De novo Discovery of Novel Long Noncoding RNAs in the North American Beaver

by
Amita Kashyap

A THESIS

submitted to

Oregon State University

Honors College

in partial fulfillment of
the requirements for the
degree of

Honors Baccalaureate of Science in BioResource Research
(Honors Scholar)

Presented November 29, 2017
Commencement June 2018

AN ABSTRACT OF THE THESIS OF

Amita Kashyap for the degree of Honors Baccalaureate of Science in BioResource Research presented on November 29, 2017. Title: *De novo* Discovery of Novel Long Noncoding RNAs in the North American Beaver.

Abstract approved:_____

Stephen A. Ramsey

Long noncoding RNAs (lncRNAs) are RNA molecules that are at least 200 nucleotides long and do not encode proteins. lncRNAs have roles in gene regulation, chromatin epigenetics, and molecular scaffolding. In light of mounting evidence implicating species-specific noncoding RNAs and gene regulatory mechanisms in species adaptations, it is reasonable to speculate that species-specific lncRNAs may underlie some of the adaptations seen in mammalian evolution. The genome and transcriptome of the North American beaver (*Castor canadensis*, a keystone species of northwest wetlands) have recently been sequenced for the first time, enabling a search for genomic adaptations of this unique semi-aquatic herbivore. The objective of this study was to identify novel lncRNAs in the beaver using a computational analysis of high-throughput sequencing data from Oregon State University's recently-released beaver genome and a pan-tissue composite transcriptome that we obtained from 16 tissues from a beaver. We found 182 novel lncRNAs and 113 lncRNAs with a known ortholog in another species. Nine candidate lncRNAs stood out for having the strongest evidence across the various performance measures. One novel lncRNA (contig62060.1) was the only putative novel multi-exonic lncRNA we detected. These novel lncRNAs may serve as a basis for hypothesis generation for targeted functional investigations.

Key Words: long noncoding RNA, *Castor canadensis*, beaver, transcriptome

Corresponding e-mail address: mitaami@gmail.com

De novo Discovery of Novel Long Noncoding RNAs in the North American Beaver

by
Amita Kashyap

A THESIS

submitted to

Oregon State University

Honors College

in partial fulfillment of
the requirements for the
degree of

Honors Baccalaureate of Science in BioResource Research
(Honors Scholar)

Presented November 29, 2015
Commencement June 2018

Honors Baccalaureate of Science in BioResource Research project of Amita Kashyap presented on November 29, 2017.

APPROVED:

_____

Stephen A. Ramsey, Mentor, representing the Department of Biomedical Sciences and the School of Electrical Engineering and Computer Science

_____

David Hendrix, Committee Member, representing the Department of Biochemistry/Biophysics and the School of Electrical Engineering and Computer Science

_____

Katharine G Field, Committee Member, Director, BioResource Research

_____

Toni Doolen, Dean, Oregon State University Honors College

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

_____

Amita Kashyap, Author

Background

Long noncoding RNAs (lncRNAs) are RNA molecules that are least 200 nucleotides (nt) long and do not encode proteins. Unlike messenger RNAs (mRNAs), which code for proteins, lncRNAs have roles in gene regulation, chromatin epigenetics, and molecular scaffolding. For example, the primary effector molecule for X chromosome inactivation is a lncRNA [1]. More broadly, various noncoding RNAs (ncRNAs) have been implicated in host defense against specific pathogens and in responses to various stressors, including hypoxia [2, 3]. In light of mounting evidence implicating species-specific ncRNAs and gene regulatory mechanisms in species adaptations [2, 4], including various species-specific responses to hypoxia [2, 3], it is reasonable to speculate that species-specific lncRNAs may underlie some of the adaptations seen in mammalian evolution.

The genome and transcriptome of the North American beaver (*Castor canadensis*, a keystone species of northwest wetlands) have recently been sequenced for the first time, enabling a search for genomic adaptations of this unique semi-aquatic herbivore. For example, the beaver can hold its breath for up to fifteen minutes [5]. This capability may imply that the beaver has adaptations in the brain and other organs to mitigate hypoxia-associated tissue damage. The lungs may also have adaptations that optimize uptake of oxygen. The beaver's unique abilities to digest tree bark [6] and certain toxic plants [7] are also the results of unique adaptations, such as enzymes that can break down the toxins [7] and gut microbiota that can process lignocellulose [8]. Such adaptations – and similar as-yet unknown ones – may be a result of novel genes or gene regulatory mechanisms. Therefore, finding novel lncRNAs unique to the beaver can provide a starting point for elucidating the biological mechanisms underlying the beaver's unique adaptations.

However, identifying a novel lncRNA poses a bioinformatics challenge because a lncRNA is defined in terms of lacking a certain property, i.e. coding for a protein, and a simple length cutoff. Therefore, it is not possible to identify a potential lncRNA by

isolating a novel protein product, as is the case with mRNA. Furthermore, lncRNAs appear to largely lack conserved structural homology across species [9]. The recent advent of computational tools, however, has yielded a number of bioinformatics tools that can perform assessments such as for coding potential [10].

