AN ABSTRACT OF THE THESIS OF

Jimmy Bell for the degree of Master of Science in Computer Science presented on March 9, 2020.

Title: Machine Learning Methods for the Discovery and Analysis of MicroRNAs.

Abstract approved:

____

David A. Hendrix

MicroRNAs are a highly conserved class of small endogenous RNA, about ~22nt in length, involved in post-transcriptional gene silencing and have prominent roles in disease and development. Though the process of microRNA discovery was once an arduous task, the advent of high throughput sequencing technology has resulted in novel microRNAs being discovered at a rapid rate. Several data-driven pipelines and machine learning-based methods have been devised so that the beginning stages of microRNA discovery can be performed *in silico*. Despite these efforts, several challenges have persisted in the computational prediction of microRNAs. These challenges include the identification of microRNAs with low expression, proper determination of the precursor span, and the precise labeling of the cleavage sites involved in their biogenesis. This thesis addresses these challenges with two new machine learning-based approaches. MiRWoods improves precursor detection and uses stacked random forests for the sensitive detection of microRNAs. We report that miRWoods has a 10% higher recall of annotated microRNAs when compared with other software. We applied this method to the genomes of human, mouse, *Felis catus* (cat) and *Bos Taurus* (cow) and identified hundreds of novel microRNAs in small RNA sequencing datasets. Our novel predictions include a microRNA in an intron of tyrosine kinase 2 (TYK2), that is present in both cat and cow, as well as a family of mirtrons with two instances in the human genome. Our predictions support a more expanded miR-2284 family in the bovine genome, a larger mir-548 family in the human

genome, and a larger let-7 family in the feline genome. DeepMirCut is a deep learning approach for identifying cleavage sites within microRNAs. This approach is inspired by site-labeling methods for natural language processing, and can accurately predict how the microRNA processing enzymes Dicer and Drosha cleave the microRNA precursor.

©Copyright by Jimmy Bell

March 9, 2020

All Rights Reserved

Machine Learning Methods for the Discovery and Analysis of MicroRNAs

by
Jimmy Bell

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Presented March 9, 2020
Commencement June 2020

Master of Science thesis of Jimmy Bell presented on March 9, 2020

APPROVED:

Major Professor, representing Computer Science

Head of the School of Electrical Engineering and Computer Science

Dean of the Graduate School

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

Jimmy Bell, Author

ACKNOWLEDGEMENTS

The author expresses sincere appreciation for the countless hours David Hendrix has spent training me, helping me improve my research, and proofreading my writing. I would also like to express my gratitude to Christiane Löhr and Maureen Larson for providing the small RNA sequencing data for cat and performing the PCR experiments on microRNAs found within the Cat genome. I'd also like to thank Massimo Bionaz who provided the small RNA sequencing data for cow and who's lab performed PCR experiments on microRNAs from the Cow genome.

CONTRIBUTION OF AUTHORS

Jimmy Bell performed computational analysis presented in this work, including software development, evaluation, and data visualization, and was the primary author of this document. Christiane Löhr and Maureen Larson provided the small RNA sequencing data for cat samples, did a PCR analysis on identified cat microRNAs, and provided the written section explaining the methodology of the PCR experiment. Michelle Kutzler and Massimo Bionaz carried out the tissue collection for cow. Massimo Bionaz provided the small RNA sequencing data for cow samples, did a PCR analysis on identified cow microRNAs and provided the written section explaining the methodology of the PCR experiment performed. David Hendrix directed the research, assisted in developing the methodology, and helped with proofreading and rewriting.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

Page

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF TABLES

1 Introduction

1.1 Biological Background

MicroRNAs (miRs) are a conserved class of small endogenous RNA around 22nt in length. Mature microRNAs are able to modulate a variety of different processes through post-transcriptional gene silencing, which result in either transcript degradation or translational inhibition [1]. MicroRNAs have a wide range of functions including cancer (both tumor-suppressor and oncogenic) [2], development [3], stress response [4], aging [5], and circadian rhythms [6]. Nucleotide positions 2 through 8 on the mature microRNA are called the seed sequence and help direct the sequence-specific activity of the RNA-induced silencing complex (RISC), where it binds to a complementary strand on the 3′- UTR of an mRNA transcript.

