consisted of direct access to fecal pellets ("Feces"), swabs or rinses of skin and/or specific glands ("Glands"), or other exposures to the scent of the "Whole Body" (e.g. through rinses or airborne currents).

| Information | Species | Scent Sources | Volatility | Reference |
| :---: | :---: | :---: | :---: | :---: |
| Species | P. cinereus <br> P. shenandoah | Substrate Feces | Both | Jaeger and Gergits, 1979 |
|  | P. glutinosus <br> P. aureolus | Substrate Whole Body | Both Volatile | Dawley, 1984a |
|  | P. glutinosus <br> P. jordani <br> P. aureolus <br> P. kentucki | Whole Body | Volatile | Dawley, 1986b |
|  | D. quadramaculatus <br> D. monticola | Substrate | Both | Roudebush and Taylor, 1987 |
|  | D. ochrophaeus <br> D. imitator | Substrate | Both | Verrell, 1989 |
|  | P. vehiculum <br> P. dunni <br> P. vandykei | Feces | Both | Ovaska and Davis, 1992 |
|  | D. ochrophaeus <br> D. fuscus | Substrate | Both | Uzendoski and Verrell, 1993 |
|  | D. ocoee <br> D. orestes <br> D. carolinensis | Substrate | Both | Verrell, 2003 |
|  | P. montanus | Substrate | Both | Palmer and Houck, 2005 |
| Sex | P. cinereus | Substrate Feces | Both | Jaeger and Gergits, 1979 |
|  | P. jordani <br> P. glutinosus | Whole Body | Volatile | Dawley, 1984a |
|  | P. vehiculum <br> P. dunni | Feces | Both | Ovaska and Davis, 1992 |
|  | P. cinereus | Substrate | Both | Karuzas et al., 2004 |
|  | P. cinereus | Glands (MG, PCG) | Both | Page and Jaeger, 2004 |
|  | P. shermani <br> P. teyahalee <br> P. montanus | Substrate | Both | Palmer and Houck, 2005 |
|  | P. cinereus | Substrate | Volatile | Dantzer and Jaeger, 2007a |
|  | P. angusticlavius | Substrate | Both | Dalton and Mathis, 2014 |
| Size | D. quadramaculatus <br> D. monticola | Substrate | Both | Roudebush and Taylor, 1987 |
|  | P. cinereus | Substrate | Both | Mathis, 1990 |
|  | P. cinereus | Glands (PCG) | Both | Mathis and Simons, 1994 |
|  | P. vehiculum <br> P. dunni | Substrate | Both | Marco et al., 1998 |
| Individual Identity | P. jordani | Whole Body | Volatile | Madison, 1975 |
|  | P. cinereus | Substrate | Both | Tristram, 1977 |
|  | P. cinereus | Substrate | Both | McGavin, 1978 |
|  | P. cinereus | Substrate | Both | Jaeger, 1981 |
|  | P. glutinosus | Substrate | Both | Dawley, 1984a |
|  | P. cinereus | Feces <br> Glands (CG) | Both | Simon and Madison, 1984 |
|  | P. cinereus | Feces | Both | Jaeger et al., 1986 |
|  | $P$. ouachitae <br> P. caddoensis | Substrate | Both | Anthony, 1993 |
|  | P. cinereus | Substrate | Both | Kohn and Jaeger, 2009 |
| Reproductive <br> Status (Female) | P. vehiculum <br> P. dunni | Substrate | Both | Marco et al., 1998 |
|  | P. cinereus | Substrate | Volatile | Dantzer and Jaeger, 2007b |
| Parasite Load | P. angusticlavius | Feces | Both | Maksimowich and Mathis, 2001 |
|  | P. angusticlavius | Substrate | Both | Dalton and Mathis, 2014 |
| Diet Quality | P. cinereus | Feces | Both | Walls et al., 1989 |
|  | P. cinereus | Whole Body | Both | Chouinard, 2012 |
| Injury | P. cinereus | Whole Body | Both | Sullivan et al., 2003 |
|  | P. cinereus | Substrate | Both | Wise et al., 2004 |



Figure 1.1. Family-level phylogeny of salamanders (Order Caudata). A time-scaled phylogeny adapted from Zhang and Wake (2009), inferred from complete mitochondrial genomes. According to these estimates, the Plethodontidae diverged from its closest relatives in the Lower Cretaceous. Other analyses differ in estimated divergence times, but this specific phylogeny was chosen because it serves as a fair representative of this variation.


Figure 1.2. Time-scaled phylogeny of the family Plethodontidae (Order Caudata) with specific reference to courtship morphology, behavior, and pheromone traits. Adapted from Pyron and Wiens (2011), with major early-diverging plethodontid clades collapsed. Other analyses differ in estimated divergence times; this particular phylogeny was chosen due to the rigor of the study and because it serves as a good representative of divergence estimates. Within the Plethodontinae (subfamily), lineages are collapsed to the genus-level, with the exception of Plethodon spp. The eastern Plethodon are depicted by its major lineages: the dermal pheromone delivery clade ( $P$. cinereus group), the intermediate group ( $P$. websteri, $P$. wehrlei and $P$. welleri groups), and the olfactory delivery clade ( $P$. glutinosus group). The intermediate and olfactory groups have been collapsed to a polytomy, to account for incongruence between phylogenies. The origin (filled) and loss (open) of traits are mapped onto their appropriate branches, separated by pheromone proteins (boxes), along with morphological and behavioral traits (circles). The placement of traits is by branch only; the time-scale for them is arbitrary. The pheromone traits mapped are: Sodefrin Precursor-like Factor (SPF), Plethodontid Modulating Factor (PMF), and Plethodontid Receptivity Factor (PRF). The morphological and behavioral courtship behaviors include: the Mental Gland (MG), Protruding Premaxillary Teeth (PPT), Dermal Pheromone Delivery (DD), and Olfactory Pheromone Delivery (OD). In the case of OD, it is mapped onto the P. glutinosus group and meant to represent stereotypical olfactory delivery behavior. Variations of some olfactory behavior are also observed in the intermediate group.


Figure 1.3. The three major proteinaceous courtship pheromones in plethodontids. Size-scaled protein models of the three major courtship pheromones: PRF-B (sensu Chouinard et al., 2013) from P. shermani, PMF-G from P. shermani (sensu Wilburn et al., 2012), and SPF-1 (sensu Doty et al., 2015) from D. ocoee. Models for PRF and SPF were built with I-TASSER (methods in Chapter 3); the PMF shown is the top model from a solved NMR structure (Wilburn et al., 2014b). Representative structures were selected as the isoforms that were typically the most abundant in their corresponding species.

## CHAPTER TWO

# INDIVIDUAL VARIATION IN PHEROMONE ISOFORM RATIOS OF THE RED-LEGGED SALAMANDER, Plethodon shermani. 

Chouinard, A.J. ${ }^{*}$, Wilburn, D.B. ${ }^{2}$, Houck, L.D. ${ }^{1}$, and Feldhoff, R.C. ${ }^{2}$

1. Department of Integrative Biology, Oregon State University, Corvallis, OR 97331
2. Department of Biochemistry, University of Louisville, Louisville, KY 40202

Published in Chemical Signals in Vertebrates 12 (pp. 99-115). Springer, New York.


#### Abstract

For more than 15 years, pheromone signaling in plethodontid salamanders has served as an important amphibian system for studying how molecules can influence reproductive behavior. In the red-legged salamander (Plethodon shermani), males utilize proteinaceous pheromones from a submandibular (mental) gland to affect female behavior and reduce the duration of courtship. The two major proteins, Plethodontid Receptivity Factor (PRF) and Plethodontid Modulating Factor (PMF), exist as multiple isoforms within the gland of an individual male. While previous research has focused on biochemical characterization of these proteins, this study explores the degree of intraspecific variability of pheromone isoform ratios among males of a single population. Biochemical analyses were performed on the mental gland extracts of individual male salamanders ( $\mathrm{n}=108$ ) via high-performance liquid chromatography to quantify components of the pheromone mixture and establish the extent of individual variability. The results of these individual analyses revealed that the two main proteins (PMF and PRF) comprise over $80 \%$ of the total mixture, with a 5:3 ratio (PMF:PRF). Numerous minor peaks also appeared in the elution range of PRF, which could account for an excess of mRNA transcripts that have not been proteomically characterized. Lastly, these data demonstrate a remarkable diversity of isoform ratios among males. Taken together, these aspects of pheromone complexity and diversity could have profound functional effects on reproductive behavior in this family of salamanders.


## InTRODUCTION

Chemical signal profiles of many species are highly complex, often containing numerous components which may differ in their proportions between individuals (Wyatt 2003). A major challenge in studying these signals is overcoming the complexity in order to identify the functional role and chemical nature of each component. Generally, the first step in this process is to fractionate the mixture and evaluate these fractions through repeated bioassays. However, reducing a blend to its constituents can lead to negative assays due to a loss of essential multicomponent interactions. A complementary approach to overcoming the complexity of chemical signals is to consider the properties of the mixture as a whole, as this is what is perceived by the receiver in nature. The first step in this process is to define the composition of the whole chemical profile for the species of interest at the individual and population levels.

The aim of this research was to quantify the complexity and intraspecific diversity in signal composition of the male-specific courtship gland of the red-legged salamander, Plethodon shermani. This salamander is known for producing one of the first characterized vertebrate pheromones (Rollmann et al. 1999). Males of this species possess a submandibular secretory gland (the mental gland) that hypertrophies each year during the breeding season. The mental gland is used during courtship, which consists of a "tail-straddling walk" (TSW; Arnold 1976) where the male slowly walks forward and the female follows him with her front legs straddling his undulating tail. During TSW, the male will periodically turn around and quickly snap its mentum downward, contacting the mental gland to the snout of the female and delivering proteinaceous pheromones to the female's nares. These proteins ultimately arrive in the vomeronasal organ (VNO; see Dawley and Bass 1988, 1989), where they activate the sensory neurons of the vomeronasal epithelium (Wirsig-Wiechmann et al. 2002, 2006). These VNO receptor neurons send electrical signals to the accessory olfactory bulb and ultimately to several regions in the brain known to be involved in reproductive behavior (Schmidt and Roth 1990; Laberge et al. 2008).

The secretion of the mental gland consists of multiple proteins, most of which exist in a variety of structural isoforms. Two proteins, Plethodontid Receptivity Factor
(PRF) and Plethodontid Modulating Factor (PMF), have been well characterized at the biochemical level, and comprise the majority of the glandular secretion (Feldhoff et al. 1999). PRF is a 22 kDa protein with sequence similarity to the IL-6 family of cytokines (Rollmann et al. 1999; Watts et al. 2004). In most males, PRF exists in a unique combination of three major isoforms (B, C1 and C2). The other major component (PMF) is a 7 kDa protein with sequence similarity to the three-finger protein superfamily (Palmer et al. 2007b). PMF persists as a large multigene complex with at least 99 mRNA haplotypes translated into at least 28 protein isoforms (Wilburn et al., 2012).

While the number of variants and degree of structural diversity between isoforms differ between protein types, PRF and PMF share two key features: (1) both protein types exist in multiple isoforms which appear to have arisen due to strong sexual selection (Watts et al. 2004; Palmer et al. 2007ab, 2010), and (2) models of molecular evolution suggest that both PRF and PMF are some of the fastest evolving proteins (compared to examples from Swanson and Vacquier 2002). A third protein (C3) is also present, though less is known about this protein than its counterparts. Other than these three abundant proteins, the other components all consistently appear in very small quantities.

The proteinaceous secretions of the mental gland modulate the timing of courtship (Rollmann et al. 1999; Houck et al. 2007). In particular, PRF decreases the duration of the courtship ritual. Reduction in courtship duration may not be the sole function of this multicomponent signal, but may be a parallel effect of other underlying physiological roles. By studying the biochemical properties of the components, coupled with their neurophysiological modes of action, we expect to reveal the broader scope of functions that are facilitated by this variable multicomponent signal.

For practical reasons, most previous biochemical and behavioral analyses of these protein secretions have utilized pooled samples obtained from many individuals. This process permitted the use of standardized preparations for purification, biochemical characterization, and use in behavioral assays. In contrast, the aim of this study was to investigate differences in pheromone composition between individual male salamanders of a single population. Because these proteins are so intimately involved in the
reproductive behavior of this species, variation among male pheromone profiles may have profound fitness effects as a result of the courtship interactions that they mediate.

## Methods

## Animal Collection and Protein Extraction

Male salamanders (n = 108) were collected from Macon County, NC ( $35^{\circ} 10^{\prime} 38^{\prime \prime} \mathrm{N}$, 83³3'47"W) in August over the course of three years: $2007(\mathrm{n}=22), 2008(\mathrm{n}=25)$ and $2010(\mathrm{n}=61)$. A preliminary analysis of the 2007 data (Houck, 2009) indicated that a larger sample size was required for the current study. Animals were housed in the laboratory in clear plastic containers ( $29 \times 14 \times 9 \mathrm{~cm}$ ) at $16^{\circ} \mathrm{C}$ on a $14: 10$ (light:dark) light cycle. Mental glands (MG) were surgically removed from each male for biochemical analysis. Males were anaesthetized in 7\% diethyl ether for 10 minutes, and once fully anaesthetized, the MG was surgically removed from the underside of the chin with iridectomy scissors. Each gland was independently rinsed in Amphibian Ringer’s solution to remove any residual mucus and then placed in $200 \mu \mathrm{~L}$ of 0.8 mM acetylcholine chloride in Amphibian Ringer's solution for extraction of protein from the tissue. After 30 minutes, glands in acetylcholine solution were centrifuged at $>10,000 \mathrm{xg}$ for 10 minutes. The supernatant was then subjected to a second centrifugation before storage at $-80^{\circ} \mathrm{C}$. The MG for each male was extracted and processed independently to allow for biochemical analysis of individual-specific profiles.

## Biochemical Analysis

Each sample was thawed and the protein concentration was determined by a bicinchoninic acid (BCA) assay (Smith et al. 1985). Samples were then analyzed by reverse-phase high-performance liquid chromatography (RP-HPLC). For each male, 25 $\mu \mathrm{g}$ of MG protein was loaded on a C18 column ( $5 \mu \mathrm{~m}$; $4.6 \times 150 \mathrm{~mm}$; Vydac, Hesperia, CA) equilibrated with $0.1 \%$ trifluoroacetic acid (TFA). Elution was accomplished on a Waters 2695 Alliance HPLC system (Waters Corp., Milford, MA) using a linear gradient ( $0 \%-100 \%$ ) of $70 \%$ acetonitrile/ $0.08 \%$ TFA at $1 \mathrm{~mL} /$ minute for 70 minutes. Absorbance was measured at a wavelength of 220 nm .

## Data Analysis

Previous analyses demonstrated that the protein secretions of the MG elute between $20-60$ minutes (Feldhoff et al. 1999). The retention times for each component were consistent between runs and verified by the use of standards. Waters Empower Software was used to integrate and quantify peak areas. The absorbance values for each peak were converted into the percent per peak of the total protein content of the male MG extract. Each run was standardized to $25 \mu \mathrm{~g}$, to allow for comparison of the relative protein composition of each male MG extract. For each male profile the proportion of PMF, PRF and C3 was calculated and compared among all males to evaluate the average composition of the MG. The number of peaks within each of these components was also counted. In the case of PRF, the relative ratios of the three major components (B, C1 and C2) were calculated and compared among all males. These data were presented in a ternary plot for compositional analysis of three components ("vcd" package in R; Meyer et al. 2006). Kernel density estimates were also calculated for these data ("adehabitatHR" package in R; Calenge, 2006). A least-squares cross validation (LSCV) method was used to increase the precision in detecting clusters within the total data. The probability function was assigned to a heat color map with blue representing low probability and red demonstrating high probability. Relative composition of PMF was not quantitatively analyzed because it is not possible to establish a reliable isoform count due to (1) the large number of isoforms, and (2) similarity of PMF isoform retention times.

## Results

## Pheromone Composition

The data from four individuals were omitted from the total sample ( $\mathrm{n}=108$ ) due to poor resolution of the RP-HPLC run, resulting in an effective sample size $\mathrm{n}=104$. A representative chromatogram (Figure 2.1) reveals retention times and absorbance values of the various proteinaceous components. The diversity of PMF can be observed as the wide range of peaks that appear between $20-40$ minutes. The protein C3 eluted at $\sim 41$ minutes. All three isoforms of PRF (PRF-B, PRF-C1 and PRF-C2) eluted between 50 55 minutes. These three components of the MG secretion (PMF, PRF, C3) constituted the
major proportion of the total mixture (Figure 2.2). The combination of PMF and PRF comprised an average of $80.8 \%$ of the total profile. Of these two major components, PMFs ( $\overline{\mathrm{x}}=48.8 \%$ ) were consistently present in higher concentrations than PRFs ( $\overline{\mathrm{x}}=$ $32.0 \%$ ). Of the remaining $\sim 20 \%$ of the total composition, C3 made up an average of $10.3 \%$. This leaves only $8.9 \%$ of the total MG secretion being composed of other minor components that were unidentified at the proteomic level.

The number of peaks of PMF varied greatly among individuals, ranging from 17 39 ( $\overline{\mathrm{x}}=27.2$ ). As noted earlier, these data are likely to be an underestimate of PMF isoform numbers because peaks may contain multiple isoforms. Although there were three major components of PRF, the number of peaks in the range of PRF also varied between individuals, ranging from $3-6(\bar{x}=4.2)$. There were also occasionally other peaks in very close proximity to the protein C 3 , ranging from $1-3(\overline{\mathrm{x}}=1.6)$. It should be noted that the protein C3 is not an additional PRF isoform. The current peak nomenclature refers to the fraction from which each protein was purified during strong anion-exchange chromatography (Wilburn et al. 2012). Three of the PRF isoforms identified by Rollmann et al. (1999) correspond to PRF-B (Isoform 1), PRF-C1 (Isoform 4) and PRF-C2 (Isoform 3). The remaining isoform (2) corresponds to a SNP variant of PRF-B.

## Minor Components

Analysis of individual pheromone profiles revealed numerous minor components in the PRF range that would normally be unobservable in pooled samples, due to the abundance of the three major isoforms overshadowing these minor components. Several of these minor components consistently appeared in the PRF range of individual males (Figure 2.3). The minor proteins were classified as such because they appeared in much lower proportion than the three major isoforms (with one exception described below), but also because they were observed less frequently. While the presence of these minor components has been established through RP-HPLC, no proteomic analyses have been conducted to ascertain the identity of these proteins. In all cases, an apostrophe after the
name of a known component indicates an unknown peak eluting near that component and is not meant to imply that it is necessarily a sequence variant of that known protein.

The most common of the minor components in the range of PRF was a peak that eluted just after PRF-C2 (C2’; Figure 2.3a). This peak occurred in ~72\% of all males, although always in small quantities (ranging from $0.4-5.9 \%$ of the total PRF for males possessing this peak). Another minor component eluted just before PRF-B (B’; Figure 2.3e); this peak was the second most common of the minor peaks in the PRF range, appearing in $\sim 13 \%$ of all males. This peak was the most prominent of the minor peaks in terms of abundance (ranging from $6.5-24.4 \%$ of the total PRF for males possessing this peak). In a single case, B' was even more abundant than PRF-B (the example shown in Figure 2.3e). Other minor peaks also occasionally eluted in the range before PRF-B, as well as between PRF-B and PRF-C1 (e.g. C1’; Figure 2.3f); these peaks were less frequent than the minor isoforms described above.

## Relative Isoform Ratios

While the relative ratios of PMF isoforms could not be calculated precisely, qualitative observation revealed that the PMF profile was different in essentially every male. In the case of PRF, the lower number of isoforms allowed for excellent resolution. As such, the relative ratios of the PRF isoforms could be quantitatively analyzed. The majority of males in the sample (85/104; 81.7\%) possessed all three major isoforms of PRF. The remaining males (19/104; 18.3\%) had only two major components. In no case did a male in the sample have only one of the major isoforms. However, there were examples of each isoform missing from the profile of a male (Figure 2.3b-d). In the 19 males lacking an isoform, the most common isoform missing was PRF-C1 (14/104 males; 13.5\%). PRF-B was rarely absent (2/104 males; 1.9\%), as was PRF-C2 (3/104 males; $2.9 \%$ ). Of the males producing all three isoforms, the relative abundance of the isoforms present differed greatly between individuals. PRF-B and PRF-C2 are typically expressed in intermediate proportions, while PRF-C1 is often expressed in lower proportions (Figure 2.4).

The above values for each of the isoforms do not convey information about the ratio of isoforms within a given male. The term "pheromone profile" will be used throughout to describe the pattern of relative component ratios that were specific to an individual male signal. The PRF profile of each male was summarized in a ternary plot (Figure 2.5). This plot displays the relative ratios of each major isoform of the total PRF. Each point within the plot demonstrates the location of one male's profile within the three-variable space of the ternary plot, revealing a large degree of diversity in PRF profiles between males of the population. Almost every region of the plot was occupied by some male profiles (with the exception of the corners - see Discussion). However, while there was a great deal of diversity between males, some PRF profiles were more common in the population (i.e. a dense clustering of points). The highest density cluster (orange - red) occurred in males that possessed intermediate amounts of PRF-B (50 60\%) and PRF-C2 (40-50\%), with low amounts of PRF-C1 (0 - 10\%). There was also a rather large region surrounding this cluster that had intermediate levels of density (aqua to yellow). This region was mostly dictated by an increasing proportion of PRF-C1, typically with PRF-B making up slightly less of the total composition (dropping to 20 50\%).

## DIscussion

In these analyses we quantified several aspects of the mental gland composition of Plethodon shermani: (1) the proportional abundance of the major protein components, (2) the prevalence of minor components in the range of PRF, and (3) the intraspecific diversity of relative isoform ratios.

## Proportional Abundance of Major Components

These data allowed for quantitative analysis of the proportions of known components that constitute the total mental gland secretion. The overall pheromone mixture of this species is composed of mainly two protein types, PMF and PRF, constituting $80.8 \%$ of the total proteinaceous secretion at a ratio of 5:3. However, because PRF is roughly three times larger than PMF, the stoichiometry ratio is closer to $\sim 5: 1$. As
the actual absorbance values for individual PMF peaks are typically less than or comparable to those of PRF, the greater abundance of PMF stems from the much greater number of isoforms (e.g. $\mathrm{n}>30$ ) than are typically found for PRF (e.g. $\mathrm{n} \sim 3$ ).

Behavioral research has revealed that the protein PRF decreases the duration of courtship (Rollmann et al. 1999). Heterologous expression of PRF later showed that a single recombinant isoform was sufficient to elicit the full capacity of the behavioral response (Houck et al. 2008b). When initially tested alone for behavioral effects, PMF increased the duration of courtship (Houck et al. 2007). These contradictory results are particularly confusing when considering that the effect of the whole mental gland secretion reduced courtship time similarly to PRF alone. If the whole extract elicited the same effect observed by PRF, yet the secretion is composed of more PMF than PRF, it seems difficult to explain the opposing behavioral effect of PMF. Recent research, however, suggests that the initial findings were due to an incomplete fraction of PMF isoforms being used in the first study (Wilburn et al., 2015). Thus, the presence of numerous isoforms may be essential for inducing an appropriate female response.

The third most abundant mental gland component was the (non-PRF) protein C3. This protein constitutes an average of $10.3 \%$ of the total extract. Little is currently known about C3 other than its (1) size ( $\sim 18 \mathrm{kDa}$ ), (2) relatively negative charge, and (3) relative hydrophobicity (RT $\sim 41$ minutes). The reason for this discrepancy is that the mRNA of C3 has not appeared in any of the previous analyses of mental gland cDNA (KiemnecTyburczy et al. 2009). The lack of C3 mRNA is presumably a function of either the timing of its expression or the size of the transcript. Recent attempts at high-throughput sequencing of the mental gland transcriptome may reveal the identity of this third major component.

In summary, $91.1 \%$ of the mental gland is composed of three protein types. The remaining unknown proteins (8.9\%) seem to be accounted for by several proteins expressed in low levels, rather than a few components with high expression. KiemnecTyburczy et al. (2009) analyzed expressed-sequence tags (ESTs) and identified several transcripts that were not classified as PRF or PMF. One of these transcripts was identified as sodefrin-like precursor factor (SPF). This protein is an ancestral plethodontid
pheromone that is more abundant in the species that deliver pheromones via dermal scratching (Palmer et al., 2007a), rather than by direct contact of the male gland to the female's nares (as in P. shermani;). SPF is thought to be undergoing evolutionary replacement by PRF, as it is greatly diminished in the mental gland mRNA of the more derived clade that utilizes olfactory-delivery of pheromones, including $P$. shermani (Palmer et al. 2007a). Thus, low levels of SPF expression in P. shermani may account for some of the unidentified proteins found in the current analyses. Despite efforts to limit contamination, cytosolic and/or blood proteins may contribute to the unidentified fraction. Transcriptome sequencing and proteomic analyses will facilitate the identification of these uncharacterized proteins.

## Minor Components in the Range of PRF

The second notable result of the individual male analysis was the appearance of numerous minor peaks in the elution range of PRF. As we have yet to characterize these peaks using mass-spectrometry, they may or may not represent PRF variants. However, the major PRFs presently account for only three of the twelve mRNA transcripts that have been identified (Watts et al. 2004; Rollmann et al. 1999). Not surprisingly, these minor components varied in their presence/absence, as well as in their proportions between individuals. Most minor components were expressed at low levels in less than $15 \%$ of the males. However, C2' and B' were present in $\sim 72 \%$ and $13 \%$ of males, respectively, with B' being the most highly expressed minor component ( $6.5-24.4 \%$ of the total PRF in males possessing B'). We speculate that B' could be the SNP variant of PRF-B identified by Rollmann et al. (1999; Isoform 2).

The reason for differences in the distribution of PRF isoform between males could lie in the realm of (1) genetic, (2) transcriptional, or (3) translational variation. The results from the current study are consistent with past findings from our laboratories and suggest that PRF represents a highly duplicated multigene family. We further hypothesize that there are likely to be several gene copies of the major isoforms, with allelic variants at different loci resulting in the expression of the minor isoforms. A similar explanation
for PMF diversity has been proposed by Wilburn et al. (2012), though with many more gene duplications resulting in the large number of observed haplotypes.

In contrast, the presumed differences in male isoform expression could be the result of either transcriptional or translational regulation. If all males possess more loci for PRF than are expressed as mRNA, differences in isoform presence between males could result from differences in transcriptional regulation. In the context of the multiple mRNA transcripts (Watts et al. 2004), variation between males could also result from differential translation rates. If there are a small number of loci or alleles that constitute the PRF gene family, it is also possible that alternative splicing could create mRNA variants, resulting in different isoforms being expressed between males. However, based on preliminary gene sequencing, alternative splicing is not a likely explanation for either PRF or PMF variability (Wilburn and Feldhoff, unpublished data). Lastly, it is possible that both genetic differences and regulatory mechanisms may be working in concert to produce the large degree of isoform diversity observed in PRF and PMF.

If additional PRF transcripts are occasionally translated in some males, the question becomes one of functional significance. Simply because a component is expressed in low levels does not mean that it could not have profound behavioral consequences. For example, the minor components in the pheromone blends of several species of moth are responsible for the entire signal's specificity and preserve its functionality (Linn et al. 1986, 1987). While it is not clear if these peaks are new isoforms of PRF, future proteomic characterization of the minor peaks in the PRF range will address this question.

## Intraspecific Diversity of Relative Isoform Ratios

Perhaps the most important aspect of this research is the demonstration of individual variation in pheromone isoform ratios among males of the same population (Figures 2.4 and 2.5). Quantitative analysis of PRF profiles revealed that a subset of profiles was more prevalent in our study population (Figure 2.5), but overall the males of this sample possessed a wide array of possible combinations of isoform proportions. In a graphical sense, the only regions of the ternary plot not occupied by any data are the
corners (Figure 2.5). The reason for this is that in order for a point to occupy a corner, a male must have very high levels of one isoform (e.g. > 80\%), with very low levels of the other two isoforms, and no instances of this were observed. The majority of male profiles ( $81.7 \%$ ) had all three isoforms present, with the remaining males expressing 2 of the 3 isoforms. PRF-C1 was the most common isoform to be absent, with PRF-B and PRF-C2 each only missing from a small number of males. In summary, the mental gland profiles of male $P$. shermani generally exist as a well-balanced blend of all three PRF isoforms.

The amount of diversity observed in these pheromone proteins leads to three important questions pertaining to their function: (1) the efficacy of individual isoforms in eliciting sensory and behavioral responses, (2) interactive or synergistic effects stemming from combinations of isoforms, and (3) the potential role of isoform ratios acting as unique neurophysiological and behavioral signals. None of these concepts are mutually exclusive and they may all be occurring simultaneously. Future investigations should permit evaluation of the roles and functions of individual pheromones and combinations of pheromone isoforms.

Genes for PRF and PMF are thought to be under rapid gene duplication and pervasive positive selection (Watts et al. 2004; Palmer et al. 2007b, 2010). As these pheromones serve to facilitate reproduction, an intuitive explanation for this genetic pattern is sexual selection (Palmer et al. 2010). Thus, we hypothesize that there is a functional benefit to having multiple isoforms in eliciting sensory stimulation of the female. The presence of numerous variants of the pheromones may elicit the maximum effect in the peripheral and/or central nervous system, making females maximally affected by the pheromone during mating. Preliminary evidence for this hypothesis may come from the molecular evolution of vomeronasal receptors. In $P$. shermani, the vomeronasal organ (VNO) responds to both PRF and PMF (Wirsig-Wiechmann et al. 2002, 2006). In the rodent VNO, V2R receptors recognize proteinaceous pheromones (Haga et al., 2010). Kiemnec-Tyburczy et al. (2011a) recently characterized cDNA sequences for numerous V 2 R receptor variants in the vomeronasal epithelium of $P$. Shermani, though the specific receptor:ligand pairs for plethodontid pheromones remain to be determined.

A further rationale for pheromone complexity comes from the field of behavior. Multicomponent pheromone mixtures are known to be present in communication scenarios that require greater specificity and complexity of information transfer. This concept has been demonstrated in insects (Hefetz and Graur 1988) as well as mammals (Hurst et al. 2001; Beynon et al. 2002; Hurst and Beynon 2004). If the pheromone mixtures of $P$. shermani are serving a signaling function in parallel with their known role in reproductive behavior, the complexity of plethodontid chemical signals may allow for sex (Dawley 1984ab), species (Dawley 1984a), population (Rollmann et al. 2000) and/or individual recognition (Jaeger 1981; Dawley 1984a; Simon and Madison 1984). Behavioral research has revealed that multiple plethodontid species are capable of sophisticated information transfer through chemical signals, which requires a high degree of complexity in the biochemical composition of the mixtures.

If males possess multiple components, with numerous isoforms of each pheromone, these variants may act in an additive or synergistic manner when perceived by females. The concept of different combinations having profound functional effects could be the case for different components (e.g. PMF + PRF), as well as between different isoforms of a single component (e.g. PRF-B $+\mathrm{C} 1+\mathrm{C} 2$ ). Interactive isoform effects may be perceived by the peripheral (VNO) or central nervous system (CNS) in a manner that ultimately contributes to their appropriate functionality. Synergistic effects of pheromone components are also known to occur in insects and mammals (Wyatt 2003, Section 2.6) and a recent behavioral experiment suggests that such interactions between PMF isoforms may be necessary for appropriate functionality (Wilburn et al., 2015). Future research efforts will continue to investigate the potential for interactive effects of different pheromone components on reproductive behavior, as well as address the mechanisms for this phenomenon at the levels of the peripheral and central nervous system.

The mental gland pheromone mixture of $P$. shermani is not only complex in terms of the number of components and the large number of structural variants; an additional level of signal complexity may arise from the actual proportions of the individual components. Several lines of evidence support the idea that relative pheromone ratios
may serve an additional aspect of functionality in communication. The cells of the vertebrate VNO are known to be exceptionally sensitive to low ligand concentrations (Leinders-Zufall et al. 2000; Kikuayama et al. 1995). In addition, different subsets of cells are finely tuned to only specific ligands (Leinders-Zufall et al. 2000; Zufall et al. 2002). Importantly, a recent study has demonstrated that different ligand concentrations can elicit distinct signals via the VNO (He et al. 2010). Cumulatively, both the presence and concentration of plethodontid pheromones may be providing a highly integrated, complex signal to female salamanders. Thus, it may be that pheromone ratios could act as individual-specific signature mixtures (Wyatt 2010), indicator mechanisms of matequality (sensu Andersson 1994), or signals of genetic compatibility (Hurst 2009).
Even though there is a large degree of variability in mental gland profiles, some profiles occurred more frequently in the population. In the current sample, the highest density of profiles occurred with males that have intermediate levels of PRF-B and PRF-C2, with lower amounts of PRF-C1. Interestingly, the most common profile also appears to differ between populations (Rollmann et al. 2000). This type of distribution could be interpreted in terms of a profile that resulted from sexual selection in this population. This hypothesis assumes that females perceive and prefer certain combinations of isoform proportions. As such, this system could serve as an indispensable tool for understanding the nature of coevolution between male traits and female preference.

## Conclusions and Future Directions

The presented biochemical analyses have provided a quantitative description of the characteristics and variability of the mental gland composition of the red-legged salamander, Plethodon shermani. Plethodontid salamanders provide an important nonmammalian system for studying the role of pheromones in reproduction and social behavior. This amphibian family also serves as a useful study system in the field of molecular evolution: pheromone genes are under rapid and pervasive positive selection in contrast to stasis exhibited by most other elements of the functional complex that serves to facilitate reproduction (Watts et al. 2004). Overall, this complex revolves largely around pheromone transfer, which involves genes, proteins, physiology, sexually
dimorphic anatomy, neurobiology and behavior. Our future research aims are to investigate the physiological processes that regulate pheromone production in addition to the molecular and neurophysiological mechanisms that mediate pheromone reception. Ultimately, this study has identified several aspects of pheromone complexity that could have profound effects on the functionality of this multicomponent chemical signal. These aspects of signal complexity include: (1) the number of components, (2) the particular isoforms present and (3) the relative ratios of the individual components and their isoforms. This description of the remarkable individual variability in pheromone composition establishes the foundation for testing hypotheses regarding the way in which the complexity and diversity of the signal may affect sensory reception and ultimately behavior. Future research efforts will address the potential fitness consequences to individuals possessing particular differences in pheromone composition, as well as the evolutionary pressures and mechanisms that may have led to the rapid diversification of structural variants in the protein pheromones present in this family of salamanders.


Figure 2.1. RP-HPLC chromatogram of mental gland extract. Representative chromatogram of a complete mental gland profile from a single male $P$. shermani. The retention time and abundance can be seen for the numerous isoforms of PMF (20 - 40 minutes), component C3 ( $\sim 41$ minutes) and the three isoforms of PRF ( $50-60$ minutes). The $y$-axis depicts absolute absorbance values, while the $x$-axis has been cropped to the range in which the proteinaceous components elute.


Figure 2.2. Proportions of major proteinaceous components of the mental gland. The two major proteins (PRF + PMF) make up over $80 \%$ of the entire secretion, with PMF regularly present in greater abundance. Of the remaining proteins, C3 makes up more than half ( $\sim 10 \%$ of the whole extract). This leaves less than $10 \%$ of the extract being composed by unknown components. The line represents the median and the circle represents the mean. The box represents the interquartile range, while the whiskers depict the total range of the data (maximum - minimum).


Figure 2.3. Example chromatograms of individual PRF profiles that contain minor components. Depicted are data for males that: (A) have a complete MG profile, (B) are missing PRF-B, (C) are missing PRF-C1, (D) are missing PRF-C2, (E) have an extra peak before PRF-B, and (F) have an extra peak before PRF-C1. The x-axis depicts the range of retention times for PRF ( $50-60 \mathrm{~min}$ ), but in some cases it has been slightly shifted to align the isoforms across all panels. The y-axis is scaled in arbitrary units.


Figure 2.4. Distribution of individual PRF isoform proportions. Isoforms PRF-B and PRF-C2 typically exist in intermediate proportions, while PRF-C1 is more likely to exist in lower proportions. Isoform PRF-C1 is also the most likely to be missing from a male's profile. The circles above the $x$-axis depict the individual data points, while the $y$-axis is a kernel function that estimates the density of the data for the corresponding $x$-value.


Figure 2.5. Variation in male PRF profiles. Ternary plot depicting the intraspecific diversity in PRF isoform ratios of P. shermani ( $\mathrm{n}=104$ ). Each point corresponds to the PRF profile of a single male. Each axis corresponds to the proportion of a single isoform and the gridlines for the appropriate isoform are in the same alignment as the units for its axis. To identify the proportions of each isoform in the total PRF composition, the axes for individual isoforms can be used to triangulate the position in the plot that corresponds to the profile of a given male. The colored contours depict a kernel-density function that estimates the clustering of data in a given area of the plot. Blue shading indicates low density of points, with an increasing gradient to red indicating higher density.

## CHAPTER THREE

# INTER- AND INTRASPECIFIC VARIATION IN PHEROMONE GLYCOSYLATION IN PLETHODONTID SALAMANDERS 

Chouinard, A.J.* ${ }^{1}$, Wilburn, D.B. ${ }^{2}$, and Weis, V. ${ }^{1}$<br>1. Department of Integrative Biology, Oregon State University, Corvallis, OR 97331<br>2. Department of Biochemistry, University of Louisville, Louisville, KY 40202


#### Abstract

Proteinaceous pheromones play a pivotal role in facilitating courtship in plethodontid salamanders. Recent proteomic observations revealed that a major component of this signal, Plethodontid Receptivity Factor (PRF), is glycosylated in some lineages but not others. Glycosylation was observed in Plethodon cinereus, a member of the ancestral clade utilizing dermal pheromone delivery, but not in $P$. shermani, a member of the derived clade with olfactory delivery. In this research, various bioinformatic tools were used to assess the likelihood that these observations apply to each species' representative clade. This is likely true based on the phylogenetic distribution of N -linked glycosylation motifs and their predicted glycosylation patterns. Variations of these motifs evolved multiple times, in different regions of the protein and in different isoform lineages. In addition, sequencing efforts in a previously unstudied population of $P$. cinereus revealed substantial intraspecific differences in isoform sequence and glycosylation potential. Intraspecific variation in glycosylation pattern are also suggested by SDS-PAGE analyses. In the broader context, there is substantial intraand interspecific variation in PRF glycosylation potential. The presence of glycosylation is highly correlated with a transition in pheromone delivery behaviors, and may be the result of selection for different target tissues and mechanisms of action between lineages. Protein structural modeling revealed that the primary N -linked sites occur in critical binding areas of a plesiomorphic receptor pathway for this protein (IL-6R $\alpha / \mathrm{gp} 130$ complex). These sites differ from the location of glycosylation in the corresponding ligand (IL-6), the binding of which is not affected by glycosylation. This suggests


selection for modulation of ancestral receptor pathways. In total, these analyses reveal key insights into the evolution of a multilevel trait complex in this rapidly-diversifying clade of vertebrates. As reproductive pheromones with substantial interspecific variation, these structural modifications may be involved in species-specific signaling.

## Introduction

Communication is an integral component of finding, assessing, and reproducing with mates. While the often elaborate courtship behaviors of vertebrates have warranted study from an ethological perspective, courtship is also important in an evolutionary context, as the aggregate of individual mate-choice decisions can have profound consequences over millennia. These interactions shape reproductive traits through time, and in some cases divergence in these traits can ultimately lead to cladogenesis, the birth of new lineages. Thus, there are many examples in which selection has acted on sexual signaling systems, to modify or elaborate them into species-specific characters that serve to establish and/or maintain reproductive isolation (Andersson, 1994). By far the oldest and most widespread communication system, chemical signaling is no exception and has long been considered to play an important role in speciation (Smadja and Butlin, 2009; Wyatt, 2014a).

## Plethodontids as a Model for Reproductive Isolation

One such mechanism has been proposed for the pheromones of plethodontid (lungless) salamanders, who rely heavily on chemical signals, and for whom other sensory modalities may prove less reliable (Dawley, 1986a). This hypothesis hinges on three main assumptions: (1) that the species in question are in fact reproductively isolated, (2) that their chemical signals have diverged enough to be distinguishable, and (3) that individuals of a given species actually perceive and act upon these signal differences in natural mate-choice scenarios.

After decades of behavioral and molecular research in this system, there is evidence that these assumptions are met in many cases. First, sexual isolation has been frequently documented for both allopatric and sympatric species pairs, but also for
allopatric populations of the same species (reviewed in Arnold et al., 1993). Secondly, chemical signals are known to be important mediators of these reproductive interactions. Among other mate-assessment information (Chouinard, 2012), plethodontid chemical signals can convey the sex and species of the sender (Jaeger and Gergits, 1979; Dawley, 1984a; 1986ab; Verrell, 1989; 2003; Palmer and Houck, 2005). Species-recognition is essential information for avoiding interspecific courtships. These behavioral observations suggest that the chemical signals involved are variable enough to convey species-specific information, and subsequent biochemical analyses have confirmed that conclusion, revealing a high degree of signal complexity and divergence.

## Courtship Pheromones in Plethodontid Lineages

To date, two protein pheromone components have been well characterized in plethodontids: Plethodontid Receptivity Factor (PRF; 22 kDa ) and Plethodontid Modulating Factor (PMF; 7 kDa ). Each component is present in numerous variants (isoforms) within an individual, resulting from a series of gene duplication events followed by rapid positive selection (Watts et al., 2004; Palmer et al., 2010). PRF is a four-helical cytokine of the interleukin-6 (IL-6) family (Rollmann et al., 1999; Watts et al., 2004), and is present in several isoforms in an individual male (Chouinard et al., 2013). PMF is a "hypervariable" member of the three-finger protein (TFP) superfamily, with individual males producing dozens of isoforms (Wilburn et al., 2012). This degree of signal divergence allows for a great deal of interspecific and intraspecific variation in overall pheromone composition.

The known plethodontid pheromones are produced by the mental gland (MG), a male-specific gland on the underside of the chin. These secretions are administered to females at various stages surrounding the "tail-straddling walk" (TSW) courtship behavior (Arnold, 1976; 1977), and significantly decrease the duration of the ritual (Rollmann et al., 1999). This effect has often been interpreted as increasing female receptivity to mating (Houck, 2009). The exact means by which males administer these pheromones depends on their position in the plethodontid phylogeny (Palmer et al., 2005; Picard, 2005).

The ancestral mode of pheromone delivery is through the skin, in which males "scratch" the female integument with specialized premaxillary teeth while rubbing their mental gland over the wound. This is thought to deliver MG secretions to the peripheral circulatory system, effectively "vaccinating" females with pheromone (Organ, 1961; Arnold, 1977). The target tissue(s) for pheromone action in this case are unknown and in need of study. Males of these species typically have small mental glands situated immediately behind the mandibular symphysis, in close proximity to the premaxillary teeth (Sever, 1976). I will refer to this group throughout as the ancestral and/or "dermaldelivery clade" (DDC).

In contrast, the most derived mode of pheromone delivery is through the olfactory pathway; in this clade, males deliver MG secretions directly to the snout of females in a "slapping" behavior (Arnold, 1977). Males of these species lack premaxillary teeth, and have large protruding mental glands, situated more posteriorly in the center of the mentum (the underside of the chin; Sever, 1976). These pheromones stimulate neurons in the vomeronasal organ (VNO; Wirsig-Wiechmann et al., 2002, 2006), which project to the accessory olfactory system and other regions of the central nervous system involved in reproduction (Laberge et al., 2008). I will refer to this group as the derived and/or "olfactory-delivery clade" (ODC). This transition marks an interesting example of a shift in the mode of action (and target tissues) for these pheromones, perhaps even constituting an exaptation of ancestral pheromones for a new function.

The transition between these two states consists of a small paraphyletic group termed the "intermediate group" (IG). This group also lacks premaxillary teeth, but the precise mechanism of pheromone transfer is unknown due to infrequent observation and less clear delivery mechanisms. The few observations that exist indicate some degree of olfactory delivery (Picard, 2005). Interestingly, unlike the pronounced positive selection evidenced in the pheromones of both the plesiomorphic and derived clades, PRF proteins in the intermediate group show signs of purifying selection (Palmer et al., 2005), perhaps indicating different selective pressures acting on MG pheromones over the course of this evolutionary transition.

## Observations of Glycosylation and Goals of the Present Study

Recently, Wilburn et al. (2014a) demonstrated that PRF is glycosylated, a common post-translational modification involving the addition of carbohydrates to protein side chains (Apweiler et al., 1999; Blom et al., 2004). Carbohydrates can be attached to asparagine residues as " N -linked" glycans, or to Serine/Threonine residues as "O-linked" glycans (see methods for more detail on the conditions required for glycosylation). This research revealed that PRF is heavily glycosylated in Plethodon cinereus, a member of the ancestral, dermal-delivery clade. In contrast, PRF is not glycosylated in P. shermani, a member of the derived, olfactory-delivery clade. As such, the presence of glycosylation may be correlated with a behavioral transition in the evolution of plethodontids: the mode of pheromone delivery. However, this hypothesis assumes that the proteomic observations for each species are representative of their respective clades.

A major goal of the present study was to further explore the phylogenetic distribution of PRF glycosylation in a genus-wide, high-throughput manner through the use of bioinformatic analyses. This tested the hypothesis that PRF glycosylation is associated with the ancestral pheromone delivery behavior. Secondly, this research aimed to identify the potential for both inter- and intraspecific variation in the patterns of glycosylation, which may provide an additional level of species-specific signal divergence. This hypothesis predicted that the patterns of glycosylation should differ among lineages in a given clade. The null hypothesis was that glycosylation does not encode the potential for a species- or population-specific signal; this predicted that if glycosylation is present in more members of the dermal-delivery clade, it should not vary in its pattern between species.

To address these hypotheses, I explored intraspecific PRF variation by: (1) identifying new PRF sequences from an unstudied allopatric population of $P$. cinereus (New Hampshire) and (2) comparing the sequences and glycosylation patterns of these PRFs to a previously-studied population of the same species (Virginia; from Wilburn et al., 2014a). To assess interspecific variation, I also (3) used bioinformatic tools to predict glycosylation potential across all known PRF sequences, (4) compared the predicted
glycosylation patterns between species, and (5) explored the evolutionary history of the sequence motifs involved. Lastly, (6) structural implications of PRF glycosylation in the ancestral clade were examined through tertiary structure prediction and analysis of receptor binding interactions.

## Methods

## Animal Collection, Gland Removal and Extraction

Mental gland extract from three species was used in this study: Desmognathus ocoee (outgroup; no PRF), Plethodon shermani (olfactory-delivery clade; ODC), and P. cinereus (dermal-delivery clade; DDC). For P. cinereus (Pcin), two populations were compared to investigate intraspecific variation: a southern population from Giles County, Virginia (VA), and a northern population from Strafford County, New Hampshire (NHPcin) for which mental gland (MG) pheromones had not previously been biochemically characterized. Mental gland samples for $D$. ocoee, $P$. shermani, and the VA population of $P$. cinereus (VA-Pcin) were obtained by other researchers, and were the same samples analyzed in Wilburn et al. (2014a). Thus, individuals from the NH-Pcin population were the only animals collected directly for this research.

Animals were collected from the University of New Hampshire college woods in Durham, NH ( $43^{\circ} 8^{\prime} 16$ "N, $70^{\circ} 56^{\prime} 53^{\prime \prime W}$ ) in November, 2010 (UNH IACUC \#080502). This corresponds to peak activity in the fall mating season (Sayler, 1966; Petranka, 1998). All other samples were collected in their corresponding ranges during their respective breeding seasons. Mental glands (MG) were removed for protein extraction under a standardized protocol (Feldhoff et al., 1999). Briefly, males were anesthetized in $7 \%$ diethyl ether for 10 minutes. Once fully anesthetized, mental glands were surgically excised with iridectomy scissors and rinsed in Amphibian Ringer's (AR) solution.

For protein analysis, glands $(\mathrm{n}=22)$ were placed into $200 \mu \mathrm{l}$ of 0.8 mM acetylcholine chloride (Ach) in AR to extract protein from the tissue. Glands remained in the Ach solution for 60 minutes and were then subjected to three rounds of centrifugation ( $10,000 \times \mathrm{G}$ ) for 10 minutes, followed by extraction of supernatant protein. Additional
glands ( $\mathrm{n}=6$ ) were placed in RNAlater (Ambion) for subsequent molecular analysis of PRF sequences.

## SDS-PAGE Glycoprotein Analysis

Electrophoretic analysis for all samples was carried out as in Wilburn et al. (2014a), and all samples (including the NH-Pcin) were run in the same gel depicted in Figure 2 of that study. Briefly, protein samples were concentrated with a YM-3 Centriprep (Millipore) and buffer-exchanged to 0.5 X phosphate-buffered saline (PBS). SDS-PAGE analysis consisted of $15 \%$ Tris-Tricine gels in duplicate (with $4 \%$ stacking gels) at 100 V for 75 minutes (after 50 V for 15 minutes). The far left lane (L1) contained a protein size ladder (M3913; Sigma-Aldrich), and the far right lane (L6) was loaded with a glycoprotein size ladder (C21852; Molecular Probes), as a positive control for subsequent glycosylation staining. The middle lanes were loaded with one of four MG samples: D. ocoee (L2), P. shermani (L3), P. cinereus from VA (L4), and P. cinereus from NH (L5). In all cases, samples were standardized to $20 \mu \mathrm{~g}$ of MG extract. One of the duplicate gels was stained for protein with Coomassie Brilliant Blue (CBB; Bio-Rad), while the second gel was stained for glycoproteins with Pro-Q 488 (Invitrogen). For easier visualization, the stains were artificially colored: protein stain (CBB) was colored green and glycan stain (Pro-Q 488) was colored red.

## Sequencing of NH P. cinereus PRFs

Total RNA was extracted from three individual mental glands using Trizol (Invitrogen), and quantified in order to create a normalized pool. Full length cDNA was prepared using the SMARTer cDNA synthesis kit (Clontech). PRF sequences were amplified using RT-PCR with Accuprime High Fidelity DNA Polymerase (Invitrogen) using oligo-dT and a gene-specific primer targeting the 5' untranslated region of PRF (5’-CGCTGGAATCTAGAATGAGG-3’). PCR products were cloned using pCR2.1-TOPO (Invitrogen), transformed into TOP10 chemically competent E. coli (Invitrogen), plated on LB/Kan, and individual clones subjected to colony PCR using M13 primers. Amplified DNA was supplied to the University of Louisville DNA Core Facility for

Sanger sequencing. Sequence data was analyzed using the DNASTAR Lasergene software.

## Inter- and Intraspecific Analyses of PRF Glycosylation

To assess the prevalence of PRF glycosylation sites throughout all previouslyidentified PRF sequences ( $n=224$ ), each available PRF mRNA sequence was downloaded from GenBank. Duplicate intraspecific coding sequences (CDS) and those from hybrid animals were removed for each species (yielding 198 PRFs). These sequences were combined with the non-redundant NH-Pcin PRFs ( $\mathrm{n}=20$ ) obtained from cDNA sequencing (above). This process resulted in a total of 218 PRF sequences for all subsequent analyses. All remaining coding sequences were translated and aligned with PROMALS3D (Pie et al., 2008), as incorporating secondary structure information into the alignment provided a better starting point than other methods tested. The alignment was manually-inspected and adjusted accordingly. All amino acid residues referenced herein refer to its position in this total alignment, which includes the signal peptide. PRF sequences were annotated for (1) signal peptide presence and cleavage site (SignalP-4.1; Petersen et al., 2011), (2) secondary structure and (3) solvent accessibility (Jpred-v4; Drozdetskiy et al., 2015), (4) N-linked glycosylation potential (NetNGlyc-1.0; Gupta et al., 2004), and (5) O-linked glycosylation potential (NetOGlyc-4.0; Steentoft et al., 2013).

The potential for N -linked glycosylation requires the presence of known motifs (sequons) of either NXT (Asn-Xxx-Thr) or NXS (Asn-Xxx-Ser), in which X can represent any amino acid other than proline (Katsuri et al., 1997; Petrescu et al., 2004; Hamby and Hirst, 2008). While these sequons are the most common, other rare motifs have been found, such as AXC (Asn-Xxx-Cys; Hamby and Hirst, 2008), or more recently NGX (Asn-Gly-Xxx) or NXV (Asn-Xxx-Val; Feng et al., 2015). O-linked glycosylation can be attached to any S (Ser) or T (Thr) residue (Caragea et al., 2007; Julenius et al., 2005). Other forms of glycosylation also exist (C-mannosylation, GPI anchors; Blom et al., 2004) but are not present in PRF. The presence of these motifs alone is not sufficient for glycosylation, which depends on the regulatory environment of the cell and aspects of
protein tertiary structure. Glycosylation has traditionally been thought to require sufficient solvent accessibility of the target residue, although Petrescu et al. (2004) demonstrate that it can still occur in cases with very low solvent accessibility. While the predictive power of this variable remains unclear, it is nevertheless included for reference when visualizing glycosylation sites.

The glycosylation prediction tools used have an overall accuracy of $>76 \%$ for N linked sites and $>88 \%$ for O-linked sites (Hamby and Hirst, 2008; Steentoft et al., 2013). In predicting O-linked glycosylation, the default NetOGlyc-4.0 score cutoff (0.5) yielded one known false positive: residue T91 of Pshe was predicted as glycosylated, whereas proteomic data suggests otherwise (Wilburn et al., 2014). Thus, the cutoff threshold was raised until this prediction was eliminated (at 0.78 ). This allowed for more conservative O-linked predictions, consisting of only sites with scores above this false positive. While imperfect, these tools provide a powerful starting point for investigating inter- and intraspecific pheromone variation in a comprehensive and high-throughput manner.

To explore relatedness between PRF isoforms, a Bayesian consensus phylogeny was constructed for all PRF amino acid sequences using MrBayes-3.2.6 (Ronquist et al., 2012) on the CIPRES Science Gateway (Miller et al., 2010). MrBayes was run with three replicates for 4,000,000 generations each, sampling every 1,000 generations. The first quarter of samples (1,000,000 generations; 1,000 samples) was discarded as burn-in; this was confirmed as sufficient through visual inspection of the log-likelihood values of the cold chain. These parameters yielded a final sample of 3,000 trees. The outgroup sequence used was the cardiotrophin-2 (Ctf2) coding sequence of Mus musculus (NM_198858). Phylogenies were plotted and explored with the "phytools" package in R (Revell, 2012).

In order to visualize the phenotypic (i.e. structural) diversity between PRF isoforms, models of tertiary structure were constructed for all Pcin sequences with ITASSER (Yang et al., 2015; Roy et al., 2010; Zhang, 2008). This program was selected because of its high accuracy and repeated top performance in the Critical Assessment of Structural Prediction (CASP) competitions (Yang and Zhang, 2015). These models allowed for qualitative inspection of the position of glycosylation sites and other aspects
of sequence divergence in the context of PRF structure. All Pcin PRF models were compared in pairwise structural alignments via TM-align to generate a matrix of percent identities and TM-scores (a more accurate representation of full-length structural alignment than alternatives; Zhang and Skolnick, 2005).

To view the extent of glycosylation expected between isoforms and its effect on pheromone structure, select isoforms were glycosylated in silico via GlyProt (BohneLang and van der Lieth, 2005). To select an appropriate glycan, the GlyProt database was searched for glycans of the appropriate size ( 30 monosaccharides per N -linked site; Wilburn et al., 2014a). Only two glycans of this size were present in the database (LinucsID 16879, 16880); both are very similar and consist of penta-antennary structures associated with a vertebrate reproductive protein (Seko et al., 1989). To be clear, however, these models show arbitrary glycosylation events; they are depicted purely for visual inspection of the effect of the observed degree of PRF glycosylation on the total pheromone structure.

Lastly, to investigate the hypothesis that glycosylation might affect ligand:receptor interactions, a representative PRF model (VA-A1) was aligned with a known cytokine receptor structure: the IL-6/IL-6R $\alpha /$ gp130 complex (PDB: 1P9M; Boulanger et al., 2003). This complex serves the basis for IL-6-type cytokine signaling, and exerts its intracellular effects through the JAK/STAT pathway (Heinrich et al., 1998). Other receptor combinations are possible (e.g. the leukemia inhibitory factor receptor; Skiniotis et al., 2008; Huyton et al., 2007), but the IL-6R $\alpha /$ gp130 complex was chosen because it serves as a representative for the stereotypical IL-6-type receptor interaction. Its structure consists of two identical heterotrimers composed of (1) an interleukin-6-type ligand (e.g. IL-6), (2) a corresponding alpha receptor (e.g. IL-6R $\alpha$ ), and (3) glycoprotein 130 (gp130). These two heterotrimers combine in an antiparallel fashion, resulting in a hexameric complex necessary for signaling (Boulanger et al., 2003). Aligning PRF to IL6 allowed for examining the proximity of glycosylation sites to the three conserved IL-6type cytokine receptor binding regions. All protein models were visualized with UCSF Chimera (Pettersen et al., 2004).

## Results

SDS-PAGE Analyses of Inter- and Intraspecific Differences in Pheromone Glycosylation
Substantial inter- and intraspecific differences in pheromone glycosylation patterns were observed via SDS-PAGE (Figure 3.1). While PRF is not present in $D$. ocoee (Figure 3.1; lane 2), the primary proteinaceous component of this species is a 20 kDa protein named Sodefrin Precursor-like Factor (SPF) that migrates to a similar size range (Palmer et al., 2007a; Doty et al., 2015). As expected, the PRF band of P. shermani is not glycosylated (Wilburn et al., 2014a); however, the overall MG extraction does contain several minor components associated with carbohydrates that have yet to be explored. In contrast, the PRF of P. cinereus is glycosylated to differing degrees, resulting in several bands eluting at various points in the PRF range, from 25 kDa to 37 kDa . The proteins migrating to these ranges were positively confirmed as PRF by Wilburn et al. (2014a), as was the presence of glycosylation. Specifically, in VA-Pcin the 37 kDa band consisted of a dual N-linked glycosylated PRF best matching PRF-A2 (herein VA-A2 to distinguish between populations), with $94 \%$ sequence coverage identified via mass spectroscopy. A second PRF in this same region of the gel ( 37 kDa ) also possessed two N-linked glycans, and best matched VA-A5 (86\% coverage). Lastly, the 25 kDa band best matched VA-B1 (84\% coverage), but contained no N-linked glycans. All three PRFs still stained for glycans after enzymatic deglycosylation of the Nlinked sites. The authors conclude that this is likely due to O-linked glycans in the Nterminus at position T21 or S23 (T2 and S4 after signal peptide cleavage in VA-Pcin respectively).

Putative PRFs from NH-Pcin are also glycosylated, but the patterns of glycosylated PRFs differ between the two populations of $P$. cinereus (Figure 3.1). This suggests that there are intraspecific differences in PRF glycosylation in addition to the observed interspecific differences. The bands corresponding to VA-Pcin PRFs at 25 kDa , 29 kDa , and 37 kDa appear less glycosylated in NH-Pcin. Furthermore, an additional glycosylated band appears at 42 kDa in NH-Pcin that has yet to be characterized. While migrating in the PRF range, these bands have not been definitively identified as PRF via mass spectroscopy; an alternative possibility is that another component might account for
the different patterns of glycosylation observed. Another protein does appear in this range, a cysteine-rich secretory protein (CRISP) homolog migrating to 25 kDa , although at a lower level of expression than PRFs. Importantly, this protein is not glycosylated in P. cinereus (Wilburn et al., 2014a). Thus, we infer that these proteins are indeed PRFs due to: (1) the absence of any other glycosylated proteinaceous components appearing in this range, (2) the conclusive identification of these bands as PRF in a different population of the same species from the same physical gel, and (3) the presence of widespread glycosylation as expected for $P$. cinereus PRFs.

## Sequence Analyses for NH P. cinereus PRFs

To further explore the potential for species-specific glycosylation patterns, PRF mRNA sequences from the NH population were identified and analyzed. Sequencing efforts of NH-Pcin yielded 20 unique coding sequences with 16 unique translations. This more than doubles the number of known PRF sequences for this species. These sequences encode 10 PRF-A isoforms and 10 PRF-B isoforms (sensu Watts et al., 2004). New sequences were named according to their phylogenetic position (Figures 3.2 and 3.3), with earlier-diverging sequences given the lowest number. In the case of polytomies, sequences were numbered according to their order in hierarchical clustering of pairwise percent identities (Figure 3.4a). These new sequences also contained two of the most diverse PRFs identified to date: two PRF-B isoforms with shortened C-termini (NH-B9 and NH-B10). The NH sequences are generally more closely-related to each other than to other PRFs (Figure 3.2). Almost all of the NH-Pcin PRF-B sequences ( $\mathrm{n}=9$ ) form a monophyletic clade, with the remaining NH-Pcin PRF-B sequence (NH-B1) related to those from $P$. serratus, P. hoffmani, and $P$. richmondi. The NH-Pcin PRF-A sequences form two monophyletic clades, each with five sequences: NH-A1 to NH-A5 are more closely-related to select VA-Pcin PRFs (VA-A6 to VA-A8), while NH-A6 to NH-A10 form the outgroup. In comparing mean pairwise protein distances between species, the VA-Pcin PRFs are more similar to those of $P$. serratus (0.182) and P. hoffmani (0.190) than to NH-Pcin PRFs (0.204). In contrast, the NH-Pcin PRFs are more similar to VAPcin PRFs than any other species.

## Glycosylation Predictions and Sequence Annotations

The NH-Pcin PRF sequences were combined with all previously identified PRFs and annotated for predicted glycosylation potential, along with other key protein features (Figure 3.5). Glycosylation predictions support the hypothesis that N -linked glycosylation is strongly correlated with pheromone delivery mode. All members of the ancestral clade (DDC) possess multiple PRF isoforms with N-linked glycosylation. Furthermore, in each species there are at least two isoforms with distinct glycosylation patterns. Different glycosylation patterns within each species are depicted with individual isoforms shown as brackets; each N -linked sequon present per isoform is listed by the residue position, followed by whether it was predicted to be glycosylated (+) or not (-). All members of the DDC possess multiple PRF glycosylation patterns: P. serratus [96-, 101+], [156+]; P. cinereus [96-, 101-], [96+, 101+], [156+]; P. hoffmani: [96-, 101-], [96+, 101+], and [156+]; and P. richmondi [96-, 101+], [156+]. PRFs from the intermediate group are largely not predicted to be glycosylated. Only two IG species possess N-linked sequons: P. websteri [96-, 101+] and P. wehrlei [96-, 149-]. Of these, only $P$. websteri (the earliest diverging intermediate species) is predicted to have a glycosylated site. Notably, P. wehrlei contains a unique N-linked sequon, although it is not predicted to be glycosylated. Like P. shermani, all members of the ODC lack Nlinked sequons and therefore have no potential for N -linked glycosylation.

The predictions for O-linked glycosylation also varied between and within species (Figure 3.5), as well as between populations of $P$. cinereus (Figure 3.6). The strongest predictions appeared in the N-terminal region of PRF in members of the DDC. This is consistent with the findings of Wilburn et al. (2014a), who concluded that O-linked glycosylation probably existed in this region for VA-Pcin on either T21 or S23. The prediction scores varied by isoform and species, but two main locations emerged: residue S23 (ancestral clade mean $=0.82$ ) and residues 29-32 (ancestral clade means $=0.83$, $0.78,0.80$ ). By integrating these predictions with mass spectroscopy data (Wilburn et al., 2014a), the best candidate for O-linked glycosylation is residue S23, at least in the VA population of $P$. cinereus.

While the extended N-terminal region of the ancestral PRFs encompassed the highest prediction scores, a few other potential O-linked sites were also identified. One such site was residue T91 for select PRF-B isoforms in members of the ancestral clade (Figure 3.5), notably for NH- but not VA-Pcin (Figure 3.6). This same site is also predicted as glycosylated for select PRF-A isoforms from a few members of the derived clade other than $P$. shermani. It is unclear whether these predictions are also false positives or if there is variation in the presence of glycosylation among members of the ODC in reality. Without further evidence, we anticipate that this site is not glycosylated in the derived clade due to (1) the known false positive at this residue in a member of this clade, and (2) the overall trend of a loss of glycosylation in the derived clade. It should be noted, however, that when predicted to occur, the scores at this site are higher in the ancestral clade ( 0.80 to 0.85 ) than those in the derived clade ( 0.78 to 0.82 ). The last O linked prediction occurred in the distal C-terminus of one of the truncated NH-Pcin PRFs (NH-B10). This site (S189) is the only exception to a $100 \%$ conserved Lysine residue present in all other PRF sequences.

As in the case of N -linked glycosylation, there is potential for intraspecific variation in the position of O-linked glycosylation between PRF-A and -B isoforms, as well as between NH-Pcin and VA-Pcin PRFs after subjecting them to site-specific Welch's t-tests (Figure 3.7). Briefly, PRF-B isoforms for both populations have significantly higher O-linked scores than PRF-A isoforms in the center of the protein (residue T91). PRF-B isoforms from NH-Pcin also have significantly higher scores in the C-terminus region (residues 189, 193-195) than the PRF-B isoforms from VA-Pcin. PRFA isoforms from NH-Pcin also have a slightly higher density of high O-linked scores in the N -terminal region.

## Phylogenetic Analysis of PRF Glycosylation Motifs

Both N -linked sequons at site 1 (Figure 3.8a) and site 2 (Figure 3.8b) are present in all PRF-A isoforms of the ancestral DDC. A monophyletic origin of these sequons is highly likely, but the sites are lost at different stages beyond the DDC. In the IG, site 1 is present in $P$. websteri and $P$. wehrlei, but not in $P$. ventralis, $P$. dorsalis, or $P$. welleri. In
contrast, site 2 exists only in $P$. websteri in the IG, the earliest-diverging species in this group. Both sequons are absent in the derived ODC. The two remaining sequons evolved independently, implying homoplasic glycosylation at multiple sites in the protein structure. Site 3 (N149) is found only in P. wehrlei PRFs (Figure 3.8c) and all sequences are monophyletic (node Pp=1.0). While present in this member of the intermediate group (IG), this site is never predicted to be glycosylated. If it is glycosylated, however, it would represent a homoplasic reversion to the ancestral state. Site 4 (N156) also evolved independently, the sole site to evolve in the PRF-B isoforms of the DDC (Figure 3.8d), and is likely monophyletic ( $\mathrm{Pp}=0.81$ ). This group includes one of the NH-Pcin sequences (NH-B1), the only P. cinereus PRF identified to date with this glycosylation site. This site is also found in P. serratus, P. hoffmani, and P. richmondi. The presence or absence of predicted glycosylation is polyphyletic for site 1 (N96), which might imply individual lineages being subjected to different selective pressures if the predictions are accurate.

## Structural Implications of PRF Glycosylation

The predictions for tertiary structure (after signal peptide cleavage) revealed a characteristic IL6-type cytokine fold (Figure 3.9). Some PRF models from P. cinereus contained an additional short alpha-helix at the C-terminus (such as the model depicted), but this is not present in all isoforms. This region is shortened in the PRFs of the ODC, and constitutes the region with the lowest degree of conservation across the phylogeny. The vast majority of polymorphic sites in PRF occur in loops or on the surface of helices, implying conservation of the cytokine fold and variation mainly among surface structures.

The locations for the "primary" (i.e. the oldest and most common) N-linked sites (N96 and N101) occur on the surface of the protein in the long loop between the first and second helix. The most likely candidate for O-linked glycosylation in the DDC (S23) occurs on the opposite side of the protein at the N-terminus. This site is highly exposed, and the second most variable region across all known PRFs. While S23 seems the most likely O-linked candidate when combining predictions with proteomic data, the residue
receiving the highest O-linked prediction score (S29) is situated nearby. The "alternative" (i.e. derived and less common) N -linked sites (N149 and N156) occur on the third helix but are located spatially near the N -terminus. The side chains for these residues form the surface of a region neighboring the likely O-linked sites, but oriented in the opposite direction.

The addition of the expected N-glycans (inferred by Wilburn et al., 2014a) leads to a remarkable change in the overall structure of PRF (Figure 3.10), increasing the pheromone's surface volume by approximately $50 \%$ (data not shown). To examine the possibility for these large glycans to affect receptor interactions, PRF was superimposed onto the IL-6/IL-6R $\alpha / \mathrm{gp} 130$ receptor complex (Boulanger et al., 2003). When aligning PRF to the IL-6 ligand, all N-linked sites occur in known receptor binding regions (Figures 3.11 and 3.12). The primary N-linked sites (N96 and N101) are situated in the "Site I" binding region, composed of the interaction between the first and fourth helices of the ligand, and the bend between the D2 and D3 domains of IL-6R $\alpha$. The first N linked site (N96) is located next to and oriented toward the IL-6R $\alpha$ D3 domain. The second N -linked site (N101) is located distally to the first and its side chain is oriented away from the receptor complex. Thus, in this model, glycosylation of PRF at N101 may (or may not) affect receptor binding, whereas this is a strong possibility in the case of N96.

The alternative N -linked sites (N149 and N156) are located on the opposite side of PRF and occur in close proximity to the "Site II" binding region, which consists of an interaction between the first and third helices of the ligand, and the bend between the D2 and D3 domains of gp130. Both side chains are located next to and oriented directly toward these domains, and thus glycosylation of these residues is also likely to affect binding interactions. The probable O-linked sites of PRF are also on this side of the protein, but occur on the N-terminal loop, which is situated above and oriented away from the "Site II" binding pocket. While the potential for these sites to affect receptor interactions will ultimately depend on the tertiary structure of this loop in reality, in this model N-terminal O-linked glycosylation is not likely to affect binding. No sites for
which there is proteomic evidence for glycosylation occur at or near the "Site III" binding region with the D1 domain of gp130 from the opposing heterotrimer.

When comparing glycosylation sites between PRF and IL-6 (Figure 3.12), there are substantial differences in their locations within the receptor complex. Human IL-6 (hIL-6) is glycosylated and can exist in multiple glycoforms (Gross et al., 1989; Cruz et al., 2004). There are two N-linked sequons: residues N73 and N172 (N45 and N144 in the model). Only the first sequon (N73) is glycosylated (Cruz et al., 2004); both sites are correctly predicted by NetNGlyc. This N-linked site (N73) occurs at the end of the first helix, at the start of the long loop. This site is located in the center of the receptor complex, oriented downward but well removed from any binding sites. There are three known O-linked glycosylation sites: T166, T170 and/or T171 (T138, T142, and T143 in the model). These sites are also predicted as glycosylated with NetOGlyc-4.0. These three O-linked sites are all located near each other, in the middle of the short loop between the third and fourth helices. When binding, they are located on the top of protein facing outward from the receptor complex, completely removed from any binding regions. Due to their locations, neither N- nor O-linked glycosylation affect binding in hIL-6 (Cruz et al., 2004).

## DIscussion

It is possible for pheromones to play a significant role in the establishment and maintenance of reproductive isolation, but in order for this to occur, individuals must be able to discern between the signals produced by the diverging lineages. In the case of chemical signals, perceptual differentiation is directly related to the biochemical composition of the signals involved, but there are multiple ways in which these signals can vary. The overall composition of plethodontid pheromones varies in terms of (1) the presence/absence of different pheromone proteins (Palmer et al., 2007a; KiemnecTyburczy et al., 2009) and (2) the primary sequence identity among pheromone homologs for each protein (Watts et al., 2004; Palmer et al., 2005, 2010), but also by (3) isoform presence/absence for each protein component and (4) the relative levels of isoform expression (Chouinard et al., 2013).

This study describes an additional level of functional signal diversity: the presence and patterns of clade-specific post-translational modifications. Specifically, glycosylation of Plethodontid Receptivity Factor (PRF) is prevalent in the plesiomorphic state (dermal-delivery; DDC), but has been lost in the derived clade (olfactory-delivery; ODC). These findings mark an intriguing example of changes in a biochemical trait occurring alongside an evolutionary transition in behavior. Much attention in this system has been paid to the "decoupling" of the levels of selection; i.e., behaviors and morphology associated with courtship have remained remarkably stable, while the proteinaceous pheromones involved in the process have undergone extensive positive selection (Watts et al., 2004; Palmer et al., 2010; Wilburn et al., 2012). In the case of glycosylation, however, changes in the proteome are strongly correlated with the behavioral transition from dermal to olfactory delivery of pheromones, the only major phenotypic shift in plethodontid courtship behavior.

Inter- and Intraspecific Differences in Pheromone Sequence and Glycosylation Potential
Sequences from the northern population of $P$. cinereus suggest that the PRFs present in these populations have diversified from different isoform ancestors, although the ancestral population likely had multiple paralogs prior to speciation (Figure 3.2). These observations reveal a unique PRF profile that differs from its sister population by a degree that is equal in some cases to interspecific distances among other members of the ancestral clade. This is not unexpected in this species, one that has been used as a prime example of the challenges and limitations of species designations (Wilson, 1999). Nevertheless, this provides insight into the degree to which reproductive pheromones in these populations have diverged.

While most members of the ancestral DDC have the potential for glycosylated PRFs, the patterns of predicted glycosylation vary between species, both within- and among clades (Figure 3.5), but also notably between the two allopatric populations of $P$. cinereus (Figure 3.6). Populations of $P$. cinereus differ in the extent of $N$-linked glycosylation in the two primary (ancestral) N -linked sites (Figure 3.6). The NH population also includes an additional, independently-derived N -linked sequon in a PRF-

B isoform (site 4). Potential for O-linked glycosylation also varies between populations at several positions in the primary and tertiary structure (Figures 3.6 and 3.7). These predictions are supported by protein expression data (Figure 3.1), which suggest differences in isoform glycosylation patterns (in addition to known sequence divergence). Further proteomic analyses are needed to definitively identify fine-scale differences in post-translational modification between and within species.

## Phylogenetic Inferences

When exploring relatedness between PRF coding sequences (Figures 3.2 and 3.3), the ability for phylogenetic inference is challenged by the large degree of inter- and intraspecific diversity. This results in several polytomies in which node probabilities did not exceed 50\% consensus. This is particularly true in the PRFs of the derived ODC. This lack of consensus is consistent with difficulties in phylogenetic inference in this family in general, in which a large number of species diverged away from each other in a short period of time (Vieites et al., 2007; Wiens et al., 2006; Kozak et al., 2006; Highton et al., 2012). The positive selection acting on plethodontid pheromones only confounds this problem further (Watts et al., 2004; Palmer et al. 2010). Nevertheless, relationships between the major clades (DDC, IG, and ODC) and isoform lineages (i.e. PRF-A and -B) are relatively clear and allow for analysis of glycosylation potential from a phylogenetic perspective.

In examining the evolutionary history of N -linked sequons, the potential for glycosylation evolved multiple times in some lineages, but was lost in others (Figure 3.8). Specifically, N -linked sequons evolved in PRF at least three times in species outside of the derived olfactory clade (i.e. the DDC and IG): twice in PRF-A, and once in PRF-B isoforms. In contrast, this potential was lost over the course of the behavioral transition to olfactory delivery. The primary N -linked sites (1 and 2) are lost independently, at different stages in the intermediate group. Selection for glycosylation in the ancestral clade is evidenced by both (1) widespread conservation within the clade and (2) homoplasic motifs arising in different regions of the protein. Together this implies a functional benefit for glycosylation in these species.

## Structure-Function Relationships

Almost all of the observed sequence divergence among PRFs has occurred in loops or on surface residues. There are notable structural differences in PRF across clades, namely the extended N - and C-termini of the ancestral lineages. Importantly, there is also substantial variation within each species. Due to the extent of positive selection observed in this protein (Watts et al., 2004) and the degree of intraspecific variation, sexual selection may be acting on individuals to diversify their PRF repertoire, possibly as a molecular bet-hedging strategy to maximize female sensory stimulation.

Structural analyses of potential N- and O-linked glycosylation sites reveal key insights about their potential consequences for functionality. The primary (i.e. oldest and most common) N -linked sequons (Figure 3.8; sites 1 and 2) occur in the long loop between the first and second helices (Figure 3.9), while the alternative (i.e. derived and less common) sites (Figure 3.8; sites 3 and 4) occur on the opposite side of the protein, near the N -terminus (Figure 3.9). These sites have been confirmed to be glycosylated through proteomic analyses (Wilburn et al,. 2014a). The most likely candidate for Olinked glycosylation (S23) also occurs on the opposite side of the protein from the primary N -linked sites (Figure 3.9) and is one of the most widespread predictions for O linked glycosylation in the DDC (Figure 3.5). A closely-associated site (S29) scored higher on average than S23, but it is unclear whether this site is glycosylated in nature. That the sole proteomic observations are drawn from a single population of $P$. cinereus, however, yields little proof for the absence of glycosylation at this site in other populations and/or species, especially considering (1) the widespread degree of predicted inter- and intraspecific variation, (2) the apparent patterns of protein-level variation between $P$. cinereus populations, and (3) the potential for multiple glycoforms in IL-6type cytokines due to combinatorial glycosylation of closely-associated O-linked sites (Gross et al., 1989; Cruz et al., 2004).

The presence of glycosylation may affect interactions with ancestral receptors, due to their location in conserved receptor binding regions in IL-6-type cytokine homologs. The primary N-linked sites (N96 and N101) are associated with "Site I" binding with the alpha receptor, while the alternative N-linked sites (N149 and N156) are
associated with "Site II" binding with gp130. Interestingly, this is not true in the case of the interleukin-6. While human IL-6 is glycosylated, all sites occur well outside the binding regions and do not affect receptor interactions (Cruz et al., 2004). Thus, the occurrence of N -glycosylation in receptor binding regions appears to be an evolved feature in PRF.

## Functional Hypotheses

The degree to which glycosylation is correlated with dermal pheromone delivery suggests an adaptive function for PRF glycosylation in the ancestral state that does not hold true in the derived olfactory clade. As PRF has different targets between lineages, it may be that glycosylation was lost during this behavioral transition due to selective pressures from different modes of delivery and/or mechanisms of action. While the exact target of PRF in the DDC is unknown, the mental gland secretions are administered directly into the bloodstream and therefore could act as allohormone pheromones (sensu Wyatt, 2014b). Many serum proteins are glycosylated (Apweiler et al., 1999), as glycoproteins have higher solubility (Bohne-Lang and von der Lieth, 2005). Glycosylation is also known to provide protection from proteolysis (Dwek, 1996). Thus, glycosylation may be favored in the dermal-delivery mode because it prolongs the pheromone's lifespan in the bloodstream and in turn its effect on the female.

It is also notable that the two glycosylated areas of PRF occur in cytokine receptor binding regions, including those of different proteins in the receptor complex. The first and most widespread occurrence of PRF glycosylation targeted the IL-6-type alpha receptors ("Site I" binding), while the second occurrence targeted gp130 interactions ("Site II" binding). Both receptor proteins are also present in soluble forms in the bloodstream that bind to their ligands and either inhibit or facilitate IL-6-type signaling (Chalaris et al., 2011). This is referred to as "trans" IL-6-type signaling, in contrast to the standard membrane-bound receptors. Some soluble IL-6-type alpha receptors exert antagonistic effects (e.g. TNF $\alpha$ ), while others enhance signaling (e.g. IL-6R $\alpha$ ). Soluble gp130 has strictly antagonistic effects on trans signaling. Thus, PRF may have evolved glycosylation to elude inhibitory effects or modulate standard cytokine receptor
interactions by attaching glycans in the regions where these receptors should bind. These interactions would not be relevant in the PRFs of the ODC which target different tissues and may stimulate different receptors, such as general vomeronasal type-2 receptors in the VNO (Kiemnec-Tyburczy et al., 2011a).

An additional hypothesis is that the patterns of differential glycosylation may be important for species-specific reproductive signaling. This hypothesis assumes isoformspecific effects on receptor interactions in order to facilitate intraspecific courtships and avoid hybridization. Findings from other systems have shown that the presence of glycosylation can drastically alter the meaning of a signal (Dwek, 1996). Thus, applying glycans to different regions of PRF could serve to differentiate signals between species and/or target different receptor pathways. Additionally, even the same sequon can have different glycan moieties attached to it (Petrescu et al., 2004), further enhancing the number of possible glycoforms of a given isoform. The result in the case of PRF is that the use of post-translational modifications of this sort greatly increases the possibility for permutations in a signal that is already extremely variable. These hypotheses are not mutually-exclusive, however; it may be that glycosylation is adaptive in the ancestral state due to the benefits of increased solubility, proteolytic resistance, and/or modulating receptor interactions, but may also carry out their effects in a species-specific fashion as a function of the presence and patterns of glycosylation (perhaps in conjunction with other aspects of signal complexity).

## Putative Roles in Speciation

These observations of inter- and intraspecific pheromone variation provide insight into the importance of population-specific courtship signaling. As populations diverge into their own PRF repertoire this could lead to reinforcement, and ultimately cause character displacement of the signals (Smadja and Butlin, 2009). Many of the best examples of this phenomenon come from the chemical signaling realm, for example in moths and Drosophila (Wyatt, 2014a). Considering the documented complexity of courtship pheromone signals it is possible for these stimuli to act as premating isolating
barriers but functional tests demonstrating the efficacy of con- versus heterospecific pheromones are needed to confirm this hypothesis.

Currently, only one study has specifically tested the effect of heterospecific pheromones on courtship in plethodontids (Rollmann et al., 2003). This work tested $P$. shermani females exposed to the MG secretions from two different species and observed that heterospecific pheromones elicited a comparable reduction in courtship duration. However, there are several problems with this study. One limitation is that the experiment consisted of courtship between $P$. shermani pairs, meaning that the stimulus ignores the myriad other multimodal signals that could indicate an appropriate conspecific pair. A more general limitation is that the species tested are allopatric, whereas under a reinforcement model of speciation the greatest selective pressure for species recognition would be predicted in sympatry (Coyne and Orr, 1997; Reagan, 1992). More importantly, the only assay used to assess heterospecific pheromone effects was the duration of courtship, which is only included as data if a pair completes courtship. While this measure has been the basis for defining the function of PRF and PMF to date, the argument for its use as a proxy for fitness is tenuous. No data were presented for the rate of failures, which is a far more important variable in assessing the role of pheromones as premating barriers. Many other behavioral studies have implicated pheromones in species recognition and/or reproductive isolation (Dawley, 1984a; 1986ab; Verrell, 1989, 2003; Palmer and Houck, 2005), but these experiments tend to rely on association preferences to non-courtship signals, which is likewise not a direct indication of courtship success or evolutionary fitness.

## Glycosylated Pheromones in Vertebrates and Others

In general, the role of post-translational modification of pheromones has not been widely studied (Wyatt, 2014b). Nevertheless, a few other examples could be found from distantly-related organisms. Another notable glycoprotein pheromone is found in salamanders (including plethodontids): Sodefrin Precursor-like Factor (SPF) is also present in multiple glycosylated isoforms in the cloacal glands of some salamandrids (Van Bocxlaer et al., 2015). There is some evidence that this may be true in the MG of
some plethodontids (Doty et al., 2015; Figure 3.1). Salamanders are not alone, however, in employing glycoproteins as pheromones. Another vertebrate example can be found in the male-specific protein of Sarotherodon galilaeus (tilapia; Machnes et al., 2008). This 22 kDa lipocalin protein is produced solely by males and is expressed at higher levels by dominant males. Other examples include select major urinary proteins (MUPs) of mice and rats (Mechref et al., 1999; 2000; Perez-Miller et al., 2010; Cavaggioni and MucignatCaretta, 2000), and aphrodisin in hamsters (Briand et al., 2000). In all of these cases, the specific function of the attached glycans is not well understood.

Other evidence for a role of glycosylation in pheromone signaling comes from studies in non-chordate systems. Glycoprotein pheromones have been documented in algae (Hallmann, 2003), rotifers (Snell et al., 2009), mollusks (Cummins et al., 2004), and arthropods (Feng et al., 2015; Lonsdale et al., 1998). Perhaps the most applicable example of the role of glycosylation comes from the mate-recognition protein (MRP) of Brachionus rotifers (Snell et al., 2009). This glycoprotein occurs on the surface of females, and is assessed by males through receptors in their corona prior to mating. Deglycosylation leads to a decrease in mating, suggesting a crucial role for the glycans in signaling processes. Rotifers are able to detect the sex, geographic origin, and age of females, and exhibit a preference to mate with young, conspecific females. This ability for species discrimination has led to the hypothesis that MRP plays a critical role in reproductive isolation in rotifers (Snell et al., 1989). As always, more work is needed to deduce the exact role of post-translational modifications in the mechanisms of pheromone action.

## Conclusions and Future Directions

Overall, there is a large degree of both inter- and intraspecific signal diversity throughout the plethodontid phylogeny. This biochemical variation may in part explain the role of chemical signals in establishing and maintaining reproductive isolation. Taken together, interactions between the multiple facets of plethodontid signal complexity allow for an enormous communication landscape in which different evolutionary lineages may occupy different sensory niches. Whether species detect and act upon these differences in
natural mate-choice scenarios remains unclear. Future work should focus on (1) establishing additional behavioral assays that move beyond association experiments or the sole measure of courtship duration, (2) identifying the target tissues and receptors for the various pheromone components across delivery modes, (3) investigating the differential neurological and/or endocrine effects of the many levels of signal complexity, and (4) examining inter- and intraspecific effects of pheromone variation on these phenomena.

(Caption on the next page)

Figure 3.1: SDS-PAGE analysis of mental gland extract from three plethodontid lineages. Mental gland extract from three species are separated electrophoretically: Desmognathus ocoee (lane two), Plethodon shermani (lane three; L3), and two populations of P. cinereus (from VA and NH; L4 and L5 respectively). Desmognathus ocoee is an outgroup within the Plethodontidae that does not produce PRF. Plethodon shermani is a representative of the derived clade that delivers PRF to the olfactory system; $P$. cinereus is a representative of the ancestral clade, which delivers PRF via the dermis. L1 consists of a standard protein size ladder (M3913), while the far right standard (L6) alternates proteins with and without glycosylation (Candy Cane). Protein is stained with CBB (panel A; colored green), and glycans are stained with Pro-Q 488 (panel B; colored red). An overlay (panel C) demonstrates glycosylated proteins. L1 through L4 are adapted and modified from Wilburn et al. (2014a), and depict the same samples from that study alongside the additional (NH) population of $P$. cinereus. All samples were run in the same gel. The arrow on the left depicts the normal migration range for unglycosylated PRF at 22 kDa (i.e. in P. shermani). The white box demonstrates the comparison of the PRF range between $P$. shermani and the two localities of $P$. cinereus. The PRFs of $P$. shermani are not glycosylated; both populations of $P$. cinereus do have glycosylated PRFs, which don't migrate as far in the gel. Furthermore, the populations of $P$. cinereus appear to differ in the extent and pattern of glycosylation.


Figure 3.2: Bayesian consensus phylogeny of PRF sequences over the course of a behavioral transition from dermal to olfactory pheromone delivery. PRF-B isoforms are restricted to members of the ancestral, dermal-delivery clade, while PRF-A isoforms are present in all lineages. The PRF isoforms from the NH population of $P$. cinereus are generally more related to each other than other PRFs, forming mostly monophyletic groups throughout the tree. The speciose olfactory clade has been collapsed for easier viewing of the dermal-delivery clade and intermediate group. In addition, branch length data has been ignored for plotting; see Figure 3.3 for the full phylogeny with all PRF sequences and branch lengths. (This is the same phylogeny used in Figure 3.8.)


Figure 3.3: Bayesian consensus phylogeny for all known PRF sequences. Phylogenetic relationships between all known PRF sequences in Plethodon spp. Three replicates of Bayesian phylogenetic inference were conducted with MrBayes-3.2 for four million generations, sampling every thousand generations. The first quarter of samples were discarded as burn-in, yielding three thousand trees per replicate. Nodes with posterior probabilities less than 0.5 are collapsed into polytomies. The first divergence in PRF species occurs between PRF-A and -B isoforms, the latter of which are only present in the ancestral, dermaldelivery species. PRF-A isoforms of the dermal-delivery clade are less divergent compared to the larger differences in PRF-A that are observed starting in members of the intermediate group. The PRF-A isoforms of the olfactory-delivery clade all diverge from those in the intermediate group, but largescale relationships of PRFs within this clade are not well-resolved.

(Caption on the next page)

Figure 3.4: Similarity matrices for primary and tertiary structure of $P$. cinereus PRFs. (A) Similarity matrix depicting the hierarchical clustering of pairwise percent identity (ID Score) of PRFs from the NH and VA populations of $P$. cinereus. (B) An analogous similarity matrix comparing three-dimensional alignments of PRF structural models constructed in I-TASSER. Structural alignments utilize the pairwise TM-scores as a value for comparing structural similarity. There is notable sequence similarity between PRF-A and -B isoforms.


Figure 3.5: Annotated sequence alignment for all known PRFs. (A) PRF sequences were aligned and annotated for basic sequence features (signal peptides, gaps, stop codons) in addition to N - and O -linked glycosylation potential. Residues predicted to be glycosylated are shown as either red ( N -linked) or dark blue (O-linked); sites predicted to be unoccupied are shown as orange (N-linked) or light blue (O-linked). Secondary structure is presented as (B) binary predictions from a PROMALS3D alignment, and (C) confidence scores from Jpred4. (D) Relative solvent accessibility predictions from Jpred4 are shown from completely buried (low) to completely accessible residues (high). The presence of N -linked glycosylation is almost exclusively limited to the ancestral, dermaldelivery clade. Intraspecific diversity in predicted PRF glycosylation patterns can also be observed within this clade and in early-diverging members of the intermediate group.


Figure 3.6: Annotated sequence alignment for two populations of $\boldsymbol{P}$. cinereus. This figure depicts the same information as Figure 3.5A, but focuses solely on the two populations of $P$. cinereus to demonstrate intraspecific differences in glycosylation potential. PRF-A isoforms between populations are predicted to differ in the degree of N linked glycosylation. A single PRF-B isoform of the NH population includes an additional, independently-derived sequon (site 4; N156) not present in the VA population. The number of isoforms exhibiting different patterns of O-linked glycosylation differs between populations at the N -terminus. The NH population also has stronger O-linked potential in the sequence center and at the C-terminus (see Figure 3.7).


Figure 3.7: Potential for $\mathbf{O}$-linked glycosylation between $\boldsymbol{P}$. cinereus PRF isoforms. Each plot depicts the O-linked glycosylation prediction score (O-linked potential from 0 - 1) at each residue along the length of the PRF alignment. Shaded barplots in the background also show relative solvent accessibility values for each residue (as in Figure 3.5D), from fully buried (0) to fully exposed (1). Four comparisons between O-linked potential are shown: differences between populations for both PRF-A and -B isoforms (A, C), and differences between isoforms for each population (B, D). Asterisks indicate statistical significance between O-linked potential at each residue for the comparison made in that panel. Significance was tested by residue-specific Welch’s t-tests, necessary due to the observation of unequal variance.


Figure 3.8: Phylogeny of PRF sequences for the three main lineages of Plethodon depicting the presence of $\mathbf{N}$-linked sequons and their predicted glycosylation. The same PRF phylogeny is shown for all four N -linked sequons. (For a version with all sequence names see Figure 3.2.) The presence of the sequon in each PRF isoform is depicted as colored branches depending on whether the site is predicted to be glycosylated (red) or not (orange) in that sequence (as in Figure 3.5). The primary (oldest) N -linked sites (1 and 2) are inherited from the ancestral PRF-A isoforms of the dermal-delivery clade, but are lost at different stages during the transition to olfactory pheromone delivery. In contrast, the alternative N-linked sequons (3 and 4) evolved independently in PRF-A isoforms of the intermediate group (site 3) and PRF-B isoforms of the dermal-delivery clade (site 4).


Figure 3.9: Representative structural model of PRF from a dermal-delivery species ( $P$. cinereus). The top I-TASSER model for VA-Pcin PRF-A1 is shown with the Nterminus on the left of the protein (left models), and rotated $180^{\circ}$ in the vertical axis (right models). This isoform was chosen as a representative model because it is a fulllength PRF. The top panels depict the backbone ribbon only, while the bottom panels include an overlay of the surface view. The color key portrays the degree of sequence conservation across all known PRF primary structures. Key glycosylation sites are shown with side-chains and identified, while all other side-chains are hidden from view. Also hidden from view are the side-chains of the alternative N -linked sites (N130/149 and N137/156), which do not contain asparagines at these positions in this isoform. Residue numbers are provided first as the position in the model (i.e. without a signal peptide), followed in parentheses by the full alignment position used throughout the manuscript.

(Caption on the next page)

Figure 3.10: Representative structural model of $\mathbf{N}$-linked glycan attachment at the primary N-linked sequons. The best I-TASSER model for PRF-VA-A6 is shown with glycans attached to one (A) or both (B) of the primary N-linked sequons. Glycans were attached in silico via GlyProt. This model was chosen because it contained both primary N -linked sequons, but yielded fewer steric interactions in the glycans added in an imperfect in silico process. The glycan structure at these sites is unknown, and therefore those selected for visualization are arbitrary. Glycans were identified by searching the GlyProt database for known structures of the appropriate size ( 30 sugar residues per N linked site) as estimated by Wilburn et al. (2014a). Only two related glycans of this size existed in the database and were used accordingly (GlyProt structures 9232 and 9233; LinucsID 16879 and 16880). The glycans are colored different shades of gray to differentiate them from one another and the PRF protein (white). Importantly, these models do not include the likely O-linked glycan occurring at the opposite side of the protein (near the N-terminus; Figure 3.9), which should also possess a glycan approximately half the size of the N -linked glycans (i.e. 15 sugar residues). These models are therefore an underestimate of the size of the most heavily-glycosylated PRF isoforms.


Figure 3.11: Structural model of a hypothetical ancestral PRF receptor interaction. A representative PRF model was structurally aligned to the IL-6 ligand in this known IL6 -type cytokine receptor structure (PDB: 1P9M; Boulanger et al., 2003). This yielded a close overlay of the cytokine fold, after which IL-6 was removed from view. The upper panel (A) shows the "top view" of the receptor, as if looking down toward the cell membrane. The middle panel $(\mathrm{B})$ is then rotated $90^{\circ}$ to reveal a side view, parallel to the membrane. Lastly, the bottom panel (C) is rotated another $90^{\circ}$ to show the "bottom view" of the complex, as if looking up from the membrane. PRF is shown in light blue, with key glycosylation sites (the same as in Figure 3.9) colored accordingly for N-linked (red) and O-linked (blue) sites. All receptor domains are labeled for both IL-6R $\alpha$ (ribbon colored dark gray) and gp130 (colored light gray). Solid wedges indicate domains that are in front of others, while domains in the background are labeled with dashed wedges. Dashed circles outline the conserved "Site I," "Site II," and "Site III" cytokine receptor binding regions that are visible in each view. Labels for each glycosylation site are present in Figure 3.12, and Figure 3.9 provides a detailed view of the ligand with the same sites labeled. The primary N-linked sites occur in the "Site I" binding region between IL-6R $\alpha$ D2 and D3, while the alternative N -linked sites occur in "Site II" binding regions between gp130 D2 and D3.