The objective of this study was to identify novel long noncoding RNAs in the North American beaver using a computational analysis of high-throughput sequencing data from Oregon State University's recently-released beaver genome (BioProject Accession: PRJEB19765; GenBank Accession: GCA_900168385.1) and a pan-tissue composite transcriptome that we obtained from 16 tissues from a beaver. We implemented a bioinformatic pipeline that filtered candidate transcriptome contigs based on novelty and evidence for coding potential.

Methods

Sample Collection

We collected sixteen tissues for an adult female near-term beaver, including: whole blood, brain, lung, liver, heart, stomach, intestine, skeletal muscle, kidney, spleen, ovaries, placenta, castor gland, tail, and toe-webbing, tongue, and placenta. Blood (200 μL), liver (10.6 mm^3), and brain (24.4 mm^3) tissues were stored in 600 μL TRI reagent (Zymo Research, Irvine, CA) per sample. All solid tissues, including liver and brain samples, were stored in 1mL RNAlater (QIAGEN, Hilden, Germany), 19.7 mm^3 per sample with four samples per tissue type.

RNA Isolation and Sequencing

We isolated RNA from each of the 16 beaver tissues using Zymo Direct-zol RNA MiniPrep (Zymo Research). RNA Integrity Number quality scores obtained by Agilent Bioanalyzer (Agilent Technologies, Santa Clara, CA) were above 6.20 for all tissues. From pooled RNA from the 16 tissues, we prepared an RNA-seq library for Illumina sequencing using the PrepX RNA-Seq for Illumina Library Kit (WaferGen Biosystems, Fremont, CA). We sequenced the pooled polyA+ transcriptome library (WaferGen Biosystems) on one lane of an Illumina MiSeq

3000 (Illumina, San Diego), obtaining approximately 30 million read pairs (2x76 nt), for a total of 60 million reads.

Transcriptome Assembly

We used RNA-seq reads with PHRED quality scores above 30 for all bases. Because the reads were of such high quality, we did not trim them. Of 12 sequences found to be overrepresented by FastQC, two were clipped using fastq_clipper [11] v534: i) polyA at the end of a sequence and ii) a likely adaptor sequence, as determined by no hit to the NCBI Nucleotide Database [12] using BLASTn [13].

Six assemblies were produced using a variety of assemblers: Velvet-Oases [14], BinPacker v1.0 [15], Maker Gene Models [16], and three Trinity assemblies [17]. Two Trinity assemblies were *de novo*, one using only those reads that were still paired after clipping and one using all reads retained after clipping. The third Trinity assembly used the Oregon State University Beaver Genome Project draft genome from a male beaver from Oregon (BioProject Accession: PRJEB19765; GenBank Accession: GCA_900168385.1) for genome-guided assembly. All assembly programs used the standard default parameters with the exception of Velvet-Oases, which had a k-mer value of 39 and minimum contig length of 201 nt.

Analysis of mammalian conserved ortholog content using BUSCO [18] showed that the individual transcriptome assemblies complemented each other. We merged the individual transcriptome assemblies into a consensus transcriptome using transfuse [19] v0.5.0 with the default *i* value of 1.0.

Novel lncRNA Discovery Pipeline

We filtered the merged transcriptome assembly to eliminate contigs that had evidence for coding potential or that had been studied before in an orthologous system (Fig. 1). None of the contigs in the merged assembly were below 200 nucleotides in length, so filtering by a length cutoff was unnecessary.

The first filter was annotation by BLASTn [13] against the NCBI Nucleotide Database [12] with an *E* value of 10^{-3}. Those contigs that did not receive an

annotation passed on to the next filter, which was analysis by the Coding Potential Assessment Test, or CPAT [10]. Using the CPAT-calculated coding probability as the *p* value, those contigs which had a False Discovery Rate less than 0.01 (FDR < 0.01) were deemed putative noncoding RNAs. These putative lncRNA contigs were then inspected for protein domains using the HMMscan online tool [20] against the Pfam database [21] with the gathering threshold [22]. The contigs that passed successfully through all these filters were deemed putative novel lncRNAs.