Biogenesis of these mature microRNAs typically begins with the transcription of a primary miRNA (pri-miRNA) transcript by RNA Polymerase II [7, 8], or in rare cases RNA Polymerase III [9]. A microprocessor complex associates with the hairpin, whereby the action of the component enzyme Drosha produces a double-stranded cleavage that results in the microRNA precursor (pre-miR) leaving a 2-nt overhang on the 3′ end [10, 11]. Exportin-5 associates with the 3′ overhang and transports the precursor from the nucleolus to the cytoplasm [12]. Here an enzyme known as Dicer produces a second double-stranded cleavage to remove the hairpin loop. Taken together, the activity of these enzymes results in four distinct cleavages of the pri-miRNA transcript and therefore a double-stranded RNA duplex. Dicer passes this RNA duplex to Argonaut, a core enzyme of RISC, which binds with only one of the strands while the other one is degraded. The RISC-bound strand is called the miR, and is involved in the core regulatory function of the microRNA, by directing RISC to complementary target sites in mRNAs. The second degraded strand is called the miR*.

1.2 Early microRNA discovery tools

An early method of microRNA discovery used lower-throughput sequencing technology (e.g. Sanger sequencing) to detect novel microRNAs [13]. However, many other types of small RNA would be detected in the same cloning library. For this reason, criteria such as sequence conservation, presence of the sequence on the stem of a hairpin fold, and the accumulation of the microRNA precursor in organisms with reduced Dicer function were used to determine if the sequence was a valid microRNA [14]. One of the major issues with low-throughput sequencing technologies was that highly expressed sequences would dominate the cloning library making it difficult to detect microRNAs with low abundance [15].

Other early software-based approaches would scan the genome for conserved hairpin forming sequences [17, 18]. miRSeeker employs this method by predicting based on features derived from patterns of stem-loop conservation within microRNAs [19]. miRScan is another conservation-based approach which slides a 21-nt window along hairpin precursors scoring them on sequence conservation and base-pairing potential [20]. However, due to the reliance of these programs on sequence conservation they would only identify homologous microRNAs [18] and may not be applicable to genomes for species with few other closely related species with sequenced genomes.

1.3 Machine Learning and Feature-based Prediction

The advent of high-throughput "deep" sequencing has provided a valuable source of feature information because these sequenced reads can be processed to provide evidence for or against Dicer and Drosha processing at a large scale. Deep sequencing technology has addressed the issue of detecting microRNAs with low expression, but is has also improved the detection of non-microRNA transcripts making software based methods of detecting them more important [15, 16]. This technique involves mapping sequenced reads to the genome, aggregating them into read stacks, and then folding the genomic sequence so their alignment to the fold can be analyzed. Boundaries of the read stacks can be used to identify cuts made by Dicer and Drosha and derived features can be used to separate microRNAs from other sequences. Dicer and Drosha tend to make clean cuts along the stem so read stacks extending from the stem onto the loop or read stacks overlapping other read stacks can be taken as evidence that the precursor is not a microRNA. Dicer and Drosha will usually cut the precursor in such a way that miR and miR* are offset by about 2 nucleotides. Read stacks representing microRNAs also tend to have a low heterogeneity on their 5′ ends because the seed sequence is important for targeting specific mRNAs.

High-throughput sequencing has allowed for large-scale analysis of small RNA sequencing data and ushered a new era of microRNA analysis. Many of the subsequent computational microRNA discovery tools created, such as miRTRAP [21], miReNA [22], miRDeep [16], miRDeep2 [23], miReap [24], and miRAnalyzer [25], use feature-based selection, score-based selection, or machine learning approaches to classify loci as microRNAs, and therefore rely heavily on feature engineering. Although these tools benefit from features which are easily interpretable, feature engineering can be laborious and relies heavily on expertise.