Figure 3.12: Comparison of glycosylation sites of interleukin-6 (IL-6) and PRF in the context of receptor binding. Left models depict the IL-6R $\alpha / \mathrm{gp} 130$ complex with the IL-6 ligand; right models show PRF in the place of IL-6 (as in Figure 3.11). The IL-6 models include labels for all receptor domains; PRF models are labeled only for individual glycosylation sites. All orientations, colors, and labels are identical to Figure 3.11. Glycosylation sites are indicated with dashed circles; the three binding regions are reiterated for the side view of PRF with solid circles. IL-6 contains one N -linked site located, oriented into the center of the complex; it also has three O-linked sites on the top of the protein, oriented outward from the receptor complex. None of these sites are in the proximity of receptor binding regions, nor affect receptor binding (Cruz et al., 2004). In contrast, PRF has evolved glycosylation sites in two key binding regions: the primary N linked sites (N96 and N101) are situated in the "Site I" binding region, while the alternative N-linked sites are situated in the "Site II" binding region.

## CHAPTER FOUR

# DE NOVO TRANSCRIPTOME ASSEMBLY OF CHEMICAL SIGNALING GLANDS IN A PLETHODONTID SALAMANDER 

Chouinard, A.J.* ${ }^{1}$, Wilburn, D.B. ${ }^{2}$, Eddy, S.L. ${ }^{1}$, and Weis, V. ${ }^{1}$<br>1. Department of Integrative Biology, Oregon State University, Corvallis, OR 97331<br>2. Department of Biochemistry, University of Louisville, Louisville, KY 40202


#### Abstract

For many years, plethodontid salamanders have served as a useful system for studying the role of chemical signals in facilitating inter- and intraspecific behaviors, including numerous social and reproductive interactions. Biochemical analyses of the molecules involved in these phenomena have almost exclusively focused on a single courtship gland: the mental gland. Yet other regions of the body are known to be important for a wide variety of information transfer associated with more general communication. This research provides the first in-depth analyses of the proteins involved in communication via three signaling glands: the cloacal glands, postcloacal gland, and the dorsal tail base. These tissues have been implicated in different aspects of reproductive and territorial behaviors. High-throughput RNA sequencing was performed for these tissues (from Plethodon shermani) and a single reference transcriptome was assembled de novo. Annotation of the resulting transcripts revealed numerous secretory proteins that may be utilized for a wide variety of behaviors, including: known pheromones, carbohydrate-binding proteins, toxins/venoms, reproductive and hormonelike proteins, protease inhibitors, and others. Mapping reads from each sample to the reference provides a preliminary look at how these glands differ in gene expression. This research is the first to document this rich landscape in a detailed manner, but more work is needed to determine proteomic expression and the behavioral roles of the proteins identified.


## Introduction

The skin of amphibians is important for many essential functions, such as respiration, osmoregulation, predator defense, protection from pathogens, and communication (Duellman and Trueb, 1994; Clarke, 1997). Despite a long history of study, researchers have only begun to scratch the surface of the vast array of chemicals produced in the skin to facilitate these functions. The primary goal of the present study was to use high-throughput sequencing technologies to explore and document the proteins responsible for this rich and largely-uncharted chemical landscape. The specific focus of this research is the communication system of salamanders, but the large number of novel genes identified in non-model transcriptome studies will simultaneously provide data for other investigations.

## Chemical Communication in Plethodontid Salamanders

Perhaps the most well-studied amphibian chemical signaling system is that of plethodontid (lungless) salamanders, particularly those of the genus Plethodon in the eastern United States. Chemical signals play the primary sensory role, and facilitate a wide variety of behavioral functions, such as: prey detection (Placyk and Graves, 2002), predator avoidance (Murray and Jenkins, 1998; Maerz et al., 2001; Sullivan et al., 2001), homing (Madison, 1969; Madison and Shoop, 1970; Kleeberger and Werner, 1982; Jaeger et al., 1993), and diverse intraspecific interactions. Key among these is the use of chemical signals as territorial advertisements. These "scent marks" modulate agonistic interactions between males (Jaeger, 1986; Simons et al., 1994; 1997) and are also likely used in mate-assessment (Walls et al., 1989; Mathis, 1990; Chouinard, 2012). Pheromones are also used directly in courtship and play an important role in facilitating reproduction (Arnold, 1977; Rollmann et al., 1999; Wilburn et al., 2015). Consequently, these signals may serve as reproductive isolating barriers (Dawley, 1986ab) and therefore also have significance in the evolution of this speciose and rapidly-radiating clade (Wiens et al, 2006).

Plethodontids produce a wide assortment of chemicals from various glandular regions of the body, and these secretions presumably communicate different aspects of
information and/or facilitate different functions. Studies of Plethodon spp. suggest that chemical signals can convey a complex suite of information, such as: species identity (Jaeger and Gergits, 1979; Dawley, 1984a; 1986b; Verrell, 1989; 2003; Palmer and Houck, 2005), sex (Jaeger and Gergits, 1979; Dawley, 1984a; Karuzas et al., 2004; Page and Jaeger, 2004; Dantzer and Jaeger, 2007a; Dalton and Mathis, 2014), size (Mathis, 1990; Marco et al., 1998), reproductive status (Marco et al., 1998; Dantzer and Jaeger, 2007b), parasite-load (Maksimowich and Mathis, 2001; Dalton and Mathis, 2014), dietquality (Walls et al., 1989; Chouinard, 2012), and individual identity (Madison, 1975; Tristram, 1977; McGavin, 1978; Jaeger, 1981; Dawley, 1984a; Simon and Madison, 1984; Jaeger et al., 1986; Anthony, 1993;). Not all of these capabilities are documented in each species, but the extent and diversity of evidence from multiple taxa indicate a large reliance on chemical signaling for inter- and intraspecific communication.

## Regions of Chemical Signal Production

Despite the large number of experiments illuminating the behavioral ecology of plethodontid chemical signaling, relatively little is understood about the molecules responsible for these behaviors. The degree of information conveyed suggests that the composition of these signals is highly complex, and likely to vary by glandular location. Salamanders have regions of the skin that take on different macroscopic anatomies and serve different behavioral functions (Woodley, 2014; von Byern et al., 2015). At the tissue level, amphibian skin contains three main gland types: serous (or granular), mucous, and lipid glands (Duellman and Trueb, 1994). The skin in select parts of the body can vary in the composition of these gland types (Simons and Felgenhauer, 1992; Hecker et al., 2003), often to apparently fulfill one or more specific functions (e.g. courtship). These modified body regions are also referred to as glands. Due to overlapping terminology, the tissue-level glands will be referenced by their specific type (e.g. serous/granular, mucus, lipid) to differentiate them from the regions of skin that compose the larger-scale, organismal-level glands. Histochemical analyses reveal that these regions stain for different chemical compositions (Hecker et al., 2003), and thus
appear to specialize in the production of distinct chemical signals or other secretory products.

The most studied plethodontid gland in regard to its biochemical composition is the mental gland (MG) on the underside of the chin in males (Sever, 1976; 2003). This gland is used directly in courtship (Arnold, 1977), and perhaps also in territorial advertisement through "chin tapping/scraping" (Jaeger, 1986). Much attention has been paid to the production and evolution of MG pheromones (reviewed in Woodley, 2010; 2015). Recent research has utilized high-throughput RNA sequencing to characterize the MG transcriptome (Wilburn, 2014). In contrast, the biochemical composition of other signaling glands has been largely unstudied at a fine scale (but see Hecker et al., 2003; Largen and Woodley, 2008), despite their known function in inter- and intraspecific communication. Thus, this research will investigate three signaling glands in plethodontids: the postcloacal gland (PCG), the dorsal tail base (DTB), and the cloacal glands (CG).

The PCG occurs on the ventral surface of the tail, starting posterior to the cloaca. It is comprised of both mucous and "S1" serous glands (sensu Hecker et al., 2003). Mucous glands are largely basophilic, with carbohydrates a dominant constituent; the secretory granules of S 1 serous glands are mildly acidophilic, containing neutral carbohydrates along with proteinaceous components (Hecker et al. 2003). These glands can be more hypertrophied in males than females for at least some species (Staub and Paladin, 1997), but not all (Largen and Woodley, 2008). Simons and Felgenhauer (1992) were the first to hypothesize that this region is responsible for territorial scent marks through the characteristic "postcloacal press" behavior, and this has been confirmed by subsequent experiments (Jaeger and Gabor, 1993; Simons et al., 1994; 1999; Largen and Woodley, 2008). Thus, much of the information contained in territorial advertisements is hypothesized to be conveyed through the secretions of the PCG.

On the other side of the tail, the DTB is also hypothesized to play a role in courtship (Sever, 2003; Sever and Siegel, 2015), although that role is poorly understood. This hypothesis is based on the observation that the snout of females spends substantial time in close proximity to the DTB during the "tail-straddling walk" courtship ritual (for
a detailed description of courtship, see Arnold, 1976; 1977). Some salamanders from other families also have elaborate structures (tubercles) in this region with enlarged mucous and serous glands (Sever, 2003). In plethodontids, the DTB is comprised of both mucous glands and "S2" serous glands (sensu Hecker et al., 2003). In contrast to S1 glands in the PCG, S2 glands of the DTB are highly proteinaceous. In addition to a putative courtship role for the DTB, this tissue may be involved in nutrient storage and predator defense through the production of sticky, noxious secretions in some species (Largen and Woodley, 2008; von Byern et al., 2015).

Lastly, the CGs are sexually-dimorphic and important for reproduction (Sever, 2003). Sever (1978a) defined five types of exocrine glands surrounding the male cloacal chamber: the pelvic, dorsal, ventral, Kingsbury, and vent glands. These glands are often hypertrophied during the breeding season (Sever, 2003) and used for spermatophore production (Sever, 2014). Females can possess up to three glands (depending on the species), including the spermatheca, the sperm-storage organ (Sever, 1978b; Sever, 2003). In males of many species, the dorsal and vent glands secrete onto the epidermis at the cloacal opening (either as the sole opening or in addition to secretion into the cloacal chamber). The vent glands are the only case in which they are equally-developed all year long (Sever, 1978a), also suggesting a more general function than reproduction alone. When integrating anatomy with behavioral observations, these two cloacal glands are implicated in pheromone production in plethodontids, including a possible role in territorial advertisement (Simon and Madison, 1984; Jaeger, 1984; 1986; Sever, 2003). This hypothesis is consistent with the function of these glands in other families in which the dorsal cloacal glands (also called "abdominal glands") secrete potent mate attractant pheromones (Kikuyama et al., 1995; Yamamoto et al., 1996; 2000; Van Bocxlaer et al., 2015; Maex et al., 2016). In this study, all cloacal glands were investigated as a part of the same sample.

## Known Plethodontid Pheromones

To date, three plethodontid pheromones have received the most study, all of which are produced in the MG. Plethodontid Receptivity Factor (PRF) is a 22 kDa
protein related to interleukin-6 (IL-6) -type cytokines (Rollmann et al., 1999; Watts et al., 2004), while Plethodontid Modulating Factor (PMF) is a 7 kDa protein of the three-finger protein superfamily (Palmer et al., 2010; Wilburn et al., 2012). Both proteins are present in numerous isoforms and together compose $>80 \%$ of MG secretions in $P$. shermani (Chouinard et al., 2013). When males apply these MG secretions to females, the effect is a reduction in the duration of courtship (Rollmann et al., 1999; Wilburn et al., 2015). Another known courtship pheromone, Sodefrin Precursor-like Factor (SPF), is a 20 kDa protein related to phospholipase A2 inhibitors (Doty et al., 2015; Van Bocxlaer et al., 2015), also derived from three-finger proteins (Janssenswillen et al., 2014). SPF mRNA is present throughout the plethodontids, but is only significantly expressed as protein in early-diverging plethodontids (e.g. Desmognathus); it appears to have been largely replaced by PRF in Plethodon in regard to MG expression (Palmer et al., 2007a). SPF is also present in the cloacal glands of salamandrids (Van Bocxlaer et al., 2015) and ambystomatids (Maex et al., 2016). In Cynops newts, the protein is post-translationally cleaved into a potent mate attractant, Sodefrin, the first vertebrate protein pheromone identified (Kikuyama et al., 1995; 2002). In other salamandrids, cleavage was not necessary to induce female following behavior (Van Bocxlaer et al., 2015); this use of full-length SPF is consistent with its expression in plethodontids (Doty et al., 2015).

Recently, proteomic and molecular studies have identified additional putative pheromones secreted by the MG of select Plethodon spp. Among these are homologs of: a tissue inhibitor of metalloproteinase (TIMP 1) termed Plethodontid TIMP-like Protein (PTP; previously referred to as C3 in Chouinard et al., 2013), a cysteine-rich secretory (CRISP-like) protein, and a poliovirus receptor-related protein (PRRP; Wilburn, 2014; Wilburn et al., 2014a). Sequencing experiments in plethodontids have also identified several hormone-like transcripts (e.g. insulin-, natriuretic-, and glucagon-like peptides, relaxin, vasoactive intestinal peptide, leptin; Kiemnec-Tyburczy et al., 2009; Wilburn, 2014; Doty et al., 2015). Taken together, these findings implicate another example of a secreted MG toxin homolog, a role for protease inhibitors, and the use of hormone-like proteins. All of the known (and putative) plethodontid pheromones documented to date provide a starting point for exploring the transcriptomes of the PCG, DTB, and CG. In
addition, bioinformatic tools will be used to identify novel pheromone candidates and/or other secretory proteins of interest.

## Goals of the Present Study

The specific goals of the present study were to construct de novo transcriptomes for the (1) cloacal glands, (2) postcloacal gland, and (3) dorsal tail base. These will serve as a reference for future proteomic work to explore differences in gland expression and function. In addition, bioinformatic analyses were used to: (1) identify all putative secretory proteins (the secretome) that may be involved in communication and/or other key behavioral functions, and (2) compare the differential expression of transcripts between gland tissues. As a precursor to subsequent experimental approaches, these analyses will serve as a tool for generating hypotheses for how these glands may differ in the production of putative pheromones, reproductive proteins, or other genes important for the behavioral ecology and evolution of salamanders.

## Methods

## Animal Collection and Gland Removal

Male salamanders $(\mathrm{n}=5)$ were collected from Macon County, NC $\left(35^{\circ} 10^{\prime} 38^{\prime \prime} \mathrm{N}\right.$, $83^{\circ} 33^{\prime} 47$ "W) in August, 2010. Animals were shipped to Oregon State University and housed in the lab in clear plastic containers ( $29 \times 14 \times 9 \mathrm{~cm}$ ) at $16^{\circ} \mathrm{C}$ on a $14: 10$ light:dark cycle (OSU IACUC 3007, 4053). Individuals were acclimated to lab conditions for one month prior to gland removal. Males were anesthetized in 7\% diethyl ether and glandular tissues were surgically excised. The cloacal gland (CG) sample contained all identifiable glandular tissue surrounding the cloacal chamber (Sever, 1978a; Sever, 2014). The postcloacal gland (PCG) sample consisted of the ventral surface of the tail (1 cm) starting immediately posterior to the cloaca (Simons et al., 1999; Hecker et al., 2003; Largen and Woodley, 2008; Chouinard, 2010). Tissue for the dorsal tail base (DTB) sample ( 1 cm ) was taken in the same region on the opposite (dorsal) side of the tail, but starting immediately above the cloaca (Sever, 1989; Hecker et al., 2003; Largen and Woodley, 2008; Sever and Siegel, 2015).

## RNA Extraction and Sequencing

Following surgery, tissues were physically homogenized; RNA was extracted using the RNeasy kit (Qiagen) and treated with DNaseI (Ambion) according to the manufacturer's instructions. RNA concentrations were estimated via absorbance at 260 nm on a NanoDrop ND-1000, and $1 \mu \mathrm{~g}$ of RNA from each of the five males was pooled ( $5 \mu \mathrm{~g}$ per gland). Pooled samples were further purified with phenol:chloroform:isoamyl alcohol (PCI) extraction followed by ethanol precipitation (EP). The SMARTer cDNA synthesis kit (Clontech) was used to prepare cDNA, which was shipped to Otogenetics (Atlanta, GA) for sequencing. Paired-end high-throughput sequencing was completed on an Illumina HiSeq2000 with 100bp read length pairs.

## Bioinformatic Analyses

Multiple parameters were used for each step in de novo transcriptome assembly. After each permutation, the quality of the resulting transcriptome was assessed via several metrics (described below). The final assembly selected was created using the following protocol.

Reads were de-duplicated and error-corrected using BBMap v35.82 (Bushnell, 2015) under default parameters. Cutadapt v1.9.1 (Martin, 2011) was used to trim low quality reads from the 5' and 3 ' ends (Phred $<5$ ), and remove (clip) sequencing adapters. Standard Illumina adapters as well as those introduced by the SMARTer cDNA kit were clipped with a $10 \%$ error rate. All reads were retained after trimming and clipping in order to maintain any single reads from broken pairs (i.e. valid reads whose mate-pairs were invalid). After trimming/clipping, a custom script was used to discard any reads under 25 bp (the kmer length used during assembly), while maintaining mate-pair information and retaining reads from broken pairs. Read quality metrics were consulted before and after trimming/clipping using FastQC v0.11.2 (Andrews, 2010).

All remaining reads were concatenated and assembled with Trinity v2.1.1 (Hass et al., 2013; Grabherr et al., 2011) using the "-run_as_paired" option. This allowed both paired and single reads to be used in the assembly. Only three other options were used in addition to default Trinity parameters ("-min_glue 1 -group_pairs_distance 800 -
path_reinforcement_distance 25"). The minimum contig length to include in the assembly was 200 bp . Basic assembly metrics were obtained using Trinity utilities and custom scripts (see Appendix Three).

Assemblies were assessed for structural quality via TransRate version (v) 1.0.1 (Smith-Unna et al., 2015) and for transcriptome completeness with BUSCO v1.1b1 (Simão et al., 2015). TransRate is a reference-free method that uses read sequences and pairing information to evaluate the quality of contigs in regard to known assembly artifacts, producing an overall assembly score. BUSCO identifies transcripts that are homologous to a clade-specific database of single-copy orthologs (the Vertebrata dataset in this case) and evaluates the length of identified orthologs as a proxy for transcriptome completeness. This yields the proportion of complete, partial, and missing single-copy orthologs expected.

Once a final assembly was selected, transcript abundance for each sample was estimated using kallisto v0.42.4 (Bray et al., 2015) through the abundance estimation pipeline in the Trinity suite. Kallisto uses a kmer-based method ("pseudo-alignment") that is both highly accurate and computationally efficient. The sequencing methods employed do not account for individual variation, and subsequent differential expression analyses should therefore be considered exploratory in nature. Differential expression was analyzed in edgeR (Robinson et al., 2010) on TMM cross-sample normalized expression values (Dillies et al., 2013), also run with tools built into the Trinity suite. Due to the lack of sequencing replicates, a highly conservative fixed dispersion estimate (0.3) was used for differential expression. Six clusters of differentially-expressed transcripts were defined after visual inspection of expression profiles.

The final transcriptome was annotated using the Trinotate program (v2.0.2) from the Trinity authors. Briefly, multiple methods were combined for functional annotation: TransDecoder v2.0.1 (part of the Trinity suite), SignalP v4.1 (Petersen et al., 2011), tmHMM v2.0 (Krogh et al., 2001), BLAST+ v2.3.0 (Altschul et al., 1990), Pfam (Bateman et al., 2004), HMMER v3.1b1 (Finn et al., 2011), and GO (Ashburner et al., 2000). GO enrichment for each expression cluster was performed with GOSeq (Young et al., 2010), also run via Trinity utilities. All BLAST searches were run on both the Swiss-

Prot database (manually-curated; fewer sequences but high confidence in annotations), as well as the UniRef90 database (more sequences but less confidence in annotations).

To identify putative secretory proteins, transcript annotations were extracted if (1) they contained an open reading frame (ORF) with a predicted signal peptide (via SignalP v4.1 under default parameters), or (2) any BLAST hits yielded GO (cellular component) terms for either "extracellular region" (GO:0005576) or "extracellular space" (GO:0005615). To explore the most highly-expressed secretory proteins, identified transcripts were extracted from the assembly and clustered with CD-HIT-EST (Li et al., 2006) at $80 \%$ sequence identity with a shortened word size ("-n 4") in the most accurate mode ("-g 1"). All cross-sample normalized expression values (TMM) were totaled per cluster, and clusters were sorted according to the highest sum expression. In a similar but more concise approach, unique BLAST hits were extracted and filtered for uncharacterized proteins; the total TMM values for each unique hit was tallied and hits were sorted in descending order of total expression (available in Appendix Three: Table A3.1). To aid in reproducibility, all custom scripts used for each step in these analyses are also provided therein.

## Results

## De Novo Transcriptome Assembly

After parameter optimization, the selected assembly was built with read deduplication, error-correction, aggressive adapter clipping and lenient quality trimming (PHRED < 5). This protocol is consistent with suggestions from recent empirical analyses of optimal de novo transcriptome assembly (MacManes, 2015). The processing of data in this manner resulted in $66,567,678$ reads to be used for assembly, of which 54,191,146 remained in valid pairs (Table 4.1; Figure A3.1).

The resulting assembly contained 73,928 "genes" and 94,221 "transcripts" (Table 4.2). This output follows the Trinity nomenclature, in which each assembled contig is labeled as a "transcript" and a "gene" refers to a cluster of related contigs. While the goal for this output is to represent alternative transcripts per true locus, in reality it is an estimate. As the transcripts per gene can include redundant/alternative sequences and
therefore artificially inflate length statistics, the longest transcript for each gene was used to calculate assembly length metrics. With this input, the assembly had a mean length of 509.95 bp , a median of 330 bp , and an N50 of 667 bp . To obtain a more meaningful length statistic (MacManes, 2015), the N50 for transcripts comprising 90\% of expression was calculated ("E90N50" sensu Trinity documentation). The majority (90\%) of expression consisted of 12,541 transcripts with an E90N50 of 1,319 bp.

The assembly obtained a raw TransRate (TR) score of 0.10 ; inspection of the distribution of individual contig scores (Figure A3.2) revealed that this value is skewed by a large number of poorly-supported contigs. After optimization (built into TR), 74,041 contigs scored above the threshold (0.03295), resulting in an optimal TR assembly score of 0.20 . The raw TR score is lower than the suggested cutoff ( 0.22 ), which places the structural quality of the assembly below the median of de novo transcriptomes submitted to the NCBI Transcriptome Shotgun Archive (Smith-Unna et al., 2015). However, higher TR scores could be obtained with other assembly parameters, but these assemblies came at the expense of substantially decreased BUSCO scores. As the goal of the study was to identify novel sequences, assembly completeness was valued over structural integrity, which is indicated by the raw TR score (and affected by inclusion of the many lowscoring contigs). The final assembly identified $23.78 \%$ of vertebrate BUSCO sequences (BUSCOs), with $14.89 \%$ of these BUSCOs found as full-length. The number of BUSCOs identified is also low, but not unexpected for de novo transcriptomes of non-model organisms.

## Transcriptome and Secretome Annotation

Each transcript was annotated using BLASTx to identify possible homologs, resulting in 19,055 (Swiss-Prot; 20.22\%) and 22,730 (UniRef90; 24.12\%) transcripts with strongly significant (e-value $<0.00001$ ) BLAST hits (after grouping hits per transcript). These transcripts also encoded 24,074 predicted ORFs, which were translated and annotated via BLASTp. This resulted in 11,308 (Swiss-Prot; 46.97\%) and 13,081 (UniRef90; 54.34\%) ORFs with significant BLAST hits (after grouping hits per ORF). The majority of BLAST hits did not constitute full-length transcripts, although many
were recovered for each method and database (Figure 4.1). These BLAST results identified 14,060 unique GO terms. As some transcripts could contain multiple likely ORFs, the total number of annotation entries was 99,454 (for 94,221 transcripts).

The identification of putative secretory proteins yielded 2,461 transcripts. After excluding uncharacterized proteins, this consisted of a secretome with 1,794 unique BLAST hits. For a full list of BLAST results for secreted proteins (sorted by total expression) see Appendix Three (Table A3.1). Briefly, these secretory transcripts included: carbohydrate binding and associated proteins, known pheromones, protease inhibitors, toxins/venoms, hormone-like proteins, reproductive proteins, immune-like proteins, lipid binding and associated proteins, proteases; antimicrobial and antifungal proteins, and others (Table 4.3).

## Differential Expression of Transcripts

Conservative differential expression analyses resulted in 1,355 differentiallyexpressed (DE) transcripts between the three samples (Figure 4.2). When analyzing the data by genes, 897 were differentially-expressed (though all downstream analyses included all transcripts). These DE transcripts were categorized into six distinct clusters after visual inspection of the heatmap (Figures 4.2 and 4.3). These clusters captured the broadest categories of differential expression: high CG but lower PCG/DTB (Clusters 1, 3), intermediate CG/PCG but lower DTB (Cluster 6), higher PCG but lower CG/DTB (Cluster 2), higher DTB but lower CG/PCG (Cluster 5), and higher PCG/DTB but lower CG (Cluster 4). All clusters contained large numbers of unidentified proteins that yielded no significant annotations. From those that did, relevant secretory proteins were identified for each cluster (Table 4.4), as were significantly enriched GO terms for all DE transcripts per cluster (false discovery rate $<0.05$; Table 4.5).

## Known Pheromone Identification and Expression

Among the many identified secretory proteins were several pheromone sequences from Plethodon spp. (in descending order of total expression): PMF, SPF, PRF, PRRP, PTP, and CRISP. These sequences were expressed at different levels in the three glands
sampled (Figure 4.4; Table 4.4). PMF was expressed most strongly in the CG and was identified as a member of DE clusters 3 and 1 (high CG). In contrast, SPF was most expressed in the PCG and DTB. Different SPF sequences were upregulated in different tissues, resulting in the inclusion of SPF variants in DE clusters 3 and 1 (high CG), 2 (high PCG), 5 (high DTB), and 4 (high PCG/DTB). PRF was most highly expressed in the CG, and was a member of DE cluster 3 (high CG). PRRP was also most highly expressed in the CG and a member of DE cluster 3 (high CG). Lastly, plethodontid CRISP-like factor was most expressed in the PCG and DTB, and was therefore a member of DE cluster 4 (high PCG/DTB).

In addition to the above results, which were restricted to sequences identified from Plethodon (Figure 4.4), many other BLAST hits implied possible homology to plethodontid pheromones. Several three-finger protein domains were identified (e.g. phospholipase A2 inhibitors, Ly6, u-PAR, PLAUR domains, Prod 1, Xenoxin, RoBo-1), which may indicate homology to either PMF or SPF. Many of these sequences were differentially-expressed, including members of DE clusters 3 (high CG), 2 (high PCG), and 4 (high PCG/DTB). Likewise, numerous (non-Plethodon) CRISP homologs were identified from each gland, including members of DE clusters 3 and 1 (high CG), and 5 (high DTB).

Transcripts encoding putative pheromone receptors were also assembled. Plethodontid interleukin-6 receptor beta (IL-6RB) sequences were expressed at low levels in all three glands. Vomeronasal receptor (VR) types I and II (with top hits from multiple vertebrates) were expressed at higher levels, mostly in the CG and PCG. Additional receptors of interest included (among others): olfactory receptors, growth hormone receptors, interleukin-1 and -17 receptors, tumor necrosis factor receptors ( $1 \mathrm{~A}, 5$, and 18), relaxin receptors, and prostaglandin F2-alpha receptor.

## Discussion

In his review of chemical signaling and territoriality in plethodontids, Jaeger (1986) identified the need for rigorous analyses of the chemicals produced by several glands implicated in inter- and intraspecific communication. At the time, behavioral
experiments had begun to reveal the importance of pheromonal scent marks (and other signals), but essentially nothing was known about the molecules involved. Since then, a great deal has been learned about the biochemistry, expression, and evolution of a few key proteins produced by the mental gland and involved in reproduction (Rollmann et al., 1999; Watts et al., 2004; Palmer et al., 2005; 2007ab; 2010; Wilburn et al., 2012; Chouinard et al., 2013; Wilburn et al., 2015). Despite this surge in exciting molecular research, little attention has been paid to the other glands long known to be important for the copious and diverse social functions of plethodontid pheromones. With the exception of a few studies (Hecker et al., 2003; Largen and Woodley, 2008), our understanding of the chemicals produced by other signaling glands in these species remains practically as uncharted as in Jaeger's day.

The results of the present study provide the first in-depth exploration of gene identity and expression in three important plethodontid pheromone glands: the cloacal glands (CG), postcloacal gland (PCG), and dorsal tail base (DTB). In exploring the transcriptomes of these three glands, this research (1) corroborated previous histochemical analyses by demonstrating differential gene expression between gland regions, (2) identified sequences for all known plethodontid pheromones, and (3) discovered numerous novel protein coding sequences that could be involved in the chemical signaling, behavioral ecology, and general biology of plethodontid salamanders.

## Identification and Expression of Known Plethodontid Pheromones

Recent biochemical research has identified several plethodontid pheromone proteins via proteomic analyses (Rollmann et al., 1999; Chouinard et al., 2013; Wilburn et al., 2014a). Thus, these proteins (PRF, PMF, PTP, SPF, PRRP, and CRISP) are known to be translated and secreted by the MG of representative plethodontid species. Sequences for each of these previously-identified pheromones were found in the secretomes of the CG, PCG and DTB, though not all pheromone transcripts were expressed equally among these three glands. Many of these sequences were full-length, but many were fragmented. The large number of paralogs with high sequence dissimilarity (Watts et al., 2004; Palmer et al., 2010; Wilburn et al., 2012) makes these
proteins especially difficult to assemble with current bioinformatic techniques. Nevertheless, these data do include many full-length sequences and provide a useful starting point for investigating the expression and function of pheromones in other (nonMG) signaling glands.

Interestingly, genes for the two primary pheromones in Plethodon (PRF and PMF) were most highly-expressed in the CG, a gland known to be involved in courtship and previously suspected of pheromone production. In addition, PRRP was also expressed almost exclusively in the CG. This protein was only recently identified (Wilburn et al., 2014a), and its role in chemical signaling is unknown. The expression of PRF and PMF by the CG is intriguing, in that it represents a new glandular source for these pheromones. Palmer et al. (2007b) did identify low levels of PMF mRNA in several tissues other than the MG (ventral/dorsal tail skin, liver, intestines, kidneys), but Fontana et al. (2007) found no PMF expression in the glands of the ventral (PCG) or dorsal tail (DTB). Neither study investigated the CG. These low (or absent) levels of PMF expression in the PCG/DTB coincides with the findings of the present study. The degree of expression in the CG, however, was several orders of magnitude higher than either of the other two tissues (Figure 4.4). Thus, it seems likely that the CG may also be a source of secreted PMF (and/or PRF) in plethodontids. No studies to my knowledge have previously identified PRF mRNA from tissues other than the MG.

In contrast to the CG, the PCG and DTB both expressed much higher levels of SPF and CRISP-like factor. As in the case of PRF/PMF, SPF had not been previously identified in these tissues (Fontana et al., 2007). In this study, expression of SPF was relatively widespread, but different tissues upregulated different SPF sequences. Even in the CG, which had the lowest level of SPF expression, it was present in higher levels than all other pheromones except PMF. In Plethodon, SPF is thought to be undergoing evolutionary replacement by PRF (Palmer et al., 2007a), but this hypothesis is based on data exclusively from the MG. SPF is an ancient salamander pheromone, and it owes its ancestral function to secretion by cloacal glands (Kikuyama et al., 1995; 2002; Janssenswillen et al., 2014; Van Bocxlaer et al., 2015; Maex et al., 2016). Thus, the role of SPF may have been underappreciated in Plethodon spp. due to a lack of attention to
other, plesiomorphic signaling glands. In other words, a reduction of SPF expression appears to be the case in the MG of select plethodontids, but it may still have other important (likely even ancestral) functions in the same species through its production in other glands.

One caveat for these analyses is that the sequences included were restricted to those with definitive top BLAST hits to the plethodontid versions of these pheromones. These were intended as conservative estimates of pheromone expression, but incomplete assembly and/or ambiguous annotations (e.g. PMF, SPF, CRISPs) may result in a limited view of expression levels in reality. In addition, all data represent mRNA expression, but this is not necessarily proportional to levels of translated protein, which is what ultimately matters for chemical signaling. As such, these data should be considered a preliminary exploration of known pheromone expression in other signaling glands. Future proteomic and behavioral experiments are needed to investigate the expression and function of pheromones from non-MG tissues.

## Identification and Expression of Putative Chemical Signaling Proteins

In addition to the detection of known plethodontid pheromones, many other proteins of interest have been identified from these signaling glands that warrant further study. Among these are many carbohydrate-binding proteins, mucins, protease inhibitors, toxins, reproductive proteins, and hormone-like proteins. As noted earlier, these data include mRNA expression only; further study is needed to conclude the extent of protein expression and any potential role in chemical signaling. Nevertheless, select cases deserve further discussion due to (1) the number of homologs identified from each category, (2) their degree of differential expression, and (3) their reasonable involvement in plethodontid chemical signaling based on prior research.

## Carbohydrate-binding Proteins and Mucins

When tallying the cumulative expression of BLAST hits across all transcripts and glands, the single most highly-expressed protein is lectin (Table A3.1). This theme is continued throughout the secretomes under study, with multiple lectin paralogs routinely
among the most abundant transcripts. Lectins are carbohydrate-binding and recognition molecules with diverse functions, including roles in innate and adaptive immunity, glycoprotein synthesis and binding, cell-recognition, regulation of cell growth, modulation of cell-cell and cell-substratum interactions, apoptosis and toxicity, spermegg fusion, and numerous other processes (Ghazarian et al., 2011). Their core functionality stems from the presence of one or more carbohydrate-recognition domains, and different lectins demonstrate varying degrees of specificity for different classes of carbohydrates. Lectins are categorized into (among others) the C-type (calciumdependent), S-type (galectins; soluble), I-type (immunoglobulin-like; siglecs), and F-type (fucolectins) lectins (Sharon and Lis, 2004; Ghazarian et al., 2011; Vasta et al., 2012). Examples of each of these major lectin groups are present and ubiquitous in the three glands studied, and different lectin paralogs are significantly upregulated in these different tissues (Table 4.4).

These proteins may be crucial components of chemical signaling in plethodontids, as inferred by: (1) the large number of lectins identified, (2) their widespread high levels of expression, (3) the extent of carbohydrate and glycoprotein components present in these signaling glands (especially the CG and PCG; Sever, 2003; Hecker et al., 2003), and (4) the diversity of lectin forms and functions. If carbohydrates themselves act as chemical signals, lectins (and other carbohydrate-binding proteins) may play an accessory role in the production, maintenance, and modulation of these glandular secretions. This hypothesized function is analogous to the accessory role in lipid signaling carried out by the major urinary proteins in the scent marks of mammals (Hurst et al., 1998; Hurst and Beynon, 2004; Hurst, 2009).

Mucin transcripts are also present in high numbers, some of which were upregulated in the two integumentary glands (PCG/DTB; Table 4.4). Mucins are heavilyglycosylated proteins (Bansil and Turner, 2006) produced by mucous glands, but in some cases may also be produced by serous (granular) or mixed glands (Fontana et al., 2006; Simons et al., 1999). Mucous is an important product of amphibian skin, helping to keep it moist and therefore aiding in respiration and water balance (Clarke, 1997). Mucins likely constitute a major structural component of both antipredator substances and scent
marks due to the sticky composition of integumentary secretions (Largen and Woodley, 2008; von Byern et al., 2015; pers. obs.) and the abundance of mucin transcripts in the DTB and PCG.

## Protease Inhibitors

Another highly-expressed group of transcripts encoded many types of protease inhibitors. A common theme was the abundance of serine protease inhibitors (serpins), such as: kazal-types, trypsin inhibitors, ovomucoid, antileukoproteinase, WAP four disulfide core, kunitz-types, and others. Metalloproteinase inhibitors were also present; this included PTP, a plethodontid TIMP-1 homolog known to be secreted from the MG (Chouinard et al., 2013; Wilburn et al., 2014a). Previous sequencing experiments in the MG have also shown an increased expression in kazal-type and cystatin transcripts (Wilburn, 2014). Taken together, these findings suggest an important function for protease inhibitors, presumably in the synthesis and maintenance of secretory products. Even when secreted to the environment (Wilburn et al., 2014a), these proteins would be important for protecting the signaling pheromones from degradation.

## Cysteine Rich Secretory Proteins

Other highly-expressed proteins of chemical signaling interest included many homologs of cysteine-rich secretory proteins (CRISPs). As discussed above, one CRISP homolog (plethodontid CRISP-like factor) is also among the proteins secreted by the MG (Wilburn et al., 2014a), but many other sequences were identified with homology to other CRISPs. These proteins have diverse phylogenetic distribution and functionality. CRISPs are ubiquitous in the venoms of snakes and other animals, exerting their toxicity by blocking calcium channels and reducing smooth muscle contraction (Yamazaki and Morita, 2004). In addition to their common inclusion in venoms, different CRISP variants are also important for reproduction; they are expressed in the testes, seminal fluids and accessory glands, form a part of the sperm acrosome and tail base, play an important role in fertilization, and act as sperm attractants (Gibbs et al., 2008; Koppers et al., 2011; Olson et al., 2001). This overlap between toxin and reproductive functions in CRISPs is
not unlike PMF, and in both cases the primary role in plethodontid chemical signaling (i.e. toxic or reproductive functions) is unclear. The plethodontid CRISP-like factor secreted by the MG shares its closest homology to CRISP2, which in mammals is produced in the testes and is involved in several aspects of sperm functionality (Gibbs et al., 2008). Many other CRISPs identified in this experiment, however, share closest homology to venoms. Both functions are likely to be important in salamanders and may be carried out by different CRISP paralogs.

## Reproductive Proteins

Proteins homologous to those with known reproductive functions were also identified. Some of these were upregulated in the CG (Table 4.4), namely betamicroseminoprotein and relaxin homologs. Beta-microseminoprotein (also called PSP94) is one of the largest components of prostate secretions in mammals, but is also present in other secretory glands (Weiber et al., 1990). This protein is hypothesized to improve fertility through antifungal effects and/or by protecting sperm from the female immune system (Laurence, 2013). Numerous insulin-like transcripts of the relaxin family were also upregulated in the CG. Relaxins have a wide range of reproductive functions, largely on the female reproductive tract prior to and during labor in mammals (Bathgate et al., 2013). However, some relaxins are also implicated in appetite-regulation, including the particular paralog upregulated in the CG (INSL5). Thus, this protein may play a metabolic role (discussed below) instead of, or in addition to, a reproductive function. Due to its use in spermatophore production (Sever, 2003) and the importance of sperm storage in this species (Eddy et al., 2015), the male CG may express proteins such as these to modulate the tissues of the spermatheca for improved fertility and/or sperm competition. Seminal fluid proteins are known to play these roles in other systems (Avila et al., 2011).

Other differentially-expressed proteins implicated in reproduction included serine protease inhibitors often found in high concentrations in seminal fluids (Perry et al., 1993; Marzoni et al., 2013), such as: acrosin/kazal-type-2 in the CG and trypsin inhibitor CITI-1 in the PCG (Table 4.4). The latter protein resulted in the PCG having significant
enrichment for male reproductive GO terms (Table 4.5). Other proteins with even clearer connections to reproductive tissues were also found, including components of the sperm acrosome and tail base, testis- and placenta-specific proteins, as well as components of amphibian eggs and foam nests. While substantial variation in expression levels exists between glands, few reproductive transcripts were significantly differentially-expressed.

## Hormone-like Proteins

Lastly, another recurring theme among highly-expressed transcripts was the presence of numerous hormone-like proteins, many of which affect appetite and/or metabolism. Examples include: insulin-like peptides, gastrointestinal growth factors, resistin, adiponectin, peptide YY, ghrelin, and appetite-regulating hormone. These findings are in conjunction with previous sequencing efforts in the MG of multiple plethodontids (Kiemnec-Tyburczy et al., 2009; Wilburn, 2014; Doty et al., 2015). While the degree of protein expression and secretion for these transcripts remains unknown, the co-option of appetite-regulating hormones as allohormone pheromones (sensu Wyatt, 2014b) would likely be adaptive in both territorial advertisement and reproduction. One of the primary functions of territorial scent marks is to aid in the establishment and defense of territories and the energetic resources they contain (Jaeger, 1986). In the context of mating, during limited bouts of activity at the surface, females are confronted with the decision to either mate or forage (Eddy, 2012). In both cases (male or female receivers), suppressing attraction to prey stimuli would be adaptive for male territory holders. Indeed, MG secretions have been shown to reduce female attention to food stimuli and increase their preference for male scents (Vaccaro et al., 2009; 2010), but more rigorous experiments are needed. Nonetheless, the production of copious appetiteregulating hormone-like transcripts in courtship (MG, CG) and scent marking glands (CG, PCG), coupled with preliminary evidence that pheromones may modulate female preference for food cues, suggest that this hypothesis merits serious consideration in future research.

## Conclusions and Future Directions

This research has provided the first extensive analysis of the transcripts produced by three important signaling glands in plethodontids: the cloacal glands, postcloacal gland, and dorsal tail base. These tissues have long been known to be important for many aspects of inter- and intraspecific communication, in addition to their roles in reproduction, and other facets of salamander biology. Despite the behavioral evidence for the importance of these glands for the social behavior of plethodontids, a significant knowledge gap existed regarding the chemicals produced to facilitate these functions. The data generated herein provide numerous novel homologs of known genes, as well as many new sequences that yield no significant similarity to known genes at present. Several of these unidentifiable proteins were highly-expressed in the glands studied, and future research is necessary to elucidate their identity and function.

The primary purpose of this research was to construct transcriptome libraries for use in future proteomic work. Ultimately, it is the production of the proteins encoded by genes that dictate functionality, either through their direct use as protein pheromones, or through their function(s) in producing and/or modulating other chemicals that serve as pheromones. Thus, the next step in this research is to carry out large-scale proteomic studies that compare protein expression and secretion between various glandular tissues. Differences between the sexes (male vs. female) and/or individuals with different reproductive status (juvenile vs. adult; gravid vs. non-gravid females) should also be examined. This will aid in identifying proteins that contribute to signal specificity that enables different signaling functions. In addition to meeting this initial goal, bioinformatic analyses provided a powerful starting point for exploring the genes present and their level of expression, and these conclusions have led to several relevant hypotheses for future study.

Table 4.1. Read count and length metrics before and after quality trimming and adapter clipping. The total number of reads is compared before and after trimming/clipping, including the number of paired and single reads. Read length data are also provided. Overall, trimming and clipping resulted in a loss of $>25 \%$ of reads and $>35 \%$ of total base pairs. For a graphical view of quality score distributions by read position before and after trimming/clipping, see Appendix Three (Figure A3.1).

| Metric | Pre Trimming \& Clipping | Post Trimming \& Clipping |
| :--- | ---: | ---: |
| Total number of reads | $90,160,158$ | $66,567,678$ |
| Number of paired reads | $(45,080,079$ pairs) $90,160,158$ | $(27,095,573$ pairs) $54,191,146$ |
| Number of single reads | 0 | $12,376,532$ |
| Average read length (bp) | $100 \pm 0.00$ (S.D.) | $85.52 \pm 17.75$ (S.D.) |
| Median read length (bp) | 100 | 97 |
| Number of base pairs (bp) | $9,016,015,800$ | $5,693,199,830$ |
| Proportion of raw reads | $100.00 \%$ | $73.83 \%$ |
| Proportion of raw base pairs | $100.00 \%$ | $63.15 \%$ |

Table 4.2. Transcriptome assembly quality metrics. Assembly count and length metrics include: the number of "genes" and "transcripts" (in Trinity nomenclature), mean and median lengths, N10-N50 measures, transcript counts by length ( $>1000$ and $>10,000$ bp ), and the number of open reading frames (ORFs). Mapping results include: the number of paired reads used, and the number and proportion mapped. Also included are the number of transcripts comprising $90 \%$ of expression and the N50 for these transcripts (E90N50). TransRate statistics include: the raw assembly score, the optimized contig score cutoff, the number of transcripts above the cutoff ("good" contigs), and the optimal score. Lastly, BUSCO results describe the number of vertebrate single copy orthologs (BUSCOs) identified as full-length (single copy and duplicated), or fragmented.

| Contig Count and Length Metrics |  |  |
| :--- | ---: | :---: |
| Trinity "genes" | 73,928 |  |
| Trinity "transcripts" | 94,221 |  |
| Mean length (longest transcript only) | 509.95 |  |
| Median length (longest transcript only) | 330 |  |
| N50 (longest transcript only) | 667 |  |
| N40 (longest transcript only) | 877 |  |
| N30 (longest transcript only) | 1,121 |  |
| N20 (longest transcript only) | 1,395 |  |
| N10 (longest transcript only) | 1,780 |  |
| Transcripts over 1,000 bp | 11,657 |  |
| Transcripts over 10,000 bp | 1 |  |
| Open Reading Frames (ORFs) | 24,074 |  |
| Abundance Estimation (Mapping) Metrics |  |  |
| Reads for mapping (paired only) | $54,191,146$ |  |
| Reads mapped | $44,028,070$ |  |
| Proportion mapped | $81.25 \%$ |  |
| Number of transcripts (90\% expression) | 12,541 |  |
| E90N50 (N50 for 90\% of expression) | 1,319 |  |
|  |  |  |
| Raw TransRate score (all contigs) | 0.10 |  |
| Optimal score cutoff | 0.03295 |  |
| Contigs above cutoff ("good") | 74,041 |  |
| Optimal TransRate score (good contigs) | 0.20 |  |
| BUSCO Metrics |  |  |
| Full-length BUSCOs (total) | 450 |  |
| Full-length BUSCOs (single copy) | 346 |  |
| Full-length BUSCOs (duplicated) | 104 |  |
| Fragmented BUSCOs | 269 |  |
| Total BUSCOs found | 719 |  |
| Vertebrate BUSCOs | 3,023 |  |
| Percentage of full-length BUSCOs | $14.89 \%$ |  |
| Percentage of BUSCOs found | $23.78 \%$ |  |
|  |  |  |

Table 4.3. Common BLAST homologies grouped by relevant secretory protein categories. Select examples of common transcript homologies are presented. Each category is listed in approximate descending order of expression; due to many transcripts sharing variations of the same homology, these are estimates (but see Table A3.1 for a quantitative ranking for each specific BLAST hit). The homologies listed are not comprehensive for each category, and in most cases variations of specific BLAST hits have been collapsed for brevity (e.g. "lectins"). Thus, this provides a summary of broad categories of homology for many of the major transcripts identified. Details about each individual BLAST hit are available in Appendix Three (Table A3.1), which includes the specific homologous hit, the number of transcripts with that homology, and its cumulative level of expression across all of those transcripts. In addition, an electronic version of the complete annotation report (including annotations for all individual transcripts) is available upon request. See the methods section for a description of the annotations included therein.

| Carbohydrate-binding and Associated Proteins |
| :--- |
| Lectins |
| Mucins |
| Galectins |
| Fucolectins |
| Lectoxin |
| Intelectin |
| Ficolins |
| Brevican/neurocan/aggrecan/versican core proteins |
| Anterior gradient protein |
| $\quad$ Known Pheromones |
| Plethodontid modulating factor (PMF) |
| Sodefrin precursor-like factor (SPF) |
| Plethodontid receptivity factor (PRF) |
| Cysteine-rich secretory protein-like factor (CRISP) |
| Plethodontid TIMP-like protein (PTP) |
| Poliovirus receptor-related protein (PRRP) |
| Mouse major urinary proteins (MUPs) |
| $\quad$ Protease Inhibitors |
| Kazal-type proteinase inhibitor variants |
| Ovomucoid |
| Pancreatic secretory trypsin inhibitors |
| Antileukoproteinase |
| Metalloproteinase inhibitors |
| Serpins |
| Ovostatin |
| Cystatin |
| Kunitz-type serine proteinase inhibitors |



## Lipid-binding and Associated Proteins

| Apolipoproteins <br> Prosaposins <br> Annexin <br> BPI-fold containing protein <br> Major urinary proteins <br> Salivary lipocalin |
| :--- |
| Proteases |
| Cathepsin <br> Antigen WC1.1 <br> Pepsin |
| Antimicrobial and Antifungal Peptides |
| Lactoperoxidase <br> Cathelicidin <br> Lysozymes <br> Defensins |

Table 4.4. Relevant secretory proteins identified from each differential-expression cluster. Secreted proteins were extracted from annotations and inspected for their potential relevance in plethodontid chemical signaling and/or social behavior. All entries represent individual transcripts and their homology, listed in approximate descending order of expression (higher expressed transcripts at the top). Transcript identifiers have been eliminated for brevity. For each differential expression cluster (DEC), the number of transcripts is listed, as are the number and proportion of identified secretory transcripts. Each transcript is accompanied by a broad functional category of interest to chemical signaling. Especially notable transcripts are identified with bolded text, namely those that have strong homology to known (or putative) chemical signaling proteins in salamanders.

| DEC | Trans. | Secreted | Relevant Secretory Protein Homology | Relevant <br> Functions |
| :---: | :---: | :---: | :---: | :---: |
| 3 | 604 | $\begin{gathered} 148 \\ (24.50 \%) \end{gathered}$ | Leukocyte cell-derived chemotaxin-2 <br> Glutathione peroxidase 3 <br> WAP-type four disulfide core <br> Kazal-type 6 serine protease inhibitor <br> Prod 1, PAR/Ly-6 domain <br> Olfactomedin-4 <br> Ovomucoid <br> Fucolectin <br> Sodefrin-like protein (Plethodon) <br> Galectin <br> Intelectin <br> Pancreatic secretory trypsin inhibitor <br> Probable cytosolic oligopeptidase A <br> Galactosidase binding lectin <br> PMF Class III-like B variant (Plethodon) <br> Antigen WC1.1 <br> Grifin <br> PMF Class III-like A variant (Plethodon) <br> Resistin <br> Brevican core protein <br> Type-2 ice-structuring protein <br> PRF A1 (Plethodon) <br> Phospholipase A2, group IIA <br> C-type lectin lectoxin-Enh7 <br> Beta-microseminoprotein <br> Kazal-type 2 serine protease inhibitor <br> Deleted in malignant brain tumors 1 <br> Immunoglobulin superfamily member 2 <br> Prokineticin Bm8-d, AVIToxin-VAR1 <br> Fibroblast growth factor 19 <br> Immunoglobulin superfamily member 3 <br> PRRP (Plethodon) <br> Cysteine-rich secretory protein 2 <br> Amyloid beta A4 protein <br> Dipeptidase 1 <br> Cysteine-rich venom protein <br> Snaclec B9 <br> C-type lectin lectoxin-Enh2 <br> BPI fold-containing family B member 3 <br> SLIT and NTRK-like protein 2 <br> Scavenger receptor cysteine-rich type 1 <br> Relaxin, Insulin-like peptide <br> Kunitz-type protease inhibitor 2 <br> C1q tumor necrosis factor-related 2 <br> Pepsin | Immune function <br> Lipid metabolism <br> Protease inhibitor <br> Protease inhibitor <br> Putative pheromone <br> Hormone-like <br> Protease inhibitor <br> Carbohydrate binding <br> Pheromone <br> Carbohydrate binding <br> Carbohydrate binding <br> Protease inhibitor <br> Protease <br> Carbohydrate binding <br> Pheromone <br> Protease <br> Carbohydrate binding <br> Pheromone <br> Hormone-like <br> Carbohydrate binding <br> Carbohydrate binding <br> Pheromone <br> Lipid metabolism, toxin <br> Carbohydrate binding <br> Reproduction <br> Protease inhibitor <br> Immune function <br> Immune function <br> Hormone-like, toxin <br> Hormone-like <br> Hormone-like, gland function <br> Putative pheromone <br> Toxin, putative pheromone <br> Protease inhibitor <br> Protease <br> Toxin, putative pheromone <br> Toxin <br> Toxin <br> Lipid binding <br> Hormone-like <br> Immune function <br> Hormone-like, reproduction <br> Protease inhibitor <br> Immune function <br> Protease |


| 6 | 55 | $\begin{gathered} 11 \\ (20.00 \%) \end{gathered}$ | Resistin <br> SPARC (kazal-type inhibitor) <br> Ig lambda-1 chain C region | Hormone-like Protease inhibitor Immune function |
| :---: | :---: | :---: | :---: | :---: |
| 1 | 209 | $\begin{gathered} 50 \\ (23.92 \%) \end{gathered}$ | Lectin <br> Kazal-type 4 serine protease inhibitor <br> Prokineticin <br> Brevican core protein <br> Fucolectin <br> Galectin <br> Sodefrin-like protein (Plethodon) <br> Insulin-like peptide <br> Pancreatic secretory trypsin inhibitor <br> Kazal-type 12 serine protease inhibitor <br> PMF Class III-like B variant (Plethodon) <br> Resistin <br> Saposin A-type, Prosaposin <br> Kazal-type 6 serine protease inhibitor <br> WAP four-disulfide core domain 18 <br> Apolipoprotein C-I <br> Kazal-type 2 serpin (acrosin) <br> Intelectin <br> Grifin <br> Pseudechetoxin-like (CRISP 2) | Carbohydrate binding <br> Protease inhibitor <br> Hormone-like, toxin <br> Carbohydrate binding <br> Carbohydrate binding <br> Carbohydrate binding <br> Pheromone <br> Hormone-like <br> Protease inhibitor <br> Protease inhibitor <br> Pheromone <br> Hormone-like <br> Lipid metabolism, gland function <br> Protease inhibitor <br> Protease inhibitor <br> Lipid metabolism <br> Protease inhibitor, reproduction <br> Carbohydrate binding <br> Carbohydrate binding <br> Toxin, putative pheromone |
| 2 | 242 | $\begin{gathered} 60 \\ (24.79 \%) \end{gathered}$ | u-PAR/Ly-6 domain <br> Avidin-related protein 4/5 <br> Ovomucoid <br> Trypsin inhibitor CITI-1 <br> Gastrointenstinal growth factor xP4 <br> Galactose-specific lectin nattectin <br> Sodefrin-like protein (Plethodon) <br> C-type lectin lectoxin-Enh7 <br> Antileukoproteinase <br> Phospholipase A2 inhibitor, Ly6/PLAUR <br> Group IIE secretory phospholipase A2 | Toxin, putative pheromone <br> Reproduction <br> Protease inhibitor <br> Protease inhibitor, reproduction <br> Hormone-like <br> Carbohydrate binding <br> Pheromone <br> Carbohydrate binding <br> Protease inhibitor <br> Toxin, putative pheromone <br> Lipid metabolism |
| 5 | 34 | $\begin{gathered} 10 \\ (29.41 \%) \end{gathered}$ | Sodefrin-like protein (Plethodon) <br> Galectin-3 <br> Cysteine-rich secretory protein 3 | Pheromone <br> Carbohydrate binding <br> Toxin, putative pheromone |
| 4 | 211 | $\begin{gathered} 47 \\ (22.27 \%) \end{gathered}$ | Galectin-3 <br> Galectin-1 <br> Sodefrin-like protein (Plethodon) <br> Galectin-9 <br> Integumentary mucin C. 1 <br> WAP-type four disulfide core <br> Antileukoproteinase <br> Phospholipase A2 inhibitor, Ly6/PLAUR <br> Ovomucoid <br> Lymphocyte antigen 6E <br> Ly6/PLAUR domain-containing 2 <br> Cathelicidin <br> Otoconin 22 <br> BPI fold-containing family C <br> CRISP-like factor 1 (Plethodon) <br> Venom peptide SjAPI-2 | Carbohydrate binding <br> Carbohydrate binding <br> Pheromone <br> Carbohydrate binding <br> Mucin, antimicrobial <br> Protease inhibitor <br> Protease inhibitor <br> Toxin, putative pheromone <br> Protease inhibitor <br> Toxin, putative pheromone <br> Toxin, putative pheromone <br> Antimicrobial <br> Lipid metabolism, toxin <br> Lipid binding <br> Putative pheromone <br> Toxin |

Table 4.5. Significantly enriched Gene Ontology Terms for each expression cluster. Annotation subsets were generated for each expression cluster (DEC; see Figures 4.2 and 4.3) and all GO terms were extracted to identify significantly-enriched terms (false discovery rate $<0.05$ ). All significant GO terms are included per cluster in descending order of statistical significance. Select GO terms with potential relevance to plethodontid chemical signaling and/or social behavior are identified with bolded text.

| DEC | Enriched | Enriched GO Terms (FDR < 0.05) | GO ID | FDR |
| :---: | :---: | :---: | :---: | :---: |
| 3 | 58 | CC extracellular region | GO:0005576 | 4.92E-48 |
|  |  | MF carbohydrate binding | GO:0030246 | 5.36E-38 |
|  |  | BP lacrimal gland development | GO:0032808 | $2.04 \mathrm{E}-10$ |
|  |  | MF enzyme inhibitor activity | GO:0004857 | $2.04 \mathrm{E}-10$ |
|  |  | MF peptidase inhibitor activity | GO:0030414 | 4.05E-10 |
|  |  | MF serine-type endopeptidase inhibitor activity | GO:0004867 | 1.36E-09 |
|  |  | MF peptidase regulator activity | GO:0061134 | $2.18 \mathrm{E}-09$ |
|  |  | MF scavenger receptor activity | GO:0005044 | 4.34E-09 |
|  |  | MF cargo receptor activity | GO:0038024 | $1.34 \mathrm{E}-08$ |
|  |  | MF endopeptidase inhibitor activity | GO:0004866 | $4.38 \mathrm{E}-08$ |
|  |  | MF endopeptidase regulator activity | GO:0061135 | 7.24E-08 |
|  |  | BP defense response | GO:0006952 | $2.18 \mathrm{E}-07$ |
|  |  | MF hyaluronic acid binding | GO:0005540 | $1.46 \mathrm{E}-05$ |
|  |  | BP acute-phase response | GO:0006953 | $2.15 \mathrm{E}-05$ |
|  |  | MF enzyme regulator activity | GO:0030234 | $4.98 \mathrm{E}-05$ |
|  |  | MF tryptamine:oxygen oxidoreductase (deaminating) activity | GO:0052593 | $1.04 \mathrm{E}-04$ |
|  |  | MF aminoacetone:oxygen oxidoreductase(deaminating) activity | GO:0052594 | $1.04 \mathrm{E}-04$ |
|  |  | MF aliphatic-amine oxidase activity | GO:0052595 | $1.04 \mathrm{E}-04$ |
|  |  | MF phenethylamine:oxygen oxidoreductase (deaminating) activity | GO:0052596 | $1.04 \mathrm{E}-04$ |
|  |  | BP acute inflammatory response | GO:0002526 | $1.21 \mathrm{E}-04$ |
|  |  | BP innate immune response | GO:0045087 | $4.58 \mathrm{E}-04$ |
|  |  | CC proteinaceous extracellular matrix | GO:0005578 | 1.53E-03 |
|  |  | MF primary amine oxidase activity | GO:0008131 | $1.94 \mathrm{E}-03$ |
|  |  | MF glycosaminoglycan binding | GO:0005539 | 2.63E-03 |
|  |  | BP cell adhesion | GO:0007155 | $2.73 \mathrm{E}-03$ |
|  |  | BP negative regulation of T cell activation | GO:0050868 | $2.84 \mathrm{E}-03$ |
|  |  | BP immune response | GO:0006955 | $3.09 \mathrm{E}-03$ |
|  |  | MF hormone activity | GO:0005179 | 3.86E-03 |
|  |  | BP biological adhesion | GO:0022610 | 3.86E-03 |
|  |  | CC cell surface | GO:0009986 | 3.86E-03 |
|  |  | BP negative regulation of leukocyte cell-cell adhesion | GO:1903038 | $4.20 \mathrm{E}-03$ |
|  |  | BP gland development | GO:0048732 | $4.75 \mathrm{E}-03$ |
|  |  | CC extracellular space | GO:0005615 | $5.64 \mathrm{E}-03$ |
|  |  | BP negative regulation of lymphocyte activation | GO:0051250 | 8.07E-03 |
|  |  | BP negative regulation of cell-cell adhesion | GO:0022408 | $9.05 \mathrm{E}-03$ |
|  |  | BP negative regulation of homotypic cell-cell adhesion | GO:0034111 | $1.07 \mathrm{E}-02$ |
|  |  | CC extracellular matrix | GO:0031012 | $1.18 \mathrm{E}-02$ |
|  |  | BP regulation of T cell activation | GO:0050863 | $1.18 \mathrm{E}-02$ |
|  |  | BP regulation of lymphocyte activation | GO:0051249 | $1.18 \mathrm{E}-02$ |
|  |  | BP regulation of immunological synapse formation | GO:2000520 | $1.18 \mathrm{E}-02$ |
|  |  | BP negative regulation of immunological synapse formation | GO:2000521 | $1.18 \mathrm{E}-02$ |
|  |  | BP regulation of T cell activation via T cell receptor contact with antigen bound to MHC molecule on antigen presenting cell | GO:2001188 | 1.18E-02 |
|  |  | BP negative regulation of T cell activation via T cell receptor contact with antigen bound to MHC molecule on antigen presenting cell BP regulation of leukocyte cell-cell adhesion | GO:2001189 | $1.18 \mathrm{E}-02$ $1.39 \mathrm{E}-02$ |
|  |  | CC spliceosomal complex | GO:0005681 | $1.45 \mathrm{E}-02$ |
|  |  | BP negative regulation of leukocyte activation | GO:0002695 | $1.67 \mathrm{E}-02$ |
|  |  | MF oxidoreductase activity, acting on the CH-NH2 group of donors, oxygen as acceptor | GO:0016641 | 1.68E-02 |
|  |  | BP negative regulation of endocytosis | GO:0045806 | $1.81 \mathrm{E}-02$ |
|  |  | BP regulation of leukocyte activation | GO:0002694 | 2.05E-02 |
|  |  | BP regulation of homotypic cell-cell adhesion | GO:0034110 | 2.40E-02 |
|  |  | BP regulation of cell-cell adhesion | GO:0022407 | $2.61 \mathrm{E}-02$ |
|  |  | BP regulation of cell activation | GO:0050865 | 3.20E-02 |


|  |  | MF phospholipase inhibitor activity BP response to stress <br> BP negative regulation of cell activation <br> MF lipase inhibitor activity <br> MF quinone binding <br> CC immunological synapse | GO:0004859 GO:0006950 GO:0050866 GO:0055102 GO:0048038 GO:0001772 | $\begin{aligned} & \hline 3.20 \mathrm{E}-02 \\ & 3.22 \mathrm{E}-02 \\ & 3.25 \mathrm{E}-02 \\ & 3.54 \mathrm{E}-02 \\ & 3.92 \mathrm{E}-02 \\ & 3.92 \mathrm{E}-02 \\ & \hline \end{aligned}$ |
| :---: | :---: | :---: | :---: | :---: |
| 6 | 1 | MF antigen binding | GO:0003823 | $1.77 \mathrm{E}-02$ |
|  |  | CC extracellular region | GO:0005576 | $0.00 \mathrm{E}+00$ |
|  |  | MF carbohydrate binding | GO:0030246 | $0.00 \mathrm{E}+00$ |
|  |  | MF serine-type endopeptidase inhibitor activity | GO:0004867 | $1.26 \mathrm{E}-07$ |
|  |  | MF enzyme inhibitor activity | GO:0004857 | $1.68 \mathrm{E}-07$ |
|  |  | MF endopeptidase inhibitor activity | GO:0004866 | 3.74E-06 |
|  |  | MF endopeptidase regulator activity | GO:0061135 | 4.32E-06 |
|  |  | MF peptidase inhibitor activity | GO:0030414 | 5.21E-06 |
|  |  | MF peptidase regulator activity | GO:0061134 | $1.09 \mathrm{E}-05$ |
|  |  | MF enzyme regulator activity | GO:0030234 | $1.09 \mathrm{E}-05$ |
|  |  | BP regulation of cellular defense response | GO:0010185 | 1.66E-05 |
|  |  | BP regulation of complement activation, lectin pathway | GO:0001868 | $2.33 \mathrm{E}-05$ |
|  |  | BP regulation of protein processing | GO:0070613 | 4.48E-05 |
|  |  | CC extracellular space | GO:0005615 | $9.55 \mathrm{E}-05$ |
|  |  | MF fucose binding | GO:0042806 | 9.93E-05 |
|  |  | BP regulation of complement activation | GO:0030449 | $1.30 \mathrm{E}-04$ |
|  |  | BP regulation of protein activation cascade | GO:2000257 | $1.30 \mathrm{E}-04$ |
| 1 | 33 | BP regulation of humoral immune response | GO:0002920 | 3.45E-04 |
|  |  | BP regulation of serine-type endopeptidase activity | GO:1900003 | 3.45E-04 |
|  |  | BP negative regulation of serine-type endopeptidase activity | GO:1900004 | 3.45E-04 |
|  |  | BP regulation of serine-type peptidase activity | GO:1902571 | 3.45E-04 |
|  |  | BP negative regulation of serine-type peptidase activity | GO:1902572 | 3.45E-04 |
|  |  | BP regulation of calcium ion import | GO:0090279 | 3.55E-04 |
|  |  | BP regulation of acute inflammatory response | GO:0002673 | 5.68E-04 |
|  |  | BP regulation of immune effector process | GO:0002697 | $1.18 \mathrm{E}-02$ |
|  |  | BP regulation of innate immune response | GO:0045088 | $1.29 \mathrm{E}-02$ |
|  |  | BP regulation of nitric oxide mediated signal transduction | GO:0010749 | $1.31 \mathrm{E}-02$ |
|  |  | BP negative regulation of nitric oxide mediated signal transduction | GO:0010751 | 1.31E-02 |
|  |  | BP regulation of vesicle-mediated transport | GO:0060627 | 2.81E-02 |
|  |  | MF monosaccharide binding | GO:0048029 | 2.81E-02 |
|  |  | BP negative regulation of cellular amine metabolic process | GO:0033239 | 2.92E-02 |
|  |  | BP defense response | GO:0006952 | 3.22E-02 |
|  |  | BP negative regulation of endocytosis | GO:0045806 | $4.18 \mathrm{E}-02$ |
|  |  | BP regulation of acrosome reaction | GO:0060046 | $4.86 \mathrm{E}-02$ |
| 2 | 30 | CC intermediate filament | GO:0005882 | $0.00 \mathrm{E}+00$ |
|  |  | CC extracellular region | GO:0005576 | 2.57E-15 |
|  |  | CC keratin filament | GO:0045095 | 1.84E-07 |
|  |  | CC hemoglobin complex | GO:0005833 | 4.51E-07 |
|  |  | MF oxygen transporter activity | GO:0005344 | $5.51 \mathrm{E}-07$ |
|  |  | MF enzyme inhibitor activity | GO:0004857 | 5.51E-07 |
|  |  | MF oxygen binding | GO:0019825 | 2.99E-05 |
|  |  | MF thioether S-methyltransferase activity | GO:0004790 | 9.66E-05 |
|  |  | MF amine N -methyltransferase activity | GO:0030748 | 9.66E-05 |
|  |  | MF serine-type endopeptidase inhibitor activity | GO:0004867 | 3.94E-04 |
|  |  | BP molting cycle | GO:0042303 | $5.96 \mathrm{E}-04$ |
|  |  | BP hair cycle | GO:0042633 | $5.96 \mathrm{E}-04$ |
|  |  | MF phospholipase inhibitor activity | GO:0004859 | 6.81E-04 |
|  |  | MF lipase inhibitor activity | GO:0055102 | 7.01E-04 |
|  |  | MF enzyme regulator activity | GO:0030234 | 7.01E-04 |
|  |  | MF peptidase inhibitor activity | GO:0030414 | 7.82E-04 |
|  |  | MF peptidase regulator activity | GO:0061134 | $1.74 \mathrm{E}-03$ |
|  |  | MF S-methyltransferase activity | GO:0008172 | $1.96 \mathrm{E}-03$ |
|  |  | BP male germ cell proliferation | GO:0002176 | $1.96 \mathrm{E}-03$ |
|  |  | BP germ cell proliferation | GO:0036093 | $1.96 \mathrm{E}-03$ |
|  |  | BP regulation of serine-type endopeptidase activity | GO:1900003 | 2.14E-03 |
|  |  | BP negative regulation of serine-type endopeptidase activity | GO:1900004 | 2.14E-03 |
|  |  | BP regulation of serine-type peptidase activity | GO:1902571 | 2.14E-03 |
|  |  | BP negative regulation of serine-type peptidase activity | GO:1902572 | 2.14E-03 |
|  |  | MF endopeptidase inhibitor activity | GO:0004866 | 2.88E-03 |
|  |  | MF endopeptidase regulator activity | GO:0061135 | 3.62E-03 |
|  |  | MF heme binding | GO:0020037 | 5.14E-03 |


|  |  | MF tetrapyrrole binding BP seminiferous tubule development MF calcium ion binding | $\begin{aligned} & \text { GO:0046906 } \\ & \text { GO:0072520 } \\ & \text { GO:0005509 } \end{aligned}$ | $\begin{aligned} & \hline 6.15 \mathrm{E}-03 \\ & 1.89 \mathrm{E}-02 \\ & 3.26 \mathrm{E}-02 \end{aligned}$ |
| :---: | :---: | :---: | :---: | :---: |
| 5 | 9 | BP eosinophil chemotaxis | GO:0048245 | $4.88 \mathrm{E}-02$ |
|  |  | BP mononuclear cell migration | GO:0071674 | $4.88 \mathrm{E}-02$ |
|  |  | BP positive regulation of mononuclear cell migration | GO:0071677 | 4.88E-02 |
|  |  | BP eosinophil migration | GO:0072677 | 4.88E-02 |
|  |  | BP positive regulation of calcium ion import | GO:0090280 | 4.88E-02 |
|  |  | BP regulation of T cell apoptotic process | GO:0070232 | 4.88E-02 |
|  |  | BP monocyte chemotaxis | GO:0002548 | 4.88E-02 |
|  |  | BP regulation of mononuclear cell migration | GO:0071675 | $4.88 \mathrm{E}-02$ |
|  |  | MF chemoattractant activity | GO:0042056 | $4.88 \mathrm{E}-02$ |
| 4 | 10 | CC extracellular region | GO:0005576 | 5.65E-17 |
|  |  | MF enzyme inhibitor activity | GO:0004857 | 1.40E-08 |
|  |  | MF phospholipase inhibitor activity | GO:0004859 | $3.91 \mathrm{E}-08$ |
|  |  | MF lipase inhibitor activity | GO:0055102 | 5.17E-08 |
|  |  | MF enzyme regulator activity | GO:0030234 | $6.73 \mathrm{E}-05$ |
|  |  | MF carbohydrate binding | GO:0030246 | 8.81E-04 |
|  |  | MF peptidase inhibitor activity | GO:0030414 | 1.97E-02 |
|  |  | BP innate immune response | GO:0045087 | $2.53 \mathrm{E}-02$ |
|  |  | BP defense response | GO:0006952 | $2.69 \mathrm{E}-02$ |
|  |  | MF peptidase regulator activity | GO:0061134 | $2.92 \mathrm{E}-02$ |



Figure 4.1. Full-length transcript reconstruction for all BLAST results. BLASTp (top) and BLASTx (bottom) results are shown for both the Swiss-Prot (left) and UniRef90 (right) databases. In each case, the number of transcripts is listed for different levels of query coverage relative to their top BLAST hit. The number on the axis below each bar is the maximum value for the bin; in other words, the first bin is from $0-10 \%$ coverage, while the final bin is for $90-100 \%$ coverage. The numbers in the top margin are the percentage of total queries corresponding to each bin. BLASTp on the UniRef90 database yielded the highest proportion of full-length transcripts, while BLASTx on the same database resulted in the highest total full-length transcripts.


Figure 4.2. Differential expression of transcripts $(\mathbf{N}=1,355)$ between the cloacal gland (CG), postcloacal gland (PCG), and dorsal tail base (DTB). Transcripts are shown as either highly (red) or lowly (blue) expressed. The units are log2(fpkm) values, median-centered for each transcript. These expression levels are also depicted for all transcripts in each sample as a trace line (solid vertical line) relative to the median (dashed vertical line). A density estimation of the same expression values for all transcripts is shown in the top left panel. Transcript expression patterns were hierarchically-clustered by similarity (tree at left) and divided into six major expression clusters (whose ranges are outlined by the boxes at right). The expression profiles for all transcripts in each cluster can be seen in Figure 4.3.


Figure 4.3. Transcript expression clusters. Expression profiles are plotted for each transcript (gray lines) across all three glands sampled, for each of the six differentialexpression clusters (see Figure 4.2). The means per sample are also plotted (black lines). The number of transcripts included in each cluster is listed in the panel titles.


Figure 4.4. Cumulative transcript production per gland for known pheromones and receptors from Plethodon spp. Each panel depicts the total expression (in TMM crosssample normalized values) for all transcripts per protein across the three glands sampled. The total TMM values per gland are shown in the margin above each bar. Data are shown for: plethodontid receptivity factor (PRF), plethodontid modulating factor (PMF), sodefrin precursor-like factor (SPF), plethodontid TIMP-like protein (PTP), CRISP-like factor (CRISP), and poliovirus receptor-related protein (PRRP). Also included are data for two putative pheromone receptor groups: interleukin-6 receptor beta (IL-6RB) and vomeronasal receptors (VR) of types I and II. These data include only those transcripts with top BLAST hits to their corresponding sequences from Plethodon spp. (with the exception of VRs, which include additional vertebrate VR homologs).

## CHAPTER FIVE: CONCLUSION

The chemical signals of plethodontid salamanders are highly complex, multicomponent mixtures that facilitate numerous behavioral interactions both within and between species. The closer we look at these signals, the more complex we realize they are. And yet, this is not surprising given the diversity of functions in which they are involved, and the high degree of signal specificity needed to convey the complex information inferred by receivers. This research has yielded new insights into the ways in which plethodontid chemical signals can vary within individuals (Chapter 2, 4), between members of the same population (Chapter 2, 3), as well as between species and the major lineages of eastern Plethodon (Chapter 3). When integrated with previous analyses, these data illustrate numerous mechanisms of signal variation at multiple levels of scale. In this conclusion, I will summarize the findings of this dissertation and attempt to synthesize their conclusions into an overall model for chemical communication in plethodontid salamanders. Lastly, I will discuss some of the most pressing residual questions in the context of directions for future research.

## The Findings of the Present Research

Chapter Two examined individual variation in mental gland pheromone composition within a single population of $P$. shermani, and demonstrated high intraspecific variability between male signals at the proteomic level. This variation was due to differences in the number and type of isoforms present, as well as their levels of expression. This research quantified the contribution of the major components, and in the case of Plethodontid Receptivity Factor (PRF), the expression of individual isoforms could also be quantified. This revealed that essentially each male in the population has a unique PRF profile, although certain profiles were more common in the population as a whole. This most common profile may differ between populations, as indicated by a previous study. Numerous minor components in the PRF range of individual males were also identified. These lowly expressed proteins have gone unreported in pooled population-level analyses, but inferences from other experimental systems suggest that
they may be functionally significant. In the case of Plethodontid Modulating Factor (PMF), the large number of isoforms meant that they could not easily be quantified at the proteomic level in the same manner. Qualitatively, the same aspects of individualspecific variation exist, but to a substantially higher degree. Overall, this research revealed that males in a single population differ markedly in the composition of courtship pheromones.

Chapter Three investigated the presence and pattern of post-translational modifications of PRF in the form of N- and O-linked glycosylation. Previous work had demonstrated this glycosylation in $P$. cinereus, a member of the ancestral clade utilizing dermal pheromone delivery, but not in P. shermani, a member of a derived clade using olfactory delivery. In the present research, bioinformatic analyses indicated that these observations extend to the entire clades represented by these species. Specifically, the potential for glycosylation (both N - and O-linked) is widespread in the PRFs of the ancestral clade, whereas all of the sequence motifs necessary for N -linked glycosylation (sequons) have been lost in the members of the derived clade. Some of these species do have slightly higher potential for O-linked glycosylation at select residues, but it's unclear at present whether these are ever glycosylated in nature. The sole species for which there is proteomic evidence implies that this is not the case. In addition, within the ancestral clade there is a diversity of patterns of glycosylation potential within and between species; that is, different PRFs have N-linked sequons in different regions of the protein, and some of these differences are pronounced between species. The potential for O-linked glycosylation was much more widespread in the ancestral clade, and likewise differed between species. Evolutionary analyses revealed that the N -linked sequons arose at different times, and protein structural predictions indicate that they may affect binding with the plesiomorphic receptor pathway. In total this work describes the properties of a higher level of signal variability that may have profound effects on specificity.

Chapter Four documented the construction and analysis of a de novo transcriptome assembly for three signaling glands involved in different aspects of social and reproductive behavior. High-throughput RNA sequencing allowed for the identification of thousands of new protein coding sequences which can be used for
functional characterization of signaling proteins. Mapping reads from individual tissues to this reference provides a first look at how these glands may differ in the expression of genes involved in communication. For decades researchers have discussed the behavioral functions of these glands and pondered the identity of the chemicals produced by them. Although restricted to mRNA, this work provides the first in-depth description of what these molecules might be. Due to the nature of protein function, this can also shed light on what other chemicals may be involved (e.g. carbohydrates). The transcripts produced by these tissues encode known plethodontid pheromones, many types of carbohydratebinding proteins, protease inhibitors, toxins, reproductive and hormone-like proteins, and many others. The three glands studied did differ in gene expression, although this should be considered a preliminary analysis in this regard. More rigorous expression methods and functional (i.e. proteomic) experimentation is needed to confidently describe the differences in secretory products produced by these tissues. Nevertheless, this analysis provides exciting new avenues for future research related to more general social and reproductive communication in plethodontids.

## An Integrated Model of Chemical Communication in Plethodontids

In general, chemical signals can vary by (1) the presence/absence of individual components, (2) structural variations of each component, of which there are numerous mechanisms depending on the molecules involved, and (3) the absolute and relative levels of expression of each component. Depending on the components included in a given secretion, the combination of these signal variability mechanisms allow for an enormous multidimensional sensory landscape within which to communicate. The features of the chemical signaling system of plethodontids characterized to date (prior to and as a result of this research) illustrate each of these mechanisms of signal variability, including levels of structural and combinatorial variation that have previously gone unnoticed or underappreciated. These mechanisms of signal variation are summarized in the following model. While informed by empirical observations, this framework is still by necessity a hypothesis, and should be critically evaluated as such.

Each individual salamander of a population produces many different proteins in various glandular regions of the integument that are adapted for specific functions. Some of these proteins will undoubtedly be involved in basic aspects of salamander biology (e.g. aiding in integumentary respiration, producing antimicrobial peptides, etc.); many others, however, will be involved in specific communicative roles, based on the adaptive specialization of individual glands. These glands are known to differ in the production of broad chemical classes (Hecker et al., 2003), as well as individual proteins (Largen and Woodley, 2008; Chapter 4). These tissues are also implicated in different behavioral phenomena (reviewed in Chapter 4). Many of the semiochemicals released will be adaptive to senders and receivers alike (signals), while others will have been produced for other reasons but are nonetheless informative and adaptive to receivers alone (cues). In either case, they play a role in communication.

In males, the mental gland (MG) produces a diverse assortment of pheromone protein variants that are used in courtship (Rollmann et al., 1999; Watts et al., 2004; Palmer et al., 2005; 2007a; 2010; Wilburn et al., 2012). The cloacal gland (CG) of males specializes in the production of spermatophores (Sever, 2003) and may be involved in intraspecific, intersexual chemical communication via seminal proteins; it may also play a broader role in chemical signaling through external secretions (e.g. via the vent gland; Jaeger, 1986), as implied by behavioral experiments (Table 1.1). The postcloacal gland (PCG) is thought to be the main source of territorial scent marks (Simons and Felgenhauer, 1992; Simons et al., 1999), through the production of proteins that act directly as signals and/or through those that affect carbohydrate signaling, the other major component of the PCG (Hecker et al., 2003). The dorsal tail base (DTB) is responsible for antipredator secretions (Largen and Woodley, 2008; von Byern et al., 2015), but might also produce proteinaceous courtship pheromones (Sever, 2003). In all cases of communication by these glands, the biochemical composition of the signal should dictate their role, as the specific chemicals produced therein are responsible and adapted for their corresponding function. Thus, within the body of an individual salamander, there is significant variation in chemical signal production by different glands that affects their ultimate function (Chapter 4).

Even in discussing a single protein produced by just one of these glands, it becomes clear that there is tremendous diversity within individuals, between individuals, and between lineages. For example, Plethodontid Receptivity Factor (PRF) is present in multiple isoforms in the (male-specific) MG (Watts et al., 2004; Palmer et al., 2005). At the scale of individual males, they differ in the number and form of PRF variants, which can also include several minor components (Chouinard et al., 2013; Chapter 2). The particular combination of isoforms and their levels of expression mean that essentially each male of a population has a unique PRF profile. In the case of species with glycosylated PRFs, different individuals may also be able to produce multiple glycoforms by attaching glycans to different residues of the protein (and different combinations of residues), and by modifying the specific carbohydrate moieties attached to a given residue. This potential for post-translational modifications adds an entire level of signal variability, one that allows for extreme plasticity in structural form and that is known to affect signaling function (Dwek, 1996).

At the level of comparisons between different species and lineages, further structural and combinatorial variations become more pronounced, and include major differences in sequence and structure. For example, PRF orthologs diverge in sequence identity as populations undergo speciation. Furthermore, continuing gene duplications yield paralogous PRF isoforms in a given species that appear after populations diverge (Watts et al., 2004; Palmer et al., 2005; Chapter 3). In addition to tertiary structural variation directly associated with the amino acid sequence, this divergence can ultimately lead to major structural differences between lineages (e.g. the presence/absence of glycosylation; Chapter 3). At even deeper levels of divergence, entire components can be co-opted from other functions and elaborated by positive selection (e.g. PRF in eastern Plethodon), or lost from a signal entirely (e.g. SPF in the MG of Plethodon; Palmer et al., 2007a).

Lastly, emergent properties of signal complexity (e.g. the relative ratios of multiple components) may allow for additional signal specificity (e.g. for species- or individual-specific odors). The interactive effects of signal components may also have consequences on the overall function of the signal. For example, the inclusion of multiple

PMF isoforms is necessary for a reduction in courtship duration (the effect of the total MG secretion; Wilburn et al., 2015). In this case, signal complexity is essential for appropriate function (assuming that is indeed the primary function of PMF; see below). Combinations of PMF isoforms also elicit a greater neural response in the vomeronasal organ than the same amount of a single isoform (Wilburn, 2014). These functional consequences of interactive effects are congruent with the working hypothesis in this system that signal complexity has been selected for in MG pheromones through widespread positive selection.

The result of these many mechanisms of signal variation is that different entities (be they individuals, populations, species, or clades) can differ in many ways regarding the form and function of chemical signals, as a consequence of the selective pressures acting on the genes involved. Obviously the degree to which this variation is functionally relevant depends upon the context for each of the aforementioned groups. Not all of these aspects of variation will necessarily have functional significance, which requires that receivers have a reliable means of discerning the compositional variability, and that it has tangible effects on downstream sensory processing and physiological/behavioral responses. In general, however, there are many ways in which each component, and characteristics of the signal as a whole, can vary. In the context of evolution, variation among entities dictates which are more likely to reproduce; in the case of plethodontid courtship and other social behavior, that variation may be manifested in part through differences in chemical communication.

## Future Directions

A great deal of fascinating research has been done in the last few decades to document the behavioral role and biochemical composition of plethodontid chemical signals, particularly in regards to the production and evolution of several mental gland pheromones (reviewed in Woodley, 2014; 2015). Despite these advances, much remains unknown. In many cases, more research is needed; in others, prior conclusions could benefit from revisitation or alternative explanations. In this section, I outline a few particular questions to consider and issues to address in future work, to supplement those
discussed in specific chapters. I provide my own perspective on these topics as informed by the evidence, and disclose particular avenues in need of further research.

## The Function of Mental Gland Pheromones

The longstanding hypothesis for the function of mental gland pheromones states that they "increase female receptivity" (Houck and Reagan, 1990; Houck, 1998; 2009), as measured by the observation of a moderate decrease in courtship duration in the laboratory. This hypothesis has been equally applied to all three MG proteins identified, but these pheromones evolved at different times, act on different target tissues between lineages, and demonstrate different patterns of selection in these lineages (Palmer et al., 2005; 2007a). The hypothesis that MG proteins evolved to increase female receptivity hinges on the assumption that they affect a female's propensity to complete courtship, pick up the spermatophore, and/or use the corresponding sperm to fertilize her eggs. In other words, it assumes that courtships with pheromone applied fail less often, but this has seldom been explicitly addressed in the literature.

In fact, the only experiment to date (to my knowledge) that provides data addressing these assumptions does not support the hypothesized effect. Eddy et al. (2012) showed that there was no difference in insemination rate between males with and without mental glands (and thus pheromone) in P. shermani. Rather, the presence of a MG had a slightly negative (but insignificant) effect on the likelihood of insemination. Secondly, the MG had no effect on the probability of transitioning from the persuasion stage to tailstraddling walk (TSW). The only effect of MG proteins was to decrease the duration of the TSW stage of courtship specifically, but even this depended on the length of the initial persuasion stage. In cases with short persuasion stages, the presence of a MG was actually associated with an increased TSW duration. Overall, male visual behavior ("foot-dancing") in the persuasion stage was the only significant determining factor in (1) transitioning from earlier to later stages of courtship, and (2) insemination success. This transition from the persuasion stage to TSW is clearly the bottleneck in female receptivity: only $22 \%$ of encounters transitioned to TSW, whereas $91 \%$ of TSW encounters resulted in insemination. As males mostly deliver pheromone during TSW, it
seems less likely that MG proteins evolved to increase female receptivity; that is, having already made it through the major receptivity bottleneck, these proteins may carry out more nuanced physiological functions relating to subsequent reproductive processes.

These observations do not provide strong evidence for the notion that MG pheromones act to increase male fitness by manipulating female receptivity, since males are no less likely to inseminate a female as a result of their MG pheromones. It may well be that the observation of reduced duration of particular courtship stages in the lab is an epiphenomenon of the primary adaptive function of these different proteins, which may differ between MG components. For these and other reasons, it is reasonable to consider alternative explanations.

Herein, I present two additional hypotheses for the potential function of MG pheromones: (1) modulation of female appetite and/or metabolism, and (2) reproductive priming for sperm uptake and storage. The former can be thought of as a refinement of the current hypothesis for MG function, in that it provides a proximate mechanism for decreasing the likelihood of females abandoning mating. This hypothesis has been proposed in previous research (Kiemnec-Tyburczy et al., 2009; Vaccaro et al., 2009), but herein I elaborate on the evolutionary pressures and clarify its rationale. The second hypothesis refers to a potential role in post-copulatory selective pressure, and has never been proposed to my knowledge in this system. These hypotheses are not mutuallyexclusive due to (1) the presence of multiple pheromone components that may differ in function, and (2) the potential for pleiotropic effects of a single protein's function.

The first hypothesis proposes that male-produced pheromones may suppress or modify female appetite, as the selective pressure to forage is perhaps far greater than the need to mate on any given night. Essentially, bouts of activity on the surface during wet weather are a precious commodity for plethodontids. Despite being largely fossorial, the vast majority of plethodontid foraging occurs on the surface, but this activity is limited by rainfall. Individuals restricted to sub-surface foraging during inevitable prolonged dry periods are subjected to negative energy budgets, sometimes with essentially no food in their guts until the next wet weather (Jaeger, 1972; 1978; 1980; Fraser, 1976).

In addition to standard energetic requirements, female fitness is mainly limited by body size and the amount of energy she can acquire during a season, and some authors even attribute this energetic limitation to the observance in some cases of a biennial female reproductive mode (Sayler, 1966; Tilley, 1968; Peacock and Nussbaum, 1973; Nagel, 1977; Lotter, 1978; Fraser, 1980; Verrell, 1995). In addition to the need for sufficient investment into egg development, females also suffer substantial energetic costs during brooding; they require ample fat reserves to aggressively guard their clutch for months on end. This behavior is a crucial component of egg survivability, and thus female fitness (Forester, 1979; Jaeger and Forester, 1993; Ng and Wilbur, 1995).

In contrast, the opportunity for females to mate is pervasive in these species, and their fitness is seldom limited by access to sperm for several reasons: (1) they tend to live in high population densities so search time for mates is low (Mathis, 1991a; Eddy, 2012), (2) reproductive females are often outnumbered by males (in a roughly $2: 1$ ratio) due to biennial cycles (Sayler, 1966), (3) mating seasons are long, moderately synchronous, and provide many chances to mate (Petranka, 1998), and (4) they can store viable sperm for well over a year, long enough to span two mating seasons (Eddy et al., 2015).

Male fitness, on the contrary, is limited in practice mostly by the number of matings they can attain (although after several matings spermatophore production can become limiting; reviewed in Houck and Verrell, 1993). As such, males should be selected to maximize the number of successful courtships (but for a less simplistic view of male mate choice see Eddy et al., 2016). As noted, the main female motivation during suitable weather may be foraging, but this time is also a prime opportunity for individuals to interact and court. As males have no intromittent organ, do not clasp, nor have means to otherwise coerce unreceptive females, they may have been selected to employ chemical persuasion to suppress the female's primary priority during bouts of surface activity: feeding.

There is already some behavioral evidence for this hypothesis, namely that MG proteins suppress female tendency to feed (Vaccaro et al., 2009), although this warrants replication and continued study. At present there is also preliminary evidence in support of a physiological mechanism for this hypothesis: (1) numerous appetite-modulating
hormone transcripts have been identified from the MG (and other glands) of several plethodontid species (Chapter 3; Kiemnec-Tyburczy et al., 2009; Wilburn, 2014; Doty et al., 2015), and (2) stimulation of vomeronasal neurons (in olfactory delivery species) by MG proteins leads to the activation of the ventromedial hypothalamus via the accessory olfactory system (Laberge et al., 2008). This fact is also cited as justification for the current hypothesis due to the hypothalamic role in modulating reproduction, but this brain region is equally important for satiety and ingestive behavior (King, 2006). These lines of evidence are anecdotal at present, but the significant evolutionary pressures in place suggest that this mechanism warrants serious consideration.

As a proximate mechanism for the female receptivity hypothesis, it again predicts that courtships with pheromone delivery fail less often. The evidence to date suggests that this is not so (Eddy et al., 2012), but future bioassays should test these interactions in more realistic courtship settings. In the laboratory, success rates of TSW may be unrealistically high; the majority (91\%) of TSW ends in insemination, and this was not affected by pheromone delivery. In the field, differences in courtship failure rates with and without pheromone delivery may be more pronounced because of the more complex habitat and the presence of other extrinsic stimuli (e.g. food cues). In contrast, it may be true that a reduction in overall time spent mating is adaptive, due to an increased net foraging time over the course of a mating season. In any case, in order for this hypothesis to stand the test of time, its assumptions should be carefully considered and explicitly addressed in future research.

The second hypothesis proposes that male pheromones may be involved in physiological priming for downstream reproductive events related to sperm uptake and storage. This hypothesis is based largely on the observations by Eddy et al., (2012) that MG pheromones are used primarily after males have passed through the main bottleneck in female receptivity. Essentially, MG pheromones are used mostly in later stages of courtship, at which point females are highly likely to finish the process and pick up the spermatophore. Due to the presence of long term sperm storage, post-copulatory sexual selection may pose a very real evolutionary pressure. In priming the female to pick up and store the spermatophore, physiological manipulation of her immune system (e.g. to
dampen the response to foreign cells) may be highly adaptive. There may also be other physiological (e.g. endocrine) factors related to sperm uptake and storage that males have evolved to manipulate. Given the complexity of post-copulatory processes and its limited investigation in plethodontids, speculating at this point is unwarranted. Instead, the point is that downstream reproductive events may pose strong post-copulatory sexual selective pressures to which males may have developed physiological adaptations via chemical signaling.

Again, these hypotheses are not mutually-exclusive. Likewise, while both address potential selective pressures, these pressures are not necessarily overcome through the use of MG pheromones. Many other glands produce relevant secretory proteins that could be involved in these processes (e.g. seminal proteins produced by cloacal glands). In establishing proposed functions of pheromones, the physiological, neurobiological, and behavioral effects of individual components need to be tested and scrutinized experimentally.

## Target Tissues and Receptors in Different Plethodon Lineages

Another major factor in contemplating the role of MG pheromones is that their function could very well have changed during the transition from dermal to olfactory delivery. The ancestral condition involved "vaccination" of the pheromone through the dermis. This behavior presumably leads to transmission into the bloodstream, but this needs to be confirmed experimentally. This hypothesis was originally posited by Organ (1961):
"The premaxillary teeth must certainly stimulate the female tactually when the male is arching and may even puncture the skin of the female. No evidence of laceration on the back of a female was seen but this does not preclude the possibility of small puncture wounds. The peripheral circulation in plethodontid salamanders is close to the surface of the skin and these blood vessels dilate before courtship. It is conceivable, therefore, that the secretions of the mental gland of the male act on the female through the circulatory system and that the enlarged premaxillary teeth of the male aid the secretions in penetrating the skin of the female. This could be verified by careful physiological and biochemical studies on courting individuals."

This hypothesis is indeed intuitive, based on the presence of sexually-dimorphic premaxillary teeth, and the various "scratching" behaviors demonstrated by the plesiomorphic species (reviewed in Arnold, 1977). And still, the verification Organ mentions has yet to occur despite decades of research on these proteins, which focused on other interesting priorities. Nevertheless, the specific tissues and receptors targeted by these pheromones need to be identified before we can even begin to understand their true function.

Compared to the ancestral state, in one specific clade (the $P$. glutinosus group) there has been a shift to delivering pheromone via the nares to the vomeronasal organ (VNO). In this case, MG pheromones stimulate neurons in the VNO (Wirsig-Wiechmann et al., 2002; 2006), which in turn act on the accessory olfactory system and downstream regions of the central nervous system (Laberge et al., 2008). The receptor proteins stimulated in the VNO are unknown (but see Kiemnec-Tyburczy et al., 2011ab). Three potential hypotheses are that (1) expression of the plesiomorphic receptors have been coopted into the dendrites of vomeronasal neurons, (2) the MG proteins have shifted to targeting generalist vomeronasal receptors (e.g. vomeronasal type-2 receptors), or (3) other unknown receptors are utilized. In any case, due to the different pathways for signal processing between dermal and olfactory delivery, the downstream effects of pheromone activation may be very different between groups. This notion is supported by the observation that dermal application of MG pheromones does not carry out the observed effect (reduced courtship duration) in $P$. shermani (Kiemnec-Tyburczy et al., 2011c).