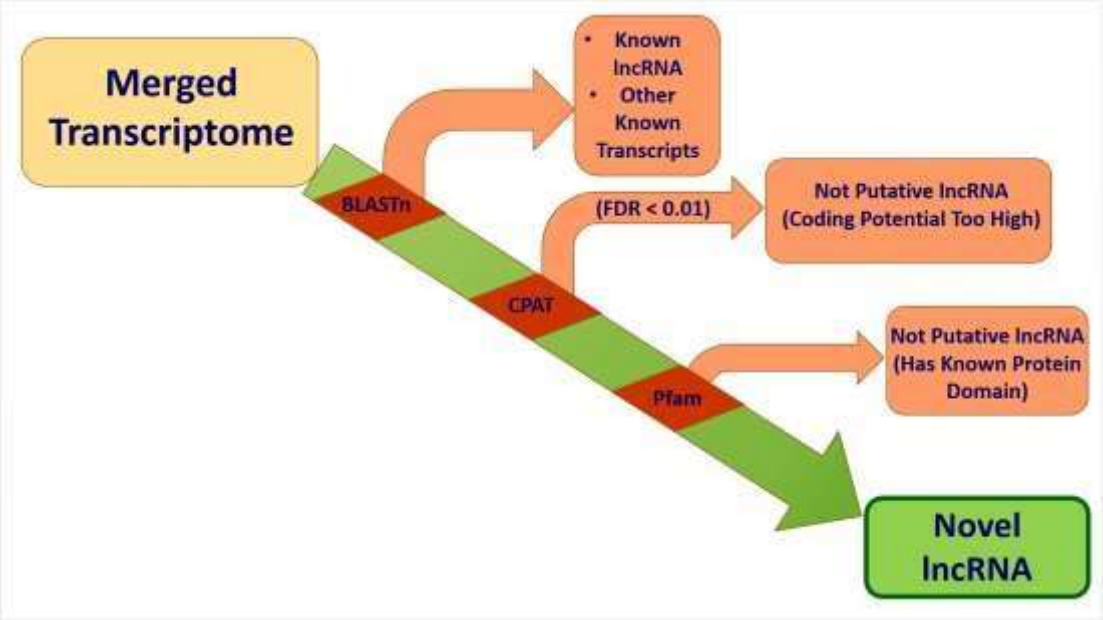

Fig. 1

Overview of the screening pipeline. Transcript contigs from the merged consensus transcriptome obtained using the transfuse program were passed through a series of consecutive filters: 1) BLASTn, 2) CPAT with an FDR < 0.01 for the coding potential, and 3) a scan against the Pfam database using the HMMscan tool. At each step, those contigs that did not meet the requirements of the filter (novelty or lack of evidence for coding characteristics) were eliminated from consideration. The remaining contigs passed on to the next filter. The contigs that successfully passed all filters are the 182 putative novel lncRNAs.

Analysis of Novel lncRNAs

The putative novel lncRNAs were aligned to the Oregon State University Beaver Genome Project draft genome (BioProject Accession: PRJEB19765; GenBank Accession: GCA_900168385.1). Secondary structures and secondary structural information were obtained using the software tool RNAfold [23]. Coverage was calculated using samtools [24] to map RNA-seq reads back to the lncRNA contigs.

Results and Discussion

Screening Pipeline

Our computational pipeline consisted of four steps, each shown in a row of Table 1: (1) joining contigs from different transcriptome assemblies into a consensus assembly; (2) identifying orthologs of known lncRNAs ("known lncRNAs") as well as contigs for which no orthologs could be identified; (3) assessing coding potential of novel contigs based on their hexamer sequence content and the length of and coverage of the transcript by the longest Open Reading Frame (ORF); and (4) testing contigs for known protein domain sequences. Table 1 lists the number of contigs that passed each stage of the pipeline.

A little over half the contigs were successfully annotated by BLASTn and thereby screened out because they were not novel. The False Discovery Rate (FDR < 0.01) on the coding probability given by CPAT was a very aggressive filter, retaining only approximately 0.6% of the remaining contigs. The HHMscan against the PFAM database did not eliminate any more contigs, leaving the final 182 putative novel lncRNAs ("novel lncRNAs").

Table 1

Contig retention through the screening pipeline

Step	# Contigs Eliminated	# Contigs Retained	% Contigs Eliminated
transfuse	N/A	86714	N/A
BLASTn	54402	32312	62.7
CPAT/FDR	32130	182	99.4
PFAM	0	182	0

The first column gives the name of the program. The remaining columns give the number or percentage of contigs obtained from the conclusion of the previous step that were retained or eliminated by each step. Since the transfuse program was merging transcriptome assemblies, the elimination of contigs is not applicable, denoted by N/A. The number of contigs "retained" by transfuse is therefore the number of contigs with which the screening pipeline began.

The putative novel lncRNAs appear to be biased towards shorter contigs

We calculated the secondary structures and Minimum Free Energies (MFE) for both the known and novel lncRNAs using RNAfold [23]. Both sets of lncRNAs had the expected inverse relationship between transcript (contig) length and MFE, though the relationship was weaker in the novel lncRNAs (Fig. 2). The weaker inverse relationship between contig length and secondary structure MFE for the novel lncRNAs may be a byproduct of the differences in length distribution between the two sets of lncRNAs (Fig. 3).