1.4 Deep learning approaches to microRNA analysis

Deep learning is a type of machine learning that uses multi-layered networks to learn relationships present in more basic input data. These approaches overcome the need for feature engineering by learning the features themselves. Two commonly used forms of deep learning methods are convolutional neural networks (CNNs) and recurrent neural networks (RNNs.) Convolutional neural networks work by learning a set of filters which perform convolution operations on the input and pass it into the next layer. RNNs are commonly used on applications such as natural language processing. RNNs iterate over sequential data, calculate their next state from their previous state and the input token, and output values at each position. In this way, output of the RNN is influenced by patterns appearing earlier in the sequence. However, ordinary RNNs suffer from the vanishing gradient problem during training [26], so variants such as Long-Short Term Memory (LSTM) and gated recurrent networks (GRU) are typically used in practice.

Although deep learning is relatively new, both convolutional neural networks (CNNs) [27] and recurrent neural networks (RNNs) [28, 29] have been used for microRNA classification. While these approaches have addressed the limitations of feature engineering, they only predict loci and don't perform site-prediction. They could potentially be used to predict other aspects of microRNA biogenesis. For instance, RNNs have been used in natural language processing applications such as named entity recognition [30] and part-of-speech tagging [31], which are similar tasks to cleavage site recognition.

1.5 Remaining Challenges to microRNA Discovery

There are still many challenges in microRNA discovery tool development. This thesis covers two software-based approaches designed to address these challenges: miRWoods and DeepMirCut.

A persistent challenge to microRNA discovery is the correct identification of precursor span to fold after mapping reads to the genome. In many cases it is possible to fold the precursor in the wrong direction relative to a microRNA. Sometimes even minor adjustments to the start and stop position can influence the fold. Detection of microRNAs with low read abundance is an additional challenge. Mapping low abundant reads to the genome usually results in a huge number of unlikely candidates that need to be folded and analyzed. Many programs avoid this problem by setting arbitrarily high cutoff and fail to detect microRNAs with very low expression.

Chapter 2 presents miRWoods, a stacked random forest strategy to predict microRNAs, both novel and known. It addresses the sequence folding problem using an approach which considers several possible folding patterns, including one that only considers RNA duplex formation without the loop, and uses the one with the highest score. MiRWoods filters down the number of candidate loci using a random forest rather than by applying an abundance cut-off score. miRWoods applies a cut-off score after this point but it is adjusted based on the size of the set and in many of our sets it still allows for 1 read while other programs applied a more stringent cut-off threshold.

Another remaining challenge for the computational analysis of microRNAs is to move beyond just prediction or classification of microRNA loci. For example, more work is needed in the characterization of positions involved in the biogenesis of microRNAs. More specifically, the accurate prediction of the positions along the fold where Dicer and Drosha make their cuts is an important remaining challenge. Some SVM-based Dicer cut-site prediction tools been developed

[32, 33]. However, no attempts have been made to identify cut-sites made by Drosha using deep learning approaches such as recurrent neural networks.

Chapter 3 presents DeepMirCut, a LSTM-based deep learning approach to identifying Dicer and Drosha cleavage sites. While other programs have use precursor sequences downloaded directly from the annotation, we've built our dataset using genomic coordinates found in miRBase to gather precursors along with a region of sequence flanking them on each side. This additional sequence is necessary for training and testing on Drosha cut-sites. We also apply random amounts of flanking region to improve training and to provide a more stringent test of our model's performance.

miRWoods: Enhanced precursor detection and stacked random forests for the sensitive detection of microRNAs

Jimmy Bell, Maureen Larson, Michelle Kutzler,
Massimo Bionaz, Christiane V. Löhr, David Hendrix

Plos Computational Biology
https://journals.plos.org/ploscompbiol/
15(10)

2 miRWoods: Enhanced precursor detection and stacked random forests for the sensitive detection of microRNAs