This transition from dermal to olfactory delivery is poorly understood in terms of the phylogenetic relationships between intermediate species (Kozak et al., 2006; Wiens et al., 2006; Pyron and Wiens, 2011; Highton et al., 2012), and the exact mode of pheromone delivery. Males in these species have morphological traits indicative of olfactory delivery (lack of premaxillary teeth and enlarged posterior mental glands), but canonical olfactory delivery is less prevalent (Palmer et al., 2007a). Differing degrees of olfactory delivery have been observed in the courtship of two intermediate species: $P$. dorsalis and P. angusticlavius (Picard, 2005; Dyal, 2006). Yet, the other intermediate species observed ( $P$. welleri) does not exhibit any olfactory pheromone delivery
behaviors (Organ, 1960). All three species have prolonged periods of contact with the mental gland and the skin of the female, which can include the head region. Picard (2005) argued that the stereotypical olfactory delivery in the P. glutinosus group via "slapping" may have evolved from these "head rubbing" behaviors observed in this transitional group.

Over the course of this transition there have been corresponding changes in the composition of MG secretions, again indicating a possible shift in functionality. When dermal delivery behaviors originated, the MG consisted of a different combination of proteins (i.e. predominantly SPF/PMF) compared to the eastern Plethodon (i.e. predominantly PMF/PRF). Due to the large (and varying) number of gene duplications observed in individual components (Watts et al., 2004; Palmer et al., 2005; 2007a; 2010), the ancestral signal also presumably differed in the number and identity of each component, as well as overall complexity. The type of selection acting on MG protein sequences also differs between components and delivery mode. For example, PRF in eastern Plethodon shows signs of widespread positive selection, while SPF shows signs of genetic drift constrained by purifying selection. In the intermediate group, this trend is reversed (Palmer et al., 2005; 2007a).

All of these issues suggest that the function of MG pheromones may have changed between species and modes of delivery, but the selective pressures involved in that transition are unclear. Future work should focus on (1) characterizing the target tissues and receptors used by each protein, and (2) identifying how these targets differ between lineages spanning the transition from dermal to olfactory delivery. Identifying the specific targets of pheromone action will inform new hypotheses concerning the physiological and behavioral functions of MG pheromones. At that point, a comparative approach will undoubtedly shed light on the changes that took place in this multilevel phenotypic complex over the course of this transition, in what was clearly a significant step in the evolution of eastern Plethodon.

## The Role of Other Signaling Glands

For practical reasons, almost all of the biochemical analyses of plethodontid pheromones to date have investigated the production of proteins from the mental gland. As far as we know, this gland is used mostly for one specific purpose (courtship), which constitutes a relatively narrow degree of the total role of chemical signals in the social behavior of plethodontids. And yet, even this one function relies on an incredibly complex mixture of proteins with captivating evolutionary histories. The same is likely true for the many other tissues involved in the great diversity of inter- and intraspecific behaviors carried out by semiochemicals, and which convey complex information. When contemplating the scope of these other functions, we have barely begun to understand the true complexity of the chemical signals produced for these ends. In addition to continued research into mental gland pheromones, complementary approaches should focus on the other glands involved in broader types of communication. In this way, we can hope to gain a more comprehensive picture of the role of chemical signals in the lives of these early-diverging terrestrial vertebrates.

The current work has made a significant first step in this direction by characterizing the transcriptome of these glands, but a great deal remains to be done. First, deeper sequencing and more rigorous expression analyses will allow for a better view of which glands produce which proteins. As protein expression is what ultimately matters for functionality, proteomic analyses are needed to confirm and elaborate upon these findings. Ideally these experiments should also include methods that are minimally invasive, so as to identify only the proteins available to receivers (in contrast to those involved in intracellular processes). Together, these approaches will: (1) confirm protein sequences by using the transcriptome as a reference for proteomic analyses, (2) determine which proteins are actively secreted by each gland to the external environment, and (3) identify differences in protein production between glands at the level of the phenotype. Lastly, chemical analyses should also focus on the other major component of these signals: carbohydrates. Other types of molecules should not be ruled out (e.g. lipids, biogenic amines, etc.). In essence, we need to continue identifying the many molecules being used to communicate, be they small organic molecules, proteins, carbohydrates, or
both (glycoproteins). In addition to these biochemical directives, behavioral and physiological assays will be needed to conclude that any identified molecules are involved in signaling. Having done so, more sophisticated questions can be posed that address the way in which different characteristics of signal complexity relate to the information encoded and/or their effect(s) on receivers.

These tasks are admittedly challenging, but having identified the core repertoire of chemical signaling, this system could prove to be a highly tractable one in which to investigate fundamental questions about communication. Individuals are locally abundant, and are easily captured, housed, and manipulated. In form, they share the core homologous vertebrate body plan and physiology, so applications to humans and other vertebrate systems are a possibility in some cases. Salamanders also rely heavily on chemical communication, and these stimuli can be tested in isolation; this can help eliminate confounding variables for experimentation. These species also provide many opportunities to study the evolution of communication systems. For those that live in sympatry, it raises great potential to investigate how different organisms in the same general niche utilize different sensory space to communicate, or in contrast, how these communication systems overlap and interact. For species in allopatry, it enables the study of how communication systems diverge. Because these signals are involved in territoriality, species recognition, and reproduction, this may allow for detailed investigations into the role of communication in speciation. Many species also hybridize with one another, affording different types of comparisons between lineages at different stages in this process. These reasons and more make this system a promising one in which to investigate chemical signaling and communication in general. Thus, it is my hope that future endeavors will continue to explore this remarkable branch in the tree of life and the field of animal behavior.

## SUMMARY

This research set out to document additional aspects of signal complexity in the semiochemicals produced by plethodontid salamanders. This was achieved by exploring (1) individual (intraspecific) variation in a specific pheromonal signal, (2) intra- and interspecific variation in post-translational structural modification (glycosylation) of a major plethodontid pheromone, and (3) organismal-level variation in the genes expressed by different regions of the body for different behavioral functions. In all cases, variation between signals may have profound proximate (mechanistic) and ultimate (evolutionary) consequences. Over the course of this work, the behavioral roles and composition of plethodontid chemical signals are reviewed, the conclusions of the present research are integrated with previous findings, several new and existing hypotheses are discussed concerning the functional implications of signal variation, and particular directions for future research are examined.

## BIBLIOGRAPHY

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., \& Lipman, D. J. (1990). Basic local alignment search tool. Journal of molecular biology, 215(3), 403-410.

AmphibiaWeb. Retrieved June, 2016, from http://www.amphibiaweb.org/
Andersson, M. B. (1994). Sexual selection. Princeton University Press.
Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Anthony, C. D. (1993). Recognition of conspecific odors by Plethodon caddoensis and P. ouachitae. Copeia, 1028-1033.

Apweiler, R., Hermjakob, H., \& Sharon, N. (1999). On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. Biochimica et Biophysica Acta (BBA)-General Subjects, 1473(1), 4-8.

Arnold, S. J. (1976). Sexual behavior, sexual interference and sexual defense in the salamanders Ambystoma maculatum, Ambystoma tigrinum and Plethodon jordani. Zeitschrift für Tierpsychologie, 42(3), 247-300.

Arnold, S. J. (1977). The evolution of courtship behavior in New World salamanders with some comments on Old World salamandrids. In The reproductive biology of amphibians (pp. 141-183). Springer US.

Arnold, S. J., Reagan, N. L., \& Verrell, P. A. (1993). Reproductive isolation and speciation in plethodontid salamanders. Herpetologica, 49(2), 216-228.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., ... \& Harris, M. A. (2000). Gene Ontology: tool for the unification of biology. Nature genetics, 25(1), 25-29.

Avila, F. W., Sirot, L. K., LaFlamme, B. A., Rubinstein, C. D., \& Wolfner, M. F. (2011). Insect seminal fluid proteins: identification and function. Annual review of entomology, 56, 21.

Bansil, R., \& Turner, B. S. (2006). Mucin structure, aggregation, physiological functions and biomedical applications. Current Opinion in Colloid \& Interface Science, 11(2), 164170.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., ... \& Studholme, D. J. (2004). The Pfam protein families database. Nucleic acids research, 32(suppl 1), D138-D141.

Bathgate, R. A. D., Halls, M. L., Van Der Westhuizen, E. T., Callander, G. E., Kocan, M., \& Summers, R. J. (2013). Relaxin family peptides and their receptors. Physiological reviews, 93(1), 405-480.

Beynon, R. J., Veggerby, C., Payne, C. E., Robertson, D. H., Gaskell, S. J., Humphries, R. E., \& Hurst, J. L. (2002). Polymorphism in major urinary proteins: molecular heterogeneity in a wild mouse population. Journal of chemical ecology, 28(7), 14291446.

Blom, N., Sicheritz-Pontén, T., Gupta, R., Gammeltoft, S., \& Brunak, S. (2004). Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. Proteomics, 4(6), 1633-1649.

Bohne-Lang, A., \& von der Lieth, C. W. (2005). GlyProt: in silico glycosylation of proteins. Nucleic acids research, 33(suppl 2), W214-W219.

Boulanger, M. J., Chow, D. C., Brevnova, E. E., \& Garcia, K. C. (2003). Hexameric structure and assembly of the interleukin-6/IL-6 $\alpha$-receptor/gp130 complex. Science, 300(5628), 2101-2104.

Bradbury, J. W., \& Vehrencamp, S. L. (1998). Principles of animal communication. Sinaeur Associates.

Bray, N., Pimentel, H., Melsted, P., \& Pachter, L. (2015). Near-optimal RNA-Seq quantification. arXiv preprint arXiv:1505.02710.

Briand, L., Huet, J. C., Perez, V., Lenoir, G., Nespoulous, C., Boucher, Y., ... \& Pernollet, J. C. (2000). Odorant and pheromone binding by aphrodisin, a hamster aphrodisiac protein. FEBS letters, 476(3), 179-185.

Burton, T. M., \& Likens, G. E. (1975). Salamander populations and biomass in the Hubbard Brook experimental forest, New Hampshire. Copeia, 541-546.

Bushnell, B. (2015). BBMap short read aligner. http://sourceforge.net/projects/bbmap.
Calenge, C. (2006). The package "adehabitat" for the R software: a tool for the analysis of space and habitat use by animals. Ecological modelling, 197(3), 516-519.

Capshaw, G., \& Soares, D. (2016). Hearing in Plethodontid Salamanders: A Review. Copeia, 104(1), 157-164.

Caragea, C., Sinapov, J., Silvescu, A., Dobbs, D., \& Honavar, V. (2007). Glycosylation site prediction using ensembles of Support Vector Machine classifiers. BMC bioinformatics, 8(1), 438.

Cavaggioni, A., \& Mucignat-Caretta, C. (2000). Major urinary proteins, a 2U-globulins and aphrodisin. Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology, 1482(1), 218-228.

Chalaris, A., Garbers, C., Rabe, B., Rose-John, S., \& Scheller, J. (2011). The soluble Interleukin 6 receptor: generation and role in inflammation and cancer. European journal of cell biology, 90(6), 484-494.

Chouinard, AJ (2010) Investigating pheromone variability in a terrestrial woodland salamander. M.S. Thesis, University of New Hampshire

Chouinard, A. J. (2012). Rapid onset of mate quality assessment via chemical signals in a woodland salamander (Plethodon cinereus). Behavioral Ecology and Sociobiology, 66(5), 765-775.

Chouinard, A. J., Wilburn, D. B., Houck, L. D., \& Feldhoff, R. C. (2013). Individual variation in pheromone isoform ratios of the red-legged salamander, Plethodon shermani. In Chemical Signals in Vertebrates 12 (pp. 99-115). Springer New York. [Chapter 4]

Clarke, B. T. (1997). The natural history of amphibian skin secretions, their normal functioning and potential medical applications. Biological Reviews of the Cambridge Philosophical Society, 72(03), 365-379.

Coyne, J. A., \& Orr, H. A. (1997). " Patterns of speciation in Drosophila" revisited. Evolution, 51(1), 295-303.

Cruz, C. S. D., Lee, Y., Viswanathan, S. R., El-Guindy, A. S., Gerlach, J., Nikiforow, S., Shedd, D., Gradoville, L., \& Miller, G. (2004). N-linked glycosylation is required for optimal function of Kaposi's sarcoma herpesvirus-encoded, but not cellular, interleukin 6. The Journal of experimental medicine, 199(4), 503-514.

Cummins, S. F., Nichols, A. E., Amare, A., Hummon, A. B., Sweedler, J. V., \& Nagle, G. T. (2004). Characterization of Aplysia enticin and temptin, two novel water-borne protein pheromones that act in concert with attractin to stimulate mate attraction. Journal of Biological Chemistry, 279(24), 25614-25622.

Dalton, B., \& Mathis, A. (2014). Identification of sex and parasitism via pheromones by the Ozark zigzag salamander. Chemoecology, 24(5), 189-199.

Dantzer, B. J., \& Jaeger, R. G. (2007a). Detection of the sexual identity of conspecifics through volatile chemical signals in a territorial salamander. Ethology, 113(3), 214-222.

Dantzer, B. J., \& Jaeger, R. G. (2007b). Male red-backed salamanders can determine the reproductive status of conspecific females through volatile chemical signals. Herpetologica, 63(2), 176-183.

Dawley, E. M. (1984a). Recognition of individual, sex and species odours by salamanders of the Plethodon glutinosus-P. jordani complex. Animal Behaviour, 32(2), 353-361.

Dawley, E. M. (1984b). Identification of sex through odors by male red-spotted newts, Notophthalmus viridescens. Herpetologica, 101-105.

Dawley, E. M. (1986a). Evolution of chemical signals as a premating isolating mechanism in a complex of terrestrial salamanders. In Chemical Signals in Vertebrates 4 (pp. 221-224). Springer US, New York.

Dawley, E. M. (1986b). Behavioral isolating mechanisms in sympatric terrestrial salamanders. Herpetologica, 156-164.

Dawley, E. M., \& Bass, A. H. (1988). Organization of the vomeronasal organ in a plethodontid salamander. Journal of Morphology, 198(2), 243-255.

Dawley, E. M., \& Bass, A. H. (1989). Chemical access to the vomeronasal organs of a plethodontid salamander. Journal of Morphology, 200(2), 163-174.

Dickens, J. C. (1999). Predator-prey interactions: olfactory adaptations of generalist and specialist predators. Agricultural and Forest Entomology, 1(1), 47-54.

Dillies, M. A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., ... \& Guernec, G. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Briefings in bioinformatics, 14(6), 671-683.

Doty, K. A., Wilburn, D. B., Bowen, K. E., Feldhoff, P. W., \& Feldhoff, R. C. (2015). Co-option and evolution of non-olfactory proteinaceous pheromones in a terrestrial lungless salamander. Journal of proteomics 135, 101-111.

Drozdetskiy, A., Cole, C., Procter, J., \& Barton, G. J. (2015). JPred4: a protein secondary structure prediction server. Nucleic acids research, 43(W1), W389-W394.

Duellman W.E., \& Trueb, L. (1994). Biology of amphibians. Johns Hopkins University Press.

Dwek, R. A. (1996). Glycobiology: toward understanding the function of sugars. Chemical Reviews, 96(2), 683-720.

Dyal, L. A. (2006). Novel courtship behaviors in three small eastern Plethodon species. Journal of herpetology, 40(1), 55-65.

Eddy, S. L. (2012). Mutual mate choice in a terrestrial salamander, Plethodon shermani, with long-term sperm storage. PhD dissertation, Department of Zoology, Oregon State University.

Eddy, S. L., Kiemnec-Tyburczy, K. M., Uyeda, J. C., \& Houck, L. D. (2012). The influence of sequential male courtship behaviors on courtship success and duration in a terrestrial salamander, Plethodon shermani. Ethology, 118(12), 1240-1250.

Eddy, S. L., Vaccaro, E. A., Baggett, C. L., Kiemnec-Tyburczy, K. M., \& Houck, L. D. (2015). Sperm Mass Longevity and Sperm Storage in the Female Reproductive Tract of Plethodon shermani (Amphibia: Plethodontidae). Herpetologica, 71(3), 177-183.

Eddy, S. L., Wilburn, D. B., Chouinard, A. J., Doty, K. A., Kiemnec-Tyburczy, K. M., \& Houck, L. D. (2016). Male terrestrial salamanders demonstrate sequential mate choice based on female gravidity and size. Animal Behaviour, 113, 23-29.

Feldhoff, R. C., Rollmann, S. M., \& Houck, L. D. (1999). Chemical analysis of courtship pheromones in a plethodontid salamander. In Advances in chemical signals in vertebrates (pp. 117-125). Springer US.

Feng, M., Fang, Y., Han, B., Xu, X., Fan, P., Hao, Y., Qi, Y., Hu, H., Huo, X., Meng, L., Wu, B., \& Li, J. (2015). In-Depth N-Glycosylation Reveals Species-Specific Modifications and Functions of the Royal Jelly Protein from Western (Apis mellifera) and Eastern Honeybees (Apis cerana). Journal of proteome research, 14(12), 5327-5340.

Finn, R. D., Clements, J., \& Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic acids research, gkr367.

Fontana, M. F., Ask, K. A., Macdonald, R. J., Carnes, A. M., \& Staub, N. L. (2006). Loss of traditional mucous glands and presence of a novel mucus-producing granular gland in the plethodontid salamander Ensatina eschscholtzii. Biological Journal of the Linnean Society, 87(3), 469-477.

Fontana, M. F., Houck, L. D., \& Staub, N. L. (2007). In situ localization of plethodontid courtship pheromone mRNA in formalin-fixed tissue. General and comparative endocrinology, 150(3), 480-485.

Forester, D. C. (1979). The adaptiveness of parental care in Desmognathus ochrophaeus (Urodela: Plethodontidae). Copeia, 332-341.

Fraser, D. F. (1976). Empirical evaluation of the hypothesis of food competition in salamanders of the genus Plethodon. Ecology, 459-471.

Fraser, D. F. (1980). On the environmental control of oocyte maturation in a plethodontid salamander. Oecologia, 46(3), 302-307.

Ghazarian, H., Idoni, B., \& Oppenheimer, S. B. (2011). A glycobiology review: carbohydrates, lectins and implications in cancer therapeutics. Acta histochemica, 113(3), 236-247.

Gibbs, G. M., Roelants, K., \& O'Bryan, M. K. (2008). The CAP superfamily: cysteinerich secretory proteins, antigen 5 , and pathogenesis-related 1 proteins-roles in reproduction, cancer, and immune defense. Endocrine reviews, 29(7), 865-897.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., ... \& Chen, Z. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology, 29(7), 644-652.

Gross, V., Andus, T., Castell, J., Vom Berg, D., Heinrich, P. C., \& Gerok, W. (1989). O -and N -glycosylation lead to different molecular mass forms of human monocyte interleukin-6. FEBS letters, 247(2), 323-326.

Gupta, R., Jung, E., \& Brunak, S. (2004). NetNGlyc 1.0 Server. Center for biological sequence analysis, technical university of Denmark available from: http://www. cbs. dtu dk/services/NetNGlyc.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., ... \& MacManes, M. D. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature protocols, 8(8), 1494-1512.

Haga, S., Hattori, T., Sato, T., Sato, K., Matsuda, S., Kobayakawa, R., ... \& Touhara, K. (2010). The male mouse pheromone ESP1 enhances female sexual receptive behaviour through a specific vomeronasal receptor. Nature, 466(7302), 118-122.

Hairston, N. G. (1983). Growth, survival and reproduction of Plethodon jordani: tradeoffs between selective pressures. Copeia, 1024-1035.

Hallmann, A. (2003). Extracellular matrix and sex-inducing pheromone in Volvox. International review of cytology, 227, 131-182.

Hamby, S. E., \& Hirst, J. D. (2008). Prediction of glycosylation sites using random forests. BMC bioinformatics, 9(1), 1.

He, J., Ma, L., Kim, S., Schwartz, J., Santilli, M., Wood, C., ... \& Yu, C. R. (2010). Distinct signals conveyed by pheromone concentrations to the mouse vomeronasal organ. The Journal of Neuroscience, 30(22), 7473-7483.

Hecker, L., Madison, D. M., Dapson, R. W., \& Holzherr, V. (2003). Presence of modified serous glands in the caudal integument of the red-backed salamander (Plethodon cinereus). Journal of herpetology, 37(4), 732-736.

Hefetz, A., \& Graur, D. (1988). The significance of multicomponent pheromones in denoting specific compositions. Biochemical systematics and ecology, 16(6), 557-566.

Heinrich, P. C., Behrmann, I., Müller-Newen, G., Schaper, F., \& Graeve, L. (1998). Interleukin-6-type cytokine signalling through the gp130/Jak/STAT pathway. Biochemical Journal, 334(2), 297-314.

Highton, R. (1999). Geographic protein variation and speciation in the salamanders of the Plethodon cinereus group with the description of two new species. Herpetologica, 43-90.

Highton, R., \& Peabody, R. B. (2000). Geographic protein variation and speciation in salamanders of the Plethodon jordani and Plethodon glutinosus complexes in the southern Appalachian Mountains with the description of four new species. In The biology of plethodontid salamanders (pp. 31-93). Springer US.

Highton, R., Hastings, A. P., Palmer, C., Watts, R., Hass, C. A., Culver, M., \& Arnold, S. J. (2012). Concurrent speciation in the eastern woodland salamanders (Genus Plethodon): DNA sequences of the complete albumin nuclear and partial mitochondrial 12s genes. Molecular phylogenetics and evolution, 63(2), 278-290.

Houck, L. D. (1998). Integrative studies of amphibians: from molecules to mating. American Zoologist, 38(1), 108-117.

Houck, L. D. (2009). Pheromone communication in amphibians and reptiles. Annual Review of Physiology, 71, 161-176.

Houck, L. D., \& Reagan, N. L. (1990). Male courtship pheromones increase female receptivity in a plethodontid salamander. Animal Behaviour, 39(4), 729-734.

Houck, L. D., \& Verrell, P. A. (1993). Studies of courtship behavior in plethodontid salamanders: a review. Herpetologica, 49(2), 175-184.

Houck, L. D., Bell, A. M., Reagan-Wallin, N. L., \& Feldhoff, R. C. (1998). Effects of experimental delivery of male courtship pheromones on the timing of courtship in a terrestrial salamander, Plethodon jordani (Caudata: Plethodontidae). Copeia, 1998(1), 214-219.

Houck, L. D., Palmer, C. A., Watts, R. A., Arnold, S. J., Feldhoff, P. W., \& Feldhoff, R. C. (2007). A new vertebrate courtship pheromone, PMF, affects female receptivity in a terrestrial salamander. Animal Behaviour, 73(2), 315-320.

Houck, L. D., Watts, R. A., Mead, L. M., Palmer, C. A., Arnold, S. J., Feldhoff, P. W., \& Feldhoff, R. C. (2008a). A candidate vertebrate pheromone, SPF, increases female receptivity in a salamander. In Chemical Signals in Vertebrates 11 (pp. 213-221). Springer New York.

Houck, L. D., Watts, R. A., Arnold, S. J., Bowen, K. E., Kiemnec, K. M., Godwin, H. A., Feldhoff, P.W., \& Feldhoff, R. C. (2008b). A recombinant courtship pheromone affects sexual receptivity in a plethodontid salamander. Chemical senses, 33(7), 623-631.

Hurst, J. L. (2009). Female recognition and assessment of males through scent. Behavioural brain research, 200(2), 295-303.

Hurst, J. L., Robertson, D. H., Tolladay, U., \& Beynon, R. J. (1998). Proteins in urine scent marks of male house mice extend the longevity of olfactory signals. Animal behaviour, 55(5), 1289-1297.

Hurst, J. L., Payne, C. E., Nevison, C. M., Marie, A. D., Humphries, R. E., Robertson, D. H., Cavaggioni, A., \& Beynon, R. J. (2001). Individual recognition in mice mediated by major urinary proteins. Nature, 414(6864), 631-634.

Hurst, J. L., \& Beynon, R. J. (2004). Scent wars: the chemobiology of competitive signalling in mice. Bioessays, 26(12), 1288-1298.

Huyton, T., Zhang, J. G., Luo, C. S., Lou, M. Z., Hilton, D. J., Nicola, N. A., \& Garrett, T. P. (2007). An unusual cytokine: Ig-domain interaction revealed in the crystal structure of leukemia inhibitory factor (LIF) in complex with the LIF receptor. Proceedings of the National Academy of Sciences, 104(31), 12737-12742.

Jaeger, R. G. (1971). Competitive exclusion as a factor influencing the distributions of two species of terrestrial salamanders. Ecology, 632-637.

Jaeger, R. G. (1972). Food as a limited resource in competition between two species of terrestrial salamanders. Ecology, 535-546.

Jaeger, R. G. (1974). Competitive exclusion: comments on survival and extinction of species. BioScience, 24(1), 33-39.

Jaeger, R. G. (1978). Plant climbing by salamanders: periodic availability of plantdwelling prey. Copeia, 1978: 686-691.

Jaeger, R. G. (1980). Fluctuations in prey availability and food limitation for a terrestrial salamander. Oecologia, 44(3), 335-341.

Jaeger, R. G. (1981). Dear enemy recognition and the costs of aggression between salamanders. American Naturalist, 962-974.

Jaeger, R. G. (1984). "Agonistic behavior of the red-backed salamander." Copeia 1984: 309-314.

Jaeger, R. G. (1986). Pheromonal markers as territorial advertisement by terrestrial salamanders. In Chemical signals in vertebrates 4 (pp. 191-203). Springer US.

Jaeger, R. G., \& Gergits, W. F. (1979). Intra-and interspecific communication in salamanders through chemical signals on the substrate. Animal Behaviour, 27, 150-156.

Jaeger, R. G., Goy, J. M., Tarver, M., \& Márquez, C. E. (1986). Salamander territoriality: pheromonal markers as advertisement by males. Animal Behaviour, 34(3), 860-864.

Jaeger, R. G., Fortune, D., Hill, G., Palen, A., \& Risher, G. (1993). Salamander homing behavior and territorial pheromones: alternative hypotheses. Journal of herpetology, 27(2), 236-239.

Jaeger, R. G., \& Forester, D. C. (1993). Social behavior of plethodontid salamanders. Herpetologica, 49(2), 163-175.

Jaeger, R. G., \& Gabor, C. R. (1993). Intraspecific chemical communication by a territorial salamander via the postcloacal gland. Copeia, 1993(4), 1171-1174.

Janssenswillen, S., Vandebergh, W., Treer, D., Willaert, B., Maex, M., Van Bocxlaer, I., \& Bossuyt, F. (2014). Origin and diversification of a salamander sex pheromone system. Molecular biology and evolution, msu316.

Julenius, K., Mølgaard, A., Gupta, R., \& Brunak, S. (2005). Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites. Glycobiology, 15(2), 153-164.

Karlson, P., \& Lüscher, M. (1959). 'Pheromones': a new term for a class of biologically active substances. Nature, 183, 55-56.

Karuzas, J. M., Maerz, J. C., \& Madison, D. M. (2004). An alternative hypothesis for the primary function of a proposed mate assessment behaviour in red-backed salamanders. Animal Behaviour, 68(3), 489-494.

Katsuri, L., Hegang, C. H. E. N., \& Shakin-Eshleman, S. H. (1997). Regulation of Nlinked core glycosylation: use of a site-directed mutagenesis approach to identify Asn-Xaa-Ser/Thr sequons that are poor oligosaccharide acceptors. Biochemical Journal, 323(2), 415-419.

Kiemnec-Tyburczy, K. M., Watts, R. A., Gregg, R. G., von Borstel, D., \& Arnold, S. J. (2009). Evolutionary shifts in courtship pheromone composition revealed by EST analysis of plethodontid salamander mental glands. Gene, 432(1), 75-81.

Kiemnec-Tyburczy, K. M., Woodley, S. K., Watts, R. A., Arnold, S. J., \& Houck, L. D. (2011a). Expression of vomeronasal receptors and related signaling molecules in the nasal cavity of a caudate amphibian (Plethodon shermani). Chemical senses, bjr105.

Kiemnec-Tyburczy, K. M., Watts, R. A., \& Arnold, S. J. (2011b). Characterization of two putative cytokine receptors, gp130 and ciliary neurotrophic factor receptor, from terrestrial salamanders. Genes \& genetic systems, 86(2), 131-137.

Kiemnec-Tyburczy, K. M., Woodley, S. K., Feldhoff, P. W., Feldhoff, R. C., \& Houck, L. D. (2011c). Dermal application of courtship pheromones does not influence receptivity in female red-legged salamanders (Plethodon shermani). Journal of Herpetology, 45(2), 169-173.

Kikuyama, S., Toyoda, F., Ohmiya, Y., Matsuda, K., Tanaka, S., \& Hayashi, H. (1995). Sodefrin: a female-attracting peptide pheromone in newt cloacal glands. Science, 267(5204), 1643-1645.

Kikuyama, S., \& Toyoda, F. (1999). Sodefrin: a novel sex pheromone in a newt. Reviews of reproduction, 4(1), 1-4.

Kikuyama, S., Yamamoto, K., Iwata, T., \& Toyoda, F. (2002). Peptide and protein pheromones in amphibians. Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology, 132(1), 69-74.

King, B. M. (2006). The rise, fall, and resurrection of the ventromedial hypothalamus in the regulation of feeding behavior and body weight. Physiology \& behavior, 87(2), 221244.

Kleeberger, S. R., \& Werner, J. K. (1982). Home range and homing behavior of Plethodon cinereus in northern Michigan. Copeia, 409-415.

Kohn, N. R., \& Jaeger, R. G. (2009). Male salamanders remember individuals based on chemical or visual cues. Behaviour, 146(11), 1485-1498.

Koppers, A. J., Reddy, T., \& O’Bryan, M. K. (2011). The role of cysteine-rich secretory proteins in male fertility. Asian J Androl, 13(1), 111-117.

Kozak, K. H., Weisrock, D. W., \& Larson, A. (2006). Rapid lineage accumulation in a non-adaptive radiation: phylogenetic analysis of diversification rates in eastern North American woodland salamanders (Plethodontidae: Plethodon). Proceedings of the Royal Society of London B: Biological Sciences, 273(1586), 539-546.

Krogh, A., Larsson, B., Von Heijne, G., \& Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. Journal of molecular biology, 305(3), 567-580.

Laberge, F., Feldhoff, R. C., Feldhoff, P. W., \& Houck, L. D. (2008). Courtship pheromone-induced c-Fos-like immunolabeling in the female salamander brain. Neuroscience, 151(2), 329-339.

Largen, W., \& Woodley, S. K. (2008). Cutaneous tail glands, noxious skin secretions, and scent marking in a terrestrial salamander (Plethodon shermani). Herpetologica, 64(3), 270-280.

Laurence, M. (2013). PSP94, what is it good for?. Shipshaw Labs.
Law, J. H., \& Regnier, F. E. (1971). Pheromones. Annual review of biochemistry, 40(1), 533-548.

Leinders-Zufall, T., Lane, A. P., Puche, A. C., Ma, W., Novotny, M. V., Shipley, M. T., \& Zufall, F. (2000). Ultrasensitive pheromone detection by mammalian vomeronasal neurons. Nature, 405(6788), 792-796.

Li, W., \& Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics, 22(13), 1658-1659.

Linn Jr, C. E., Campbell, M. G., \& Roelofs, W. L. (1986). Male moth sensitivity to multicomponent pheromones: critical role of female-released blend in determining the functional role of components and active space of the pheromone. Journal of Chemical Ecology, 12(3), 659-668.

Linn, C. E., Campbell, M. G., \& Roelofs, W. L. (1987). Pheromone components and active spaces: what do moths smell and where do they smell it?. Science, 237(4815), 650652.

Linn, C. E., \& Roelofs, W. L. (1989). Response specificity of male moths to multicomponent pheromones. Chemical Senses, 14(3), 421-437.

Lonsdale, D. J., Frey, M. A., \& Snell, T. W. (1998). The role of chemical signals in copepod reproduction. Journal of Marine Systems, 15(1), 1-12.

Lotter, F. (1978). Reproductive ecology of the salamander Plethodon cinereus (Amphibia, Urodela, Plethodontidae) in Connecticut. Journal of Herpetology, 231-236.

Machnes, Z., Avtalion, R., Shirak, A., Trombka, D., Wides, R., Fellous, M., \& Don, J. (2008). Male-specific protein (MSP): a new gene linked to sexual behavior and aggressiveness of tilapia males. Hormones and behavior, 54(3), 442-449.

MacManes, M. D. (2015). An opinionated guide to the proper care and feeding of your transcriptome. bioRxiv, 035642.

Madison, D. M. (1969). Homing behaviour of the red-cheeked salamander, Plethodon jordani. Animal Behaviour, 17, 25-39.

Madison, D. M. (1975). Intraspecific odor preferences between salamanders of the same sex: dependence on season and proximity of residence. Canadian Journal of Zoology, 53(9), 1356-1361.

Madison, D. M., \& Shoop, C. R. (1970). Homing behavior, orientation, and home range of salamanders tagged with tantalum-182. Science, 168(3938), 1484-1487.

Maerz, J. C., Panebianco, N. L., \& Madison, D. M. (2001). Effects of predator chemical cues and behavioral biorhythms on foraging, activity of terrestrial salamanders. Journal of chemical ecology, 27(7), 1333-1344.

Maex, M., Van Bocxlaer, I., Mortier, A., Proost, P., \& Bossuyt, F. (2016). Courtship Pheromone Use in a Model Urodele, the Mexican Axolotl (Ambystoma mexicanum). Scientific reports, 6.

Maksimowich, D. S., \& Mathis, A. (2001). Pheromonal markers as indicators of parasite load: parasite-mediated behavior in salamanders (Plethodon angusticlavius). acta ethologica, 3(2), 83-87.

Marco, A., Chivers, D. P., Kiesecker, J. M., \& Blaustein, A. R. (1998). Mate choice by chemical cues in western redback (Plethodon vehiculum) and Dunn's (P. dunni) salamanders. Ethology, 104(9), 781-788.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal, 17(1), pp-10.

Marvin, G. A. (1996). Life history and population characteristics of the salamander Plethodon kentucki with a review of Plethodon life histories. American Midland Naturalist, 385-400.

Marzoni, M., Castillo, A., Sagona, S., Citti, L., Rocchiccioli, S., Romboli, I., \& Felicioli, A. (2013). A proteomic approach to identify seminal plasma proteins in roosters (Gallus gallus domesticus). Animal reproduction science, 140(3), 216-223.

Mason, R. T., Fales, H. M., Jones, T. H., Pannell, L. K., Chinn, J. W., \& Crews, D. (1989). Sex pheromones in snakes. Science, 245(4915), 290-293.

Mathis, A. (1990). Territorial salamanders assess sexual and competitive information using chemical signals. Animal Behaviour, 40(5), 953-962.

Mathis, A. (1991a). Territories of male and female terrestrial salamanders: costs, benefits, and intersexual spatial associations. Oecologia, 86(3), 433-440.

Mathis, A. (1991b). Large male advantage for access to females: evidence of male-male competition and female discrimination in a territorial salamander. Behavioral Ecology and Sociobiology, 29(2), 133-138.

Mathis, A., \& Simons, R. R. (1994). Size-dependent responses of resident male redbacked salamanders to chemical stimuli from conspecifics. Herpetologica, 335-344.

Maynard Smith, J, \& Harper, D. (2003). Animal signals. Oxford University Press.
McGavin, M. (1978). Recognition of conspecific odors by the salamander Plethodon cinereus. Copeia, 1978(2), 356-358.

Mechref, Y., Ma, W., Hao, G., \& Novotny, M. V. (1999). N-linked oligosaccharides of vomeromodulin, a putative pheromone transporter in rat. Biochemical and biophysical research communications, 255(2), 451-455.

Mechref, Y., Zidek, L., Ma, W., \& Novotny, M. V. (2000). Glycosylated major urinary protein of the house mouse: characterization of its N -linked oligosaccharides. Glycobiology, 10(3), 231-235.

Meyer, D., Zeileis, A., \& Hornik, K. (2006). The strucplot framework: visualizing multiway contingency tables with vcd. Journal of Statistical Software, 17(3), 1-48.

Miller, M. A., Pfeiffer, W., \& Schwartz, T. (2010, November). Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In Gateway Computing Environments Workshop (GCE), 2010 (pp. 1-8). IEEE.

Mueller, R. L., Macey, J. R., Jaekel, M., Wake, D. B., \& Boore, J. L. (2004).
Morphological homoplasy, life history evolution, and historical biogeography of plethodontid salamanders inferred from complete mitochondrial genomes. Proceedings of the National Academy of Sciences of the United States of America, 101(38), 1382013825.

Murray, D. L., \& Jenkins, C. L. (1999). Perceived predation risk as a function of predator dietary cues in terrestrial salamanders. Animal Behaviour, 57(1), 33-39.

Nagel, J. W. (1977). Life history of the red-backed salamander, Plethodon cinereus, in northeastern Tennessee. Herpetologica, 13-18.

Ng, M. Y., \& Wilbur, H. M. (1995). The cost of brooding in Plethodon cinereus. Herpetologica, 1-8.

Olson, J. H., Xiang, X., Ziegert, T., Kittelson, A., Rawls, A., Bieber, A. L., \& Chandler, D. E. (2001). Allurin, a 21-kDa sperm chemoattractant from Xenopus egg jelly, is related to mammalian sperm-binding proteins. Proceedings of the National Academy of Sciences, 98(20), 11205-11210.

Organ, J. A. (1960). Studies on the life history of the salamander, Plethodon welleri. Copeia, 1960(4), 287-297.

Organ, J. A. (1961). Studies of the local distribution, life history, and population dynamics of the salamander genus Desmognathus in Virginia. Ecological Monographs, 31(2), 189-220.

Organ, J. A., \& Lowenthal, L. A. (1963). Comparative studies of macroscopic and microscopic features of spermatophores of some plethodontid salamanders. Copeia, 659669.

Ovaska, K., \& Davis, T. M. (1992). Faecal pellets as burrow markers: intra-and interspecific odour recognition by western plethodontid salamanders. Animal behaviour, 43(6), 931-939.

Page, R. B., \& Jaeger, R. G. (2004). Multimodal signals, imperfect information, and identification of sex in red-backed salamanders (Plethodon cinereus). Behavioral Ecology and Sociobiology, 56(2), 132-139.

Palmer, C. A., \& Houck, L. D. (2005). Responses to sex-and species-specific chemical signals in allopatric and sympatric salamander species. In Chemical Signals in Vertebrates 10 (pp. 32-41). Springer US.

Palmer, C. A., Watts, R. A., Gregg, R. G., McCall, M. A., Houck, L. D., Highton, R., \& Arnold, S. J. (2005). Lineage-specific differences in evolutionary mode in a salamander courtship pheromone. Molecular biology and evolution, 22(11), 2243-2256.

Palmer, C. A., Watts, R. A., Houck, L. D., Picard, A. L., \& Arnold, S. J. (2007a). Evolutionary replacement of components in a salamander pheromone signaling complex: more evidence for phenotypic-molecular decoupling. Evolution, 61(1), 202-215.

Palmer, C. A., Hollis, D. M., Watts, R. A., Houck, L. D., McCall, M. A., Gregg, R. G., Feldhoff, P.W., Feldhoff, R.C., \& Arnold, S. J. (2007b). Plethodontid modulating factor,
a hypervariable salamander courtship pheromone in the three-finger protein superfamily. FEBS Journal, 274(9), 2300-2310.

Palmer, C. A., Watts, R. A., Hastings, A. P., Houck, L. D., \& Arnold, S. J. (2010). Rapid evolution of plethodontid modulating factor, a hypervariable salamander courtship pheromone, is driven by positive selection. Journal of molecular evolution, 70(5), 427440.

Peacock, R. L., \& Nussbaum, R. A. (1973). Reproductive biology and population structure of the western red-backed salamander, Plethodon vehiculum (Cooper). Journal of Herpetology, 215-224.

Perez-Miller, S., Zou, Q., Novotny, M. V., \& Hurley, T. D. (2010). High resolution X-ray structures of mouse major urinary protein nasal isoform in complex with pheromones. Protein Science, 19(8), 1469-1479.

Perry, A. C., Jones, R., \& Hall, L. (1993). Sequence analysis of monkey acrosin-trypsin inhibitor transcripts and their abundant expression in the epididymis. Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression, 1172(1), 159-160.

Petersen, T. N., Brunak, S., von Heijne, G., \& Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature methods, 8(10), 785786.

Petranka, J. W. (1998). Salamanders of the United States and Canada. Smithsonian Institution Press, Washington, D.C.

Petrescu, A. J., Milac, A. L., Petrescu, S. M., Dwek, R. A., \& Wormald, M. R. (2004). Statistical analysis of the protein environment of N -glycosylation sites: implications for occupancy, structure, and folding. Glycobiology, 14(2), 103-114.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., \& Ferrin, T. E. (2004). UCSF Chimera - a visualization system for exploratory research and analysis. Journal of computational chemistry, 25(13), 1605-1612.

Picard, A. L. (2005). Courtship in the Zig-Zag Salamander (Plethodon dorsalis): Insights into a Transition in Pheromone-Delivery Behavior. Ethology, 111(9), 799-809.

Pie, J., Kim, B. H., \& Grishin, N. V. (2008). PROMALS3D: a tool for multiple sequence and structure alignment. Nucleic Acids Research, 36(7), 2295-300.

Placyk Jr, J. S., \& Graves, B. M. (2002). Prey detection by vomeronasal chemoreception in a plethodontid salamander. Journal of chemical ecology, 28(5), 1017-1036.

Pyron, R. A., \& Wiens, J. J. (2011). A large-scale phylogeny of Amphibia including over 2800 species, and a revised classification of extant frogs, salamanders, and caecilians. Molecular Phylogenetics and Evolution, 61(2), 543-583.

Quinn, V. S., \& Graves, B. M. (1999). Space use in response to conspecifics by the red-backed salamander (Plethodon cinereus, Plethodontidae, Caudata). Ethology, 105(11), 993-1002.

Reagan, N. L. (1992). Evolution of sexual isolation in salamanders of the genus Plethodon. PhD dissertation, Department of Ecology and Evolution, University of Chicago.

Revell, L. J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution, 3(2), 217-223.

Robinson, M. D., McCarthy, D. J., \& Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26(1), 139-140.

Rollmann, S. M., Houck, L. D., \& Feldhoff, R. C. (1999). Proteinaceous pheromone affecting female receptivity in a terrestrial salamander. Science, 285(5435), 1907-1909.

Rollmann, S. M., Houck, L. D., \& Feldhoff, R. C. (2000). Population variation in salamander courtship pheromones. Journal of Chemical Ecology, 26(12), 2713-2724.

Rollmann, S. M., Houck, L. D., \& Feldhoff, R. C. (2003). Conspecific and heterospecific pheromone effects on female receptivity. Animal Behaviour, 66(5), 857-861.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., \& Huelsenbeck, J. P. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic biology, 61(3), 539-542.

Roudebush, R. E., \& Taylor, D. H. (1987). Chemical communication between two species of desmognathine salamanders. Copeia, 744-748.

Roy, A., Kucukural, A., \& Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. Nature protocols, 5(4), 725-738.

Ruben, J. A., \& Boucot, A. J. (1989). The origin of the lungless salamanders (Amphibia: Plethodontidae). American Naturalist, 161-169.

Sayler, A. (1966). The reproductive ecology of the red-backed salamander, Plethodon cinereus, in Maryland. Copeia, 183-193.

Schmidt, A., \& Roth, G. (1990). Central olfactory and vomeronasal pathways in salamanders. Journal fur Hirnforschung, 31(5), 543-553.

Seko, A., Kitajima, K., Iwasaki, M., Inoue, S., \& Inoue, Y. (1989). Structural studies of fertilization-associated carbohydrate-rich glycoproteins (hyosophorin) isolated from the fertilized and unfertilized eggs of flounder, Paralichthys olivaceus. Presence of a novel penta-antennary N -linked glycan chain in the tandem repeating glycopeptide unit of hyosophorin. Journal of Biological Chemistry, 264(27), 15922-15929.

Selby, M. F., Winkel, S. C., \& Petranka, J. W. (1996). Geographic uniformity in agonistic behaviors of Jordon's salamander. Herpetologica, 108-115.

Sever, D. M. (1976). Morphology of the mental hedonic gland clusters of plethodontid salamanders (Amphibia, Urodela, Plethodontidae). Journal of Herpetology, 227-239.

Sever, D. M. (1978a). Male cloacal glands of Plethodon cinereus and Plethodon dorsalis (Amphibia: Plethodontidae). Herpetologica, 1-20.

Sever, D. M. (1978b). Female cloacal anatomy of Plethodon cinereus and Plethodon dorsalis (Amphibia, Urodela, Plethodontidae). Journal of Herpetology, 397-406.

Sever, D. M. (1989). Caudal hedonic glands in salamanders of the Eurycea bislineata complex (Amphibia: Plethodontidae). Herpetologica, 322-329.

Sever, D. M. (2003). Courtship and mating glands. Reproductive Biology and Phylogeny of Urodela (Amphibia). DM Sever (ed.). Science Publishers, Inc., Enfield, New Hampshire, 323-381.

Sever, D. M. (2014). Ultrastructure of the Male Cloacal Glands of the Eastern Redbacked Salamander, Plethodon cinereus. Herpetologica, 70(1), 1-13.

Sever, D. M., \& Siegel, D. S. (2015). Histology and ultrastructure of the caudal courtship glands of the red-backed salamander, Plethodon cinereus (Amphibia: Plethodontidae). Journal of morphology, 276(3), 319-330.

Sharon, N., \& Lis, H. (2004). History of lectins: from hemagglutinins to biological recognition molecules. Glycobiology, 14(11), 53R-62R.

Shen, X. X., Liang, D., Chen, M. Y., Mao, R. L., Wake, D. B., \& Zhang, P. (2015). Enlarged Multilocus Data set Provides Surprisingly Younger Time of Origin for the Plethodontidae, the Largest Family of Salamanders. Systematic biology, syv061.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., \& Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with singlecopy orthologs. Bioinformatics, 31(19), 3210-3212.

Simon, G. S., \& Madison, D. M. (1984). Individual recognition in salamanders: cloacal odours. Animal Behaviour, 32(4), 1017-1020.

Simons, R. R., \& Felgenhauer, B. E. (1992). Identifying areas of chemical signal production in the red-backed salamander, Plethodon cinereus. Copeia, 776-781.

Simons, R. R., Felgenhauer, B. E., \& Jaeger, R. G. (1994). Salamander scent marks: site of production and their role in territorial defence. Animal Behaviour, 48(1), 97-103.

Simons, R. R., Jaeger, R. G., \& Felgenhauer, B. E. (1997). Competitor assessment and area defense by territorial salamanders. Copeia, 70-76.

Simons, R. R., Felgenhauer, B. E., \& Thompson, T. (1999). Description of the postcloacal glands of Plethodon cinereus, the red-backed salamander, during bouts of scent marking. Journal of morphology, 242(3), 257-269.

Skiniotis, G., Lupardus, P. J., Martick, M., Walz, T., \& Garcia, K. C. (2008). Structural organization of a full-length gp130/LIF-R cytokine receptor transmembrane complex. Molecular cell, 31(5), 737-748.

Smadja, C., \& Butlin, R. K. (2009). On the scent of speciation: the chemosensory system and its role in premating isolation. Heredity, 102(1), 77-97.

Smith, P., Krohn, R. I., Hermanson, G. T., Mallia, A. K., Gartner, F. H., Provenzano, M., ... \& Klenk, D. C. (1985). Measurement of protein using bicinchoninic acid. Analytical biochemistry, 150(1), 76-85.

Smith-Unna, R. D., Boursnell, C., Patro, R., Hibberd, J. M., \& Kelly, S. (2015). TransRate: reference free quality assessment of de-novo transcriptome assemblies. BioRxiv, 021626.

Snell, T. W. (1989). Systematics, reproductive isolation and species boundaries in monogonont rotifers. Hydrobiologia, 186(1), 299-310.

Staub, N. L., \& Paladin, J. (1997). The presence of modified granular glands in male and female Aneides lugubris (Amphibia: Plethodontidae). Herpetologica, 339-344.

Steentoft, C., Vakhrushev, S. Y., Joshi, H. J., Kong, Y., Vester-Christensen, M. B., Scholdager, K. T. B. G., Lavrsen, K., Dabelsteen, S., Pedersen, N.B., Marcos-Silva, L., Gupta, R., Bennett, E.P., Mandel, U., Brunak, S., Wandall, H.H., Levery, S.B., \& Clausen, H. (2013). Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. The EMBO journal, 32(10), 1478-1488.

Sullivan, A. M., Madison, D. M., \& Rohr, J. R. (2003). Behavioural responses by redbacked salamanders to conspecific and heterospecific cues. Behaviour, 140(4), 553-564.

Sullivan, A. M., Picard, A. L., \& Madison, D. M. (2005). To avoid or not to avoid? Factors influencing the discrimination of predator diet cues by a terrestrial salamander. Animal behaviour, 69(6), 1425-1433.

Swanson, W. J., \& Vacquier, V. D. (2002). The rapid evolution of reproductive proteins. Nature Reviews Genetics, 3(2), 137-144.

Taub, F. B. (1961). The distribution of the red-backed salamander, Plethodon cinereus, within the soil. Ecology, 42(4), 681-698.

Thomas, J. S., Jaeger, R. G., \& Horne, E. A. (1989). Are all females welcome? Agonistic behavior of male red-backed salamanders. Copeia, 915-920.

Thurow, G. (1976). Aggression and competition in eastern Plethodon (Amphibia, Urodela, Plethodontidae). Journal of Herpetology, 277-291.

Tilley, S. G. (1968). Size-fecundity relationships and their evolutionary implications in five desmognathine salamanders. Evolution, 806-816.

Tristram, D. A. (1977). Intraspecific olfactory communication in the terrestrial salamander Plethodon cinereus. Copeia, 1977(3), 597-600.

Uzendoski, K., \& Verrell, P. (1993). Sexual incompatibility and mate-recognition systems: a study of two species of sympatric salamanders (Plethodontidae). Animal behaviour, 46(2), 267-278.

Vaccaro, E. A., Feldhoff, P. W., Feldhoff, R. C., \& Houck, L. D. (2009). Male courtship pheromones suppress female tendency to feed but not to flee in a plethodontid salamander. Animal Behaviour, 78(6), 1421-1425.

Vaccaro, E. A., Feldhoff, P. W., Feldhoff, R. C., \& Houck, L. D. (2010). A pheromone mechanism for swaying female mate choice: enhanced affinity for a sexual stimulus in a woodland salamander. Animal Behaviour, 80(6), 983-989.

Van Bocxlaer, I., Treer, D., Maex, M., Vandebergh, W., Janssenswillen, S., Stegen, G., Kok, P., Willaert, B., Matthijs, S., Martens, E., Mortier, A., de Greve, H., Proost, P., \& Bossuyt, F. (2015). Side-by-side secretion of Late Palaeozoic diverged courtship pheromones in an aquatic salamander. Proceedings of the Royal Society of London B: Biological Sciences, 282(1803), 20142960.

Verrell, P. A. (1989). An experimental study of the behavioral basis of sexual isolation between two sympatric plethodontid salamanders, Desmognathus imitator and D. ochrophaeus. Ethology, 80(1-4), 274-282.

Verrell, P. A. (1995). Males choose larger females as mates in the salamander Desmognathus santeetlah. Ethology, 99(1-2), 162-171.

Verrell, P. (2003). Population and species divergence of chemical cues that influence male recognition of females in desmognathine salamanders. Ethology, 109(7), 577-586.

Vieites, D. R., Min, M. S., \& Wake, D. B. (2007). Rapid diversification and dispersal during periods of global warming by plethodontid salamanders. Proceedings of the National Academy of Sciences, 104(50), 19903-19907.
von Byern, J., Dicke, U., Heiss, E., Grunwald, I., Gorb, S., Staedler, Y., \& Cyran, N. (2015). Morphological characterization of the glandular system in the salamander Plethodon shermani (Caudata, Plethodontidae). Zoology, 118(5), 334-347.

Wake, D. B. (1966). Comparative osteology and evolution of the lungless salamanders, family Plethodontidae (Vol. 4). Southern California Academy of Sciences.

Walls, S. C., Mathis, A., Jaeger, R. G., \& Gergits, W. F. (1989). Male salamanders with high-quality diets have faeces attractive to females. Animal Behaviour, 38(3), 546-548.

Wanner, K. W., Nichols, A. S., Allen, J. E., Bunger, P. L., Garczynski, S. F., Linn Jr, C. E., Robertson, H.M., \& Luetje, C. W. (2010). Sex pheromone receptor specificity in the European corn borer moth, Ostrinia nubilalis. PLoS One, 5(1), e8685.

Watts, R. A., Palmer, C. A., Feldhoff, R. C., Feldhoff, P. W., Houck, L. D., Jones, A. G., Pfrender, M.E., Rollmann, S.M., \& Arnold, S. J. (2004). Stabilizing selection on behavior and morphology masks positive selection on the signal in a salamander pheromone signaling complex. Molecular biology and evolution, 21(6), 1032-1041.

Weiber, H., Andersson, C., Murne, A., Rannevik, G., Lindström, C., Lilja, H., \& Fernlund, P. (1990). Beta microseminoprotein is not a prostate-specific protein. Its identification in mucous glands and secretions. The American journal of pathology, 137(3), 593.

Wiens, J. J., Engstrom, T. N., \& Chippindale, P. T. (2006). Rapid diversification, incomplete isolation, and the "speciation clock" in North American salamanders (genus Plethodon): testing the hybrid swarm hypothesis of rapid radiation. Evolution, 60(12), 2585-2603.

Wilburn, D. B. (2014). Mechanisms of Action and Co-optive Evolution for Hypervariable Courtship Phermones in Plethodontid Salamanders. PhD dissertation, Department of Biochemistry and Molecular Biology, University of Louisville.

Wilburn, D. B., Bowen, K. E., Gregg, R. G., Cai, J., Feldhoff, P. W., Houck, L. D., \& Feldhoff, R. C. (2012). Proteomic and UTR analyses of a rapidly evolving hypervariable family of vertebrate pheromones. Evolution, 66(7), 2227-2239.

Wilburn, D. B., Bowen, K. E., Feldhoff, P. W., \& Feldhoff, R. C. (2014a). Proteomic analyses of courtship pheromones in the redback salamander, Plethodon cinereus. Journal of chemical ecology, 40(8), 928-939.

Wilburn, D. B., Bowen, K. E., Doty, K. A., Arumugam, S., Lane, A. N., Feldhoff, P. W., \& Feldhoff, R. C. (2014b). Structural insights into the evolution of a sexy protein: novel topology and restricted backbone flexibility in a hypervariable pheromone from the redlegged salamander, Plethodon shermani. PloS one, 9(5), e96975.

Wilburn, D. B., Eddy, S. L., Chouinard, A. J., Arnold, S. J., Feldhoff, R. C., \& Houck, L. D. (2015). Pheromone isoform composition differentially affects female behaviour in the red-legged salamander, Plethodon shermani. Animal Behaviour, 100, 1-7.

Wilson, E. O. (1970). Chemical communication within animal species. Chemical ecology, 9, 133-155.

Wilson, E. O. (1999). The diversity of life. WW Norton \& Company, New York.
Wirsig-Wiechmann, C. R., Houck, L. D., Feldhoff, P. W., \& Feldhoff, R. C. (2002). Pheromonal activation of vomeronasal neurons in plethodontid salamanders. Brain research, 952(2), 335-344.

Wirsig-Wiechmann, C. R., Houck, L. D., Wood, J. M., Feldhoff, P. W., \& Feldhoff, R. C. (2006). Male pheromone protein components activate female vomeronasal neurons in the salamander Plethodon shermani. BMC neuroscience, 7(1), 26.

Wise, S. E., Verret, F. D., \& Jaeger, R. G. (2004). Tail autotomy in territorial salamanders influences scent marking by residents and behavioral responses of intruders to resident chemical cues. Copeia, 2004(1), 165-172.

Woodley, S. K. (2010). Pheromonal communication in amphibians. Journal of Comparative Physiology A, 196(10), 713-727.

Woodley, S. K. (2014). Chemical Signaling in Amphibians. In: Mucignat-Caretta, C. (Ed.), Neurobiology of chemical communication, 255. CRC Press, Boca Raton, Florida.

Woodley, S. (2015). Chemosignals, hormones, and amphibian reproduction. Hormones and behavior, 68, 3-13.

Wyatt, T. D. (2003). Pheromones and animal behaviour: communication by smell and taste. Cambridge University Press, Cambridge, England.

Wyatt, T. D. (2010). Pheromones and signature mixtures: defining species-wide signals and variable cues for identity in both invertebrates and vertebrates. Journal of Comparative Physiology A, 196(10), 685-700.

Wyatt, T. D. (2014a). Pheromones and animal behavior: chemical signals and signatures. Cambridge University Press, Cambridge, England.

Wyatt, T. D. (2014b). Proteins and peptides as pheromone signals and chemical signatures. Animal Behaviour, 97, 273-280.

Yamamoto, K., Toyoda, F., Tanaka, S., Hayashi, H., \& Kikuyama, S. (1996). Radioimmunoassay of a newt sex pheromone, sodefrin, and the influence of hormones on its level in the abdominal gland. General and comparative endocrinology, 104(3), 356363.

Yamamoto, K., Kawai, Y., Hayashi, T., Ohe, Y., Hayashi, H., Toyoda, F., Kawahara, G., Iwata, T., \& Kikuyama, S. (2000). Silefrin, a sodefrin-like pheromone in the abdominal gland of the sword-tailed newt, Cynops ensicauda. FEBS letters, 472(2), 267-270.

Yamazaki, Y., \& Morita, T. (2004). Structure and function of snake venom cysteine-rich secretory proteins. Toxicon, 44(3), 227-231.

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., \& Zhang, Y. (2015). The I-TASSER Suite: protein structure and function prediction. Nature methods, 12(1), 7-8.

Yang, J., \& Zhang, Y. (2015). Protein Structure and Function Prediction Using I-TASSER. Current Protocols in Bioinformatics, 5-8.

Young, M. D., Wakefield, M. J., Smyth, G. K., \& Oshlack, A. (2010). Method Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol, 11, R14.

Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. BMC bioinformatics, 9(1), 40.

Zhang, Y., \& Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic acids research, 33(7), 2302-2309.

Zhang, P., \& Wake, D. B. (2009). Higher-level salamander relationships and divergence dates inferred from complete mitochondrial genomes. Molecular Phylogenetics and Evolution, 53(2), 492-508.

Zufall, F., Kelliher, K. R., \& Leinders-Zufall, T. (2002). Pheromone detection by mammalian vomeronasal neurons. Microscopy research and technique, 58(3), 251-260.

## APPENDICES

Contents Page
Appendix One: Individual Male Profiles (Chapter Two). ..... 161
Appendix Two: Bioinformatic Analyses (Chapter Three) ..... 175
Appendix Three: Bioinformatic Analyses (Chapter Four).... ..... 251

## APPENDIX ONE

Contents Page
Figure A1.1: Individual Male Mental Gland Profiles....................................... 162














## APPENDIX TWO

Contents Page
Custom Bioinformatic Scripts (Chapter Three) ..... 176
PRF-Glycosylation.pl ..... 176
README: Previous PRF Sequences ..... 182
Find-Duplicates.pl ..... 183
ParseFasta.pl ..... 185
Rename-Previous-PRFs.pl ..... 186
Translate.pl ..... 187
Untranslate-Alignment.pl ..... 188
Re-Gap.pl ..... 190
Parse-Nglyc.pl ..... 192
Rename-Long-to-Short.pl ..... 194
Rename-Short-to-Long.pl ..... 196
Parse-Oglyc.pl ..... 198
SubsetSeqs.pl ..... 200
Parse-Oglyc-version4.pl ..... 201
Aligned Protein Sequences ..... 204
Aligned Coding Sequences (mRNA) ..... 217

## PRF-Glycosylation.pl

\#!/usr/bin/perl

```
use strict;
use warnings;
```

\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\# \#\#\#
\#\#\# This script will run an automated pipeline to analyze glycosylation potential of all PRF sequences \#\#\# \#\#\#
\#\#\# The input file was prepped manually, as it required close scrutiny on behalf of the researcher \#\#\# \#\#\# To do so:
\#\#\#
\#\#\#
\#\#\#
\#\#\#
\#\#\#
\#\#\#
\#\#\#
\#\#\#
\#\#\#
\#\#\#
\#\#\#

- All previous PRF coding sequences (CDS) were downloaded from Genbank
- Duplicate sequences within any species were removed (i.e. intraspecifically non-redunant)
- Use "FindDuplicates.pl" to identify any redundancies for each species
- Any unwanted sequences (e.g. hybrids) were also removed
- Population variation was condensed for each species (while eliminating duplicates, as above)
- Sequences were renamed to include:

Species_gi|NUMBER|_PRF_ISOFORM (e.g. ">P.serratus_gi|60499226|_PRF_A1"

- Use "Rename-Previous-PRFs.pl" to change the names, and adjust as needed
- Previous sequences were combined with P.cinereus sequences from New Hampshire (NH)
- These differ from the P.cinereus files in Genbank, obtained from Virginia (VA)
- All coding sequences (CDS) were then translated into amino acids ("Translate.pl")
- Amino acid (AA) sequences were then aligned (via PROMALS3D web server) and manually adjusted
- Aligned AA sequences were converted to aligned CDS ("Untranslate-Alignment.pl") and adjusted
- These can be used for subsequent analyses (e.g. phylogenetics) using nucleotides
- Sequences were then ready for glycosylation (and other) analyses
lycosylation analyses (a.k.a. this script) will do the following:
- Predict (1) signal peptides, (2) N-linked glycosylation, and (3) O-linked glycosylation
- Parse the raw (and annoying) output of each of those programs and save it in a custom format
$\# \# \#$
$\# \# \#$ - Parse the raw (and annoying) output of each of those programs and save it in a custom format
$\# \# \#$
$\# \# \#$ - Parse the raw (and annoying) output of each of those programs and save it in a custom format
\#\#\# - Parse the manually-generated 0-linked predictions (from NetOglyc 4.0) - only if it exists
Note on 0-linked predictions: at present NetOglyc 4.0 does not have a standalone program,
but must be run on a server. The script itself runs the current standalone version (3.1d)
and then parses/processes the 4.0 results for plotting if it exists. This must be made
by manually submitting 50 -sequence subsets (use "SubsetSeqs.pl") to the server. The
results then need to be copied and concatenated manually. As in the standalone version, $\begin{array}{ll}\# & \text { results then need to be copied and concatenated manually. As in the standalone version } \\ \# & \text { sequence names need to be shortened before submitting to the server, but they will be } \\ \# & \text { re-converted to the full names as a part of the processing in this script. Likewise, the }\end{array}$
re-converted to the full names as a part of the processing in this script. Likewise, they
at each position (with "NA" in the place of residues other than $S$ or $T$.
Arguments :
\#\#\#
\#\#\#
\#\#\#
\#\#\#
\#\#\#
\#\#\#
\#\#\#
\#\#\#
\#\#\#
\#\#\#
\#\#\#
\#\#\#
\#\#\#
\#\#\#
\#\#\#
\#\#\#
\#\#\#
\#\#\#
\#\#\#
\#\#\#
\#\#\#
\#\#\#
\#\#\#
\#\#\#
\#\#\#
\#\#\#
\#\#\#
\#\# \#\#\#
\#\#\#
\#\#\#
\#\#\#
\#\#\#
\#\#\#
\#\#\#
\#\#\#
\#\#\#


## \#\#\#

Additional files to prepare:
\#\#\# - netOglyc 4.0 results (input as "\$Oglyc_v4_raw")
\#\#\#
\#\#\# - jnet secondary structure and solvent accessibility predictions (input as "\$jnet" variable)
\#\#\# Dependencies:
\#\#\# - Signalp version 4.1
\#\#\# - NetNglyc version 1.0
\#\#\# - NetOglyc version 3.1d
Re-Gap.pl (custom script)
\#\#\#
\#\#\#
\#\#\#
\#\#\# - Parse-SignalP.pl (custom script) \#\#\#

\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#
\#\#\# USER-CUSTOMIZABLE OPTIONS and LOCAL INSTALLATIONS OF ALL DEPENDENCIES \#\#\#
\# Set path to dependencies \#

| my \$signalp | = "/nfs0/IB/Weis_Lab/chouinad/Tools/signalp-4.1/signalp"; |
| :---: | :---: |
| my \$netNglyc | = "/nfs0/IB/Weis_Lab/chouinad/Tools/netNglyc-1.0/netNglyc"; |
| my \$netOglyc | = "/nfs0/IB/Weis_Lab/chouinad/Tools/netOglyc-3.1d/net0glyc"; |
| my \$parseNglyc | = "/nfs0/IB/Weis_Lab/chouinad/Tools/Parse-Nglyc.pl"; |
| my \$parseOglyc | = "/nfs0/IB/Weis_Lab/chouinad/Tools/Parse-Oglyc.pl"; |
| my \$parse0_v4 | = "/nfs0/IB/Weis_Lab/chouinad/Tools/Parse-Oglyc-version4.pl"; |
| my \$parseSP | = "/nfs0/IB/Weis_Lab/chouinad/Tools/Parse-SignalP.pl"; |
| my \$rename | = "/nfs0/IB/Weis_Lab/chouinad/Tools/Rename-Long-to-Short.pl"; |
| my \$regap | = "/nfs0/IB/Weis_Lab/chouinad/Tools/Re-Gap.pl"; |



```
# Parse results #
my $Oglyc_tmp = "./Oglyc.tmp";
my $Oglyc_results = "./Oglyc.results";
system("$parseOglyc $Oglyc_raw $Oglyc_tmp");
# Reference the hash created in the beginning to convert output back to full names #
open(TMP, "<$Oglyc_tmp") or die("Cannot open 0-linked temp file because: $!\n");
open(OGLYC, ">$Oglyc_results");
while(<TMP>){
        if($_ =~ /^>(\d{8})/){
            my $shorty = $1;
            my $newout = $_;
            $newout =~ s/$shorty/$HeaderKey{$shorty}/;
                    print OGLYC $newout;
        elsif($_ = / /^>(NH_PRF_\w+)/){
            my $shorty = $1;
            my $newout = $_;
                    $newout =~ s/$shorty/$HeaderKey{$shorty}/;
                    print OGLYC $newout;
                else{die("Error processing name changes for the sequence: $_\n")}
        },
close(TMP);
close(OGLYC);
system("rm $Oglyc_tmp");
%HeaderKey=();
```

\#\#\# Read in the unaligned file, store headers/sequences in a hash and the order of headers in an array \#\#\#
open(UNALIGNED, "<\$unaligned") or die("Cannot open unaligned input file because: \$! \n");
my \$ID = "";
my \$SEQ = "";
my @SeqOrder;
my \%SEQUENCES;
while (<UNALIGNED>) \{
\$ =~ s/\s+\$//;
unless(\$ID eq "") \{
\$SEQUENCES $\{\$ I D\}=\$ S E Q ;$
\}
\$ID = \$ ; ;
push(@SeqOrder, \$ID);
else\{
\$SEQ . = \$_;
\}
\$SEQUENCES\{\$ID\} = \$SEQ;
close(UNALIGNED);
\#\#\# Open each of the results files (sequentially) and annotate the raw sequences \#\#\#

| \#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\# |  |
| :--- | :--- |
| \#\#\# |  |
| \#\#\# The annotation code used for plotting is as follows: |  |
| \#\#\# |  |
| \#\#\# | $0=$ Gap |
| \#\#\# | $1=$ Any other residue (i.e. non-annotated) |
| \#\#\# | $2=N$-Linked Sequon (predicted -) |
| \#\#\# | $3=N$-Linked Sequon (predicted +) |
| \#\#\# | $4=0$-Linked Site (predicted -) |
| \#\#\# | $5=0$-Linked Site (predicted +) |
| \#\#\# | $6=$ Signal Peptide |
| \#\#\# | $7=$ Stop Codon |
| \#\#\# |  |
| \#\#\#\#\#\#\#\#\#\# |  |
| \#\#\#\#\#\#\#\# |  |

system("date");
print "Annotating sequences...\n";
\# Annotate 0-linked glycosylation results - must be first to avoid over-writing other annotations \#
my \$Oglyc_results2 = \$Oglyc_results;
open(OGLYC, "<\$Oglyc_results2");
my @0_ARRAY;
my \%O_HASH;
while(<0GLYC>) \{
chomp;
@O_ARRAY = split(/八t/, \$_);
my \$name = shift(@0_ARRAY);
if(@0_ARRAY == 0)\{
\$0_ARRAY[0] = "NONE";

```
        }
    $O_HASH{$name} = [ @O_ARRAY ];
    }
foreach(@SeqOrder){
    my $name = $_;
    if($O_HASH{$name}[0] eq "NONE"){
            next;
            }
    foreach(@{$0_HASH{$name}}){
            my $Osite = "';
            my $prediction = ""
            my $annotation = "";
            if($_ =~ /(\d+)[ST]([\+-])/){
                    $Osite = ($1 - 1)
                                    $prediction = $2;
                                }
                                    else{die("Cannot read the O-linked prediction results for: $name\n")}
                if($prediction eq "-"){
                    $annotation = "4";
                    }
            elsif($prediction eq "+"){
                                    $annotation = "5";
                                    }
                                    else{die("Cannot interpret O-linked prediction code for: $name\n")}
                substr($SEQUENCES{$name}, $Osite, 1, $annotation)
                }
    }
close(OGLYC)
# Annotate N-linked glycosylation results #
my $Nglyc_results2 = $Nglyc_results;
open(NGLYC, "<$Nglyc_results2")
my @N_ARRAY;
my %N_HASH;
while(<NGLYC>){
    chomp;
    @N_ARRAY = split(/\t/, $_);
    my $name = shift(@N_ARRAY);
    if(@N_ARRAY == 0){
            $N_ARRAY[0] = "NONE";
            }
    $N_HASH{$name} = [ @N_ARRAY ];
    }
    my $name = $_;
    if($N_HASH{$name}[0] eq "NONE"){
                next;
    foreach(@{$N_HASH{$name}}){
            my $Nsite = "";
                my $prediction = "";
                my $annotation = "";
                if($_ =~ /(\d+)([\+-])/){
                    $Nsite = ($1 - 1);
                        $prediction = $2;
                        }pr
                            else{die("Cannot read the N-linked prediction results for: $name\n")}
            if($prediction eq "-"){
                    $annotation = "2" x 3;
                    }
            elsif($prediction eq "+"){
                    $annotation = "3" x 3;
                            }
                            else{die("Cannot interpret N-linked prediction code for: $name\n")}
                substr($SEQUENCES{$name}, $Nsite, 3, $annotation)
                }
    }
undef(%N_HASH);
undef(@N_ARRAY);
close(NGLYC);
# Annotate signal peptide results #
my $SP_results2 = $SP_results;
open(S\overline{P}, "<$SP_results2") or die("Cannot open signal peptide results file because: $!\n");
my @SP_ARRAY;
my %SP_HASH;
while(<SP>){
    chomp;
    @SP_ARRAY = split(/\t/, $_);
    $SP_HASH{$SP_ARRAY[0]} = $SP_ARRAY[1];
    }
foreach(@SeqOrder){
    my $name = $_;
    my $length = ($SP_HASH{$name} - 1);
```

```
    my $replacement = "6" x $length;
    substr($SEQUENCES{$name}, 0, $length, $replacement);
    }
undef(%SP_HASH)
undef(@SP_ARRAY);
close(SP);
# Annotate all other (non-annotated) residues and the stop codon #
foreach(@SeqOrder){
my $name = $_;
$SEQUENCES{$name} =~ s/[a-zA-Z]/1/g;
}
```

\#\#\# Print annotations to file and re-gap them according to the aligned file \#

```
system("date");
print "Re-gapping the unaligned annotated file...\n";
    # Print annotated sequences to file #
    my $unaligned_ann_tmp = "./Unaligned_Annotated.tmp";
    open(UNALIGNED_ANN_TMP, ">$unaligned_ann_tmp")
    foreach(@SeqOrder){
            print UNALIGNED_ANN_TMP $_ . "\n";
            print UNALIGNED_ANN_TMP $SEQUENCES{$_} . "\n";
            }
    close(UNALIGNED_ANN_TMP);
    # Print a second one for posterity with stop codons - these cause problems when re-gapping #
    my $unaligned_ann = "./Unaligned_Annotated.fasta";
    open(UNALIGNED_ANN, ">$unaligned_ann");
    foreach(@SeqOrder){
            print UNALIGNED_ANN $_ . "\n";
            my $seq = $SEQUENCES{$_};
            $seq =~ s/$/7/;
            print UNALIGNED_ANN $seq . "\n";
            }
    close(UNALIGNED_ANN);
    # Re-gap the unaligned annotations #
    my $aligned_ann_tmp = "./Aligned_Annotated.tmp";
    system("$regap $unaligned_ann_tmp $aligned $aligned_ann_tmp");
    # Add stop codons - these cause problems with re-gapping - and annotate gaps #
    my $aligned_ann = "./Aligned_Annotated.fasta";
    open(ALIGNED_ANN_TMP, "<$aligned_ann_tmp");
    open(ALIGNED_ANN, ">$aligned_ann");
    while(<ALIGNED_ANN_TMP>) {
            chomp;
            my $line = $_;
            if($line =~ /^^>/){
                    print ALIGNED_ANN $line . "\n";
                    }
            else{
                    if($line =~ /-+$/){
                                    $line =~ s/(-+)$/7$1/;
                                    $line =~ s/-/0/g;
                                    print ALIGNED_ANN $line . "\n";
                    else{
                                    $line =~ s/$/7/;
                                    $line =~ s/-/0/g;
                                    print ALIGNED_ANN $line . "\n";
                                    }r
                            }
            }
    close(ALIGNED_ANN_TMP);
    close(ALIGNED_ANN);
    system("rm $unaligned_ann_tmp $aligned_ann_tmp");
\#\#\# Set up annotations to be input to the plot \#\#\#
open(ALIGNED_ANN, "<\$aligned_ann");
my \$annotations = "./Annotations.csv";
open(ANNOTATIONS, ">\$annotations");
my \$header = "";
while(<ALIGNED_ÁNN>) \{
```


## chomp;

```
my \$line = \$_;
if(\$line \(=\sim\) /^>/) \(\{\)
\$header = \$line;
\$header \(=\sim\) s/^>//;
\}
else\{
print ANNOTATIONS \$header . ",";
my @SEQ = split(//,\$line);
```

```
        foreach(@SEQ){
            print ANNOTATIONS $_ . "\,";
            }
        print ANNOTATIONS "\n";
        }
    }
close(ALIGNED_ANN)
close(ANNOTATIONS);
### Parse manually-generated O-linked prediction (from NetOglyc 4.0) if it exists ###
my $Oglyc_v4_raw = "./Oglyc-v4.raw";
my $Oglyc_v4_unaligned = "./Oglyc-v4-unaligned.results";
my $Oglyc_v4_aligned = "./Oglyc-v4-aligned.results";
if(-e $Oglyc_v4_raw){
    system("date");
    print "Parsing NetOglyc 4.0 results...\n";
    # Parse results #
    system("$parseO_v4 $Oglyc_v4_raw $unaligned $Oglyc_v4_unaligned");
    # Read in aligned sequences to serve as a reference for re-gapping #
    open(ALIGNED_SEQS, "<$aligned") or die("Can't open aligned reference file because: $!\n");
    my $ID = "";
    my $SEQ = "";
    my %Gaps = (),
    while (<ALIGNED_SEQS>) {;
    $ =~ s/\s+$//;'
                unless($ID eq ""){
                $ID = $ ;
                push(@SeqOrder, $ID);
                $SEQ = "";
            else{
                        $SEQ .= $_;
                        }
            }
    $Gaps{$ID} = $SEQ;
    close(ALIGNED_SEQS);
    # Read in the unaligned results to re-gap #
    open(UNALIGNED_V4, "<$Oglyc_v4_unaligned") or die(
                            "Can't open unaligned NetOglyc 4.0 results because: $!\n");
    my %NoGaps = ();
    while(<UNALIGNED_V4>){
    chomp;
    my @OSITES = split(/\t/, $_);
            my $seq_name = shift(@OSITES);
            $NoGaps{$seq_name} = [ @OSITES ];
            }
    close(UNALIGNED_V4);
        # Re-gap the unaligned results according to the reference #
        open(ALIGNED_V4, ">$Oglyc_v4_aligned");
        foreach(@SeqOrder){
            my @Aligned_Seq = split(//, $Gaps{$_});
            my @Unaligned_Seq = @{$NoGaps{$_}};
            print ALIGNED_V4 $_;
            my $i = "";
            my $x = 0;
            foreach $i (@Aligned_Seq){
                if($i eq '-'){
                        print ALIGNED_V4 "\tNA";
                else{
                        print ALIGNED_V4 "\t" . $Unaligned_Seq[$x];
                                    $x++;
                                    }
                }
            print ALIGNED_V4 "\tNA\n";
            # This makes length compatible with alignment including stop codon #
            }
        close(ALIGNED_V4);
        }
else{print "NetOglyc 4.0 predictions were not made, so they will not be processed.\n";}
```

\#\#\# All Done \#\#\#
system("date");
print "Done. \n";

## README: Previous PRF Sequences

A quick note about the pruning of previous PRF sequences for glycosylation and evolutionary analysis:

- All previous PRF sequences from Plethodon were downloaded from Genbank ( $\mathrm{N}=224$ ). - Resulting file: "All-Previous-PRFs-CDS-Raw.fasta"
- Intraspecific duplicates were identified ( $\mathrm{N}=10$ ).
- Total duplicate matrix: "Previous-PRFs-Duplicates.tab"
- Resulting file: "All-Previous-PRFs-CDS-NR.fasta"

Intraspecific duplicates:
gi|42566655|gb|AY499372.1| Plethodon jordani receptivity factor PRFA1 mRNA, complete cds
gi|60499462|gb|AY927006.1| Plethodon shermani clone P.sherDGPRFA13 receptivity factor PRFA13 (PRF) mRNA, complete cds
gi|60499468|gb|AY927009.1| Plethodon shermani clone P.sherDGPRFA16 receptivity factor PRFA16 (PRF) mRNA,
complete cds
gi|60499338|gb|AY926944.1| Plethodon teyahalee clone P.teyaMPPRFA2 receptivity factor PRFA2 (PRF) mRNA, complete cds
gi|60499342|gb|AY926946.1| Plethodon teyahalee clone P.teyaMPPRFA4 receptivity factor PRFA4 (PRF) mRNA, complete cds
gi|60499286|gb|AY926918.1| Plethodon metcalfi clone P.metcCWPRFA11 receptivity factor PRFA11 (PRF) mRNA, complete cds

- Duplicates also included the original P. jordani sequences, which is a relic of the species name change - These sequences are from the Rollmann et al. (1999) paper, which actually sampled what would become "P. shermani"
- They are identical to the corresponding $P$. shermani isoforms (sequenced under the new species name)
P. jordani/shermani duplicates:
gi|6007745|gb|AF181480.1| Plethodon jordani receptivity factor isoform 1 precursor (PRF) mRNA, complete cds gi|6007747|gb|AF181481.1| Plethodon jordani receptivity factor isoform 2 precursor (PRF) mRNA, complete cds gi|6007749|gb|AF181482.1| Plethodon jordani receptivity factor isoform 3 precursor (PRF) mRNA, complete cds gi|6007751|gb|AF181483.1| Plethodon jordani receptivity factor isoform 4 precursor (PRF) mRNA, complete cds
- Hybrids were thrown out ( $\mathrm{N}=16$ )
- Hybrids identified were "Plethodon shermani x Plethodon teyahalee"
- Resulting file: "All-Previous-PRFs-CDS-NR-No-Hybrids.fasta"
- The final file ("Previous-PRFs-CDS.fasta") therefore contains only unique sequences per species ( $N=198$ )
- No hybrids are included but any population sampling variation yielding unique sequences is preserved
- In this case, the population variation is simply housed under the same species name
- The exception to this is "P. cinereus" as the goal is to examine these two populations
- The sequence headers in this file were renamed to a more succinct nomenclature

Example: ">Species_gi|\#\#\#\#\#\#\#\#|_isoform"

- Again the only exception is "P. cinereus" in which the header contains "VA" or "NH" before the isoform


## FindDuplicates.pl

```
#!/usr/bin/perl
```

```
use strict;
use warnings;
```


\#\#\# This script will check FASTA files for duplicates and print out the ID for duplicates
\#\#\# Input must be in FASTA format, with one line for sequence ID and a second line for the entire sequence \#\#\#
\#\#\# Use the script "ParseFasta.pl" to convert the input file if needed
\#\#\#
\#\#\# Input must be in FASTA format, with one line for sequence ID and a second line for the entire sequence \#\#\#
\#\#\# Use the script "ParseFasta.pl" to convert the input file if needed
\#\#\# Example Input File Format:
\#\#\#
\#\#\#
\#\#\# ACGTACGTACGTACGT
>Sequence2
>Sequence2
\# >Etc.
\#\#\#
\#\#\#
\#\#\#
\#\#\# 1. Input file [REQUIRED]
- Must be in fasta format (".fa", ".fas", ".fasta")
- Must be parsed (as mentioned above)
\#\#\#
>Sequence1
\#\#\#
\#\#\#
———\#\#

```
                                    push(@results, $ref2);
                    }
        }
    unless(@results == 0){
        print ">" . $ref1 . " matches...\n".
        foreach(@results){
            print "\t" . $_ . "\n";
        print "\n";
            }
    }
}
else{
my $ref1 = "";
my $ref2 = "";
my $IDcounter = 0;
my $resultCounter = 0;
my @results = "";
print "Sequence\t";
foreach $ref1 (@seqOrder){
    $IDcounter++;
        unless($IDcounter == 1){print "\t"};
        print $ref1;
print "\n";
foreach $ref1 (@seqOrder){
        @results = "";
        foreach $ref2'(@seqOrder){
            if ($hash{$ref1} eq $hash{$ref2})
                {
                push(@results, "1");
            else{
                        push(@results, "0");
                        }
            }
        print $ref1
        $resultCounter = 0;
        foreach(@results){
            $resultCounter++;
            unless($resultCounter == 1){print "\t";}
            print $_;
        print "\n";
        pri
}
```


## ParseFasta.pl

```
#!/usr/bin/perl
use warnings;
use strict;
```



```
## Identify IN/OUT files (determined by Bash calling script) ##
my $InFileName = $ARGV[0];
my $OutFileName = $ARGV[1];
unless(@ARGV == 2){die("Usage: ParseFasta-Old.pl <IN-FILE.fasta> <OUT-FILE.fasta>\n");}
## Check if the output file exists and delete it if it does to avoid appending to the end ##
if (-e $OutFileName) {unlink($OutFileName) or die("Can't unlink old file because: $!\n");}
## Opens the appropriate files ##
open (INFILE, "<$InFileName") or die ("Can't open input file because: $!");
open (OUTFILE, ">>$OutFileName") or die ("Can't open output file because: $!\n");
## Reads each line of input file and deals with it accordingly ##
my $ID = "";
my $SEQ = "";
while (<INFILE>) {;
    $ =~ s/\s+$//;
    if ($_ =~ /^>/) {
                print OUTFILE ("$SEQ\n") unless ($SEQ eq "");
            $ID = $_;
            print OUTFILE "$ID\n";
    $SEQ = "";
    }
        else {
            $SEQ .= $_;
    }
}
print OUTFILE "$SEQ\n";
```


## Rename-Previous-PRFs.pl

```
#!/usr/bin/perl
#############################################################################################################
### Extract the accession number, species, and isoform of all previously-discovered PRF sequences
###
### Arguments:
### 1. Input File
    2. Output File
###
### Open dependencies, check for proper usage ###
use strict; use warnings; use Bio::SeqIO;
unless(@ARGV == 2){die("Usage: Rename-Previous-PRFs.pl <IN-FILE> <OUT-FILE>\n");}
### Open input and output files ###
my $infile = $ARGV[0];
my $outfile = $ARGV[1];
open(INFILE, "<$infile") or die("Can't open input file because:$!\n");
open(OUTFILE, ">$outfile");
### Read in sequences, save ID/sequences as a hash, and store order in an array ###
my $ID = ""; my $SEQ = ""; my %sequences; my @seqOrder;
while (<INFILE>){
    $_ =~ s/\s+$//;
    if ($_ =~ /^>/) {
                unless ($SEQ eq ""){
                        $sequences{$ID} = $SEQ;
                        }
            $ID = $_;
            push (@seqOrder, $ID);
            $SEQ = "";
            }
    els
    $sequences{$ID} = $SEQ;
close(INFILE);
### Identify accession number, species name, and PRF isoform - print them all to screen ###
$ID = ""; my $spp = ""; my $isoform = ""; my $accession = "";
print "These are the new headers...\n";
foreach $ID (@seqOrder){
    # Identify accession number or die #
    if ($ID =~ /gi\|(\d+)\|/){
            $accession = $1;
            $\mp@code{ac}
    else {die("Species ID not recognizable for: $ID\n");}
    # Identify species or die #
    if ($ID =~ /Plethodon\s([a-zA-Z]+)\s/){
            $spp = "P." . $1;
            }
    else {die("Species ID not recognizable for: $ID\n");}
    # Identify isoform or die #
    if($ID =~ /isoform\s(\d+)\s/){
            $isoform = $1;
    elsif($ID =~ /PRF([abAB]\d+)\s/){
            $isoform = $1;
            }
        else{die("Isoforms not recognizable for: $ID\n");}
        $isoform = "PRF_" . $isoform;
    if ($spp eq "P.cinereus"){$isoform = "VA_" . $isoform}
    # Print out headers with full new information #
    my $new_header = ">" . $spp . "_" . "gi|" . $accession . "|_" . $isoform;
    print $new_header . "\n";
    # Print out sequences with new headers #
    print OUTFILE $new_header . "\n";
    print OUTFILE $sequences{$ID} . "\n";
    }
```


## Translate.pl

```
#!/usr/bin/perl
```



```
unless(@ARGV == 2){die("Usage: Translate.pl <IN-FILE> <OUTFILE>\n");}
## Open dependencies ##
use warnings,
use Bio::SeqIO;
## Identify IN/OUT files (determined by Bash calling script) ##
$InFileName = $ARGV[0];
$OutFileName = $ARGV[1]
## Check if the output file exists and delete it to avoid appending to the end of the old file ##
if (-e $OutFileName) {unlink($OutFileName) or die("Can't unlink old file because: $!\n");}
## Open the input and output file ##
$seqIO_obj = Bio::SeqIO->new(-file => $InFileName, -format => "fasta");
open (OUTFILE, ">>$OutFileName") or die ("Can't open output file because: $!\n");
## Read in sequences from file, translate, print sequence to outfile ##
while($seq_obj = $seqIO_obj->next_seq){
    $prot_obj = $seq_obj->translate(-complete => 1, -throw => 1);
    print OUTFILE ">" . $seq_obj->display_id . "\n" . $prot_obj->seq . "\n";
}
close OUTFILE;
```


## Untranslate-Alignment.pl

\#!/usr/bin/perl
use strict; use warnings;
\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#
\#\#\# $\quad$ \#\#\#
\#\#\# Take a protein alignment file and untranslate it based on the nucleotide sequences (from reference) \#\#\# \#\#\# This is helpful after translating mRNA sequences for a protein-based alignment
\#\#\# Notes:
\#\#\# - If the headers are not identical in the two files, the script will not work
\#\#\# - The reference file (pre-translated mRNA sequences) needs to contain the coding sequence only! \#\#\#
\#\#\# - Depending on the alignment, it may result in sequences with an extra codon \#\#\#
Due to the addition of stop codons)
\#\#\# - As such, all results should be inspected manually (stop codons especially) \#\#\#
\#\#\#
\#\#\#
\#\#\# 1. Aligned protein file (fasta)
\#\#\# 2. Original mRNA sequence file (fasta)
\#\#\# 3. New, aligned mRNA sequence file (output)
\#\#\#\#\#\#\#\#\#\#\#\# \#
unless(@ARGV == 3)\{
die("Usage: Untranslate-Alignment.pl <ALIGNED-AA-FILE> <UNALIGNED-mRNA-FILE> <OUTPUT:ALIGNED-
mRNA>\n") ;
\#\#\# Determine inputs and outputs \#\#\#

```
my $protAlignment = $ARGV[0];
my $original_mRNA = $ARGV[1];
my $outfile = $ARGV[2];
open(OUTFILE, ">$outfile");
my $ID = "";
my $SEQ = ""';
my @SeqOrder;
my%AAseqs;
while (<INFILE1>) {;
    $_ =~ s/\s+$//;
    if ($_ =~ /^>/){
        unless($ID eq ""){
            $AAseqs{$ID} = $SEQ;
            }
            $ID = $_;
            push(@SeqOrder, $ID);
            $SEQ = "";
            }
            else{
            $SEQ .= $_;
        }
}
$AAseqs{$ID} = $SEQ;
```

open(INFILE1, "<\$protAlignment") or die("Cannot open \$protAlignment because: \$!");
open(INFILE2, "<\$original_mRNA") or die("Cannot open \$original_mRNA because: \$!");
\#\#\# Read in AA file, save the order of sequence IDs in an array, store ID/sequence pair in a hash \#\#\#
\#\#\# Read in (unaligned) mRNA file (original sequences), store ID/sequence pair in a hash \#\#\#
\$ID = "";
\$SEQ = "";
my\%RNAseqs;
while (<INFILE2>) \{;
\$ $=\sim \mathrm{s} / \backslash \mathrm{s}+\$ / / ;$
if (\$_ =~/^>/)\{
unless(\$ID eq "") \{
\$RNAseqs $\{\$ I D\}=\$ S E Q ;$
\}
\$ID = \$_;
\$SEQ = "'"
\}
else\{
\$SEQ . = \$_;
\}
\}
\$RNAseqs $\{\$ I D\}=\$ S E Q$;
\#\#\# Read in aligned AA sequences, print out the appropriate mRNA sequence including gaps \#\#\#

```
foreach my $header (@SeqOrder){
    # Print the sequence ID to get things started
    print OUTFILE $header . "\n";
    # Split protein sequence into characters in array
    my @ProteinSeq = split(//, $AAseqs{$header});
    # Do the same for the nucleotide sequence
    my @NucleotideSeq = split(//, $RNAseqs{$header});
    # Initialize nucleotide codon position counters
    my $codonPos1 = 0; my $codonPos2 = 1; my $codonPos3 = 2;
    # Identify stop codon sequence (last three nucleotides)
    my $stopCodon = "";
    if($RNAseqs{$header} =~ /(\w{3})$/){
            $stopCodon = $1;
            }
    my $newSeq = "".
    foreach my $residue (@ProteinSeq){
            # Loop through protein sequences...
            # If gap in protein position, print out three nucleotide gaps
            if($residue eq "-"){
                                    $newSeq = $newSeq . "---"
            # If residue in protein position, print out corresponding codon
            else{
                $newSeq = $newSeq . $NucleotideSeq[$codonPos1] . $NucleotideSeq[$codonPos2] .
$NucleotideSeq[$codonPos3]
                                    # Increment codon positions for next iteration
                                    # Only when no gap in protein sequence
                                    $codonPos1 = $codonPos3 + 1;
                                    $codonPos2 = $codonPos1 + 1;
                                    $codonPos3 = $codonPos2 + 1;
            }
    }
    if($newSeq =~ /\w$/){
            $newSeq = $newSeq . $stopCodon;
            }
    elsif($newSeq =~ /[\-]$/){
    $newSeq =~ s/([\-]{3})([\-]*)$/$stopCodon$2/;
    }
}
```


## Re-Gap.pl

```
#!/usr/bin/perl
use strict;
use warnings:
```


\#\#\# Get arguments or die \#\#\#
unless(@ARGV == 3)\{die("Usage: Re-Gap.pl <UNALIGNED-FILE> <ALIGNED-REFERENCE> <OUTPUT-FILE>\n")\}
my \$unaligned = \$ARGV[0];
my \$aligned = \$ARGV[1];
my \$outfile = \$ARGV[2];
\#\#\# Read in the unaligned file, store ID/sequence pair in a hash \#\#\#
open(INFILE, "<\$unaligned") or die("Cannot open unaligned file (\$unaligned) because: \$! \n");
my \$ID =
my \$SEQ = ""
my @SeqOrder;
my \%NoGaps;
while (<INFILE>) \{
$\$_{-}=\sim s / \backslash s+\$ / / ;$
if (\$_ =~ / $>/$ ) $\{$
unless(\$ID eq "")\{
\$NoGaps $\{\$ I D\}=\$ S E Q ;$
\$ID = \$_;
\$SEQ = "'"; push(@SeqOrder, \$ID); \}
else\{
\$SEQ . = \$_;
\}
\}
\$NoGaps\{\$ID\} = \$SEQ;
close(INFILE);
\#\#\# Read in the aligned reference file, store ID/sequence pair in a hash \#\#\#
open(INFILE, "<\$aligned") or die("Cannot open reference file (\$aligned) because: \$!\n");

```
$ID = "";
$SEQ = "";
my %Gaps;
while (<INFILE>) {;
        $_ =~ s/\s+$//;
        if ($_ =~ /^>/){
            unless($ID eq ""){
                $Gaps{$ID} = $SEQ;
            TID = $ }
            $ID = $_;
    }
    else{
            $SEQ .= $_;
    }
}
$Gaps{$ID} = $SEQ;
close(INFILE);
```

\#\#\# Index gaps in aligned file, insert them into unaligned sequence for output \#\#\#
open(OUTFILE, ">\$outfile");
foreach(@SeqOrder) \{
my @Aligned_Seq = split(//, \$Gaps\{\$_\});
my @Unaligned_Seq = split(//, \$NoGaps\{\$_\});
print OUTFILE \$_ . "\n";
my \$i = "";
my $\$ x=0$
foreach \$i (@Aligned_Seq) \{
if(\$i eq '-')\{
print OUTFILE '-';
else\{
print OUTFILE \$Unaligned_Seq[\$x]; \$x++;
\}
print OUTFILE "\n";
\}
close(OUTFILE);

## Parse-Nglyc.pl




```
\begin{tabular}{|c|c|c|}
\hline \#\#\# & & \#\#\# \\
\hline \#\#\# & \multirow[t]{2}{*}{START SCRIPT} & \#\#\# \\
\hline \#\#\# & & \#\#\# \\
\hline
\end{tabular}
### Read in arguments or show usage information ###
unless(@ARGV == 2){die("Usage: Parse-Nglyc.pl <INPUT-FILE> <OUTPUT-FILE>\n")}
my $infile = $ARGV[0];
my $outfile = $ARGV[1];
open(INFILE, "<$infile");
open(OUTFILE, ">$outfile");
my $species = "";
my $spp = "";
my $spp_check = "";
my @sites;
while(<INFILE>) {
    if($_ =~ /Name:(?:\s+)(.+)(?:\s+)Length/){
    unless($species eq ""){
                        print OUTFILE ">" . $species;
                        unless(@sites == 0){
                                    foreach(@sites){
                                    print OUTFILE "\t" . $_;
                                    }
                                    print OUTFILE "\n";
                                    }
                                    $species =~ s/^\s+//;
                                    $species =~ s/\s+$//;
                                    $spp = quotemeta($species);
                    @sites=();
        elsif($_ =~ /No\ssi
                next;
        elsif($_ = = /^($spp)\s+(\d+)\s+\w+\s+.*([\+-])$/){
            $spp_check = $1;
                my $site = $2 . $3;
                push(@sites, $site);
                unless($spp_check eq $species){die("Error parsing results for sequence: $species\n")}
                }
        }
print OUTFILE ">" . $species;
foreach(@sites){
            unless(@sites == 0){
                print OUTFILE "\t" . $_;
            }
print OUTFILE "\n";
close(INFILE);
close(OUTFILE);
```