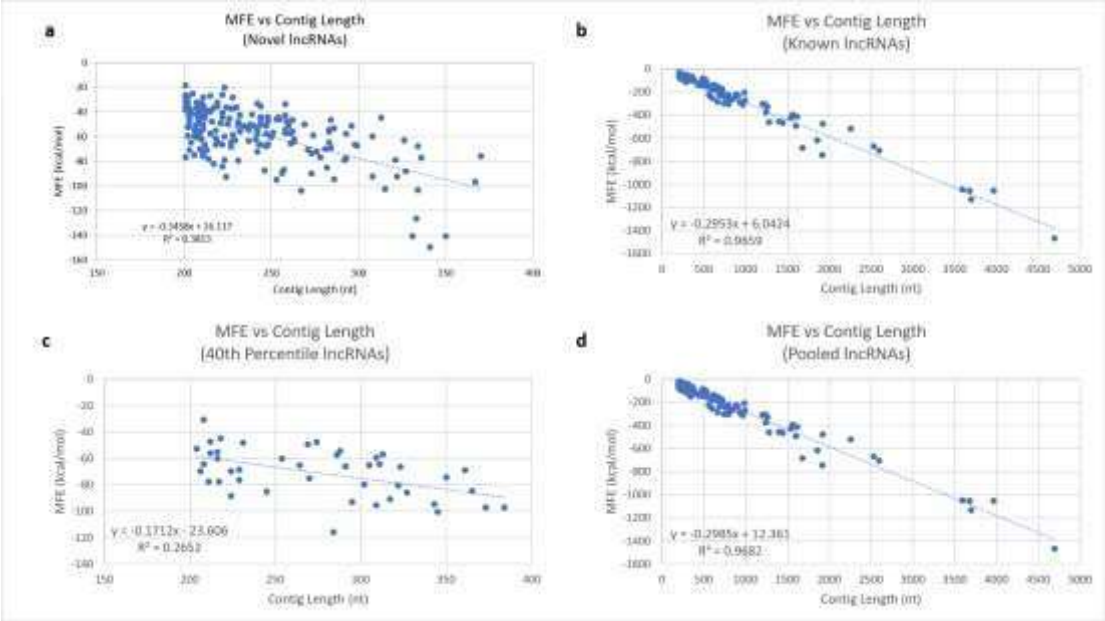

Fig. 2

Minimum Free Energy (MFE) as a function of contig length for the various subsets of lncRNA. **a** Putative novel lncRNAs ("novel lncRNAs"). **b** Contigs found to be orthologous to previously-studied lncRNAs in the NCBI Nucleotide database ("known lncRNAs"). **c** The lower 40th percentile by contig length of known lncRNAs ("40th percentile lncRNAs"), corresponding to the same length range as the novel lncRNA. **d** The set of both known and novel lncRNA, comprising all of the lncRNA in the beaver as per current standards ("pooled lncRNAs").

The length distributions of both the known and novel lncRNA contigs appeared to be exponential (Fig. 3). Whereas the annotated lncRNAs were in the range of 204 - 4691 nt in length, the putative novel lncRNA contigs were all below 400 nt. The putative novel lncRNAs therefore represented the lower 40th (38.94) percentile of the length distribution of the annotated lncRNAs ("40th percentile lncRNA"). The length distribution of the 40th percentile lncRNAs also exhibited an exponential distribution (Fig. 3), though not identical to that of the novel lncRNAs.

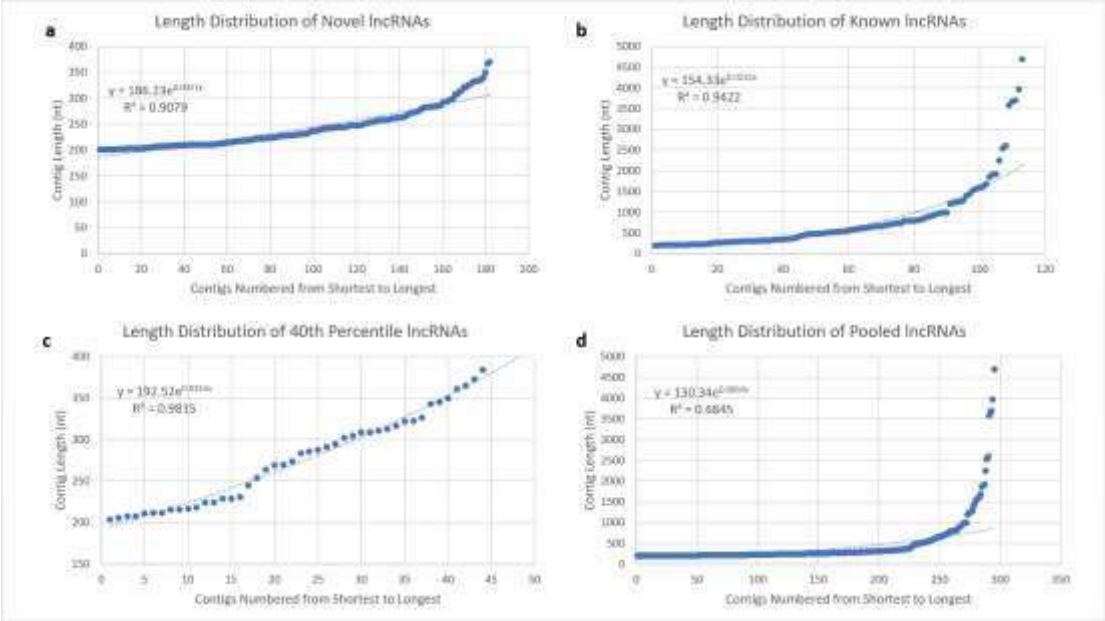

Fig. 3

Length distributions of the various subsets of lncRNA. The lengths are plotted from shortest contig to longest contig along the x-axis to facilitate visual inspection and comparison across subsets. **a** Novel lncRNAs length distribution. **b** Known lncRNAs length distribution. **c** 40th percentile lncRNAs length distribution. **d** Pooled lncRNAs length distribution.