2.1 Abstract

MicroRNAs are conserved, endogenous small RNAs with critical post-transcriptional regulatory functions throughout eukaryota, including prominent roles in development and disease. Despite much effort, microRNA annotations still contain errors and are incomplete due especially to challenges related to identifying valid miRs that have small numbers of reads, to properly locating hairpin precursors and to balancing precision and recall. Here, we present miRWoods, which solves these challenges using a duplex-focused precursor detection method and stacked random forests with specialized layers to detect mature and precursor microRNAs and has been tuned to optimize the harmonic mean of precision and recall. We trained and tuned our discovery pipeline on data sets from the well-annotated human genome and evaluated its performance on data from mouse. Compared to existing approaches, miRWoods better identifies precursor spans, and can balance sensitivity and specificity for an overall greater prediction accuracy, recalling an average of 10% more annotated microRNAs, and correctly predicts substantially more microRNAs with only one read. We apply this method to the under-annotated genomes of *Felis catus* (domestic cat) and *Bos taurus* (cow). We identified hundreds of novel microRNAs in small RNA sequencing data sets from muscle and skin from cat, from 10 tissues from cow and also from human and mouse cells. Our novel predictions include a microRNA in an intron of tyrosine kinase 2 (TYK2) that is present in both cat and cow, as well as a family of mirtrons with two instances in the human genome. Our predictions support a more expanded miR-2284 family in the bovine genome, a larger mir-548 family in the human genome, and a larger let-7 family in the feline genome.

2.2 Introduction

MicroRNAs (miRNAs, miRs) are a highly-conserved class of small endogenous RNA molecules that are involved in post-transcriptional gene silencing by acting as a guide RNA for the RNA-induced silencing complex (RISC). The biogenesis of microRNAs begins with the generation of a primary transcript (pre-miR), which folds into a structure containing one or more ~70-nt hairpins. These hairpin precursors (pre-miRs) are cut at the base by Drosha [10]. After export from the nucleus, the loop of the hairpin is cut by Dicer. The resultant double-stranded RNA duplex is unwound to produce two mature ~22-nt microRNAs (miRs), named 5′ and 3′ after the arm of the hairpin from which they derive. Typically, only one of the mature microRNAs is incorporated into RISC, and the other microRNA is degraded and designated miR-star or miR*. The seed sequence at positions 2-8 of RISC-bound mature microRNAs binds to complementary sequences in the 3′ untranslated regions (UTRs) of mRNAs.

The advent of deep sequencing data has enabled the high-throughput discovery and annotation of novel microRNAs. Most microRNA prediction approaches begin by aligning size-selected deep sequenced RNA (small RNA-seq) reads to the genome, and then the identification of overlapping aligned reads, “read stacks”. These read stacks correspond to mature microRNA products, as well as other sequenced fragments including microRNA offset RNAs (moRs) [34], hairpin loops, and spurious RNA fragments. The RNA secondary structures for the genomic sequences surrounding the read stacks are predicted and reads overlapping predicted hairpin structures are analyzed for arrangements consistent with microRNA processing. The prediction methods vary in the specifics of how the data are processed, and relevant features are quantified, as well as what classification techniques are used. Methods employing this strategy include miRTRAP [21], the software upon which miRWoods was built, along with miRDeep [16], the improved miRDeep2 [35] and other variants [36, 37], miReap [24], and miRAnalyzer [25].

Several challenges remain in the computational prediction of microRNAs. Current approaches have strengths and weaknesses; while some approaches focus on higher precision at the expense of false negatives, others focus on higher recall at the expense of false positives. Most approaches require a minimum number of mapped reads at a given locus, meaning that many valid lowly expressed microRNAs are missed. Also, hairpin precursor detection is challenging because slight changes in the boundaries can shift the secondary structure prediction away from the hairpin. Our analysis of the predictions from available methods identifies many cases that partially overlap with or are shifted from annotated loci, and mistake 5′ for 3′ mature miRs.