## Rename-Long-to-Short.pl

```
#!/usr/bin/perl
##############################################################################################################
###
### Extracts the accession number, species, and isoform of all previously-discovered PRF sequences
### It will then print out a new header consisting of only the GI number
###
### Some programs need a shorter header name (e.g. PROMALS3d, netOglyc) so use this to shorten the names
### Then use "Rename-Short-to-Long.pl" (using the original input file from this script) as a reference
### This allows you to rename the headers to their full length (as they were before running this script)
###
### For the record: this is utterly ridiculous!
###
### Arguments:
### 1. Input file to shorten (this very file will then be the reference for "Rename-Short-to-Long.pl") ###
### 2. Output file (this will then be the input file for "Rename-Short-to-Long.pl")
###
####################################################################################################################
### Open dependencies, check for proper usage ###
use strict;
use warnings;
use Bio::SeqIO;
unless(@ARGV == 3){
    die("Usage: Rename-Previous-PRFs.pl <IN-FILE> <OUT-FILE> <IDENTIFIER: \"--gi\" OR \"--acc\">\n");
    }
### Open input and output files ###
my $infile = $ARGV[0];
my $outfile = $ARGV[1];
open(INFILE, "<$infile") or die("Can't open input file because:$!\n");
open(OUTFILE, ">$outfile");
### Read in sequences, save ID/sequences as a hash, and store order in an array ###
my $ID = ""; my $SEQ = ""; my %sequences; my @seqOrder;
while (<INFILE>){
    $_ =~ s/\s+$//;
    if ($_ =~ /^>/) {
                unless ($SEQ eq ""){
                        $sequences{$ID} = $SEQ;
                }
                $ID = $_;
                push (@seqOrder, $ID);
                $SEQ = "";
    else {
                $SEQ .= $_;
        }
    }
    $sequences{$ID} = $SEQ;
close(INFILE);
### Identify accession number, species name, and PRF isoform - print them all to screen ###
$ID = ""; my $spp = ""; my $isoform = ""; my $identifier = "";
print "These are the new headers...\n";
foreach \$ID (@seqOrder) \{
```

```
# Identify either accession number of GI number, depending on argument #
```


# Identify either accession number of GI number, depending on argument

if("--gi" ~~ @ARGV){
if("--gi" ~~ @ARGV){
\# Identify gi number or die \#
\# Identify gi number or die \#
if ($ID =~ /gi\|(\d+)\|/){
    if ($ID =~ /gi\|(\d+)\|/){
\$identifier = \$1;
\$identifier = \$1;
}
}
else {
else {
print "GI number not recognizable for: \$ID\n";
print "GI number not recognizable for: \$ID\n";
\$identifier = \$ID;
\$identifier = \$ID;
\$identifier =~ s/>//;
$identifier =~ s/>//;
            }
            }
    }
    }
elsif("--acc" ~~ @ARGV){
elsif("--acc" ~~ @ARGV){
    # Identify accession number or die #
    # Identify accession number or die #
    if ($ID =~ /gb\|(\w{2}\d{6}\.\d)\|/){

```
    if ($ID =~ /gb\|(\w{2}\d{6}\.\d)\|/){
```

```
                        $identifier = $1;
                    }
else {
print "Accession number not recognizable for: $ID\n";
            $identifier = $ID;
    $identifier =~ s/>//;
}
```

\# Print gi number only - a short name is necessary for some programs \# my \$new_header = ">" . \$identifier;
print \$new_header . "\n";
\# Print out sequences with new headers \#
print OUTFILE \$new_header . "\n";
print OUTFILE \$sequences\{\$ID\} . "\n";
\}
close(OUTFILE);

## Rename-Short-to-Long.pl

```
#!/usr/bin/perl
#############################################################################################################
###
### Re-create the full header name from a file that was shortened to only the GI number
###
### "Rename-Long-to-Short.pl" shortens names to the index number, and this will regain the full header ###
###
### Arguments
### 1. Reference file (with full header names) - the input to "Rename-Long-to-Short.pl"
### 2. Input file with short names (gi number only) - the output of "Rename-Long-to-Short.pl")
### 3. Output File
###
##############################################################################################################
### Open dependencies, check for proper usage ###
use strict;
use warnings
use Bio::SeqIO;
unless(@ARGV == 4){
    die("Usage: Rename-Previous-PRFs.pl <REFERENCE-FILE> <IN-FILE> <OUT-FILE> <IDENTIFIER: \"--gi\" OR
\"--acc\">\n");
    }
### Open input and output files ###
my $infile_ref = $ARGV[0];
my $infile_short = $ARGV[1];
my $outfile = $ARGV[2];
### Read in reference file, save full header as a value in a hash (with the gi number as the key) ###
open(INFILE1, "<$infile_ref") or die("Can't open reference file because:$!\n");
my $identifier1 = ""; my %header_hash;
while (<INFILE1>){
    $_ =~ s/\s+$//;
    if ($_ =~ /^>/) {
                if("--gi" ~~ @ARGV){
                        if ($_ =~ /gi\|(\d+)\|/){
                        $identifier1 = $1;
                            $header_hash{$identifier1} = $_;
                            }
                elsif("--acc" ~~ @ARGV){
                    if ($_ =~ /gb\|(\w{2}\d{6}\.\d)\|/){
                        $identifier1 = $1;
                        $header_hash{$identifier1} = $_;
                        }
                }
    else {next;}
    }
close(INFILE1);
### Read in file (with short names) and save sequences in a hash (current header is the key) ###
open(INFILE2, "<$infile_short") or die("Can't open renaming file because:$!\n");
my $ID = ""; my $SEQ = ""; my $identifier2; my %sequences; my @seqOrder;
while (<INFILE2>){
    $_ =~ s/\s+$//
            if ($_ =~ /^>/) {
                unless ($SEQ eq ""){
                    $sequences{$ID} = $SEQ;
            $ID = $_;
        push (@seqOrder, $ID);
        $SEQ = "";
    else {
            SEQ .= $_;
            }
    }
    $sequences{$ID} = $SEQ;
```

```
close(INFILE2);
open(OUTFILE, ">$outfile");
foreach(@seqOrder){
    if("--gi" ~~ @ARGV){
        if ($_ =~ /^>(\d|*)$/){
                                $identifier2 = $1;
                                print OUTFILE $header_hash{$identifier2} . "\n";
                                print OUTFILE $sequences{$_} . "\n";
            else{
                                print OUTFILE $_ . "\n";
                                print OUTFILE $sequences{$_} . "\n";
                                pri
        elsif("--acc" ~~ @ARGV){
            if ($_ =~ /^>(\w{2}\d{6}\.\d)$/){
                                $identifier2 = $1;
                                print OUTFILE $header_hash{$identifier2} . "\n";
                                print OUTFILE $sequences{$_} . "\n";
            else{
                        print OUTFILE $_. "\n";
                                print OUTFILE $sequences{$_} . "\n";
                }
            }
    }
```

close(OUTFILE);

## Parse-Oglyc.pl

```
#!/usr/bin/perl
use strict;
use warnings:
#######################################################################################################################
### ###
### Parse the output of standalone netOglyc (3.1d)
###
####
### This script will print out a tab-delimited file with the sequence name, followed by the locations ###
### of each S/T site with a "+" or "-" indicating that glycosylation is predicted or not (respectively). ###
###
### Note:
###
###
### - This extracts only the minimal essential results for plotting annotations ###
### - For more information visit the raw output file or modify the script to print more information ###
###
\#\#\#
###
1. Input File (in fasta format) ###
2. Output File (tab-delimited file with short sequence names and all S/T sites +/- glycosylation) ###
### ###
### EXAMPLE OUTPUT ENTRY FOR NetOGlyc 3.1d ###
```



```
#################################################################################################################
\#\#\#
\begin{tabular}{ll} 
\#\#\# START SCRIPT \\
\#\#\# & \\
\#\#\#
\end{tabular}
##############################################################################################################
### Read in arguments or show usage information ###
unless(@ARGV == 2){die("Usage: Parse-Oglyc.pl <INPUT-FILE> <OUTPUT-FILE>\n")}
my $infile = $ARGV[0];
my $outfile = $ARGV[1];
### Read in results and print to output file ###
open(INFILE, "<$infile");
open(OUTFILE, ">$outfile");
my $species = "";
my $spp = "";
my $spp_check = "";
my @sites;
while(<INFILE>) {
    if($_ =~ /Name:(?:\s+)(.+)(?:\s+)Length/) {
                unless($species eq ""){
                    print OUTFILE ">" . $species;
                    unless(@sites == 0){
                                    foreach(@sites){
                                    print OUTFILE "\t" . $_;
                                    }
                    }
                    print OUTFILE "\n";
                $species \stackrel{}}{=}$1;
                    $species =~ s/^\s+//;
                $species =~ s/\s+$//;
                    $spp = quotemeta($species);
                @sites=();
```

```
    elsif($_ = =~ /^($spp)\s+([ST])\s+(\d+)\s+(\d\.\d{3})\s+(\d\.\d{3})\s+([\.ST])/){
        $spp_check = $1;
        my $prediction = "";
        if(($6 eq "S") || ($6 eq "T")){
                        $prediction = "+";
            $pr
            eq "."){
            $prediction = "-";
                }
        my $site = $3 . $2 . $prediction;
        push(@sites, $site);
        unless($spp_check eq $species){die("Error parsing results for sequence: $species\n")}
        }
    }
print OUTFILE ">" . $species;
foreach(@sites){
        unless(@sites == 0){
        print OUTFILE "\t" . $_;
    }
print OUTFILE "\n";
close(INFILE);
close(OUTFILE);
```


## SubsetSeqs.pl

```
#!/usr/bin/perl
use warnings; use strict;
##############################################################################################################
```



```
\#\#\# Split fasta file into multiple subsets \#\#\#
### Arguments: ###
    1. Input file (must be parsed, two-line fasta format - e.g. after ParseFasta.pl) ###
    2. Number of sequences per file (i.e. how many sequences to output per subset) ###
###
#############################################################################################################
unless (@ARGV == 2){die("Usage: SubsetSeqs.pl <IN-FILE> <#-SEQUENCES-PER-FILE>\n")}
my $infile = $ARGV[0];
my $num_seqs = $ARGV[1]
open(INFILE, "<$infile");
my $count = 0;
my $file_num = 1;
my $outfile = "";
my $id = "";
my $seq = "";
while(<INFILE>) {
    chomp;
    $outfile = "Subset" . $file_num . ".fasta";
    open(OUTFILE, ">>$outfile");
    if($_ =~ /^>/){
        $count++;
        $id = $_;
        }
    else{
        $seq = $_;
        print OUTFILE $id . "\n" . $seq . "\n".
        if($count == $num_seqs){
                        $count = 0;
                        $file_num++
                }
    close(OUTFILE);
close(INFILE);
```


# Parse-Oglyc-version4.pl 

```
#!/usr/bin/perl
use strict;
use warnings;
#################################################################################################################
### ###
### Parse the output of netOglyc 4.0
#### This script will print out a tab-delimited file with the sequence name, followed by the locations
###
###
### This script will print out a tab-delimited file with the sequence name, followed by the locations ###
### of each S/T site with a "+" or "-" indicating that glycosylation is predicted or not (respectively). ###
###
### Note:
###
### - This extracts only the minimal essential results for plotting annotations ###
###
### - For more information visit the raw output file or modify the script to print more information
###
### Arguments:
###
### 1. Input File (concatenated output of NetOglyc 4.0)
### 3. Reference File - the sequences predicted (parsed two-line fasta - e.g after "Parse-Fasta,pl") ###
### 2. Output File (tab-delimited file with short sequence names and all S/T sites +/- glycosylation) ###
### ###
### EXAMPLE OUTPUT ENTRY FOR NetOGlyc 4.0 - manually concatenate all sequences into file with one header ###
\begin{tabular}{|c|c|c|c|c|c|c|c|}
\hline \# \#seqname & source feature start & end score & strand & frame & comment & & \\
\hline \# NH_PRF_30 & netOGlyc-4.0.0.13 & CARBOHYD & 3 & 3 & 0.0443054 & & \\
\hline \# NH_PRF_30 & netOGlyc-4.0.0.13 & CARBOHYD & 13 & 13 & 0.353207 & & \\
\hline \# NH_PRF_30 & netOGlyc-4.0.0.13 & CARBOHYD & 18 & 18 & 0.284046 & & \\
\hline \# NH_PRF_30 & netOGlyc-4.0.0.13 & CARBOHYD & 19 & 19 & 0.547452 & . & \\
\hline \#POSITIVE & & & & & & & \\
\hline \# etc.. & & & & & & & \\
\hline \# 60499218 & netOGlyc-4.0.0.13 & CARBOHYD & 3 & 3 & 0.123729 & . & \\
\hline \# 60499218 & netOGlyc-4.0.0.13 & CARBOHYD & 4 & 4 & 0.0676288 & & \\
\hline \# 60499218 & netOGlyc-4.0.0.13 & CARBOHYD & 13 & 13 & 0.347222 & & \\
\hline \# 60499218 & netOGlyc-4.0.0.13 & CARBOHYD & 16 & 16 & 0.0940582 & . & . \\
\hline
\end{tabular}
### END EXAMPLE OUTPUT FORMAT ###
\begin{tabular}{|c|c|c|}
\hline \multicolumn{3}{|l|}{\#\#\# \#\#\#} \\
\hline \#\#\# & START SCRIPT & \#\#\# \\
\hline \#\#\# & & \#\#\# \\
\hline
\end{tabular}
### Read in arguments or show usage information ###
unless(@ARGV == 3){
    die("Usage: Parse-Oglyc-version4.pl <INPUT-FILE> <SEQUENCES-PREDICTED.fasta> <OUTPUT-FILE>\n")
    }
my $infile = $ARGV[0];
my $reference = $ARGV[1];
my $outfile = $ARGV[2];
### Read in reference file, store headers and sequences as a hash, sequence order as an array ###
my $ID = "";
my $SEQ = """.
my %Seqs = (')
my @SeqOrder = ();
open(REFERENCE, "<$reference");
while (<REFERENCE>) {;
        $ =~ s/\s+$//;
            if ($_ =~ /^>/){
                unless($ID eq ""){
                                    $Seqs{$ID} = $SEQ;
            $ID = $_;
            push(@SeqOrder, $ID);
            push(@SeqOr
            }
            else{
                $SEQ .= $_;
            }
            D}
$Seqs{$ID} = $SEQ;
close(REFERENCE);
### Read in and parse results, print results for each sequence to output file ###
```

```
open(INFILE, "<$infile");
my $line_count = 0;
my $seq_new = "";
my $seq_current = ""; 
my $seq_length = "";
my $start_site = "";
my $stop_site = "";
my $score = "";
my %sites = ();
my %PREDICTIONS = ();
while(<INFILE>){
    chomp;
    $line_count++;
    if($_ =~ /seqname\s+source\s+feature\s+start\s+end\s+score\s+strand\s+frame\s+comment/){next;}
    elsif($_ =~ /^(\w+)\s+.*CARBOHYD\s+(\d+)\s+(\d+)\s+(\d\.\d+)/){
        # Get fields #
        $seq_new = $1;
        $start_site = $2;
        $stop_site = $3;
        $score = $4;
        # Save the first sequence as the current one and get its full sequence name #
        if($seq_current eq ""){
            $seq_current = $seq_new;
            my $match_check = "FALSE";
            foreach(keys(%Seqs)){
                                    if($_ =~ /$seq_current/){
                                    $full_seq_name = $_;
                                    $seq_length = length($Seqs{$full_seq_name});
                                    $match_check = "TRUE";
                                    }
                    }
if($match_check eq "FALSE"){
                                    die("Short name couldn't be converted to full name for: $seq_current\n");
                                    }
}
# Save old sequence info if moving onto a new sequence, and... #
# Set current sequence and get its full name
if($seq_new ne $seq_current){
            # Save results in array for last sequence #
            my $i;
            my @KEYS = keys(%sites);
            my @SITE_ARRAY = ();
            for($i = 1; $i <= $seq_length; $i++){
                    if($i ~~ @KEYS){
                                    push(@SITE_ARRAY, $sites{$i})
                    else{
                            push(@SITE_ARRAY, "NA")
                                    }
                    }
                $PREDICTIONS{$full_seq_name} = [ @SITE_ARRAY ];
                    # Set new sequence info #
                    $seq_current = $seq_new;
                %sites = ();
                $sites{$start_site} = $score;
                my $match_check = "FALSE";
                foreach(keys(%Seqs)){
                    if($_ =~ /$seq_current/){
                        $full_seq_name = $_;
                        $seq_length = length($Seqs{$full_seq_name});
                            $match_check = "TRUE"
                                    }
                        }
                            if($match_check eq "FALSE"){
                    die("Short name couldn't be converted to full name for: $seq_current\n");
}
# Continue saving sites if still on the same sequence #
elsif($seq_new eq $seq_current){
                        # Save score for each residue if the site is longer than one residue #
                        # Save score for each residue if
                        my $i = "";
                        for($i = $start_site; $i <= $stop_site; $i++){
                    $sites{$i} = $score;
                        }
```

```
                    }
                # Save score for this residue if only one residue long #
                elsif($start_site == $stop_site){
                            $sites{$start_site} = $score;
        else{die("Could not differentiate start and stop sites at line: $line_count\n")}
                    }
    else{die("Check input file for blank or incompatible lines at line: $line_count\n")}
    }
# Save results in array for final sequence #
my $i;
my @KEYs = keys(%sites);
my @SITE_ARRAY = ();
for($i = 1; $i <= $seq_length; $i++){
    if($i ~~ @KEYS){
        push(@SITE_ARRAY, $sites{$i})
            }
    else{
            push(@SITE_ARRAY, "NA")
            }
    }
$PREDICTIONS{$full_seq_name} = [ @SITE_ARRAY ];
close(INFILE);
open(OUTFILE, ">$outfile");
foreach my $final_id (@SeqOrder){
            print OUTFILE $final_id;
            foreach my $final_site (@{$PREDICTIONS{$final_id}}){
            print OUTFILE "\t" . $final_site
    } print OUTFILE "\n";
close(OUTFILE);
```


## Aligned Protein Sequences

>P.serratus_gi|60499226|_PRF_A1
MRSTALLIFLVVSVRRATSLTIPAPVKRSTSSDVAELSETTIVLFSETQKFAESMKDNAD SLLTTYLSFQGAPLSDPEYHLPEVKIDNLPTAAMDNATFLNQTDEIRLKNNLYFYSAIVQ FLQVAMTEQEDLNPAEESLKAKFQEAMANSNTLISKISAIMTQMGMSVTVTFPDPLVVPY EGSAIFKQKLRGGVICKEYKERVLLTKRDFEFFAKKYHEQLIQQQEQEE
$>P$.serratus_gi|60499228|_PRF_A2
MRSTALLIFLVVSVRSATSLTISAPVKRSTSSDVAELSETTIVLFSETQKFAESMKDNAD SLLTTYLSFQGAPLSDPEYHLPEVKIDNLPTAAMDNATFLNQTDEIRLKNNLYFYSAIVQ FLQVAMTEQEDLNPAEESLKAKFQEAMANSNTLISKISAIMTQMGMSVTVTFPDPLVVPY EGSAIFKQKLRGGVICKEYKERVLLTKRDFEFFAKKYHEQLIQQQEQEE
$>P$.serratus_gi|60499222|_PRF_A3
MRSTALLIFLVVSVRSATSLTISAPVKRSTSSDVAELSKTTIVLFSETQKFVESMKDNAD SLLTTYLSFQGAPLSDPEYHLPEVKIDNLPTATMDNATFLNQTDEIRLKNNLYFYSAIVQ FLQVAMTEQEDLNPAEESLKAKFQEAMANSNTLISKISAIMTQMGMSVTVTFPDPLVVPY EGSAIFKQKLRGGVICKEYKERVLLTKWDFEFFAKKYHEQLIQHQEQEE
>P.serratus_gi|60499230|_PRF_A4
MRSTALLIFLVVSVRRATSLTIPAPVKRSTSSDVAELSKTTIVLFSETQKFVESMKDNAD SLLTTYLSFQGAPLSDPEYHLPEVKIDNLPTATMDNATFLNQTDEIRLKNNLYFYSAIVQ FLQVAMTEQEDLNPAEESLKAKFQEAMANSNTLISKISAIMTQMGMSVTVTFPDPLVVPY EGSAIFKQKLRGGVICKEYKERVLLTKRDFEFFAKKYHEQLIQQQEQEE
>P.serratus_gi|60499224|_PRF_A5
MRSTALLIFLVVSVRSATSLTISAPVKRSTSSDVAELSETTIVLFSETQKFAESMKDNAD SLLTTYLSFQGAPLSDPEYHLPEVKIDNLPTAAMDNATFLNQTDEIRLKNNLYFYSAIVQ FLQVAMTEQEDLNPAEESLKAKFQEAMANSNTLISKISAIMTQMGMSVTVTFPDPLVVPY EGSAIFKQKLRGGVICKEYKERVLLTKRDFEFFAKKYHEQLIQQQEQEE
>P.serratus_gi|60499220|_PRF_B1
MRSTALLIFLVVSVRSATSLTISAPVKRSTSSDVADLSETTIVLFSETQKFAESMKDNAD SLLTTYLSLQGEPYNDPDFSLPDIRIANLPTATMNYDTFMRQTDEIRLKNNLYFYSAIIK FLQLTMDEQKDLNPADPSLNAKFEQVLSNINSLIDNISTIMTQMDFSVKIEFRTPMVESY IGSAIFKKKLRGYVVCREYKERVALTMRDFEFLAKKYEGHLQRERP---
>P.serratus_gi|60499218|_PRF_B2
MRSTALLIFLVVSVRSATSLTISAPVKRSTSSDVADLSETTIVLFSETQEFAESMKDNAD SLLTTYLSLQGEPYNDPDFSLPDIRIANLPTASMNYDTFMRQTDEIRLKNNLYFYSAIIK FLQLTMDEQKDLNPADPSLNAKFEQVLSNINSLIDNISTIMTQMDFSVKIEFRTPMVESY IGSAIFKKKLRGYVVCREYKERVALTMRDFEFLAKKYEGHLQRERP--
>P.cinereus_gi|42566697|_VA_PRF_A1
MRSTALLIFLVVSVRRATSLTISAPVKRSTSSDVAELSETTIVLFSETQKFVESMKDNAD SLLTTYLSLQGAPLSDPEYHLPEVKIDNLPTATMDNATFFNQTDEIRLKNNLYFYSAIVQ FLQVAMTEQEDLNPAEESLKTKFQEALANSNTLISKISAIMTQMGMSVTVTFPDPLVVPY EGSAIFKQKLRGGVICKEYKERVLLTKRDFEFFAKKYHEQLVQQQEQEE
>P.cinereus_gi|42566675|_VA_PRF_A2
MRSTALLIFLVVSVRRATSLTISAPVKRSTSSDVAELSETTIVLFSETQKFVESMKDNAD SLLTTYLSFQGAPLSDPEYHLPEVKIDNLPTATMDNATFFNQTDEIRLKNNLYFYSAIVQ FLQVAMTEQEDLNPAEESLKTKFQEALANSNTLISKISAIMTQMGMSVTVTFPDPLVVPY EGSAIFKQKLRGGVICKEYKERVLLTKRDFEFFAKKYHEQLVQQQEQEE
>P.cinereus_gi|42566685|_VA_PRF_A3
MRSTALLIFLVVSVRRATSLTISAPVKRSTSSDVAELSETTIVLFSETQKFVESMKDNAD SLLTTYLSFQGAPLSDPEYHLPEVKIDNLPTATMDNATFFNQTDEIRLKNNLYFYSAIVQ FLQVAMTEQEDLNPAEESLKTKFQEALANSNTLISKISAIMTQMGMSVTVTFPDPLVVPY EGSAIFKQKLRGGVICKEYKERVLLTKRDFEFFAKKYHEQLVQQQEQEE
>P.cinereus_gi|42566669|_VA_PRF_A4
MRSTALLIFLVVSVRRATSLTISAPVKRSTSSDVAELSETTIVLFSETQKFVESMKDNAD SLLTTYLSFQGAPLSDPEYHLPEVKIDNLPTATMDNATFLNQTDEIRLKNNLYFYSAIVQ FLQVAMTEQEDLNPAEESLKAKFQEALANSNTLISKISAIMTQMGMSVTVTFPDPLVVPY EGSAIFKQKLRGGVICKEYKERVLLTKRDFEFFAKKYHEQLVQQQEQEE
>P.cinereus_gi|42566681|_VA_PRF_A5
MRSTALLIFLVVSVRRATSLTISAPVKRSTSSDVADLSKTTIVLFSETHKFVESMKDNAD SLLTTYLSFQGAPLSDPEYHLPEVKIDNLPTATMDNATFLNQTDEIRLKNNLYFYSAIVQ FLQVAMTEQEDLNPAEESLKAKFQEALANSNTLISKISAIMTQMGMSVTVTFPDPLVVPY EGSAIFKQKLRGGVICKEYKERVLLTKRDFEFFAKKYHEQLIQQQEQEE
>P.cinereus_gi|42566695|_VA_PRF_A6
MRSTALLIFLVVSVRRATSLTISAPVKRSTSSDVADLSKTTIVLFSETHKFVESMKDNAD SLLTTYLSFQGAPLSDPEYHLPEVKIDNLPTATMDNATFFNQTDEIRLKNNLYFYSAIVQ FLQVAMTEQEDLNPAEESLKTKFQEALANSNTLISKISAIMTQMGMSVTVTFPDPLVVPY EGSAIFKQKLRGGVICKEYKERVLLTKRDFEFFAKKYHEQLVQQQEQEE
>P.cinereus_gi|42566689|_VA_PRF_A7
MRSTALLIFLVVSVRRATSLTISAPVKRSTSSDVADLSKTTIVLFSETHKFVESMKDNAD SLLTTYLSFQGAPLSDPEYHLPEVKIDNLPTATMDNATFLNQTDEIRLKNNLYFYSAIVQ FLQVAMTEQEDLNPAEESLKAKFQEALANSNTLISKISAIMTQMGMSVTVTFPDPLVVPY EGSAIFKQKLRGGVICKEYKERVLLTKRDFEFFAKKYHEQLIQQQEQEE
>P.cinereus_gi|42566693|_VA_PRF_A8
MRSTALLIFLVVSVRRATSLTISAPVKRSTSSDVADLSKTTIVLFSETHKFVESMKDNAD SLLTTYLSFQGAPLSDPEYHLPEVKIDNLPTATMDNATFLNQTDEIRLKNNLYFYSAIVQ FLQVAMTEQEDLNPAEESLKAKFQEALANSNTLISKISAIMTQMGMSVTVTFPDPLVVPY EGSAIFKQKLRGGVICKEYKERVLLTKRDFEFFAKKYHEQLVQQQEQEE
>P.cinereus_gi|42566709|_VA_PRF_B1
MRSTALLIFLVVSVRRATSLTISAPVKRSTSSDVADLSETTIVLFSETQKFAESMKDNAD SLLTTYLSLQGEPYNNPDFSLRDIQIANLPTATMNYDTFMRQTDEIRLKNNLYFYSAIIK FLQLTMDEQKDLNPEDPSLNAKFEQVLSNINSLINKISTIMTQMDFSVKIEFRSPMVESY IGSAIFKKKLRGYVVCREYKERVALTMRDFEFLAKKYEGHLHRARP---
>P.cinereus_gi|42566707|_VA_PRF_B2
MRSTALLIFLVVSVRRATSLTISAPVKRSTSSDVADLSETTIVLFSETQKFAESMKDNAD

SLLTTYLSLQGEPYNNPDFSLRDIQIANLPTATMNYDTFMRQTDEIRLKNNLYFYSAIIK FLQLTMDEQKDLNPEDPSLNAKFEQVLSNINSLINKISTIMTQMDFSVKIEFRTPMVESY IGSAIFKKKLRGYIVCREYKERVALTMRDFEFLAKKYEGHLHRARP---
>P.cinereus_gi|42566711|_VA_PRF_B3
MRSTALLIFLVVSVRRATSLTISAPVKRSTSSDVADLSETTIVLFSETQKFAESMKDNAD SLLTTYLSLQGEPYNNPDFSLRSIQIANLPTATMNYDTFMRQTDEIRLKNNLYFYSAIIK FLQLTMDEQKDLNPEDPSLNAKFEQVLSNINSLINKISTIMTQMDFSVKIEFRTPMVESY IGSAIFKKKLRGYIVCREYKERVALTMRDFEFLAKKYEGHLHRARP---
>P.cinereus_gi|42566699|_VA_PRF_B4
MRSTALLIFLVVSVRRATSLTISAPVKRSTSSDVADLSETTIVLFSETQKFAESMKDNAD SLLTTYLSLQGEPYNNPDFSLRDIQIANLPTATMNYDTFMRQTDEIRLKNNLYFYSAIIK FLQLTMDEQKDLNPEDPSLNAKFEQVLSNINSLINKISTIMTQMDFSVKIEFRSPMVESY IGSAIFKKKLRGYVVCREYKERVLLTKRDFEFFAKKYHEQLVQQQEQEE
>P.cinereus_gi|42566703|_VA_PRF_B5
MRSTALLIFLVVSVRRATSLTISAPVKRSTSSDVAELSETTIVLFSETQKFVESMKDNAD SLLTTYLSLQGEPYNNPDFSLRDIQIANLPTATMNYDTFMRQTDEIRLKNNLYFYSAIIK FLQLTMDEQKDLNPEDPSLNAKFEQVLSNINSLINKISTIMTQMDFSVKIEFRSPMVESY IGSAIFKKKLRGYVVCREYKERVALTMRDFEFLAKKYEGHLHRARP---
>P.cinereus_gi|42566705|_VA_PRF_B6
MRSTALLIFLVVSVRRATSLTISAPVKRSTSSDVAELSETTIVLFSETQKFAESMKDNAD SLLTTYLSLQGEPYNNPDFSLRDIQIANLPTATMNYDTFMRQTDEIRLKNNLYFYSAIIK FLQLTMDEQKDLNPEDPSLNAKFEQVLSNINSLINKISTIMTQMDFSVKIEFRSPMVESY IGSAIFKKKLRGYVVCREYKERVALTMRDFEFLAKKYEGHLHRARP - -
>P.cinereus_NH_PRF_A1
MRSTALLIFLVVSVRRATSLTISSPVKRSTSSDVADLSKTTIVLFSETHKFVESMKDNAD SLLTTYLSFQGAPLSDPEYHLPEVKIDNLPTATMDNTTFLNQTDEIRLKNNLYFYSAIVQ FLQVAMTEQEDLNPVEESLKTKFQEALANSNTLISKISAIMTQMGMSATVTFPDPLVVPY EGSAIFKQKLRGGVICKEYKERVLLTKRDFEFFAQKYHEQLIQQQEQEE
>P.cinereus_NH_PRF_A2
MRSTALLIFLVVSVRRATSLTISSPVKRSTSSDVADLSKTTIVLFSETHKFMESMKDNAD SLLTTYLSFQGAPLSDPEYHLPEVKIDNLPTATMDNTTFLNQTDEIRLKNNLYFYSAIVQ FLQVAMTEQEDLNPAEESLKTKFQEALANSNTLISKISAIMTQMGMSATVTFPDPLVVPY EGSAIFKQKLRGGVICKEYKERVLLTKRDFEFFAQKYHEQLIQQQEQEE
>P.cinereus_NH_PRF_A3
MRSTALLIFLVVSVRRATSLTISSPVKRSTSSDVADLSKTTIVLFSETHKFVESMKDNAD SLLTTYLSFQGAPLSDPEYHLPEVKIDNLPTATMDNTTFLNQTDEIRLKNNLYFYSAIVQ FLQVAMTEQEDLNPAEESLKTKFQEALANSNTLISKISAIMTQMGMSATVTFPDPLVVPY EGSAIFKQKLRGGVICKEYKERVLLTKRDFEFFAQKYHEQLIQQQEQEE
>P.cinereus_NH_PRF_A4
MRSTALLIFLVVSVRRATSLTISSPVKRSTSSDVADLSKTTIVLFSETHKFVESMKDNAD SLLTTYLSFQGAPLSDPEYHLPEVKIDNLPTATMDNTTFLNQTDEIRLKNNLYFYSAIVQ FLQVAMTEQEDLNPAEESLKTKFQEALANSNTLISKISAIMTQMGMSATVTFPDPLVVPY EGSAIFKQKLRGGVICKEYKERVLLTKRDFEFFFAQKYHEQLIQQQEQEE
>P.cinereus_NH_PRF_A5
MRSTALLIFLVVSVRRATSLTISSPVKRSTSSDVADLSKTTIVLFSETHKFVESMKDNAD SLLTTYLSFQGAPLSDPEYHLPEVKIDNLPTATMDNTTFLNQTDEIRLKNNLYFYSAIVQ FLQVAMTEQEDLNPAEESLKTKFQEALANSNTLISKISAIMTQMGMSATVTFPDPPLVVPY EGSAIFKQKLRGGVICKEYKERVLLTKRDFEFFFAQKYHEQLIQQQEQEE
>P.cinereus_NH_PRF_A6
MRSTVLLILLVVSVRRATSLTISAPVKRSTSSDVADLSETTIVLFSETQKFVESMKDNAD SLLTTYLSFQGAPLSDPEYHLPEVKIDNLPTATMDNATFLNQTDEIRLKNNLYFYSAIVQ FLQVAMTEQEDLNPAEESLKAKFQEALANSNTLISKISAIMTQMGMSATVTFPDPLVVPY EGSAIFKQKLRGGVICKEYKERVLLTKHDFEFFFAKKYHEQLIQQQEQEE
>P.Cinereus_NH_PRF_A7
MRSTVLLIFLVVSVRRATSLTISAPVKRSTSSDVADLSETTIVLFSETQKFVESMKDNAD SLLTTYLSFQGAPLSDPEYHLPEVKIDNLPTATMDNATFLNQTDEIRIKNNLYFYSAIVQ FLQVAMTEQEDLNPAEESLKAKFQEALANSNTLISKISAIMTQMGMSATVTFPDPLVVPY EGSAIFKQKLRGGVICKEYKERVLLTKQDFEFFFAKKYHEQLIQQQEQEE
>P.cinereus_NH_PRF_A8
MRSTVLLIFLVVSVRRATSLTISAPVKRSTSSDVADLSETTIVLFSETQKFVESMKDNAD SLLTTYLSFQGAPLSDPEYHLPEVKIDNLPTATMDNATFLNQTDEIRLKNNLYFYSAIVQ LLQVAMTEQEDLNPAEESLKAKFQEALANSNTLISKISAIMTQMGMSATVTFPDPLVVPY EGSAIFKQKLRGGVICKEYNERVLLTKQDFEFFAKKYHEQLIQQQGQEE >P.cinereus_NH_PRF_A9
MRSTVLLIFLVVSVRRATSLTISAPVKRSTSSDVADLSETTIVLFSETQKFVESMKDNAD SLLTTYLSFQGAPLSDPEYHLPEVKIDNLPTATMDNATFLNQTDEIRLKNNLYFYSAIVQ FLQVAMTEQEDLNPAEESLKAKFQEALANSNTLISKISAIMTQMGMSATVTFPDPLVVPY EGSAIFKQKLRGGVICKEYKERVLLTKQDFEFFAKKYHEQLIQQQEQEE
>P.cinereus_NH_PRF_A10
MRSTVLLIFLVVSVRRATSLTISAPVKRSTSSDVADLSETTIVLFSETQKFVESMKDNAD SLLTTYLSFQGAPLSDPEYHLPEVKIDNLPTATMDNATFLNQTDEIRLKNNLYFYSAIVQ FLQVAMTEQEDLNPAEESLKAKFQEALANSNTLISKISAIMTQMGMSATVTFPDPLVVPY EGSAIFKQKLRGGVICKEYKERVLLTKQDFEFFAKKYHEQLIQQQEQEE
>P.cinereus_NH_PRF_B1
MRSTALLIFLVVSVRRVTSLTISAPVKRSTSSDVADLSETTIVLFSETQEFAESMKDNAD SLLTTYLSLQGEPYNDPDFSLPDIRIANLPTATMNYDTFMRQTDEIRLKNNLYFYSAIIK FLQLTMDEQKELNPADPSLNAKFEQVLSNINSLIDNISTIMTQMDFSVKIELRTPMVESY IGSAIFKKKLRGYVVCREYKERVALTMRDFEFLAKKYEGHLQRARP---
>P.cinereus_NH_PRF_B2
MRSTVLLIFLVVSVRRATSLTISAPVKRSTSSDVADLSETTIVLFSETQKFAESMKDNAD SLLTTYLSLQGEPYNNPDFSLRDIQIANLPTATMNYDTFMRQTDEIRLKNNLYFYSAIIK FLQLTMDEQKDLNPEDPSLNAKFEQVLSNINSLIHKISTIMTQMDFSVKIEFRTPMVESY IGSAIFKKKLRGYVVCREYKERVALTMWDFEFLAKKYEGHLHRARP---
>P.cinereus_NH_PRF_B3

MRSIVLLIFLVVSVRRATSLTISAPVKRSTSSDVADLSETTIVLFSETQKFAESMKDNAD SLLTTYLSLQGEPYNNPDFSLRDIQIANLPTATMNYDTFMRQTDEIRLKNNLYFYSAIIK FLQLTMDEQKDLNPEDPSLNAKFEQVLSNINSLIHKISTIMTQMDFSVKIEFRTPMVELY IGSAIFKKKLRGYVVCREYKERVALTMRDFEFLAKKYEGHLHRARP--
>P.cinereus_NH_PRF_B4
MRSTVLLIFLVVSVRRATSLTISAPVKRSTSSDVADLSETTIVLFSETQKFAESMKDNAD SLLTTYLSLQGEPYNNPDFSLRDIQIANLPTATMNYDTFMRQTDEIRLKNNLYFYSAIIK FLQLTMDEQKDLNPEDPSLNAKFEQVLSNINSLIHKISTIMTQMDFSVKIEFRTPMVESY IGSAIFKKKLRDYVVCREYKERVALTMRDFEFLAKKYEGHLHRARP--
>P.cinereus_NH_PRF_B5
MRSTVLLIFLVVSVRRATSLTISAPVKRSTSSDVADLSETTIVLFSETQKFAESMKDNAD SLLTTYLSLQGEPYNNPDFSLRDIQIANLPTAAMNYDTFMRQTDEIRLKNNLYFYSAIIK FLQLTMDEQKDLNPEDPSLNAKFEQVLSNINSLIHKISTIMTQMDFSVKIEFRTPMVESY IGSAIFKKKLRGYVVCREYKERVALTMRDFEFLAKKYEGHLHRARP--
>P.cinereus_NH_PRF_B6
MRSTVLLIFLVVSVRRATSLTISAPVKRSTSSDVADLSETTIVLFSETQKFAESMKDNAD SLLTTYLSLQGEPYNNPDFSLRDIQIANLPTATMNYDTFMRQTDEIRLKNNLYFYSAIIK FLQLTMDEQKDLNPEDPSLNAKFEQVLSNINSLIHKISTIMTQMDFSVKIEFRTPMVEPY IGSAIFKKKLRGYVVCREYKERVALTMRDFEFLAKKYEGHLHRARP---
>P.cinereus_NH_PRF_B7
MRSTVLLIFLVVSVRRATSLTISAPVKRSTSSDVADLSETTIVLFSETQKFAESMKDNAD
SLLTTYLSLQGEPYNNPDFSLRDIQIANLPTATMNYDTFMRQTDEIRLKNNLYFYSAIIK
FLQLTMDEQKDLNPEDPSLNAKFEQVLSNINSLIHKISTIMTQMDFSVKIEFRTPMVESY IGSAIFKKKLRGYVVCREYKERVALTMRDFEFLAKKYEGHLHRARP--
>P.cinereus_NH_PRF_B8
MRSTVLLIFLVVSVRRATSLTISAPVKRSTSSDVADLSETTIVLFSETQKFAESMKDNAD
SLLTTYLSLQGEPYNNPDFSLRDIQIANLPTATMNYDTFMRQTDEIRLKNNLYFYSAIIK
FLQLTMDEQKDLNPEDPSLNAKFEQVLSNINSLIHKISTIMTQMDFSVKIEFRTPMVESY IGSAIFKKKLRGYVVCREYKERVALTMRDFEFLAKKYEGHLHRARP-. -
>P.cinereus_NH_PRF_B9
MRSTVLLIFLVVSVRRATSLTISAPVKRSTSSDVADLSETTIVLFSETQKFAESMKDNAD SLPTTYLSLQGEPYNNPDFSLRDIQIANLPTATMNYDTFMRQTDEIRLKNNLYFYSTIIK FLQLTMDEQKDLNPEDPSLNAKFEQVLSNINSLIHKISTIMTQMDFSVKIEFRTPMVESY IGSAIFKKKLRGYVVCREQREGDTDDAGF--------------------
>P.cinereus_NH_PRF_B10
MRSTVLLIFLVVSVRRATSLTISAPVKRSTSSDVADLSETTIVLFSETQKFAESMKDNAD SLLTTYLSLQGEPYNNPDFSLRDIQIANLPTATMNYDTFMRQTDEIRLKNNLYFYSAIIK FLQLTMDEQKDLNPEDPSLNAKFEQVLSNINSLIHKISTIMTQMDFSVKIEFRTPMVESY IGSAIFKRSCGATSSAGNTKRGWH------------------------
>P. hoffmani_gi|42566671|_PRF_A1
MRSTALLIFLVVSVSRATSLTISAPVKRSTSSDVAELSETTIVLFSETQKFVESMKDSAD SLLTTYLSFQGAPLSDPEYHLPEVKVDNLPTAAMDNATFLNQTDEIRLKNNLYFYSAIVQ FLQVAMTEQEDLNPAEESLKAKFQEALANSNTLISKISAIMTQMGMSVTVTFPDPLVVPY EGSAIFKQKLRGGVICKEYKERVLLTKRDFEFFAKKYHEQLVQQQEE--
>P. hoffmani_gi|42566683|_PRF_A2
MRSTALLIFLVVSVSRATSLTISAPVKRSTSSDVADLSETTIVLFSETQKFVESMKDNAD SLLTTYLSFQGAPLSDPEYHLPEVKIDNLPTAAMDNATFFNQTDEIRLKNNLYFYSAIVQ FLQVAMTEQEDLNPAEESLKTKFQEAMANSNTLISKISAIMTQMGMSVTVTFPDPLVVPY EGSAIFKQKLRGGVICKEYKERVLLTKRDFEFFAQKYHEQLVQQQEQEE
>P.hoffmani_gi|42566719|_PRF_B1
MRSTALLIFLVVSVSRATSLTISAPVKRSTSSDVADLSETTIVLFSETQEFAESMKDNAD SLLTTYLSLQGEPYSDPDFGLPDIRIANLPTATMNYDTFMRQTDEIRLKNNLYFYTAIIK FLQLTMDEQKDLNPADPSLNAKFEQVLSNINALIDNISTIMTQMDFSVKIEFRTPMVESY IGSAIFKKKLRGYVVCREYKERVALTMRDFEFLAKKYEGHLQRARP---
>P.richmondi_gi|42566679|_PRF_A1
MRSTALLIFLVVSVSRATSLTISAPVKRSTSSDVAELSETTIVLFSETQKFVESMKDNAD SLLTTYLSFQGAPLSDPEYHLPEVKVDNLPTAAMDNATFLNQTDEIRLKNNLYFYSAIVQ FLQVAMTEQEDLNPAEESLKAKFQEAMANINTLISKISAIMTQMGMSVTVTFPDPLVVPY EGSAIFKQKLRGGVICKEYKERVLLTKRDFEFFAKKYHEQLIQQQEQEE
>P.richmondi_gi|42566677|_PRF_A2
MRSTALLIFLVVSVSRATSLTISAPVKRSTSSDVAELSETTIVLFSETQKFVESMKDNAD SLLTTYLSFQGAPLSDPEYHLPEVKVDNLPTAAMDNATFLNQTDEIRLKNNLYFYSAIVQ FLQVAMTEQEDLNPAEESLKAKFQEAMANINTLISKISAIMTQMGMSVTVTFPDPLVVPY EGSAIFKQKLRGGVICKEYKERVLLTKRDFEFFAKKYHEQLIQQQEQEE
>P.richmondi_gi|42566691|_PRF_A3
MRSTALLIFLVVSVSRATSLTISAPVKRSTSSDVAELSETTIVLFLETQKFAESMKDNAD SLLTTYLSFQGAPLSDPEYHLPEVKVDNLPTAAMDNATFLNQTDEIRLKNNLYFYSAIVQ FLQVTMTEQEDLNPAEESLKAKFQEAMANINTLISKISAIMTQMGMSVTVTFPDPLVVPY EGSAIFKQKLRGGVICQEYKERVLLTKRDFEFFAKKYHEQLVQQQEQEE
>P.richmondi_gi|42566673|_PRF_A4
MRSTALLIFLVVSVSRATSLTISAPVKRSTSSDVAELSETTIVLFLETQKFAESMKDNAD SLLTTYLSFQGAPLSDPEYHLPEVKVDNLPTAAMDNATFLNQTDEIRLKNNLYFYSAIVQ FLQVAMTEQEDLNPAEESLKAKFQEAMANINTLISKISAIMTQMGMSVTVTFPDPLVVPY EGSAIFKQKLRGGVICKEYKERVLLTKRDFEFFAKKYHEQLIQQQEQEE
>P.richmondi_gi|42566687|_PRF_A5
MRATALLIFLVVSVSRATSLTISAPVKRSTSSDMAELSETTIVLFLETQKFAESMKDNAD SLLTTYLSFQGAPLSDPEYHLPEVKVDNLPTAAMDNATFLNQTDEIRLKNNLYFYSAIVQ FLQVAMTEQEDLNPAEESLKAKFQEAMANINTLISKISAIMTQMGMSVTVTFPDPLVVPY EGSAIFKQKLRGGVICKEYKERVLLTKRDFEFFAKKYHEQLIQQQEQEE
>P.richmondi_gi|60499234|_PRF_A8
MRSTALLIFLVVSVSRATSLTISAPVKGSTSSDVAELSETTIVLFSETQKFVESMKDNAD SLLTIYLSFQGAPLSDPEYHLPEVKVDNLPTAAMDNATFLNQTDEIRLKNNLYFYSAIVQ FLQVAMTEQEDLNPAEESLKAKFQEAMANINTLISKISAIMTQMGMSVTVTFPDPLVVPY EGSAIFKQKLRGGVICKEYKERVLLTKRDFEFFAKKYHEQLIQQQEQEE
>P.richmondi_gi|60499236|_PRF_A9
MRSTALLIFLVVSVSRATSLTISAPVKRSTSSDVAELSETTIVLFSETQKFVESMKDNAD SLLTIYLSFQGAPLSDPEYHLPEVKVDNLPTAAMDNATFLNQTDEIRLKNNLYFYSAIVQ FLQVAMTEQEDLNPAEESLKAKFQEAMANINTLISKISAIMTQMGMSVTVTFPDPLVVPY EGSAIFKQKLRGGVICKEYKERVLLTKRDFEFFAKKYHEQLIQQQEQEE
>P.richmondi_gi|60499238|_PRF_A10
MRSTALLIFLVVSVSRATSLTISAPVKRSTSSDVAELSETTIVLFSETQKFVESMKDNAD SLLTTYLSFQGAPLSDPEYHLPEVKVDNLPTAAMDNATFLNQTDEIRLKNNLYFYSAIVQ FLQVAMTEQEDLNPAEESLKAKFQEAMANINTLISKISAIMTQMGMSVTVTFPDPLVVPY EGSAIFKQKLRGGVICKEYKERVLLTKRDFEFFAKKYHEQLIQQQEQEE
>P.richmondi_gi|42566715|_PRF_B1
MRSTALLIFLVVSVSRATSLTISTPVKRSTSSDVAELSATTIVLFSETQKFAESMKDNAD SLLTTYLSLQGEPYSNPDFSFLDIQIANLPTAAMNYDTFMRQTDEIRLKNNLYFYSAIIK FLQLTMDEQKDLNPADQSITAKFEQVLSNINALIDNISTIMTQMDFSVKIEFRTPMVESY IGSAIFRKKLRGYIVCREYKERVALTMRDFEFLAKKYEGHLQRARP---
>P.richmondi_gi|42566713|_PRF_B2
MRSTALLIFLVVSVSRATSLTISTPVKRSTSSDVAELSATTIVLFSETQKFAESMKDNAD SLLTTYLSLQGEPYSNPDFSFLAIQIANLPTAAMNYDTFMRQTDEIRLKNNLYFYSAIIK FLQLTMDEQKDLNPADQSITAKFEQVLSNINALIDNISTIMTQMDFSVKIEFRTPMVESY IGSAIFRKKLRGYIVCREYKERVALTMRDFEFLAKKYEGHLQRARP--
>P.richmondi_gi|42566701|_PRF_B3
MRSTALLIFLVVSVSRATSLTISAPVKRSTSSDVAELSETTIVLFLETQKFAESMKDNAD SLLTTYLSLQGEPYSNPDFSFLAIQIANLPTAAMNYDTFMRQTDEIRLKNNLYFYSAIIK FLQLTMDEQKDLNPADQSITAKFEQVLSNINALIDNISTIMTQMDFSVKIEFRTPMVESY IGSAIFRKKLRGYIVCREYKERVALTMRDFEFFAKKYHEQLVQQQEQEE
>P.richmondi_gi|42566717|_PRF_B4
MRSTALLIFLVVSVSRATSLTISTPVKRSTSSDVAELSATTIVLFSETQKFAESMKDNAD SLLTTYLSLQGEPYSNPDFSFLAIQIANLPTAAMNYDTFMRQTDEIRLKNNLYFYSAIIK FLQLTMDEQKDLNPADQSITAKFEQVLSNINALIDNISTIMTQMDLSVKIEFRTPMVESY IGSAIFRKKLRGYIVCREYKERVALTMRDFEFLAKKYEGHLQRARP--
>P.richmondi_gi|60499232|_PRF_B5
MRSTALLIFLVVSVRRATSLTISAPVKRSTSSDVADLSETTIVLFSETQKFAESMKDNAD SLLTTYLSLQGEPYNNPDFSLRDIQIANLPTATMNYDTFMRQTDEIRLKNNLYFYSAIIK FLQLTMDEQKDLNPEDPSLNAKFEQVLSNINSLINKISTIMTQMDFSVKIEFRSPMVESY IGSAIFKKKLRGYVVCREYKERVALTMRDFEFLAKKYEGHLHRARP - -
>P.websteri_gi|60499262|_PRF_A1
MRSTALLTFLVVFVSTATSMAMADT------ DVAELSKETIVLFSETQKIVEDMQSSAA
SLLPTYLSFQGVPLSNPEYQLPNIKIENLPTAAMDNDTFLNHTDEIRLKNNLYFYSAIVE FLKTTMAEQEDLNP--QSLTAKFEEAMANSNTLISKISAIMTQMGLSVTITFPEPLVVPY VGSAFFKKKLRGGVICKEYKERVLLTKQEFEFFAKKYHEQLVQPQEE--
$>P$.websteri_gi|60499264|_PRF_A2
MRSTALLTFLVVFLSTATSLAMADT------DVAELSKETIVLFSETQKIVEDMQSSAA SLLPTYLSFQGVPLSNPEYQLPNIKIENLPTAAMDNDTFLNHTDEIRLKNNLYFYSAIVE FLKTTMAEQEDLNP--QSLTAKFEEAMANSNTLISKISAIMTQMGLSVTITFPEPLVVPY VGSAFFKKKLRGGVICKEYKERVLLTKREFEFFAKKYHEQLVQPQEE--
>P.wehrlei_gi|60499240|_PRF_A1
MRSTALLTFLVVFVSTAIDLTMADTDEHSTSSDVAELSKDTIILFSETQKIVENIQSNAG SLLATYLSFQGAPLSNPEYHLPDVKIDNLPTAAMDNDTFLSQTDETRLKNNLYFYSAIVE FLQIVKTEQEDLNPFALDLTAKLEEAVANSTTLSNKISAIMTQMGMSVTIEFPKPLVVPY EGSSYFRKKLRGGIICKEYKDRVFLTKRDFEFLAKKYQGQLIQQQEE--
>P.wehrlei_gi|60499242|_PRF_A2
MRSTALLTFLVVFVSTAIDLTMADTDEHSTSSDVAELSKDTIILFSETQKIVENIQSNAG SLLATYLSFQGAPLSNPEYHLPDVKIDNLPTAAMDNDTFLSQTDETRLKNNLYFYSAIVE FLQIVKTEQEDLNPFALDLTAKLEEAVANSTTLSNKISAIMTQMGMSVTIEFPKPLVVPY EGSSYFRKKLRGGVICKEYKDRVFLTKRDFEFLAKKYQGQLIQQQEE--
>P.wehrlei_gi|60499244|_PRF_A3
MRSTALLTFLVVFVSTAIGLTMADTDEHSTSSDVAELSKDTIILFSETQKIVENIQSNAG SLLATYLSFQGAPLSNPEYHLPDVKIDNLPTAAMDNDTFLSQTDETRLKNNLYFYSAIVE FLQIVKTEQEDLNPFALDLTAKLEEAVANSTTLSNKISAIMTQMGMSVTIEFPKPLVVPY EGSSYFRKKLRGGVICKEYKDRVFLTKRDFEFLAKKYQGQLIQQQEE--
>P.wehrlei_gi|60499246|_PRF_A4
MRSTALLTFL-VFVSTAIGLTMADTDEHSTSSDVAELSKDTIILFSETQNIVENIQSNAG SLLTTYLSFQGAPLSNPDYHLPDVKIDNLPTAAMDNDTFLSQTDETRLKNNLYFYSAIVE FLQIVKTEQEDLNPFALDLTAKLEEAVANSTTLSNKISAIMTQMGMSVTIEFPKPLVVPY EGSSYFRKKLRGGVICKEYKDRVFLTKRDFEFLAKKYQGQLIQQQEE--
>P.wehrlei_gi|60499248|_PRF_A5
MRSTALLTFL-VFVSTAIGLTMADTDEHSTSSDVAELSKDTIILFSETQKIVENIQSNAG SLLTTYLSFQGAPLSNPDYHLPDVKIDNLPTAAMDNDTFLSQTDETRLKNNLYFYSAIVE FLQIVKTEQEDLNPFALDLTAKLEEAVANSTTLSNKISAIMTQMGMSVTIEFPKPLVVPY EGSSYFRKKLRGGVICKEYKDRVFLTKRDFEFLAKKYQGQLIQQQEE--
>P.welleri_gi|60499250|_PRF_A1
MRSTALLTFLVLSI--ATGLTMADTDEHS-----ELSTDTIALFSETQKIAEDMQSNAD SLPTTYLSFQGPPLSHPEYHLPDIKIENLPTAAMDYDTFLSQTDETRLKNNLYFYSAIVE FLQIVKTEQEDLNPFALDLTAKLEEAVANSNTLISKISAIMTHMGMSVTIEFPDPLVVPF KGTAYFQKKLRGGVICKEFKDRVDLTKRDFEFFAKKYEGHL------
>P.welleri_gi|60499252|_PRF_A2
MRSTALLTFLVVSI--ATGLTMADTDEHS-----ELSTDTIALFSETQKIAEDMQSNAD SLPTTYLSFQGPPLSHPEYHLPDIKIENLPTAAMDYDTFLSQTDETRLKNNLYFYSAIVE FLQIVKTEQEDLNPFALDLTAKLEEAVANSNTLISKISAIMTQMGMSVTIEFPDPLVVPF KGTAYFQKKLRGGVICKEFKERVDLTKRDFEFFAKKYEGRL--.......
>P.welleri_gi|60499254|_PRF_A3
MRSTALLTFLVLSI--ATGLTMADTDEHS------ELSTDTIALFSETQKIAEDMQSNAD SLPTTYLSFQGPPLSHPEYHLPDIKIENLPTAAMDYDTFLSQTDETRLKNNLYFYSAIVE FLQIVKTEQEDLNPFALDLTAKLEEAVANSNTLISKISAIMTQMGMSVTIEFPDPLVVPF

KGTSYFQKKLRGGVICKEFKERVDLTKRDFEFFAKKYEGRL------- -
$>$ P.dorsalis_gi|60499260|_PRF_A1
MRSTALLTFLLVSI--ATGLTMADTDEHS------ELSTDTIALFSEIQKIAEDMQSNAD SLPTTYLSFQGPPLSNADYHLPDIKIENLPTAAMDYDTFLSQTDETRLKNNLYFYSAIVE FLQIVKTEQEDLNPFALDLTAKLEEAVANSNTLISKISAIMTQMGMSETIEFPDPLVVPF KGTAYFQKKLRGGVICKEFKERVDLTKRDFEFFAKKYEGHL
>P.dorsalis_gi|60499258|_PRF_A2
MRSTALLTFLLVSI--ATGLTMADTDEHS------ELSTDTIALFSEIQKIAEDMQSNAD SLPTTYLSFQGPPLSNADYHLPDIKIENLPTAAMDYDTFLSQTDETCLKNNLYFYSAIVE FLQIVKTEQEDLNPFALDLTAKLEEAVANSNTLISKISAIMTQMGMSETIEFPDPLVVPF KGTAYFQKKLRGGVICKEFKERVDLTKRDFEFFAKKYEGHL
>P.ventralis_gi|60499256|_PRF_A1
MRSTALLTFLVVSI--ATGLTMADTDEHS------ELSTDTIALFSETQKIAEDMQSNAD SLPTTYLSFQGPPLSNADYHLPDIKIENLPTAAMDYDTFLSQTDETRLKNNLYFYSAIVE FLQIVKTEQEDLNPFALDLTAKLEEAVANSNTLISKISAIMTQMGMSVTIEFPDPLVVPF KGTAYFQKKLRGGVICKEFKERVDLTKRDFEFFAKKYEGHL
>P.yonahlossee_gi|42566629|_PRF_A1
MRSTSLLTFLLVSLSTATSLAMAE-----INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLNNNLYFYSAIVE FLKVAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPPLVVPF KGSAFFSKKLRGGVVCKEYKERVLLTKRDFELLAKKYQGPL--------
>P.yonahlossee_gi|42566623|_PRF_A2
MRSTSLLTFLLVSLSTATSLAMAE-----INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLNNNLYFYSAIVE FLKVAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAFFSKKLRGGVVCKEYKERVLLTKRDFELLAKKYQGPL--------
>P.yonahlossee_gi|42566615|_PRF_A3
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLNNNLYFYSAIVE FLKVAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAFFSKKLRGGVVCKEYKERVLLTKRDFELLAKKYQGPL------ -
>P.yonahlossee_gi|42566647|_PRF_A4
MRSTSLLTFLLVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLNNNLYFYSAIVE FLKVAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAFFSKKLRGGVVCKEYKERVLLTKQDFELLAKKYQGPL-------
>P.kentucki_gi|42566665|_PRF_A1
MRSTSLLTFLVVSLSTASSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAALDLKAKFEEAMANSNTLISKISDIMTQMGMSVTVTLPKPLVVPY KGSAYFSKKLRGGVVCKEYKERVFLTKRDFELLAKKYQGPF-------
>P. kentucki_gi|42566631|_PRF_A2
MRSTSLLTFLVVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD ALLPTYLSFQGAPLSDPDYRLPHIKVVNLPTAAMDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFELLAKKYQGPF--.-.-. -
>P.kentucki_gi|42566659|_PRF_A3
MRSTSLLTFLVVSLSTASSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPPIKVVNLPTAAMDYDTFMKQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKKPLVVPY EGSAYFRKKLRGGVVCKEYKERVLLTKRDFEFLAKKYQGHL
>P.ouachitae_gi|60499362|_PRF_A1
MRSTSLLTFLVVSLSTATSLAMAE-----TNDVAELSSDTIVLFSEAQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVANLPTAAMDYDTFMSQNDETRLNNNLYFYSAIVE FLKVAMTEQEDLNPAELSLKAKFEEAMANTNTLISKISDIMTQMGMSVTITLPKPLVVPY KGSAIFGKKLRGGVVCKEYKEMVLLTKRDFEFLAKKYQGHL
>P. ouachitae_gi|60499364|_PRF_A2
MRSTALLTFLVVSLSTATSRTMAE------TNDVAELSSDTIVLFSETQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVANLPTAAMDYDTFMSQNDETRLNNNLYFYSAIVE FLKVAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPY KGSAIFRKKLRGGVVCKEYKEMVLLTKRDFEFLAKKYQGHL
>P. ouachitae_gi|60499366|_PRF_A3
MRSTALLTFLVVSLSTATSLAMAE------TNDVAELSSDTIVLFSEAQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVANLPTAAMDYDTFMSQNDETRLNNNLYFYSAIVE FLKVAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPY KGSAIFRKKLRGGVVCKEYKEMVLLTKRDFEFLAKKYQGHL
>P.ouachitae_gi|60499368|_PRF_A4
MRSTALLTFLVVSLSTATSLAMAE------TNDVAELSSDTIVLFSETQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVANLPTAAMDYDTFMSQNDETRLNNNLYFYSAIVE FLKVAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPY KGSAIFRKKLRGGVVCKEYKEMVLLTKRDFEFLAKKYQGHL--------
>P.ouachitae_gi|60499370|_PRF_A5
MRSTSLLTFLVVSLSTATSLAMAE------TNDVAELSSDTIVLFSEAQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVANLPTAAMDYDTFMSQNDETRLNNNLYFYSAIVE FLKVAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPY KGSAIFRKKLRGGVVCKEYKEMVLLTKRDFEFLAKKYQGHL-------
>P.fourchensis_gi|60499372|_PRF_A1
MRSTSLLTFLVVSLSTATSLAMAE------TNDVAELSSDTIVLFSETQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVANLPTAAMDYDTFMSQNDETRLNNNLYFYSAIVE FLKVAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPY KGSAIFRKKLRGGVVCKEYKEMVLLTKRDFEFLAKKYQGHL------ --
>P.fourchensis_gi|60499374|_PRF_A2
MRSTSLLTFLVVSLSTATSLAMAE------TNDVAELSSDTIVLFSETQKFAEDIQSSAD
SLLPTYLSFQGAPLSDPDYQLPHIKVANLPTAAMDYDTFMSQNDETRLNNNLYFYSAIVE

FLKVAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPY KGSAIFRKKLRGGVVCKEYKEMVLLTKRDFEFLAKKYQGHL--------
>P.aureolus_gi|42566605|_PRF_A1
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD
SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFMLLAEKYQGPL--------
>P.aureolus_gi|42566621|_PRF_A2
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELTLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFVLLAQKYQGPL
>P.aureolus_gi|42566627|_PRF_A3
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFMLLAEKYQGPL
>P.aureolus_gi|42566625|_PRF_A4
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFVLLAQKYQGPL
>P.aureolus_gi|60499540|_PRF_A5
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELTLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFVLLAQKYQGPL
>P.mississippi_gi|60499530|_PRF_A1
MRSTSLLTFLVVSLSTATSLAMAE-----INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITFPKPLVVPF KGSAFFSKKLRGGVVCKEYKERVFLTKRDFELLAKKYQGHL--------
>P.mississippi_gi|60499532|_PRF_A2
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVVNLPTAAMDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITFPKPLVVPF KGSAFFSKKLRGGVVCKEYKERVFLTKRDFEFLAKKYQGHL-------
>P.mississippi_gi|60499534|_PRF_A3
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITFPKPLVVPF KGSAFFSKKLRGGVVCKEYKERVFLTKRDFELLAKKYQGHL--------
>P.mississippi_gi|60499536|_PRF_A4
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITFPKPLVVPF KGSAFFSKKLRGGVVCKEYKERVFLTKRDFEFLAKKYQGHL------ -
>P.mississippi_gi|60499538|_PRF_A5
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPPLVVPY EGSAYFRKKLRGGVVCKEYKERVFLTKRDFEFLAKKYQGHL-------
>P.grobmani_gi|42566651|_PRF_A1
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPPIKVVNLPTAAMDYDTFMRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPY EGSAYFSKKLRGGVVCKEYKERVFLTKRDFELLAKKYQGHL----- --
>P.grobmani_gi|42566641|_PRF_A2
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPPIKVVNLPTAAMDYDTFMRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPY EGSAYFRKKLRGGVVCKEYKERVFLTKRDFEFLAKKYQGHL
>P.savannah_gi|60499376|_PRF_A1
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFMRQTDETRLNNNLYFYSAIVE FLKVAMTEQEDLNPAALTLKAKFEEAMANSNTLISKISDIMTQMGMSVTITFPKPLVVPF KGSAFFSKKLRGGVVCKEYKERVFLTKRDFELLAKKYQGHL
>P.savannah_gi|60499378|_PRF_A2
MRSTSLLTFLVVSLSTATSLAMAE-----INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLNNNLYFYSAIVE FLKVAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITFPKPLVVPY EGSAYFSKKLRGSVVCKEYKERVFLTKRDFELLAKKYQGHL
>P.savannah_gi|60499380|_PRF_A3
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFMRQTDETRLNNNLYFYSAIVE FLKVAMTEQEDLNPAALTLKAKFEEAMANSNTLISKISDIMTQMGMSVTITFPKPLVVPF KGSAFFSKKLRGGVVCKEYKERVFLTKRDFELLAKKYQGHL
>P.savannah_gi|60499382|_PRF_A4
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLNNNLYFYSAIVE FLKVAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITFPKPLVVPY EGSAYFSKKLRGGVVCKEYKERVFLTKRDFELLAKKYQGHL--------
>P.savannah_gi|60499384|_PRF_A5
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD

SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLNNNLYFYSAIVE FLKVAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITFPKPLVVPY EGSAYFSKKLRGGVVCKEYKERVFLTKRDFELLAKKYQGHL------
>P.cheoah_gi|42566633|_PRF_A1
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVRFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVANLPTAAMDYDTFMKQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF EGSAYFRKKLRGGVVCKEYKERVLLTKRDFEFLAKKYQGHL--------
>P.cheoah_gi|60499386|_PRF_A1
MRSTSLLTFLLVSLSTATSLAMAE------INDVADLSSDTIVRFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTAAMDYDTFMKQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSATITLPKPLVVPF EGSAYFRKKLRGGVVCKEYKERVLLTKRDFEFLAKKYQGHL-------
>P.cheoah_gi|42566613|_PRF_A2
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD
SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTAAMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFELLAKKYQGPF-------
>P.cheoah_gi|60499388|_PRF_A2
MRSTSLLTFLVVSLSTATSLAMAE------TNDVADLSSDTIVRFSEAQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVANLPTAAMDYDTFMKQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPPLVVPF EGSAYFRKKLRGGVVCKEYKERVLLTKRDFELLAKKYYQGHL
>P.cheoah_gi|42566653|_PRF_A3
MRSTSLLTFLLVSLSTATSLAMAE------INDVADLSSDTIVRFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVANLPTAAMDYDTFMKQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF EGSAYFRKKLRGGVVCKEYKERVLLTKRDFEFLAKKYQGHL
>P.cheoah_gi|60499390|_PRF_A3
MRSTSLLTFLVVSLSTATSLAMAE------TNDVADLSSDTIVRFSEAQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVANLPTAAMDYDTFMKQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF EGSAYFRKKLRGGVVCKEYKERVLLTKRDFEFLAKKYQGHL
>P.cheoah_gi|60499392|_PRF_A4
MRSTSLLTFLVVSLSTATSLAMAE------TNDVAELSSDTIVRFSEAQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVANLPTAAMDYDTFMKQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF EGSAYFRKKLRGGVVCKEYKERVLLTKRDFEFLAKKYQGHL-------
>P.cheoah_gi|60499394|_PRF_A5
MRSTSLLTFLVVSLSTATSLAMAE------TNDVADLSSDTIVRFSEAQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVANLPTAAMDYDTFMKQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF EGSAYFRKKLRGGVVCKEYKERVLLTKRDFEFLAKKYQGHL-------
>P.cheoah_gi|60499396|_PRF_A6
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVRFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVANLPTAAMDYDTFMKQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF EGSAYFRKKLRGGVVCKEYKERVLLTKRDFEFLAKKYQGHL--------
>P. Cheoah_gi|60499398|_PRF_A7
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVANLPTAAMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF EGSAYFRKKLRGGVVCKEYKERVLLTKRDFEFLAKKYQGHL-------
>P.cheoah_gi|60499400|_PRF_A8
MRSTSLLTFLLVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFVLLAQKYQGPF--------
>P.cheoah_gi|60499402|_PRF_A9
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTAAMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFELLAKKYQGPL------
>P.cheoah_gi|60499404|_PRF_A10
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD
SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTAAMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFELLAKKYQGPF--------
>P.chattahoochee_gi|42566607|_PRF_A1
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFMLLAEKYQGPL
>P.chattahoochee_gi|60499480|_PRF_A2
MRSTSLLTFLVVSLSTATSLAMAE-----INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTANMDYDAFIRQTDETRLKNNLYFCSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKECKERVFLTKRDFMLLAEKYQGPL--------
>P.chattahoochee_gi|60499482|_PRF_A3
MRSTSLLTFLVVSLSTATSLAMAE-----INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPAAAMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFMLLAEKYQGPL--------
>P.chattahoochee_gi|60499484|_PRF_A4

MRSTSLLTFLVVSLSTATSLAMAE------ INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKECKERVFLTKRDFMLLAEKYQGPL--------
>P.chattahoochee_gi|60499486|_PRF_A5
MRPTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTANMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFMLLAEKYQGPL-------
>P.chattahoochee_gi|60499488|_PRF_A6
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYPSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFMLLAEKYQGPL------ -
>P.chattahoochee_gi|60499490|_PRF_A7
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTANMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKKLRGGVVCKEYKERVFLTRRDFMLLAEKYQGPL--------
>P.chattahoochee_gi|60499492|_PRF_A8
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLKNNLYFYSAIAE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISGIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFMLLAEKYQGPL-------
>P.cylindraceus_gi|60499314|_PRF_A1
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELTLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFVLLAKKYQGPL-------
>P.cylindraceus_gi|60499316|_PRF_A2
MRSTSLLTFLVVSLSTATSLAMAE-----INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELTLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFELLAKKYQGPL--------
>P.cylindraceus_gi|60499318|_PRF_A3
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELTLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFVLLAKKYQGPL
>P.cylindraceus_gi|60499320|_PRF_A4
MRSTSLLTFLVVSLSTATSLAMAE-----INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELTLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFELLAKKYQGPL
>P.cylindraceus_gi|60499322|_PRF_A5
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFELLAKKYQGPL-------
>P.teyahalee_gi|60499324|_PRF_A1
MRSTSLLTFLLVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFVLLAQKYQGPF
>P.teyahalee_gi|60499336|_PRF_A1
MRSTSLLTFLLVSLSTATSLAMAD-----INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFVLLAQKYQGPF-------
>P.teyahalee_gi|60499326|_PRF_A2
MRSTSLLTFLLVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFVLLAQKYQGPF------ -
>P.teyahalee_gi|60499328|_PRF_A3
MRSTSLLTFLLVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFVLLAQKYQGPF-------
>P.teyahalee_gi|60499340|_PRF_A3
MRSTSLLTFLVVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFVLLAQKYQGPF------->P.teyahalee_gi|60499330|_PRF_A4
MRSTSLLTFLVVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKKLRGGVVCKEYKERVFLTKRDFVLLLAQKYQGPF
>P.teyahalee_gi|60499332|_PRF_A5
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFVLLAQKYQGPF
>P.teyahalee_gi|60499334|_PRF_A6
MRSTSLLTFLVVSLSTATSLAMAE-----INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFMLLAEKYQGPL
>P.shermani_gi|42566635|_PRF_A1
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVRFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVANLPTAAMDYDTFMKQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF EGSAYFRKKLRGGVVCKEYKERVLLTKRDFEFLAKKYQGHL-------
>P.shermani_gi|60499438|_PRF_A1
MRSTSLLTFLVVSLSTATSLAMAE----- INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFMLLAEKYQGPF -
>P.shermani_gi|42566643|_PRF_A2
MRSTSLLTFLVVSLSTATSLAMAE----- INDVAGLSSDTIVRFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVANLPTAAMDYDTFMKQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF EGSAYFRKKLRGGVVCKEYKERVLLTKRDFEFLAKKYQGHL-----
>P.shermani_gi|60499440|_PRF_A2
MRSTSLLTFLVVSLSTATSLAMAE----- INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELGLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFMLLAEKYQGPF--.-. --
>P.shermani_gi|42566609|_PRF_A3
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFMLLAEKYQGPL------
>P.shermani_gi|60499442|_PRF_A3
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTANMDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELGLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFMLLAEKYQGPF - $-\ldots-$
$>P$.shermani_gi|42566619|_PRF_A4
MRSTSLLTFLVVSLSTATSLAMAD----- INDVADLSSDTIVLFSEVQKFAEDIQSSAD
SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTAAMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFMLLAEKYQGPL
>P.shermani_gi|60499444|_PRF_A4
MRSTSLLTFLVVSLSTATSLAMAD-----INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTAAMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFMLLAEKYQGPL
>P.shermani_gi|42566657|_PRF_A5
MRSTSLLTFLVVSLSTATSLAMAE-----INDVADLSSDTIVRFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVANLPTAAMDYDTFMKQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSETITLPKPLVVPF EGSAYFRKKLRGGVVCKEYKERVLLTKRDFEFLAKKYQGHL
>P.shermani_gi|60499446|_PRF_A5
MRSTSLLTFLVVSLSTATSLAMAE-----INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTANMDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFMLLAEKYQGPL------
>P.shermani_gi|42566649|_PRF_A6
MRSTSLLTFLVVSLSTATSLAMAE-----INDVADLSSDTIVRFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVANLPTAAMDYDTFMKQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAALSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF EGSAYFRKKLRGGVVCKEYKERVLLTKRDFEFLAKKYQGHL-------
>P.shermani_gi|60499448|_PRF_A6
MRSTSLLTFLVVSLSTATSLAMAE----- INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTANMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELGLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFMLLAQKYQGPL------
>P.shermani_gi|60499450|_PRF_A7
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTANIDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELGLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFMLLAQKYQGPF------
>P.shermani_gi|42566611|_PRF_A7
MRSTSLLTFLVVSLSTATSLAMAE----- INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVICKEYKERVFLTKRDFMLLAEKYQGPL-------
>P.shermani_gi|42566617|_PRF_A8
MRSTSLLTFLVVSLSTATSLAMAD----- INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTAAMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFMLLAEKYQGPF---- -
>P.shermani_gi|60499452|_PRF_A8
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTANMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF

KGSAYFSKKLRGGVVCKEYKERVFLTKRDFMLLAEKYQGPL--------
>P.shermani_gi|42566645|_PRF_A9
MRSTSLLTFLVVSLSTATSLAMAD----- INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTANIDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELGLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFMLLAQKYQGPF
>P.shermani_gi|60499454|_PRF_A9
MRSTSLLTFLVVSLSTATSLAMAD-----INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTANMDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELGLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFMLLAEKYQGPF
>P.shermani_gi|42566661|_PRF_A10
MRSTSLLTFLVVSLSTATSLAMAD-----INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTANIDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELGLKAKFEEAMANSNTLISKIYDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFMLLAQKYQGPF
>P.shermani_gi|60499456|_PRF_A10
MRSTSLLTFLVVSLSTATSLAMAD-----INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTANMDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELGLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFMLLAEKYQGPL-------
>P.shermani_gi|42566663|_PRF_A11
MRSTSLLTFLVVSLSTATSLAMAD----- INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTANIDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELGLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFMLLAQKYQGPF
>P.shermani_gi|60499458|_PRF_A11
MRSTSLLTFLVVSLSTATSLAMAD----- INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTANMDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELGLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFMLLAQKYQGPF----- -
>P.shermani_gi|42566667|_PRF_A12
MRSTSLLTFLVVSLSTATSLAMAE----- INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVICKEYKERVFLTKRDFMLLAEKYQGPF----- -
>P.shermani_gi|60499460|_PRF_A12
MRSTSLLTFLVVSLSTATSLAMAD----- INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTANIDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELGLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFMLLAQKYQGPF $-\ldots-{ }^{-}$
>P.shermani_gi|60499470|_PRF_A12
MRSTSLLTFLVVSLSTATSLAMAE----- INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELTLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFMLLAEKYQGPL-.........
>P.shermani_gi|60499472|_PRF_A13
MRSTSLLTFLVVSLSTATSLAMAE----- INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELTLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFMLLAEKYQGPL
>P.shermani_gi|60499464|_PRF_A14
MRSTSLLTFLVVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTANIDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELGLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFMLLAEKYQGPL
>P.shermani_gi|60499474|_PRF_A14
MRSTSLLTFLVVSLSTATSLAMAE-----INDVADLSSDTIVRFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVANLPTAAMDYDTFMKQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF EGSAYFRKKLRGGVVCKEYKERVLLTKRDFEFLAKKYQGHL
>P.shermani_gi|60499466|_PRF_A15
MRSTSLLTFLVVSLSTATSLAMAD-----INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTANIDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELGLKAKFEEAMANSNTLISKISDIMTQIGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFMLLAQKYQGPF
>P.shermani_gi|60499476|_PRF_A15
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVRFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFMLLAEKYQGPF -
>P.shermani_gi|60499478|_PRF_A16
MRSTSLLTFLLVSLSTATSLAMAD----- INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFVLLAQKYQGPF----- -
>P.jordani_gi|60499308|_PRF_A1
MRSTSLLTFLVVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTAAVDYDIFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKSLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFELLSKKYQGPF----- -
>P.jordani_gi|60499310|_PRF_A2
MRSTSLLTFLVVSLSTATSLAMAD----- INDVADLSSDTIVLFSEVQKFAEDIQSSAD
SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTAAVDYDIFIRQTDETRLKNNLYFYSAIVE

FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFELLSKKYQGPF-------
>P.jordani_gi|60499312|_PRF_A3
MRSTSLLTFLVVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTAAVDYDIFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFELLSKKYQGPF--------
>P.meridianus_gi|42566639|_PRF_A1
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTANMDYDTFIRQTDEARLNNNLYFYSAIVE FLKEAMTEQEDLNPAELGLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFRKKLRGGVVCKEYKERVFLTKRDFVLLTNKYQGPL
>P.meridianus_gi|42566637|_PRF_A2
MRSTSLLTFLVVSLSTATSLAMAE-----INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTANMDYDTFIRQTDEARLNNNLYFYSAIVE FLKEAMTEQEDLNPAELGLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPPLVVPF KGSAYFRKKKLRGGVVCKEYKERVFLTKRDFVLLANKYQGPL
>P.meridianus_gi|60499494|_PRF_A3
MRSTSLLTFLVVSLSTATSLAMAE-----INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTASMDYDTFIRQTDEARLNNNLYFYSAIVE FLKEAMTEQEDLNPAELGLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFRKKLRGGVVCKEYKERVFLTKRDFVLLANKYQGPL
>P.amplus_gi|60499344|_PRF_A1
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLNNNPYFYSAIVE FLKEAMTEQEDLNPAELGLKAKFEEAMANSNTLTSKISDIMTQMGMSVTITLPKPLVVPF KGSAYFRKKLRGGVVCKEYKERVFLTKRDFELLANKYQGPL-------
>P.amplus_gi|60499346|_PRF_A2
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELGLKAKFEEAMANSNTLTIKISDIMTQMGMSVTITLPKPLVVPF KGSAYFRKKLRGGVVCKEYKERVFLTKRDFELLANKYQGPL--------
>P.amplus_gi|60499348|_PRF_A3
MRSTSLLTFLVVSLSTATSLAMAE----- INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAVDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELGLKAKFEEAMANSNTLTSKISDIMTQMGMSVTITLPKPLVVPF KGSAYFRKKLRGGVVCKEYKERVFLTKRDFELLANKYQGPL--------
>P.amplus_gi|60499350|_PRF_A4
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLSNNLYFYSAIVE FLKEAMTEQEDLNPAELGLKAKFEEAMANSNTLTSKISDIMTQMGMSVTITLPKPLVVPF KGSAYFRKKLRGGVVCKEYKERVFLTKRDFELLANKYQGPL------ -
>P.amplus_gi|60499352|_PRF_A5
MRSISLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAVDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELGLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFRKKLRGGVVCKEYKERVFLTKRDFELLANKYQGPL--------
>P.amplus_gi|60499354|_PRF_A6
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELGLKAKFEEAMANSNTLTSKISDIMTQMGMSVTITLPKPLVVPF KGSAYFRKKLRGGVVCKEYKERVFLTKRDFELLANKYQGPL-------
>P.amplus_gi|60499356|_PRF_A7
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPIAAMDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELGLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFRKKLRGGVVCKEYKERVFLTKRDFELLANKYQGPL--------
>P.amplus_gi|60499358|_PRF_A8
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAVDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELGLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFRKKLRGGVVCKEYKERVFLTKRDFELLANKYQGPL
>P.amplus_gi|60499360|_PRF_A9
MRSTSLLTFLVVSLSTATSLAMAE------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLSNNLYFYSAIVE FLKEAMTEQEDLNPAELGLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVVPF KGSAYFRKKKLRGGVVCKEYKERVFLTKRDFELLANKYQGPL
>P.metcalfi_gi|60499266|_PRF_A1
MRSTSLLTFLVVSLSTATSLAMAE------TNDVADLSRDTILLFSETQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTANIDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFELLAKKYQGPF
>P.metcalfi_gi|60499300|_PRF_A1
MRSTSLLTFLVVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTANIDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELGLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFELLAQKYQGPF
>P.metcalfi_gi|60499268|_PRF_A2
MRSTSLLTFLVVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTAAVDYDIFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFELLSKKYQGPF-
>P.metcalfi_gi|60499302|_PRF_A2
MRSTSLLTFLVVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD

SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTANMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELGLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFELLAQKYQGPF--------
>P.metcalfi_gi|60499270|_PRF_A3
MRSTSLLTFLVVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLKNNLYFYSAIVE FLKVAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFELLAKKYQGPF--------
>P.metcalfi_gi|60499304|_PRF_A3
MRSTSLLTFLVVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTANMDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELGLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFELLAQKYQGPF------
>P.metcalfi_gi|60499272|_PRF_A4
MRSTSLLTFLVVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD
SLLPTYLSFQGAPLSDPDYRLPHIKVVVNLPTANIDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFELLAKKYQGPF-------
>P.metcalfi_gi|60499306|_PRF_A4
MRSTSLLTFLVVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTANMDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELGLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFELLAQKYQGPF
>P.metcalfi_gi|60499274|_PRF_A5
MRSTSLLTFLVVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTANMDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAFFSKKLRGGVVCKEYKERVLLTKRDFEFLAKKYQGHL
>P.metcalfi_gi|60499276|_PRF_A6
MRSTSLLTFLVVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTANMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFVLLAKKYQGPF
>P.metcalfi_gi|60499278|_PRF_A7
MRSTSLLTFLVVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTANMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFELLSKKYQGPF
>P.metcalfi_gi|60499280|_PRF_A8
MRSTSLLTFLVVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTANMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFVLLAKKYQGPF-------
>P.metcalfi_gi|60499282|_PRF_A9
MRSTSLLTFLVVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTANMDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFVLLAKKYQGPF--------
>P.metcalfi_gi|60499284|_PRF_A10
MRSTSLLTFLVVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTANMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFVLLAKKYQGPF-------
>P.metcalfi_gi|60499288|_PRF_A12
MRSTSLLTFLVVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFVLLAKKYQGPF--------
>P.metcalfi_gi|60499290|_PRF_A13
MRSTSLLTFLVVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTAAMDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFELLAKKYQGPF--------
>P.metcalfi_gi|60499292|_PRF_A14
MRSTSLLTFLVVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFELLAKKYQGPF-------
>P.metcalfi_gi|60499294|_PRF_A15
MRSTSLLTFLVVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTANMDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELALKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFVLLSKKYQGPF
>P.metcalfi_gi|60499296|_PRF_A16
MRSTSLLTFLVVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTANMDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFELLSKKKYQGPF
>P.metcalfi_gi|60499298|_PRF_A17
MRSTSLLTFLVVSLSTATSLAMAE------TNDVADLSRDTILLFSETQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFMRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAFFSKKLRGGVVCKEYKERVLLTKRDFEFLAKKYQGHL--------
>P.montanus_gi|60499496|_PRF_A1

MRSTSLLTFLVVSLSTATSLAMAE------TNDVADLSRDTILLFSETQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDIFLRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFELLAKKYQGPF-
>P.montanus_gi|60499498|_PRF_A2
MRSTSLLTFLVVSLSTATSLAMAE------TNDVADLSRDTILLFSETQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFMRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAFFSKKLRGGVVCKEYKERVLLTKRDFEFLAKKYQGHL-------
>P.montanus_gi|60499500|_PRF_A3
MRSTSLLTFLVVSLSTATSLAMAE------TNDVADLSRDTILLFSETQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDIFLRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFELLSKKYQGPF--------
>P.montanus_gi|60499502|_PRF_A4
MRSTSLLTFLVVSLSTATSLAMAE------TNDVADLSRDTILLFSETQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFMRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFELLAKKYQGPF-------
>P.montanus_gi|60499504|_PRF_A5
MRSTSLLTFLVVSLSTATSLAMAD------INDVVDLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTAAMDYDIFLRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFELLAKKYQGPF-------
>P.montanus_gi|60499506|_PRF_A6
MRSTSLLTFLVVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTADVDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFELLSKKYQGPF-------
>P.montanus_gi|60499508|_PRF_A7
MRSTSLLTFLVVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTADVDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFELLSKKYQGLF
>P.montanus_gi|60499510|_PRF_A8
MRSTSLLTFLVVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTADVDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFELLAKKYQGPF
>P.montanus_gi|60499512|_PRF_A9
MRSTSLLTFLVVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTADVDYDTFIRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKKPLVVPF KGSAFFSKKLRGGVVCKEYKERVLLTKRDFEFLAKKYQGHL
>P.montanus_gi|60499514|_PRF_A10
MRSTSLLTFLVVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTADVDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKKLRGGVVCKEYKERVFLTKRDFELLAKKYQGPF--------
>P.montanus_gi|60499516|_PRF_A11
MRSTSLLTFLVVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTADVDYDTFIRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFELLAKKYQGPF-
>P.montanus_gi|60499518|_PRF_A12
MRSTSLLTFLVVSLSTATSLAMAE------TNDVADLSRDTILLFSETQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVVNLPTAAMDYDTFMRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFELLAKKYQGPF-------
>P.montanus_gi|60499520|_PRF_A13
MRSTSLLTFLVVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTAAMDYDIFLRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAFFSKKLRGGVVCKEYKERVLLTKRDFEFLAKKYQGHL------ -
>P.montanus_gi|60499522|_PRF_A14
MRSTSLLTFLVVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTAAMDYDTFMRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFELLSKKYQGPF------ -
>P.montanus_gi|60499524|_PRF_A15
MRSTSLLTFLVVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTAAMDYDIFLRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKLRGGVVCKEYKERVFLTKRDFELLSKKKYQGPF-------
>P.montanus_gi|60499526|_PRF_A16
MRSTSLLTFLVVSLSTATSLAMAD------INDVADLSSDTIVLFSEVQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYRLPHIKVVNLPTAAMDYDIFLRQTDETRLKNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPLVVPF KGSAYFSKKKLRGGVVCKEYKERVFLTKRDFELLAKKYQGPF
>P.montanus_gi|60499528|_PRF_A17
MRSTSLLTFLVVSLSTATSLAMAD------INDVADLSSDTIVLFSETQKFAEDIQSSAD SLLPTYLSFQGAPLSDPDYQLPHIKVVNLPTAAMDYDTFMRQTDETRLNNNLYFYSAIVE FLKEAMTEQEDLNPAELSLKAKFEEAMANSNTLISKISDIMTQMGMSVTITLPKPPLVVPF KGSAFFSKKLRGGVVCKEYKERVLLTKRDFEFLAKKYQGHL

## Aligned Coding Sequences (mRNA)

>P.serratus_gi|60499226|_PRF_A1
ATGAGGTCAACTGCGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGAGCCACCAGCCTG ACAATCCCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGAACTGAGCGAAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGCGGAGAGCATGAAGGACAATGCGGAC TCССТССTСACTACCTATCTTAGCTTTCAGGGAGCACCCTTGAGTGACCCAGAGTACCAC CTCCCTGAGGTTAAGATTGATAACCTGCCCACCGCCGCCATGGACAATGCTACATTTTTG AACCAAACTGATGAAATCCGCCTAAAGAACAACCTCTACTTCTACAGTGCCATTGTCCAA TTCCTGCAGGTCGCCATGACAGAACAGGAAGATCTAAACCC - AGCAGAGGAGTCCCTCAA AGCCAAGTTTCAGGAAGCCATGGCCAACAGCAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGTAACAGTTACATTTCCCGATCCGCTAGTGGTGCCATA CGAAGGTAGTGCCATCTTTAAACAGAAGCTGCGGGGCGGTGTCATCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTTGCCAAGAAGTA-TCATGAGC AGCTCATTCAG-CAACAGGAACAGGAAGAATGA
>P.serratus_gi|60499228|_PRF_A2
ATGAGGTCAACTGCGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGCGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGAACTGAGCGAAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGCGGAGAGCATGAAGGACAATGCGGAC TCCCTCCTCACTACCTATCTTAGCTTTCAGGGAGCACCCTTGAGTGACCCAGAGTACCAC CTCCCTGAGGTTAAGATTGATAACCTGCCCACCGCCGCCATGGACAATGCTACATTTTTG AACCAAACTGATGAAATCCGCCTAAAGAACAACCTCTACTTCTACAGTGCCATTGTCCAA TTCCTGCAGGTCGCCATGACAGAACAGGAAGATCTAAACCC-AGCAGAGGAGTCCCTCAA AGCCAAGTTTCAGGAAGCCATGGCCAACAGCAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGTAACAGTTACATTTCCCGATCCGCTAGTGGTGCCATA CGAAGGTAGTGCCATCTTTAAACAGAAGCTGCGGGGCGGTGTCATCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTTGCCAAGAAGTA-TCATGAGC AGCTCATTCAG-CAACAGGAACAGGAAGAATGA
>P.serratus_gi|60499222|_PRF_A3
ATGAGGTCAACTGCGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGCGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGAACTGAGCAAAACT ACAATTGTCCTTTTCTCAGAGACACAGAAATTTGTGGAGAGCATGAAGGACAATGCGGAC TCCCTCCTCACTACCTATCTTAGCTTTCAGGGAGCACCCTTGAGTGACCCAGAGTACCAC CTCCCTGAGGTTAAGATTGATAACCTGCCCACCGCCACCATGGACAATGCTACATTTTTG AACCAAACTGATGAAATCCGCCTAAAGAACAACCTCTACTTCTACAGTGCCATTGTCCAA TTCCTGCAGGTCGCCATGACAGAACAGGAAGATCTAAACCC-AGCAGAGGAGTCCCTCAA AGCCAAGTTTCAGGAAGCCATGGCCAACAGCAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGTAACAGTTACATTTCCCGATCCGCTAGTGGTGCCATA CGAAGGTAGTGCCATCTTTAAACAGAAGCTGCGGGGCGGTGTCATCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGTGGGATTTTGAGTTTTTTGCCAAGAAGTA-TCATGAGC AGCTCATTCAG-CACCAGGAACAGGAAGAATGA
>P.serratus_gi|60499230|_PRF_A4
ATGAGGTCAACTGCGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGAGCCACCAGCCTG ACAATCCCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGAACTGAGCAAAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGTGGAGAGCATGAAGGACAATGCGGAC TCССTССTСАСTACCTATCTTAGCTTTCAGGGAGCACCCTTGAGTGACCCAGAGTACCAC CTCCCTGAGGTTAAGATTGATAACCTGCCCACCGCCACCATGGACAATGCTACATTTTTG AACCAAACTGATGAAATCCGCCTAAAGAACAACCTCTACTTCTACAGTGCCATTGTCCAA TTCCTGCAGGTCGCCATGACAGAACAGGAAGATCTAAACCC-AGCAGAGGAGTCCCTCAA AGCCAAGTTTCAGGAAGCCATGGCCAACAGCAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGTAACAGTTACATTTCCCGATCCGCTAGTGGTGCCATA CGAAGGTAGTGCCATCTTTAAACAGAAGCTGCGGGGCGGTGTCATCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTTGCCAAGAAGTA-TCATGAGC AGCTCATTCAG-CAACAGGAACAGGAAGAATGA
>P.serratus_gi|60499224|_PRF_A5
ATGAGGTCAACTGCGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGCGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGAACTGAGCGAAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGCGGAGAGCATGAAGGACAATGCGGAC TCCCTCCTCACTACCTATCTTAGCTTTCAGGGAGCACCCTTGAGTGACCCAGAGTACCAC CTCCCTGAGGTTAAGATTGATAACCTGCCCACCGCCGCCATGGACAATGCTACATTTTTG AACCAAACTGATGAAATCCGCCTAAAGAACAACCTCTACTTCTACAGTGCCATTGTCCAA TTTCTGCAGGTCGCCATGACAGAACAGGAAGATCTAAACCC-AGCAGAGGAGTCCCTCAA AGCCAAGTTTCAGGAAGCCATGGCCAACAGCAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGTAACAGTTACATTTCCCGATCCGCTAGTGGTGCCATA CGAAGGTAGTGCCATCTTTAAACAGAAGCTGCGGGGCGGTGTCATCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTTGCCAAGAAGTA-TCATGAGC AGCTCATTCAG-CAACAGGAACAGGAAGAATGA
>P.serratus_gi|60499220|_PRF_B1
ATGAGGTCAACTGCGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGCGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGATCTGAGCGAAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGCGGAGAGCATGAAGGACAATGCGGAC TCCCTCCTCACTACCTATCTTAGCTTACAGGGAGAACCCTACAACGATCCGGATTTCAGC CTCCCCGATATCCGGATTGCTAACCTGCCCACAGCCACCATGAACTATGATACATTCATG AGGCAAACTGATGAAATCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATCATCAAA TTCCTGCAGCTCACCATGGATGAGCAGAAAGACCTGAACCC-GGCGGACCCATCCCTCAA TGCCAAGTTTGAGCAAGTCCTGAGCAACATAAATTCTCTAATCGACAACATCTCTACCAT CATGACCCAAATGGATTTCTCAGTAAAAATTGAATTTCGCACGCCGATGGTGGAGTCATA CATAGGCAGTGCCATCTTTAAAAAGAAGTTGCGGGGCTACGTCGTCTGCAGGGAATACAA AGAGAGGGTGGCACTGACGATGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TGAAGGGC ACCTTCAGCGGGAAAGACCATGA-............
>P.serratus_gi|60499218|_PRF_B2
ATGAGGTCAACTGCGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGCGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGATCTGAGCGAAACT ACAATTGTCCTTTTCTCGGAGACACAGGAATTTGCGGAGAGCATGAAGGACAATGCGGAC