The length distributions of the lower 40th percentile and novel lncRNAs were also differentiated by sample size. We found 182 putative lncRNAs whereas only 44 annotated lncRNAs lie within the same range. It is not unreasonable, however, that we found so many short putative lncRNAs. As much as 68% of genes can encode lncRNA [25] and our genome has approximately 86,714 predicted genes, as roughly estimated by the number of transcript contigs produced by transcriptome assembly. This corresponds to an estimated 58,966 lncRNA transcripts, 23,586 of which would be in the lower 40th percentile by length. However, considering that our putative novel lncRNAs only fell into the lower 40th percentile of lengths for their annotated counterparts, it does seem that our screening pipeline has biased results towards shorter contigs.

The putative novel lncRNAs have coverage comparable to that of the known lncRNAs

We calculated coverage for both the novel and known lncRNAs and compared the distributions and the relationship between coverage and contig length. Coverage is calculated as:

$$Coverage = \frac{\#\ reads\ mapped\ to\ the\ contig \times read\ length}{contig\ length}$$

Coverage distribution, like contig length distribution and the MFE relationships, was also consistent across known and predicted lncRNA subsets. Coverage appeared to have a polynomial distribution (Fig. 4).

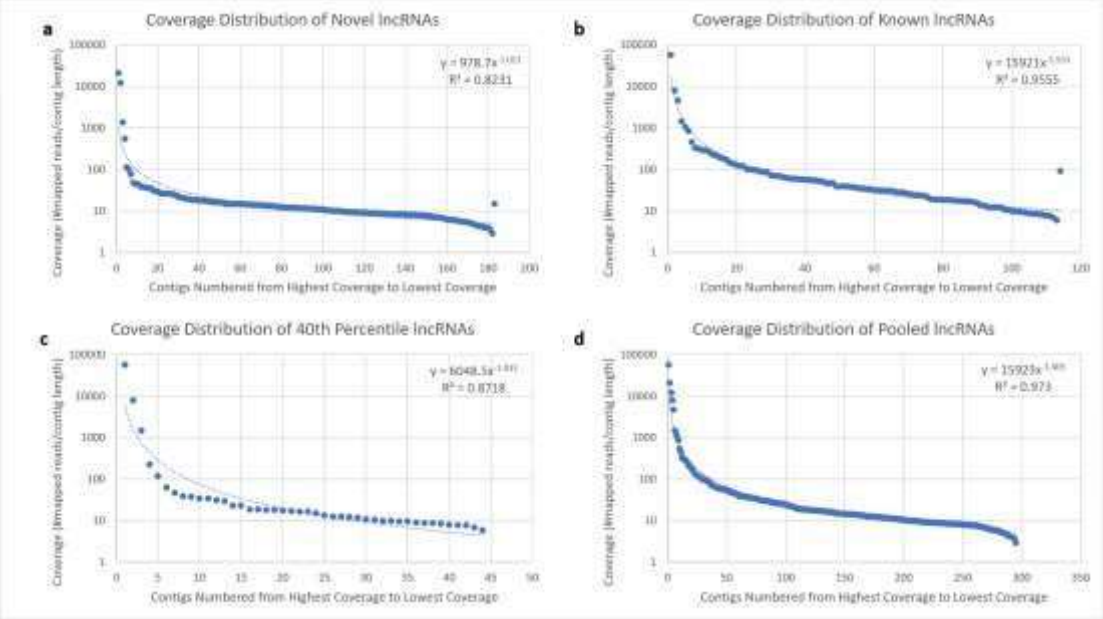

Fig. 4

Coverage distributions of the various subsets of lncRNA. The coverages are plotted from highest coverage contig to lowest coverage contig along the x-axis to facilitate visual inspection and comparison across subsets. **a** Novel lncRNAs coverage

distribution. **b** Known lncRNAs coverage distribution. **c** 40th percentile lncRNAs coverage distribution. **d** Pooled lncRNAs coverage distribution.

Once contigs with exceptionally high coverage were removed from the analysis, average coverage was highest in the known lncRNAs, followed by the 40th percentile lncRNAs, and then the novel lncRNAs. The novel lncRNAs had approximately 75% the average coverage that the 40th percentile lncRNAs had, whereas the 40th percentile lncRNAs had only approximately 26% the average coverage of the entire set of known lncRNAs. Therefore, the coverage of the putative lncRNAs was comparable to their annotated counterparts in terms of contig length. However, the disparity in average coverage between the entire set of known lncRNAs and the lower 40th percentile by length of known lncRNAs indicates that longer contigs may have greater coverage. To investigate this possibility, we compared coverage to contig length and found that they were exponentially related (Fig. 5). Longer contigs also tended to have more coverage, with the relationship, like coverage itself, strongest in the known lncRNA, followed by the 40th percentile lncRNA, and then the novel lncRNA. Unlike with the differences between groups in coverage, the differences between groups in coverage versus contig length were smaller (Table 2).