These remaining challenges to microRNA discovery motivated us to create miRWoods, a microRNA discovery pipeline using stacked random forests with an improved method for determining hairpin precursor span (Figure 2.1). The miRWoods pipeline consists of a mature product random forest (MPRF) for mature product detection, and a hairpin precursor random forest (HPRF) for hairpin precursor identification. For balancing precision *versus* recall, we tuned miRWoods to optimize F-score, which is the harmonic mean of precision and recall. We trained and tuned miRWoods on well-annotated human data sets, evaluated cross-species performance using mouse data and used the pipeline to subsequently identify novel microRNAs in the feline and bovine genomes.

2.3 Results

2.3.1 Overview of strategy behind miRWoods

Because current approaches impose a threshold for the read abundance for a locus to be evaluated as a putative microRNA, many low-abundant miRs are missed. To avoid this, we have added a machine learning classifier to identify read stacks that are plausible mature microRNA loci, thereby enabling miRWoods to detect microRNAs with a single read. This RF evaluates read abundance-related features in the context of other features to classify plausible mature products (Table 2.1). To avoid the sensitive-dependence on precursor span for secondary structure prediction, we examine several putative precursors for each read stack, including one derived from the boundaries of the optimal duplex between the read stack and surrounding genomic region (duplex-focused spans) and those derived from the boundaries with other products (product-focused spans). Through extensive feature-engineering, we have added several novel features to help classify the microRNA precursors, which are listed in Table 2.2. Finally, we have tuned parameters of our model to optimize F-score, the harmonic mean of precision and recall, to result in improved performance that doesn’t sacrifice precision, and recalls 10% more annotated microRNAs on average.

2.3.2 Stacked random forest approach

As with other microRNA discovery tools, miRWoods begins by analyzing genomic loci where small RNA reads mapped. A distinguishing feature of miRWoods is the use of an additional RF layer (the MRPF) to classify reads stacks as plausible mature microRNAs rather than rely only on the number of reads mapping to that genomic locus. The MPRF is trained on a balanced set of positive and negative examples, whereas the HPRF is trained with a much larger negative set. This results in the MPRF being a more lenient predictor allowing the HPRF to make a more stringent prediction later on. The features used in the MRPF are summarized in Table 2.1. The MRPF also leverages basic sequence features previously shown to be effective in detecting precursors such as GC-content and dinucleotide frequencies [22, 25, 38-40]. In addition, we introduce some novel features such as the duplex energy between the read stack's most frequent read sequence and the surrounding genomic locus. This quantity is distinct from miR:miR* duplex energy because the input is a read stack, and miR/miR* designations have not been assigned at this point. Also included are the observed frequencies of 5′ read ends relative to the most abundant position.

The HRPF also uses several novel features, summarized in Table 2.2. Novel features include 11 "overlap" features, corresponding to the degree of overlap between different identified products (e.g. 5′ moR, 5′ miR, loop, 3′ miR). We also introduced several features describing destabilizing structures, such as bulges and loops, and several features describing the regions duplexed with the most abundant product. We also analyzed what features were most important for miRWoods, and summarized feature importance in Figure 2.2. We found that the frequency of reads in the start position of the read stack and the duplex energy to be highest in importance for the MPRF (Figure 2.2a). We found that the decision value from the MPRF, the reads per million in the sense and anti-sense strands, the product base pairing, and the duplex energy to be the most important features for the HPRF (Figure 2.2b). Because the value of some features showed correlation, we also examined feature importance for RFs trained with correlated features removed. We identified features with an R^2 greater than or equal to 0.5 and removed the feature with the highest importance for each correlated pair. We saw an increase in feature importance for some features in the HRPF, such as totalSenseRPM, dupLoopDistance, ARV, wARV, dupPBP, and afh (Figure 2.3). We also examined the change in importance when correlated features are removed (Figure 2.4a-b). In some cases, features gained the importance after removal of their correlated partner. In other cases, such as “dupLoopDistance” and “dupPBP”, features showed a substantial increase in importance despite not having correlated features removed. We did not observe a significant decrease or consistent change in performance with the most correlated features removed (Figure 2.4c).