TCССТССТСАСТАССТАТСTTAGCTTACAGGGAGAACCCTACAACGATCCGGATTTCAGC CTCCCCGATATCCGGATTGCTAACCTGCCCACAGCCTCCATGAACTATGATACATTCATG AGGCAAACTGATGAAATCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATCATCAAA TTCCTGCAGCTCACCATGGATGAGCAGAAAGACCTGAACCC-GGCGGACCCATCCCTCAA TGCCAAGTTTGAGCAAGTCCTGAGCAACATAAATTCTCTAATCGACAACATCTCTACCAT CATGACCCAAATGGATTTCTCAGTAAAAATTGAATTTCGCACGCCGATGGTGGAGTCATA CATAGGCAGTGCCATCTTTAAAAAGAAGTTGCGGGGCTACGTCGTCTGCAGGGAATACAA AGAGAGGGTGGCACTGACGATGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TGAAGGGC ACCTTCAGCGGGAAAGACCATGA---------
>P.cinereus_gi|42566697|_VA_PRF_A1
ATGAGGTCAACTGCGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGAACTGAGCGAAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGTGGAGAGCATGAAGGACAATGCGGAC TCССТССTCACTACCTATCTTAGCTTACAGGGAGCACCCTTGAGTGACCCAGAGTACCAC CTCCCTGAGGTTAAGATTGATAACCTGCCCACCGCCACCATGGACAATGCTACATTTTTT AACCAAACTGATGAAATCCGCCTAAAGAACAACCTCTACTTCTACAGTGCCATTGTCCAA TTCCTGCAGGTCGCCATGACAGAACAGGAAGATCTAAACCC-AGCAGAGGAGTCCCTCAA AACCAAGTTTCAGGAAGCCCTTGCCAACAGCAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGTAACAGTTACATTTCCCGATCCGCTAGTGGTGCCATA CGAAGGTAGTGCCATCTTTAAACAGAAGCTGCGGGGCGGTGTCATCTGCAAGGAATACAA GGAGAGGGTGTTGCTGACGAAGCGGGATTTTGAGTTTTTTGCCAAGAAGTA-TCATGAGC AGCTCGTTCAG-CAACAGGAACAGGAAGAATGA
>P.cinereus_gi|42566675|_VA_PRF_A2
ATGAGGTCAACTGCGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGAACTGAGCGAAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGTGGAGAGCATGAAGGACAATGCGGAC TCССТССТСАСТАССТАTCTTAGCTTTCAGGGAGCACCCTTGAGTGACCCAGAGTACCAC CTCCCTGAGGTTAAGATTGATAACCTGCCCACCGCCACCATGGACAATGCTACATTTTTT AACCAAACTGATGAAATCCGCCTAAAGAACAACCTCTACTTCTACAGTGCCATTGTCCAA TTCCTGCAGGTCGCCATGACAGAACAGGAAGATCTAAACCC-AGCAGAGGAGTCCCTCAA AACCAAGTTTCAGGAAGCCCTTGCCAACAGCAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGTAACAGTTACATTTCCCGATCCGCTAGTGGTGCCATA CGAAGGTAGTGCCATCTTTAAACAGAAGCTGCGGGGCGGTGTCATCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTTGCCAAGAAGTA-TCATGAGC AGCTCGTTCAG-CAACAGGAACAGGAAGAATGA
>P.cinereus_gi|42566685|_VA_PRF_A3
ATGAGGTCAACTGCGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGAACTGAGCGAAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGTGGAGAGCATGAAGGACAATGCGGAC TCССТССТСАСТАССТАТСTTAGCTTTCAGGGAGCACCCTTGAGTGACCCAGAGTACCAC CTCCCTGAGGTTAAGATTGATAACCTGCCCACCGCCACCATGGACAATGCTACATTTTTT AACCAAACTGATGAAATCCGCCTAAAGAACAACCTCTACTTCTACAGTGCCATTGTCCAA TTCCTGCAGGTCGCCATGACAGAACAGGAAGATCTAAACCC-AGCAGAGGAGTCCCTCAA AACCAAGTTTCAGGAAGCCCTTGCCAACAGCAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGTAACAGTTACATTTCCCGATCCGCTAGTGGTGCCATA CGAAGGTAGTGCCATCTTTAAACAGAAGCTGCGGGGCGGTGTCATCTGCAAGGAATACAA GGAGAGGGTGTTGCTGACGAAGCGGGATTTTGAGTTTTTTGCCAAGAAGTA-TCATGAGC AGCTCGTTCAG-CAACAGGAACAGGAAGAATGA
>P.cinereus_gi|42566669|_VA_PRF_A4
ATGAGGTCAACTGCGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGAACTGAGCGAAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGTGGAGAGCATGAAGGACAATGCGGAC TCCCTCCTCACTACCTATCTTAGCTTTCAGGGAGCACCCTTGAGTGACCCAGAGTACCAC CTCCCTGAGGTTAAGATTGATAACCTGCCCACCGCCACCATGGACAATGCTACATTTTTG AACCAAACTGATGAAATCCGCCTAAAGAACAACCTCTACTTCTACAGTGCCATTGTCCAA TTCCTGCAGGTCGCCATGACAGAACAGGAAGATCTAAACCC-AGCAGAGGAGTCCCTCAA AGCCAAGTTTCAGGAAGCCCTTGCCAACAGCAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGTAACAGTTACATTTCCCGATCCGCTAGTGGTGCCATA CGAAGGTAGTGCCATCTTTAAACAGAAGCTGCGGGGCGGTGTCATCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTTGCCAAGAAGTA-TCATGAGC AGCTCGTTCAG-CAACAGGAACAGGAAGAATGA
>P.cinereus_gi|42566681|_VA_PRF_A5
ATGAGGTCAACTGCGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGATCTGAGCAAAACT ACAATTGTCCTTTTCTCGGAGACACACAAATTTGTGGAGAGCATGAAGGACAATGCGGAC TCССТССТСАСТАССТАТСTTAGCTTTCAGGGAGCACCCTTGAGTGACCCAGAGTACCAC CTCCCTGAGGTTAAGATTGATAACCTGCCCACCGCCACCATGGACAATGCTACATTTTTG AACCAAACTGATGAAATCCGCCTAAAGAACAACCTCTACTTCTACAGTGCCATTGTCCAA TTCCTGCAGGTCGCCATGACAGAACAGGAAGATCTAAACCC-AGCAGAGGAGTCCCTCAA AGCCAAGTTTCAGGAAGCCCTTGCCAACAGCAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGTAACAGTTACATTTCCCGATCCGCTAGTGGTGCCATA CGAAGGTAGTGCCATCTTTAAACAGAAGCTGCGGGGCGGTGTCATCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTTGCCAAGAAGTA-TCATGAGC AGCTCATTCAG-CAACAGGAACAGGAAGAATGA
>P.cinereus_gi|42566695|_VA_PRF_A6
ATGAGGTCAACTGCGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGATCTGAGCAAAACT ACAATTGTCCTTTTCTCGGAGACACACAAATTTGTGGAGAGCATGAAGGACAATGCGGAC TCССТССТСACTACCTATCTTAGCTTTCAGGGAGCACCCTTGAGTGACCCAGAGTACCAC CTCCCTGAGGTTAAGATTGATAACCTGCCCACCGCCACCATGGACAATGCTACATTTTTT AACCAAACTGATGAAATCCGCCTAAAGAACAACCTCTACTTCTACAGTGCCATTGTCCAA TTCCTGCAGGTCGCCATGACAGAACAGGAAGATCTAAACCC-AGCAGAGGAGTCCCTCAA AACCAAGTTTCAGGAAGCCCTTGCCAACAGCAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGTAACAGTTACATTTCCCGATCCGCTAGTGGTGCCATA

CGAAGGTAGTGCCATCTTTAAACAGAAGCTGCGGGGCGGTGTCATCTGCAAGGAATACAA GGAGAGGGTGTTGCTGACGAAGCGGGATTTTGAGTTTTTTGCCAAGAAGTA-TCATGAGC AGCTCGTTCAG-CAACAGGAACAGGAAGAATGA
>P.cinereus_gi|42566689|_VA_PRF_A7
ATGAGGTCAACTGCGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGATCTGAGCAAAACT ACAATTGTCCTTTTCTCGGAGACACACAAATTTGTGGAGAGCATGAAGGACAATGCGGAC TCCCTCCTCACTACCTATCTTAGCTTTCAGGGAGCACCCTTGAGTGACCCAGAGTACCAC CTCCCTGAGGTTAAGATTGATAACCTGCCCACCGCCACCATGGACAATGCTACATTTCTG AACCAAACTGATGAAATCCGCCTAAAGAACAACCTCTACTTCTACAGTGCCATTGTCCAA TTCCTGCAGGTCGCCATGACAGAACAGGAAGATCTAAACCC-AGCAGAGGAGTCCCTCAA AGCCAAGTTTCAGGAAGCCCTTGCCAACAGCAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGTAACAGTTACATTTCCCGATCCGCTAGTGGTGCCATA CGAAGGTAGTGCCATCTTTAAACAGAAGCTGCGGGGCGGTGTCATCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTTGCCAAGAAGTA-TCATGAGC AGCTCATTCAG-CAACAGGAACAGGAAGAATGA
>P.cinereus_gi|42566693|_VA_PRF_A8
ATGAGGTCAACTGCGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGATCTGAGCAAAACT ACAATTGTCCTTTTCTCGGAGACACACAAATTTGTGGAGAGCATGAAGGACAATGCGGAC TCCCTCCTCACTACCTATCTTAGCTTTCAGGGAGCACCCTTGAGTGACCCAGAGTACCAC CTCCCTGAGGTTAAGATTGATAACCTGCCCACCGCCACCATGGACAATGCTACATTTTTG AACCAAACTGATGAAATCCGCCTAAAGAACAACCTCTACTTCTACAGTGCCATTGTCCAA TTCCTGCAGGTCGCCATGACAGAACAGGAAGATCTAAACCC-AGCAGAGGAGTCCCTCAA AGCCAAGTTTCAGGAAGCCCTTGCCAACAGCAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGTAACAGTTACATTTCCCGATCCGCTAGTGGTGCCATA CGAAGGTAGTGCCATCTTTAAACAGAAGCTGCGGGGCGGTGTCATCTGCAAGGAATACAA GGAGAGGGTGTTGCTGACGAAGCGGGATTTTGAGTTTTTTGCCAAGAAGTA-TCATGAGC AGCTCGTTCAG-CAACAGGAACAGGAAGAATGA
>P.cinereus_gi|42566709|_VA_PRF_B1
ATGAGGTCAACTGCGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGATCTGAGCGAAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGCGGAGAGCATGAAGGACAATGCGGAC TCCCTCCTCACTACCTATCTTAGCTTACAGGGAGAACCCTACAACAATCCGGATTTCAGC CTCCGCGATATCCAGATTGCTAACCTGCCCACAGCCACCATGAACTATGATACATTCATG AGGCAAACTGATGAAATCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATCATCAAA TTCCTGCAGCTCACCATGGATGAGCAGAAAGACCTGAACCC-GGAGGACCCGTCCCTCAA TGCCAAGTTTGAGCAAGTCCTGAGCAACATAAATTCTCTAATCAACAAAATCTCTACCAT CATGACCCAAATGGATTTCTCAGTAAAAATTGAATTTCGTTCGCCGATGGTGGAGTCATA CATAGGCAGTGCCATCTTTAAAAAGAAGCTGCGGGGCTACGTCGTCTGCAGGGAATACAA AGAGAGGGTGGCACTGACGATGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TGAAGGGC ACCTTCATCGGGCAAGACCATGA---------
>P.cinereus_gi|42566707|_VA_PRF_B2
ATGAGGTCAACTGCGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGATCTGAGCGAAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGCGGAGAGCATGAAGGACAATGCGGAC TCCCTCCTCACTACCTATCTTAGCTTACAGGGAGAACCCTACAACAATCCGGATTTCAGC CTCCGCGATATCCAGATTGCTAACCTGCCCACAGCCACCATGAACTATGATACATTCATG AGGCAAACTGATGAAATCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATCATCAAA TTCCTGCAGCTCACCATGGATGAGCAGAAAGACCTGAACCC-GGAGGACCCGTCCCTCAA TGCCAAGTTTGAGCAAGTCCTGAGCAACATAAATTCTCTAATCAACAAAATCTCTACCAT CATGACCCAAATGGATTTCTCAGTAAAAATTGAATTTCGCACGCCGATGGTGGAGTCATA CATAGGCAGTGCCATCTTTAAAAAGAAGCTGCGGGGCTACATCGTCTGCAGGGAATACAA AGAGAGGGTGGCACTGACGATGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TGAAGGGC ACCTTCATCGGGCAAGACCATGA---------
>P.cinereus_gi|42566711|_VA_PRF_B3
ATGAGGTCAACTGCGCTGCTCATCTTCTTGGTGGTGTCTGTACGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGATCTGAGCGAAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGCGGAGAGCATGAAGGACAATGCGGAC TCCCTCCTCACTACCTATCTTAGCTTACAGGGAGAACCCTACAACAATCCGGATTTCAGC CTCCGCAGTATCCAGATTGCTAACCTGCCCACAGCCACCATGAACTATGATACATTCATG AGGCAAACTGATGAAATCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATCATCAAA TTCCTGCAGCTCACCATGGATGAGCAGAAAGACCTGAACCC-GGAGGACCCGTCCCTCAA TGCCAAGTTTGAGCAAGTCCTGAGCAACATAAATTCTCTAATCAACAAAATCTCTACCAT CATGACCCAAATGGATTTCTCAGTAAAAATTGAATTTCGCACGCCGATGGTGGAGTCATA CATAGGCAGTGCCATCTTTAAAAAGAAGCTGCGGGGCTACATCGTCTGCAGGGAATACAA AGAGAGGGTGGCACTGACGATGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TGAAGGGC ACCTTCATCGGGCAAGACCATGA--------
>P.cinereus_gi|42566699|_VA_PRF_B4
ATGAGGTCAACTGCGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGATCTGAGCGAAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGCGGAGAGCATGAAGGACAATGCGGAC TCССТССТСАСТАССТАТСTTAGCTTACAGGGAGAACCCTACAACAATCCGGATTTCAGC CTCCGCGATATCCAGATTGCTAACCTGCCCACAGCCACCATGAACTATGATACATTCATG AGGCAAACTGATGAAATCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATCATCAAA TTCCTGCAGCTCACCATGGATGAGCAGAAAGACCTGAACCC-GGAGGACCCGTCCCTCAA TGCCAAGTTTGAGCAAGTCCTGAGCAACATAAATTCTCTAATCAACAAAATCTCTACCAT CATGACCCAAATGGATTTCTCAGTAAAAATTGAATTTCGTTCGCCGATGGTGGAGTCATA CATAGGCAGTGCCATCTTTAAAAAGAAGCTGCGGGGCTACGTCGTCTGCAGGGAATACAA AGAGAGGGTGTTGCTGACGAAGCGGGATTTTGAGTTTTTTGCCAAGAAGTA-TCATGAGC AGCTCGTTCAG-CAACAGGAACAGGAAGAATGA
>P.cinereus_gi|42566703|_VA_PRF_B5
ATGAGGTCAACTGCGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGAGCCACCAGCCTG
ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGAACTGAGCGAAACT

ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGTGGAGAGCATGAAGGACAATGCGGAC TCCCTCCTCACTACCTATCTTAGCTTACAGGGAGAACCCTACAACAATCCGGATTTCAGC CTCCGCGATATCCAGATTGCTAACCTGCCCACAGCCACCATGAACTATGATACATTCATG AGGCAAACTGATGAAATCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATCATCAAA TTCCTGCAGCTCACCATGGATGAGCAGAAAGACCTGAACCC-GGAGGACCCGTCCCTCAA TGCCAAGTTTGAGCAAGTCCTGAGCAACATAAATTCTCTAATCAACAAAATCTCTACCAT CATGACCCAAATGGATTTCTCAGTAAAAATTGAATTTCGTTCGCCGATGGTGGAGTCATA CATAGGCAGTGCCATCTTTAAAAAGAAGCTGCGGGGCTACGTCGTCTGCAGGGAATACAA AGAGAGGGTGGCACTGACGATGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TGAAGGGC ACCTTCATCGGGCAAGACCATGA
>P.cinereus gil42566705| VA PRF B6
ATGAGGTCAACTGCGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGAACTGAGCGAAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGCGGAGAGCATGAAGGACAATGCGGAC TCCCTCCTCACTACCTATCTTAGCTTACAGGGAGAACCCTACAACAATCCGGATTTCAGC CTCCGCGATATCCAGATTGCTAACCTGCCCACAGCCACCATGAACTATGATACATTCATG AGGCAAACTGATGAAATCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATCATCAAA TTCCTGCAGCTCACCATGGATGAGCAGAAAGACCTGAACCC-GGAGGACCCGTCCCTCAA TGCCAAGTTTGAGCAAGTCCTGAGCAACATAAATTCTCTAATCAACAAAATCTCTACCAT CATGACCCAAATGGATTTCTCAGTAAAAATTGAATTTCGTTCGCCGATGGTGGAGTCATA CATAGGCAGTGCCATCTTTAAAAAGAAGCTGCGGGGCTACGTCGTCTGCAGGGAATACAA AGAGAGGGTGGCACTGACGATGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TGAAGGGC ACCTTCATCGGGCAAGACCATGA-------
>P.cinereus_NH_PRF_A1
ATGAGGTCAACTGCGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGAGCCACCAGCCTG ACAATCTCGTCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGATCTGAGCAAAACT ACAATTGTCCTTTTCTCGGAGACACACAAATTTGTGGAGAGCATGAAGGACAATGCGGAC TCCCTCCTCACTACCTATCTTAGCTTTCAGGGAGCACCCTTGAGTGACCCAGAGTACCAC CTCCCTGAGGTTAAGATTGATAACCTGCCCACCGCCACCATGGACAATACTACATTTTTG AACCAAACTGATGAAATCCGCCTAAAGAACAACCTCTACTTCTACAGTGCCATTGTCCAA TTCCTGCAGGTCGCCATGACAGAACAGGAAGATCTAAACCC-AGTAGAGGAGTCCCTCAA AACCAAGTTTCAGGAAGCCCTTGCCAACAGCAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGCAACAGTTACATTTCCCGATCCGCTAGTGGTGCCATA CGAAGGTAGTGCCATCTTTAAACAGAAGCTGCGGGGCGGTGTCATCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTTGCCCAGAAGTA-TCATGAGC AGCTCATTCAG-CAGCAGGAACAGGAAGAATGA
>P.cinereus_NH_PRF_A2
ATGAGGTCAACTGCGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGAGCCACCAGCCTG ACAATCTCGTCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGATCTGAGCAAAACT ACAATTGTCCTTTTCTCGGAGACACACAAATTTATGGAGAGCATGAAGGACAATGCGGAC TCCCTCCTCACTACCTATCTTAGCTTTCAGGGAGCACCCTTGAGTGACCCAGAGTACCAC CTCCCTGAGGTTAAGATTGATAACCTGCCCACCGCCACCATGGACAATACTACATTTTTG AACCAAACTGATGAAATCCGCCTAAAGAACAACCTCTACTTCTACAGTGCCATTGTCCAA TTCCTGCAGGTCGCCATGACAGAACAGGAAGATCTAAACCC-AGCAGAGGAGTCCCTCAA AACCAAGTTTCAGGAAGCCCTTGCCAACAGCAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGCAACAGTTACATTTCCCGATCCGCTAGTGGTGCCATA CGAAGGTAGTGCCATCTTTAAACAGAAGCTGCGGGGCGGTGTCATCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTTGAGTTTTTTGCCCAGAAGTA-TCATGAGC AGCTCATTCAG-CAACAGGAACAGGAAGAATGA
>P.cinereus_NH_PRF_A3
ATGAGGTCAACTGCGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGAGCCACCAGCCTG ACAATCTCGTCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGATCTGAGCAAAACT ACAATTGTCCTTTTCTCGGAGACACACAAATTTGTGGAGAGCATGAAGGACAATGCGGAC TCCCTCCTCACTACCTATCTTAGCTTTCAGGGAGCACCCTTGAGTGACCCAGAGTACCAC CTCCCTGAGGTTAAGATTGATAACCTGCCCACCGCCACCATGGACAATACTACATTTTTG AACCAAACTGATGAAATCCGCCTAAAGAACAACCTCTACTTCTACAGTGCCATTGTCCAA TTCCTGCAGGTCGCCATGACAGAACAGGAAGATCTAAACCC-AGCAGAGGAGTCCCTCAA AACCAAGTTTCAGGAAGCCCTTGCCAACAGCAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGCAACAGTTACATTTCCCGATCCGCTAGTGGTGCCATA CGAAGGTAGTGCCATCTTTAAACAGAAGCTGCGGGGCGGTGTCATCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTTGCCCAGAAGTA-TCATGAGC AGCTCATTCAG-CAACAGGAGCAGGAAGAATGA
>P.cinereus_NH_PRF_A4
ATGAGGTCAACTGCGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGAGCCACCAGCCTG ACAATCTCGTCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGATCTGAGCAAAACT ACAATTGTCCTTTTCTCGGAGACACACAAATTTGTGGAGAGCATGAAGGACAATGCGGAC TCCCTCCTCACTACCTATCTTAGCTTTCAGGGAGCACCCTTGAGTGACCCAGAGTACCAC CTCCCTGAGGTTAAGATTGATAACCTGCCCACCGCCACCATGGACAATACTACATTTTTTG AACCAAACTGATGAAATCCGCCTAAAGAACAACCTCTACTTCTACAGTGCCATTGTCCAA TTCCTGCAGGTCGCCATGACAGAACAGGAAGATCTAAACCC-AGCAGAGGAGTCCCTCAA AACCAAGTTTCAGGAAGCCCTTGCCAACAGCAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGCAACAGTTACATTTCCCGATCCGCTAGTGGTGCCATA CGAAGGTAGTGCCATCTTTAAACAGAAGCTGCGGGGCGGTGTCATCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTTGCCCAGAAGTA-TCATGAGC AGCTCATTCAG-CAGCAGGAACAGGAAGAATGA
>P.cinereus_NH_PRF_A5
ATGAGGTCAACTGCGCTGCTCATCTTCCTGGTGGTGTCCGTACGCAGAGCCACCAGCCTG ACAATCTCGTCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGATCTGAGCAAAACT ACAATTGTCCTTTTCTCGGAGACACACAAATTTGTGGAGAGCATGAAGGACAATGCGGAC TCCCTCCTCACTACCTATCTTAGCTTTCAGGGAGCACCCTTGAGTGACCCAGAGTACCAC CTCCCTGAGGTTAAGATTGATAACCTGCCCACCGCCACCATGGACAATACTACATTTTTG AACCAAACTGATGAAATCCGCCTAAAGAACAACCTCTACTTCTACAGTGCCATTGTCCAA TTCCTGCAGGTCGCCATGACAGAACAGGAAGATCTAAACCC-AGCAGAGGAGTCCCTCAA AACCAAGTTTCAGGAAGCCCTTGCCAACAGCAATACTCTTATCAGCAAAATCTCTGCCAT

CATGACCCAGATGGGTATGTCAGCAACAGTTACATTTCCCGATCCGCTAGTGGTGCCATA CGAAGGTAGTGCCATCTTTAAACAGAAGCTGCGGGGCGGTGTCATCTGCAAGGAATACAA AGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTTGCCCAGAAGTA-TCATGAGC AGCTCATTCAG-CAACAGGAACAGGAAGAATGA
>P.cinereus_NH_PRF_A6
ATGAGGTCAACTGTGCTGCTCATCCTCCTGGTGGTGTCTGTACGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGATCTGAGCGAAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGTGGAGAGCATGAAGGACAATGCGGAC TCCCTCCTCACTACCTATCTTAGCTTTCAGGGAGCACCCTTGAGTGACCCAGAGTACCAC CTCCCTGAGGTTAAGATTGATAACCTGCCCACCGCCACCATGGACAATGCTACATTTTTG AACCAAACTGATGAAATCCGCCTAAAGAACAACCTCTACTTCTACAGTGCCATTGTCCAA TTCCTGCAGGTCGCCATGACAGAACAGGAAGATCTAAACCC-AGCAGAGGAGTCCCTCAA AGCCAAGTTTCAGGAAGCCCTTGCCAACAGCAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGCAACAGTTACATTTCCCGATCCGCTAGTGGTGCCATA CGAAGGTAGTGCCATCTTTAAACAGAAGCTGCGGGGCGGTGTCATCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCACGATTTTGAGTTTTTTGCCAAGAAGTA-TCATGAGC AGCTCATTCAG-CAACAGGAACAGGAAGAATGA
>P.cinereus_NH_PRF_A7
ATGAGGTCAACTGTGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGATCTGAGCGAAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGTGGAGAGCATGAAGGACAATGCGGAC TCCCTCCTCACTACCTATCTTAGCTTTCAGGGAGCACCCTTGAGTGACCCAGAGTACCAC CTCCCTGAGGTTAAGATTGATAACCTGCCCACCGCCACCATGGACAATGCTACATTTTTG AACCAAACTGATGAAATCCGCATAAAGAACAACCTCTACTTCTACAGTGCCATTGTCCAA TTCCTGCAGGTCGCCATGACAGAACAGGAAGATCTAAACCC-AGCAGAGGAGTCCCTCAA AGCCAAGTTTCAGGAAGCCCTTGCCAACAGCAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGCAACAGTTACATTTCCCGATCCGCTAGTGGTGCCATA CGAAGGTAGTGCCATCTTTAAACAGAAGCTGCGGGGCGGTGTCATCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCAGGATTTTGAGTTTTTTGCCAAGAAGTA-TCATGAGC AGCTCATTCAG-CAACAGGAACAGGAAGAATGA
>P.cinereus_NH_PRF_A8
ATGAGGTCAACTGTGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGATCTGAGCGAAACT ACAATTGTCCTCTTCTCGGAGACACAGAAATTTGTGGAGAGCATGAAGGACAATGCGGAC TCCCTCCTCACTACCTATCTTAGCTTTCAGGGAGCACCCTTGAGTGACCCAGAGTACCAC CTCCCTGAGGTTAAGATTGATAACCTGCCCACCGCCACCATGGACAATGCTACATTTTTG AACCAAACTGATGAAATCCGCCTAAAGAACAACCTCTACTTCTACAGTGCCATTGTCCAA CTCCTGCAGGTCGCCATGACAGAACAGGAAGATCTAAACCC-AGCAGAGGAGTCCCTCAA AGCCAAGTTTCAGGAAGCCCTTGCCAACAGCAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGCAACAGTTACATTTCCCGATCCGCTAGTGGTGCCATA CGAAGGTAGTGCCATCTTTAAACAGAAGCTGCGGGGCGGTGTCATCTGCAAGGAATACAA TGAGAGGGTGTTACTGACGAAGCAGGATTTTGAGTTTTTTGCCAAGAAGTA-TCATGAGC AGCTCATTCAG-CAACAGGGACAGGAAGAATGA
>P.cinereus_NH_PRF_A9
ATGAGGTCAACTGTGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGATCTGAGCGAAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGTGGAGAGCATGAAGGACAATGCGGAC TCCCTCCTCACTACCTATCTTAGCTTTCAGGGAGCACCCTTGAGTGACCCAGAGTACCAC CTCCCTGAGGTTAAGATTGATAACCTGCCCACCGCCACCATGGACAATGCTACATTTTTG AACCAAACTGATGAAATCCGCCTAAAGAACAACCTCTACTTCTACAGTGCCATTGTCCAA TTCCTGCAGGTCGCCATGACAGAACAGGAAGATCTAAACCC-AGCAGAGGAGTCCCTCAA AGCCAAGTTTCAGGAAGCCCTTGCCAACAGCAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGCAACAGTTACATTTCCCGATCCGCTAGTGGTGCCATA CGAAGGTAGTGCCATCTTTAAACAGAAGCTGCGGGGCGGTGTCATCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCAGGATTTTGAGTTTTTTGCCAAGAAGTA-TCATGAGC AGCTCATTCAG-CAACAGGAACAGGAAGAATGA
>P.cinereus_NH_PRF_A10
ATGAGGTCAACTGTGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGATCTGAGCGAAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGTGGAGAGCATGAAGGACAATGCGGAC TCCCTCCTCACTACCTATCTTAGCTTTCAGGGAGCACCCTTGAGTGACCCAGAGTACCAC CTCCCTGAGGTTAAGATTGATAACCTGCCCACCGCCACCATGGACAATGCTACATTTTTG AACCAAACTGATGAAATCCGCCTAAAGAACAACCTCTACTTCTACAGTGCCATTGTCCAA TTCCTGCAGGTCGCCATGACAGAACAGGAAGATCTAAACCC-AGCAGAGGAGTCCCTCAA AGCCAAGTTTCAGGAAGCCCTTGCCAACAGCAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCGGCAACAGTTACATTTCCCGATCCGCTAGTGGTGCCATA CGAAGGTAGTGCCATCTTTAAACAGAAGCTGCGGGGCGGTGTCATCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCAGGATTTTGAGTTTTTTGCCAAGAAGTA-TCATGAGC AGCTCATTCAG-CAACAGGAACAGGAAGAATGA
>P.cinereus_NH_PRF_B1
ATGAGGTCAACTGCGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGAGTCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGATCTGAGCGAAACT ACAATTGTCCTTTTCTCGGAGACACAGGAATTTGCGGAGAGCATGAAGGACAATGCGGAC TCCСTCСTCACTACCTATCTTAGCTTACAGGGAGAACCCTACAACGATCCGGATTTCAGC CTCCCCGATATCCGGATTGCTAACCTGCCCACAGCCACCATGAACTATGATACATTCATG AGGCAAACTGATGAAATCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATCATCAAA TTCCTGCAGCTCACCATGGATGAGCAGAAAGAGCTGAACCC-GGCGGACCCATCCCTCAA TGCCAAGTTTGAGCAAGTCCTGAGCAACATAAATTCTCTAATCGACAACATCTCTACCAT CATGACCCAAATGGATTTCTCAGTAAAAATTGAACTTCGCACGCCGATGGTGGAGTCATA CATAGGCAGTGCCATCTTTAAAAAGAAGCTGCGGGGCTACGTCGTCTGCAGGGAATACAA AGAGAGGGTGGCACTGACGATGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TGAAGGGC ACCTTCAGCGGGCAAGACCATGA---------
>P.cinereus_NH_PRF_B2
ATGAGGTCAACTGTGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGAGCCACCAGCCTG

ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGATCTGAGCGAAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGCGGAGAGCATGAAGGACAATGCGGAC TCССТССТСАСТАССТАТСTTAGCTTACAGGGAGAACCCTACAACAATCCGGATTTCAGC CTCCGCGATATCCAGATTGCTAACCTGCCCACAGCCACCATGAACTATGATACATTCATG AGGCAAACTGATGAAATCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATCATCAAA TTCCTGCAGCTCACCATGGATGAGCAGAAAGACCTGAACCC-GGAGGACCCGTCCCTCAA TGCCAAGTTTGAGCAAGTCCTGAGCAACATAAATTCTCTAATCCACAAAATCTCTACCAT CATGACCCAAATGGATTTCTCAGTAAAAATTGAATTTCGCACGCCGATGGTGGAGTCATA CATAGGCAGTGCCATCTTTAAAAAGAAGCTGCGGGGCTACGTCGTCTGCAGGGAATACAA AGAGAGGGTGGCACTGACGATGTGGGATTTTGAGTTTTTAGCCAAGAAGTA-TGAAGGGC ACCTTCATCGGGCAAGACCATGA--------
>P.cinereus_NH_PRF_B3
ATGAGGTCAATTGTGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGATCTGAGCGAAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGCGGAGAGCATGAAGGACAATGCGGAC TCССТССТСАСТАССТАТСTTAGCTTACAGGGAGAACCCTACAACAATCCGGATTTCAGC CTCCGCGATATCCAGATTGCTAACCTGCCCACAGCCACCATGAACTATGATACATTCATG AGGCAAACTGATGAAATCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATCATCAAA TTCCTGCAGCTCACCATGGATGAGCAGAAAGACCTGAACCC-GGAGGACCCGTCCCTCAA TGCCAAGTTTGAGCAAGTCCTGAGCAACATAAATTCTCTAATCCACAAAATCTCTACCAT CATGACCCAAATGGATTTCTCAGTAAAAATTGAATTTCGCACGCCGATGGTGGAGTTATA CATAGGCAGTGCCATCTTTAAAAAGAAGCTGCGGGGCTACGTCGTCTGCAGGGAATACAA AGAGAGGGTGGCACTGACGATGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TGAAGGGC ACCTTCATCGGGCAAGACCATGA-...........
>P. cinereus_NH_PRF_B4
ATGAGGTCAACTGTGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGATCTGAGCGAAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGCGGAGAGCATGAAGGACAATGCGGAC TCССТССТСАСТАССТАTCTTAGCTTACAGGGAGAACCCTACAACAATCCGGATTTCAGC CTCCGCGATATCCAGATTGCTAACCTGCCCACAGCCACCATGAACTATGATACATTCATG AGGCAAACTGATGAAATCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATCATCAAA TTCCTGCAGCTCACCATGGATGAGCAGAAAGACCTGAACCC-GGAGGACCCGTCCCTCAA TGCCAAGTTTGAGCAAGTCCTGAGCAACATAAATTCTCTAATCCACAAAATCTCTACCAT CATGACCCAAATGGATTTCTCAGTAAAAATTGAATTTCGCACGCCGATGGTGGAGTCATA CATAGGCAGTGCCATCTTTAAAAAGAAGCTGCGGGACTACGTCGTCTGCAGGGAATACAA AGAGAGGGTGGCACTGACGATGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TGAAGGGC ACCTTCATCGGGCAAGACCATGA--------
>P.cinereus_NH_PRF_B5
ATGAGGTCAACTGTGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGATCTGAGCGAAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGCGGAGAGCATGAAGGACAATGCGGAC TCССTCСTCACTACCTATCTTAGCTTACAGGGAGAACCCTACAACAATCCGGATTTCAGC CTCCGCGATATCCAGATTGCTAACCTGCCCACAGCCGCCATGAACTATGATACATTCATG AGGCAAACTGATGAAATCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATCATCAAA TTCCTGCAGCTCACCATGGATGAGCAGAAAGACCTGAACCC-GGAGGACCCGTCCCTCAA TGCCAAGTTTGAGCAAGTCCTGAGCAACATAAATTCTCTAATCCACAAAATCTCTACCAT CATGACCCAAATGGATTTCTCAGTAAAAATTGAATTTCGCACGCCGATGGTGGAGTCATA CATAGGCAGTGCCATCTTTAAAAAGAAGCTGCGGGGCTACGTCGTCTGCAGGGAATACAA AGAGAGGGTGGCACTGACGATGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TGAAGGGC ACCTTCATCGGGCAAGACCATGA----------
>P.cinereus_NH_PRF_B6
ATGAGGTCAACTGTGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGATCTGAGCGAAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGCGGAGAGCATGAAGGACAATGCGGAC TCССТССТСАСТАССТАТСTTAGCTTACAGGGAGAACCCTACAACAATCCGGATTTCAGC CTCCGCGATATCCAGATTGCTAACCTGCCCACAGCCACCATGAACTATGATACATTCATG AGGCAAACTGATGAAATCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATCATCAAA TTCCTGCAGCTCACCATGGATGAGCAGAAAGACCTGAACCC-GGAGGACCCGTCCCTCAA TGCCAAGTTTGAGCAAGTCCTGAGCAACATAAATTCTCTAATCCACAAAATCTCTACCAT CATGACCCAAATGGATTTCTCAGTAAAAATTGAATTTCGCACGCCGATGGTGGAGCCATA CATAGGCAGTGCCATCTTTAAAAAGAAGCTGCGGGGCTACGTCGTCTGCAGGGAATACAA AGAGAGGGTGGCACTGACGATGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TGAAGGAC ACCTTCATCGGGCAAGACCATGA-.............
>P.cinereus_NH_PRF_B7
ATGAGGTCAACTGTGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGATCTGAGCGAAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGCGGAGAGCATGAAGGACAATGCGGAC TCCCTCCTCACTACCTATCTTAGCTTACAGGGAGAACCCTACAACAATCCGGATTTCAGC CTCCGCGATATCCAGATTGCTAACCTGCCCACAGCCACCATGAACTATGATACATTCATG AGGCAAACTGATGAAATCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATCATCAAA TTCCTGCAGCTCACCATGGATGAGCAGAAAGACCTGAACCC-GGAGGACCCGTCCCTCAA TGCCAAGTTTGAGCAAGTCCTGAGCAACATAAATTCTCTAATCCACAAAATCTCTACCAT CATGACCCAAATGGATTTCTCAGTAAAAATTGAATTTCGCACGCCGATGGTGGAGTCATA CATAGGCAGTGCCATCTTTAAAAAGAAGCTGCGGGGCTACGTCGTCTGCAGGGAATACAA AGAGAGGGTGGCACTGACGATGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TGAAGGAC ACCTTCATCGGGCAAGACCATGA---------
>P.cinereus_NH_PRF_B8
ATGAGGTCAACTGTGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGATCTGAGCGAAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGCGGAGAGCATGAAGGACAATGCGGAC TCCCTCCTCACTACCTATCTTAGCTTACAGGGAGAACCCTACAACAATCCGGATTTCAGC CTCCGCGATATCCAGATTGCTAACCTGCCCACAGCCACCATGAACTATGATACATTCATG AGGCAAACTGATGAAATCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATCATCAAA TTCCTGCAGCTCACCATGGATGAGCAGAAAGACCTGAACCC-GGAGGACCCGTCCCTCAA

TGCCAAGTTTGAGCAAGTCCTGAGCAACATAAATTCTCTAATCCACAAAATCTCTACCAT CATGACCCAAATGGATTTCTCAGTAAAAATTGAATTTCGCACGCCGATGGTGGAGTCATA CATAGGCAGTGCCATCTTTAAAAAGAAGCTGCGGGGCTACGTCGTCTGCAGGGAATACAA AGAGAGGGTGGCACTGACGATGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TGAAGGGC ACCTTCATCGGGCAAGACCATGA---------
>P.cinereus_NH_PRF_B9
ATGAGGTCAACTGTGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGATCTGAGCGAAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGCGGAGAGCATGAAGGACAATGCGGAC TCCCTCCCCACTACCTATCTTAGCTTACAGGGAGAACCCTACAACAATCCGGATTTCAGC CTCCGCGATATCCAGATTGCTAACCTGCCCACAGCCACCATGAACTATGATACATTCATG AGGCAAACTGATGAAATCCGCCTGAAAAACAACCTCTACTTCTACAGTACCATCATCAAA TTCCTGCAGCTCACCATGGATGAGCAGAAAGACCTGAACCC-GGAGGACCCGTCCCTCAA TGCCAAGTTTGAGCAAGTCCTGAGCAACATAAATTCTCTAATCCACAAAATCTCTACCAT CATGACCCAAATGGATTTCTCAGTAAAAATTGAATTTCGCACGCCGATGGTGGAGTCATA CATAGGCAGTGCCATCTTTAAAAAGAAGCTGCGGGGCTACGTCGTCTGCAGGGAA--CAA AGAGAGGGTGACACTGACGATGCGGGATTTTGA

## >P.cinereus_NH_PRF_B10

ATGAGGTCAACTGTGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGATCTGAGCGAAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGCGGAGAGCATGAAGGACAATGCGGAC TCCCTCCTCACTACCTATCTTAGCTTACAGGGAGAACCCTACAACAATCCGGATTTCAGC CTCCGCGATATCCAGATTGCTAACCTGCCCACAGCCACCATGAACTATGATACATTCATG AGGCAAACTGATGAAATCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATCATCAAA TTCCTGCAGCTCACCATGGATGAGCAGAAAGACCTGAACCC-GGAGGACCCGTCCCTCAA TGCCAAGTTTGAGCAAGTCCTGAGCAACATAAATTCTCTAATCCACAAAATCTCTACCAT CATGACCCAAATGGATTTCTCAGTAAAAATTGAATTTCGCACGCCGATGGTGGAGTCATA CATAGGCAGTGCCATCTTTAAA-AGAAGCTGCGGGGCTACGTCGTCTGCAGGGAATACAA AGAGAGGGTGGCACTGA-
>P.hoffmani_gi|42566671|_PRF_A1
ATGAGGTCAACTGCGCTGCTTATCTTCTTGGTGGTGTCTGTAAGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGAACTGAGCGAAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGTGGAGAGCATGAAGGACAGTGCGGAC TCCCTCCTCACTACCTATCTTAGCTTTCAGGGAGCACCCTTGAGTGACCCAGAGTACCAC CTCCCTGAGGTTAAGGTTGATAACCTGCCCACCGCCGCCATGGACAATGCTACATTTTTG AACCAAACTGATGAAATCCGCCTAAAGAACAACCTCTACTTCTACAGTGCCATTGTCCAA TTCCTGCAGGTCGCCATGACAGAACAGGAAGATCTAAACCC-AGCAGAGGAGTCCCTCAA AGCCAAGTTTCAGGAAGCCCTTGCCAACAGCAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGTAACAGTTACATTTCCCGATCCGCTAGTGGTGCCATA CGAAGGTAGTGCCATCTTTAAACAGAAGCTGCGGGGCGGTGTCATCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTTGCCAAGAAGTA-TCATGAGC AGCTCGTTCAG-CAACAGGAAGAATGA-----
>P. hoffmani_gi|42566683|_PRF_A2
ATGAGGTCAACTGCGCTGCTTATCTTCTTGGTGGTGTCTGTAAGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGATCTGAGCGAAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGTGGAGAGCATGAAGGACAATGCGGAC TCCCTCCTCACTACCTATCTTAGCTTTCAGGGAGCACCCTTGAGTGACCCAGAGTACCAC CTCCCTGAGGTTAAGATTGATAACCTGCCCACCGCCGCCATGGACAATGCTACATTTTTT AACCAAACTGATGAAATCCGCCTAAAGAACAACCTCTACTTCTACAGTGCCATTGTCCAA TTCCTGCAGGTCGCCATGACAGAACAGGAAGATCTAAACCC-AGCAGAGGAGTCCCTCAA AACCAAGTTTCAGGAAGCCATGGCCAACAGCAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGTAACAGTTACATTTCCCGATCCGCTAGTGGTGCCATA CGAAGGTAGTGCCATCTTTAAACAGAAGCTGCGTGGCGGTGTCATCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTTGCCCAGAAGTA-TCATGAGC AGCTCGTTCAG-CAACAAGAACAGGAAGAATGA
>P. hoffmani_gi|42566719|_PRF_B1
ATGAGGTCAACTGCGCTGCTTATCTTCTTGGTGGTGTCTGTAAGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGATCTGAGCGAAACT ACAATTGTCCTTTTCTCGGAGACACAGGAATTTGCGGAGAGCATGAAGGACAATGCGGAC TCCCTCСTCACTACCTATCTTAGCTTACAGGGAGAACCCTACAGCGATCCGGATTTCGGC CTCCCCGATATCCGAATTGCTAACCTGCCCACCGCCACCATGAACTATGATACATTCATG AGGCAAACTGATGAAATCCGCCTGAAAAACAACCTCTACTTCTACACTGCCATCATCAAA TTCCTGCAGCTCACCATGGATGAGCAGAAAGACCTGAACCC-GGCGGACCCATCCCTCAA TGCCAAGTTTGAGCAAGTCCTGAGCAACATAAATGCTCTTATCGACAACATCTCTACCAT CATGACCCAGATGGATTTCTCAGTAAAAATTGAATTTCGCACGCCGATGGTGGAGTCATA CATAGGCAGTGCCATCTTTAAAAAGAAGCTGCGGGGCTACGTCGTCTGCAGGGAATACAA AGAGAGGGTGGCACTGACGATGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TGAAGGGC ACCTCCAGCGGGCAAGACCATGA---------
>P.richmondi_gi|42566679|_PRF_A1
ATGAGGTCAACTGCGCTGCTTATCTTCTTGGTGGTGTCTGTAAGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGAACTAAGCGAAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGTGGAGAGCATGAAGGACAATGCAGAC TCTCTCCTCACTACCTATCTTAGCTTTCAGGGAGCACCCTTGAGTGACCCAGAGTACCAC CTCCCTGAGGTTAAGGTTGATAACCTGCCCACCGCCGCCATGGACAATGCTACATTTTTG AACCAAACTGATGAAATCCGCCTAAAGAACAACCTCTACTTCTACAGTGCCATTGTCCAA TTCCTGCAGGTCGCCATGACAGAACAGGAAGATCTAAACCC-AGCAGAGGAGTCCCTGAA AGCCAAGTTTCAGGAAGCCATGGCCAACATCAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGTAACAGTTACATTTCCCGATCCGCTAGTGGTGCCATA CGAAGGTAGTGCCATCTTTAAACAGAAGCTGCGGGGCGGTGTCATCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTTGCCAAGAAGTA-TCATGAGC AGCTCATTCAG-CAACAGGAACAGGAAGAATGA
>P.richmondi_gi|42566677|_PRF_A2

ATGAGGTCAACTGCGCTGCTTATCTTCTTGGTGGTGTCTGTAAGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGAACTAAGCGAAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGTGGAGAGCATGAAGGACAATGCAGAC TCTCTCСTCACTACCTATCTTAGCTTTCAGGGAGCACCCTTGAGTGACCCAGAGTACCAC CTCCCTGAGGTTAAGGTTGATAACCTGCCCACCGCCGCCATGGACAATGCTACATTTTTG AACCAAACTGATGAAATCCGCCTAAAGAACAACCTCTACTTCTACAGTGCCATTGTCCAA TTCCTGCAGGTCGCCATGACAGAACAGGAAGATCTAAACCC-AGCAGAGGAGTCCCTCAA AGCCAAGTTTCAGGAAGCCATGGCCAACATCAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGTAACAGTTACATTTCCCGATCCGCTAGTGGTGCCATA CGAAGGTAGTGCCATCTTTAAACAGAAGCTGCGTGGCGGTGTCATCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTTGCCAAGAAGTA-TCATGAGC AGCTCATTCAG-CAACAGGAACAGGAAGAATGA >P.richmondi_gi|42566691|_PRF_A3
ATGAGGTCAACTGCGCTGCTTATCTTCTTGGTGGTGTCTGTAAGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGAACTGAGCGAAACT ACAATTGTCCTTTTCTTGGAGACACAGAAATTTGCGGAGAGCATGAAGGACAATGCGGAC TCCCTCCTCACTACCTATCTTAGCTTTCAGGGAGCACCCTTGAGTGACCCAGAGTACCAC CTCCCTGAGGTTAAGGTTGATAACCTGCCCACCGCCGCCATGGACAATGCTACATTTTTG AACCAAACTGATGAAATCCGCCTAAAGAACAACCTCTACTTCTACAGTGCCATTGTCCAA TTCCTGCAGGTCACCATGACAGAACAGGAAGATCTAAACCC-AGCAGAGGAGTCCCTCAA AGCCAAGTTTCAGGAAGCCATGGCCAACATCAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGTAACAGTTACATTTCCCGATCCGCTAGTGGTGCCATA CGAAGGTAGTGCCATCTTTAAACAGAAGCTGCGGGGCGGTGTCATCTGCCAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTTGCCAAGAAGTA-TCATGAGC AGCTCGTTCAG-CAACAGGAACAGGAAGAATGA
>P.richmondi_gi|42566673|_PRF_A4
ATGAGGTCAACTGCGCTGCTTATCTTCTTGGTGGTGTCTGTAAGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGAACTGAGCGAAACT ACAATTGTCCTTTTCTTGGAGACACAGAAATTTGCGGAGAGCATGAAGGACAATGCGGAC TCCCTCCTCACTACCTATCTTAGCTTTCAGGGAGCACCCTTGAGTGACCCAGAGTACCAC CTCCCTGAGGTTAAGGTTGATAACCTGCCCACCGCCGCCATGGACAATGCTACATTTTTG AACCAAACTGATGAAATCCGCCTAAAGAACAACCTCTACTTCTACAGTGCCATTGTCCAA TTCCTGCAGGTCGCCATGACAGAACAGGAAGATCTAAACCC-AGCAGAGGAGTCCCTCAA AGCCAAGTTTCAGGAAGCCATGGCCAACATCAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGTAACAGTTACATTTCCCGATCCGCTAGTGGTGCCATA CGAAGGTAGTGCCATCTTTAAACAGAAGCTGCGTGGCGGTGTCATCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTTGCCAAGAAGTA-TCATGAGC AGCTCATTCAG-CAACAGGAACAGGAAGAATGA
>P.richmondi_gi|42566687|_PRF_A5
ATGAGGGCAACTGCGCTGCTTATCTTCTTGGTGGTGTCTGTAAGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACATGGCAGAACTGAGCGAAACT ACAATTGTCCTTTTCTTGGAGACACAGAAATTTGCGGAGAGCATGAAGGACAATGCGGAC TCССТССTСАСТАССТАTCTTAGCTTTCAGGGAGCACCCTTGAGTGACCCAGAGTACCAC CTCCCTGAGGTTAAGGTTGATAACCTGCCCACCGCCGCCATGGACAATGCTACATTTTTG AACCAAACTGATGAAATCCGCCTAAAGAACAACCTCTACTTCTACAGTGCCATTGTCCAA TTCCTGCAGGTCGCCATGACAGAACAGGAAGATCTAAACCC-AGCAGAGGAGTCCCTCAA AGCCAAGTTTCAGGAAGCCATGGCCAACATCAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGTAACAGTTACATTTCCCGATCCGCTAGTGGTGCCATA CGAAGGTAGTGCCATCTTTAAACAGAAGCTGCGTGGCGGTGTCATCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTTGCCAAGAAGTA-TCATGAGC AGCTCATTCAG-CAACAGGAACAGGAAGAATGA
>P.richmondi_gi|60499234|_PRF_A8
ATGAGGTCAACTGCGCTGCTTATCTTCTTGGTGGTGTCTGTAAGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAAGGTTCAACCAGTTCCGACGTGGCAGAACTGAGCGAAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGTGGAGAGCATGAAGGACAATGCAGAC TCССТССТСАСТАТСТАTCTTAGCTTTCAGGGAGCACCCTTGAGTGACCCAGAATACCAC CTCCCTGAGGTTAAGGTTGATAACCTGCCCACCGCCGCCATGGACAATGCTACATTTTTG AACCAAACTGATGAAATCCGCCTAAAGAACAACCTCTACTTCTACAGTGCCATTGTCCAA TTCCTGCAGGTCGCCATGACAGAACAGGAAGATCTAAACCC-AGCAGAGGAGTCCCTGAA AGCCAAGTTTCAGGAAGCCATGGCCAACATCAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGTAACAGTTACATTTCCCGATCCGCTAGTGGTGCCATA CGAAGGTAGTGCCATCTTTAAACAGAAGCTGCGTGGCGGTGTCATCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTTGCCAAGAAGTA-TCATGAGC AGCTCATTCAG-CAACAGGAACAGGAAGAATGA
>P.richmondi_gi|60499236|_PRF_A9
ATGAGGTCAACTGCGCTGCTTATCTTCTTGGTGGTGTCTGTAAGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGAACTGAGCGAAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGTGGAGAGCATGAAGGACAATGCAGAC TCCCTCCTCACTATCTATCTTAGCTTTCAGGGAGCACCCTTGAGTGACCCAGAATACCAC CTCCCTGAGGTTAAGGTTGATAACCTGCCCACCGCCGCCATGGACAATGCTACATTTTTG AACCAAACTGATGAAATCCGCCTAAAGAACAACCTCTACTTCTACAGTGCCATTGTCCAA TTCCTGCAGGTCGCCATGACAGAACAGGAAGATCTAAACCC-AGCAGAGGAGTCCCTCAA AGCCAAGTTTCAGGAAGCCATGGCCAACATCAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGTAACAGTTACATTTCCCGATCCGCTAGTGGTGCCATA CGAAGGTAGTGCCATCTTTAAACAGAAGCTGCGGGGCGGTGTCATCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTTGCCAAGAAGTA-TCATGAGC AGCTCATTCAG-CAACAGGAACAGGAAGAATGA
>P.richmondi_gi|60499238|_PRF_A10
ATGAGGTCAACTGCGCTGCTTATCTTCTTGGTGGTGTCTGTAAGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGAACTAAGCGAAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGTGGAGAGCATGAAGGACAATGCAGAC TCTCTCCTCACTACCTATCTTAGCTTTCAGGGAGCACCCTTGAGTGACCCAGAGTACCAC CTCCCTGAGGTTAAGGTTGATAACCTGCCCACCGCCGCCATGGACAATGCTACATTTTTG AACCAAACTGATGAAATCCGCCTAAAGAACAACCTCTACTTCTACAGTGCCATTGTCCAA

TTCCTGCAGGTCGCCATGACAGAACAGGAAGATCTAAACCC-AGCAGAGGAGTCCCTGAA AGCCAAGTTTCAGGAAGCCATGGCCAACATCAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGTAACAGTTACATTTCCCGATCCGCTAGTGGTGCCATA CGAAGGTAGTGCCATCTTTAAACAGAAGCTGCGTGGCGGTGTCATCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTTGCCAAGAAGTA-TCATGAGC AGCTCATTCAG-CAACAGGAACAGGAAGAATGA
>P.richmondi_gi|42566715|_PRF_B1
ATGAGGTCAACTGCGCTGCTTATCTTCTTGGTGGTGTCTGTAAGCAGAGCCACCAGCCTG ACAATCTCGACCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGAACTGAGCGCAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGCGGAGAGCATGAAGGACAATGCGGAC TCCCTCCTCACTACCTATCTTAGCTTACAGGGAGAACCCTACAGCAATCCGGATTTCAGC TTCCTCGATATCCAGATTGCTAACCTGCCCACCGCCGCCATGAACTATGATACATTCATG AGGCAAACTGATGAAATCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATCATCAAA TTCCTGCAGCTCACCATGGATGAGCAGAAAGACCTGAACCC-GGCAGACCAATCCATCAC TGCCAAGTTTGAGCAAGTCCTGAGCAACATAAATGCTCTTATCGACAACATCTCTACCAT CATGACCCAGATGGATTTCTCAGTAAAAATTGAATTTCGCACGCCGATGGTGGAGTCATA CATAGGCAGTGCCATCTTTAGAAAGAAGCTGCGGGGCTACATCGTCTGCAGGGAATACAA AGAGAGGGTGGCACTGACGATGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TGAAGGGC ACCTCCAGCGGGCAAGACCATGA-----.-. -
>P.richmondi_gi|42566713|_PRF_B2
ATGAGGTCAACTGCGCTGCTTATCTTCTTGGTGGTGTCTGTAAGCAGAGCCACCAGCCTG ACAATCTCGACCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGAACTGAGCGCAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGCGGAGAGCATGAAGGACAATGCGGAC TCССТССТСАСТАССТАТСTTAGCTTACAGGGAGAACCCTACAGCAATCCGGATTTCAGC TTCCTCGCTATCCAGATTGCTAACCTGCCCACCGCCGCCATGAACTATGATACATTCATG AGGCAAACTGATGAAATCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATCATCAAA TTCCTGCAGCTCACCATGGATGAGCAGAAAGACCTGAACCC-GGCAGACCAATCCATCAC TGCCAAGTTTGAGCAAGTCCTGAGCAACATAAATGCTCTTATCGACAACATCTCTACCAT CATGACCCAGATGGATTTCTCAGTAAAAATTGAATTTCGCACGCCGATGGTGGAGTCATA CATAGGCAGTGCCATCTTTAGAAAGAAGCTGCGGGGCTACATCGTCTGCAGGGAATACAA AGAGAGGGTGGCACTGACGATGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TGAAGGGC ACCTCCAGCGGGCAAGACCATGA---------
>P.richmondi_gi|42566701|_PRF_B3
ATGAGGTCAACTGCGCTGCTTATCTTCTTGGTGGTGTCTGTAAGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGAACTGAGCGAAACT ACAATTGTCCTTTTCTTGGAGACACAGAAATTTGCGGAGAGCATGAAGGACAATGCGGAC TCССТССTСACTACCTATCTTAGCTTACAGGGAGAACCCTACAGCAATCCGGATTTCAGC TTCCTCGCTATCCAGATTGCTAACCTGCCCACCGCCGCCATGAACTATGATACATTCATG AGGCAAACTGATGAAATCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATCATCAAA TTCCTGCAGCTCACCATGGATGAGCAGAAAGACCTGAACCC-GGCAGACCAATCCATCAC TGCCAAGTTTGAGCAAGTCCTGAGCAACATAAATGCTCTTATCGACAACATCTCTACCAT CATGACCCAGATGGATTTCTCAGTAAAAATTGAATTTCGCACGCCGATGGTGGAGTCATA CATAGGCAGTGCCATCTTTAGAAAGAAGCTGCGGGGCTACATCGTCTGCAGGGAATACAA AGAGAGGGTGGCACTGACGATGCGGGATTTTGAGTTTTTTGCCAAGAAGTA-TCATGAGC AGCTCGTTCAG-CAACAGGAACAGGAAGAATGA
>P.richmondi_gi|42566717|_PRF_B4
ATGAGGTCAACTGCGCTGCTTATCTTCTTGGTGGTGTCTGTAAGCAGAGCCACCAGCCTG ACAATCTCGACCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGAACTGAGCGCAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGCGGAGAGCATGAAGGACAATGCGGAC TCCCTCCTCACTACCTATCTTAGCTTACAGGGAGAACCCTACAGCAATCCGGATTTCAGC TTCCTCGCTATCCAGATTGCTAACCTGCCCACCGCCGCCATGAACTATGATACATTCATG AGGCAAACTGATGAAATCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATCATCAAA TTCCTGCAGCTCACCATGGATGAGCAGAAAGACCTGAACCC-GGCAGACCAATCCATCAC TGCCAAGTTTGAGCAAGTCCTGAGCAACATAAATGCTCTTATCGACAACATCTCTACCAT CATGACCCAGATGGATCTCTCAGTAAAAATTGAATTTCGCACGCCGATGGTGGAGTCATA CATAGGCAGTGCCATCTTTAGAAAGAAGCTGCGGGGCTACATCGTCTGCAGGGAATACAA AGAGAGGGTGGCACTGACGATGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TGAAGGGC ACCTCCAGCGGGCAAGACCATGA-
>P.richmondi_gi|60499232|_PRF_B5
ATGAGGTCAACTGCGCTGCTCATCTTCCTGGTGGTGTCTGTACGCAGAGCCACCAGCCTG ACAATCTCGGCCCCTGTCAAACGTTCAACCAGTTCCGACGTGGCAGATCTGAGCGAAACT ACAATTGTCCTTTTCTCGGAGACACAGAAATTTGCGGAGAGCATGAAGGACAATGCGGAC TCССТССТСАСТАССТАТСTTAGCTTACAGGGAGAACCCTACAACAATCCGGATTTCAGC CTCCGCGATATCCAGATTGCTAACCTGCCCACAGCCACCATGAACTATGATACATTCATG AGGCAAACTGATGAAATCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATCATCAAA TTCCTGCAGCTCACCATGGATGAGCAGAAAGACCTGAACCC-GGAGGACCCGTCCCTCAA TGCCAAGTTTGAGCAAGTCCTGAGCAACATAAATTCTCTAATCAACAAAATCTCTACCAT CATGACCCAAATGGATTTCTCAGTAAAAATTGAATTTCGTTCGCCGATGGTGGAGTCATA CATAGGCAGTGCCATCTTTAAAAAGAAGCTGCGGGGCTACGTCGTCTGCAGGGAATACAA AGAGAGGGTGGCACTGACGATGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TGAAGGGC ACCTTCATCGGGCAAGACCATGA--------
>P.websteri_gi|60499262|_PRF_A1
ATGAGGTCAACTGCGCTGCTTACCTTCTTGGTGGTGTTTGTAAGCACAGCCACCAGCATG
 ACCATTGTCCTTTTCTCTGAGACACAGAAAATTGTGGAGGACATGCAGAGCAGTGCTGCG TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGTACCTTTGAGTAATCCAGAGTACCAG СТСССТАATATTAAGATTGAAAACCTGCCCACCGCCGCCATGGACAATGATACATTTTTG AACCATACTGATGAAATCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGACCACCATGGCAGAGCAGGAAGATCTAAACCC-G----- - CAATCCCTCAC TGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTCTGTCAGTAACAATTACATTTCCCGAGCCGCTAGTGGTGCCTTA CGTAGGCAGTGCCTTCTTTAAAAAGAAGCTGCGGGGCGGTGTCATTTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCAGGAGTTTGAGTTTTTTGCCAAGAAGTA-TCATGAGC AGCTCGTTCAG-CCACAGGAAGAATGA---- -
>P.websteri_gi|60499264|_PRF_A2
ATGAGGTCAACTGCGCTGCTTACCTTCTTGGTGGTGTTTTTAAGCACAGCCACCAGCCTG GCAATGGCGGACACTGA $-\cdots-\cdots-\cdots-\cdots$ - $-\cdots$ - $-\cdots$ ACCATTGTCCTTTTCTCTGAGACACAGAAAATTGTGGAGGACATGCAGAGCAGTGCTGCG TCССТССТСССТАССТАTCTTAGCTTTCAGGGAGTACCCTTGAGTAATCCAGAGTACCAG CTCCCTAATATTAAGATTGAAAACCTGCCCACCGCCGCCATGGACAATGATACATTTTTG AACCATACTGATGAAATCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGACCACCATGGCAGAGCAGGAAGATCTAAACCC-G-----CAATCCCTCAC TGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTCTGTCAGTAACAATTACATTTCCCGAGCCGCTAGTGGTGCCTTA CGTAGGCAGTGCCTTCTTTAAAAAGAAGCTGCGGGGCGGTGTCATTTGCAAAGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGAGTTTGAGTTTTTTGCCAAGAAGTA-TCATGAGC AGCTTGTTCAG-CCACAGGAAGAATGA-....-
>P.wehrlei_gi|60499240|_PRF_A1
ATGAGGTCAACCGCGCTGCTCACCTTCTTGGTGGTGTTTGTAAGCACAGCCATCGATCTG ACAATGGCGGACACTGACGAACATTCAACCAGTTCCGATGTGGCAGAATTGAGCAAAGAT ACCATTATCCTTTTCTCTGAGACACAGAAAATTGTAGAGAACATTCAGAGCAATGCTGGT TCCCTCCTCGCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTAATCCAGAGTACCAC CTCCCTGATGTTAAGATTGATAACCTGCCCACCGCCGCCATGGACAATGATACATTTTTG AGCCAAACTGATGAAACTCGCCTGAAGAACAACCTCTACTTCTACAGCGCCATTGTCGAA TTCCTGCAGATTGTCAAGACAGAGCAGGAAGATCTGAACCC-ATTCGCGCTCGACCTCAC TGCCAAGCTTGAGGAAGCCGTGGCCAACTCAACTACTCTTAGCAACAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGTAACAATTGAATTTCCCAAGCCGCTGGTGGTGCCTTA CGAAGGCAGTTCCTACTTTAGAAAGAAGCTGCGGGGCGGTATCATCTGCAAGGAATACAA GGACAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TCAAGGGC AGCTCATTCAG-CAACAGGAAGAATGA-
>P.wehrlei_gi|60499242|_PRF_A2
ATGAGGTCAACCGCGCTGCTCACCTTCTTGGTGGTGTTTGTAAGCACAGCCATCGATCTG ACAATGGCGGACACTGACGAACATTCAACCAGTTCCGATGTGGCAGAATTGAGCAAAGAT ACCATTATCCTTTTCTCTGAGACACAGAAAATTGTAGAGAACATTCAGAGCAATGCTGGT TCCCTCCTCGCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTAATCCAGAGTACCAC CTCCCTGATGTTAAGATTGATAACCTGCCCACCGCCGCCATGGACAATGATACATTTTTG AGCCAAACTGATGAAACTCGCCTGAAGAACAACCTCTACTTCTACAGCGCCATTGTCGAA TTCCTGCAGATTGTCAAGACAGAGCAGGAAGATCTGAACCC-ATTCGCGCTCGACCTCAC TGCCAAGCTTGAGGAAGCCGTGGCCAACTCAACTACTCTTAGCAACAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGTAACAATTGAATTTCCCAAGCCGCTGGTGGTGCCTTA CGAAGGCAGTTCCTACTTTAGAAAGAAGCTGCGGGGCGGTGTCATCTGCAAGGAATACAA GGACAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TCAAGGGC AGCTCATTCAG-CAACAGGAAGAATGA-----
>P.wehrlei_gi|60499244|_PRF_A3
ATGAGGTCAACCGCGCTGCTCACCTTCTTGGTGGTGTTTGTAAGCACAGCCATCGGTCTG ACAATGGCGGACACTGACGAACATTCAACCAGTTCCGATGTGGCAGAATTGAGCAAAGAT ACCATTATCCTTTTCTCTGAGACACAGAAAATTGTAGAGAACATTCAGAGCAATGCTGGT TCCCTCCTCGCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTAATCCAGAGTACCAC CTCCCTGATGTTAAGATTGATAACCTGCCCACCGCCGCCATGGACAATGATACATTTTTG AGCCAAACTGATGAAACTCGCCTGAAGAACAACCTCTACTTCTACAGCGCCATTGTCGAA TTCCTGCAGATTGTCAAGACAGAGCAGGAAGATCTGAACCC-ATTCGCGCTCGACCTCAC TGCCAAGCTTGAGGAAGCCGTGGCCAACTCAACTACTCTTAGCAACAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGTAACAATTGAATTTCCCAAGCCGCTGGTGGTGCCTTA CGAAGGCAGTTCCTACTTTAGAAAGAAGCTGCGGGGCGGTGTCATCTGCAAGGAATACAA GGACAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TCAAGGGC AGCTCATTCAG-CAACAGGAAGAATGA-....-
>P.wehrlei_gi|60499246|_PRF_A4
ATGAGGTCAACCGCGCTGCTCACCTTCTTG---GTGTTTGTAAGCACAGCCATCGGTCTG ACAATGGCGGACACTGACGAACATTCAACCAGTTCCGATGTGGCAGAATTGAGCAAAGAT ACCATTATCCTTTTCTCTGAGACACAGAACATTGTAGAGAACATTCAGAGCAATGCTGGT TCCCTCCTCACTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTAATCCAGACTACCAC CTCCCTGATGTTAAGATTGATAACCTGCCCACCGCCGCCATGGACAATGATACATTTTTG AGCCAAACTGATGAAACTCGCCTGAAGAACAACCTCTACTTCTACAGCGCCATTGTCGAA TTCCTGCAGATTGTCAAGACAGAGCAGGAAGATCTGAACCC-ATTCGCGCTCGACCTCAC TGCCAAGCTTGAGGAAGCCGTGGCCAACTCAACTACTCTTAGCAACAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGTAACAATTGAATTTCCCAAGCCGCTGGTGGTGCCTTA CGAAGGCAGTTCCTACTTTAGAAAGAAGCTGCGGGGCGGTGTCATCTGCAAGGAATACAA GGACAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TCAAGGGC AGCTCATTCAG-CAACAGGAAGAATGA-----
>P.wehrlei_gi|60499248|_PRF_A5
ATGAGGTCAACCGCGCTGCTCACCTTCTTG---GTGTTTGTAAGCACAGCCATCGGTCTG ACAATGGCGGACACTGACGAACATTCAACCAGTTCCGATGTGGCAGAATTGAGCAAAGAT ACCATTATCCTTTTCTCTGAGACACAGAAAATTGTAGAGAACATTCAGAGCAATGCTGGT TCССTCCTCACTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTAATCCAGACTACCAC CTCCCTGATGTTAAGATTGATAACCTGCCCACCGCCGCCATGGACAATGATACATTTTTG AGCCAAACTGATGAAACTCGCCTGAAGAACAACCTCTACTTCTACAGCGCCATTGTCGAA TTCCTGCAGATTGTCAAGACAGAGCAGGAAGATCTGAACCC-ATTCGCGCTCGACCTCAC TGCCAAGCTTGAGGAAGCCGTGGCCAACTCAACTACTCTTAGCAACAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGTAACAATTGAATTTCCCAAGCCGCTGGTGGTGCCTTA CGAAGGCAGTTCCTACTTTAGAAAGAAGCTGCGGGGCGGTGTCATCTGCAAGGAATACAA GGACAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TCAAGGGC AGCTCATTCAG-CAACAGGAAGAATGA-----
>P.welleri_gi|60499250|_PRF_A1
ATGAGGTCAACCGCGCTGCTCACCTTCTTGGTGCTGTCTATA------ GCCACCGGTCTG
 ACCATTGCCCTTTTCTCTGAGACACAGAAAATTGCAGAGGACATGCAGAGCAATGCTGAC TCССТССССАСТАССТАТСTTAGCTTTCAGGGACCACCCTTAAGCCATCCCGAGTACCAC CTCCCTGATATTAAGATTGAAAATCTGCCCACCGCCGCCATGGACTATGATACATTTTTG

AGCCAAACTGATGAAACCCGCCTGAAGAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGCAGATTGTTAAGACAGAGCAGGAAGATCTGAACCC-GTTTGCGCTCGACCTCAC TGCCAAGCTTGAGGAAGCCGTGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCACATGGGTATGTCAGTAACAATTGAATTTCCCGATCCGCTAGTGGTGCCATT CAAAGGCACTGCCTACTTTCAAAAGAAGCTGCGCGGCGGTGTCATCTGCAAGGAATTCAA GGATAGGGTGGATCTGACGAAGCGGGATTTTGAGTTTTTTGCCAAGAAGTA-TGAAGGGC ACCTATAA-
>P.welleri_gi|60499252|_PRF_A2
ATGAGGTCAACCGCGCTGCTCACCTTCTTGGTGGTGTCTATA------ GCCACCGGTCTG ACAATGGCGGACACAGACGAACATTCA------------------GAATTGAGCACAGAT ACCATTGCCCTTTTCTCTGAGACACAGAAAATTGCAGAGGACATGCAGAGCAATGCTGAC TCCCTCСССАСТАССТАТСТTAGCTTTCAGGGACCACCCTTAAGCCATCCCGAGTACCAC CTCCCTGATATTAAGATTGAAAATCTGCCCACCGCCGCAATGGACTATGATACATTTTTG AGCCAAACTGATGAAACCCGCCTGAAGAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGCAGATTGTTAAGACAGAGCAGGAAGATCTGAACCC-GTTTGCGCTCGACCTCAC TGCCAAGCTTGAGGAAGCCGTGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGTAACAATTGAATTTCCCGATCCGCTAGTGGTGCCATT CAAAGGCACTGCCTACTTTCAAAAGAAGCTGCGCGGCGGTGTCATCTGCAAGGAATTCAA GGAGAGGGTGGATCTGACGAAGCGGGATTTTGAGTTTTTTGCCAAGAAGTA-TGAAGGGC GCCTATAA-
>P.welleri_gi|60499254|_PRF_A3
ATGAGGTCAACCGCGCTGCTCACCTTCTTGGTGCTGTCTATA------GCCACCGGTCTG ACAATGGCGGACACAGACGAACATTCA-----------------GAATTGAGCACAGAT ACCATTGCCCTTTTCTCTGAGACACAGAAAATTGCAGAGGACATGCAGAGCAATGCTGAC TCCCTCCCCACTACCTATCTTAGCTTTCAGGGACCACCCTTAAGCCATCCCGAGTACCAC CTCCCTGATATTAAGATTGAAAATCTGCCCACCGCCGCCATGGACTATGATACATTTTTG AGCCAAACTGATGAAACCCGCCTGAAGAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGCAGATTGTTAAGACAGAGCAGGAAGATCTGAACCC-GTTTGCGCTCGACCTCAC TGCCAAGCTTGAGGAAGCCGTGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGTAACAATTGAATTTCCCGATCCGCTAGTGGTGCCATT CAAAGGCACTTCCTACTTTCAAAAGAAGCTGCGCGGCGGTGTCATCTGCAAGGAATTCAA GGAGAGGGTGGATCTGACGAAGCGGGATTTTTGAGTTTTTTGCCAAGAAGTA-TGAAGGGC GCCTATAA -
>P.dorsalis_gi|60499260|_PRF_A1
ATGAGGTCAACCGCGCTGCTCACCTTCTTGCTGGTGTCTATA------GCCACCGGTCTG ACAATGGCGGACACAGACGAACATTCA-----------------GAATTGAGCACAGAT ACCATTGCCCTTTTCTCTGAGATACAGAAAATTGCAGAGGACATGCAGAGCAATGCTGAC TCCCTCCCCACTACCTATCTTAGCTTTCAGGGACCACCCTTAAGTAATGCGGACTACCAC CTCCCTGATATTAAGATTGAAAATCTGCCCACCGCCGCCATGGACTATGATACATTTTTG AGCCAAACTGATGAAACCCGCCTGAAGAACAACCTCTACTTCTACAGTGCCATTGTAGAA TTCCTGCAGATTGTTAAGACAGAGCAGGAAGATCTGAACCC-GTTTGCACTCGACCTCAC TGCCAAGCTTGAGGAAGCCGTGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGAAACAATTGAATTTCCCGATCCGCTAGTGGTGCCATT CAAAGGCACTGCCTACTTTCAAAAGAAGCTGCGCGGCGGTGTTATCTGCAAGGAATTCAA GGAGAGGGTGGATCTGACGAAGCGGGATTTTGAGTTTTTTGCCAAGAAGTA-TGAAGGGC ACCTATAA -
>P.dorsalis_gi|60499258|_PRF_A2
ATGAGGTCAACCGCGCTGCTCACCTTCTTGCTGGTGTCTATA----- -GCCACCGGTCTG ACAATGGCGGACACAGACGAACATTCA------------------GAATTGAGCACAGAT ACCATTGCCCTTTTCTCTGAGATACAGAAAATTGCAGAGGACATGCAGAGCAATGCTGAC TCCCTCCCCACTACCTATCTTAGCTTTTCAGGGACCACCCTTAAGTAATGCGGACTACCAC CTCCCTGATATTAAGATTGAAAATCTGCCCACCGCCGCCATGGACTATGATACATTTTTG AGCCAAACTGATGAAACCTGCCTGAAGAACAACCTCTACTTCTACAGTGCCATTGTAGAA TTCCTGCAGATTGTTAAGACAGAGCAGGAAGATCTGAACCC-GTTTGCACTCGACCTCAC TGCCAAGCTTGAGGAAGCCGTGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGAAACAATTGAATTTCCCGATCCGCTAGTGGTGCCATT CAAAGGCACTGCCTACTTTCAAAAGAAGCTGCGCGGCGGTGTTATCTGCAAGGAATTCAA GGAGAGGGTGGATCTGACGAAGCGGGATTTTGAGTTTTTTGCCAAGAAGTA-TGAAGGGC ACCTATAA-
>P.ventralis_gi|60499256|_PRF_A1
ATGAGGTCAACCGCGCTGCTCACCTTCTTGGTGGTGTCTATA------GCCACCGGTCTG ACAATGGCGGACACAGACGAACATTCA-----------------GAATTGAGCACAGAT ACTATTGCCCTTTTCTCTGAGACACAGAAAATTGCAGAGGACATGCAGAGCAATGCTGAC TCCCTCCCCACTACCTATCTTAGCTTTCAGGGACCACCCTTAAGTAATGCGGACTACCAC CTCCCTGATATTAAGATTGAAAATCTGCCCACCGCCGCCATGGACTATGATACATTTTTG AGCCAAACTGATGAAACCCGCCTGAAGAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGCAGATTGTTAAGACAGAGCAGGAAGATCTGAACCC-GTTTGCGCTCGACCTCAC TGCCAAGCTTGAGGAAGCCGTGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGCCAT CATGACCCAGATGGGTATGTCAGTAACAATTGAATTTCCCGATCCGCTAGTGGTGCCATT CAAAGGCACTGCCTACTTTCAAAAGAAGCTGCGCGGCGGTGTCATCTGCAAGGAATTCAA GGAGAGGGTGGATCTGACGAAGCGGGATTTTTGAGTTTTTTGCCAAGAAGTA-TGAAGGGC ACCTATAA
>P.yonahlossee_gi|42566629|_PRF_A1
ATGAGGTCAACTTCGCTGCTTACCTTCTTGCTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA-----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGTCGCCATGACAGAGCAGGAAGATCTAAACCC-AGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTTCTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTGTTAGCCAAGAAGTA-TCAAGGGC

CCCTCTAA
>P.yonahlossee_gi|42566623|_PRF_A2
ATGAGGTCAACTTCGCTGCTTACCTTCTTGCTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCССТССТСССТАССТАTCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCССTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGTCGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTTCTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTGTTAGCCAAGAAGTA-TCAAGGGC CCCTCTAA -
>P.yonahlossee_gi|42566615|_PRF_A3
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA---------------- - ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCССТССТСССТАССТАTСTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGTCGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTTCTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTGTTAGCCAAGAAGTA-TCAAGGGC CCCTCTAA-
>P.yonahlossee_gi|42566647| PRF A4
ATGAGGTCAACTTCGCTGCTTACCTTCTTGCTGGTGTCTTTAAGCACAGCCACCAGCCTG
 ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGTCGCCATGACAGAGCAGGAAGATCTAAACCC-AGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTTCTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCAGGATTTTGAGTTGTTAGCCAAGAAGTA-TCAAGGGC CCCTCTAA-
>P.kentucki_gi|42566665|_PRF_A1
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCTCCAGCCTG GCAATGGCGGAA-----------------ATTAACGATGTTGCAGATTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTGCAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCССТССТСССТАССТАТСTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGCGCTCGACCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAGTTACATTGCCCAAGCCGCTAGTGGTGCCATA CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTTCTGACGAAGCGGGATTTTGAGTTGTTAGCCAAGAAGTA-TCAAGGGC CCTTCTAA-
>P.kentucki_gi|42566631|_PRF_A2
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAT--------------- - ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC GCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTTATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTAGCCAAGAAGTA-TCAAGGGC CCTTCTAA-
>P.kentucki_gi|42566659|_PRF_A3
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCTCCAGCCTG
 ACCATTGTCCTTTTCTCGGAGGTGCAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCСTCССТАССТATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCCTATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATG AAGCAAACTGATGAAACTCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATA CGAAGGCAGTGCCTACTTTAGAAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA
>P. ouachitae_gi|60499362|_PRF_A1
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA-----------------ACTAACGATGTTGCAGAATTGAGCAGTGAT ACCATTGTCCTTTTCTCAGAGGCACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCССТССТСССТАССТАТСTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG

CTCCCTCATATTAAGGTTGCTAACCTGCCCACCGCCGCCATGGACTATGACACATTCATG AGTCAAAATGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGTCGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACACAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATA CAAAGGCAGTGCCATCTTTGGAAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGATGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA-
>P. ouachitae_gi|60499364|_PRF_A2
ATGAGGTCAACTGCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCGG ACAATGGCGGAA-----------------ACTAACGATGTTGCAGAATTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGACACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCССТССТСССТАССТАТСTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGCTAACCTGCCCACCGCCGCCATGGACTATGACACATTCATG AGTCAAAATGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGTCGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATA CAAAGGCAGTGCCATCTTTAGAAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGATGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA-
>P. ouachitae_gi|60499366|_PRF_A3
ATGAGGTCAACTGCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA-----------------ACTAACGATGTTGCAGAATTGAGCAGTGAT ACCATTGTCCTTTTCTCAGAGGCACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCССТСATATTAAGGTTGCTAACCTGCCCACCGCCGCCATGGACTATGACACATTCATG AGTCAAAATGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGTCGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATA CAAAGGCAGTGCCATCTTTAGAAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGATGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA
>P. ouachitae_gi|60499368|_PRF_A4
ATGAGGTCAACTGCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA---------------ACTAACGATGTTGCAGAATTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGACACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG СТСССТСАTATTAAGGTTGCTAACCTGCCCACCGCCGCCATGGACTATGACACATTCATG AGTCAAAATGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGTCGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATA CAAAGGCAGTGCCATCTTTAGAAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGATGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA -
>P. ouachitae_gi|60499370|_PRF_A5
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA--------------- - ACTAACGATGTTGCAGAATTGAGCAGTGAT ACCATTGTCCTTTTCTCAGAGGCACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCССТССТСССТАССТАTСTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGCTAACCTGCCCACCGCCGCCATGGACTATGACACATTCATG AGTCAAAATGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGTCGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATA CAAAGGCAGTGCCATCTTTAGAAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGATGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA-
>P.fourchensis_gi|60499372|_PRF_A1
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG
 ACCATTGTCCTTTTCTCGGAGACACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCССТСАTATTAAGGTTGCTAACCTGCCCACCGCCGCCATGGACTATGACACATTCATG AGTCAAAATGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGTCGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCACTAGTGGTGCCATA CAAAGGCAGTGCCATCTTTAGAAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGATGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA
>P.fourchensis_gi|60499374|_PRF_A2
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA---------------ACTAACGATGTTGCAGAATTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGACACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG СТСССТСАTATTAAGGTTGCTAACCTGCCCACCGCCGCCATGGACTATGACACATTCATG AGTCAAAATGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGTCGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATA CAAAGGCAGTGCCATCTTTAGAAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA

GGAGATGGTGTTACTGACGAAGCGGGATTTTTGAGTTTTTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA-------------------------
>P.aureolus_gi|42566605|_PRF_A1
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA-----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCССTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTATGTTGTTAGCCGAGAAGTA-TCAAGGGC CTCTCTAA-
>P.aureolus gi| $42566621 \mid$ PRF A2
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCCATGGCGGAA-----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCACCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGTGTTGTTAGCCCAGAAGTA-TCAAGGGC CTCTCTAA
>P.aureolus_gi|42566627|_PRF_A3
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCCATGGCGGAA----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTTCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG СTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTATGTTGTTAGCCGAGAAGTA-TCAAGGGC

ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCCATGGCGGAA-----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTTCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGTGTTGTTAGCCCAGAAGTA-TCAAGGGC CTCTCTAA-
>P.aureolus_gi|60499540|_PRF_A5
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA------------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCACCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGTGTTGTTAGCCCAGAAGTA-TCAAGGGC СТСТСТАА-
>P.mississippi_gi|60499530|_PRF_A1
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA -----------------ATTAACGATGTTGCAGATTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTGCAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTTCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTTCTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTTCTGACGAAGCGGGATTTTGAGTTATTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA
>P.mississippi_gi|60499532|_PRF_A2
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA------------------ATTAACGATGTTGCAGATTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTGCAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC

TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTTCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTTCTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTTCTGACGAAGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA-
>P.mississippi_gi|60499534|_PRF_A3
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA-----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTGCAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCСTСССTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTTCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTTCTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTTCTGACGAAGCGGGATTTTGAGTTATTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA -
>P.mississippi_gi|60499536|_PRF_A4
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTGCAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTTCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTTCTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTTCTGACGAAGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA
>P.mississippi_gi|60499538|_PRF_A5
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA-----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTGCAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCСTСССTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATA CGAAGGCAGTGCCTACTTTAGAAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTTCTGACGAAGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA-
>P.grobmani_gi|42566651|_PRF_A1
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA-----------------ATTAACGATGTTGCAGATTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTGCAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCССTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCCTATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATG AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATA CGAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTTCTGACGAAGCGGGATTTTTGAGTTATTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA-
>P.grobmani_gi|42566641|_PRF_A2
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA------------------ATTAACGATGTTGCAGATTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTGCAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCCTATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATG AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATA CGAAGGCAGTGCCTACTTTAGAAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTTCTGACGAAGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA
>P.savannah_gi|60499376|_PRF_A1
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA------------------ATTAACGATGTTGCAGATTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTGCAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATG AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGTCGCCATGACAGAGCAGGAAGATCTAAACCC-AGCAGCGCTCACCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTTCCCAAGCCGCTAGTGGTGCCATT

CAAAGGCAGTGCCTTCTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTTCTGACGAAGCGGGATTTTGAGTTATTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA-
>P.savannah_gi|60499378|_PRF_A2
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA------------------ATTAACGATGTTGCAGATTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTGCAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGTCGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTTCCCAAGCCGCTAGTGGTGCCATA CGAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCAGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTTCTGACGAAGCGGGATTTTGAGTTATTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA-
>P.savannah_gi|60499380|_PRF_A3
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA-----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTGCAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATG AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGTCGCCATGACAGAGCAGGAAGATCTAAACCC-AGCAGCGCTCACCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTTCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTTCTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTTCTGACGAAGCGGGATTTTGAGTTATTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA
>P.savannah_gi|60499382|_PRF_A4
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA-----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTGCAGAAATTTGCTGAGGACATTCAGAGTAGTGCTGAC TCCCTCСTСССTACCTATCTTAGCTTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGTCGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTTCCCAAGCCGCTAGTGGTGCCATA CGAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTTCTGACGAAGCGGGATTTTGAGTTATTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA
>P.savannah_gi|60499384|_PRF_A5
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA-----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTGCAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGTCGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTTCCCAAGCCGCTAGTGGTGCCATA CGAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTTCTGACGAAGCGGGATTTTGAGTTATTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA
>P.cheoah_gi|42566633|_PRF_A1
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA------------------ATTAACGATGTTGCAGATTTGAGCAGTGAT ACCATTGTCCGTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGCTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATG AAGCAAACTGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CGAAGGCAGTGCCTACTTTAGAAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA-
>P. cheoah_gi|60499386|_PRF_A1
ATGAGGTCAACTTCGCTGCTTACCTTCTTGCTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA-----------------ATTAACGATGTTGCAGATTTGAGCAGTGAT ACCATTGTCCGTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCСTСССTACCTATCTTAGCTTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATG AAGCAAACTGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGCAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CGAAGGCAGTGCCTACTTTAGAAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA-
>P.cheoah_gi|42566613|_PRF_A2
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG
GCAATGGCGGAA----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT

ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAGAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTAGCCAAGAAGTA-TCAAGGGC CCTTCTAA-
>P.cheoah_gi|60499388|_PRF_A2
ATGAGGTCAACTTCGCTGCTTACCTTC̄TTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG
 ACCATTGTCCGTTTCTCGGAGGCACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCССТССТСССТАССТАТСTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCССТСATATTAAGGTTGCTAACCTGCCCACCGCTGCCATGGACTATGATACATTCATG AAGCAAACTGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CGAAGGCAGTGCCTACTTTAGAAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTGTTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA
$>P$. cheoah_gi|42566653|_PRF_A3
ATGAGGTCAACTTCGCTGCTTACCTTCTTGCTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA--------------- -- ATTAACGATGTTGCAGATTTGAGCAGTGAT ACCATTGTCCGTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCССTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGCTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATG AAGCAAACTGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CGAAGGCAGTGCCTACTTTAGAAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA-
------------------------
>P.cheoah_gi|60499390|_PRF_A3
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA-----------------ACTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCGTTTCTCGGAGGCACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCСTCССTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCССТСATATTAAGGTTGCTAACCTGCCCACCGCTGCCATGGACTATGATACATTCATG AAGCAAACTGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CGAAGGCAGTGCCTACTTTAGAAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA
>P.cheoah_gi|60499392|_PRF_A4
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA-----------------ACTAACGATGTTGCAGAATTGAGCAGTGAT ACCATTGTCCGTTTCTCGGAGGCACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG СTСССТСАTATTAAGGTTGCTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATG AAGCAAACTGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CGAAGGCAGTGCCTACTTTAGAAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA-
>P. cheoah_gi|60499394|_PRF_A5
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA----------------ACTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCGTTTCTCGGAGGCACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG СТСССТСАTATTAAGGTTGCTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATG AAGCAAACTGATGAAACCCGCCTGAAGAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CGAAGGCAGTGCCTACTTTAGAAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA -
>P. cheoah_gi|60499396|_PRF_A6
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG
 ACCATTGTCCGTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCССТССТСССТАССТАTСTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGCTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATG AAGCAAACTGATGAAACCCGCCTGAAGAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT

CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CGAAGGCAGTGCCTACTTTAGAAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA-
>P.cheoah_gi|60499398|_PRF_A7
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG
 ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGCTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAGAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CGAAGGCAGTGCCTACTTTAGAAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA-
>P.cheoah_gi|60499400|_PRF_A8
ATGAGGTCAACTTCGCTGCTTACATTCTTGCTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAT--------------- - ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG СТСССТСАTATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGTGTTGTTAGCCCAGAAGTA-TCAAGGGC CCTTCTAA
>P. cheoah_gi|60499402|_PRF_A9
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAGAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTAGCCAAGAAGTA-TCAAGGGC CCCTCTAA-
>P.cheoah_gi|60499404|_PRF_A10
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG
 ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAGAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTAGCCAAGAAGTA-TCAAGGGC CCTTCTAA-
>P.chattahoochee_gi|42566607|_PRF_A1
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA-----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG СТСССТСАTATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTATGTTGTTAGCCGAGAAGTA-TCAAGGGC CTCTCTAA
>P. chattahoochee_gi|60499480|_PRF_A2
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCAACATGGACTATGATGCATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTGCAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTGCGAGGCGGTGTCGTCTGCAAGGAATGCAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTATGTTGTTAGCCGAGAAGTA-TCAAGGGC CTCTCTAA-------------------------
>P.chattahoochee_gi|60499482|_PRF_A3
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG

GCAATGGCGGAA-----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCССТССTСССTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCGCCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTATGTTGTTAGCCGAGAAGTA-TCAAGGGC СТСТСТАА -
>P.chattahoochee_gi|60499484|_PRF_A4
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA-----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCGTCTGCAAGGAATGCAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTATGTTGTTAGCCGAGAAGTA-TCAAGGGC СТСТСТАА-
>P.chattahoochee_gi|60499486|_PRF_A5
ATGAGGCCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA------------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCAACATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTATGTTGTTAGCCGAGAAGTA-TCAAGGGC CTCTCTAA -
>P.chattahoochee_gi|60499488|_PRF_A6
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA-----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCССТССТСССТАССТАТССТАGСTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTATGTTGTTAGCCGAGAAGTA-TCAAGGGC СТСТСТАА
>P.chattahoochee_gi|60499490|_PRF_A7
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA------------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCAACATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAGGCGGGATTTTATGTTGTTAGCCGAGAAGTA-TCAAGGGC СТСТСТАА-
>P.chattahoochee_gi|60499492|_PRF_A8
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA-----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCСTСССTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGCCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAACTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGGCAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTATGTTGTTAGCCGAGAAGTA-TCAAGGGC CTCTCTAA-
>P.cylindraceus_gi|60499314|_PRF_A1
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCСTСССTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAACAGGAAGATCTAAACCC-GGCAGAGCTCACCCTCAA

AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGTGTTGTTAGCCAAGAAGTA-TCAAGGGC CTCTCTAA-
>P.cylindraceus_gi|60499316|_PRF_A2
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG
 ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCСTCССТACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCACCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTAGCCAAGAAGTA-TCAAGGGC CTCTCTAA -
>P.cylindraceus_gi|60499318|_PRF_A3
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA---------------- - ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCССТССТСССТАССТАТСTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCACCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGTGTTGTTAGCCAAGAAGTA-TCAAGGGC CTCTCTAA-
>P.cylindraceus_gi|60499320|_PRF_A4
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA---------------- -ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCACCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTAGCCAAGAAGTA-TCAAGGGC CTCTCTAA-
>P.cylindraceus_gi|60499322|_PRF_A5
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG
 ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCСТСССТАССТАTCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTAGCCAAGAAGTA-TCAAGGGC CTCTCTAA-
>P. teyahalee_gi|60499324|_PRF_A1
ATGAGGTCAACTTCGCTGCTTACATTCTTGCTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAT---------------- - ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCСTCССТACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGTAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGTGTTGTTAGCCCAGAAGTA-TCAAGGGC CTTTCTAA
>P. teyahalee_gi|60499336|_PRF_A1
ATGAGGTCAACTTCGCTGCTTACATTCTTGCTGGTGTCTTTAAGCACAGCCACCAGCCTG
 ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTTCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGTGTTGTTAGCCCAGAAGTA-TCAAGGGC CTTTCTAA-
>P.teyahalee_gi|60499326|_PRF_A2

ATGAGGTCAACTTCGCTGCTTACATTCTTGCTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAT-----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGTGTTGTTAGCCCAGAAGTA-TCAAGGGC CTTTCTAA-
>P.teyahalee_gi|60499328|_PRF_A3
ATGAGGTCAACTTCGCTGCTTACCTTCTTGCTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAT------------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCСTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGTGTTGTTAGCCCAGAAGTA-TCAAGGGC CTTTCTAA-
>P.teyahalee_gi|60499340|_PRF_A3
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAT-----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATCGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGTGTTGTTAGCCCAGAAGTA-TCAAGGGC CTTTCTAA-
>P.teyahalee gil60499330| PRF A4
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAT-----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGTGTTGTTAGCCCAGAAGTA-TCAAGGGC CTTTCTAA
>P.teyahalee_gi|60499332|_PRF_A5
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA-----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCСTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGTGTTGTTAGCCCAGAAGTA-TCAAGGGC CTTTCTAA-
>P.teyahalee_gi|60499334|_PRF_A6
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA------------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCСTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTATGTTGTTAGCCGAGAAGTA-TCAAGGGC CTCTCTAA -
>P. shermani_gi|42566635|_PRF_A1
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA-----------------ATTAACGATGTTGCAGATTTGAGCAGTGAT ACCATTGTCCGTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGCTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATG AAGCAAACTGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA

TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CGAAGGCAGTGCCTACTTTAGAAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA-
>P.shermani_gi|60499438|_PRF_A1
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA---------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTATGTTGTTAGCCGAGAAGTA-TCAAGGGC CCTTCTAA
>P. shermani_gi|42566643|_PRF_A2
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG
 ACCATTGTCCGTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCСTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGCTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATG AAGCAAACTGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CGAAGGCAGTGCCTACTTTAGAAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA-
>P. shermani_gi|60499440|_PRF_A2
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG
 ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCССТСATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGGCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTATGTTGTTAGCCGAGAAGTA-TCAAGGGC CCTTCTAA-
>P shermani gil4256060 | PRF
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG СТСССТСАTATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTATGTTGTTAGCCGAGAAGTA-TCAAGGGC CTCTCTAA
>P. shermani_gi|60499442|_PRF_A3
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA-----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCССТССТСССТАССТАTCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCAACATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGGCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTATGTTGTTAGCCGAGAAGTA-TCAAGGGC CCTTCTAA-
>P. shermani_gi|42566619|_PRF_A4
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG
 ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCССТССТСССTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG СТСССТСАTATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTATGTTGTTAGCCGAGAAGTA-TCAAGGGC CTCTCTAA-
>P.shermani_gi|60499444|_PRF_A4
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG
 ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCСTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTATGTTGTTAGCCGAGAAGTA-TCAAGGGC CTCTCTAA -
P. sherma

ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA----------------ATTAACGATGTTGCAGATTTGAGCAGTGAT ACCATTGTCCGTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGCTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATG AAGCAAACTGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGAAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CGAAGGCAGTGCCTACTTTAGAAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA-
>P.shermani_gi|60499446|_PRF_A5
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG
 ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCССТССТСССТАССТАТСTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCAACATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTATGTTGTTAGCCGAGAAGTA-TCAAGGGC CTCTCTAA-
>P. shermani_gi|42566649|_PRF_A6
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA---------------ATTAACGATGTTGCAGATTTGAGCAGTGAT ACCATTGTCCGTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCССТССТСССTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCССТСATATTAAGGTTGCTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATG AAGCAAACTGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGCGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CGAAGGCAGTGCCTACTTTAGAAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA
>P.shermani_gi|60499448|_PRF_A6
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA-----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCAACATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGGCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTATGTTGTTAGCCCAGAAGTA-TCAAGGGC CTCTCTAA-
>P. shermani_gi|60499450|_PRF_A7
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG
 ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCAACATAGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGGCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTATGTTGTTAGCCCAGAAGTA-TCAAGGGC CCTTCTAA-
>P. shermani_gi|42566611|_PRF_A7
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA----------------- ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCССТССТСССТАССТАTCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT

AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCATCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTTATGTTGTTAGCCGAGAAGTA-TCAAGGGC СТСТСТАА-
>P. shermani_gi|42566617|_PRF_A8
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAC------------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCССТССТСССТАССТАТСТTAGCTTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTATGTTGTTAGCCGAGAAGTA-TCAAGGGC CCTTCTAA -
>P.shermani_gi|60499452|_PRF_A8
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA-----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCAACATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTTATGTTGTTAGCCGAGAAGTA-TCAAGGGC CTCTCTAA -
>P. shermani_gi|42566645|_PRF_A9
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAT-----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCAACATAGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGGCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTATGTTGTTAGCCCAGAAGTA-TCAAGGGC CCTTCTAA
>P. shermani_gi|60499454|_PRF_A9
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAC------------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCСTСССTACCTATCTTAGCTTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCAACATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGGCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTATGTTGTTAGCCGAGAAGTA-TCAAGGGC CCTTCTAA-
>P.shermani_gi|42566661|_PRF_A10
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAT-----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCAACATAGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGGCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTATGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTTATGTTGTTAGCCCAGAAGTA-TCAAGGGC CCTTCTAA-
>P.shermani_gi|60499456|_PRF_A10
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAC-----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCAACATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGGCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTATGTTGTTAGCCGAGAAGTA-TCAAGGGC

CTCTCTAA-
>P.shermani_gi|42566663|_PRF_A11
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAT----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCССТССТСССТАСTTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG СТСССТСАTATTAAGGTTGTTAACCTGCCCACCGCCAACATAGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGGCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTATGTTGTTAGCCCAGAAGTA-TCAAGGGC CCTTCTAA-
>P.shermani_gi|60499458|_PRF_A11
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAC----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCССТССТСССТАССТАTСTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCAACATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGGCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTATGTTGTTAGCCCAGAAGTA-TCAAGGGC CCTTCTAA-
>P.shermani gi|42566667| PRF A12
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG
 ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCATCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTATGTTGTTAGCCGAGAAGTA-TCAAGGGC CCTTCTAA-
>P.shermani_gi|60499460|_PRF_A12
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAC----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCССТССТСССТАССТАТСTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG СТСССТСАTATTAAGGTTGTTAACCTGCCCACCGCCAACATAGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGGCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTATGTTGTTAGCCCAGAAGTA-TCAAGGGC CCTTCTAA-
>P.shermani_gi|60499470|_PRF_A12
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA---------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCACCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTATGTTGTTAGCCGAGAAGTA-TCAAGGGC CTCTCTAA-
>P.shermani gi|60499472| PRF A13
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG
 ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCACCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTATGTTGTTAGCCGAGAAGTA-TCAAGGGC CTCTCTAA-
>P.shermani_gi|60499464|_PRF_A14
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAT-----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG

CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCAACATAGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGGCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTATGTTGTTAGCCGAGAAGTA-TCAAGGGC CTCTCTAA
>P.shermani_gi|60499474|_PRF_A14
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGTACAGCCACCAGCCTG GCAATGGCGGAA---------------- -ACCATTGTCCGTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCССТССТСССТАССТАТСTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCССTCATATTAAGGTTGCTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATG AAGCAAACTGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CGAAGGCAGTGCCTACTTTAGAAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA-
>P.shermani gi|60499466| PRF A15
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG
 ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG СТСССТСАTATTAAGGTTGTTAACCTGCCCACCGCCAACATAGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGGCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATTGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTATGTTGTTAGCCCAGAAGTA-TCAAGGGC CCTTCTAA-
>P. shermani_gi|60499476|_PRF_A15
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA----------------ATTAACGATGTTGCAGATTTGAGCAGTGAT ACCATTGTCCGTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTATGTTGTTAGCCGAGAAGTA-TCAAGGGC CCTTCTAA -
>P. shermani_gi|60499478|_PRF_A16
ATGAGGTCAACTTCGCTGCTTACATTCTTGCTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAT--------------- --ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCССТССТСССTAССТАTСTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGTGTTGTTAGCCCAGAAGTA-TCAAGGGC CTTTCTAA-
>P.jordani_gi|60499308|_PRF_A1
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG
 ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCССТСАTATTAAGGTTGTTAACCTGCCCACCGCCGCCGTGGACTATGATATATTCATA AGGCAAACAGATGAAACCCGCCTGAAGAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGTCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTATCCAAGAAGTA-TCAAGGGC CCTTCTAA-
>P.jordani_gi|60499310|_PRF_A2
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAC---------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCTGCCGTGGACTATGATATATTCATA AGGCAAACAGATGAAACCCGCCTGAAGAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA

GGAGAGGGTGTTCCTGACGAAGCGGGATTTTTGAGTTGTTATCCAAGAAGTA-TCAAGGGC CCTTCTAA--------------------------
>P.jordani_gi|60499312|_PRF_A3
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAC-----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCGTGGACTATGATATATTCATA AGGCAAACAGATGAAACCCGCCTGAAGAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTATCCAAGAAGTA-TCAAGGGC CCTTCTAA-
>P.meridianus_gi|42566639|_PRF_A1
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA------------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCAACATGGACTATGATACATTCATT AGGCAAACAGATGAAGCCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGGCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGAAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGTGTTGTTAACCAATAAGTA-TCAAGGGC CCCTCTAA-
>P.meridianus_gi|42566637|_PRF_A2
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG СTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCAACATGGACTATGATACATTCATT AGGCAAACAGATGAAGCCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGGCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGAAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGTGTTGTTAGCCAATAAGTA-TCAAGGGC CCCTCTAA
>P.meridianus_gi|60499494|_PRF_A3
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA-----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCAGCATGGACTATGATACATTCATT AGGCAAACAGATGAAGCCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGGCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGAAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGTGTTGTTAGCCAATAAGTA-TCAAGGGC CCCTCTAA-
>P.amplus_gi|60499344|_PRF_A1
ATGAGGTCAACTTCTCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA-----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCCCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGGCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTACCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGAAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTAGCCAATAAGTA-TCAAGGGC CCCTCTAA-
>P.amplus_gi|60499346|_PRF_A2
ATGAGGTCAACTTCTCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA -----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGGCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTACCATCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGAAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTAGCCAATAAGTA-TCAAGGGC CCCTCTAA----------------------
>P.amplus_gi|60499348|_PRF_A3
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA-----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC

TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCGTGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGGCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTACCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGAAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTAGCCAATAAGTA-TCAAGGGC CCCTCTAA-
>P.amplus_gi|60499350|_PRF_A4
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA-----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCСTСССTACCTATCTTAGCTTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAGCAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGGCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTACCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGAAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTAGCCAATAAGTA-TCAAGGGC CCCTCTAA-
>P.amplus_gi|60499352|_PRF_A5
ATGAGGTCAATTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA-----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCGTGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGGCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGAAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTAGCCAATAAGTA-TCAAGGGC СССТСТАА-
>P.amplus_gi|60499354|_PRF_A6
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA------------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TСССТССТСССТАССТАТСТTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGGCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTACCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGAAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTAGCCAATAAGTA-TCAAGGGC CCCTCTAA-
>P.amplus_gi|60499356|_PRF_A7
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA-----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCATCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGGCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGAAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTTGAGTTGTTAGCCAATAAGTA-TCAAGGGC CCCTCTAA-
>P.amplus_gi|60499358|_PRF_A8
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCGTGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGGCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGAAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTAGCCAATAAGTA-TCAAGGGC СССТСТАА-------------------------
>P.amplus_gi|60499360|_PRF_A9
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA------------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAGCAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGGCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT

CAAAGGCAGTGCCTACTTTAGAAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTAGCCAATAAGTA-TCAAGGGC CCCTCTAA -
>P.metcalfi_gi|60499266|_PRF_A1
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA-------------- - ACTAACGATGTTGCAGATTTGAGCCGTGAT ACCATTCTCCTTTTCTCGGAGACACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCAACATAGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTAGCCAAGAAGTA-TCAAGGGC CCTTCTAA-
>P.metcalfi_gi|60499300|_PRF_A1
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG
 ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCССТССТСССТАССТАTCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCAACATAGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGGCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTAGCCCAGAAGTA-TCAAGGGC CCTTCTAA -
>P.metcalfi_gi|60499268|_PRF_A2
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAC----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCССТССТСССТАССТАТСTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCGTGGACTATGATATATTCATA AGGCAAACAGATGAAACCCGCCTGAAGAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTATCCAAGAAGTA-TCAAGGGC CCTTCTAA-
>P.metcalfi_gi|60499302|_PRF_A2
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAC-------------- - ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCССТССТСССТАССТАТСTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCAACATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGGCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTAGCCCAGAAGTA-TCAAGGGC CCTTCTAA
>P.metcalfi_gi|60499270|_PRF_A3
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG
 ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAGAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGTCGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAACGGGATTTTGAGTTGTTAGCCAAGAAGTA-TCAAGGGC CCTTCTAA-
>P.metcalfi_gi|60499304|_PRF_A3
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAT---------------- - ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCССТССТСССТАССТАТСTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCAACATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGGCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTAGCCCAGAAGTA-TCAAGGGC CCTTCTAA-
>P.metcalfi_gi|60499272| PRF_A4
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG
GCAATGGCGGAC---------------- - ATTAACGATGTTGCAGACTTGAGCAGTGAT

ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCAACATAGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTAGCCAAGAAGTA-TCAAGGGC CCTTCTAA-
>P.metcalfi_gi|60499306|_PRF_A4
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAC----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCAACATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGGCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTAGCCCAGAAGTA-TCAAGGGC CCTTCTAA
>P.metcalfi_gi|60499274|_PRF_A5
ATGAGGTCAACTTCGCTGCTTACCTTCTTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAC-----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCСTСССTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCAACATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTTCTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA-
>P.metcalfi_gi|60499276|_PRF_A6
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAC------------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG СTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCAACATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGTGTTGTTAGCCAAGAAGTA-TCAAGGGC CCTTCTAA-
>P.metcalfi_gi|60499278|_PRF_A7
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAC------------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCAACATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTATCCAAGAAGTA-TCAAGGGC CCTTCTAA-
>P.metcalfi_gi|60499280|_PRF_A8
ATGAGGTCAACTTCGCTGCTTACCTTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAT-----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCAACATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGTGTTGTTAGCCAAGAAGTA-TCAAGGGC CCTTCTAA -
>P.metcalfi_gi|60499282|_PRF_A9
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAC------------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCССТССТСССTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG СTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCAACATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT

CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGTGTTGTTAGCCAAGAAGTA-TCAAGGGC CCTTCTAA-
>P.metcalfi_gi|60499284|_PRF_A10
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG
 ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCAACATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGTGTTGTTAGCCAAGAAGTA-TCAAGGGC CCTTCTAA-
>P.metcalfi_gi|60499288|_PRF_A12
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAT--------------- - ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG СТСССТСАTATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGTGTTGTTAGCCAAGAAGTA-TCAAGGGC CCTTCTAA-
>P.metcalfi_gi|60499290|_PRF_A13
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAC----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCССТССTСССTAССTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTAGCCAAGAAGTA-TCAAGGGC CCTTCTAA-
>P.metcalfi gi|60499292| PRF A14
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG
 ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTAGCCAAGAAGTA-TCAAGGGC CCTTCTAA-
>P.metcalfi_gi|60499294|_PRF_A15
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAC----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG СТСССТСАTATTAAGGTTGTTAACCTGCCCACCGCCAACATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAAAAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCGCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGAGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGTGTTGTTATCCAAGAAGTA-TCAAGGGC CCTTCTAA
>P.metcalfi_gi|60499296|_PRF_A16
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAC--------------- --ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCAACATGGACTATGATACATTCATT AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTATCCAAGAAGTA-TCAAGGGC CCTTCTAA-
>P.metcalfi gi|60499298| PRF A17
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG

GCAATGGCGGAA-----------------ACTAACGATGTTGCAGATTTGAGCCGTGAT ACCATTCTCCTTTTCTCGGAGACACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCСTСССTACCTATCTTAGCTTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATG AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTTCTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TCAAGGGC AССТСТАА-
>P.montanus_gi|60499496|_PRF_A1
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA-----------------ACTAACGATGTTGCAGATTTGAGCCGTGAT ACCATTCTCCTTTTCTCGGAGACACAGAAATTTGGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCСTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATATATTCTTA AGGCAAACAGATGAAACCCGCCTGAAGAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTGCGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTAGCCAAGAAGTA-TCAAGGGC CCTTCTAA
>P.montanus il
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA-----------------ACTAACGATGTTGCAGATTTGAGCCGTGAT ACCATTCTCCTTTTCTCGGAGACACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TСССТССТСССТАССТАТСTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATG AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTTCTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTAGCCAAGAAGTA - TCAAGGGC ACCTCTAA-
>P.montanus_gi|60499500|_PRF_A3
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA-----------------ACTAACGATGTTGCAGATTTGAGCCGTGAT ACCATTCTCCTTTTCTCGGAGACACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATATATTCTTA AGGCAAACAGATGAAACCCGCCTGAAGAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTATCCAAGAAGTA-TCAAGGGC CCTTCTAA-
>P.montanus_gi|60499502|_PRF_A4
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAA------------------ACTAACGATGTTGCAGATTTGAGCCGTGAT ACCATTCTCCTTTTCTCGGAGACACAGAAATTTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATG AGGCAAACAGATGAAACCCGCCTGAAGAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTGCGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTAGCCAAGAAGTA-TCAAGGGC CCTTCTAA-------------------------
>P.montanus_gi|60499504|_PRF_A5
ATGAGGTCAACTTCGCTGCTTACCTTTTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAC------------------ATTAACGATGTTGTAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG СTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATATATTCTTA AGGCAAACAGATGAAACCCGCCTGAAGAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTGCGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTAGCCAAGAAGTA-TCAAGGGC CCTTCTAA-
>P.montanus_gi|60499506|_PRF_A6
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAC----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGACGTGGACTATGATACATTCATA AGGCAAACAGATGAAACCCGCCTGAAGAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA

AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTATCCAAGAAGTA-TCAAGGGC CCTTCTAA-
>P.montanus_gi|60499508|_PRF_A7
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG
 ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCСTCССТACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGACGTGGACTATGATACATTCATA AGGCAAACAGATGAAACCCGCCTGAAGAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTATCCAAGAAGTA-TCAAGGGC TCTTCTAA-
>P.montanus_gi|60499510|_PRF_A8
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAC---------------- - ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCСTСССТАССТАТСTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGACGTGGACTATGATACATTCATA AGGCAAACAGATGAAACCCGCCTGAAGAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTGCGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTAGCCAAGAAGTA-TCAAGGGC CCTTCTAA-
>P.montanus_gi|60499512|_PRF_A9
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAC---------------- - ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCСTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGACGTGGACTATGATACATTCATA AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTTCTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA-
>P.montanus gil60499514| PRF 10
>P.montanus gi|60499514| PRF A10
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG
 ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCССТССТСССТАССТАТСTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGACGTGGACTATGATACATTCATA AGGCAAACAGATGAAACCCGCCTGAAGAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT TATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTAGCCAAGAAGTA-TCAAGGGC CCTTCTAA -
>P.montanus_gi|60499516|_PRF_A11
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAC----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAATCTGCCCACCGCCGACGTGGACTATGATACATTCATA AGGCAAACAGATGAAACCCGCCTGAAGAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTGCGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTAGCCAAGAAGTA-TCAAGGGC CCTTCTAA-
>P.montanus gi|60499518| PRF A12
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG
 ACCATTCTCCTTTTCTCGGAGACACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCССТССТСССТАССТАТСTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATG AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTGCGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTAGCCAAGAAGTA-TCAAGGGC CCTTCTAA-
>P.montanus_gi|60499520|_PRF_A13

ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG
 ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATATATTCTTA AGGCAAACAGATGAAACCCGCCTGAAGAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTTCTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA -
>P.montanus_gi|60499522|_PRF_A14
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAC----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATG AGGCAAACAGATGAAACCCGCCTGAAGAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTATCCAAGAAGTA-TCAAGGGC CCTTCTAA-----------------------
$>$ P.montanus gi| $60499524 \mid$ PRF_A15
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAC----------------ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATATATTCTTA AGGCAAACAGATGAAACCCGCCTGAAGAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTATCCAAGAAGTA-TCAAGGGC CCTTCTAA-
>P.montanus gi|60499526| PRF A16
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG
 ACCATTGTCCTTTTCTCGGAGGTACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCCCTCCTCCCTACCTATCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCGG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATATATTCTTA AGGCAAACAGATGAAACCCGCCTGAAGAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTACTTTAGTAAGAAGCTGCGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTCCTGACGAAGCGGGATTTTGAGTTGTTAGCCAAGAAGTA-TCAAGGGC CCTTCTAA-
>P.montanus_gi|60499528|_PRF_A17
ATGAGGTCAACTTCGCTGCTTACCTTCTTGGTGGTGTCTTTAAGCACAGCCACCAGCCTG GCAATGGCGGAC---------------- ATTAACGATGTTGCAGACTTGAGCAGTGAT ACCATTGTCCTTTTCTCGGAGACACAGAAATTTGCTGAGGACATTCAGAGCAGTGCTGAC TCССТССТСССТАССТАTCTTAGCTTTCAGGGAGCACCCTTAAGTGATCCAGACTACCAG CTCCCTCATATTAAGGTTGTTAACCTGCCCACCGCCGCCATGGACTATGATACATTCATG AGGCAAACAGATGAAACCCGCCTGAACAACAACCTCTACTTCTACAGTGCCATTGTCGAA TTCCTGAAGGAAGCCATGACAGAGCAGGAAGATCTAAACCC-GGCAGAGCTCTCCCTCAA AGCCAAGTTTGAGGAAGCCATGGCCAACTCAAATACTCTTATCAGCAAAATCTCTGACAT CATGACCCAGATGGGTATGTCAGTAACAATTACATTGCCCAAGCCGCTAGTGGTGCCATT CAAAGGCAGTGCCTTCTTTAGTAAGAAGCTACGGGGCGGTGTCGTCTGCAAGGAATACAA GGAGAGGGTGTTACTGACGAAGCGGGATTTTGAGTTTTTAGCCAAGAAGTA-TCAAGGGC ACCTCTAA--------------------------

## APPENDIX THREE

Contents Page
Figure A3.1: Read quality trimming and adapter clipping ..... 252
Figure A3.2: Distribution of TransRate contig scores ..... 253
Table A3.1: The frequency and total expression of each unique BLAST hit ..... 254
Custom Bioinformatic Scripts (Chapter Four). ..... 284
De-duplicate.sh ..... 284
Error-correct.sh. ..... 284
Trim-Clip-Assembly-Pipeline.pl. ..... 285
Filter-Reads.pl. ..... 290
Parse-Assembly-Stats.pl ..... 294
ReadLengths.R. ..... 296
Expression-Pipeline.pl. ..... 297
Annotate-Pipeline.pl ..... 301
Trinotate-Report.pl ..... 304
Calculate_EXN50.pl ..... 306
GO-enrichment.sh. ..... 307
Identify-Expression-Results.pl ..... 308
ExtractContigs.pl ..... 310
ExtractLongest.pl ..... 312
Find-Secreted.pl ..... 313
Find-Keywords.pl ..... 314
Cluster-Annotations.pl ..... 315


Figure A3.1. Read quality trimming and adapter clipping. The PHRED quality score distributions for each position in reads (from $0-100 \mathrm{bp}$ ) are summarized before trimming and clipping (panel A), compared to after trimming and clipping (panel B). The distribution of quality scores at each position are indicated by the median (solid line), inner quartiles (boxes), outer quartiles (whiskers), and the minimum value (open circles). The horizontal dashed line (at $\mathrm{Q}=5$ ) represents the quality trimming cutoff. The right (red) axis and its corresponding line demonstrate the number of reads at each corresponding read length (base position). Quality trimming, and adapter clipping in particular, led to improved and less variable quality scores, but comes at the cost of a reduction in read count and length.


Figure A3.2. Distribution of TransRate contig scores. The number of contigs (out of $94,221)$ per bin ( 0.01 ) that were assigned the corresponding contig score $(0-1)$. The distribution is skewed by a large number of low-scoring contigs. 74,041 contigs scored higher than the cutoff ( 0.03295 ; vertical dashed line) after optimization.

Table A.3.1. The frequency and total expression of each unique BLAST homology. Each unique BLAST hit (from all BLAST searches) was extracted from the transcriptome annotation report. For each homology, the number of transcripts with a significant hit to it was tallied ("Transcripts"), and the TMM value for the corresponding transcript was added to a cumulative tally ("Expression"). In this way, the total expression of each unique homology could be compared across multiple transcripts. This table is sorted in descending order of the total level of expression per BLAST hit.

| Expression | Transcripts | BLAST Hit |
| :---: | :---: | :---: |
| 21782.38 | 10 | Lectin |
| 19885.09 | 47 | Galectin \{ECO:0000256\|RuleBase:RU102079\} |
| 16943.29 | 36 | Galectin-3 |
| 14526.85 | 12 | Avidin-related protein 4/5 |
| 14301.26 | 45 | phospholipase A2 inhibitor and Ly6/PLAUR domain-containing protein |
| 11762.09 | 7 | Mucin-19 |
| 11735.99 | 5 | Serine protease inhibitor Kazal-type 4 |
| 11123.84 | 1 | Hemoglobin subunit beta-1 |
| 11123.84 | 1 | Beta-globin \{ECO:0000313\|EMBL:BAE47565.1\} |
| 9792.12 | 3 | Avidin-related protein 3 \{ECO:0000313\|EMBL:AFP11326.1\} |
| 8299.84 | 7 | Trypsin inhibitor CITI-1 |
| 8079.94 | 1 | Pancreatic secretory trypsin inhibitor \{ECO:0000313\|EMBL:ACO13237.1\} |
| 7399.50 | 1 | Mucin apoprotein \{ECO:0000313\|EMBL:DAA05596.1\} |
| 7115.01 | 3 | Prokineticin 2 variant 1S/2/4 \{ECO:0000313\|EMBL:AAV73831.1\} |
| 7069.27 | 6 | Kazal-type serine proteinase inhibitor \{ECO:0000313\|EMBL:CAI46283.1\} |
| 6624.41 | 11 | Ovomucoid |
| 6047.29 | 4 | Sodefrin-like protein \{ECO:0000313\|EMBL:ACB54667.1\} |
| 6037.85 | 4 | Prokineticin Bm8-d |
| 4950.21 | 1 | Creatine kinase, muscle \{ECO:0000313\|EMBL:AAH64186.1\} |
| 4950.21 | 1 | Creatine kinase M-type |
| 4343.85 | 1 | Submandibular gland protein C \{ECO:0000313\|Ensembl:ENSMUSP00000085907\} |
| 4343.85 | 1 | Integumentary mucin B. 1 |
| 4207.49 | 4 | Insulin-like peptide INSL5 |
| 3618.78 | 6 | Fucolectin-4 |
| 3586.54 | 2 | Mgc108129 protein \{ECO:0000313\|EMBL:AAI57779.1\} |
| 3586.54 | 2 | Beta-enolase |
| 3288.20 | 7 | Brevican core protein |
| 3153.19 | 1 | Translationally-controlled tumor protein homolog |
| 3124.82 | 4 | Pancreatic secretory trypsin inhibitor |
| 3056.86 | 5 | Xenoxin-1 |
| 2999.52 | 1 | Histone H2AX |
| 2999.52 | 1 | Histone H2A \{ECO:0000256\|RuleBase:RU003556\} |
| 2932.00 | 2 | Sodefrin-like protein \{ECO:0000313\|EMBL:ACB54665.1\} |
| 2901.00 | 4 | PMF Class III-like B variant 1 \{ECO:0000313\|EMBL:AHL39271.1\} |
| 2855.11 | 2 | Protein S100-A11 |
| 2839.99 | 1 | P26olf \{ECO:0000313\|EMBL:BAA34388.1\} |
| 2830.75 | 11 | Pancreatic secretory trypsin inhibitor \{ECO:0000313\|EMBL:ACM09634.1\} |
| 2768.63 | 6 | Galectin-1 |
| 2732.95 | 6 | Glutathione peroxidase \{ECO:0000256\|RuleBase:RU000499\} |
| 2729.87 | 5 | Glutathione peroxidase 3 |
| 2704.79 | 6 | Lysozyme C |
| 2623.47 | 4 | Phospholipase A2 inhibitor subunit gamma B |
| 2611.98 | 2 | Ranaspumin-4 \{ECO:0000313\|EMBL:AAP48833.1\} |
| 2506.87 | 2 | Serine protease inhibitor Kazal-type 4 \{ECO:0000313\|EMBL:EFV59031.1\} |
| 2402.03 | 2 | Fucolectin-5 |
| 2331.51 | 4 | Annexin \{ECO:0000256\|RuleBase:RU003540\} |
| 2293.61 | 5 | Serine protease inhibitor Kazal-type 6 |
| 2289.73 | 2 | Annexin A1 |
| 2250.62 | 7 | Triosephosphate isomerase \{ECO:0000256\|RuleBase:RU000517\} |
| 2250.62 | 7 | Triosephosphate isomerase |
| 2187.68 | 5 | WAP four-disulfide core domain protein 5 |
| 2120.91 | 5 | Putative gastrointestinal growth factor xP4 |
| 2108.96 | 9 | Serine protease inhibitor Kazal-type 2 |
| 2026.09 | 3 | C-type lectin lectoxin-Enh7 |
| 1930.41 | 3 | Sodefrin-like factor \{ECO:0000313\|EMBL:AAZ06331.1\} |
| 1913.19 | 5 | Sodefrin-like protein \{ECO:0000313\|EMBL:ACB54672.1\} |
| 1870.74 | 1 | Serine protease inhibitor Kazal-type 12 |
| 1799.70 | 2 | Protein S100 \{ECO:0000256\|RuleBase:RU361184\} |
| 1784.88 | 20 | Metalloproteinase inhibitor 1 |
| 1784.58 | 1 | Protein S100-A4 |
| 1743.37 | 2 | Avidin-related protein 4/5 \{ECO:0000313\|EMBL:KGL82208.1\} |
| 1740.10 | 1 | Protein S100-A16 |
| 1664.01 | 1 | Troponin T type 3 (Skeletal, fast) \{ECO:0000313\|EMBL:AEZ53893.1\} |


| 1664.01 | 1 | Troponin T, fast skeletal muscle |
| :---: | :---: | :---: |
| 1634.40 | 6 | UPF0764 protein C16orf89 homolog |
| 1585.35 | 7 | Neurocan core protein |
| 1576.88 | 2 | Keratin, type I cytoskeletal 12 |
| 1556.09 | 8 | Sodefrin-like factor \{ECO:0000313\|EMBL:AAZ06333.1\} |
| 1526.81 | 14 | Antileukoproteinase |
| 1395.98 | 1 | Prokineticin Bm8-a |
| 1395.98 | 1 | Prokineticin-1 \{ECO:0000313\|EMBL:AHH39740.1\} |
| 1384.36 | 1 | Protein C3orf33 homolog |
| 1384.36 | 1 | Leukocyte cell-derived chemotaxin-2 \{ECO:0000313\|EMBL:EHB13820.1\} |
| 1384.36 | 1 | Leukocyte cell-derived chemotaxin-2 |
| 1376.35 | 3 | Sodefrin-like factor \{ECO:0000313\|EMBL:AAZ06327.1\} |
| 1364.86 | 1 | Elongation factor 1-beta \{ECO:0000313\|EMBL:EPY73586.1\} |
| 1364.86 | 1 | Elongation factor 1-beta |
| 1342.44 | 2 | Pvalb protein \{ECO:0000313\|EMBL:AAH91031.1\} |
| 1342.44 | 2 | Parvalbumin beta |
| 1329.14 | 10 | Deleted in malignant brain tumors 1 protein |
| 1323.31 | 2 | Antileukoproteinase \{ECO:0000313\|EMBL:ELW61671.1\} |
| 1266.63 | 2 | C-type lectin mannose-binding isoform |
| 1262.64 | 1 | F-type lectin \{ECO:0000313\|EMBL:BAK38714.1\} |
| 1259.52 | 5 | Protein RoBo-1 |
| 1238.13 | 5 | Prosaposin |
| 1224.04 | 10 | Lectin-Cha-1 \{ECO:0000313\|EMBL:AG197186.1\} |
| 1199.71 | 1 | Alpha- N -acetylgalactosamine-specific lectin \{ECO:0000303\|PubMed:11886841\} |
| 1198.50 | 4 | Intelectin-1b |
| 1194.10 | 1 | Protein S100-A13 |
| 1148.00 | 2 | Parvalbumin alpha |
| 1139.39 | 1 | FBP32 \{ECO:0000313\|EMBL:ABB29997.1\} |
| 1114.74 | 1 | Serine peptidase inhibitor, Kazal type 2 (Acrosin-trypsin inhibitor) \{ECO:0000313\|EMBL:AAI09720.1\} |
| 1096.65 | 4 | Galectin-4 |
| 1092.62 | 5 | Sodefrin-like protein \{ECO:0000313\|EMBL:ACB54666.1\} |
| 1088.42 | 60 | Titin |
| 1086.54 | 2 | Apolipoprotein C-I |
| 1076.77 | 1 | Anterior gradient protein 2 \{ECO:0000313\|EMBL:AGH15795.1\} |
| 1076.77 | 1 | Anterior gradient protein 2 \{ECO:0000312\|EMBL:CAJ82848.1\} |
| 1055.29 | 1 | Serine protease inhibitor Kazal-type 6 \{ECO:0000313\|EMBL:EHB11446.1\} |
| 1025.35 | 1 | Fructose-bisphosphate aldolase \{ECO:0000256\|RuleBase:RU003994\} |
| 1025.35 | 1 | Fructose-bisphosphate aldolase A |
| 969.69 | 4 | WAP four-disulfide core domain protein 18 |
| 959.18 | 3 | Grifin |
| 932.49 | 4 | Phospholipase A2 inhibitor subunit gamma B \{ECO:0000313\|EMBL:EMP42220.1\} |
| 910.46 | 1 | 60S ribosomal protein L3 |
| 903.18 | 4 | Secretory leukocyte peptidase inhibitor \{ECO:0000313\|EMBL:ABO52972.1\} |
| 903.02 | 2 | Larval beta-globin \{ECO:0000313\|EMBL:BAA86390.1\} |
| 903.02 | 2 | Hemoglobin subunit beta |
| 901.65 | 11 | Titin \{ECO:0000313\|EMBL:EMP26366.1\} |
| 887.13 | 1 | Hypothetical LOC496414 \{ECO:0000313\|EMBL:AAH82718.1\} |
| 887.13 | 1 | Collagen alpha-1(I) chain |
| 886.87 | 1 | Serpin B5 |
| 871.27 | 11 | Scavenger receptor cysteine-rich type 1 protein M130 |
| 857.94 | 2 | Protein RoBo-1 \{ECO:0000313\|EMBL:KFO37849.1\} |
| 840.73 | 1 | Predicted protein \{ECO:0000313\|EMBL:EDO34925.1\} |
| 833.88 | 3 | Thioredoxin |
| 832.09 | 3 | SPARC |
| 831.14 | 2 | Thioredoxin \{ECO:0000256\|PIRNR:PIRNR000077\} |
| 806.01 | 1 | Galectin-3 \{ECO:0000313\|EMBL:KFP31188.1\} |
| 766.13 | 2 | Sodefrin-like factor \{ECO:0000313\|EMBL:AAZ06336.1\} |
| 763.92 | 1 | Ovomucoid \{ECO:0000313\|EMBL:KGL87148.1\} |
| 743.40 | 1 | Calreticulin \{ECO:0000313\|EMBL:AAS49610.1\} |
| 743.40 | 1 | Calreticulin |
| 732.74 | 7 | Beta-microseminoprotein |
| 726.61 | 2 | Protein notum homolog |
| 688.76 | 4 | Cofilin-1 |
| 675.28 | 1 | Protein notum-like protein \{ECO:0000313\|EMBL:AFO95640.1\} |


| 670.08 | 1 | Putative extracellular matrix protein 1 \{ECO:0000313\|EMBL:JAA48164.1\} |
| :---: | :---: | :---: |
| 670.08 | 1 | Extracellular matrix protein 1 \{ECO:0000313\|EMBL:EMP27941.1\} |
| 670.08 | 1 | Extracellular matrix protein 1 |
| 654.57 | 1 | Keratin, type I cytoskeletal 24 |
| 645.39 | 3 | Integumentary mucin C. 1 |
| 624.51 | 1 | Nascent polypeptide-associated complex subunit alpha \{ECO:0000313\|EMBL:ACQ57932.1\} |
| 624.51 | 1 | Nascent polypeptide-associated complex subunit alpha |
| 614.55 | 3 | Galectin-9 |
| 613.86 | 1 | Sodefrin-like factor \{ECO:0000313\|EMBL:AAZ06325.1\} |
| 605.82 | 6 | Sodefrin-like protein \{ECO:0000313\|EMBL:ACB54668.1\} |
| 582.95 | 1 | Serine protease inhibitor Kazal-type 2-like protein \{ECO:0000313\|EMBL:AFP12929.1\} |
| 550.59 | 2 | Fucolectin-1 |
| 541.96 | 2 | Collagen alpha-1(III) chain |
| 541.69 | 5 | LOC100145442 protein \{ECO:0000313\|EMBL:AAI61058.1\} |
| 540.46 | 1 | Translocon-associated protein subunit delta |
| 529.33 | 1 | Collagen alpha-1(III) chain \{ECO:0000313\|EMBL:EMP26256.1\} |
| 529.28 | 1 | Tetraspanin \{ECO:0000256\|RuleBase:RU361218\} |
| 529.28 | 1 | CD9 antigen |
| 529.19 | 4 | Olfactomedin-4 |
| 526.41 | 8 | Aggrecan core protein |
| 517.96 | 2 | Mitochondrial ATP synthase FO complex subunit c isoform 3 \{ECO:0000313\|EMBL:ABF22459.1\} |
| 517.96 | 2 | ATP synthase F(0) complex subunit C2, mitochondrial |
| 514.98 | 1 | Prod 1 \{ECO:0000313\|EMBL:ABV29331.1\} |
| 513.16 | 4 | Cysteine-rich venom protein pseudechetoxin-like |
| 510.36 | 1 | Kidney androgen-regulated protein |
| 510.36 | 1 | 60 r ribosomal protein L21 |
| 508.89 | 2 | Ovostatin \{ECO:0000313\|EMBL:EMP24416.1\} |
| 500.19 | 1 | Galectin-3 \{ECO:0000313\|EMBL:KGL84865.1\} |
| 487.94 | 1 | Cysteine-rich secretory protein 2 \{ECO:0000313\|EMBL:KFM14003.1\} |
| 485.44 | 6 | Putative gastrointestinal growth factor xP1 |
| 472.16 | 3 | Deleted in malignant brain tumors 1 protein \{ECO:0000313\|EMBL:KFV53424.1\} |
| 471.81 | 1 | Phospholipase A2 inhibitor 25 kDa subunit \{ECO:0000313\|EMBL:KFQ32492.1\} |
| 458.19 | 1 | Ovostatin \{ECO:0000313\|EMBL:EMP41115.1\} |
| 458.19 | 1 | Ovostatin |
| 453.38 | 3 | Beta-microseminoprotein \{ECO:0000313\|EMBL:KFV58775.1\} |
| 415.34 | 3 | Cystatin |
| 410.61 | 2 | Glucose-6-phosphate isomerase \{ECO:0000256\|RuleBase:RU000612\} |
| 410.61 | 2 | Glucose-6-phosphate isomerase |
| 404.98 | 1 | Carbonic anhydrase 6 |
| 389.99 | 2 | Proteinase inhibitor PSKP-2 |
| 382.12 | 3 | Intelectin-2 |
| 381.96 | 4 | Lactoperoxidase |
| 378.19 | 1 | 14-3-3 protein epsilon |
| 369.86 | 2 | Ovomucoid \{ECO:0000313\|EMBL:KFW89191.1\} |
| 369.65 | 1 | Alpha-actinin-2 |
| 352.71 | 2 | Intelectin-1a \{ECO:0000313\|EMBL:EMP30013.1\} |
| 349.45 | 2 | Cofilin-2 \{ECO:0000313\|EMBL:ELK33269.1\} |
| 349.45 | 2 | Cofilin-2 |
| 345.58 | 1 | Interleukin enhancer-binding factor 2 |
| 345.37 | 4 | Beta-2-microglobulin |
| 336.43 | 2 | Antigen WC1.1 |
| 335.50 | 1 | Probable cytosolic oligopeptidase A |
| 332.34 | 1 | LOC100036914 protein \{ECO:0000313\|EMBL:AAI29646.1\} |
| 332.34 | 1 | Cysteine-rich venom protein LEI1 |
| 332.34 | 1 | Cysteine-rich venom protein ablomin |
| 331.03 | 4 | Fish-egg lectin |
| 325.98 | 2 | Beta-2 microglobulin \{ECO:0000313\|EMBL:AAC64996.1\} |
| 321.06 | 2 | Collagen alpha-1(XII) chain |
| 318.82 | 2 | AVIToxin-VAR1 \{ECO:0000313\|EMBL:KFP85430.1\} |
| 310.72 | 1 | T-complex protein 1 subunit epsilon |
| 302.40 | 1 | Phospholipase A2 inhibitor subunit gamma B \{ECO:0000313\|EMBL:KFV87582.1\} |
| 302.38 | 2 | Phospholipase A2, membrane associated |
| 302.07 | 2 | Cystatin-C |
| 297.91 | 5 | Cysteine-rich venom protein |


| 293.53 | 2 | C-type lectin lectoxin-Enh2 |
| :---: | :---: | :---: |
| 288.51 | 1 | Translocon-associated protein subunit alpha \{ECO:0000313\|EMBL:JAB53080.1\} |
| 288.51 | 1 | Translocon-associated protein subunit alpha \{ECO:0000313\|EMBL:AFH33747.1\} |
| 288.51 | 1 | Translocon-associated protein subunit alpha |
| 287.92 | 3 | Scavenger receptor cysteine-rich type 1 protein M130 \{ECO:0000313\|Ensembl:ENSP00000445438\} |
| 286.68 | 2 | Cathepsin L1 |
| 281.83 | 4 | Malate dehydrogenase, cytoplasmic |
| 278.16 | 1 | Translocon-associated protein subunit beta \{ECO:0000256\|PIRNR:PIRNR016400\} |
| 278.16 | 1 | Translocon-associated protein subunit beta |
| 277.44 | 1 | PMF Class III-like A variant 4 \{ECO:0000313\|EMBL:AHL39267.1\} |
| 274.81 | 7 | Cathelicidin 1 \{ECO:0000313\|EMBL:ABV01938.1\} |
| 274.81 | 7 | Cathelicidin-1 |
| 272.28 | 1 | Calsequestrin \{ECO:0000256\|RuleBase:RU000648\} |
| 272.28 | 1 | Calsequestrin-1 |
| 271.30 | 2 | Cysteine and glycine-rich protein 1 |
| 269.80 | 1 | Protein disulfide-isomerase \{ECO:0000256\|RuleBase:RU361130\} |
| 269.80 | 1 | Protein disulfide-isomerase A3 |
| 268.65 | 2 | Plethodontid modulating factor \{ECO:0000313\|EMBL:AEO22653.1\} |
| 261.59 | 3 | Poly(U)-specific endoribonuclease |
| 259.78 | 3 | Cathepsin B |
| 258.90 | 1 | Malate dehydrogenase \{ECO:0000256\|RuleBase:RU003405\} |
| 257.43 | 5 | Immunoglobulin superfamily member 3 |
| 252.33 | 2 | Cathepsin B \{ECO:0000313\|EMBL:EMP27213.1\} |
| 251.71 | 2 | Galectin-9C |
| 241.28 | 3 | Cardiotrophin-2 |
| 238.75 | 1 | Type-2 ice-structuring protein |
| 237.34 | 1 | Poly(U)-specific endoribonuclease \{ECO:0000313\|EMBL:KFV81547.1\} |
| 236.88 | 4 | Kunitz-type protease inhibitor 2 |
| 234.79 | 3 | Resistin |
| 232.22 | 2 | Serine protease inhibitor Kazal-type 6 \{ECO:0000313\|EMBL:KFV03117.1\} |
| 232.04 | 1 | Transmembrane emp24 domain trafficking protein 2 \{ECO:0000313\|EMBL:AAH50165.1\} |
| 232.04 | 1 | Transmembrane emp24 domain-containing protein 2 |
| 232.02 | 2 | Putative Kunitz-type serine protease inhibitor |
| 226.92 | 11 | C3 variant 3 \{ECO:0000313\|EMBL:AHL39255.1\} |
| 224.36 | 1 | Receptivity factor PRFA1 \{ECO:0000313\|EMBL:AAX21912.1\} |
| 224.28 | 1 | LOC100145517 protein \{ECO:0000313\|EMBL:AAI61188.1\} |
| 221.47 | 1 | Superoxide dismutase [Cu-Zn] \{ECO:0000256\|RuleBase:RU000393\} |
| 221.47 | 1 | Extracellular superoxide dismutase [ $\mathrm{Cu}-\mathrm{Zn}$ ] |
| 221.15 | 2 | WAP four-disulfide core domain protein 8 \{ECO:0000313\|EMBL:ELK04241.1\} |
| 219.76 | 1 | Cysteine-rich secretory protein 3 \{ECO:0000313\|EMBL:ELK00219.1\} |
| 219.76 | 1 | Cysteine-rich secretory protein 3 |
| 219.67 | 1 | Phospholipase A2, group IIA (Platelets, synovial fluid) \{ECO:0000313\|EMBL:AAI22707.1\} |
| 218.79 | 1 | Splicing factor, proline- and glutamine-rich |
| 215.07 | 2 | Complement component 1 Q subcomponent-binding protein, mitochondrial |
| 214.98 | 4 | Galectin-8 |
| 213.91 | 8 | Perilipin \{ECO:0000256\|PIRNR:PIRNR036881\} |
| 213.31 | 1 | Beta-microseminoprotein \{ECO:0000313\|EMBL:KFZ62961.1\} |
| 211.63 | 1 | Lactoperoxidase \{ECO:0000313\|EMBL:EGV96616.1\} |
| 208.21 | 1 | Myosin heavy chain, cardiac muscle isoform |
| 205.22 | 1 | Myosin-binding protein C, fast-type |
| 198.82 | 2 | Transmembrane emp24 domain-containing protein 7 |
| 198.82 | 2 | Transmembrane emp24 domain containing 3 \{ECO:0000313\|EMBL:AAH64279.1\} |
| 197.05 | 3 | Amyloid beta A4 protein |
| 195.33 | 4 | Pro-cathepsin H |
| 195.33 | 4 | LOC100036949 protein \{ECO:0000313\|EMBL:AAI29704.1\} |
| 195.32 | 1 | Immunoglobulin superfamily member 2 |
| 190.59 | 1 | Tacstd1-prov protein \{ECO:0000313\|EMBL:AAH77845.1\} |
| 190.59 | 1 | Epithelial cell adhesion molecule |
| 190.00 | 2 | Lumican \{ECO:0000313\|EMBL:AAH82719.1\} |
| 190.00 | 2 | Lumican |
| 187.93 | 1 | Mesencephalic astrocyte-derived neurotrophic factor |
| 187.93 | 1 | Armet protein \{ECO:0000313\|EMBL:AAH82888.1\} |
| 186.97 | 3 | Heterogeneous nuclear ribonucleoprotein D-like \{ECO:0000313\|Ensembl:ENSMUSP00000118555\} |
| 186.97 | 3 | Heterogeneous nuclear ribonucleoprotein D-like |


| 184.06 | 3 | Fibroblast growth factor-binding protein 1 |
| :---: | :---: | :---: |
| 183.97 | 1 | Novel protein similar to spint2 (Serine protease inhibitor, Kunitz type, 2) \{ECO:0000313\|EMBL:CAJ82013.1\} |
| 183.97 | 1 | MGC81165 protein \{ECO:0000313\|EMBL:AAH72344.1\} |
| 183.84 | 1 | Adenylate kinase isoenzyme 6 \{ECO:0000256\|HAMAP-Rule:MF_03173\} |
| 183.84 | 1 | Adenylate kinase isoenzyme 6 \{ECO:0000255\|HAMAP-Rule:MF_03173\} |
| 178.01 | 1 | Proteinase inhibitor PSKP-1 \{ECO:0000313\|EMBL:AFY11407.1\} |
| 177.68 | 2 | High mobility group protein B1 |
| 177.53 | 2 | T-cell immunoreceptor with Ig and ITIM domains |
| 177.53 | 2 | Poliovirus receptor-related protein-like factor variant 1 \{ECO:0000313\|EMBL:AHL39279.1\} |
| 177.37 | 5 | C-factor |
| 176.23 | 2 | NADH dehydrogenase [ubiquinone] 1 beta subcomplex subunit 10 |
| 173.52 | 1 | Proteasome subunit beta type \{ECO:0000256\|RuleBase:RU004203\} |
| 173.52 | 1 | Proteasome subunit beta type-3 |
| 172.25 | 2 | Galectin-3 \{ECO:0000313\|EMBL:KFO74617.1\} |
| 171.91 | 3 | Cytochrome P450 2 J 2 |
| 171.64 | 3 | 60 kDa heat shock protein, mitochondrial |
| 170.04 | 5 | Clusterin |
| 168.20 | 1 | Nucleobindin-2 |
| 165.69 | 4 | Group IIE secretory phospholipase A2 |
| 164.89 | 2 | Nuclear receptor coactivator 5 |
| 164.02 | 3 | Fibroblast growth factor \{ECO:0000256\|RuleBase:RU049442\} |
| 162.49 | 2 | Fibrillin-1 |
| 158.45 | 2 | Toxin 3FTx-Tri2 |
| 158.38 | 1 | Fibrillin-1 \{ECO:0000313\|EMBL:AFE71298.1\} |
| 158.32 | 1 | Protein FAM3B |
| 156.90 | 2 | Cysteine-rich secretory protein 2 \{ECO:0000313\|EMBL:KGL88693.1\} |
| 156.10 | 1 | Fibroblast growth factor 19 |
| 155.86 | 1 | Plethodontid modulating factor \{ECO:0000313\|EMBL:ABI48789.1\} |
| 155.86 | 1 | Plethodontid modulating factor \{ECO:0000313\|EMBL:ABI48656.1\} |
| 154.81 | 2 | Antifreeze protein \{ECO:0000313\|EMBL:AFK10488.1\} |
| 153.92 | 1 | Adiponectin \{ECO:0000313\|EMBL:AFF19461.1\} |
| 153.92 | 1 | Adiponectin |
| 152.08 | 1 | High mobility group-T protein |
| 152.08 | 1 | High-mobility group box 1 \{ECO:0000313\|EMBL:AAH63332.1\} |
| 151.96 | 1 | Zonadhesin |
| 151.67 | 1 | Peroxiredoxin-4 |
| 150.35 | 1 | Lysosome-associated membrane glycoprotein 1 \{ECO:0000313\|EMBL:KFV75938.1\} |
| 150.35 | 1 | Lysosome-associated membrane glycoprotein 1 |
| 149.22 | 1 | Endoplasmic reticulum resident protein 44 |
| 148.27 | 2 | Peptidyl-prolyl cis-trans isomerase \{ECO:0000256\|RuleBase:RU000493\} |
| 148.27 | 2 | Peptidyl-prolyl cis-trans isomerase B |
| 145.30 | 2 | Syndecan \{ECO:0000256\|RuleBase:RU000649\} |
| 143.77 | 3 | LOC100144981 protein \{ECO:0000313\|EMBL:AAI58319.1\} |
| 141.65 | 1 | Dolichyl-diphosphooligosaccharide--protein glycosyltransferase subunit STT3A \{ECO:0000313\|EMBL:ELK09784.1\} |
| 141.65 | 1 | Dolichyl-diphosphooligosaccharide--protein glycosyltransferase subunit STT3A |
| 138.25 | 1 | cDNA FL59360, moderately similar to Biglycan \{ECO:0000313\|EMBL:BAG60276.1\} |
| 138.25 | 1 | Biglycan |
| 137.22 | 1 | Truncated sodefrin-like protein \{ECO:0000313\|EMBL:ACB54671.1\} |
| 136.63 | 2 | Secreted prophospholipase A2 group IIA \{ECO:0000313\|EMBL:ADG46025.1\} |
| 136.55 | 2 | Galectin-related protein |
| 133.97 | 2 | Ketosamine-3-kinase |
| 133.75 | 1 | Antigen WC1.1 \{ECO:0000313\|EMBL:KFV62824.1\} |
| 132.06 | 5 | Perilipin-2 |
| 130.91 | 5 | Pepsin A |
| 130.88 | 1 | NADH dehydrogenase [ubiquinone] 1 beta subcomplex subunit 3 \{ECO:0000313\|EMBL:KGL84084.1\} |
| 130.88 | 1 | NADH dehydrogenase [ubiquinone] 1 beta subcomplex subunit 3 |
| 129.45 | 1 | Eukaryotic translation initiation factor 3 subunit G \{ECO:0000256\|HAMAP-Rule:MF_03006\} |
| 129.45 | 1 | Eukaryotic translation initiation factor 3 subunit G \{ECO:0000255\|HAMAP-Rule:MF_03006\} |
| 129.08 | 2 | Ovomucoid \{ECO:0000313\|EMBL:KFQ82109.1\} |
| 128.79 | 1 | Syndecan-2 |
| 127.53 | 3 | Protein-L-isoaspartate O-methyltransferase \{ECO:0000256\|RuleBase:RU003802\} |
| 127.53 | 3 | Protein-L-isoaspartate(D-aspartate) O-methyltransferase |
| 127.18 | 4 | Ethanolamine-phosphate phospho-lyase |


| 125.88 | 1 | Collagen alpha-2(V) chain |
| :---: | :---: | :---: |
| 125.68 | 2 | Ryanodine receptor 1 \{ECO:0000313\|EMBL:ELK12861.1\} |
| 125.68 | 2 | Ryanodine receptor 1 |
| 123.91 | 1 | Cystatin JZTX-75 |
| 123.83 | 3 | Butyrophilin subfamily 1 member A1 |
| 123.64 | 1 | Ig alpha-2 chain C region |
| 122.42 | 3 | Cathelicidin-OH antimicrobial peptide |
| 122.12 | 1 | Inorganic pyrophosphatase |
| 121.34 | 1 | Microfibrillar-associated protein 5 |
| 119.70 | 1 | Kelch repeat and BTB (POZ) domain containing 10 \{ECO:0000313\|EMBL:AAH63333.1\} |
| 119.70 | 1 | Kelch-like protein 41 |
| 118.57 | 4 | WAP four-disulfide core domain protein 8 |
| 117.42 | 1 | PDZ and LIM domain protein 7 |
| 117.42 | 1 | MGC85532 protein \{ECO:0000313\|EMBL:AAH81236.1\} |
| 116.86 | 1 | Zgc:112397 \{ECO:0000313\|EMBL:AAH95803.1\} |
| 116.06 | 2 | Peroxidasin homolog |
| 116.06 | 2 | Immunoglobulin superfamily containing leucine-rich repeat protein |
| 115.94 | 1 | BPI fold-containing family B member 4 |
| 115.82 | 1 | Transmembrane protein 42 |
| 115.82 | 1 | Palmitoyltransferase \{ECO:0000256\|RuleBase:RU079119\} |
| 115.28 | 3 | Sodefrin-like factor \{ECO:0000313\|EMBL:AAZ06282.1\} |
| 114.82 | 1 | MGC80906 protein \{ECO:0000313\|EMBL:AAH73425.1\} |
| 114.17 | 1 | Epithelial membrane protein 2 \{ECO:0000313\|EMBL:ACO52019.1\} |
| 114.17 | 1 | Epithelial membrane protein 2 |
| 112.35 | 2 | Grifin \{ECO:0000313\|EMBL:KFW84439.1\} |
| 110.50 | 4 | Dipeptidase \{ECO:0000256\|RuleBase:RU341113\} |
| 110.50 | 4 | Dipeptidase 1 |
| 109.96 | 1 | Galectin \{ECO:0000313\|EMBL:ACO36044.1\} |
| 109.73 | 1 | C-factor \{ECO:0000313\|EMBL:KFO71389.1\} |
| 109.30 | 1 | Grifin \{ECO:0000313\|EMBL:KFV60199.1\} |
| 109.30 | 1 | Grifin \{ECO:0000313\|EMBL:KFV10199.1\} |
| 108.73 | 1 | Putative membrane protein C19orf24 \{ECO:0000313\|EMBL:KFP05021.1\} |
| 108.28 | 4 | Laminin subunit alpha-2 |
| 108.10 | 3 | Complement C1s subcomponent |
| 107.90 | 3 | Sodefrin-like factor \{ECO:0000313\|EMBL:AAZ06330.1\} |
| 106.76 | 1 | Vesicular integral-membrane protein VIP36 |
| 105.42 | 1 | Clusterin \{ECO:0000313\|EMBL:KFO78294.1\} |
| 104.63 | 3 | Ly6/PLAUR domain-containing protein 2 |
| 103.43 | 1 | Phospholipase A2 inhibitor subunit gamma B \{ECO:0000313\|EMBL:KFV42288.1\} |
| 102.20 | 1 | UPF0556 protein C19orf10 homolog |
| 102.20 | 1 | UPF0556 protein C19orf10 \{ECO:0000313\|EMBL:KGL82794.1\} |
| 96.99 | 1 | Cortical granule lectin \{ECO:0000313\|EMBL:CAA57946.1\} |
| 96.83 | 4 | Serine protease inhibitor Kazal-type 9 |
| 96.14 | 1 | 17-beta-hydroxysteroid dehydrogenase 13 |
| 95.61 | 3 | Serpin B10 |
| 95.50 | 1 | Putative calmodulin \{ECO:0000313\|EMBL:JAA44932.1\} |
| 95.04 | 2 | Transforming growth factor-beta-induced protein ig-h3 |
| 94.16 | 2 | Basigin |
| 93.65 | 1 | Transmembrane emp24 domain-containing protein 10 |
| 93.24 | 4 | Ig kappa chain C region |
| 93.08 | 1 | Lymphocyte antigen 6E \{ECO:0000313\|EMBL:KFP17033.1\} |
| 92.96 | 1 | Kunitz-type serine protease inhibitor ShPI-1 |
| 92.84 | 1 | Proliferation-associated protein 2G4 |
| 92.31 | 1 | Fibronectin \{ECO:0000313\|EMBL:KGL80627.1\} |
| 92.31 | 1 | Fibronectin |
| 91.90 | 1 | Plethodontid modulating factor \{ECO:0000313\|EMBL:ABI48542.1\} |
| 91.27 | 3 | Serpin H1 |
| 90.85 | 3 | Cathelicidin \{ECO:0000313\|EMBL:AHF22104.1\} |
| 89.39 | 3 | Bone marrow proteoglycan |
| 88.91 | 3 | Complement C1r subcomponent |
| 88.81 | 1 | Sperm acrosome membrane-associated protein 4 \{ECO:0000313\|EMBL:EMP25750.1\} |
| 88.81 | 1 | Sperm acrosome membrane-associated protein 4 |
| 88.01 | 1 | Gamma-glutamylcyclotransferase \{ECO:0000313\|EMBL:EMP29484.1\} |
| 88.01 | 1 | Gamma-glutamylcyclotransferase |


| 86.76 | 1 | Gelsolin (Amyloidosis, Finnish type) \{ECO:0000313\|EMBL:CAJ82431.1\} |
| :---: | :---: | :---: |
| 86.76 | 1 | Gelsolin |
| 85.75 | 3 | Leucine-rich glioma-inactivated protein 1 |
| 85.72 | 1 | NADH dehydrogenase [ubiquinone] iron-sulfur protein 7, mitochondrial |
| 85.43 | 6 | Keratin, type I cytoskeletal 10 |
| 85.06 | 2 | Cocaine esterase |
| 84.81 | 4 | Avidin \{ECO:0000313\|EMBL:KFV71220.1\} |
| 84.59 | 1 | Nuclear receptor coactivator 5 \{ECO:0000313\|EMBL:EMP28280.1\} |
| 83.39 | 3 | Protein FAM3C |
| 83.39 | 1 | Testis cDNA clone: QtsA-15330, similar to human coatomer protein complex, subunit alpha (COPA) \{ECO:0000313\|EMBL:BAE02266.1\} |
| 83.39 | 1 | Coatomer subunit alpha |
| 83.29 | 2 | Augurin |
| 83.09 | 2 | Polyglutamine-binding protein 1 |
| 82.90 | 2 | Methionine aminopeptidase 2 \{ECO:0000256\|HAMAP-Rule:MF_03175\} |
| 82.90 | 2 | Methionine aminopeptidase 2 \{ECO:0000255\|HAMAP-Rule:MF_03175\} |
| 82.71 | 1 | Group IIE secretory phospholipase A2 \{ECO:0000313\|EMBL:KFR16347.1\} |
| 82.66 | 1 | Kelch-like ECH-associated protein 1 |
| 82.40 | 2 | Growth hormone receptor |
| 82.18 | 7 | Ficolin-1 |
| 82.17 | 4 | Transmembrane 9 superfamily member 2 |
| 81.85 | 3 | Perilipin-3 |
| 80.91 | 1 | Peroxiredoxin-5, mitochondrial |
| 80.85 | 3 | von Willebrand factor |
| 80.72 | 1 | CD99 \{ECO:0000313\|EMBL:AAI69812.1\} |
| 80.72 | 1 | CD99 antigen-like protein 2 |
| 80.38 | 1 | Interferon-inducible double-stranded RNA-dependent protein kinase activator A homolog A |
| 80.38 | 1 | Interferon-inducible double stranded RNA-dependent protein kinase activator A \{ECO:0000313\|EMBL:EMP26370.1\} |
| 80.30 | 1 | Nuclear receptor coactivator 5 \{ECO:0000313\|EMBL:EMP35936.1\} |
| 79.89 | 1 | SLIT and NTRK-like protein 2 |
| 78.78 | 1 | Snaclec B9 |
| 78.52 | 1 | BPI fold-containing family B member 3 |
| 78.40 | 2 | Transketolase |
| 78.25 | 1 | Serpin H1-like \{ECO:0000313\|EMBL:AFJ51394.1\} |
| 78.21 | 1 | GHR protein \{ECO:0000313\|EMBL:AAI36497.1\} |
| 77.90 | 1 | Laminin subunit alpha-2 \{ECO:0000313\|EMBL:KFP71983.1\} |
| 77.73 | 3 | Carboxylesterase 5A |
| 77.22 | 2 | Aminoacyl tRNA synthase complex-interacting multifunctional protein 1 |
| 76.86 | 1 | Otoraplin |
| 76.30 | 3 | Phospholipase A2 homolog otoconin-22 |
| 76.30 | 3 | Otoconin-22 \{ECO:0000313\|EMBL:BAC67169.1\} |
| 76.29 | 2 | Lysozyme C, milk isozyme \{ECO:0000313\|Ensembl:ENSECAP00000015931\} |
| 76.01 | 2 | Cathelicidin-5 |
| 74.50 | 4 | ES1 protein homolog, mitochondrial |
| 73.58 | 2 | Sodefrin-like protein \{ECO:0000313\|EMBL:ACB54670.1\} |
| 73.32 | 1 | Sialomucin core protein 24 \{ECO:0000313\|EMBL:ETE69945.1\} |
| 73.32 | 1 | Sialomucin core protein 24 |
| 72.78 | 4 | Alcohol dehydrogenase [NADP(+)] |
| 72.20 | 4 | Laminin subunit beta-1 |
| 71.51 | 2 | Protein LYRIC |
| 71.50 | 2 | Vacuolar protein sorting-associated protein 13C |
| 71.47 | 3 | Insulin-degrading enzyme |
| 71.12 | 1 | Pentraxin-related protein PTX3 \{ECO:0000313\|EMBL:EMP28954.1\} |
| 71.12 | 1 | Pentraxin-related protein PTX3 |
| 70.93 | 1 | Mediator of RNA polymerase II transcription subunit 28 |
| 70.82 | 1 | Collagen alpha-1(XI) chain \{ECO:0000313\|EMBL:KFO13872.1\} |
| 70.82 | 1 | Collagen alpha-1(XI) chain |
| 70.57 | 2 | Macrophage migration inhibitory factor |
| 69.75 | 2 | Secreted frizzled-related protein 2 |
| 69.09 | 1 | Apolipoprotein Eb \{ECO:0000313\|EMBL:EMP25261.1\} |
| 69.09 | 1 | Apolipoprotein Eb |
| 68.88 | 2 | Immunoglobulin J chain |
| 68.61 | 2 | Avidin-related protein 6 |


| 67.38 | 2 | tRNA selenocysteine 1-associated protein 1 |
| :---: | :---: | :---: |
| 66.24 | 1 | Macrophage migration inhibitory factor \{ECO:0000313\|EMBL:AFN21496.1\} |
| 66.04 | 4 | Venom peptide SjAPI-2 |
| 65.38 | 1 | Stromal cell-derived factor 2-like protein 1 |
| 65.38 | 1 | MGC79547 protein \{ECO:0000313\|EMBL:AAH80914.1\} |
| 64.27 | 3 | Keratin, type I cytoskeletal 18 |
| 63.17 | 1 | ATP-dependent RNA helicase DDX24 \{ECO:0000313\|EMBL:EMP32404.1\} |
| 63.17 | 1 | ATP-dependent RNA helicase DDX24 |
| 63.05 | 3 | Palmitoyl-protein thioesterase 1 |
| 62.52 | 9 | NXPE family member 3 |
| 61.81 | 1 | Serine protease inhibitor Kazal-type 2 \{ECO:0000313\|Ensembl:ENSP00000425961\} |
| 61.70 | 7 | C-type lectin 2 \{ECO:0000313\|EMBL:JAC96487.1\} |
| 61.65 | 4 | Ferritin \{ECO:0000256\|RuleBase:RU361145\} |
| 61.49 | 2 | Fucose mutarotase |
| 61.39 | 1 | Pigment epithelium-derived factor \{ECO:0000313\|EMBL:EMP37826.1\} |
| 61.39 | 1 | Pigment epithelium-derived factor |
| 61.27 | 2 | Heterochromatin-associated protein MENT \{ECO:0000313\|EMBL:KFO99938.1\} |
| 61.11 | 1 | 40S ribosomal protein S19 \{ECO:0000313\|EMBL:ELW71038.1\} |
| 61.11 | 1 | 40S ribosomal protein S19 |
| 61.02 | 1 | Olfactomedin-like protein 3 |
| 60.96 | 3 | Ervatamin-B |
| 60.24 | 1 | UPF0480 \{ECO:0000313\|EMBL:JAB52965.1\} |
| 60.24 | 1 | ER membrane protein complex subunit 7 |
| 60.16 | 3 | Secretory leukocyte protease inhibitor \{ECO:0000313\|EMBL:AAD34035.1\} |
| 59.76 | 6 | L-amino-acid oxidase |
| 59.70 | 1 | Ig lambda-1 chain V region S43 |
| 59.70 | 1 | Ig lambda-1 chain C region |
| 59.42 | 1 | Alpha-2-macroglobulin-like protein 1 \{ECO:0000313\|EMBL:AFE69079.1\} |
| 59.42 | 1 | Alpha-2-macroglobulin-like protein 1 |
| 59.09 | 1 | Beta-microseminoprotein \{ECO:0000313\|EMBL:KFW92582.1\} |
| 58.97 | 1 | Multiple coagulation factor deficiency protein 2 homolog |
| 58.10 | 1 | Cathelicidin antimicrobial peptide |
| 57.40 | 1 | 39S ribosomal protein L47, mitochondrial |
| 57.00 | 2 | Ceruloplasmin |
| 56.25 | 2 | Laminin subunit beta-2 |
| 56.14 | 1 | Cystatin C \{ECO:0000313\|EMBL:ABG48755.1\} |
| 55.57 | 2 | Collagen alpha-2(IV) chain |
| 55.50 | 2 | Serum paraoxonase/arylesterase 2 \{ECO:0000313\|EMBL:EMP33942.1\} |
| 55.50 | 2 | Serum paraoxonase/arylesterase 2 |
| 54.88 | 1 | A-kinase anchor protein 8-like |
| 54.63 | 2 | Perlwapin \{ECO:0000250\|UniProtKB:P84811\} |
| 54.37 | 2 | Insulin-like growth factor-binding protein 3 |
| 53.77 | 1 | Nuclear factor erythroid 2-related factor 2 \{ECO:0000313\|EMBL:KFO89273.1\} |
| 53.77 | 1 | Nuclear factor erythroid 2-related factor 2 |
| 52.46 | 2 | Kinase suppressor of Ras 1 \{ECO:0000313\|EMBL:EGV94352.1\} |
| 52.26 | 1 | Resistin-like beta \{ECO:0000313\|EMBL:ELV12719.1\} |
| 52.26 | 1 | Resistin \{ECO:0000313\|EMBL:EMP23808.1\} |
| 52.26 | 1 | Charged multivesicular body protein 4b |
| 52.26 | 1 | Anionic trypsin |
| 52.23 | 3 | Activity-dependent neuroprotector homeobox protein |
| 52.06 | 4 | Di-N-acetylchitobiase |
| 51.66 | 5 | Lysozyme g |
| 51.46 | 2 | CRISP-like factor variant 1 \{ECO:0000313\|EMBL:AHL39260.1\} |
| 51.44 | 1 | Insulin-like growth factor-binding protein 3 \{ECO:0000313\|EMBL:KFV86244.1\} |
| 51.37 | 1 | Lysozyme \{ECO:0000313\|EMBL:AIA57285.1\} |
| 51.33 | 1 | Protein notum like protein \{ECO:0000313\|EMBL:EMP31531.1\} |
| 50.70 | 1 | Murinoglobulin-1 |
| 50.27 | 1 | Methyltransferase-like protein 9 |
| 50.27 | 1 | DORA reverse strand protein 1 \{ECO:0000313\|EMBL:CAJ82774.1\} |
| 50.26 | 2 | Peptidyl-prolyl cis-trans isomerase A |
| 49.90 | 2 | Transmembrane protein 9B \{ECO:0000313\|EMBL:AEC13102.1\} |
| 49.90 | 2 | Transmembrane protein 9B |
| 49.36 | 1 | Pepsin A-5 \{ECO:0000313\|Ensembl:ENSP00000441981\} |
| 49.26 | 1 | Podoplanin \{ECO:0000313\|EMBL:KFV78802.1\} |


| 48.58 | 1 | Anterior gradient protein 2 homolog |
| :---: | :---: | :---: |
| 48.52 | 1 | Secreted frizzled-related protein 2 \{ECO:0000313\|EMBL:AAH61311.1\} |
| 48.15 | 1 | Target of Nesh-SH3 |
| 48.15 | 1 | Fibronectin type III domain-containing protein 1 \{ECO:0000313\|EMBL:KFO10663.1\} |
| 48.05 | 1 | Putative Kunitz-type proteinase inhibitor \{ECO:0000313\|EMBL:EMP28254.1\} |
| 47.95 | 1 | Beta-sarcoglycan |
| 47.63 | 1 | Snaclec rhodocytin subunit beta |
| 47.63 | 1 | Galactose binding lectin \{ECO:0000313\|EMBL:BAN82148.1\} |
| 47.62 | 3 | Neuroserpin |
| 47.54 | 1 | T-complex protein 1 subunit gamma \{ECO:0000256\|RuleBase:RU004191\} |
| 47.54 | 1 | T-complex protein 1 subunit gamma |
| 47.51 | 2 | Urokinase plasminogen activator surface receptor |
| 47.49 | 2 | Serglycin |
| 47.44 | 2 | Protein YIPF \{ECO:0000256\|RuleBase:RU361264\} |
| 47.44 | 2 | Protein YIPF1 |
| 47.44 | 1 | Thioredoxin domain-containing secreted protein Agr3 \{ECO:0000313\|EMBL:AGG20201.1\} |
| 47.44 | 1 | Anterior gradient protein 3 homolog |
| 46.83 | 1 | MGC81165 protein \{ECO:0000313\|EMBL:AAH99302.1\} |
| 46.43 | 1 | Thyroid receptor-interacting protein 11 \{ECO:0000313\|EMBL:ELW55803.1\} |
| 46.43 | 1 | Fibulin-5 |
| 46.40 | 1 | Transmembrane protein 147 |
| 46.36 | 5 | C-reactive protein |
| 46.17 | 2 | BPI fold-containing family C protein |
| 46.13 | 3 | 2-phosphoxylose phosphatase 1 \{ECO:0000250\|UniProtKB:Q8TE99\} |
| 46.10 | 3 | Centrosomal protein of 164 kDa |
| 45.59 | 3 | Lysyl oxidase homolog 1 |
| 45.25 | 2 | Complement C1q tumor necrosis factor-related protein 2 \{ECO:0000313\|EMBL:KFQ85433.1\} |
| 45.25 | 2 | Complement C1q tumor necrosis factor-related protein 2 |
| 45.16 | 1 | Cathelicidin \{ECO:0000313\|EMBL:ACJ76797.1\} |
| 44.98 | 1 | C-type lectin domain family 11 member A |
| 44.61 | 1 | Thioredoxin domain-containing protein 15 |
| 44.56 | 3 | Inter-alpha-trypsin inhibitor heavy chain H5 |
| 44.54 | 1 | Pre-mRNA 3'-end-processing factor FIP1 |
| 44.54 | 1 | Glutamine synthetase \{ECO:0000313\|EMBL:KFV42666.1\} |
| 44.54 | 1 | Glutamine synthetase |
| 44.45 | 3 | Clusterin \{ECO:0000256\|RuleBase:RU000629\} |
| 43.88 | 1 | Plasma protease C1 inhibitor |
| 43.76 | 1 | Fumarylacetoacetate hydrolase domain-containing protein 2 |
| 43.41 | 4 | Lysozyme g \{ECO:0000313\|EMBL:KFO07954.1\} |
| 43.34 | 1 | Protein Red |
| 43.34 | 1 | MGC83922 protein \{ECO:0000313\|EMBL:AAH77589.1\} |
| 43.06 | 1 | Tissue inhibitor of metalloproteinase 1 \{ECO:0000313\|EMBL:ABB88702.1\} |
| 42.84 | 2 | Tenascin-X |
| 42.20 | 4 | Collagen triple helix repeat-containing protein 1 |
| 42.14 | 1 | Twisted gastrulation protein homolog 1 |
| 41.78 | 2 | Annexin A2 |
| 41.54 | 3 | ESF1 homolog |
| 41.23 | 4 | Serpin B6 |
| 41.05 | 1 | Activity-dependent neuroprotector homeobox protein \{ECO:0000313\|EMBL:KFQ35549.1\} |
| 41.01 | 7 | NXPE family member 1 |
| 41.01 | 3 | Ankyrin repeat and KH domain-containing protein 1 |
| 40.46 | 2 | ATP-binding cassette sub-family F member 1 |
| 40.41 | 1 | Legumain \{ECO:0000313\|EMBL:AAH75316.1\} |
| 40.41 | 1 | Legumain |
| 40.40 | 3 | Snake venom metalloprotease inhibitor 02A10 |
| 40.40 | 3 | Predicted protein \{ECO:0000313\|EMBL:EDO28433.1\} |
| 40.36 | 1 | NFKB activating protein \{ECO:0000313\|EMBL:AFP00720.1\} |
| 40.36 | 1 | NF-kappa-B-activating protein |
| 40.29 | 3 | Ferritin light chain 1 |
| 39.85 | 1 | Tiarin \{ECO:0000313\|EMBL:BAB85495.1\} |
| 39.85 | 1 | Olfactomedin |
| 39.78 | 2 | Trefoil factor \{ECO:0000313\|EMBL:CAB65516.1\} |
| 39.68 | 1 | TAR DNA-binding protein 43 \{ECO:0000313\|Ensembl:ENSP00000479758\} |
| 39.68 | 1 | TAR DNA-binding protein 43 |


| 39.48 | 3 | Sarcolemmal membrane-associated protein |
| :---: | :---: | :---: |
| 39.34 | 4 | Plastin-2 |
| 39.27 | 1 | Transmembrane protein 213 |
| 39.05 | 2 | Probable U3 small nucleolar RNA-associated protein 11 |
| 38.92 | 1 | MGC79016 protein \{ECO:0000313\|EMBL:AAH68626.1\} |
| 38.92 | 1 | Apolipoprotein O |
| 38.91 | 1 | Proheparin-binding EGF-like growth factor |
| 38.66 | 1 | Peptidyl-prolyl cis-trans isomerase A \{ECO:0000313\|EMBL:EGW10562.1\} |
| 38.64 | 3 | Nucleotide exchange factor SIL1 |
| 38.36 | 3 | Ribonuclease P protein subunit p14 |
| 38.36 | 3 | Hydroxyacyl-thioester dehydratase type 2, mitochondrial \{ECO:0000303\|PubMed:17898086\} |
| 38.26 | 2 | Centrosomal protein of 164 kDa \{ECO:0000313\|EMBL:KFV78845.1\} |
| 38.16 | 1 | Oncorhynchus mykiss genomic scaffold, scaffold_89 \{ECO:0000313\|EMBL:CDQ60634.1\} |
| 37.50 | 1 | Kunitz-type protease inhibitor 1 |
| 37.38 | 1 | Pyruvate dehydrogenase kinase, isoenzyme 2 \{ECO:0000313\|EMBL:CAJ82883.1\} |
| 37.38 | 1 | [Pyruvate dehydrogenase (acetyl-transferring)] kinase isozyme 4, mitochondrial |
| 36.57 | 2 | Complement C1q tumor necrosis factor-related protein 1 |
| 36.54 | 2 | Peptidyl-tRNA hydrolase ICT1, mitochondrial |
| 36.39 | 3 | Ig lambda chain V-I region BL2 |
| 36.36 | 2 | Keratin, type I cytoskeletal 18 \{ECO:0000313\|Ensembl:ENSXETP00000012361\} |
| 36.14 | 2 | Cochlin |
| 36.08 | 1 | Dnas like protein subfamily B member 11 \{ECO:0000313\|EMBL:EPQ02836.1\} |
| 36.08 | 1 | DnaJ homolog subfamily B member 11 |
| 35.87 | 1 | Complement C1r subcomponent-like protein \{ECO:0000313\|EMBL:JAA97381.1\} |
| 35.74 | 8 | CD109 antigen |
| 35.38 | 2 | 39S ribosomal protein L18, mitochondrial |
| 35.10 | 2 | Apoptosis regulator R1 |
| 35.09 | 1 | Fructose-1,6-bisphosphatase isozyme 2 |
| 34.88 | 1 | Cytochrome c oxidase assembly factor 6 homolog |
| 34.77 | 2 | Ig kappa chain V region Mem5 |
| 34.72 | 1 | Thrombospondin-4 \{ECO:0000313\|EMBL:KFO80158.1\} |
| 34.72 | 1 | Thrombospondin-4 \{ECO:0000313\|EMBL:ADX36090.1\} |
| 34.72 | 1 | Thrombospondin-4 |
| 34.61 | 1 | Chorionic proteinase inhibitor \{ECO:0000313\|EMBL:BAC00855.1\} |
| 34.44 | 1 | Serine protease inhibitor Kazal-type 2-like protein \{ECO:0000313\|EMBL:AHJ79063.1\} |
| 34.15 | 2 | lg mu chain C region secreted form |
| 34.07 | 2 | Synaptotagmin-like protein 4 |
| 33.97 | 1 | Deleted in malignant brain tumors 1 protein \{ECO:0000313\|EMBL:ELK34636.1\} |
| 33.96 | 1 | Pepsinogen A \{ECO:0000313\|EMBL:BAB20798.1\} |
| 33.68 | 2 | Ig kappa chain V-IV region |
| 33.64 | 3 | Pleiotrophin |
| 33.45 | 2 | Inter-alpha-trypsin inhibitor heavy chain H5 \{ECO:0000313\|EMBL:EMP38828.1\} |
| 33.44 | 1 | 28 r ribosomal protein S5, mitochondrial |
| 33.37 | 1 | Splicing factor U2AF 26 kDa subunit \{ECO:0000313\|Ensembl:ENSMUSP00000129385\} |
| 33.37 | 1 | Splicing factor U2AF 26 kDa subunit |
| 33.29 | 1 | MGC68532 protein \{ECO:0000313\|EMBL:AAH60400.1\} |
| 33.29 | 1 | Far upstream element-binding protein 3 |
| 33.16 | 3 | Transmembrane protein 87A |
| 33.09 | 1 | Gag-Pol polyprotein |
| 33.07 | 4 | Krev interaction trapped protein 1 |
| 33.07 | 2 | Serotransferrin-A |
| 32.93 | 1 | Zygote arrest protein 1 |
| 32.69 | 2 | Family with sequence similarity 3, member D \{ECO:0000313\|EMBL:CAJ83001.1\} |
| 32.39 | 2 | Serine/threonine-protein kinase RIO2 |
| 32.29 | 2 | Aminopeptidase B |
| 32.25 | 5 | Angiopoietin-related protein 3 |
| 32.04 | 1 | MGC107878 protein \{ECO:0000313\|EMBL:AAH89660.1\} |
| 32.04 | 1 | Coagulation factor X |
| 32.02 | 1 | Laminin subunit gamma-2 |
| 31.83 | 1 | Ectonucleotide pyrophosphatase/phosphodiesterase family member 5 |
| 31.81 | 1 | cDNA FL39956 fis, clone SPLEN2024990, highly similar to Plastin-2 \{ECO:0000313\|EMBL:BAG53443.1\} |
| 31.64 | 1 | ICT1 \{ECO:0000313\|PDB:4CE4\} |
| 31.62 | 1 | Transducin beta-like protein 2 |
| 31.56 | 1 | Heterochromatin-associated protein MENT \{ECO:0000313\|EMBL:KFQ53181.1\} |


| 31.38 | 1 | Group XIIA secretory phospholipase A2 |
| :---: | :---: | :---: |
| 31.21 | 2 | Alpha-1-antiproteinase 2 |
| 31.07 | 3 | Guanine nucleotide-binding protein G(I)/G(S)/G(T) subunit beta-2 |
| 31.04 | 1 | Zinc finger protein 460 |
| 30.90 | 1 | Sarcolemmal membrane-associated protein \{ECO:0000313\|EMBL:KFW82164.1\} |
| 30.56 | 1 | Beta-1,4-galactosyltransferase 1 |
| 29.72 | 5 | Complement component C7 |
| 29.71 | 1 | Plasminogen activator inhibitor 2 |
| 29.71 | 1 | Heterochromatin-associated protein MENT \{ECO:0000313\|EMBL:KFV82937.1\} |
| 29.21 | 2 | Dipeptidyl peptidase 4 |
| 28.95 | 2 | LOC100125128 protein \{ECO:0000313\|EMBL:AAI35282.1\} |
| 28.95 | 2 | 39 S ribosomal protein L46, mitochondrial |
| 28.93 | 1 | Platelet-derived growth factor receptor-like protein |
| 28.93 | 1 | MGC82397 protein \{ECO:0000313\|EMBL:AAH72269.1\} |
| 28.89 | 2 | Zinc finger MYND domain-containing protein 11 |
| 28.83 | 3 | Alpha-1-antiproteinase F |
| 28.78 | 1 | Oncorhynchus mykiss genomic scaffold, scaffold_8130 \{ECO:0000313\|EMBL:CDQ91870.1\} |
| 28.74 | 1 | CD157 \{ECO:0000313\|EMBL:BAL72805.1\} |
| 28.74 | 1 | ADP-ribosyl cyclase/cyclic ADP-ribose hydrolase 2 |
| 28.62 | 1 | Urea transporter 2 |
| 28.46 | 2 | Galactokinase |
| 28.46 | 1 | UNQ655/PRO1286 \{ECO:0000313\|EMBL:ACQ58460.1\} |
| 28.46 | 1 | Ubiquinol-cytochrome-c reductase complex assembly factor 3 \{ECO:0000250\|UniProtKB:Q6UW78\} |
| 28.15 | 2 | Hypothetical LOC496762 \{ECO:0000313\|EMBL:AAH88031.1\} |
| 28.14 | 4 | Peroxiredoxin-1 |
| 28.13 | 1 | Keratin-3, type I cytoskeletal 51 kDa |
| 28.08 | 2 | Protein PBMUCL2 |
| 27.82 | 3 | Ubiquitin carboxyl-terminal hydrolase \{ECO:0000256\|RuleBase:RU004435\} |
| 27.79 | 3 | Glypican-3 |
| 27.70 | 1 | Ribonucleoprotein, PTB-binding 2 \{ECO:0000313\|EMBL:AAI22908.1\} |
| 27.70 | 1 | Ribonucleoprotein PTB-binding 2 |
| 27.65 | 3 | Glia-derived nexin |
| 27.47 | 4 | C-C motif chemokine \{ECO:0000256\| RuleBase:RU361150\} |
| 27.47 | 1 | WD repeat-containing protein 44 \{ECO:0000313\|EMBL:KFZ62824.1\} |
| 27.47 | 1 | WD repeat-containing protein 44 |
| 27.41 | 1 | Syntenin-1 |
| 27.31 | 1 | Interleukin-6 signal transducer \{ECO:0000313\|EMBL:ACW83267.1\} |
| 27.31 | 1 | Interleukin-6 receptor subunit beta |
| 27.18 | 3 | Protein phosphatase 1 regulatory subunit 14B |
| 27.04 | 2 | Semaphorin-3D |
| 27.01 | 1 | Precursor polypeptide (AA -10 to 121) (413 is 1st base in codon) \{ECO:0000313\|EMBL:CAA36013.1\} |
| 27.01 | 1 | Ig kappa chain V-II region RPMI 6410 |
| 26.85 | 1 | Transmembrane protein 2 \{ECO:0000313\|EMBL:ELW63127.1\} |
| 26.85 | 1 | Alpha/beta hydrolase domain-containing protein 17B |
| 26.77 | 1 | UPF0762 protein C6orf58 homolog |
| 26.77 | 1 | UPF0762 protein C6orf58 \{ECO:0000313\|EMBL:KFM09062.1\} |
| 26.51 | 2 | Alcohol dehydrogenase (NADP+)-like protein \{ECO:0000313\|EMBL:AEH58609.1\} |
| 26.37 | 1 | Calcium-binding protein 39-like |
| 26.31 | 1 | E3 UFM1-protein ligase 1 \{ECO:0000313\|EMBL:KFP41050.1\} |
| 26.31 | 1 | E3 UFM1-protein ligase 1 |
| 26.16 | 2 | Leucine-rich repeats and immunoglobulin-like domains protein 3 |
| 26.02 | 1 | Protein FAM3D |
| 25.97 | 1 | Ubiquitin-40S ribosomal protein S27a |
| 25.96 | 5 | Trimethylguanosine synthase |
| 25.92 | 2 | Plasma alpha-L-fucosidase |
| 25.69 | 2 | Alpha-1,4 glucan phosphorylase \{ECO:0000256\|RuleBase:RU000587\} |
| 25.65 | 3 | Dickkopf-related protein 2 |
| 25.65 | 1 | Trefoil factor 2 |
| 25.60 | 1 | Ac2-008 \{ECO:0000313\|EMBL:AAP86257.1\} |
| 25.57 | 1 | Intracellular hyaluronan-binding protein 4 |
| 25.37 | 1 | MGC107932 protein \{ECO:0000313\|EMBL:AAH89683.1\} |
| 25.37 | 1 | Cathepsin K |
| 25.26 | 2 | MGC85345 protein \{ECO:0000313\|EMBL:AAH78523.1\} |
| 25.22 | 3 | CRiSP-Den-3 \{ECO:0000313\|EMBL:JAA74735.1\} |


| 25.20 | 4 | Major urinary proteins 11 and 8 |
| :---: | :---: | :---: |
| 25.10 | 3 | Tissue factor pathway inhibitor \{ECO:0000256\|PIRNR:PIRNR001620\} |
| 25.00 | 3 | tRNA (guanine(26)-N(2))-dimethyltransferase |
| 25.00 | 3 | LOC100158298 protein \{ECO:0000313\|EMBL:AAI60728.1\} |
| 24.94 | 1 | T-cell immunomodulatory protein |
| 24.83 | 1 | Extracellular sulfatase Sulf-2 |
| 24.59 | 3 | BolA-like protein 1 |
| 24.53 | 1 | Primary amine oxidase, liver isozyme |
| 24.53 | 1 | Membrane primary amine oxidase \{ECO:0000313\|EMBL:KFO97554.1\} |
| 24.30 | 1 | Obg-like ATPase 1 \{ECO:0000255\|HAMAP-Rule:MF_03167\} |
| 24.15 | 1 | Chitinase domain-containing protein 1 |
| 24.10 | 1 | ORM1-like protein 2 \{ECO:0000313\|EMBL:JAB07118.1\} |
| 24.10 | 1 | ORM1-like protein 2 |
| 24.09 | 3 | Avidin \{ECO:0000313\|EMBL:KFV71219.1\} |
| 23.80 | 1 | UDP-N-acetylglucosamine--peptide N -acetylglucosaminyltransferase 110 kDa subunit |
| 23.77 | 1 | Aminopeptidase B \{ECO:0000313\|EMBL:EMP41007.1\} |
| 23.76 | 3 | Protein Mup9 \{ECO:0000313\|Ensembl:ENSMUSP00000113461\} |
| 23.70 | 2 | Tissue factor pathway inhibitor 2 |
| 23.34 | 1 | Serpin B6 \{ECO:0000313\|EMBL:KFR09243.1\} |
| 23.30 | 4 | Transforming growth factor beta-2 |
| 23.08 | 1 | Thrombospondin-4 \{ECO:0000313\|EMBL:KFQ24790.1\} |
| 23.08 | 1 | Cartilage oligomeric matrix protein |
| 22.90 | 2 | Peptidase inhibitor 16 \{ECO:0000313\|Ensembl:ENSMUSP00000110347\} |
| 22.90 | 2 | Peptidase inhibitor 16 |
| 22.88 | 3 | Cholesteryl ester transfer protein \{ECO:0000303\|PubMed:17574888\} |
| 22.87 | 2 | Cartilage-associated protein |
| 22.86 | 4 | Laminin subunit alpha-4 |
| 22.81 | 2 | Nucleotide exchange factor SIL1 \{ECO:0000313\|EMBL:KFM03393.1\} |
| 22.55 | 2 | von Willebrand factor D and EGF domain-containing protein |
| 22.43 | 2 | BolA-like protein 3 |
| 22.41 | 1 | MHC class IA alpha chain \{ECO:0000313\|EMBL:AAC60108.1\} |
| 22.41 | 1 | Class I histocompatibility antigen, F10 alpha chain |
| 22.38 | 1 | Succinate dehydrogenase assembly factor 4, mitochondrial \{ECO:0000250\|UniProtKB:Q5VUM1\} |
| 22.38 | 1 | MGC89089 protein \{ECO:0000313\|EMBL:AAH82498.1\} |
| 22.36 | 1 | V(Kappa) gene product \{ECO:0000313\|EMBL:CAA29301.1\} |
| 22.25 | 2 | Renalase |
| 22.22 | 5 | Interleukin-17D |
| 22.12 | 1 | Apoptogenic protein 1, mitochondrial \{ECO:0000313\|EMBL:JAB13401.1\} |
| 22.12 | 1 | Apoptogenic protein 1, mitochondrial |
| 22.10 | 3 | La-related protein 7 |
| 22.04 | 1 | Dynamin-1-like protein |
| 22.04 | 1 | Dynamin-1-like \{ECO:0000313\|EMBL:KFW81991.1\} |
| 21.93 | 1 | Golgin subfamily A member 4 \{ECO:0000250\|UniProtKB:Q13439\} |
| 21.93 | 1 | Centrosome-related protein F46 \{ECO:0000313\|EMBL:AIJ28482.1\} |
| 21.88 | 2 | Sclerostin domain-containing protein 1 \{ECO:0000313\|EMBL:ETE64115.1\} |
| 21.88 | 2 | Sclerostin domain-containing protein 1 |
| 21.85 | 2 | Ubiquitin carboxyl-terminal hydrolase 1 |
| 21.85 | 1 | Sushi, von Willebrand factor type A, EGF and pentraxin domain-containing protein 1 |
| 21.80 | 5 | LINE-1 retrotransposable element ORF2 protein |
| 21.60 | 1 | 60S ribosomal protein L30 |
| 21.36 | 1 | Ferritin heavy chain |
| 21.29 | 1 | Bone morphogenetic protein receptor type-2 \{ECO:0000313\|EMBL:KFQ06517.1\} |
| 21.29 | 1 | Bone morphogenetic protein receptor type-2 |
| 21.27 | 1 | Beta-galactoside alpha-2,6-sialyltransferase 1 \{ECO:0000313\|EMBL:KFR08047.1\} |
| 21.27 | 1 | Beta-galactoside alpha-2,6-sialyltransferase 1 |
| 21.26 | 2 | Cholesteryl ester transfer protein \{ECO:0000313\|EMBL:KFU85826.1\} |
| 21.26 | 1 | Major urinary protein 1 |
| 21.23 | 1 | LOC100037908 protein \{ECO:0000313\|EMBL:AAI35252.1\} |
| 20.83 | 3 | Phosphatidylcholine-sterol acyltransferase |
| 20.83 | 3 | MGC82035 protein \{ECO:0000313\|EMBL:AAH81072.1\} |
| 20.80 | 3 | Protein FAM172A |
| 20.80 | 1 | Glycogen phosphorylase, brain form |
| 20.74 | 2 | Galactose-specific lectin nattectin |
| 20.66 | 1 | Ras-related protein Rab-40B |


| 20.66 | 1 | MGC85390 protein \{ECO:0000313\|EMBL:AAH78543.1\} |
| :---: | :---: | :---: |
| 20.46 | 1 | Serum paraoxonase/arylesterase 2 \{ECO:0000313\|EMBL:KFV62202.1\} |
| 20.46 | 1 | Basic phospholipase A2 DsM-S1 |
| 20.45 | 1 | Lens fiber membrane intrinsic protein |
| 20.42 | 2 | Mitotic spindle assembly checkpoint protein MAD2B |
| 20.36 | 1 | Vasculin \{ECO:0000313\|EMBL:ELW65670.1\} |
| 20.36 | 1 | Vasculin |
| 20.21 | 1 | Scaffold attachment factor B1 |
| 20.19 | 2 | Cysteine-rich with EGF-like domain protein 2-B |
| 20.17 | 3 | Avidin |
| 20.17 | 1 | Clusterin \{ECO:0000313\|EMBL:AAG31162.1\} |
| 19.99 | 1 | Transporter \{ECO:0000256\|RuleBase:RU003732\} |
| 19.99 | 1 | Sodium- and chloride-dependent neutral and basic amino acid transporter $\mathrm{B}(0+$ ) |
| 19.95 | 2 | LOC100126637 protein \{ECO:0000313\|EMBL:AAI53784.1\} |
| 19.87 | 1 | Oncorhynchus mykiss genomic scaffold, scaffold_1961 \{ECO:0000313\|EMBL:CDQ83694.1\} |
| 19.83 | 1 | Mgc53960 protein \{ECO:0000313\|EMBL:AAI24909.1\} |
| 19.58 | 1 | LOC100125115 protein \{ECO:0000313\|EMBL:AAI53335.1\} |
| 19.58 | 1 | ESF1 like protein \{ECO:0000313\|EMBL:EMP31753.1\} |
| 19.58 | 1 | ESF1 homolog \{ECO:0000313\|EMBL:CCP80336.1\} |
| 19.58 | 1 | C-X-C motif chemokine 14 |
| 19.48 | 3 | Interleukin-1 receptor-associated kinase 4 |
| 19.47 | 1 | Ubiquitin-like protein ISG15 |
| 19.39 | 2 | Beta-2-microglobulin \{ECO:0000256\|SAAS:SAAS00098961\} |
| 19.36 | 1 | Phosphatidylinositol 3-kinase \{ECO:0000256\|PIRNR:PIRNR000587\} |
| 19.36 | 1 | Phosphatidylinositol 3-kinase catalytic subunit type 3 |
| 19.32 | 1 | Receptor-type tyrosine-protein phosphatase mu |
| 19.31 | 4 | Radixin |
| 19.31 | 1 | Coatomer subunit gamma \{ECO:0000256\|PIRNR:PIRNR037093\} |
| 19.31 | 1 | Coatomer subunit gamma-1 |
| 19.19 | 1 | 39S ribosomal protein L45, mitochondrial |
| 19.19 | 1 | 39S ribosomal protein L45 \{ECO:0000313\|EMBL:EMP33799.1\} |
| 19.17 | 1 | Threonyl-tRNA synthetase-like 2 \{ECO:0000313\|EMBL:AAH67949.1\} |
| 19.17 | 1 | Threonine-tRNA ligase, cytoplasmic |
| 19.06 | 1 | Glutamate--cysteine ligase catalytic subunit \{ECO:0000313\|EMBL:EMP40238.1\} |
| 19.06 | 1 | Glutamate--cysteine ligase catalytic subunit |
| 18.99 | 2 | Protein Gm15299 \{ECO:0000313\|Ensembl:ENSMUSP00000131618\} |
| 18.98 | 1 | DNA replication complex GINS protein PSF2 \{ECO:0000256\|PIRNR:PIRNR028998\} |
| 18.98 | 1 | DNA replication complex GINS protein PSF2 |
| 18.96 | 2 | Complement C3 |
| 18.96 | 2 | Carboxypeptidase B |
| 18.95 | 2 | Corticotropin-releasing factor-binding protein |
| 18.88 | 1 | Cysteine-rich with EGF-like domain protein 2-A |
| 18.87 | 2 | Carboxypeptidase N catalytic chain |
| 18.84 | 1 | 40S ribosomal protein S3 |
| 18.73 | 1 | Prostaglandin reductase 1 |
| 18.69 | 2 | Peptidyl-prolyl cis-trans isomerase G |
| 18.67 | 1 | Lysyl oxidase 1 \{ECO:0000313\|EMBL:KFV99147.1\} |
| 18.54 | 1 | Transmembrane gamma-carboxyglutamic acid protein 2 |
| 18.54 | 1 | Proline rich Gla (G-carboxyglutamic acid) 4 (Transmembrane) \{ECO:0000313\|EMBL:AAH75566.1\} |
| 18.51 | 2 | Collagen alpha-1(II) chain |
| 18.36 | 1 | Melanoma-derived growth regulatory protein |
| 18.28 | 6 | Alpha-2-macroglobulin receptor-associated protein |
| 18.10 | 3 | Extracellular peptidase inhibitor \{ECO:0000313\|EMBL:ELR61914.1\} |
| 18.04 | 1 | Igk protein \{ECO:0000313\|EMBL:AAH80787.1\} |
| 17.78 | 2 | Secretogranin-3 |
| 17.73 | 1 | Lysozyme C II |
| 17.67 | 3 | Ficolin B \{ECO:0000313\|EMBL:JAC95001.1\} |
| 17.60 | 2 | Avd protein \{ECO:0000313\|EMBL:AAI52705.1\} |
| 17.59 | 1 | Tumor necrosis factor receptor superfamily member 5 |
| 17.59 | 1 | Single-pass membrane and coiled-coil domain-containing protein 3 |
| 17.58 | 1 | 5'-AMP-activated protein kinase subunit gamma-3 |
| 17.35 | 1 | Protein TSSC1 \{ECO:0000313\|EMBL:KFQ46550.1\} |
| 17.35 | 1 | Protein TSSC1 \{ECO:0000313\|EMBL:KFP82785.1\} |
| 17.35 | 1 | Protein TSSC1 |