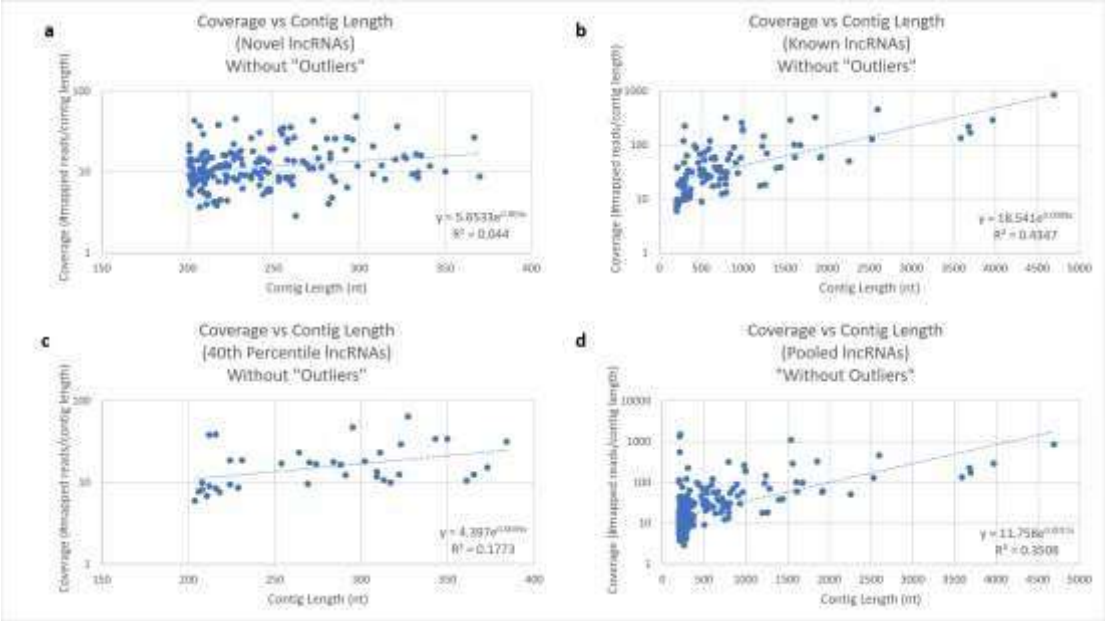

Fig. 5

Coverage vs contig length for the various subsets of lncRNA. “Outlier” contigs were identified by inspection and hence may not be true statistical outliers. **a** Novel lncRNAs coverage vs contig length distribution. **b** Known lncRNAs coverage vs contig length distribution. **c** 40th percentile lncRNAs coverage vs contig length distribution. **d** Pooled lncRNAs coverage vs contig length distribution.

Table 2

Differences in coverage across lncRNA subsets

	Known lncRNAs	**40th Percentile lncRNAs**	**Novel lncRNAs**	**Pooled lncRNAs**
Average Coverage	705.21	1532.65	207.81	398.34
Average Coverage without "Outliers"	70.07	18.27	13.6	50.65
Average Coverage vs Length Ratio (Coverage/Length)	2.5	6.05	0.83	1.47
Average Coverage vs Length Ratio without "Outliers"	0.09	0.07	0.06	0.13
Number of "Outliers"	5	5	7	5

Coverage vs length ratio is calculated as contig coverage/contig length and quantifies the propensity for a contig to be supported by RNA-seq reads as a function of its length. Averages in this table are averages of the individually calculated values for each contig in each subset. “Outlier” contigs were identified by inspection and hence may not be true statistical outliers.

When all of the lncRNAs – both annotated and novel – were examined as a whole (pooled lncRNAs), average coverage fell between that of the known lncRNAs and the 40th percentile lncRNAs, decreasing by approximately 30% from the average coverage calculated for the known lncRNAs. The strength of the relationship between coverage and contig length also fell between that for the entire set of known lncRNAs and the lower 40th percentile of annotated lncRNAs (Table 2).

The similarity between the coverage measures for the novel and 40th percentile lncRNAs and the similarity between the coverage measures for the known and pooled lncRNAs indicate that the novel lncRNAs had comparable coverage to their

annotated counterparts. This evidence supports the hypothesis that the novel lncRNAs are true positives and not falsely assembled contigs.

The putative novel lncRNAs map back to the draft beaver genome

Every novel lncRNA was successfully mapped back to the Oregon State University draft beaver genome (BioProject Accession: PRJEB19765; GenBank Accession: GCA_900168385.1) with upwards of 90% identity using a BLASTn gapped alignment [13]. We calculated the length of the longest alignment as a percentage of the lncRNA contig length (Fig. 6) and found that 91.2% of putative novel lncRNA contigs had an alignment equivalent to at least 70% of the contig's length. This concordance between the novel lncRNA transcript sequences and the genetic sequence indicates that neither is likely to be a false assembly. Furthermore, the mapping serves as a preliminary step in examining the sequence and chromatin context of the putative lncRNA gene. Confirming placement between a transcriptional start site and transcriptional end site would be a next step in confirming or rejecting the putative novel lncRNAs.

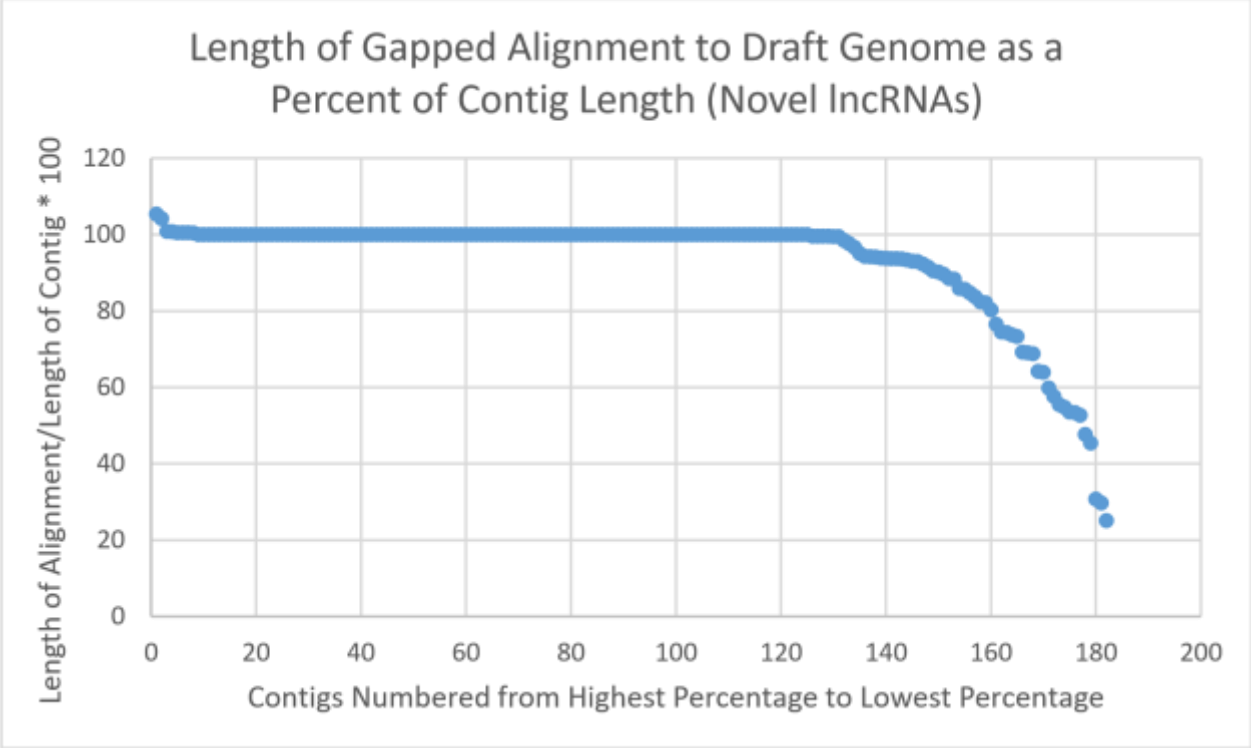

Fig. 6

Gapped genome alignment length as a percentage of contig length. The percentage can be over 100% because the gapped alignment allows intervening unpaired bases in either sequence (transcript contig or draft genome scaffold).

One contig (contig62060.1) had two non-overlapping alignments within 33 nucleotides of each other on the draft genome. This seems indicative of excision of an intron. As a putative two-exon lncRNA, this putative lncRNA would be of particular interest to confirm experimentally.

Novel lncRNAs of particular interest

The novel lncRNAs as a group performed similarly to their annotated counterparts on the measures we used to determine biological plausibility. Nine candidate lncRNAs stood out, however, for having the strongest evidence across the various measures (Table 3). Six of these contigs were among the top ten contigs in terms of at least

length and MFE. This concordance between length and MFE is not surprising in light of the inverse relationship between length of a transcript and secondary structural stability. One novel lncRNA (contig62060.1) was the only putative novel multi-exonic lncRNA we have detected.

Table 3

Novel lncRNA contigs with strongest performance across measurements

	Measure				
Contig	**Length**	**MFE**	**Coverage**	**BLASTn alignment length**	**Intronic**
contig41254.1	Yes	Yes	No	Yes	No
contig43610.1	Yes	Yes	No	No	No
contig44966.1	Yes	Yes	No	No	No
contig46102.1	Yes	Yes	No	No	No
contig45799.1	Yes	No	No	Yes	No
contig46542.1	Yes	Yes	No	No	No
contig46174.1	Yes	Yes	No	No	No
contig59927.1	No	Yes	No	Yes	No
contig62060.1	No	No	No	No	Yes

Contigs of particular interest were determined by having the strongest supportive evidence in more than one category. Cells labeled “Yes” denote that the contig was among the top ten scoring contigs in that category. Cells labeled “No” denote that the contig was not among the top ten scoring contigs in the respective category. The fourth column (BLASTn alignment length) is the length of the gapped alignment as a percentage of the contig length. The last column (Intronic) refers to contigs having a gapped alignment with a gap indicative of a potential splicing event (i.e. the contig may be multi-exonic). Only one putative lncRNA was in this category and hence it is

a contig of particular interest despite not scoring among the highest on any of the other measures.