| 17.32 | 4 | Niemann-Pick C 1 protein |
| :---: | :---: | :---: |
| 17.32 | 3 | Interferon alpha 8/6 \{ECO:0000313\|EMBL:AAO63595.1\} |
| 17.31 | 2 | Prenylcysteine oxidase-like |
| 17.30 | 1 | Integral membrane protein 2B |
| 17.22 | 1 | Angiogenic factor with G patch and FHA domains 1 |
| 17.16 | 3 | NADPH-dependent diflavin oxidoreductase 1 \{ECO:0000256\|HAMAP-Rule:MF_03178\} |
| 17.16 | 3 | NADPH-dependent diflavin oxidoreductase 1 \{ECO:0000255\|HAMAP-Rule:MF_03178\} |
| 17.12 | 2 | Bone morphogenetic protein 7 |
| 17.09 | 2 | Stabilin-2 \{ECO:0000313\|EMBL:EMP23972.1\} |
| 17.09 | 2 | Stabilin-2 |
| 16.99 | 3 | Coxsackievirus and adenovirus receptor homolog |
| 16.94 | 2 | Cthrc1 protein \{ECO:0000313\|EMBL:AAI44746.1\} |
| 16.76 | 2 | Matrilin-4 |
| 16.69 | 1 | Probable pancreatic secretory proteinase inhibitor |
| 16.69 | 1 | Chymotrypsin inhibitor \{ECO:0000313\|EMBL:KFP71820.1\} |
| 16.68 | 2 | C-X-C motif chemokine 10 |
| 16.63 | 2 | Vesicle transport through interaction with t-SNAREs homolog 1B |
| 16.54 | 1 | Interleukin-8 \{ECO:0000313\|EMBL:AFK10667.1\} |
| 16.54 | 1 | C-X-C motif chemokine 13 |
| 16.52 | 1 | Retinol-binding protein 1 |
| 16.51 | 1 | Vacuolar-sorting protein SNF8 |
| 16.51 | 1 | Syndecan-1 |
| 16.41 | 1 | Ankyrin repeat and KH domain-containing protein 1 \{ECO:0000313\|EMBL:KFP65664.1\} |
| 16.40 | 1 | Transferrin \{ECO:0000313\|EMBL:AEZ53879.1\} |
| 16.39 | 1 | Phospholipase A2 \{ECO:0000313\|EMBL:BAP39946.1\} |
| 16.39 | 1 | Acidic phospholipase A2 Ts-A3 |
| 16.36 | 1 | Ovomucoid \{ECO:0000313\|EMBL:KFO93738.1\} |
| 16.35 | 1 | Cardiotrophin-1 |
| 16.32 | 4 | Deoxyribonuclease \{ECO:0000256\|PIRNR:PIRNR000988\} |
| 16.32 | 4 | Deoxyribonuclease-1-like 2 \{ECO:0000313\|EMBL:KFV67432.1\} |
| 16.32 | 4 | Deoxyribonuclease-1-like 2 |
| 16.21 | 1 | Protein canopy homolog 1 |
| 16.17 | 3 | Transmembrane 9 superfamily member 4 |
| 16.16 | 1 | Cthrc1-prov protein \{ECO:0000313\|EMBL:AAH89073.1\} |
| 16.11 | 1 | Latent-transforming growth factor beta-binding protein 3 |
| 15.99 | 1 | LOC398774 protein \{ECO:0000313\|EMBL:AAH97629.1\} |
| 15.99 | 1 | Ig mu chain C region |
| 15.96 | 1 | Ig kappa chain V-IV region B17 \{ECO:0000313\|EMBL:EMP41492.1\} |
| 15.92 | 1 | HIV-1 Rev binding protein \{ECO:0000313\|EMBL:CAJ81718.1\} |
| 15.92 | 1 | Arf-GAP domain and FG repeat-containing protein 1 |
| 15.83 | 1 | Zinc finger protein 3 |
| 15.83 | 1 | Nucleotide exchange factor SIL1 \{ECO:0000313\|EMBL:KFZ57603.1\} |
| 15.81 | 1 | F-box only protein 21 |
| 15.77 | 1 | Bystin |
| 15.64 | 1 | Methenyltetrahydrofolate synthase domain-containing protein \{ECO:0000313\|EMBL:KFQ76690.1\} |
| 15.64 | 1 | Methenyltetrahydrofolate synthase domain-containing protein |
| 15.62 | 1 | Phosphatidylethanolamine-binding protein 4 \{ECO:0000313\|EMBL:JAB53705.1\} |
| 15.62 | 1 | Phosphatidylethanolamine-binding protein 4 |
| 15.61 | 1 | Antileukoproteinase-like 2 \{ECO:0000313\|EMBL:CAE51403.1\} |
| 15.58 | 2 | 1-phosphatidylinositol phosphodiesterase |
| 15.56 | 1 | Gliomedin |
| 15.23 | 2 | MLK-like mitogen-activated protein triple kinase beta \{ECO:0000313\|EMBL:BAM36483.1\} |
| 15.11 | 1 | EGF domain-specific O-linked N -acetylglucosamine transferase |
| 14.99 | 2 | Insulin-like growth factor-binding protein-like 1 |
| 14.98 | 1 | Zinc finger protein 32 |
| 14.94 | 3 | Neudesin |
| 14.90 | 1 | Collagen alpha-1(XIII) chain \{ECO:0000313\|Ensembl:ENSP00000388774\} |
| 14.90 | 1 | Collagen alpha-1(XIII) chain |
| 14.84 | 1 | Cathelicidin-3 |
| 14.81 | 2 | GATA zinc finger domain-containing protein 1 |
| 14.79 | 3 | LOC495510 protein \{ECO:0000313\|EMBL:AAH85209.1\} |
| 14.62 | 1 | Prokineticin-1 |
| 14.62 | 1 | Enterin neuropeptides |
| 14.58 | 1 | Neugrin |


| 14.58 | 1 | LOC779519 protein \{ECO:0000313\|EMBL:AAI67664.1\} |
| :---: | :---: | :---: |
| 14.51 | 2 | Neural cell adhesion molecule 1 |
| 14.51 | 2 | NCAM-140 \{ECO:0000313\|EMBL:BAA24033.1\} |
| 14.50 | 2 | Glutathione S-transferase P |
| 14.40 | 1 | ER degradation-enhancing alpha-mannosidase-like protein 2 |
| 14.40 | 1 | alpha-1,2-Mannosidase \{ECO:0000256\|RuleBase:RU361193\} |
| 14.35 | 1 | Protein Igkv19-93 \{ECO:0000313\|Ensembl:ENSMUSP00000139272\} |
| 14.35 | 1 | Ig kappa chain V-I region Rei |
| 14.35 | 1 | Goannatyrotoxin-Vere1 |
| 14.34 | 2 | C-C motif chemokine 28 |
| 14.32 | 1 | Serine protease HTR4 |
| 14.21 | 1 | UPF0669 protein C6orf120 homolog |
| 13.91 | 2 | Malate dehydrogenase, cytoplasmic \{ECO:0000313\|Ensembl:ENSP00000386719\} |
| 13.77 | 1 | Coiled-coil domain-containing protein 173 |
| 13.72 | 2 | Ig kappa chain V-VI region NQ2-17.4.1 |
| 13.69 | 2 | Succinate dehydrogenase assembly factor 3, mitochondrial \{ECO:0000250\|UniProtKB:Q04401\} |
| 13.69 | 1 | Mediator of RNA polymerase II transcription subunit 9 \{ECO:0000313\|EMBL:CCP79924.1\} |
| 13.69 | 1 | Mediator of RNA polymerase II transcription subunit 9 |
| 13.66 | 2 | Alcohol dehydrogenase class VII \{ECO:0000313\|EMBL:CCQ25770.1\} |
| 13.66 | 2 | Alcohol dehydrogenase 1 |
| 13.60 | 3 | Major urinary protein 4 |
| 13.50 | 1 | Sodium/bile acid cotransporter 4-like protein \{ECO:0000313\|EMBL:AFP09893.1\} |
| 13.50 | 1 | Sodium/bile acid cotransporter 4 |
| 13.46 | 2 | Riddle 2 \{ECO:0000313\|EMBL:AAO15687.1\} |
| 13.44 | 1 | Alpha-defensin 2 |
| 13.38 | 1 | Titin \{ECO:0000313\|EMBL:AGQ56078.1\} |
| 13.35 | 1 | UPF0696 protein C11orf68 homolog |
| 13.33 | 3 | Brorin |
| 13.33 | 1 | Meteorin-like protein |
| 13.33 | 1 | Meteorin-like \{ECO:0000313\|EMBL:KFQ46667.1\} |
| 13.32 | 2 | Phospholipase A2, major isoenzyme |
| 13.32 | 2 | Phospholipase A2 \{ECO:0000313\|EMBL:ELK15525.1\} |
| 13.30 | 2 | CDGSH iron-sulfur domain-containing protein 3, mitochondrial |
| 13.26 | 1 | Sin3 histone deacetylase corepressor complex component SDS3 |
| 13.20 | 3 | Serpin B9 |
| 13.19 | 2 | Zinc finger C3H1 domain-containing protein |
| 13.15 | 2 | Isthmin-1 \{ECO:0000313\|EMBL:EMP31749.1\} |
| 13.14 | 1 | Prominin-1.s13 splice variant \{ECO:0000313\|EMBL:ABB89202.1\} |
| 13.14 | 1 | Prominin-1 |
| 13.13 | 2 | C-C motif chemokine 19 |
| 13.10 | 1 | Probable phospholipid-transporting ATPase IH |
| 12.99 | 2 | Protein phosphatase inhibitor 2 |
| 12.97 | 1 | Retrovirus-related Pol polyprotein from type-1 retrotransposable element R2 |
| 12.95 | 3 | Angiopoietin-related protein 5 |
| 12.91 | 3 | Serotriflin |
| 12.89 | 1 | Periphilin-1 \{ECO:0000313\|EMBL:ELK08455.1\} |
| 12.89 | 1 | Periphilin-1 |
| 12.86 | 1 | Ryncolin-4 |
| 12.86 | 1 | Ficolin-1 \{ECO:0000313\|EMBL:EMP31096.1\} |
| 12.83 | 1 | Major urinary protein 17 |
| 12.78 | 1 | Prostate stem cell antigen |
| 12.74 | 1 | Membrane-associated protein \{ECO:0000313\|EMBL:EAP91103.1\} |
| 12.74 | 1 | Lon protease homolog 2, peroxisomal \{ECO:0000255\|HAMAP-Rule:MF_03121\} |
| 12.73 | 1 | Thioredoxin domain-containing secreted protein Ag1 \{ECO:0000313\|EMBL:AGG20198.1\} |
| 12.73 | 1 | Anterior gradient protein 1 |
| 12.69 | 2 | Peroxisomal leader peptide-processing protease \{ECO:0000313\|EMBL:EMP40605.1\} |
| 12.69 | 2 | Peroxisomal leader peptide-processing protease |
| 12.68 | 1 | Transforming growth factor beta-2 \{ECO:0000313\|Ensembl:ENSSSCP00000029638\} |
| 12.66 | 2 | Gamma-glutamyltranspeptidase 1 |
| 12.63 | 1 | LOC100124865 protein \{ECO:0000313\|EMBL:AAI35585.1\} |
| 12.59 | 1 | Lymphocyte antigen 6E |
| 12.59 | 1 | Ly6/PLAUR domain-containing protein 2 \{ECO:0000313\|EMBL:KFQ66330.1\} |
| 12.56 | 2 | Surfeit locus protein 1 |
| 12.50 | 2 | G-protein coupled receptor 143 |


| 12.48 | 2 | Cytoplasmic dynein 1 light intermediate chain 2 \{ECO:0000313\|EMBL:KFQ29862.1\} |
| :---: | :---: | :---: |
| 12.48 | 2 | Cytoplasmic dynein 1 light intermediate chain 2 |
| 12.47 | 2 | OX-2 membrane glycoprotein |
| 12.46 | 2 | Complement factor B |
| 12.41 | 3 | Nicotinamide phosphoribosyltransferase |
| 12.41 | 1 | LOC594871 protein \{ECO:0000313\|EMBL:AAH82898.1\} |
| 12.41 | 1 | Laminin subunit beta-3 |
| 12.39 | 2 | Protein Ighv9-1 \{ECO:0000313\|Ensembl:ENSMUSP00000100252\} |
| 12.39 | 2 | Ig heavy chain V-I region HG3 |
| 12.39 | 1 | Arylsulfatase G \{ECO:0000313\|EMBL:EHB10232.1\} |
| 12.39 | 1 | Arylsulfatase G |
| 12.38 | 1 | Zinc finger protein 417 |
| 12.36 | 3 | Cathelicidin-2 |
| 12.36 | 2 | Prolactin-2C3 |
| 12.35 | 2 | Chordin-like protein 1 |
| 12.25 | 2 | Apolipoprotein A-I \{ECO:0000313\|EMBL:AAH77663.1\} |
| 12.25 | 2 | Apolipoprotein A-I |
| 12.23 | 1 | Glypican-3 \{ECO:0000313\|EMBL:ELK04462.1\} |
| 12.17 | 2 | Placenta-specific protein 9 |
| 12.15 | 1 | ST3 beta-galactoside alpha-2,3-sialyltransferase 2 \{ECO:0000313\|EMBL:CAJ81569.1\} |
| 12.15 | 1 | CMP-N-acetylneuraminate-beta-galactosamide-alpha-2,3-sialyltransferase 2 |
| 12.13 | 2 | Probable G-protein coupled receptor 173 |
| 12.12 | 1 | Membrane-spanning 4-domains subfamily A member 8 |
| 12.11 | 4 | Cyclin-dependent kinase 13 |
| 12.11 | 2 | Steryl-sulfatase |
| 11.98 | 1 | Olfactomedin-like protein 2B |
| 11.80 | 3 | NXPE family member 4 |
| 11.79 | 1 | von Willebrand factor D and EGF domain-containing protein \{ECO:0000313\|EMBL:KFZ58890.1\} |
| 11.79 | 1 | Cytochrome P450 2A6 |
| 11.79 | 1 | Cytochrome P450 2A13 \{ECO:0000313\|EMBL:EPQ09308.1\} |
| 11.77 | 2 | Fucolectin-7 |
| 11.75 | 1 | Peroxiredoxin-1 \{ECO:0000313\|EMBL:ELK06103.1\} |
| 11.73 | 2 | Complement C2 isoform 1 \{ECO:0000313\|EMBL:ERE88814.1\} |
| 11.73 | 2 | Complement C2 |
| 11.65 | 1 | Endoplasmic reticulum aminopeptidase 1 |
| 11.65 | 1 | Adipocyte-derived leucine aminopeptidase variant \{ECO:0000313\|EMBL:BAD96593.1\} |
| 11.63 | 1 | MYC-induced nuclear antigen \{ECO:0000313\|EMBL:EMP37877.1\} |
| 11.63 | 1 | Bifunctional lysine-specific demethylase and histidyl-hydroxylase MINA |
| 11.62 | 1 | Fucolectin-3 |
| 11.53 | 3 | Niemann-Pick C1 protein \{ECO:0000313\|EMBL:KFZ53542.1\} |
| 11.51 | 2 | NADH dehydrogenase [ubiquinone] complex I, assembly factor 7 |
| 11.49 | 2 | Complement C1q tumor necrosis factor-related protein 3 |
| 11.46 | 1 | Pterin-4-alpha-carbinolamine dehydratase 2 |
| 11.44 | 1 | Solute carrier family 35 member F5 |
| 11.43 | 1 | Isthmin-1 |
| 11.38 | 1 | Oncorhynchus mykiss genomic scaffold, scaffold_439 \{ECO:0000313\|EMBL:CDQ70327.1\} |
| 11.37 | 2 | Trefoil factor 2 \{ECO:0000313\|EMBL:EMP31461.1\} |
| 11.30 | 1 | Lymphocyte antigen 6E \{ECO:0000313\|EMBL:KFW70679.1\} |
| 11.29 | 1 | Alpha-galactosidase A |
| 11.26 | 2 | Nucleostemin \{ECO:0000313\|EMBL:BAF31324.1\} |
| 11.26 | 2 | Guanine nucleotide-binding protein-like 3 |
| 11.20 | 2 | Endonuclease domain-containing 1 protein |
| 11.16 | 1 | Glucose-fructose oxidoreductase domain-containing protein 1 \{ECO:0000313\|EMBL:EOA97957.1\} |
| 11.16 | 1 | Glucose-fructose oxidoreductase domain-containing protein 1 |
| 11.03 | 1 | Liprin-beta-2 |
| 10.97 | 3 | Ribonuclease T2 |
| 10.94 | 2 | Alpha-1-antiproteinase F \{ECO:0000313\|EMBL:KFO74898.1\} |
| 10.87 | 1 | Serine protease inhibitor A3C |
| 10.86 | 2 | Ly6/PLAUR domain-containing protein 3 \{ECO:0000313\|EMBL:CCP76859.1\} |
| 10.86 | 2 | Ly6/PLAUR domain-containing protein 3 |
| 10.83 | 1 | Receptivity factor PRFA3 \{ECO:0000313\|EMBL:AAX21883.1\} |
| 10.75 | 2 | Protein Pri2c1 \{ECO:0000313\|Ensembl:ENSMUSP00000105964\} |
| 10.75 | 2 | Prolactin-2C2 |
| 10.75 | 1 | Dickkopf homolog 2 \{ECO:0000313\|EMBL:AAI69368.1\} |


| 10.74 | 2 | Titin \{ECO:0000313\|EMBL:AFM84561.1\} |
| :---: | :---: | :---: |
| 10.73 | 3 | Carboxypeptidase Q |
| 10.72 | 2 | Prolyl 3-hydroxylase OGFOD1 |
| 10.72 | 1 | Bone morphogenetic protein 2 |
| 10.69 | 1 | Hyaluronidase \{ECO:0000256\|RuleBase:RU610713\} |
| 10.69 | 1 | Hyaluronidase-1 |
| 10.67 | 3 | Amiloride-sensitive amine oxidase [copper-containing] |
| 10.67 | 1 | Protein TMEPAI |
| 10.65 | 3 | Angiopoietin-4 |
| 10.61 | 3 | Pro-neuregulin-1, membrane-bound isoform |
| 10.58 | 1 | Growth arrest-specific protein 6 |
| 10.58 | 1 | Growth arrest-specific 6 \{ECO:0000313\|EMBL:ABC55062.1\} |
| 10.55 | 1 | Peroxiredoxin-6 |
| 10.54 | 1 | Interferon alpha-inducible protein 6 |
| 10.49 | 2 | Testis, prostate and placenta-expressed protein |
| 10.46 | 1 | Trafficking kinesin-binding protein 1 |
| 10.46 | 1 | Myosin heavy chain \{ECO:0000313\|EMBL:AAH67305.1\} |
| 10.46 | 1 | Myosin-1 |
| 10.46 | 1 | LOC100135152 protein \{ECO:0000313\|EMBL:AAI55399.1\} |
| 10.46 | 1 | Cystatin \{ECO:0000313\|EMBL:KFV97917.1\} |
| 10.38 | 3 | Integumentary mucin B. 1 \{ECO:0000313\|EMBL:CAA69604.1\} |
| 10.35 | 2 | NAD-dependent protein deacetylase sirtuin-6 |
| 10.29 | 1 | ADP-ribose pyrophosphatase, mitochondrial |
| 10.22 | 2 | Proprotein convertase subtilisin/kexin type 5 \{ECO:0000313\|EMBL:KFQ11461.1\} |
| 10.22 | 2 | Proprotein convertase subtilisin/kexin type 5 |
| 10.21 | 1 | Apolipoprotein D |
| 10.19 | 1 | Laminin subunit alpha-1 \{ECO:0000313\|EMBL:KFP40866.1\} |
| 10.19 | 1 | Laminin subunit alpha-1 |
| 10.17 | 1 | Prenylcysteine oxidase-like \{ECO:0000313\|EMBL:AFE69874.1\} |
| 10.13 | 1 | Titin \{ECO:0000313\|EMBL:AAV93233.2\} |
| 10.05 | 1 | Acidic mammalian chitinase |
| 10.04 | 1 | Adenylate kinase 7 \{ECO:0000313\|EMBL:EMP30980.1\} |
| 10.04 | 1 | Adenylate kinase 7 |
| 10.01 | 1 | Protein FAM55D \{ECO:0000313\|EMBL:EGV93271.1\} |
| 10.00 | 2 | Noelin |
| 9.94 | 3 | MGC108277 protein \{ECO:0000313\|EMBL:AAH91048.1\} |
| 9.94 | 3 | Dermatopontin |
| 9.93 | 1 | Cofilin 1 (Non-muscle), isoform CRA_a \{ECO:0000313\|EMBL:EAW74448.1\} |
| 9.92 | 1 | Salivary lipocalin \{ECO:0000313\|EMBL:EHB13722.1\} |
| 9.90 | 1 | Dickkopf-related protein 1 \{ECO:0000313\|EMBL:KFQ97749.1\} |
| 9.90 | 1 | Dickkopf-related protein 1 |
| 9.89 | 1 | C-X-C motif chemokine \{ECO:0000256\|RuleBase:RU361149\} |
| 9.88 | 1 | Beta-1,3-galactosyl-O-glycosyl-glycoprotein beta-1,6-N-acetylglucosaminyltransferase \{ECO:0000313\|EMBL:ETE71302.1\} |
| 9.88 | 1 | Beta-1,3-galactosyl-O-glycosyl-glycoprotein beta-1,6-N-acetylglucosaminyltransferase |
| 9.87 | 1 | C-type lectin 3 \{ECO:0000313\|EMBL:AEU60005.1\} |
| 9.84 | 2 | Interferon alpha-5 |
| 9.80 | 1 | Coagulation factor VII \{ECO:0000313\|EMBL:AAI21657.1\} |
| 9.80 | 1 | Coagulation factor VII |
| 9.73 | 1 | Complement factor D \{ECO:0000313\|EMBL:ELK19287.1\} |
| 9.73 | 1 | Complement factor D |
| 9.71 | 1 | Fibronectin type III domain-containing protein 1 \{ECO:0000313\|EMBL:ETE68866.1\} |
| 9.71 | 1 | Fibronectin type III domain-containing protein 1 |
| 9.63 | 3 | Plasma kallikrein |
| 9.59 | 2 | Mitochondrial-processing peptidase subunit alpha |
| 9.56 | 4 | Protein FAM184A |
| 9.55 | 1 | Ryncolin-1 |
| 9.55 | 1 | E3 ubiquitin-protein ligase MARCH6 \{ECO:0000313\|EMBL:KFV48775.1\} |
| 9.55 | 1 | E3 ubiquitin-protein ligase MARCH6 \{ECO:0000313\|EMBL:KFQ83675.1\} |
| 9.55 | 1 | E3 ubiquitin-protein ligase MARCH6 |
| 9.44 | 2 | LOC100127729 protein \{ECO:0000313\|EMBL:AAI54911.1\} |
| 9.41 | 1 | Scavenger receptor cysteine-rich type 1 protein M130 \{ECO:0000250\|UniProtKB:Q86VB7\} |
| 9.41 | 1 | Inactive serine protease PAMR1 \{ECO:0000313\|EMBL:KFQ29677.1\} |
| 9.41 | 1 | Inactive serine protease PAMR1 |


| 9.40 | 2 | Sialoadhesin |
| :---: | :---: | :---: |
| 9.39 | 1 | Acylglycerol kinase, mitochondrial |
| 9.33 | 2 | Platelet-derived growth factor subunit A |
| 9.30 | 3 | Mitochondrial ribosomal protein L3 \{ECO:0000313\|EMBL:CAJ81305.1\} |
| 9.30 | 3 | 395 ribosomal protein L3, mitochondrial |
| 9.29 | 2 | LOC496762 protein \{ECO:0000313\|EMBL:AAH87749.1\} |
| 9.28 | 2 | von Willebrand factor A domain-containing protein 8 |
| 9.23 | 2 | Prostatic acid phosphatase |
| 9.23 | 1 | Acyl-CoA-binding protein |
| 9.23 | 1 | Ac1-130 \{ECO:0000313\|EMBL:AAP86251.1\} |
| 9.21 | 2 | D-3-phosphoglycerate dehydrogenase \{ECO:0000313\|EMBL:EMP34557.1\} |
| 9.21 | 2 | D-3-phosphoglycerate dehydrogenase |
| 9.21 | 1 | MGC131153 protein \{ECO:0000313\|EMBL:AAI08609.1\} |
| 9.21 | 1 | Lethal(3)malignant brain tumor-like protein 1 |
| 9.18 | 1 | Peptide YY |
| 9.17 | 1 | CDGSH iron-sulfur domain-containing protein 3, mitochondrial \{ECO:0000313\|EMBL:ETE68874.1\} |
| 9.13 | 1 | U2 snRNA gene \{ECO:0000313\|EMBL:CAA24954.2\} |
| 9.11 | 1 | Interleukin \{ECO:0000256\|RuleBase:RU003453\} |
| 9.11 | 1 | Interleukin-15 |
| 9.10 | 2 | Ectonucleotide pyrophosphatase/phosphodiesterase family member 2 |
| 9.10 | 1 | Collagen triple helix repeat-containing protein 1 \{ECO:0000313\|EMBL:KFV85123.1\} |
| 9.08 | 1 | Lectoxin-Vind1 \{ECO:0000313\|EMBL:ADK39270.1\} |
| 9.05 | 2 | DnaJ homolog subfamily C member 1 \{ECO:0000313\|EMBL:BAK63174.1\} |
| 9.05 | 2 | DnaJ homolog subfamily C member 1 |
| 9.02 | 1 | Macaca fascicularis brain cDNA clone: QorA-13060, similar to human malate dehydrogenase 1, NAD (soluble) (MDH1), mRNA, RefSeq: NM_005917.2 \{ECO:0000313\|EMBL:BAE88227.1\} |
| 8.99 | 1 | Radixin \{ECO:0000313\|EMBL:ELW64231.1\} |
| 8.96 | 2 | LOC398643 protein \{ECO:0000313\|EMBL:AAH54293.1\} |
| 8.94 | 2 | Amine oxidase \{ECO:0000256\|RuleBase:RU000672\} |
| 8.93 | 3 | Thrombospondin-2 \{ECO:0000313\|EMBL:KFW95348.1\} |
| 8.93 | 3 | Thrombospondin-2 |
| 8.88 | 1 | Tumor necrosis factor ligand superfamily member 13B |
| 8.88 | 1 | Retinol binding protein 4, plasma, isoform CRA_b \{ECO:0000313\|EMBL:EAW50068.1\} |
| 8.88 | 1 | Retinol-binding protein 4 |
| 8.85 | 1 | MGC114896 protein \{ECO:0000313\|EMBL:AAH97637.1\} |
| 8.83 | 1 | MHC class I heavy chain \{ECO:0000313\|EMBL:AAD44930.1\} |
| 8.83 | 1 | Major histocompatibility complex class I-related gene protein |
| 8.82 | 1 | MGC81714 protein \{ECO:0000313\|EMBL:AAH81034.1\} |
| 8.80 | 1 | tRNA (guanine-N(7)-)-methyltransferase non-catalytic subunit WDR4 \{ECO:0000256\|HAMAP-Rule:MF_03056\} |
| 8.80 | 1 | tRNA (guanine-N(7)--)-methyltransferase non-catalytic subunit WDR4 \{ECO:0000255\|HAMAP-Rule:MF_03056\} |
| 8.72 | 2 | Neurturin \{ECO:0000313\|EMBL:BAF46118.1\} |
| 8.72 | 2 | Neurturin |
| 8.70 | 1 | Interferon-induced guanylate-binding protein 1 |
| 8.66 | 2 | Membrane metallo-endopeptidase-like 1 |
| 8.66 | 2 | Hepatocyte growth factor activator |
| 8.66 | 1 | Midkine \{ECO:0000313\|EMBL:EGW03925.1\} |
| 8.66 | 1 | Midkine |
| 8.62 | 2 | Fibrillin-2 |
| 8.54 | 2 | Serine/threonine-protein phosphatase 6 regulatory subunit 1 |
| 8.54 | 2 | MGC69001 protein \{ECO:0000313\|EMBL:AAH59317.1\} |
| 8.54 | 1 | Iduronate 2-sulfatase |
| 8.52 | 1 | Aminopeptidase_b \{ECO:0000313\|EMBL:BAN89416.1\} |
| 8.50 | 2 | Apolipoprotein A-IV \{ECO:0000313\|EMBL:EMP35116.1\} |
| 8.50 | 2 | Apolipoprotein A-IV |
| 8.48 | 1 | Neurotensin/neuromedin N |
| 8.46 | 2 | Chemokine-like factor |
| 8.43 | 1 | Putative cation-transporting ATPase 13A4 \{ECO:0000313\|EMBL:KFV54901.1\} |
| 8.43 | 1 | Probable cation-transporting ATPase 13A4 |
| 8.33 | 2 | RING finger and SPRY domain-containing protein 1 |
| 8.31 | 1 | Protein Prl2c5 \{ECO:0000313\|Ensembl:ENSMUSP00000117522\} |
| 8.29 | 1 | Complement factor I |
| 8.28 | 1 | Retinol dehydrogenase 7 |
| 8.28 | 1 | MCG17030 \{ECO:0000313\|EMBL:EDL24526.1\} |
| 8.25 | 1 | Myomegalin \{ECO:0000313\|EMBL:KFV46184.1\} |


| 8.25 | 1 | Myomegalin |
| :---: | :---: | :---: |
| 8.25 | 1 | G-type lysozyme \{ECO:0000313\|EMBL:AEQ98812.1\} |
| 8.22 | 1 | ADM2 |
| 8.19 | 2 | Trimethylguanosine synthase \{ECO:0000313\|EMBL:ELW71973.1\} |
| 8.17 | 1 | Protein canopy 4 |
| 8.17 | 1 | Noelin \{ECO:0000313\|Ensembl:ENSP00000340318\} |
| 8.17 | 1 | Lipopolysaccharide-binding protein |
| 8.16 | 2 | Oviduct protein p20 \{ECO:0000313\|EMBL:AAS68624.1\} |
| 8.12 | 2 | Serpin B9 \{ECO:0000313\|EMBL:EGV99244.1\} |
| 8.11 | 1 | Tripartite motif-containing protein 59 |
| 8.11 | 1 | Transmembrane gamma-carboxyglutamic acid protein 4 \{ECO:0000313\|EMBL:KFW84618.1\} |
| 8.11 | 1 | Transmembrane gamma-carboxyglutamic acid protein 4 |
| 8.10 | 2 | Probable tRNA N6-adenosine threonylcarbamoyltransferase \{ECO:0000256\|HAMAP-Rule:MF_03180\} |
| 8.10 | 2 | Probable tRNA N6-adenosine threonylcarbamoyltransferase \{ECO:0000255\|HAMAP-Rule:MF_03180\} |
| 8.08 | 1 | Heat shock protein 47 \{ECO:0000313\|EMBL:AAF78479.1\} |
| 8.04 | 3 | Pancreatic alpha-amylase |
| 8.04 | 3 | Alpha-amylase \{ECO:0000256\|RuleBase:RU361134\} |
| 8.02 | 4 | Predicted protein \{ECO:0000313\|EMBL:EDO37701.1\} |
| 8.01 | 1 | G protein-coupled receptor 143 \{ECO:0000313\|EMBL:AAH82732.1\} |
| 7.93 | 2 | Cyclin-T2 \{ECO:0000313\|Ensembl:ENSP00000387436\} |
| 7.93 | 2 | Cyclin-T2 |
| 7.92 | 2 | MGC84260 protein \{ECO:0000313\|EMBL:AAH74366.1\} |
| 7.89 | 3 | Coagulation factor XIII B chain |
| 7.87 | 1 | C-factor \{ECO:0000313\|EMBL:KFO93365.1\} |
| 7.85 | 2 | Leukocyte elastase inhibitor |
| 7.85 | 2 | Alpha-1-antitrypsin |
| 7.78 | 1 | Metalloproteinase inhibitor 4 \{ECO:0000313\|EMBL:KFZ55204.1\} |
| 7.78 | 1 | Metalloproteinase inhibitor 4 |
| 7.78 | 1 | ATP-dependent RNA helicase DDX51 |
| 7.77 | 2 | Ficolin-2 |
| 7.77 | 2 | Cerebral dopamine neurotrophic factor |
| 7.74 | 3 | Olfactory receptor \{ECO:0000256\|RuleBase:RU000688\} |
| 7.73 | 1 | Cathelicidin-BF antimicrobial peptide |
| 7.73 | 1 | Arginase, non-hepatic 2 |
| 7.73 | 1 | Arginase \{ECO:0000256\|RuleBase:RU361159\} |
| 7.70 | 1 | Cell division cycle 2-like protein kinase 5 \{ECO:0000313\|EMBL:EOB04792.1\} |
| 7.69 | 2 | Gamma-interferon-inducible lysosomal thiol reductase |
| 7.67 | 3 | Acid sphingomyelinase-like phosphodiesterase 3a \{ECO:0000313\|EMBL:EMP29810.1\} |
| 7.67 | 3 | Acid sphingomyelinase-like phosphodiesterase 3a |
| 7.65 | 1 | Ectonucleotide pyrophosphatase/phosphodiesterase family member 6 \{ECO:0000313\|EMBL:KGL75977.1\} |
| 7.65 | 1 | Ectonucleotide pyrophosphatase/phosphodiesterase family member 6 |
| 7.60 | 1 | Cullin-3 \{ECO:0000313\|EMBL:KFQ03734.1\} |
| 7.60 | 1 | Cullin-3 |
| 7.59 | 1 | Prostaglandin-H2 D-isomerase |
| 7.58 | 1 | UBX domain-containing protein 7 |
| 7.57 | 1 | Desumoylating isopeptidase 1 |
| 7.57 | 1 | Angiotensin-converting enzyme |
| 7.57 | 1 | Ace protein \{ECO:0000313\|EMBL:AAH52531.1\} |
| 7.54 | 3 | Titin \{ECO:0000313\|EMBL:EOB05726.1\} |
| 7.51 | 1 | Chymotrypsinogen A |
| 7.50 | 1 | Putative methyltransferase NSUN3 |
| 7.50 | 1 | Coiled-coil domain-containing protein 147 |
| 7.49 | 1 | Ghrelin \{ECO:0000313\|EMBL:BAM29300.1\} |
| 7.49 | 1 | Appetite-regulating hormone |
| 7.48 | 1 | Interferon alpha-13 |
| 7.48 | 1 | Glioma pathogenesis-related protein 1 |
| 7.46 | 1 | Myomodulin neuropeptides |
| 7.45 | 2 | Bromodomain and PHD finger-containing protein 3 |
| 7.45 | 1 | Cathepsin B \{ECO:0000313\|EMBL:EGW10359.1\} |
| 7.44 | 1 | Otopetrin-1 |
| 7.43 | 2 | Angiopoietin-related protein 1 |
| 7.36 | 2 | Stromal cell-derived factor 2 |
| 7.36 | 1 | Nucleolar protein 12 \{ECO:0000313\|EMBL:EMP33125.1\} |
| 7.36 | 1 | Nucleolar protein 12 |


| 7.31 | 1 | BRCA1-associated ATM activator 1 |
| :---: | :---: | :---: |
| 7.27 | 1 | Alpha-1-antitrypsin-like protein GS55-LT |
| 7.24 | 1 | Secreted frizzled-related protein 1 \{ECO:0000313\|EMBL:EHB14559.1\} |
| 7.24 | 1 | Secreted frizzled-related protein 1 |
| 7.20 | 2 | Disintegrin and metalloproteinase domain-containing protein 17 |
| 7.17 | 2 | LOC100144991 protein \{ECO:0000313\|EMBL:AAI58332.1\} |
| 7.17 | 2 | Bile salt-activated lipase |
| 7.12 | 1 | Protein Ighv10-3 \{ECO:0000313\|Ensembl:ENSMUSP00000100276\} |
| 7.12 | 1 | Ig heavy chain V-III region VH26 |
| 7.11 | 2 | Protease-associated domain-containing protein 1 |
| 7.10 | 2 | Whey acidic protein |
| 7.03 | 1 | Sodium channel subunit beta-4 |
| 7.03 | 1 | Protein Gm21286 \{ECO:0000313\|Ensembl:ENSMUSP00000137167\} |
| 7.03 | 1 | Major urinary protein 5 |
| 6.96 | 2 | Beta-microseminoprotein \{ECO:0000313\|EMBL:EMP27053.1\} |
| 6.96 | 1 | Protein LYRIC \{ECO:0000313\|EMBL:ETE68625.1\} |
| 6.95 | 2 | Epididymis-specific alpha-mannosidase |
| 6.95 | 2 | Alpha-mannosidase \{ECO:0000256\|RuleBase:RU361199\} |
| 6.93 | 1 | Immunoglobulin superfamily member 10 |
| 6.88 | 1 | T-cell surface glycoprotein CD3 delta chain |
| 6.88 | 1 | CD3 gamma/delta chain protein \{ECO:0000313\|EMBL:AAL13251.1\} |
| 6.85 | 4 | Coagulation factor XIII A chain |
| 6.84 | 1 | Magnesium transporter protein 1 |
| 6.80 | 1 | Sulfated glycoprotein 1 \{ECO:0000313\|EMBL:EGV98207.1\} |
| 6.75 | 1 | Ras GTPase-activating-like protein IQGAP2 |
| 6.74 | 1 | Serine/threonine-protein phosphatase \{ECO:0000256\|RuleBase:RU004273\} |
| 6.74 | 1 | Serine/threonine-protein phosphatase 2B catalytic subunit beta isoform |
| 6.66 | 1 | Serpin B8 |
| 6.64 | 1 | Vitamin K-dependent protein S |
| 6.64 | 1 | Protein S (Alpha) \{ECO:0000313\|EMBL:CAJ83353.1\} |
| 6.58 | 3 | Granzyme A \{ECO:0000313\|EMBL:KGL85525.1\} |
| 6.55 | 1 | Inhibin beta A chain |
| 6.55 | 1 | Activin beta-A subunit \{ECO:0000313\|EMBL:BAA12693.1\} |
| 6.54 | 1 | Telethonin \{ECO:0000313\|EMBL:ETE72874.1\} |
| 6.54 | 1 | Telethonin |
| 6.52 | 1 | LOC734198 protein \{ECO:0000313\|EMBL:AAI60743.1\} |
| 6.51 | 2 | Heparanase |
| 6.51 | 1 | Laminin subunit alpha-3 |
| 6.51 | 1 | High mobility group protein B2 |
| 6.50 | 1 | Heterochromatin-associated protein MENT \{ECO:0000313\|EMBL:KFM07484.1\} |
| 6.47 | 2 | R-spondin-3 \{ECO:0000313\|EMBL:KFO62372.1\} |
| 6.47 | 2 | R-spondin-3 |
| 6.46 | 1 | Cytochrome P450 2G1 |
| 6.40 | 2 | Actin-related protein 2 |
| 6.40 | 1 | Probable RNA-binding protein 19 |
| 6.39 | 1 | CRS1C-5 \{ECO:0000313\|EMBL:ACF04795.1\} |
| 6.39 | 1 | Alpha-defensin-related sequence 1 |
| 6.38 | 1 | Osteocalcin \{ECO:0000256\|RuleBase:RU361261\} |
| 6.38 | 1 | Osteocalcin |
| 6.38 | 1 | Myeloid-associated differentiation marker like protein \{ECO:0000313\|EMBL:EMP25559.1\} |
| 6.38 | 1 | Myeloid-associated differentiation marker homolog |
| 6.35 | 1 | NXPE family member 2 |
| 6.33 | 1 | Extracellular sulfatase Sulf-1 \{ECO:0000313\|EMBL:KFV06832.1\} |
| 6.33 | 1 | Extracellular sulfatase Sulf-1 |
| 6.32 | 1 | SUMO-specific isopeptidase USPL1 |
| 6.30 | 2 | Protein FAM92A1-B |
| 6.29 | 1 | Stromal cell-derived factor 2 \{ECO:0000313\|EMBL:CCP78400.1\} |
| 6.25 | 1 | Inactive serine protease 35 \{ECO:0000313\|EMBL:KFV15493.1\} |
| 6.25 | 1 | Inactive serine protease 35 |
| 6.23 | 1 | LOC398864 protein \{ECO:0000313\|EMBL:AAI24852.1\} |
| 6.21 | 1 | C-C motif chemokine 25 |
| 6.18 | 1 | Fibrillin-2 \{ECO:0000313\|EMBL:KFQ12662.1\} |
| 6.16 | 1 | Serpin A3-4 |
| 6.15 | 1 | Glypican-3 \{ECO:0000313\|EMBL:KFR16127.1\} |


| 6.14 | 2 | Bone morphogenetic protein 1 |
| :---: | :---: | :---: |
| 6.09 | 1 | Tachylectin-2 |
| 6.09 | 1 | Receptivity factor PRFB5 \{ECO:0000313\|EMBL:AAX21847.1\} |
| 6.08 | 2 | MGC116563 protein \{ECO:0000313\|EMBL:AAH99367.1\} |
| 6.07 | 2 | Transcription activator BRG1 |
| 6.07 | 1 | Protein FAM198A |
| 6.07 | 1 | Ectonucleotide pyrophosphatase/phosphodiesterase family member 2 \{ECO:0000313\|EMBL:KFZ60494.1\} |
| 6.07 | 1 | Carboxypeptidase N catalytic chain \{ECO:0000313\|EMBL:KFW88670.1\} |
| 6.06 | 2 | Radixin \{ECO:0000313\|EMBL:KFO14586.1\} |
| 6.04 | 1 | Interleukin 17B \{ECO:0000313\|EMBL:AAH75405.1\} |
| 6.04 | 1 | Interleukin-17B |
| 6.03 | 2 | Protein FAM184A \{ECO:0000313\|EMBL:KFQ60354.1\} |
| 6.03 | 1 | Dehydrogenase/reductase SDR family member 7C |
| 5.97 | 1 | Xaa-Pro aminopeptidase 2 |
| 5.97 | 1 | Ubiquitin carboxyl-terminal hydrolase 16 \{ECO:0000255\|HAMAP-Rule:MF_03062\} |
| 5.95 | 2 | Angiopoietin-related protein 7 |
| 5.95 | 1 | Keratin, type I cytoskeletal 42 |
| 5.92 | 1 | PCTP-like protein |
| 5.90 | 1 | Protein Frmd7 \{ECO:0000313\|Ensembl:ENSRNOP00000037504\} |
| 5.90 | 1 | FERM domain-containing protein 7 |
| 5.88 | 1 | Prolyl endopeptidase FAP \{ECO:0000250\|UniProtKB:Q12884\} |
| 5.81 | 2 | Serine/threonine-protein kinase Chk1 |
| 5.81 | 2 | Noelin-2 |
| 5.81 | 2 | LOC100124837 protein \{ECO:0000313\|EMBL:AAI35504.1\} |
| 5.81 | 2 | CHEK1 protein \{ECO:0000313\|EMBL:AAI40692.1\} |
| 5.77 | 2 | Alpha-actinin-4 |
| 5.73 | 1 | Mitogen-activated protein kinase kinase kinase MLT |
| 5.72 | 2 | Membrane-bound transcription factor site-2 protease \{ECO:0000313\|EMBL:KFW12494.1\} |
| 5.72 | 2 | Membrane-bound transcription factor site-2 protease |
| 5.71 | 1 | Selenoprotein P \{ECO:0000313\|Ensembl:ENSMUSP00000125505\} |
| 5.71 | 1 | Selenoprotein P |
| 5.68 | 2 | Ribonuclease T2 \{ECO:0000313\|EMBL:AC167553.1\} |
| 5.68 | 1 | Tolloid-like protein 2 |
| 5.67 | 1 | Trafficking protein particle complex subunit 12 |
| 5.66 | 1 | Peroxiredoxin-1 \{ECO:0000313\|Ensembl:ENSP00000361150\} |
| 5.64 | 1 | Putative lysosomal acid lipase/cholesteryl ester hydrolase |
| 5.64 | 1 | Lysosomal acid lipase/cholesteryl ester hydrolase \{ECO:0000313\|EMBL:KFW90192.1\} |
| 5.63 | 2 | Kelch-like protein 34 |
| 5.62 | 1 | Pre-rRNA processing protein FTSJ3 \{ECO:0000313\|EMBL:KFV81839.1\} |
| 5.62 | 1 | pre-rRNA processing protein FTSJ3 \{ECO:0000255\|HAMAP-Rule:MF_03163\} |
| 5.55 | 1 | Alpha-defensin 5 |
| 5.54 | 1 | Prostatic spermine-binding protein \{ECO:0000313\|Ensembl:ENSMUSP00000138338\} |
| 5.54 | 1 | Prostatic spermine-binding protein |
| 5.52 | 1 | P-selectin |
| 5.50 | 1 | Serotransferrin-B |
| 5.50 | 1 | Ovotransferrin \{ECO:0000313\|EMBL:KFQ88597.1\} |
| 5.47 | 1 | Protein Igkv15-103 \{ECO:0000313\|Ensembl:ENSMUSP00000100125\} |
| 5.47 | 1 | Ig kappa chain V-I region Roy |
| 5.45 | 2 | Protein NDNF |
| 5.44 | 1 | Ectonucleotide pyrophosphatase/phosphodiesterase family member 3 |
| 5.42 | 1 | Follistatin-related protein 1 \{ECO:0000313\|EMBL:EHB00602.1\} |
| 5.42 | 1 | Follistatin-related protein 1 |
| 5.42 | 1 | Coagulation factor XI |
| 5.40 | 3 | Granzyme A |
| 5.40 | 2 | Olfactory receptor 1537 |
| 5.39 | 1 | Myosin-9 |
| 5.37 | 1 | Titin \{ECO:0000313\|EMBL:AGQ56068.1\} |
| 5.37 | 1 | Titin \{ECO:0000313\|EMBL:AFM84557.1\} |
| 5.36 | 1 | LOC100145320 protein \{ECO:0000313\|EMBL:AAI60516.1\} |
| 5.33 | 1 | Fibroblast growth factor 7 |
| 5.33 | 1 | Chromogranin-A \{ECO:0000313\|EMBL:KFP62931.1\} |
| 5.33 | 1 | Chromogranin-A |
| 5.28 | 1 | Serpina1 protein \{ECO:0000313\|EMBL:AAH62869.1\} |
| 5.27 | 1 | Succinate dehydrogenase assembly factor 1, mitochondrial \{ECO:0000250\|UniProtKB:A6NFY7\} |


| 5.25 | 1 | Ig kappa chain C region \{ECO:0000313\|Ensembl:ENSP00000478196\} |
| :---: | :---: | :---: |
| 5.20 | 1 | Sterol regulatory element-binding protein 1 |
| 5.15 | 1 | Collagen alpha-1(V) chain \{ECO:0000313\|EMBL:CCP84683.1\} |
| 5.15 | 1 | Collagen alpha-1(V) chain |
| 5.11 | 1 | Leukocyte elastase inhibitor A \{ECO:0000313\|Ensembl:ENSMUSP00000089257\} |
| 5.11 | 1 | Leukocyte elastase inhibitor A |
| 5.11 | 1 | Krev interaction trapped protein 1 \{ECO:0000313\|EMBL:KFQ40278.1\} |
| 5.08 | 1 | Protein Gm11397 \{ECO:0000313\|Ensembl:ENSMUSP00000049819\} |
| 5.08 | 1 | Apolipoprotein B-100 |
| 5.06 | 2 | Steel factor \{ECO:0000313\|EMBL:AAD17253.1\} |
| 5.06 | 2 | Kit ligand |
| 5.04 | 1 | Slit homolog 2 protein |
| 5.02 | 1 | Phospholipid transfer protein |
| 5.01 | 2 | Death associated protein 3 \{ECO:0000313\|EMBL:AAI66942.1\} |
| 5.01 | 2 | 285 ribosomal protein S29, mitochondrial |
| 5.00 | 2 | Anthrax toxin receptor 2 \{ECO:0000313\|EMBL:EMP25861.1\} |
| 5.00 | 2 | Anthrax toxin receptor 2 |
| 4.96 | 1 | Protein CYR61 |
| 4.96 | 1 | Cysteine-rich angiogenic inducer 61 \{ECO:0000313\|EMBL:AFP43734.1\} |
| 4.94 | 1 | Tumor necrosis factor receptor superfamily member 1A \{ECO:0000313\|EMBL:EMP33636.1\} |
| 4.94 | 1 | Tumor necrosis factor receptor superfamily member 1A |
| 4.94 | 1 | Keratin, type II cytoskeletal 6A |
| 4.94 | 1 | Keratin 5 \{ECO:0000313\|EMBL:AFN68290.1\} |
| 4.92 | 1 | Protease-associated domain-containing protein 1 \{ECO:0000313\|EMBL:KFW87745.1\} |
| 4.90 | 1 | Trifunctional purine biosynthetic protein adenosine-3 |
| 4.90 | 1 | Glycinamide ribonucleotide synthetase-aminoimidazole ribonucleotide synthetase-glycinamide ribonucleotide transformylase \{ECO:0000313\|EMBL:BAD17892.1\} |
| 4.90 | 1 | Desmin |
| 4.89 | 1 | Glycogen phosphorylase, liver form |
| 4.88 | 1 | NAD(P)(+)--arginine ADP-ribosyltransferase \{ECO:0000256\|RuleBase:RU361228\} |
| 4.88 | 1 | NAD(P)(+)--arginine ADP-ribosyltransferase 2 |
| 4.87 | 1 | Oncorhynchus mykiss genomic scaffold, scaffold_10903 \{ECO:0000313\|EMBL:CDQ93435.1\} |
| 4.86 | 1 | Zinc finger CCHC domain-containing protein 3 |
| 4.85 | 1 | Lymphocyte antigen 96 \{ECO:0000313\|EMBL:KFP38876.1\} |
| 4.85 | 1 | Lymphocyte antigen 96 |
| 4.82 | 1 | Protein Gm14744 \{ECO:0000313\|Ensembl:ENSMUSP00000109673\} |
| 4.82 | 1 | Odorant-binding protein |
| 4.82 | 1 | L-amino acid oxidase \{ECO:0000313\|EMBL:JAC95028.1\} |
| 4.79 | 1 | Semaphorin-4D |
| 4.79 | 1 | Semaphorin-3D \{ECO:0000313\|EMBL:KFO13189.1\} |
| 4.78 | 1 | Steryl-sulfatase \{ECO:0000313\|EMBL:ELK30206.1\} |
| 4.77 | 1 | LOC100145251 protein \{ECO:0000313\|EMBL:AAI59335.1\} |
| 4.77 | 1 | Biotinidase \{ECO:0000313\|EMBL:EHB10484.1\} |
| 4.77 | 1 | Biotinidase |
| 4.77 | 1 | Amphiregulin |
| 4.76 | 1 | Ig lambda chain V region 4A \{ECO:0000313\|EMBL:ELK35251.1\} |
| 4.76 | 1 | Ig lambda chain V region 4A |
| 4.72 | 1 | Vespryn |
| 4.72 | 1 | Ig lambda chain V-VI region SUT |
| 4.70 | 1 | Trafficking protein particle complex subunit 8 \{ECO:0000313\|EMBL:EMP36723.1\} |
| 4.70 | 1 | Trafficking protein particle complex subunit 8 |
| 4.70 | 1 | Bone morphogenetic protein 1 \{ECO:0000313\|EMBL:ETE62418.1\} |
| 4.69 | 1 | Plastin-2 \{ECO:0000313\|EMBL:KFQ27057.1\} |
| 4.68 | 1 | Thrombospondin-1 |
| 4.68 | 1 | Tetratricopeptide repeat protein 27 |
| 4.66 | 2 | Laminin, beta 1 \{ECO:0000313\|EMBL:EPY80804.1\} |
| 4.66 | 1 | Neuferricin |
| 4.65 | 2 | Vomeronasal type-2 receptor 116 \{ECO:0000312\|MGI:MGI:3646674\} |
| 4.65 | 2 | Protein Vmn2r89 \{ECO:0000313\|Ensembl:ENSMUSP00000124065\} |
| 4.64 | 2 | EMILIN-2 |
| 4.64 | 1 | NGF-VAR1 \{ECO:0000313\|EMBL:AAZ75634.1\} |
| 4.64 | 1 | Beta-nerve growth factor |
| 4.63 | 2 | Cathelicidin-AL \{ECO:0000313\|EMBL:AEI69698.1\} |
| 4.63 | 1 | Urokinase-type plasminogen activator \{ECO:0000313\|EMBL:ETE66008.1\} |


| 4.63 | 1 | Urokinase-type plasminogen activator |
| :---: | :---: | :---: |
| 4.61 | 1 | Scavenger receptor cysteine-rich type 1 protein M160 |
| 4.60 | 1 | Ubinuclein-2 \{ECO:0000313\|EMBL:KFM02226.1\} |
| 4.60 | 1 | Ubinuclein-2 |
| 4.59 | 1 | Oncorhynchus mykiss genomic scaffold, scaffold_344 \{ECO:0000313\|EMBL:CDQ68175.1\} |
| 4.58 | 1 | Xanthine dehydrogenase/oxidase |
| 4.57 | 1 | Mucin-6 \{ECO:0000313\|Ensembl:ENSMUSP00000140483\} |
| 4.57 | 1 | Mucin-6 |
| 4.56 | 1 | Regulator of nonsense transcripts 3B |
| 4.55 | 1 | ARP2 actin-related protein 2 homolog (Yeast), isoform CRA_d \{ECO:0000313\|EMBL:EAW99914.1\} |
| 4.54 | 1 | FLYWCH-type zinc finger-containing protein 1 |
| 4.48 | 1 | Collagen alpha-1(XXIV) chain |
| 4.48 | 1 | Apolipoprotein B |
| 4.46 | 2 | Alpha-2-macroglobulin receptor-associated protein \{ECO:0000313\|EMBL:KFO09232.1\} |
| 4.45 | 1 | Polypeptide N-acetylgalactosaminyltransferase \{ECO:0000256\|RuleBase:RU361242\} |
| 4.45 | 1 | Polypeptide N -acetylgalactosaminyltransferase 1 |
| 4.43 | 2 | YALIOC04136p \{ECO:0000313\|EMBL:CAG81726.1\} |
| 4.43 | 2 | Sodium channel protein type 8 subunit alpha |
| 4.43 | 2 | Radial spoke head protein 3 homolog |
| 4.40 | 1 | Choline transporter-like protein 3 \{ECO:0000313\|EMBL:AFP00439.1\} |
| 4.40 | 1 | Choline transporter-like protein 3 |
| 4.38 | 2 | Myocilin |
| 4.38 | 1 | Renin |
| 4.37 | 1 | Lipase member M |
| 4.34 | 1 | MKIAA0548 protein \{ECO:0000313\|EMBL:BAC41429.1\} |
| 4.34 | 1 | Attractin |
| 4.33 | 1 | Solute carrier family 2, facilitated glucose transporter member 11 \{ECO:0000313\|EMBL:KFO23713.1\} |
| 4.33 | 1 | NACHT, LRR and PYD domains-containing protein 12 |
| 4.31 | 2 | Butyrophilin subfamily 3 member A3 |
| 4.31 | 2 | Butyrophilin subfamily 3 member A1 |
| 4.28 | 1 | Immnogloblin light chain variable region \{ECO:0000313\|EMBL:BAA32080.1\} |
| 4.28 | 1 | Ig kappa chain V-V region MOPC 41 |
| 4.27 | 1 | cDNA FL50376, highly similar to Interleukin-1 receptor-associated kinase 4 (EC 2.7.11.1) \{ECO:0000313\|EMBL:BAG65371.1\} |
| 4.26 | 1 | Oncorhynchus mykiss genomic scaffold, scaffold_89 \{ECO:0000313\|EMBL:CDQ60620.1\} |
| 4.26 | 1 | LOC100135075 protein \{ECO:0000313\|EMBL:AAI55681.1\} |
| 4.25 | 1 | Ig lambda chain V-III region SH |
| 4.25 | 1 | Complement component C7 \{ECO:0000313\|EMBL:KFV84312.1\} |
| 4.24 | 1 | Pancreatic secretory trypsin inhibitor \{ECO:0000313\|EMBL:ELK17240.1\} |
| 4.23 | 1 | Zinc finger CCCH domain-containing protein 13 |
| 4.22 | 1 | Collagen, type II, alpha 1 \{ECO:0000313\|EMBL:AEZ53727.1\} |
| 4.19 | 1 | Heregulin-gamma protein isoform 1 \{ECO:0000313\|EMBL:ABY66348.1\} |
| 4.19 | 1 | Growth hormone receptor \{ECO:0000313\|EMBL:AlI23392.1\} |
| 4.19 | 1 | Cingulin-like protein 1 |
| 4.18 | 1 | Zinc finger and BTB domain-containing protein 24 |
| 4.17 | 1 | Radial spoke head 10 like protein B \{ECO:0000313\|EMBL:EMP34980.1\} |
| 4.17 | 1 | Radial spoke head 10 homolog B |
| 4.15 | 1 | Trypsin-4 |
| 4.15 | 1 | Insulin-like growth factor-binding protein 5 |
| 4.13 | 1 | Titin \{ECO:0000313\|EMBL:KFO27993.1\} |
| 4.12 | 2 | Transcobalamin-1 |
| 4.12 | 2 | Cystatin-F |
| 4.12 | 1 | tRNA (Adenine(58)-N(1))-methyltransferase non-catalytic subunit TRM6 \{ECO:0000313\|EMBL:EPY79157.1\} |
| 4.12 | 1 | tRNA (adenine(58)-N(1))-methyltransferase non-catalytic subunit TRM6 |
| 4.12 | 1 | Insulin-like growth factor I |
| 4.11 | 1 | Peroxisome proliferator-activated receptor gamma \{ECO:0000313\|EMBL:ELW46949.1\} |
| 4.11 | 1 | Peroxisome proliferator-activated receptor gamma |
| 4.10 | 1 | Otospiralin |
| 4.09 | 1 | Nerve growth factor beta polypeptide \{ECO:0000313\|EMBL:ACF32199.1\} |
| 4.09 | 1 | Nerve growth factor |
| 4.06 | 1 | Ovalbumin \{ECO:0000303\|PubMed:21058653\} |
| 4.06 | 1 | Leukocyte elastase inhibitor B \{ECO:0000313\|EMBL:KFP30396.1\} |
| 4.05 | 1 | Putative tRNA (Uracil-O(2)-)-methyltransferase \{ECO:0000313\|EMBL:EMP26433.1\} |
| 4.05 | 1 | Protein Prl2c1 \{ECO:0000313\|Ensembl:ENSMUSP00000078221\} |


| 4.05 | 1 | Probable tRNA (uracil-O(2)-)-methyltransferase |
| :---: | :---: | :---: |
| 4.05 | 1 | Collagen alpha-6(V) chain |
| 4.03 | 1 | Neuregulin 1 \{ECO:0000313\|EMBL:ADN26625.1\} |
| 4.02 | 1 | Actn4-prov protein \{ECO:0000313\|EMBL:AAH77918.1\} |
| 4.01 | 1 | Adipocyte plasma membrane-associated protein |
| 3.99 | 1 | Zinc finger protein 81 |
| 3.99 | 1 | Protein Gm13248 \{ECO:0000313\|Ensembl:ENSMUSP00000121301\} |
| 3.99 | 1 | Crustapain |
| 3.99 | 1 | Counting factor associated protein D-like protein \{ECO:0000313\|EMBL:AFP04248.1\} |
| 3.97 | 1 | YLP motif-containing protein 1 \{ECO:0000313\|EMBL:EGW13062.1\} |
| 3.96 | 1 | Putative endothelial lipase |
| 3.96 | 1 | Endothelial lipase \{ECO:0000313\|EMBL:KFR16037.1\} |
| 3.96 | 1 | Collagen alpha-1(XX) chain \{ECO:0000313\|EMBL:KFM03908.1\} |
| 3.96 | 1 | Collagen alpha-1(XX) chain |
| 3.96 | 1 | Cartilage matrix protein |
| 3.95 | 1 | Ig lambda chain V-IV region Bau |
| 3.93 | 1 | Small inducible cytokine SCYA107 \{ECO:0000313\|EMBL:AAO21210.1\} |
| 3.93 | 1 | MGC130869 protein \{ECO:0000313\|EMBL:AAI08508.1\} |
| 3.93 | 1 | Complement C1r-B subcomponent |
| 3.93 | 1 | CC chemokine \{ECO:0000313\|EMBL:AFP13593.1\} |
| 3.91 | 1 | Tyrosine-protein kinase receptor UFO |
| 3.91 | 1 | Collagen alpha-1(XXVIII) chain |
| 3.91 | 1 | Axl protein \{ECO:0000313\|EMBL:AAI27292.1\} |
| 3.91 | 1 | Angiopoietin-4 \{ECO:0000313\|EMBL:ELV13579.1\} |
| 3.90 | 1 | Potassium voltage-gated channel subfamily S member 1 |
| 3.88 | 2 | Tissue factor pathway inhibitor |
| 3.88 | 1 | Prostaglandin F2-alpha receptor |
| 3.87 | 1 | Nucleobindin-1 |
| 3.86 | 1 | Tyrosine-protein kinase Mer |
| 3.83 | 1 | Angiopoietin-4 \{ECO:0000313\|EMBL:EMP33264.1\} |
| 3.83 | 1 | 3-ketodihydrosphingosine reductase |
| 3.82 | 1 | Titin \{ECO:0000313\|Ensembl:ENSMUSP00000116031\} |
| 3.81 | 1 | Vascular endothelial growth factor C \{ECO:0000313\|EMBL:KFP43744.1\} |
| 3.81 | 1 | Vascular endothelial growth factor C |
| 3.81 | 1 | Angiotensin-converting enzyme 2 |
| 3.79 | 2 | Angiopoietin-1 receptor \{ECO:0000313\|EMBL:ELW70235.1\} |
| 3.79 | 2 | Angiopoietin-1 receptor |
| 3.79 | 1 | Frizzled-6 |
| 3.78 | 2 | Coiled-coil domain-containing protein 124-A \{ECO:0000313\|EMBL:KFQ36476.1\} |
| 3.78 | 2 | Coiled-coil domain-containing protein 124 |
| 3.78 | 1 | Serine incorporator 3 \{ECO:0000313\|EMBL:KFP04108.1\} |
| 3.78 | 1 | Serine incorporator 3 |
| 3.77 | 1 | Hemopexin \{ECO:0000313\|EMBL:ACP30429.1\} |
| 3.77 | 1 | Hemopexin \{ECO:0000250\|UniProtKB:P20058\} |
| 3.76 | 2 | Dickkopf-related protein 3 \{ECO:0000313\|EMBL:KFM13907.1\} |
| 3.76 | 2 | Dickkopf-related protein 3 |
| 3.73 | 1 | Prolactin-3D1 |
| 3.73 | 1 | Placental lactogen-I beta \{ECO:0000313\|EMBL:ABF65993.1\} |
| 3.73 | 1 | Oncorhynchus mykiss genomic scaffold, scaffold_3739 \{ECO:0000313\|EMBL:CDQ87580.1\} |
| 3.73 | 1 | Neuronal acetylcholine receptor subunit alpha-7 |
| 3.73 | 1 | LOC446301 protein \{ECO:0000313\|EMBL:AAI28918.1\} |
| 3.73 | 1 | General vesicular transport factor p115 |
| 3.73 | 1 | Fukutin-related protein \{ECO:0000313\|EMBL:EPQ02470.1\} |
| 3.73 | 1 | Fukutin-related protein |
| 3.72 | 1 | Smarca4 protein \{ECO:0000313\|EMBL:AAH61214.1\} |
| 3.71 | 1 | Novel protein containing SPRY domain-containing SOCS box protein SSB-3 \{ECO:0000313\|EMBL:CAJ81703.1\} |
| 3.71 | 1 | Nidogen-2 |
| 3.71 | 1 | Murinoglobulin-2 |
| 3.70 | 1 | Chemokine-like factor-like protein \{ECO:0000313\|EMBL:ERE77304.1\} |
| 3.68 | 2 | MCG140531 \{ECO:0000313\|EMBL:EDL02195.1\} |
| 3.68 | 1 | Leukocyte elastase inhibitor \{ECO:0000313\|EMBL:EMP36655.1\} |
| 3.67 | 1 | Secretogranin-1 \{ECO:0000313\|EMBL:EMP27348.1\} |
| 3.67 | 1 | Secretogranin-1 |
| 3.65 | 1 | Myocilin \{ECO:0000313\|EMBL:EQB78941.1\} |


| 3.60 | 1 | Histo-blood group ABO system transferase |
| :---: | :---: | :---: |
| 3.59 | 2 | Transmembrane protein 19 \{ECO:0000313\|EMBL:KFQ30134.1\} |
| 3.59 | 2 | Transmembrane protein 19 |
| 3.58 | 2 | Peroxidasin |
| 3.58 | 1 | Galectin-7 |
| 3.57 | 1 | Hepatocyte growth factor activator \{ECO:0000313\|EMBL:KFW91820.1\} |
| 3.57 | 1 | Chymotrypsin-C |
| 3.53 | 1 | Pro-neuregulin-4, membrane-bound isoform |
| 3.49 | 2 | Coxsackievirus and adenovirus receptor \{ECO:0000313\|EMBL:KFQ24077.1\} |
| 3.49 | 1 | Interleukin-23 subunit alpha |
| 3.48 | 2 | Complement component C8 beta chain \{ECO:0000313\|EMBL:KFV48722.1\} |
| 3.48 | 2 | Complement component C8 beta chain |
| 3.48 | 1 | Carboxylesterase 4A |
| 3.46 | 2 | Serum albumin \{ECO:0000313\|EMBL:AFM52325.1\} |
| 3.46 | 2 | Serum albumin B |
| 3.46 | 1 | Cysteine-rich secretory protein 2 \{ECO:0000313\|EMBL:KGL79382.1\} |
| 3.44 | 1 | Tripartite motif-containing protein 7 \{ECO:0000313\|EMBL:EMP25059.1\} |
| 3.44 | 1 | Disks large homolog 3 \{ECO:0000313\|Ensembl:ENSRNOP00000068233\} |
| 3.44 | 1 | Disks large homolog 3 |
| 3.44 | 1 | Acidic repeat-containing protein |
| 3.41 | 1 | Group 10 secretory phospholipase A2 |
| 3.40 | 1 | Granzyme M |
| 3.39 | 1 | Semaphorin-3E |
| 3.39 | 1 | Pleckstrin homology domain-containing family H member 3 \{ECO:0000313\|EMBL:ELW54985.1\} |
| 3.39 | 1 | Pleckstrin homology domain-containing family H member 3 |
| 3.39 | 1 | Coagulation factor XIII A chain \{ECO:0000313\|EMBL:EMP26970.1\} |
| 3.38 | 1 | Zinc finger CCHC domain-containing protein 4 |
| 3.38 | 1 | Interleukin-17 receptor A |
| 3.37 | 1 | Disintegrin and metalloproteinase domain-containing protein 28 |
| 3.36 | 1 | Ribonucleases P/MRP protein subunit POP1 |
| 3.33 | 1 | Neudesin \{ECO:0000313\|EMBL:EMP37434.1\} |
| 3.33 | 1 | Cyclin-dependent kinase 7 \{ECO:0000313\|Ensembl:ENSP00000422737\} |
| 3.33 | 1 | Cyclin-dependent kinase 7 |
| 3.31 | 1 | Protein CTLA-2-alpha |
| 3.31 | 1 | Prostatic acid phosphatase \{ECO:0000313\|EMBL:KFQ98040.1\} |
| 3.30 | 1 | Interferon alpha 11 \{ECO:0000313\|EMBL:AAO63596.1\} |
| 3.30 | 1 | Interferon alpha-1 |
| 3.30 | 1 | Cathelin-related peptide SC5 |
| 3.27 | 1 | Kinesin-like protein KIF2C |
| 3.27 | 1 | Kinesin-like protein \{ECO:0000256\|RuleBase:RU000394\} |
| 3.26 | 1 | Chordin-like protein 2 |
| 3.25 | 2 | Protein phosphatase methylesterase 1 \{ECO:0000256\|PIRNR:PIRNR022950\} |
| 3.25 | 2 | Protein phosphatase methylesterase 1 |
| 3.24 | 1 | Tumor necrosis factor |
| 3.22 | 2 | Podocan |
| 3.22 | 1 | Angiopoietin-2 |
| 3.20 | 1 | Laminin subunit alpha-4 isoform 2 \{ECO:0000313\|EMBL:EQB78454.1\} |
| 3.20 | 1 | 78 kDa glucose-regulated protein |
| 3.19 | 1 | Mediator of RNA polymerase II transcription subunit 20 \{ECO:0000313\|EMBL:AEJ84141.1\} |
| 3.19 | 1 | Mediator of RNA polymerase II transcription subunit 20 |
| 3.18 | 1 | Leucine-rich repeat serine/threonine-protein kinase 2 \{ECO:0000313\|EMBL:ETE56782.1\} |
| 3.18 | 1 | Leucine-rich repeat serine/threonine-protein kinase 2 |
| 3.17 | 2 | Transforming growth factor beta-3 \{ECO:0000313\|EMBL:KFM08521.1\} |
| 3.17 | 2 | Transforming growth factor beta-3 |
| 3.16 | 1 | Neurotrophin-3 |
| 3.16 | 1 | Cryptic protein |
| 3.13 | 1 | Cystatin-B |
| 3.12 | 1 | Ninein |
| 3.08 | 1 | Glutathione peroxidase 7 |
| 3.07 | 1 | Zinc finger protein 106 \{ECO:0000313\|EMBL:KFM06373.1\} |
| 3.07 | 1 | Zinc finger protein 106 |
| 3.03 | 2 | RNA-directed DNA polymerase from mobile element jockey |
| 3.03 | 1 | Spermatogenesis-associated protein 6 |
| 3.03 | 1 | Ectonucleotide pyrophosphatase/phosphodiesterase family member 2 \{ECO:0000313\|EMBL:KFW07303.1\} |


| 3.03 | 1 | BolA-like protein 3 \{ECO:0000313\|EMBL:AER95134.1\} |
| :---: | :---: | :---: |
| 3.02 | 2 | Uromodulin |
| 3.01 | 1 | Condensin-2 complex subunit G2 |
| 2.98 | 1 | Basic immunoglobulin-like variable motif-containing protein |
| 2.94 | 1 | Major urinary protein 3 |
| 2.93 | 1 | Insulin-like growth factor binding protein-3 \{ECO:0000313\|EMBL:ACJ66735.1\} |
| 2.92 | 2 | Lung adenoma susceptibility protein 2 |
| 2.91 | 1 | Angiopoietin-4 \{ECO:0000313\|EMBL:EPQ13030.1\} |
| 2.88 | 1 | Titin \{ECO:0000313\|EMBL:CAA45941.1\} |
| 2.88 | 1 | Protein Igkv4-80 \{ECO:0000313\|Ensembl:ENSMUSP00000100142\} |
| 2.88 | 1 | Anti-dectin-1 15E2 light chain \{ECO:0000313\|EMBL:AIU56801.1\} |
| 2.85 | 1 | Tenascin |
| 2.84 | 2 | Plastin-2 \{ECO:0000313\|EMBL:KFQ07139.1\} |
| 2.82 | 1 | Neuroendocrine protein 7B2 |
| 2.82 | 1 | Leucine-rich repeat flightless-interacting protein 2 |
| 2.80 | 1 | Chymotrypsin-like elastase family member 1 |
| 2.79 | 1 | Pro-interleukin-16 \{ECO:0000313\|EMBL:KFV84631.1\} |
| 2.79 | 1 | Pro-interleukin-16 |
| 2.79 | 1 | Fatty acyl-CoA hydrolase, medium chain \{ECO:0000313\|EMBL:EMP36848.1\} |
| 2.79 | 1 | Arachidonate 5-lipoxygenase |
| 2.78 | 1 | Aminopeptidase N |
| 2.77 | 1 | Protein-cysteine N -palmitoyltransferase HHAT |
| 2.77 | 1 | Platelet-activating factor acetylhydrolase \{ECO:0000313\|EMBL:EMP34007.1\} |
| 2.77 | 1 | Platelet-activating factor acetylhydrolase |
| 2.76 | 1 | Versican core protein \{ECO:0000313\|Ensembl:ENSRNOP00000045935\} |
| 2.76 | 1 | Versican core protein |
| 2.76 | 1 | Protein Gm20782 \{ECO:0000313\|Ensembl:ENSMUSP00000094556\} |
| 2.74 | 1 | Thioredoxin \{ECO:0000313\|Ensembl:ENSRNOP00000016447\} |
| 2.71 | 1 | Transmembrane and ubiquitin-like domain-containing protein 1 |
| 2.70 | 1 | Laminin subunit beta-1 \{ECO:0000313\|EMBL:KGL97946.1\} |
| 2.70 | 1 | Interferon alpha-9 \{ECO:0000313\|EMBL:EGV95891.1\} |
| 2.70 | 1 | Interferon alpha-12 |
| 2.69 | 1 | Regenerating islet-derived protein 3 beta \{ECO:0000313\|EMBL:EGW03567.1\} |
| 2.69 | 1 | Regenerating islet-derived protein 3-beta |
| 2.69 | 1 | Little elongation complex subunit 2 |
| 2.69 | 1 | Fucolectin-6 |
| 2.69 | 1 | Fucolectin |
| 2.68 | 1 | Ladderlectin |
| 2.66 | 1 | Thioredoxin-related transmembrane protein 1 \{ECO:0000313\|EMBL:KFQ37635.1\} |
| 2.66 | 1 | Thioredoxin-related transmembrane protein 1 |
| 2.64 | 1 | Pepsinogen A \{ECO:0000313\|EMBL:BAB20092.1\} |
| 2.63 | 1 | WNT1 inducible signaling pathway protein 1 \{ECO:0000313\|EMBL:AFP43736.1\} |
| 2.63 | 1 | WNT1-inducible-signaling pathway protein 1 |
| 2.63 | 1 | Gastrokine-1 |
| 2.61 | 2 | Ufm1-specific protease 1 |
| 2.61 | 1 | Prolactin-7A2 |
| 2.59 | 2 | IgGFc-binding protein |
| 2.59 | 1 | Fibroblast growth factor 6 |
| 2.57 | 1 | Fibroblast growth factor 13 |
| 2.57 | 1 | Alpha-1-antitrypsin \{ECO:0000313\|EMBL:KGL74436.1\} |
| 2.56 | 1 | Transmembrane protein 87B |
| 2.55 | 1 | Laminin subunit gamma-1 |
| 2.55 | 1 | Endothelial lipase |
| 2.55 | 1 | Cyclin-dependent kinase 13 \{ECO:0000313\|EMBL:EPQ13559.1\} |
| 2.54 | 1 | Terminal uridylyltransferase 4 |
| 2.52 | 1 | Cholinesterase |
| 2.51 | 1 | Lactation elevated protein 1 |
| 2.48 | 1 | Tissue factor pathway inhibitor (Lipoprotein-associated coagulation inhibitor) variant \{ECO:0000313\|EMBL:BAD93103.1\} |
| 2.47 | 1 | Zinc finger protein 513 \{ECO:0000313\|EMBL:KFP64546.1\} |
| 2.47 | 1 | Zinc finger protein 513 |
| 2.44 | 1 | Fibrillin 2 \{ECO:0000313\|EMBL:CAM59484.1\} |
| 2.43 | 1 | Calcium-activated chloride channel regulator 2 |
| 2.42 | 1 | Protein Aimp1 \{ECO:0000313\|Ensembl:ENSRNOP00000015213\} |


| 2.38 | 1 | Slit homolog 3 protein |
| :---: | :---: | :---: |
| 2.37 | 1 | Zinc finger protein 347 |
| 2.37 | 1 | Protein Gpr98 \{ECO:0000313\|Ensembl:ENSRNOP00000029273\} |
| 2.37 | 1 | Protein Gm13235 \{ECO:0000313\|Ensembl:ENSMUSP00000116281\} |
| 2.37 | 1 | G-protein coupled receptor 98 |
| 2.35 | 1 | Putative global transcription activator SNF2L2 \{ECO:0000313\|EMBL:AFO94279.1\} |
| 2.34 | 1 | Olfactory receptor 10AG1 |
| 2.31 | 1 | Cysteine-rich secretory protein 2 \{ECO:0000313\|EMBL:KFP74139.1\} |
| 2.29 | 1 | ITLN1 protein \{ECO:0000313\|EMBL:AAH77059.1\} |
| 2.28 | 1 | Trypsin-3 \{ECO:0000313\|EMBL:EMP27701.1\} |
| 2.28 | 1 | Serpin B6 \{ECO:0000313\|EMBL:EMP36654.1\} |
| 2.28 | 1 | Retrovirus-related Pol polyprotein from type-2 retrotransposable element R2DM |
| 2.28 | 1 | Ovostatin homolog |
| 2.28 | 1 | Kallikrein-14 |
| 2.28 | 1 | BC048546 protein \{ECO:0000313\|EMBL:AAH34132.1\} |
| 2.27 | 1 | Calcitonin-related peptide \{ECO:0000313\|EMBL:ACP19071.1\} |
| 2.27 | 1 | Calcitonin gene-related peptide |
| 2.25 | 1 | Probable 28S rRNA (cytosine-C(5))-methyltransferase |
| 2.24 | 2 | Pulmonary surfactant-associated protein D |
| 2.24 | 1 | Tumor necrosis factor ligand superfamily member 6 |
| 2.24 | 1 | Noelin-3 \{ECO:0000313\|EMBL:EPQ15334.1\} |
| 2.24 | 1 | Noelin-3 |
| 2.22 | 1 | Granzyme A \{ECO:0000313\|EMBL:KFO93227.1\} |
| 2.19 | 1 | F-box only protein 41 \{ECO:0000313\|EMBL:EGW02929.1\} |
| 2.18 | 1 | Polymeric immunoglobulin receptor |
| 2.16 | 1 | Sodium channel protein \{ECO:0000256\|RuleBase:RU361132\} |
| 2.16 | 1 | Cysteine protease ATG4C |
| 2.11 | 1 | Protein FAM92A1 \{ECO:0000313\|EMBL:EMP42258.1\} |
| 2.11 | 1 | Protein FAM92A1-A |
| 2.08 | 1 | Solute carrier family 35 member E1 \{ECO:0000313\|EMBL:KFR04818.1\} |
| 2.08 | 1 | Solute carrier family 35 member E1 |
| 2.08 | 1 | Calumenin \{ECO:0000313\|Ensembl:ENSMUSPO0000134708\} |
| 2.08 | 1 | Calumenin |
| 2.07 | 1 | Interleukin-1 receptor-associated kinase 4 \{ECO:0000313\|EMBL:EGW10212.1\} |
| 2.06 | 1 | Collagen alpha-1(XXI) chain |
| 2.01 | 1 | SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily E member 1 \{ECO:0000313\|Ensembl:ENSP00000466608\} |
| 2.01 | 1 | SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily E member 1 |
| 1.99 | 1 | RING finger protein 207 |
| 1.99 | 1 | Ankyrin repeat and SOCS box protein 18 \{ECO:0000313\|EMBL:KFQ59013.1\} |
| 1.99 | 1 | Ankyrin repeat and SOCS box protein 18 |
| 1.96 | 1 | Ly6/PLAUR domain-containing protein 6 |
| 1.96 | 1 | Laminin subunit alpha-2 \{ECO:0000313\|EMBL:KFO85806.1\} |
| 1.94 | 1 | Metalloendopeptidase \{ECO:0000256\|RuleBase:RU361183\} |
| 1.94 | 1 | Meprin A subunit alpha |
| 1.92 | 1 | RPGR 1-19 isoform \{ECO:0000313\|EMBL:ABB03733.1\} |
| 1.92 | 1 | KH domain-containing, RNA-binding, signal transduction-associated protein 3 |
| 1.92 | 1 | Alternative protein KHDRBS3 \{ECO:0000313\|EMBL:CCQ43363.1\} |
| 1.89 | 1 | Leucine-rich, glioma inactivated 1 \{ECO:0000313\|EMBL:AAI21469.1\} |
| 1.88 | 1 | Lymphotoxin-alpha |
| 1.88 | 1 | Elongation factor 1-alpha |
| 1.87 | 1 | Platelet-derived growth factor D \{ECO:0000313\|EMBL:KFO25640.1\} |
| 1.87 | 1 | Platelet-derived growth factor D |
| 1.87 | 1 | N -acetylmuramoyl-L-alanine amidase |
| 1.86 | 1 | 60 S ribosomal protein L7 |
| 1.85 | 1 | Alpha-glucosidase |
| 1.83 | 1 | Cysteine-rich venom protein ENH1 |
| 1.83 | 1 | Cysteine-rich secretory protein 2 \{ECO:0000313\|EMBL:KFQ10034.1\} |
| 1.82 | 1 | Uromodulin-like protein \{ECO:0000313\|EMBL:ERE80014.1\} |
| 1.82 | 1 | Bone morphogenetic protein 3 \{ECO:0000256\|PIRNR:PIRNR037403\} |
| 1.82 | 1 | Bone morphogenetic protein 3 |
| 1.78 | 1 | Insulin-like growth factor-binding protein 6 \{ECO:0000313\|EMBL:EMP36961.1\} |
| 1.78 | 1 | Insulin-like growth factor-binding protein 6 |
| 1.75 | 1 | Alpha-actinin-4 \{ECO:0000313\|Ensembl:ENSP00000398393\} |


| 1.73 | 1 | Tumor necrosis factor ligand superfamily member 13 |
| :---: | :---: | :---: |
| 1.73 | 1 | PSv-2 \{ECO:0000313\|EMBL:AAM83196.1\} |
| 1.73 | 1 | APRIL \{ECO:0000313\|EMBL:AGC24709.1\} |
| 1.72 | 1 | Isthmin |
| 1.72 | 1 | Corticoliberin \{ECO:0000313\|EMBL:EMP33293.1\} |
| 1.72 | 1 | Corticoliberin |
| 1.71 | 1 | Trypsin \{ECO:0000313\|EMBL:KFP04945.1\} |
| 1.71 | 1 | Trypsin |
| 1.66 | 1 | Titin \{ECO:0000313\|EMBL:KFV82010.1\} |
| 1.65 | 1 | Zinc finger protein with KRAB and SCAN domains 1 |
| 1.65 | 1 | Zinc finger protein 781 |
| 1.65 | 1 | Zinc finger protein 167 \{ECO:0000313\|EMBL:ELK10134.1\} |
| 1.65 | 1 | Macaca fascicularis brain cDNA clone: QmoA-10946, similar to human similar to Zinc finger protein 429 (LOC388522), mRNA, RefSeq: XM_371154.2 \{ECO:0000313\|EMBL:BAE90955.1\} |
| 1.63 | 1 | Cholecystokinin |
| 1.60 | 1 | Titin \{ECO:0000313\|EMBL:AAV93231.1\} |
| 1.60 | 1 | Predicted protein \{ECO:0000313\|EMBL:EDO26299.1\} |
| 1.60 | 1 | Circumsporozoite protein |
| 1.59 | 1 | Chordin-like protein 1 \{ECO:0000313\|EMBL:EMP35434.1\} |
| 1.57 | 1 | Striatin-interacting protein 1 |
| 1.57 | 1 | LOC100158439 protein \{ECO:0000313\|EMBL:AAI61720.1\} |
| 1.57 | 1 | Ancylostoma secreted protein |
| 1.54 | 1 | Osteopontin \{ECO:0000313\|EMBL:CAI65407.1\} |
| 1.54 | 1 | Osteopontin |
| 1.54 | 1 | Acyl-CoA synthetase family member 4 \{ECO:0000313\|EMBL:KFV76493.1\} |
| 1.54 | 1 | Acyl-CoA synthetase family member 4 |
| 1.51 | 1 | Peptidylaminoacyl-L/D-isomerase \{ECO:0000313\|EMBL:AAX55674.1\} |
| 1.48 | 1 | Histo-blood group ABO system transferase 2 |
| 1.47 | 2 | Lectin \{ECO:0000313\|EMBL:ACO14384.1\} |
| 1.46 | 1 | Tumor necrosis factor receptor superfamily member 18 \{ECO:0000313\|EMBL:EMP25995.1\} |
| 1.46 | 1 | Tumor necrosis factor receptor superfamily member 18 |
| 1.46 | 1 | LOC100037012 protein \{ECO:0000313\|EMBL:AAI30108.1\} |
| 1.46 | 1 | Coagulation factor XII |
| 1.44 | 1 | Procollagen C-proteinase 3 \{ECO:0000313\|EMBL:BAA75639.1\} |
| 1.44 | 1 | LON peptidase N -terminal domain and RING finger protein 1 |
| 1.41 | 1 | Eukaryotic translation initiation factor 2A |
| 1.31 | 1 | E3 ubiquitin-protein ligase RFWD2 \{ECO:0000313\|EMBL:KFQ53406.1\} |
| 1.31 | 1 | E3 ubiquitin-protein ligase RFWD2 |
| 1.30 | 1 | Lysozyme C-1 |
| 1.29 | 1 | Protein angel-like 2 \{ECO:0000313\|EMBL:ETE70463.1\} |
| 1.29 | 1 | Protein angel homolog 2 |
| 1.29 | 1 | Protein angel 2 \{ECO:0000313\|EMBL:KFM07759.1\} |
| 1.26 | 1 | Lambda-recombinase-like protein \{ECO:0000313\|EMBL:AAL35359.2\} |
| 1.22 | 1 | WAP four-disulfide core domain protein 12 |
| 1.22 | 1 | Coagulation factor V |
| 1.22 | 1 | Claudin \{ECO:0000256\|RuleBase:RU060637\} |
| 1.22 | 1 | Claudin-8 |
| 1.22 | 1 | Alpha-2-macroglobulin receptor-associated protein \{ECO:0000313\|EMBL:KFP90247.1\} |
| 1.20 | 1 | Uromodulin \{ECO:0000313\|EMBL:EGW12970.1\} |
| 1.19 | 1 | Venom allergen 5 \{ECO:0000313\|EMBL:ERG86032.1\} |
| 1.19 | 1 | Venom allergen 5 |
| 1.15 | 1 | Heparanase \{ECO:0000313\|EMBL:ERE88251.1\} |
| 1.08 | 1 | Peptidylaminoacyl-L/D-isomerase \{ECO:0000313\|EMBL:AAX55673.1\} |
| 1.06 | 1 | GPI mannosyltransferase 2 |
| 1.05 | 1 | Ig lambda chain V-I region MEM |
| 1.04 | 1 | Proenkephalin \{ECO:0000313\|EMBL:ABM05721.1\} |
| 1.04 | 1 | Proenkephalin-A |
| 0.93 | 1 | Sorting nexin-9 |
| 0.91 | 1 | Ras-related protein Rab-1A |
| 0.90 | 1 | Tomoregulin-2 |
| 0.88 | 1 | Putative DMBT1-like protein \{ECO:0000313\|EMBL:EHB14981.1\} |
| 0.83 | 1 | Complement component C7 \{ECO:0000313\|EMBL:JAB54517.1\} |
| 0.80 | 1 | Transmembrane protein 245 \{ECO:0000313\|EMBL:KFV66544.1\} |
| 0.80 | 1 | Transmembrane protein 245 |


| 0.74 | 1 | Zinc finger protein 252 |
| :--- | :--- | :--- |
| 0.68 | 1 | GLIPR1-like protein 1 |
| 0.68 | 1 | Glioma pathogenesis-related protein 1 \{ECO:0000313\|EMBL:ETE667855.1\} |
| 0.64 | 1 | Liver carboxylesterase 4 |
| 0.47 | 1 | Arylsulfatase K-like protein \{ECO:0000313\|EMBL:AEP99345.1\} |
| 0.47 | 1 | Arylsulfatase K |
| 0.45 | 1 | Cartilage intermediate layer protein 1 |
| 0.35 | 1 | Macaca fascicularis brain CDNA clone: QmoA-10650, similar to human O-sialoglycoprotein endopeptidase <br> (OSGEP), mRNA, RefSeq: NM_017807.1 \{ECO:0000313\|EMBL:BAE87456.1\} |
| 0.24 | 1 | Cyclin-dependent kinase 13 \{ECO:0000313\|EMBL:KFQ96733.1\} |
| 0.24 | 1 | Cyclin-dependent kinase 13 \{ECO:0000313\|EMBL:KFO06794.1\} |
| 0.16 | 1 | Fatty acid desaturase 6 \{ECO:0000313\|EMBL:KFZ63933.1\} |
| 0.16 | 1 | Fatty acid desaturase 6 |
| 0.07 | 1 | Alpha-1-antitrypsin homolog |

## De-duplicate.sh

```
#!/bin/bash
# ESTABLISH PROGRAM PATHS #
dedupe=/nfs0/IB/Weis_Lab/chouinad/Tools/bbmap/dedupe2.sh
reformat=/nfs0/IB/Weis_Lab/chouinad/Tools/bbmap/reformat.sh
fastqc=/nfs0/IB/Weis_Lab/chouinad/Tools/fastqc
# CHANGE TO LOCAL RUN DIRECTORY - CHECK THIS FOR THE MACHINE BEING USED #
cd /data/chouinad/Reads/Dedup-Corrected
# DE-DUPLICATE SEQUENCES FOR EACH SAMPLE #
touch dedup.log
$dedupe in1=../CG_R1.fastq in2=../CG_R2.fastq out=CG_interleaved.fastq ac=f threads=22 1>> dedup.log 2>>
dedup.log
$dedupe in1=../PCG_R1.fastq in2=../PCG_R2.fastq out=PCG_interleaved.fastq ac=f threads=22 1>> dedup.log 2>>
dedup.log
$dedupe in1=../DTB_R1.fastq in2=../DTB_R2.fastq out=DTB_interleaved.fastq ac=f threads=22 1>> dedup.log 2>>
dedup.log
# COMBINE THEM ALL AND RUN FASTQC ON THE TOTAL FILE #
cat *interleaved.fastq > ALL_interleaved.fastq
$fastqc ALL_interleaved.fastq
# DE-INTERLEAVE FILES #
echo "De-interleaving CG"
$reformat in=CG_interleaved.fastq out1=CG_R1_dedup.fastq out2=CG_R2_dedup.fastq
echo "De-interleaving PCG"
$reformat in=PCG_interleaved.fastq out1=PCG_R1_dedup.fastq out2=PCG_R2_dedup.fastq
echo "De-interleaving DTB"
$reformat in=DTB_interleaved.fastq out1=DTB_R1_dedup.fastq out2=DTB_R2_dedup.fastq
# REMOVE INTERLEAVED #
if [ -f CG_R1_dedup.fastq ] && [ -f CG_R2_dedup.fastq ]
    then
        rm CG_interleaved.fastq
fi
if [ -f PCG_R1_dedup.fastq ] && [ -f PCG_R2_dedup.fastq ]
    then
        rm PCG_interleaved.fastq
fi
if [ -f DTB_R1_dedup.fastq ] && [ -f DTB_R2_dedup.fastq ]
        then
    rm DTB_interleaved.fastq
fi
```


## Error-correct.sh

## \#!/bin/bash

ecc=/nfs0/IB/Weis_Lab/chouinad/Tools/bbmap/ecc.sh
cd /data2/chouinad/Reads/Dedup-Corrected
echo "Error-correcting for CG"
\$ecc in=CG_R1_dedup.fastq in2=CG_R2_dedup.fastq out=CG_R1.fastq out2=CG_R2.fastq threads=22 echo "Error-correcting for PCG"
\$ecc in=PCG_R1_dedup.fastq in2=PCG_R2_dedup.fastq out=PCG_R1.fastq out2=PCG_R2.fastq threads=22
echo "Error-correcting for DTB"
\$ecc in=DTB_R1_dedup.fastq in2=DTB_R2_dedup.fastq out=DTB_R1.fastq out2=DTB_R2.fastq threads=22