Interestingly, none of the nine lncRNAs were among those contigs with the highest coverage. This may be explained by the weakness of the relationship between length and observed coverage of novel lncRNA transcripts (Table 2). Furthermore, among the novel transcripts, the seven contigs with exceptionally high coverage had coverage that was, on average, 372-fold greater than that of the rest of the contigs. Additionally, all of these contigs with exceptionally high coverage were under 250 nt long, while the ten longest novel lncRNAs were over 300 nt. Those lncRNAs with the highest coverage may be lncRNAs that are expressed across multiple tissues, though it is possible that they could be extremely strongly expressed in select tissues.

Currently, the lncRNAs do not have associated tissue-specificity. The addition of this information with the planned individual sequencing of RNA from all 16 beaver tissues collected will provide both an opportunity to replicate the current results and a basis for hypothesis generation for targeted functional investigations.

Conclusion

We found 182 potential novel lncRNAs that are expressed in beaver tissues. Nine of these contigs have especially strong evidence across performance measures and may be the most promising contigs to use as a basis for hypothesis generation for targeted functional investigations. To the best of our knowledge, this work represents the first pan-tissue transcriptome analysis of the beaver. Furthermore, this analysis provides a foundation on which to base future work elucidating the biological mechanisms underlying the beaver's unique adaptations.

Following the planned sequencing of RNA from all 16 collected beaver tissues, we will be able to determine the tissue specificity of each of these lncRNAs. Consequently, we will be able to identify lncRNAs that are specific to organs where molecular adaptations might be expected in the beaver, such as lung, liver, and brain.

This will also allow us to produce replicates of this pan-tissue lncRNA discovery pipeline.

Acknowledgments

Thank you to the Oregon Zoo for their contributions. I also thank my many collaborators in the Center for Genome Research and Biocomputing and across Oregon State University. This work was funded in part by OSU College of Veterinary Medicine Biomedical Sciences Student Summer Research Program.

References

1. Lee JT. Epigenetic Regulation by Long Noncoding RNAs. Science. 2012;338:1435–9.
2. Amaral PP, Dinger ME, Mattick JS. Non-coding RNAs in homeostasis, disease and stress responses: an evolutionary perspective. Briefings in Functional Genomics. 2013;12:254–78.
3. Yang F, Huo X, Yuan S, Zhang L, Zhou W, Wang F, et al. Repression of the Long Noncoding RNA-LET by Histone Deacetylase 3 Contributes to Hypoxia-Mediated Metastasis. Molecular Cell. 49:1083–96.
4. Paralkar VR, Mishra T, Luan J, Yao Y, Kossenkov AV, Anderson SM, et al. Lineage and species-specific long noncoding RNAs during erythromegakaryocytic development. Blood. 2014;123:1927–37.
5. National Geographic. Beaver: *Castor canadensis*. http://animals.nationalgeographic.com/animals/mammals/beaver/. Accessed (unknown)
6. Nature Works. Beaver – *Castor canadensis*. http://www.nhptv.org/natureworks/beaver.htm. Accessed (unknown)
7. Müller-Schwarze, Dietland. Arms Race in the Woods: How Beavers Recycle Tree Defenses. 2014.

http://northernwoodlands.org/outside_story/article/beavers. Accessed (unknown)

8. Wong MT, Wang W, Lacourt M, Couturier M, Edwards EA, Master ER. Substrate-Driven Convergence of the Microbial Community in Lignocellulose-Amended Enrichments of Gut Microflora from the Canadian Beaver (*Castor canadensis*) and North American Moose (Alces americanus). Frontiers in Microbiology. 2016;7:961.
9. Ulitsky I. Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. Nature Reviews Genetics. 2016;17:601.
10. Wang L, Park HJ, Dasari S, Wang S, Kocher J-P, Li W. CPAT: CodingPotential Assessment Tool using an alignment-free logistic regression model.
    Nucleic Acids Research. 2013;41:e74–e74.
11. Erik Aronesty (2011). ea-utils : "Command-line tools for processing biological sequencing data"; https://github.com/ExpressionAnalysis/ea-utils
12. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research. 2016;44 Database issue:D7–19.
13. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990;215:403–10.
14. Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNAseq assembly across the dynamic range of expression levels. Bioinformatics.
    2012;28:1086–92.
15. Liu J, Li G, Chang Z, Yu T, Liu B, McMullen R, et al. BinPacker: PackingBased De Novo Transcriptome Assembly from RNA-seq Data. PLoS Computational Biology. 2016;12:e1004772.
16. Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, et al. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Research. 2008;18:188–96.

17. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology. 2011;29:644.
18. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2.
19. Chris Boursnell. transfuse; https://github.com/cboursnell/transfuse.git
20. Finn RD, Clements J, Arndt W, Miller BL, Wheeler TJ, Schreiber F, et al. HMMER web server: 2015 update. Nucleic Acids Research. 2015;43:W30–8.
21. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Research. 2016;44:D279–85.
22. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, et al. The Pfam protein families database. Nucleic Acids Research. 2012;40 Database issue:D290–301.
23. Gruber AR, Lorenz R, Bernhart SH, Neuböck R, Hofacker IL. The Vienna RNA Websuite. Nucleic Acids Research. 2008;36 Web Server issue:W70–4.
24. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
25. Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, et al. The landscape of long noncoding RNAs in the human transcriptome. Nature Genetics. 2015;47:199.