# Trim-Clip-Assembly-Pipeline.sh 


\#\#\# User-defined sample prefixes (raw reads must end in "_R1.fastq" and "_R2.fastq") \#\#\#
SAMPLES=( CG PCG DTB )
\#\#\# User-defined dependency paths \#\#\#

```
this_script=./TC-Assembly-SAMPLES.sh
job_home=/nfs0/IB/Weis_Lab/chouinad/Assembly
    cutadapt=/home/zoo/chouinad/.local/bin/cutadapt
    filter=/nfs0/IB/Weis_Lab/chouinad/Tools/Filter-Reads.pl
    fastqc=/nfs0/IB/Weis_Lab/chouinad/Tools/fastqc
    read_stats=/nfs0/IB/Weis_Lab/chouinad/Tools/ReadLengths.R
    trinity_dir=/nfs0/IB/Weis_Lab/chouinad/Tools/trinityrnaseq-2.1.1
    trinity=${trinity_dir}/Trinity
    trinity_stats=${trinity_dir}/util/TrinityStats.pl
    mapreads1=${trinity_dir}/util/bowtie_PE_separate_then_join.pl
    mapreads2=${trinity_dir}/util/SAM_nameSorted_to_uniq_count_stats.pl
    transdecoder1=${trinity_dir}/trinity-plugins/TransDecoder-2.0.1/TransDecoder.LongOrfs
    transdecoder2=${trinity_dir}/trinity-plugins/TransDecoder-2.0.1/TransDecoder.Predict
    transrate=/nfs0/IB/Weis_Lab/chouinad/Tools/transrate-1.0.1-linux-x86_64/transrate
    bedtools=/nfs0/IB/Weis_Lab/chouinad/Tools/bedtools2-2.25.0/bin/bedtools
    parse_stats=/nfs0/IB/Weis_Lab/chouinad/Tools/Parse-Assembly-Stats.pl
```

\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\# \#\#\# The remainder of the script doesn't need to be customized, but can be tweaked for usage as necessary \#\#\# \#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#
\#\#\# Get arguments and check usage \#\#\#
if [[ \$\# -ne 4 ]]; then
echo "USAGE: Trim-Clip-Assembly-Pipeline.sh <PATH_TO_ READS> <QUALITY_TRIMMING_CUTOFF> <MEMORY>
<THREADCOUNT>"
exit 1
fi
READS=\$1
cutoff=\$2 \# The script was called with a cutoff of 5 for the selected assembly \#
max_mem=\$3
threadcount=\$4
\# Check for valid threadcount request \#
need=\$(( \$\{\#SAMPLES[@]\} * 2 ))
if [ ! \$threadcount -ge \$need ]
then

```
        echo "Error: you must request at least $need processors with this many samples!"
        exit 1
    fi
# Make README file for posterity #
if [ ! -f $this_script ]
        then
            echo "Error: TC-Assembly script is not present - check path before running"
            exit 1
fi
echo "Here is the total pipeline used: " > ./README
cat $this_script >> README
```

\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#

\#\#\# Run cutadapt 1.9 .1 to quality trim and clip adapters
Adapters:
\#\#\#
3'-Illumina : AGATCGGAAGAGC
$\begin{array}{lll}\text { \#\#\# } & 5^{\prime}-\text { Illumina } & \text { : AGATCGGAAGAGC } \\ \text { \#\# } & \text { SMARTer } & \text { AAGCAGTGGTATCAACGCAGAGTACATGGG }\end{array}$
\#\#\#
\#
3'-Illumina : AGATCGGAAGAGC
\#\#\# $5^{\prime}-$ SMARTer $-\mathrm{RC} \quad:$ CCCATGTACTCTGCGTTGATACCACTGCTT $\quad$ \#\#
\#\#\# $3^{\prime}-$ SMARTer : AAGCAGTGGTATCAACGCAGAGTACTTTTTTT $\quad$ \#\#\#
$\begin{array}{lll}\text { \#\#\# } & 3^{\prime}-\text { SMARTer-RC } & \text { : AAAAAAAGTACTCTGCGTTGATACCACTGCTT } \\ \# \# \# & \text { Poly-A-tail } & \text { : AAAAAAAAAA }\end{array}$
\#\#\#
\#\#\#
\#\#\#
\#\#\#
\#\#\#
$\begin{array}{ll}\# \# \# & -\quad \text { This needs to be run individually for each pair wITHOUT read discarding! } \\ \# \# \# & -T h e ~ r e a d s ~ a r e ~ f i l t e r e d ~ i n ~ t h e ~ n e x t ~ s t e p, ~ i n ~ o r d e r ~ t o ~ r e t a i n ~ b r o k e n ~ p a i r s ~\end{array}$
Notes:

- This needs to be run individually for each pair WITHOUT read discarding!
\#\#\#
The reads are filtered in the next step, in order to retain broken pairs
\#\#\#
5'-SMARTer-RC : CCCATGTACTCTGCGTTGATACCACTGCTT
\#\#\#
\#\#\#
Poly-A-tail : AAAAAAAAAA
Poly-A-tail-RC : TTTTTTTTTT
\#\#\#

\#\#\#
Notes:
- This needs to be run individually for each pair WITHOUT read discarding!
\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#
\#\#\# Trim/Clip reads for each sample in parallel \#\#\#
for prefix in "\$\{SAMPLES[@]\}"
do
\# Trim and clip first read file \#
echo -n "Trimming and clipping first read file for $\$\{p r e f i x\}:$ "; date
R1=\$\{READS\}/\$\{prefix\}_R1.fastq
R1_tc_outfile=\$\{prefix\}_R1_tc_unmerged.fastq
\$cutadapt -q \$cutoff,\$cutoff -a AGATCGGAAGAGC -g AAGCAGTGGTATCAACGCAGAGTACATGGG - a $\backslash$
CCCATGTACTCTGCGTTGATACCACTGCTT -g AAGCAGTGGTATCAACGCAGAGTACTTTTTTT - a AAAAAAAGTACTCTGCGTTGATACCACTGCTT
-e 0.1 -n 10 -o \$R1_tc_outfile \$R1 1> ./\$\{prefix\}_R1_TC.out 2> ./\$\{prefix\}_R1_TC.err \&
\# Trim and clip second read file \#
echo -n "Trimming and clipping second read file for $\$\{p r e f i x\}: ~ " ; ~ d a t e ~$
R2=\$\{READS\}/\$\{prefix\}_R2.fastq
R2_tc_outfile=\$\{prefix\}_R2_tc_unmerged.fastq
\$cutadapt - $q$ \$cutoff, \$cutoff - a AGATCGGAAGAGC - g AAGCAGTGGTATCAACGCAGAGTACATGGG - a \}
CCCATGTACTCTGCGTTGATACCACTGCTT -g AAGCAGTGGTATCAACGCAGAGTACTTTTTTT - a AAAAAAAGTACTCTGCGTTGATACCACTGCTT $\backslash$
-e 0.1 -n 10 -o \$R2_tc_outfile \$R2 1> ./\$\{prefix\}_R2_TC.out 2> ./\$\{prefix\}_R2_TC.err \&
done
wait
\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#
\#\#\# $\quad$ Run "Filter-Reads.pl" to discard reads under a minimum length threshold after trimming/clipping \#\#\#
\#\#\#
$\begin{array}{ll}\# \# \# & \text { Run "Filter-Reads.pl to discard reads under a minimum length threshold after trimming/clipping } \\ \# \# \# & \text { The script will retain proper mate pairing as well as singletons/unpaired (from broken pairs) }\end{array}$
\#\#\#
\#\#\# USAGE: Filter-Reads.pl <R1.fastq> <R2.fastq> <Length-Cutoff> <0utput-Prefix> <OPTIONAL: Adapters>
\#\#\#
\#\#\# Adapters to search reads for can be added after the first four arguments, separated by spaces \#\#\#
Reads will be searched for perfect adapter matches and discarded if a match is found \#\#\#

\#\#\# Filter reads for each sample in parallel \#\#\#
for prefix in "\$\{SAMPLES[@]\}"
do
echo -n "Filtering reads for $\$\{p r e f i x\}: ~ " ; ~ d a t e$
\# Run "Filter-Reads.pl" to discard short reads and retain singletons \#
\# Add adapters to the end of the filter script if you want to discard any residual contamination \#
\# Add adapters to the end of the filter scrip
R1_tc_outfile $=\$\{$ prefix\}_R1_tc_unmerged.fastq
R2_tc_outfile $=\$\{$ prefix\}_R2_tc_unmerged.fastq
\$filter \$R1_tc_outfile \$R2_tc_outfile 25 \$prefix GCAGTGGTATCAA \&
done


## wait

```
### Clean up the unmerged reads ###
for prefix in "${SAMPLES[@]}"
do
    R1_tc_outfile=${prefix}_R1_tc_unmerged.fastq
    R2_tc_outfile=${prefix}_R2_tc_unmerged.fastq
    rm -f $R1_tc_outfile $R2_tc_outfile
done
```


\#\#\#
\#\#\# Combine reads from each sample into total read files (with prefix "ALL")
\#\#\#
\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#
\#\#\# Establish filtered read file names as variables \#\#\#
\# Establish valid paired output file names \#
R1_val=ALL_R1_valid.fastq
R2_val=ALL_R2_valid.fastq
\# Establish broken pair output file names \#
R1_up=ALL_R1_unpaired.fastq
R2_up=ALL_R2_unpaired.fastq
\#\#\# Combine files \#\#\#
touch \$R1_val \$R2_val \$R1_up \$R2_up
for prefix in "\$\{SAMPLES[@]\}"
do
cat \$\{prefix\}_R1_valid.fastq >> \$R1_val
cat \$\{prefix\}_R2_valid.fastq >> \$R2_val
cat \$\{prefix\}_R1_unpaired.fastq >> \$R1_up
cat \$\{prefix\}_R2_unpaired.fastq >> \$R2_up
done
\#\#\# Combine files yet further into a single input for QC and assembly (with "--run_as_paired" option) \#\#\#
TOTAL_FILE=ALL.fastq
cat \$R1_val \$R2_val \$R1_up \$R2_up > \$TOTAL_FILE
\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#
\#\#\#
\#\#\#
\#\#\#

echo -n "Counting reads and generating summary statistics: "; date
\# Count reads per file \#
Nr1=\$(grep -c "\@HWI" \$R1_val)
Nr2=\$(grep -c "\@HWI" \$R2_val)
Nup1=\$(grep -c "\@HWI" \$R1_up)
Nup2=\$(grep -c "\@HWI" \$R2_up)
\# Get read length stats \#
awk '\{ if(NR\%4==2) print length(\$0) \}' \$TOTAL_FILE > ./ALL-ReadLengths.txt
Nbp=\$(cat./ALL-ReadLengths.txt | awk '\{ sum += \$0 \} END \{ print sum \}')
R CMD BATCH '--args ./ALL-ReadLengths.txt ./ALL-LengthStats.tab' \$read_stats
\# Get adapter contamination - may need to be adjusted by user \#
Adapt=\$(grep -c "GCAGTGGTATCAA" \$TOTAL_FILE)
polyA=\$(grep -c "AAAAAAAAAA" \$TOTAL_FILE)
polyT=\$(grep -c "TTTTTTTTTT" \$TOTAL_FILE)
\# Print output to summary file \#
echo "Reads1-Paired Reads2-Paired Reads1-Broken Reads2-Broken N-bp Evil-13-mers
$\begin{array}{clllll}\text { A\{10\} } & \text { T\{10\}" }> & \text {./ALL-ReadCount.tab } & & \\ \text { echo "\$Nr1 } & \text { \$Nr2 } & \text { \$Nup1 \$Nup2 }\end{array}$
paste ./ALL-ReadCount.tab ./ALL-LengthStats.tab > ./ALL-ReadSummary.tab
rm -f ./ALL-ReadCount.tab ./ALL-LengthStats.tab

| \#\#\# |  |  |
| :---: | :---: | :---: |
| \#\#\# | Run fastqc for quality assurance |  |
| \#\#\# |  |  |

```
echo -n "Running fastqc for quality assurance: "; date
$fastqc $TOTAL_FILE
```

```
############################################################################################################
###
### Run Trinity to assemble transcriptome and run assembly metrics ###
### Ran Trinity
    ###
################################################################################################################
### Run Trinity to assemble reads ###
    # Run the program #
    echo -n "Running Trinity: "; date
    $trinity --seqType fq --single $TOTAL_FILE --run_as_paired --max_memory $max_mem --CPU $threadcount \
    --min_glue 1 --group_pairs_distance 800 --path_reinforcement_distance 25
    # Clean up the single input file to save some space #
    if [ ! -f ./trinity_out_dir/Trinity.fasta ]
        then
            echo "Error: Trinity did not complete!"
            exit 1
    fi
    rm $TOTAL_FILE
```

\#\#\# Run Trinity contig count and length statistics - generates stats file \#\#\#
echo -n "Running Trinity assembly statistics: "; date
\$trinity_stats ./trinity_out_dir/Trinity.fasta > ./Stats.txt
\#\#\# Run TransDecoder to predict ORFs \#\#\#
\# Predict ORFs \#
echo -n "Running TransDecoder to predict ORFs at least 50 AA long: "; date
\$transdecoder1 -t ./trinity_out_dir/Trinity.fasta
\$transdecoder2 -t ./trinity_out_dir/Trinity.fasta --retain_long_orfs 150
ORF_count=\$(grep -c "^>" ./Trinity.fasta.transdecoder.pep)
\# Annotate stats file \#
echo "\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#" >> ./Stats.txt
echo "\#\#\# ORFs predicted by TransDecoder \#\#\#" >> ./Stats.txt
echo "\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#" >> ./Stats.txt
echo "Candidate ORFs: \$\{ORF_count\}" >> ./Stats.txt
\#\#\# Run read mapping on valid pairs only - use bowtie \#\#\#

```
    # Assess read coverage, including proper pairs and singletons - part 1 of Trinity mapping protocol #
    echo -n "Mapping reads: "; date
    $mapreads1 --seqType fq --left $R1_val --right $R2_val --target ./trinity_out_dir/Trinity.fasta \
    --aligner bowtie -- -p $threadcount --all --best --strata -m 300 -X 800
    # Annotate stats file #
    echo "" >> ./Stats.txt
    echo "###############################" >> ./Stats.txt
    echo "### Read mapping statistics ###" >> ./Stats.txt
    echo "###############################" >> ./Stats.txt
    echo "" >> ./Stats.txt
    # Run part 2 of Trinity mapping protocol - send output to stats file #
    echo -n "Generating read mapping statistics: "; date
    $mapreads2 ./bowtie_out/bowtie_out.nameSorted.bam >> ./Stats.txt
```

\#\#\# Determine depth of coverage \#\#\#
echo -n "Calculating depth of coverage: "; date
\# Annotate stats file \#
echo "" >> ./Stats.txt
echo "\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#" >> ./Stats.txt
echo "\#\#\# Average depth of coverage \#\#\#" >> ./Stats.txt
echo "\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#" >> ./Stats.txt
echo "" >> ./Stats.txt
\# Calculate the average and standard deviation for the depth of coverage \#
\$bedtools genomecov -ibam ./bowtie_out/bowtie_out.coordSorted.bam -d | awk '\{sum += \$3; sumsq += \$3*\$3\} END
\{ print "Average depth of coverage: ", sum/NR; print "Coverage standard deviation: ", sqrt(sumsq/NR -
(sum/NR)^2) $\}^{\prime} \gg$./Stats.txt

```
echo -n "Summarizing all trimming/clipping/assembly metrics for ${1}: "; date
    # Parse the stats and combine with read summary #
    $parse_stats --custom ./Stats.txt > ./Assembly-Stats.tmp.tab
    paste ./ALL-ReadSummary.tab ./Assembly-Stats.tmp.tab > ./Assembly-Stats.tab
    # Clean up #
    rm ./Assembly-Stats.tmp.tab
```


echo -n "Running TransRate to assess assembly quality: "; date
\$transrate --assembly ./trinity_out_dir/Trinity.fasta --left \$R1_val --right \$R2_val \}
--threads \$threadcount 1> TR.out 2> TR.err

\# Get the job name and make the storage directory if it needs to be created \#
wd=\$(basename $\$(p w d))$
mkdir -p \$\{job_home\}/\$\{wd\}
\# Loop through important files and copy them back to the primary storage path \#
for f in ./README \$this_script ./\$\{wd\}.out ./\$\{wd\}.err ./*.html./Stats.txt ./Assembly-Stats.tab ./TR.out
./transrate_results/assemblies.csv ./trinity_out_dir/Trinity.fasta
./transrate_results/Trinity/good.Trinity.fasta
do
if [ ! -f \$f ]
then
echo "Error: file $\$\{f\}$ doesn't exist and thus cannot be moved!"
continue
fi
file=\$(basename \$f)
newname=\$\{wd\}-\$\{file
if [ ! -f \$\{job_home\}/\$\{wd\}/\$\{newname\} ]
then
echo "Copying \$file to \$\{job_home\}/\$\{wd\}/\$\{newname\}"
cp \$f \$\{job_home\}/\$\{wd\}/\$\{newname\}
else
echo "File \$\{file\} already exists in \$\{job_home\}/\$\{wd\} and thus was not moved!"
fi
done
\# Rename the few with different names \#
outlog=\$\{job_home\}/\$\{wd\}/\$\{wd\}-\$\{wd\}. out
errlog=\$\{job_home\}/\$\{wd\}/\$\{wd\}-\$\{wd\}. err
filtered=\$\{job_home\}/\$\{wd\}/\$\{wd\}-good.Trinity.fasta
if [! -f \$outlog ]
then
echo "Error: file \$outlog doesn't exist and thus cannot be renamed!"
else
mv \$outlog \$\{job_home\}/\$\{wd\}/\$\{wd\}.out
fi
if [! -f \$errlog ]
then
echo "Error: file \$errlog doesn't exist and thus cannot be renamed!"
else
mv \$errlog \$\{job_home\}/\$\{wd\}/\$\{wd\}.err
fi
if [ ! -f \$filtered ]
then
echo "Error: file \$filtered doesn't exist and thus cannot be renamed!"
else
mv \$filtered \$\{job_home\}/\$\{wd\}/\$\{wd\}-Trinity-Filtered.fasta
fi

## Filter-Reads.pl

```
#!/usr/bin/perl
```

use strict; use warnings;

\#\#\# Check usage and get arguments \#\#\#
unless(@ARGV >= 4) \{
die("USAGE: Filter-Reads.pl <R1.fastq> <R2.fastq> <Length-Cutoff> <Output-Prefix> <OPTIONAL: Adapter
1> <OPTIONAL: Adapter 2> <ETC...>\n")
\}
foreach (@ARGV) \{chomp(\$_)\}
my \$r1 = shift(@ARGV)
my $\$ \mathrm{r} 2=$ shift(@ARGV);
my \$cutoff = shift(@ARGV);
my \$prefix = shift(@ARGV);
my @ADAPTERS = @ARGV,
my \$adapter_pattern = join('|', @ADAPTERS);
\#\#\# Read in files, make index files that reference whether or not each read is valid \#\#\#

```
# Make index for R1 file #
my @index1;
open(R1, "<$r1");
my $counter = 0;
my $seq_num = 0;
my $id; my $seq; my $other; my $qual;
while(<R1>){
$counter++;
chomp($_);
if($counter == 1){
                                    $id = $_
                                    }
elsif($counter == 2){
            $seq = $_;
            }
elsif($counter == 3){
                $other = $_
                $ot
elsif($counter == 4){
                    $qual = $_;
                $seq_num++;
                if(@ADAPTERS == 0){
                        if(length($seq) >= $cutoff){
                                    push(@index1, 0);
            elsif(length($seq) < $cutoff){
                                    push(@index1, 1);
                                    pus
                    else{die("Error: couldn't read length for R1 of sequence $seq_num\n")}
                    }
                elsif(@ADAPTERS > 0){
                    if((length($seq) >= $cutoff) && ($seq !~ /$adapter_pattern/)){
                        push(@index1, 0);
                    elsif((length($seq) < $cutoff) || ($seq =~ /$adapter_pattern/)){
                        push(@index1, 1);
                            else{die("Error: couldn't read length for R1 of sequence $seq_num\n")}
                    }
                else{die("Error: could not read adapter input correctly for R1\n")}
                $counter = 0;
```

```
            else{die("Error: something is wrong with your loop for R1!\n")}
            }
# Make index for R2 file #
    my @index2;
    open(R2, "<$r2");
    $counter = 0;
    $seq_num = 0;
    $id = "";
    $seq = "";
    $other = "";
    $qual = "";
    while(<R2>){
        $counter++;
        chomp($_);
        if($counter == 1){
            $id = $_;
        elsif($counter == 2){
            $seq = $_;
            }
        elsif($counter == 3){
            $other = $_
            }
        elsif($counter == 4){
            $qual = $_;
            $seq_num++;
            if(@ADAPTERS == 0){
                if(length($seq) >= $cutoff){
                    push(@index2, 0);
                    }
                    elsif(length($seq) < $cutoff){
                        push(@index2, 1);
                    pus
                    else{die("Error: couldn't read length for R2 of sequence $seq_num\n")}
                    }
                elsif(@ADAPTERS > 0){
                    if((length($seq) >= $cutoff) && ($seq !~ /$adapter_pattern/)){
                        push(@index2, 0);
                        }
                    elsif((length($seq) < $cutoff) || ($seq =~ /$adapter_pattern/)){
                        push(@index2, 1);
                        }
                            else{die("Error: couldn't read length for R2 of sequence $seq_num\n")}
                }
            else{die("Error: could not read adapter input correctly for R1\n")}
            $counter = 0;
            }
    else{die("Error: something is wrong with your loop for R2!\n")}
    }
### Read in files, output reads to relevant file based on whether one or both reads are valid ###
    # Establish output files #
    my $r1_val = $prefix . "_R1_valid.fastq";
    my $r2_val = $prefix . "_R2_valid.fastq";
    my $r1_up = $prefix . "_R1_unpaired.fastq";
    my $r2_up = $prefix . "_R2_unpaired.fastq";
    my $r1_discard = $prefix . "_R1_discards.fastq";
    my $r2_discard = $prefix . "_R2_discards.fastq";
    # Initialize read stat counters for summary output #
    my $val_count = 0;
    my $r1_up_count = 0;
    my $r2_up_count = 0;
    my $r1_discard_count = 0;
    my $r2_discard_count = 0;
    # Filter R1 file #
        # Open output files #
        open(R1_VAL, ">$r1_val");
            open(R1_UP, ">$r1_up");
            open(R1_DISCARD, ">$r1_discard");
            # Read input file #
```

```
    open(R1, "<$r1");
    # Get each read, check where it goes (based on indices), and write it there #
    $counter = 0;
    $seq_num = 0;
    $id = "";
    $seq = "'";
    $other = "";
    $qual = "";
    while(<R1>){
        $counter++;
        if($counter == 1){
                        $id = $_;
        elsif($counter == 2){
            $seq = $_;
            }
        elsif($counter == 3){
                $other = $_
        }
            $qual == $_;
                $seq_num++;
                if($index1[$seq_num - 1] == 1){
                    $r1_discard_count++;
                        print R1_DISCARD $id;
                        print R1_DISCARD $seq;
                        print R1_DISCARD $other;
                                    print R1_DISCARD $qual;
                    }
                elsif(($index1[$seq_num - 1] == 0) && ($index2[$seq_num - 1] == 1)){
                    $r1_up_count++;
                    print R1_UP $id;
                    print R1_UP $seq;
                print R1_UP $other;
                print R1_UP $qual;
                elsif(($index1[$seq_num - 1] == 0) && ($index2[$seq_num - 1] == 0)){
                        $val_count++;
                        print R1_VAL'$id
                print R1_VAL $seq;
                print R1_VAL $other;
                                    print R1_VAL $qual;
                else{die("Error: could not interpret index for R1 read $seq_num\n")}
                $counter = 0;
                else{die("Error: something is wrong with your loop for R1!\n")}
        }
    # Close output files #
    close(R1_VAL);
    close(R1_UP);
    close(R1_DISCARD);
# Filter R2 file #
    # Open output files #
    open(R2_VAL, ">$r2_val");
    open(R2_UP, ">$r2_up");
    open(R2_DISCARD, ">$r2_discard");
    # Read input file #
    open(R2, "<$r2");
    # Get each read, check where it goes (based on indices), and write it there #
    $counter = 0;
    $seq_num = 0;
    $id = "";
    $seq = "";
    $other = """;
    $qual = "";
    while(<R2>){
        $counter++;
        if($counter == 1){
            $id = $_;
            }
```

```
    elsif($counter == 2){
    $seq = $_;
            }
        elsif($counter == 3){
        $other = $_
        elsif($counter == 4){
        $qual = $_
        $seq_num++;
        if($index2[$seq_num - 1] == 1){
                $r2_discard_count++;
                print R2_DISCARD $id;
                print R2_DISCARD $seq;
                print R2_DISCARD $other;
                print R2_DISCARD $qual;
            elsif(($index2[$seq_num - 1] == 0) && ($index1[$seq_num - 1] == 1)){
                    $r2_up_count++;
                print R2_UP $id
                print R2_UP $seq;
                print R2_UP $other
                print R2_UP $qual;
            elsif(($index2[$seq_num - 1] == 0) && ($index1[$seq_num - 1] == 0)){
                print R2_VAL $id;
                print R2_VAL $seq;
                print R2_VAL $other;
                print R2_VAL $qual;
                pr
            else{die("Error: could not interpret index for R2 read $seq_num\n")}
            $counter = 0;
            }
    else{die("Error: something is wrong with your loop for R1!\n")}
close(R2);
# Close output files #
close(R2_VAL);
close(R2_UP);
close(R2_DISCARD);
```

```
### Done ###
print "-------------------\n";
print "Done filtering reads: "'. $prefix . "\n";
print "--------------------\n";
print "Total reads: " . (($val_count * 2) + $r1_up_count + $r2_up_count + $r1_discard_count +
$r2_discard_count) . "\n";
print "R1 discards: $r1_discard_count\n";
print "R2 discards: $r2_discard_count\n";
print "Number left: " . (($val_count * 2) + $r1_up_count + $r2_up_count) . "\n";
print "Number left: " . (($val_count * 2) + $r1_up_count + $r2_up_cou
print "R1 unpaired: $r1_up_count\n";
print "R2 unpaired: $r2_up_count\n";
print "---------------------\n";
```


## Parse-Assembly-Stats.pl

\#!/usr/bin/perl
use strict; use warnings;
\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\# \#\#\# \#\#\# This script will parse the output of Trinity assembly statistics (and stop there in "--normal" mode) \#\#\# \#\#\# - These metrics are created by running "\$trinity_home/util/TrinityStats.pl" \#\#\#
\#\#\#
\#\#\# Some custom scripts add other post-assembly metrics into this file (e.g. ORFs, read mapping, etc.) \#\#\#
\#\#\# - Thus, it can also parse the specific custom metrics added to stats file (with "--custom" mode) \#\#\#
\#\#\# - Be sure to check compatibility of the regular expression workhorses if using custom metrics! \#\#\#
\#\#\# As output by the TrinityStats.pl script, the file actually contains two sets of stats per assembly: \#\#\#
\#\#\# - This is redundant; it inflates stats proportionally to the number of alternative transcripts \#\#\#
\# - Consequently, this is also the first set given (e.g. N50a) by this script $\quad$ \#\#\#
\#\#\# 2. Stats for only the longest transcript per trinity "gene" \#\#\#

- This is much more informative for most assembly quality assurance purposes. \#\#\#
\#\#\# - It is also the second set of stats given (e.g. N50b) by this script \#\#\#
\#\#\# \#n
\#\#\# Arguments: \#\#\#
\#\#\# 1 = "--normal" OR "--custom" mode $\quad$ \#\#\#
\#\#\# $\quad 2-\mathrm{N}=$ Input File(s) \#\#\#
ent

```
foreach(@ARGV){
    chomp;
unless(@ARGV >= 2) {
            die("Usage 1: Parse-Assembly-Stats.pl <--normal OR --custom> <TrinityStatsFile.txt>\n");
            }
unless(($ARGV[0] eq "--normal") || ($ARGV[0] eq "--custom")){
    die("Usage 2: Parse-Assembly-Stats.pl <--normal OR --custom> <TrinityStatsFile.txt>\n")
    }
my $mode = shift(@ARGV);
my $custom_header;
if($mode eq "--custom"){
    $custom_header = "\tORFs\tTotal-Mapped\tProper-Pairs\tImproper-Pairs\tLeft-Only\tRight-
Only\tCoverage\tCoverage-SD"
    cover
elsif($mode eq "--normal"){
    $custom_header = "";
    $cu
else{die("Script mode not recognized: use --normal or --custom\n")}
print
"File\tGenes\tTranscripts\t%GC\tMean\tMedian\tN10a\tN20a\tN30a\tN40a\tN50a\tBases\tMean\tMedian\tN10b\tN20b\tN
30b\tN40b\tN50b\tBases" . $custom_header . "\n";
foreach my $infile (@ARGV) {
```

my \$genes = "NA";
my \$transcripts $\stackrel{\prime}{=}$ "NA";
my \$gc = "NA";
my @N10;
my @N20;
my @N30;
my @N40;
my @N50;
my @Mean;
my @Median;
my @Bases;
my \$orfs = "NA";
my \$total_mapped = "NA";
my \$map_proper = "NA";
my \$map_improper = "NA";
my \$left_only = "NA";
my \$right_only = "NA";
my \$coverage = "NA";
my \$cov_sd = "NA";
my \$custom_stats = "";
open (INFILE, "<\$infile") or die("Can't open in-file because: \$! \n");
while(<INFILE>)
if(\$_ =~ /Total\strinity\s'genes':\s+(.+)\$/)
\{

```
            $genes = $1;
            elsif($_ =~ /Total\strinity\stranscripts:\s+(.+)$/)
                $transcripts = $1;
            elsif($_ =~ /Percent\sGC:\s+(.+)$/){
                $gc = $1;
            elsif($_ 
            push(@N10, $1);
            elsif($_ =~~
                push(@N20, $1);
            elsif($_ =~ /Contig\sN30:\s+(.+)$/)
            push(@N30, $1);
            elsif($_ 
            push(@N40, $1);
            elsif($_ _ {~ /Contig\sN50:\s+(.+)$/)
            push(@N50, $1);
            elsif($_ =~~
            push(@Median, $1);
            elsif($_ 
            push(@Mean, $1);
            elsif($_ =~~
            push(@Bases, $1);
            elsif($_
            $orfs = $1;
            elsif($_ = =~ /^improper_pairs\s+(\d+)\s.*$/){
            $map_improper = $1;
            }
            elsif($_ =~ /^left_only\s+(\d+)\s.*$/){
            $left_only = $1;
            elsif($_ =~ /^proper_pairs\s+(\d+)\s.*$/){
            $map_proper = $1;
            elsif($_ }\underset{~}{=~
            elsif($_ =~ /Total\saligned\sreads:\s(\d+)$/){
                $total_mapped = $1;
            elsif($- =~~
            $coverage = $1;
            elsif($_ = /Coverage\sstandard\sdeviation:\s+(.+)$/){
                $cov_sd = $1;
            }
            else{next;}
                    }
    if($mode eq "--custom"){
    $custom_stats = "\t" . $orfs . "\t" . $total_mapped . "\t" . $map_proper . "\t" .
$map_improper . "\t" $. $left_only . "\t" . $right_only . "\t" . $coverage . "\t" . $cov_sd;
    elsif($mode eq "--normal"){
            $custom_stats = "";
            }
    else{die("Script mode not recognized: use --normal or --custom\n")}
    print "$infile" . "\t" . $genes . "\t". $transcripts . "\t" . $gc. "\t" . $Mean[0] " "\t" .
$Median[0] . "\t" . $N10[0] . "\t". $N20[0] . "\t" . $N30[0] . "\t" . $N40[0] . "\t" . $N50[0] . "\t"
$Bases[0]."\t" . $Mean[1]. "\t". $Median[1] . "\t" . $N10[1] . "\t" . $N20[1] . "\t" . $N30[1] . "\t" .
$N40[1] . "\t" . $N50[1]. "\t" . $Bases[1] . $custom_stats . "\n";
    close(INFILE)
    clo
```


## ReadLengths. R



## Expression-Pipeline.pl

\#!/usr/bin/perl

```
use strict; use warnings; use File::Basename;
###############################################################################################
```



```
### This script runs the Differential Expression pipeline for Trinity output ###
###
###
### Arguments required consist of:
###
### 1.) Assembly (e.g. "Trinity.fasta")
### 2.) Transcript Abundance Estimation Method ("--RSEM" OR "--kallisto")
### 3.) Threadcount (i.e. the number of threads to use in parallel processing)
    4.) OPTIONAL: "--no_align" OR "--no_analysis"
    - As the alignment can be done using multiple threads, this allows you to
        do the alignment and analysis in separate runs to maximize CPU efficiency
###
## Note:
    - The "analyze_diff_exp.pl" script ($analyze_DE) was modified (after backing up) ###
    The following arguments were added to the "PtR" command of the
    "cluster_diff_expressed_transcripts" subroutine (line 235 as of trinity-2.1.1) ###
    --heatmap_colorscheme \"blue,white,red\" ###
    --sample_cor_scale_limits \"-1,1\" # Scale for sample correlation ###
    - You may want to customize this between analyses, or redo upon a new install ###
    - You can also change other options, or run "PtR" directly for graphical analyses ###
    - It is found in: $trinity_dir/Analysis/DifferentialExpression/PtR ###
############################################################################################
### User-customized program paths ###
    # Relevant directories #
    my $home_dir = "/nfs0/IB/Weis_Lab/chouinad";
    my $file_dir = ".";
    my $wd = ".";
    # File naming system - used to generate the file names for each sample #
    # Provide sample prefixes - different per sample, the same for different reads #
    my @SAMPLES = qw(
            CG
            PCG
            DTB
            );
        # Provide file suffixes - different per read, the same for different samples #
        my $left_suffix = "_R1_valid.fastq";
        my $right_suffix = "_R2_valid.fastq";
        # Program paths - may need changing if versions or install locations change #
        my $trinity_dir = $home_dir . "/Tools/trinityrnaseq-2.1.1"; # Local install
        my $align_and_estimate = $trinity_dir . "/util/align_and_estimate_abundance.pl";
        my $abundance_matrix = $trinity_dir . "/util/abundance_estimates_to_matrix.pl";
        my $DE trinity util = $trinity_dir . "/Analysis/DifferentialExpression/run_DE'analysis.pl";
        my $analyze_DE = $trinity_dir . "/Analysis/DifferentialExpression/analyze_diff_expr.pl";
        my $define_clusters = $trinity_dir .
"/Analysis/DifferentialExpression/define_clusters_by_cutting_tree.pl";
##############################################################################################
### Read arguments and check usage ###
#################################################################################################
unless(@ARGV>=3){die("Usage: Expression-Pipeline.pl <Assembly.fasta> <--RSEM OR --kallisto> <Threads>
<OPTIONAL: --no_align OR --no_analysis>\n")};
my $assembly = shift(@ARGV); chomp($assembly);
my $name = basename($assembly); $name =~ s/\.fasta//;
my $est_method = shift(@ARGV); chomp($est_method); $est_method =~ s/--//;
unless(($est_method eq "RSEM") || ($est_method eq "kallisto")){die("Error: estimation method $est_method not
recognized\n")}
my $threadcount = shift(@ARGV); chomp($threadcount); # Request the same number when submitting #
my $no_align = 1;
my $no_analysis = 1;
if($ARGV[0] eq "--no_align") {
    $no_align = 0;
    $
elsif($ARGV[0] eq "--no_analysis"){
    $no_analysis = 0;
    $
my $run_count = @SAMPLES;
unless($assembly =~ ハ.fasta$/){die("Warning: input file must end in \".fasta\"\n");}
```

```
###############################################################################################
\#\#\# Run Expression Analyses \(\quad\) \#\#\#
#############################################################################################
# SKIP LOGIC #
unless($no_align == 0){
### Prep the reference ###
    print "Prepping the reference: "; system("date");
    if($est_method eq "RSEM"){
        system("$align_and_estimate --transcripts $assembly --est_method $est_method --aln_method
bowtie --trinity_mode --prep_reference --output_dir $wd");
        }
        elsif($est_method eq "kallisto"){
            system("$align_and_estimate --transcripts $assembly --est_method $est_method --trinity_mode
--prep_reference --output_dir $wd");
        }
    else{die("Error: Estimation method $est_method not recognized\n")}
    print "Finished at: "; system("date");
\#\#\# Align and estimate reads for each sample in series \#\#\#
print "Aligning and estimating abundance for all samples: "; system("date");
foreach(@SAMPLES) \{
```

```
                my $left = $file_dir . "/" . $_ . $left_suffix;
```

                my $left = $file_dir . "/" . $_ . $left_suffix;
                my $right = $file_dir . "/" . $_ . $right_suffix;
                my $output_dir = $wd . "/" . $_;
                print "Aligning $_" . "\n";
                if($est_method eq "RSEM"){
                                    system("$align_and_estimate --transcripts $assembly --left $left --right $right --
    seqType fq --est_method \$est_method --aln_method bowtie --trinity_mode --thread_count \$threadcount --
output_dir $output_dir");
                elsif($est_method eq "kallisto"){
system("\$align_and_estimate --transcripts \$assembly --left \$left --right \$right --
seqType fq --est_method \$est_method --trinity_mode --thread_count \$threadcount --output_dir \$output_dir");
}
else{die("Error: Estimation method \$est_method not recognized\n")}
};
print "Finished at: "; system("date");
\#\#\# Convert results into abundance matrix for GENES \#\#\#

```
```

print "Converting abundance estimates into a matrix for GENES: "; system("date");

```
print "Converting abundance estimates into a matrix for GENES: "; system("date");
# Expand all sample names into variable for the system call #
my @SAMPLE_GENES_RESULTS;
for(my $i=0; $i<@SAMPLES; $i++){
            if($est_method eq "RSEM"){
                        $SAMPLE_GENES_RESULTS[$i] = $wd . "/" . $SAMPLES[$i] . "/RSEM.genes.results";
            elsif($est_method eq "kallisto"){
                        $SAMPLE_GENES_RESULTS[$i] = $wd . "/" . $SAMPLES[$i] . "/abundance.tsv.genes";
                    }
            }
my $genes_results_all_samples = "@SAMPLE_GENES_RESULTS";
# Run the system call #
my $genes_prefix = $name . "_genes";
system("$abundance_matrix --est_method $est_method --name_sample_by_basedir --out_prefix
$genes_prefix $genes_results_all_samples");
print "Finished at: "; system("date");
### Convert results into abundance matrix for ISOFORMS ###
print "Converting abundance estimates into a matrix for ISOFORMS: "; system("date");
# Expand all sample names into variable for the system call #
my @SAMPLE_ISOFORMS_RESULTS;
for(my $i=0; $i<@SAMPLES; $i++){
            if($est_method eq "RSEM"){
                    $SAMPLE_ISOFORMS_RESULTS[$i] = $wd . "/" . $SAMPLES[$i] . "/RSEM.isoforms.results";
                    }
    elsif($est_method eq "kallisto"){
            $SAMPLE_ISOFORMS_RESULTS[$i] = $wd . "/" . $SAMPLES[$i] . "/abundance.tsv";
            }
    }
my $isoforms_results_all_samples = "@SAMPLE_ISOFORMS_RESULTS";
    # Run the system call #
```

```
    my $trans_prefix = $name . "_isoforms";
    system("$abundance_matrix --est_method'$est_method --name_sample_by_basedir --out_prefix
$trans_prefix $isoforms_results_all_samples");
    print "Finished at: "; system("date");
} # END SKIP LOGIC #
# SKIP LOGIC #
unless($no_analysis == 0){
### Run edgeR via Trinity utility for GENES ###
    print "Running edgeR for GENES: "; system("date");
    my $genes_prefix = $name . "_genes";
    my $genes_counts_matrix = $wd . "/" . $genes_prefix . ".counts.matrix";
    my $edgeR_genes_dir = $wd . "/edgeR_genes";
    system("$DE_trinity_util --matrix $genes_counts_matrix --method edgeR --dispersion 0.4 --output
$edgeR_genes_dir");
    print "Finished at: "; system("date");
### Run edgeR via Trinity utility for ISOFORMS ###
    print "Running edgeR for ISOFORMS: "; system("date");
    my $trans_prefix = $name . "_isoforms";
    my $trans_counts_matrix = $w\overline{d . "/" . $trans_prefix . ".counts.matrix";}
    my $edgeR_trans_dir = $wd . "/edgeR_trans";
    system("$DE_trinity_util --matrix $trans_counts_matrix --method edgeR --dispersion 0.4 --output
$edgeR_trans_dir");
    print "Finished at: "; system("date");
```



```
\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#
\#\#\# Analyze edgeR results - extracts differentially-expressed GENES \#\#\#
print "Analyzing edgeR results for GENES: "; system("date");
my $genes_expr_matrix = $wd . "/". $genes_prefix . ".TMM.EXPR.matrix";
my $DE_genes_output = $genes_prefix . "_DE";
system("cd $edgeR_genes_dir; $analyze_DE --matrix ../$genes_expr_matrix --output $DE_genes_output");
print "Finished at: "; system("date");
\#\#\# Analyze edgeR results - extracts differentially-expressed ISOFORMS \#\#\#
```

```
print "Analyzing edgeR results for ISOFORMS: "; system("date");
```

print "Analyzing edgeR results for ISOFORMS: "; system("date");
my \$trans_expr_matrix = \$wd . "/" . \$trans_prefix . ".TMM.EXPR.matrix";
my \$trans_expr_matrix = \$wd . "/" . \$trans_prefix . ".TMM.EXPR.matrix";
my \$DE_trans_output = \$trans_prefix . "_DE";
my \$DE_trans_output = \$trans_prefix . "_DE";
system("cd \$edgeR_trans_dir; $analyze_DE --matrix ../$trans_expr_matrix --output \$DE_trans_output");
system("cd \$edgeR_trans_dir; $analyze_DE --matrix ../$trans_expr_matrix --output \$DE_trans_output");
print "Finished at: "; system("date");

```
print "Finished at: "; system("date");
```

\#\#\# Define expression clusters for GENES \#\#\#

```
print "Defining expression clusters for GENES: "; system("date");
```

my \$DE_RData_genes = \$DE_genes_output . ".matrix.RData";
\# Define clusters by percentage of tree height \#
\#system("cd \$edgeR_genes_dir; \$define_clusters --Ptree 30 -R \$DE_RData_genes");
\# Define clusters into K number of clusters \#
system("cd \$edgeR_genes_dir; \$define_clusters --Ktree 3 -R \$DE_RData_genes");
system("cd \$edgeR_genes_dir; \$define_clusters --Ktree 4 -R \$DE_RData_genes");
system("cd \$edgeR_genes_dir; \$define_clusters --Ktree 5 -R \$DE_RData_genes");
system("cd \$edgeR_genes_dir; \$define_clusters --Ktree 6 -R \$DE_RData_genes");
system("cd \$edgeR_genes_dir; \$define_clusters - -Ktree 7 -R \$DE_RData_genes");
system("cd \$edgeR_genes_dir; \$define_clusters --Ktree 8 -R \$DE_RData_genes");
system("cd \$edgeR_genes_dir; \$define_clusters --Ktree 9 -R \$DE_RData_genes");
print "Finished at: "; system("date");
\#\#\# Define expression clusters for ISOFORMS \#\#\#

```
print "Defining expression clusters for ISOFORMS: "; system("date");
my $DE_RData_trans = $DE_trans_output . ".matrix.RData";
# Define clusters by percentage of tree height #
#system("cd $edgeR_trans_dir; $define_clusters --Ptree 30 -R $DE_RData_trans");
# Define clusters into K number of clusters #
system("cd $edgeR_trans_dir; $define_clusters --Ktree 3 -R $DE_RData_trans");
system("cd $edgeR_trans_dir; $define_clusters --Ktree 4 -R $DE_RData_trans");
system("cd $edgeR_trans_dir; $define_clusters --Ktree 5 -R $DE_RData_trans");
```

```
    system("cd $edgeR_trans_dir; $define_clusters --Ktree 6 -R $DE_RData_trans");
    system("cd $edgeR_trans_dir; $define_clusters --Ktree 7-R $DE_RData_trans");
    system("cd $edgeR_trans_dir; $define_clusters --Ktree 8 -R $DE_RData_trans");
    system("cd $edgeR_trans_dir; $define_clusters --Ktree 9 -R $DE_RData_trans");
print "Finished at: "; system("date");
```

\} \# END SKIP LOGIC \#

## Annotate-Pipeline.pl

\#!/usr/bin/perl

\#\#\# User-customized program paths \#\#\#

```
my $home_dir = "/nfs0/IB/Weis_Lab/chouinad";
my $wd = ".";
my $db_dir = "/data/chouinad/DB";
# Program paths - may need changing if versions or install locations change #
my $trinity_dir = $home_dir . "/Tools/trinityrnaseq-2.1.1"; # Local install
my $trinity = $trinity_dir . "/Trinity";
my $trinotate_dir = "/local/cluster/Trinotate_r20140708";
my $trinotate = $trinotate_dir . "/Trinotate";
my $transdecoder1 = $trinity_dir . "/trinity-plugins/TransDecoder-2.0.1/TransDecoder.LongOrfs";
my $transdecoder2 = $trinity_dir . "/trinity-plugins/TransDecoder-2.0.1/TransDecoder.Predict";
my $blastp = "/nfs0/IB/Weis_Lab/chouinad/Tools/ncbi-blast-2.3.0+/bin/blastp";
my $blastx = "/nfs0/IB/Weis_Lab/chouinad/Tools/ncbi-blast-2.3.0+/bin/blastx";
my $signalp = $home_dir. "/Tools/signalp-4.1/signalp";
my $hmmscan = "/local/cluster/hmmer-3.1b1/bin/hmmscan";
my $tmhmm = "/local/cluster/bin/tmhmm";
my $rnammer = $trinotate_dir . "/util/rnammer_support/RnammerTranscriptome.pl";
my $rnammer_path = "/nfs0/IB/Weis_Lab/chouinad/Tools/rnammer_1.2/rnammer";
my $gene_trans_map = $trinity_dir_. "/util/support_scripts/get_Trinity_gene_to_trans_map.pl";
my $group_blast = $trinity_dir : "/util/misc/blast_outfmt6_group_segments.pl";
my $blast_coverage = $trinity_dir . "/util/misc/blast_outfmt6_group_segments.tophit_coverage.pl";
```

\#\#\# Databases \#\#\#

```
my $db1 = $db_dir . "/uniprot_sprot.trinotate.pep";
my $db2 = $db_dir . "/uniprot_uniref90.trinotate.pep";
my $pfamA = $db_dir . "/Pfam-A.hmm";
```

\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#
\#\#\# Read arguments and check usage $\quad$ \#\#\#
\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#

```
unless(@ARGV==2){die("Usage: Annotate-Pipeline.pl <Assembly.fasta> <Threads>\n")};
my $assembly = $ARGV[0]; chomp($assembly);
my $threadcount = $ARGV[1]; chomp($threadcount); # Request the same number when submitting #
unless($assembly =~ ハ.fasta$/){die("Warning: input file must end in \".fasta\"\n");}
my $name = basename($assembly); if($name =~ /(.+)\.fasta/){$name = $1;}
#############################################################################################
\begin{tabular}{|c|c|c|}
\hline \multirow[t]{2}{*}{\#\#\#} & \multirow[t]{2}{*}{Run Annotation Analyses} & \multirow[t]{2}{*}{\#\#\#} \\
\hline & & \\
\hline
\end{tabular}
\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#
```

\#\#\# Run TransDecoder - retain all ORFs >= 150 BP (i.e. 50 AA) - can change value below \#\#\#
print "Running TransDecoder to predict ORFs: "; system("date");
system("\$transdecoder1 - t \$assembly; \$transdecoder2 - t \$assembly --retain_long_orfs 150");
print("Finished: "); system("date");
my \$transD_pep = \$wd . "/" . \$assembly . ".transdecoder.pep";
\#\#\# Run BLASTX on Assembly with Swiss-Prot (manually-curated) database \#\#\#
print "Running BLASTX on Swiss-Prot database: "; system("date");
system("\$blastx -query \$assembly -db \$db1 -num_threads \$threadcount -max_target_seqs 1 -outfmt $6>$
\$wd/sprot.blastx.out");
print("Finished: "); system("date");
\#\#\# Run BLASTX on Assembly with Unifref90 database (non-curated, more comprehensive) \#\#\#

```
print "Running BLASTX on Uniref90 database: "; system("date");
system("$blastx -query $assembly -db $db2 -num_threads $threadcount -max_target_seqs 1 -outfmt 6 >
$wd/uniref90.blastx.out");
print("Finished: "); system("date");
\#\#\# Run BLASTP on TransDecoder output with Swiss-Prot (manually-curated) database \#\#\#
```

```
print "Running BLASTP on Swiss-Prot database: "; system("date");
system("$blastp -query $transD_pep -db $db1 -num_threads $threadcount -max_target_seqs 1 -outfmt 6 >
$wd/sprot.blastp.out");
print("Finished: "); system("date");
### Run BLASTP on TransDecoder output with Uniref90 database ###
print "Running BLASTP on Uniref90 database: "; system("date");
system("$blastp -query $transD_pep -db $db2 -num_threads $threadcount -max_target_seqs 1 -outfmt 6 >
$wd/uniref90.blastp.out");
print("Finished: "); system("date");
### Run HMMSCAN ###
print "Running HMMSCAN: "; system("date");
system("$hmmscan --cpu $threadcount --domtblout $wd/TrinotatePFAM.out $pfamA $transD_pep > $wd/pfam.log");
print("Finished: "); system("date");
### Run SignalP ###
print "Running signalP: "; system("date");
system("$signalp -f short -n $wd/signalp.out $transD_pep");
print("Finished: "); system("date");
### Run TMHMM ###
print "Running TMHMM: "; system("date");
system("$tmhmm --short < $transD_pep > $wd/tmhmm.out");
print("Finished: "); system("date");
### Run RNAmmer ###
print "Running RNAmmer: "; system("date");
system("$rnammer --transcriptome $assembly --path_to_rnammer $rnammer_path");
print("Finished: "); system("date");
```

\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#
\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#1-length transcript analysis
\#\#\# Group BLAST hits to improve coverage \#\#\#
print "Grouping BLAST hits to improve coverage: "; system("date");

```
print "Grouping BLASTX with SWISS-PROT database";
system("$group_blast $wd/sprot.blastx.out $assembly $db1 > $wd/sprot.blastx.out.grouped");
print "Grouping BLASTX with UNIREF90 database";
system("$group_blast $wd/uniref90.blastx.out $assembly $db2 > $wd/uniref90.blastx.out.grouped");
print "Grouping BLASTP with SWISS-PROT database";
system("$group_blast $wd/sprot.blastp.out $transD_pep $db1 > $wd/sprot.blastp.out.grouped");
print "Grouping BLASTP with UNIREF90 database";
system("$group_blast $wd/uniref90.blastp.out $transD_pep $db2 > $wd/uniref90.blastp.out.grouped");
print("Finished: "); system("date");
```

\#\#\# Full-length transcript analysis \#\#\#
print "Analyzing full-length transcripts: "; system("date");
print "Running full-length transcript analysis on BLASTX with SWISS-PROT database"; system("\$blast_coverage \$wd/sprot.blastx.out.grouped >
\$wd/full_length_transcripts_blastx_sprot.tab");
print "Running full-length transcript analysis on BLASTX with UNIREF90 database";
system("\$blast_coverage \$wd/uniref90.blastx.out.grouped >
\$wd/full_length_transcripts_blastx_uniref90.tab");
print "Running full-length transcript analysis on BLASTP with SWISS-PROT database"; system("\$blast_coverage \$wd/sprot.blastp.out.grouped >
\$wd/full_length_transcripts_blastp_sprot.tab");
print "Running full-length transcript analysis on BLASTP with UNIREF90 database"; system("\$blast_coverage \$wd/uniref90.blastp.out.grouped > \$wd/full_length_transcripts_blastp_uniref90.tab");

```
print("Finished: "); system("date");
```


## Trinotate-Report.pl

\#!/usr/bin/perl
use strict; use warnings; use File::Basename;
\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\# \#\#\# \#\#\# Load annotation ("Annotate-Pipeline.pl") and expression ("Expression-Pipeline.pl") \#\#\# \#\#\# data into Trinotate sqlite database and print out an annotation report . \#\#\#
\#\#\#
\#\#\# Arguments required consist of:
\#\#\#
\#\#\#
\#\#\# - Assembly (e.g. "Trinity.fasta")
\#\#\#
\#\#\#
\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#
\#\#\# User-customized program paths \#\#\#

```
my $home_dir = "/nfs0/IB/Weis_Lab/chouinad";
my $db_dir = "/data/chouinad/DB";
my $wd = ".";
# Program paths - may need changing if versions or install locations change #
    my $trinity_dir = $home_dir . "/Tools/trinityrnaseq-2.1.1"; # Local install
    my $trinity = $trinity_dir . "/Trinity";
    my $trinotate_dir = $home_dir . "/Tools/trinotate-2.0.2";
    my $trinotate = $trinotate_dir . "/Trinotate";
    my $gene_trans_map = $trinity_dir .
"/util/support_scripts/get_Trinity_gene_to_trans_map.pl";
    my $import_DE = $trinotate_dir .
"/util/transcript_expression/import_expression_and_DE_results.pl";
    my $import_clusters = $trinotate_dir.
"/util/transcript_expression/import_transcript_clusters.pl";
```

\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#
\#\#\# Read arguments and check usage $\quad$ \#\#\#
\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#
unless(@ARGV==1) \{die("Usage: Trinotate-Report.pl <Assembly.fasta>\n")\};
my \$assembly = \$ARGV[0]; chomp(\$assembly);
unless(\$assembly =~ ハ.fasta\$/)\{die("Warning: input file must end in \".fasta\"\n");\}
my \$name = basename(\$assembly); if(\$name =~ /(.+) \.fasta/)\{\$name = \$1; \}
\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\# \#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#
\#\#\# Retrieve the Trinotate pre-generated resource SQLite database \#\#\#
my \$sql_db = "\$wd/Trinotate.sqlite";
if(-e \$sql_db) \{print "Trinotate.sqlite file already exists. The file will not be downloaded. ln "; $\}$
else\{
print "Retrieving the Trinotate boilerplate database: "; system("date");
system("wget
\"https://data.broadinstitute.org/Trinity/Trinotate_v2.0_RESOURCES/Trinotate.sprot_uniref90.20150131.boilerpla
te.sqlite.gz\" -0 \$wd/Trinotate.sqlite.gz");
print "Unpacking the Trinotate boilerplate database: "; system("date");
system("gunzip \$wd/Trinotate.sqlite.gz");
print "Finished at: "; system("date");
\}
\#\#\# Load transcripts and coding regions \#\#\#
print "Loading transcripts and coding region information into Trinotate: "; system("date");
my \$gtmap_name = \$wd . "/" . \$name . ". gene_trans_map";
my \$transD_pep $=\$ w d$ : "/". \$assembly . ".transdecoder.pep";
if(-e \$gtmap_name) \{print "Genes-Transcripts map already exists and will not be created. ln "; $\}$
else\{
system("\$gene_trans_map \$assembly > \$gtmap_name");
\}
system("\$trinotate \$wd/Trinotate. sqlite init --gene_trans_map \$gtmap_name --transcript_fasta
\$assembly --transdecoder_pep \$transD_pep");
print "Finished at: "; system("date");
\#\#\# Load all BLAST results \#\#\#
print "Loading BLASTP / Swiss-Prot results into Trinotate: "; system("date");
system("\$trinotate \$wd/Trinotate.sqlite LOAD_swissprot_blastp \$wd/sprot.blastp.out");
print "Finished at: "; system("date");

```
print "Loading BLASTP / Uniref90 results into Trinotate: "; system("date");
system("$trinotate $wd/Trinotate.sqlite LOAD_trembl_blastp $wd/uniref90.blastp.out");
print "Finished at: "; system("date");
print "Loading BLASTX / Swiss-Prot results into Trinotate: "; system("date");
system("$trinotate $wd/Trinotate.sqlite LOAD_swissprot_blastx $wd/sprot.blastx.out");
print "Finished at: "; system("date");
print "Loading BLASTX / Uniref90 results into Trinotate: "; system("date");
system("$trinotate $wd/Trinotate.sqlite LOAD_trembl_blastx $wd/uniref90.blastx.out");
print "Finished at: "; system("date");
\#\#\# Load pfam domains \#\#\#
```

```
print "Loading Pfam results into Trinotate: "; system("date");
```

print "Loading Pfam results into Trinotate: "; system("date");
system("\$trinotate \$wd/Trinotate.sqlite LOAD_pfam $wd/TrinotatePFAM.out");
system("$trinotate \$wd/Trinotate.sqlite LOAD_pfam \$wd/TrinotatePFAM.out");
print "Finished at: "; system("date");
print "Finished at: "; system("date");
\#\#\# Load transmembrane domains \#\#\#

```
```

print "Loading TMHMM results into Trinotate: "; system("date");

```
print "Loading TMHMM results into Trinotate: "; system("date");
system("$trinotate $wd/Trinotate.sqlite LOAD_tmhmm $wd/tmhmm.out");
system("$trinotate $wd/Trinotate.sqlite LOAD_tmhmm $wd/tmhmm.out");
print "Finished at: "; system("date");
print "Finished at: "; system("date");
\#\#\# Load signal peptide predictions \#\#\#
```

```
print "Loading SignalP results into Trinotate: "; system("date");
```

print "Loading SignalP results into Trinotate: "; system("date");
system("\$trinotate \$wd/Trinotate.sqlite LOAD_signalp $wd/signalp.out");
system("$trinotate \$wd/Trinotate.sqlite LOAD_signalp \$wd/signalp.out");
print "Finished at: "; system("date");
print "Finished at: "; system("date");
\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#
\#\#\# Load expression analyses
\#\#\#
Load expression analyses

# START SKIP

unless(1 == 1){

### Load expression results for GENES

    print "Loading expression results for GENES into Trinotate: "; system("date");
    my $genes_counts = $wd . "/" . $name . "_genes.counts.matrix";
    my $genes_expr = $wd . "/" . $name . "_genes.TMM.EXPR.matrix";
    my $genes_DE_dir = $wd . "/edgeR_genes";
    system("$import_DE --sqlite $wd/Trinotate.sqlite --samples_file $wd/samples.txt --count_matrix
    \$genes_counts --fpkm_matrix \$genes_expr --DE_dir \$genes_DE_dir --component_mode");
print "Finished at: "; system("date");

### Load expression results for ISOFORMS

    print "Loading expression results for ISOFORMS into Trinotate: "; system("date");
    my $trans_counts = $wd . "/" . $name . "_isoforms.counts.matrix";
    my $trans_expr = $wd. "/". $name . "_isoforms.TMM.EXPR.matrix";
    my $trans_DE_dir = $wd . "/edgeR_trans";
    system("$import_DE --sqlite $wd/Trinotate.sqlite --samples_file $wd/samples.txt --count_matrix
    \$trans_counts --fpkm_matrix \$trans_expr --DE_dir \$trans_DE_dir --transcript_mode");
print "Finished at: "; system("date");

### Load expression clusters for ISOFORMS

    print "Loading expression clusters for ISOFORMS into Trinotate: "; system("date");
    my $trans_clusters = $wd . "/" . $trans_DE_dir .
    "/TestAssembly_trans_DE_analysis.matrix.RData.clusters_fixed_P_25";
my \$analysis_name = \$trans_DE_dir . "/" . \$name .
"/_trans_DE_analysis.matrix.RData.clusters_fixed_P_25";
my \$clusters = \$analysis_name . "/*matrix";
my $group_name = "edgeR_DE_analysis";
    system("$import_clusters --group_name \$group_name --analysis_name \$analysis_name --sqlite
\$wd/Trinotate.sqlite \$clusters");
print "Finished at: "; system("date");
} \# END SKIP
\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#

### Output Trinotate annotation report

\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#

```
```

print "Creating Trinotate annotation report: "; system("date");

```
print "Creating Trinotate annotation report: "; system("date");
my $trinotate_report = "$wd/trinotate_annotation_report.xls";
my $trinotate_report = "$wd/trinotate_annotation_report.xls";
system("$trinotate $wd/Trinotate.sqlite report > $trinotate_report");
system("$trinotate $wd/Trinotate.sqlite report > $trinotate_report");
print "Finished at: "; system("date");
```

print "Finished at: "; system("date");

```

\section*{Calculate-EXN50.pl}
```

\#!/usr/bin/perl
use strict; use warnings;
\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#

### 

### This script will calculate the ExN50 contig length statistics of a Trinity assembly

### Arguments: \#\# Matrix File (e g. "genes.TMM EXPR matrix")

### 1. Expression Matrix File (e.g. "genes.TMM.EXPR.matrix")

2. Assembly
\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#

### User-customized program paths

my \$hd = "/nfs0/IB/Weis_Lab/chouinad";
my $trinity_dir = "$hd/Tools/trinityrnaseq-2.1.1";
\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#

### Get arguments and check usage

\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#
unless(@ARGV == 2){
die("USAGE: Calculate_EXN50.pl <Expression-Matrix> <Assembly>\n")
}
my \$expr_file = \$ARGV[0];
my \$assembly = \$ARGV[1];
\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#

### Run the ExN50 script

\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#
system("\$trinity_dir/util/misc/contig_ExN50_statistic.pl \$expr_file \$assembly");

```

\section*{GO-enrichment.sh}
\#!/bin/bash
\begin{tabular}{|c|c|c|}
\hline \#\#\# & & \#\#\# \\
\hline \#\#\# & Run GO-enrichment analysis for the gene list files specified & \#\#\# \\
\hline \#\#\# & & \#\#\# \\
\hline \#\#\# & Arguments: & \#\#\# \\
\hline \#\#\# & & \#\#\# \\
\hline \#\#\# & 1. Trinotate Annotation Report (".xls" file) & \#\#\# \\
\hline \#\#\# & 2. Factors File (e.g. "factors.tsv") & \#\#\# \\
\hline \#\#\# & 3. Gene Lengths File (e.g. "gene-lengths.tsv") & \#\#\# \\
\hline \#\#\# & & \#\#\# \\
\hline \#\#\# & Notes: & \#\#\# \\
\hline \#\#\# & & \#\#\# \\
\hline \#\#\# & - The "Factors File" identifies the enrichment conditions (e.g. differential expression) & \#\#\# \\
\hline \#\#\# & - It needs to be in the format: (condition)(tab)(gene_id) & \#\#\# \\
\hline \#\#\# & - Check this, as the order varies by Trinity version & \#\#\# \\
\hline \#\#\# & - For expression, one way to get this is from the per isoform expression ".tsv" file & \#\#\# \\
\hline \#\#\# & - This is produced by "Identify-Expression-Results.pl" & \#\#\# \\
\hline \#\#\# & - Get the factors file by running the comand below on the output of the above script & \#\#\# \\
\hline \#\#\# & & \#\#\# \\
\hline \#\#\# & awk -v OFS='\t' 'NR >1 \{ print \$5, \$1 \}' expression-results.tsv | awk '\$1 !~ 八./' | \ & \#\#\# \\
\hline \#\#\# & sort -u -k2,2 | sort -k1 > factors.tsv & \#\#\# \\
\hline \#\#\# & & \#\#\# \\
\hline \#\#\# & - The "Gene Lengths File" can be generated from the EXN50 input file, for example: & \#\#\# \\
\hline \#\#\# & & \#\#\# \\
\hline \#\#\# & cat \$name_genes.TMM.EXPR.matrix.E-inputs | cut -f 1,2| awk 'NR > 1' > gene-lengths.tsv & \#\#\# \\
\hline \#\#\# & & \#\#\# \\
\hline \#\#\# & - This file needs to correspond to the GENES file, not ISOFORMS, even if running for isoforms & \#\#\# \\
\hline \#\#\# & - Either generate this if it doesn't exist, or get gene lengths file another way & \#\#\# \\
\hline \#\#\# & - You also need a header on this file, e.g.: transcript(tab)length & \#\#\# \\
\hline \#\#\# & - This needs to be done beforehand and the file provided as an argument (as above) & \#\#\# \\
\hline \#\#\# & & \#\#\# \\
\hline
\end{tabular}
\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#
\#\#\# Check usage and get arguments \#\#\#
if \(\underset{\text { then }}{\text { ! }}\) \# -eq 3 ]
then
echo "USAGE: GO-enrichment.sh <ANNOTATION-REPORT-XLS> <FACTOR-FILE> <GENE-LENGTHS-FILE>" exit 1
fi
home=/nfs0/IB/Weis_Lab/chouinad
xls=\$1
factors=\$2
lengths=\$3
\#\#\# Extract GO Terms \#\#\#
\$\{home\}/Tools/trinotate-2.0.2/util/extract_GO_assignments_from_Trinotate_xls.pl --Trinotate_xls \$xls
--trans --include_ancestral_terms > ./GO-terms.txt
\#\#\# Run GOSeq \#\#\#
\$\{home\}/Tools/trinityrnaseq-2.1.1/Analysis/DifferentialExpression/run_GOseq.pl --factor_labeling \$factors \} --GO_assignments ./GO-terms.txt --lengths \$lengths

\section*{Identify-Expression-Results.pl}
```

\#!/usr/bin/perl

```
use strict; use warnings; use File::Basename;
\begin{tabular}{|c|c|c|}
\hline \#\#\# & & \#\#\# \\
\hline \#\#\# & Identify the expression results for each transcript in a Trinotate Annotation Report & \#\#\# \\
\hline \#\#\# & & \#\#\# \\
\hline \#\#\# & Arguments: & \#\#\# \\
\hline \#\#\# & & \#\#\# \\
\hline \#\#\# & 1. Trinotate Annotation Report File & \#\#\# \\
\hline \#\#\# & - e.g. "\$dir/trinotate_annotation_report.xls" & \#\#\# \\
\hline \#\#\# & 2. Isoform TMM-Normalized Expression Matrix File & \#\#\# \\
\hline \#\#\# & - e.g. "\$dir/Trinity_isoforms.TMM.EXPR.matrix" & \#\#\# \\
\hline \#\#\# & 3. Isoform Expression Cluster Directory & \#\#\# \\
\hline \#\#\# & - e.g. "\$dir/edgeR_trans/Trinity_isoforms_DE.matrix.RData.clusters_fixed_Ktree_5" & \#\#\# \\
\hline \#\#\# & & \#\#\# \\
\hline \#\#\# & Notes: & \#\#\# \\
\hline \#\#\# & & \#\#\# \\
\hline \#\#\# & - The script prints to standard output, so redirect the output to save to a file & \#\#\# \\
\hline \#\#\# & & \#\#\# \\
\hline
\end{tabular}
```


### Check usage and get arguments

unless(@ARGV == 3){
die("USAGE: Identify-Expression-Results.pl <Trinotate-Report-File> <TMM-EXPR-Matrix> <Isoform-
Expression-Cluster-Directory>\n")
}

```
my \$report = \$ARGV[0];
my \$matrix = \$ARGV[1];
my \$clusters = \$ARGV[2]
\#\#\# Get a list of all the isoforms in the Trinotate report and add character to indicate an end of the id \#\#\#
system("awk '\{ if(NR>1) print \(\backslash \$ 2\) \}' \$report > ID-list.txt");
system("sed -i 'sハ\$/:/' ID-list.txt");
\#\#\# Read each sequence, save the order, and save the header as keys for cluster and TMM hashes \#\#\#
my @SEQ_ORDER;
my \%CLUSTER_HASH;
my \%TMM_HASH;
open(LIST, "<", "ID-list.txt");
while(<LIST>) \{
    \# Read the sequence ID - make sure it is in the right format so as to not hit any unintended matches
\#
    my \$id = \$_;
    chomp(\$id);
    \$id \(=\sim\) s \(/\) s \(+\$ / /\)
    unless(\$id \(\left.=\sim / T R I N I T Y \_D N \backslash d+\_c \backslash d+\_g \backslash d+\_i \backslash d+: \$ /\right)\{\)
        die("Error: the transcript ID file does not match the Trinity transcript header format
used! \n");
                                    \}
\# Save the sequence order \#
push(@SEQ_ORDER, \$id);
\# Save the transcript header as a hash key with the default value (cluster number) as having no
cluster \#
    \$CLUSTER_HASH\{\$id\} = ".";
    \$TMM_HASH\{\$id\} = "ERROR!";
    \}
close(LIST);
\#\#\# Read each expression cluster file and save the cluster number as the value for that header in the cluster hash \#\#\#
my @cluster_files = <\$clusters/*fpkm.matrix>;
foreach(@cluster_files) \{
```


# Get the cluster number from the input file

```
my \$cluster_name = basename(\$_)
my \$cluster_number;
```

    if($cluster_name =~ /subcluster_(\d+)_/){
            $cluster_number = $1;
            $\mp@code{$}
        else{die("Error: Expression cluster number could not be read - check that the file names meet the
    matching criteria!\n")}
\# Read the cluster file and save headers while ensuring unique matches via the same header
modification \#
open(CLUSTER_ID_LIST, "<\$_");
my $line_counter=1;
    while(<CLUSTER_ID_LIST>){
            if($line_counter == 1){
\$line_counter++;
next;
else{
my @line = split(/\t/, \$_);
my \$header = $line[0];
                chomp($header);
$header =~ s/\s+$//;
$header =~ s/$/:/;
$CLUSTER_HASH{$header} = \$cluster_number;
\$line_counter++;
}
}
close(CLUSTER_ID_LIST);
}

```
\#\#\# Read the TMM matrix file and save results in a hash \#\#\#
```

open(TMM_ID_LIST, "<\$matrix");
my \$line_counter=1;
my $tmm_header;
while(<TMM_ID_LIST>) {
    if($line_counter == 1){
my @line = split(/\t/, \$_);
my \$seq = shift(@line);
$tmm_header = join("\t", @line);
                    chomp($tmm_header);
\$line_counter++;
next;
else{
my @line = split(/\t/, \$_);
my \$header = shift(@line);
my $tmm_result = join("\t",@line);
            chomp($tmm_result);
chomp(\$header);
$header =~ s/\s+$//;
$header =~ s/$/:/;
$TMM_HASH{$header} = \$tmm_result;
\$line_counter++;
}
}
close(TMM_ID_LIST);
\#\#\# Print out the sequences with their expression cluster and TMM results \#\#\#
print "Transcript" . "\t" . \$tmm_header . "\t" . "Cluster" . "\n";
foreach(@SEQ_ORDER) \{
my \$cluster = \$CLUSTER_HASH\{\$_\};
my \$tmm_result = \$TMM_HASH\{\$_\};
my \$header = \$_;
\$header $=\sim$ s/:\$//;
print \$header . "\t" . \$tmm_result . "\t" . \$cluster . "\n";
\}
\#\#\# Clean up \#\#\#
system("rm ID-list.txt");

```

\section*{ExtractContigs.pl}
```

\#!/usr/bin/perl
use warnings; use Bio::SeqIO;

| \#\#\# |  |
| :---: | :---: |
| \#\#\# | Extract contigs (genes/isoforms) as specified by a file listing their headers |
| \#\#\# |  |
| \#\#\# | Arguments: |
| \#\#\# |  |
| \#\#\# | 1. Full Assembly File (e.g. "Trinity.fasta") |
| \#\#\# | 2. Output File (e.g. "Subset.fasta") |
| \#\#\# | 3. File listing the Trinity-formatted headers for the contigs of interest |
| \#\#\# | 4. "--genes" OR "--isoforms" mode (this is necessary to specify to avoid erroneous matches) |
| \#\#\# |  |
| \#\#\# | Current Trinity format used (YOU WILL NEED TO MODIFY THE SCRIPT IF THIS CHANGES): |
| \#\#\# |  |
| \#\#\# | GENES FORMAT: TRINITY_DN21111_c0_g1 |
| \#\#\# | ISOFORMS FORMAT: TRINITY_DN21111_c0_g1_i1 |
| \#\#\# |  |
| \#\#\# | Note: The "CONTIG-FILE" must be a single header (for the contig to be extracted) per line |
| \#\#\# |  |

### Check usage and get arguments

unless(@ARGV == 4){
die("Usage: ExtractContigs.pl <ASSEMBLY> <OUT-FILE> <CONTIG-FILE> <--genes OR --isoforms>\n")
}
my \$assembly = \$ARGV[0]
my \$outfile = \$ARGV[1];
my \$contig_file = \$ARGV[2];
my \$mode = $ARGV[3];
unless(($mode eq "--genes") || ($mode eq "--isoforms")){
    die("Error: Mode \"$mode\" is not understood - please use \"--genes\" or \"--isoforms\"\n")
}

```
\#\#\# Open the requested contigs file and save the headers as a hash \#\#\#
my \%requests;
```

open(REQUESTS, "<\$contig_file")
while(<REQUESTS> ) {
my \$request = $_;
    chomp($request);
\$request =~ s/^>//;
$request =~ s/\s+$//;
\# Use magic to make sure the header will match only the single gene/isoform in question \#
if($mode eq "--genes"){
        if($request =~ /(TRINITY_DN\d+_c\d+_g\d+)\$/){
$request =~ s/$/_/;
}
else{die("Error: In \"--genes\" mode the request must end with the gene number!\n")}
}
elsif($mode eq "--isoforms"){
            if($request =~ /(TRINITY_DN\d+_c\d+_g\d+_i\d+)\$/){
$request =~ s/$/_/;
}
else{die("Error: In \"--isoforms\" mode the request must end with the isoform number!\n")}
else{die("Error: Mode \$mode is not understood!\n")}
\# Load it into the hash for reference later \#
$requests{$request} = 1;
}
close(REQUESTS);

```
\#\#\# Take in each sequence and print it out if it is present in the requested contigs file \#\#\#
```

$seq_in = Bio::SeqIO->new(-file => "<$assembly", -format => "fasta")
$seq_out = Bio::SeqIO->new(-file => ">$outfile", -format => "fasta");
while(\$contig = \$seq_in->next_seq){

```
```


# Identify ID and magically modify it for the appropriate mode

my \$cseq = $contig->display_id;
chomp($cseq);

```
```

\$cseq =~ s/^>//;
$cseq =~ s/\s+$//;
if(\$mode eq "--genes"){
\$cseq =~ s/(TRINITY_DN\d+_c\d+_g\d+_)i\d+/$1/;
elsif($mode eq "--isoforms"){
$cseq =~ s/$/_/;
}
else{die("Error: Mode \$mode is not understood!\n")}

# Print it out if it is present in the contig requests file

if(exists($requests{$cseq})){
$seq_out->write_seq($contig);
\$\mp@code{\}
}

```

\section*{ExtractLongest.pl}
```

\#!/usr/bin/perl
use warnings; use Bio::SeqIO;
\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#
\#\#\#
\#\#\# Extract the longest "transcript" for each "gene" from Trinity output \#\#\#

```

```

Man

### 1. Assembly file (e.g. Trinity.fasta)

### 2. Output file

### 

### Notes:

    - Make sure the Trinity output still matches the regex used
    
## Open dependencies, require arguments and explain usage

use warnings;
use Bio::SeqIO;
unless(@ARGV == 2){die("Usage: ExtractLongest.pl IN-FILE.fasta OUT-FILE.fasta\n")}

## Open the input and output file

\$seqIO_obj = Bio::SeqIO->new(-file => $InFileName, -format => "fasta");
open (OUTFILE, ">>$OutFileName") or die ("Can't open output file because: \$!\n");

## Read in sequences from file, identify and store longest isoform per gene

while(\$seq_obj = \$seqIO_obj->next_seq){
\#\# Initialize loop \#\#
\$Short_ID = "";
\$Full_ID = "";
\$ID = "";
\$Gene = "";
\#\# Identify ID and sequence \#\#
\$Short_ID = \$seq_obj->display_id;
\$Full_ID = \$seq_obj->desc;
$ID = "$Short_ID \$Full_ID";
\$SEQ = \$seq_obj->seq;
\$SEQ_LEN = $seq_obj->length;
    ## Read ID to determine gene and transcript ##
    if ($ID =~ /(TRINITY_DN\d+_c\d+_g\d+)_i.*\$/){
\$Gene = $1;
                if(! exists($lengths{\$Gene})){
push (@GeneOrder, \$Gene)
}
}
else {die("Could not parse Trinity output\n")}

## Determine longest transcript for each gene, store its sequence and full ID

if ((! exists($lengths{$Gene})) || (\$SEQ_LEN > $lengths{$Gene})){
$lengths{$Gene} = \$SEQ_LEN;
$seqs{$Gene} = \$SEQ;
$isoform{$Gene} = \$ID;
}
}
foreach \$i (@GeneOrder){
print OUTFILE ">" . $isoform{$i} . "\n";
print OUTFILE $seqs{$i} . "\n";
}
close OUTFILE;

```

\section*{Find-Secreted.pl}
\#!/usr/bin/perl
use strict; use warnings;

```


### Check usage and get arguments

unless(@ARGV == 2){
die("USAGE: Find-Secreted.pl <Trinotate-Annotation-Report.xls> <Output-File>\n");
}
my \$report = \$ARGV[0];
my \$output = \$ARGV[1];

```
\#\#\# Initialize tallies \#\#\#
my \(\$ \mathrm{sp}=0\); \# signal peptide counter \#
my \$es = 0; \# extracellular space counter \#
my \$ST = 0; \# total secreted counter
\#\#\# Loop through lines and check each for signal peptides and/or extracellular space GO components \#\#\#
open(INFILE, "<\$report") or die("Error: cannot open the trinotate annotation report!\n");
open(OUTFILE, ">\$output") or die("Error: cannot open the trinotate annotation report!\n");
while(<INFILE>) \{
    my \$entry = \$_;
    my \$secreted = 0;
    \# Check for signal peptide \#
    if(\$entry =~ /sigP: \d+\^/)\{
            \$sp++;
            \$secreted = 1 ;
            \$se
\(\}\)
    \# Check for extracellular region \#
    if(\$entry =~ /G0:0005576/) \{
            \$es++;
            \$secreted = 1;
            \}
    \# Check for extracellular space \#
    if(\$entry =~ /GO:0005615/) \{
            \$es++;
            \$secreted = 1;
            \}
    if(\$secreted == 1) \{
            \$ST++;
            print OUTFILE \$entry;
            \}
    \}
close(INFILE);
close(OUTFILE);

print " Total signal peptide entries : " . \$sp . "\n";
print " Total extracellular GO terms : " . \$es . "\n"
print " Total secretion annotations : " . \$ST . "\n"';


\section*{Find-Keywords.pl}
```

\#!/usr/bin/perl
use strict; use warnings;
\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#

### 

### Find, tally, and output entries in a Trinotate Annotation Report matching search terms

\#\#\#

```

```


### (\#\# 1. Trinotate Annotation Report file (xls)

### 2. Output file

### 3. Term to search (can repeat, as long as they start at argument 3)

### Notes:

### Notes:

### - The terms are all concatenated, so each term will be searched for each entry

    - Any match causes the entry to be printed to the output file
    - Some entries in the annotation file enter incorrectly
        - This can cause single entries to occupy multiple lines
    
### - Thus, it's best to visually-inspect the annotation report before downstream analyses

        ###
    
### - Thus, it's best to visually-inspect the annotation report before downstream analyses

\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#

### Check usage and get arguments

unless(@ARGV >= 2){
die("USAGE: Find-Keywords.pl <Trinotate-Annotation-Report.xls> <0utput-File> <Term-1> <Etc...>\n");
}
my \$report = shift(@ARGV);
my \$output = shift(@ARGV);
my \$terms = join('|',@ARGV);

### Initialize tallies

my \$count = 0;

### Loop through lines and check each for signal peptides and/or extracellular space GO components

open(INFILE, "<$report") or die("Error: cannot open the trinotate annotation report!\n");
open(OUTFILE, ">$output") or die("Error: cannot open the trinotate annotation report!\n");
while(<INFILE>) {
my \$entry = \$_;
my $match = 0;
    if($entry =~ /\$terms/i){
$match = 1;
        }
    if($match == 1){
\$count++;
print OUTFILE \$entry;
}
close(INFILE);
close(OUTFILE);
print "---------------------------------------------------------
print " Total matching entries: " . \$count . "\n";
print "-------------------------------------------------

```

\section*{Cluster-Annotations.pl}
\#!/usr/bin/perl
use strict; use warnings;
\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#
\begin{tabular}{lll}
\(\# \# \#\) \\
\(\# \# \#\) & Run a pipeline to extract and cluster sequences from a Trinotate Annotation Report
\end{tabular}


Arguments:
\#
\#\#\#
\#\#
1. Trinotate Annotation Report file (as ".xls" or ".csv")
2. Assembly file (e.g. "Trinity.fasta")
3. Prefix to be used in outputs (e.g. "Secreted") \#\#\#

Notes:
\#\#\#
\#\#\#
This refers - The latter is incorporated with annotation reports via "Identify-Expression-Results.pl" \#\#\#
- Some entries in the annotation file enter incorrectly
\#\#\#
- Thus, it's best to visually-inspect the annotation report before downstream analyses
\#\#\#
- This is also designed for annotation reports with total expression values included \#\#\#
- Run "Identify-Expression-Results.pl" and appending these to the annotation report
\#\#\#
\#\#\#
- The script will work on any trinotate report (or subset) that does not include a header line \#\#\#

Pipeline: \#\#\#
1. Get transcript headers, in Trinity format: \$header =~ /(TRINITY_DN\d+_c\d+_g\d+_i\d+)/ \#\#\#

This identifies the transcripts in the report necessary for subsets _ _ _
This identifies the transcripts in the report, necessary for subsets
\#\#\#
2. Extract the appropriate contigs (from the assembly) via "ExtractContigs.pl" \#\#\#
3. Convert them into a short-name version (for CD-HIT) and save a reference key \#\#\#
4. Run CD-HIT-EST to cluster contigs at \(80 \%\)
\#\#\#
5. Parse the CD-HIT-EST results (using the name reference key)
\#\#\#
6. Extract annotation reports and print it after the transcript cluster number (append cluster) \#\#\#
7. Get the (unique) BLAST hit names per transcript, tally and sort total cluster expression \#\#\#

Output:
\#\#\#
Output: \#\#\#
\$prefix-Headers.txt \(\quad=\quad\) List of contig headers in the annotation report \(\quad\) \#\#\#
\$prefix-Contigs.fasta \(\quad=\quad\) Contigs corresponding to the annotation report \#\#\#
\$prefix-Contigs.fasta.shortnames \(\quad=\quad\) Contigs converted to shortened naming system \#\#\#
\$prefix-Contigs.fasta.namesindex \(\quad=\quad\) A reference file for converting back to old names \(\# \# \#\)
\$prefix-Clusters.fasta \(\quad=\) Cluster representatives from the CD-HIT-EST output
\$prefix-Clusters.txt \(\quad=\) List of clusters and their transcripts (headers)
\$prefix-Clusters-Annotations.csv \(=\) Annotation report with cluster number as first entry
\$prefix-Clusters-Expression-Hits.tsv \(=\) File containing a line for each transcript:
its cluster number, total expression (TMM), the
\#\#\#
umulative cluster expression, and all unique full \#\#\#
names from any BLAST hits in the annotation report \#\#\#
\$prefix-BLASTs-Sorted-Expression.tsv = File containing a line for each BLAST hit:
the total expression for transcripts with that hit, \#\#\#
the number of transcripts with that hit, and \#\#\#
the name of the unique BLAST hit \#\#\#
\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#\#
\#\#\# User-defined options and program paths \#\#\#
\begin{tabular}{ll} 
my \$extract_contigs & \(=\) "/nfs0/IB/Weis_Lab/chouinad/Tools/ExtractContigs.pl"; \\
my \$cdhitEST & \(=\) "/nfs0/IB/Weis_Lab/chouinad/Tools/cd-hit-v4.6.1-2012-08-27/cd-hit-est";
\end{tabular}

```

        my $header_list = $prefix . "-Headers.txt";
        my %reporthash;
    open(INFILE, "<$report") or die("Error: cannot open the trinotate annotation report!\n");
    open(OUTFILE, ">$header_list") or die("Error: cannot open the transcript header output file!\n");
    while(<INFILE>){
        my $entry = $_;
        if($entry =~ /(TRINITY_DN\d+_c\d+_g\d+_i\d+),/){
            my $trans = $1;
            print OUTFILE $trans . "\n";
            $reporthash{$trans} = $entry';
            }
    }
    close(INFILE);
    close(OUTFILE);
    
### Use "ExtractContigs.pl" to, well, extract the contigs

    # Usage: ExtractContigs.pl <ASSEMBLY> <OUT-FILE> <CONTIG-FILE> <--genes OR --isoforms> #
        my $contig_file = $prefix . "-Contigs.fasta";
        system("$extract_contigs $assembly $contig_file $header_list --isoforms");
    
### Convert contig names into short version for CD-HIT - this is seriously absurd

    my $short_names = $prefix . "-Contigs.fasta.shortnames";
        my $ref = $prefix . "-Contigs.fasta.namesindex";
    open(INFILE, "<$contig_file") or die("Error: cannot open the extracted contig file!\n");
    open(OUTFILE, ">$short_names") or die("Error: cannot open the contig shortname output file!\n");
    open(REF, ">$ref") or die("Error: cannot open the name conversion reference file!\n");
    my $count = 1;
    while(<INFILE>){
    my $line = $_;
        if($line =~ /(TRINITY_DN\d+_c\d+_g\d+_i\d+)\s/){
                        my $name = $1;
                        print OUTFILE ">" . $count . "\n";
                        print REF $count . "\t" . $name . "\n";
                $count++;
            }
            else{
            print OUTFILE $line;
            }
        }
    close(INFILE);
close(OUTFILE);
close(REF);
\#\#\# Cluster with CD-HIT-EST at $80 \%$ - the lowest identity possible to date \#\#\#

```
```

my \$cluster_out = \$prefix . "-Clusters.fasta";

```
my $cluster_out = $prefix . "-Clusters.fasta";
system("$cdhitEST -i $short_names -o $cluster_out -c 0.8 -n 4 -g 1");
system("$cdhitEST -i $short_names -o $cluster_out -c 0.8 -n 4 -g 1");
\#\#\# Parse CD-HIT-EST clustering results \#\#\#
```

```
# Read in reference file and save a hash of the short names and the full headers #
```


# Read in reference file and save a hash of the short names and the full headers

open(REF, "<$ref") or die("Error: cannot open the name conversion reference file!\n");
open(REF, "<$ref") or die("Error: cannot open the name conversion reference file!\n");
my %hash;
my %hash;
while(<REF>){
while(<REF>){
my \$line = \$_;
my \$line = $_;
    chomp($line);
chomp($line);
    my @NAMES = split(/\t/,$line);
my @NAMES = split(/\t/,\$line);
my \$short = \$NAMES[0];
my \$short = \$NAMES[0];
my \$long = \$NAMES[1];
my \$long = \$NAMES[1];
$hash{$short}=\$long;
$hash{$short}=\$long;
}
}
close(REF);
close(REF);

# Open cluster results file, write a file with clusters and full Trinity headers

```
# Open cluster results file, write a file with clusters and full Trinity headers #
```

```
my $cluster_results = $cluster_out . ".clstr";
my $cluster_headers = $prefix . "-clusters.txt";
open(INFILE, "<$cluster_results") or die("Error: cannot open the CD-HIT-EST cluster results!\n");
open(OUTFILE, ">$cluster_headers") or die("Error: cannot open the parsed cluster result output!\n");
while(<INFILE>){
    my $line = $_;
        if($line =~ />Cluster\s(\d+)$/){
                            my $cluster = $1;
            print OUTFILE "### CLUSTER " . $cluster . " ###\n";
            }
        elsif($line =~ />(\d+)\.{3}/){
            my $shortname = $1;
            print OUTFILE $hash{$shortname} . "\n";
            @}
    }
close(INFILE);
close(OUTFILE);
\#\#\# Extract annotations after clustering, ordered and numbered according to clusters \#\#\#
```

```
my $clustered_anns = $prefix . "-Clusters-Annotations.csv";
```

my \$clustered_anns = \$prefix . "-Clusters-Annotations.csv";

# Read the cluster results header file and print out the full line with cluster numbers

# Read the cluster results header file and print out the full line with cluster numbers

open(RESULTS, "<$cluster_headers") or die("Error: cannot open the cluster results headers file!\n");
open(RESULTS, "<$cluster_headers") or die("Error: cannot open the cluster results headers file!\n");
open(OUTFILE, ">$clustered_anns") or die("Error: cannot open the clustered annotation output!\n");
open(OUTFILE, ">$clustered_anns") or die("Error: cannot open the clustered annotation output!\n");
my \$clust;
my \$clust;
while(<RESULTS>){
while(<RESULTS>){
my \$line = \$_;
my \$line = $_;
    chomp($line);
chomp($line);
    if($line =~ /CLUSTER\s(\d+)\s/){
if(\$line =~ /CLUSTER\s(\d+)\s/){
\$clust = \$1;
\$clust = $1;
            }
            }
    elsif($line =~ /^TRINITY/){
elsif(\$line =~ /^TRINITY/){
print OUTFILE \$clust . "," . $reporthash{$line};
print OUTFILE \$clust . "," . $reporthash{$line};
}
}
}
}
close(RESULTS);
close(RESULTS);
close(OUTFILE);
close(OUTFILE);
\#\#\# Get unique BLAST hit names, tally total cluster expression, print out the final cluster annotations \#\#\#

# Open clustered annotation file, print cluster, expression, and BLAST hit names

open(INFILE, "<\$clustered_anns") or die("Error: cannot open the clustered annotation report!\n");
open(TMPOUT, ">", "tmp.ann.csv") or die("Error: cannot open the temporary output!\n");
\$clust = "";
my \$exp = "";
my @BLASTs = ();
while(<INFILE>){
my \$entry = \$_
$entry =~ s/\s+$//;
if(\$entry =~ /^(\d+'),/){
\$clust = $1;
        if($entry }}=~ /(\d+\.\d+)\$/)
\$exp = \$1;
$exp = sprintf("%.2f",$exp);
}
print TMPOUT \$clust . "\t" . \$exp;
my @captures = ( \$entry =~ /\^.{3}Name:\sFull=(.+?);\^/g );
my @names = ();
foreach(@captures){
my \$name = $_;
            unless($name~~ @names){
push(@names, \$name);
}
foreach (@ @names){
print TMPOUT "\t" . $_;
            unless($_ ~~ @BLASTs){
push(@BLASTs, \$_);
}

```
```

        print TMPOUT "\n";
    }
    close(INFILE);
close(TMPOUT);

# Tally cluster expression and save it (value) in a hash per cluster number (key)

open(TMPIN, "<", "tmp.ann.csv") or die("Error: cannot open the temporary annotation file!\n");
my \$newclust = "";
my \$oldclust = -1;
my %clusthash = ();
my \$ann;
\$exp = "";
\$count = 1;
while(<TMPIN>) {
my \$entry = $_
        if($entry =~ /(\d+)\s+(\d+\.\d{2})(\s*.*\$)/){
\$newclust = \$1;
\$exp = \$2;
\$ann = $3;
        if($newclust != \$oldclust){
$clusthash{$newclust} = \$exp;
\$oldclust = $newclust;
        }
        elsif($newclust == \$oldclust){
$clusthash{$newclust} += \$exp;
else{die("Error: cannot compare cluster numbers!\n")}
\$count++;
}
else{die("Error: cannot read clustered annotation line \$count\n");}
}

# Read in file - print cluster number, transcript expression, cluster expression, and annotation

my \$cluster_hits = $prefix . "-Clusters-Expression-Hits.tsv";
open(TMPIN, "<", "tmp.ann.csv") or die("Error: cannot open the temporary annotation file!\n");
open(OUTFILE, ">$cluster_hits") or die("Error: cannot open the output file!\n");
\$clust = "";
\$exp = "";
\$ann = "";
while(<TMPIN>){
my \$entry = \$_;
my \$tmp = $entry;
    if($entry =~ /(\d+)\s+(\d+\.\d{2})(.*)\$/){
\$clust = \$1;
\$exp = \$2;
\$ann = \$3;
else{die("Error: cannot read clustered annotation line \$count\n");}
print OUTFILE \$clust . "\t" . \$exp . "\t" . $clusthash{$clust} . "\t" . \$ann . "\n";
}
close(TMPIN);
close(OUTFILE);
unlink("tmp.ann.csv");
\#\#\# Print out unique BLAST hits \#\#\#

```
```

my \$blasts = \$prefix . "-BLAST-Hits.txt";

```
my $blasts = $prefix . "-BLAST-Hits.txt";
open(BLAST_TMP, ">", "blast-hits.tmp");
open(BLAST_TMP, ">", "blast-hits.tmp");
foreach(@BLASTs){
foreach(@BLASTs){
    chomp;
    chomp;
    unless($_ =~ /uncharacterized/i){
    unless($_ =~ /uncharacterized/i){
                print BLAST_TMP $_ . "\n";
                print BLAST_TMP $_ . "\n";
            }
            }
    }
    }
close(BLAST_TMP);
close(BLAST_TMP);
system("cat blast-hits.tmp | sort | uniq > $blasts");
system("cat blast-hits.tmp | sort | uniq > $blasts");
unlink("blast-hits.tmp");
unlink("blast-hits.tmp");
@BLASTs = "";
```

@BLASTs = "";

```
\#\#\# Tally transcript expression for each unique BLAST hit \#\#\#
```


# Get each unique BLAST hit into an array

my @HITS;
open(HITFILE, "<\$blasts");
while(<HITFILE>) {
my \$hit = $_;
    chomp($hit);
unless(\$hit ~~ @HITS){
push(@HITS, \$hit);
}
}

# For each hit, loop through report, tally expression, and print out total expression and hit name

my %tally;
my %hitcount;
open(HIT_EXP_TMP, ">", "hit-exp.tmp");
foreach(@HITS){
\# Initialize values \#
my \$count = 0;
my \$hit = \$_;
\# Make an entry in the hash to tally expression \#
$tally{$hit} = 0;
$hitcount{$hit} = 0;
\# Loop through and check each annotation \#
open(REPORT2, "<\$cluster_hits");
while(<REPORT2>){
\$count++
my \$line'= $_;
            # Parse entry #
            my @ENTRY = split('\t',$line);
my \$cluster = shift(@ENTRY);
my \$exp = shift(@ENTRY);
my \$clust_exp = shift(@ENTRY);
\# Find if it matches the hit \#
my $match = 0;
                    foreach(@ENTRY) {
                    chomp;
                                    if($_ eq \$hit){
$match = 1;
                                    }
                    if($match == 1){
$tally{$hit} += \$exp;
$hitcount{$hit}++
}
}
close(REPORT2);
my \$final_exp = sprintf("%.2f", $tally{$hit});
my \$ntrans = $hitcount{$hit};
print HIT_EXP_TMP \$final_exp . "\t" . \$ntrans . "\t" . \$hit . "\n";
}
close(HIT_EXP_TMP);
my \$hit_exp_out = \$prefix . "-BLAST-Hits-Sorted-Expression.tsv";
system("cat hit-exp.tmp | sort -nrk1,1 > \$hit_exp_out");
unlink("hit-exp.tmp");

```
```