We examined the role of read-abundance on the performance of miRWoods. The histograms of true positive predictions from miRWoods, miRDeep2, and miRWoods demonstrate that miRWoods correctly identifies more single-read miRs (Figure 2.5a-f). We observed that consistently in both predictions trained and tested in human (same-species) and trained on human, tested on mouse (cross-species), miRWoods consistently makes more valid positive predictions for loci supported by only one read (Figure 2.5g). While these predictions illustrate the power of miRWoods, in practice any single-read predictions are not proof and would require further validation. We analyzed the effect of removing read-abundance-related features and found that while performance does reduce with the removal of these features (Figure 2.5h), the overall greater performance on low-abundance loci demonstrates that these features do not impair performance.

2.3.3 Accurate mapping of hairpin precursor span

Proper identification of hairpin precursor span is critical for microRNA prediction, because methods typically rely on secondary structure prediction, which can significantly depend on the defined window. The labeling of 5′ vs 3′ products requires accurate identification of the hairpin precursor. We imposed stringent requirements for predictions for the hairpin span of a locus to be considered a true positive when compared to miRBase annotations. Predicted loci where the hairpin folded in the wrong direction and/or overlapped less than 50% of the annotation were counted as false predictions. To address these stringent criteria, we developed an approach that focuses on strong miR/miR* duplex energy, rather than secondary structure of the hairpin. While most approaches focus on the predicted structure in a region around the most abundant product (i.e. major product), our duplex-focused method selects the span of hairpin regions using the optimal duplex pairing with the most abundant product (Figure 2.6a). Alternatively, product-focused spans covering the major product and any product 4 nt or more away from the major product are also considered. Each of these options are considered as putative loci, and evaluated in subsequent steps. We found that miRWoods uses the duplex-focused span an average of 88.4% of the time in its final predictions for all human sets (Table 2.3).

The percentage of predictions that matched an annotation well enough to be considered a valid hairpin precursor was computed for miRWoods, miRDeep2 and miReap and summarized in Table 2.4. miRWoods predictions used the proper fold an average of 99.1% of the time for human samples and 99.9% of the time for mouse samples. miReap was able to predict the proper fold 98.2% of the time for human and 97.8% of the time for mouse. miRDeep2 was able to predict the proper fold 98.9% of the time for Human and 97.7% of the time for mouse. In some examples, miRWoods corrects errors in the miRBase annotations. In Figure 2.6b we show the current annotation for hsa-mir-4721. While miRWoods predicts a hairpin precursor that directly matches with intron splice junctions (a mirtron), the miRBase annotation only overlaps one mature product. Similarly, Figure 2.7a shows hsa-mir-6860, which miRWoods predicts to be a half-mirtron and the current miRBase annotation does not. In both cases the miRWoods predicted hairpin span lines up with the intron splice site, even though miRWoods does not use splice junction locations in its predictions, thereby providing independent support to the predictions. In other examples, such as mmu-let-7c-2, the miRDeep2 hairpin span is offset, assigning the 5′ product as the 3′ product (Figure 2.6c). A similar scenario is observed for hsa-mir-431 (Figure 2.7b).

2.3.4 Evaluation of prediction performance

The repertoire of expressed microRNAs can vary considerably between tissue types in the same organism; therefore, we tested miRWoods against different cell types and conditions. We tested miRWoods on 9 samples from 4 small RNA sequencing experiments and provide performance metrics compared to other methods in Table 2.5. We compared the performance of miRWoods, miRDeep2, and miReap on several small RNA data sets from human and mouse downloaded from GEO [41]. In each evaluation, the same RF models trained on human data were tested on small RNA data collected from different tissues including human MCF-7 total cell content (GSE31069), MCF-7 cytoplasmic fractions (GSE31069), human cancer cell lines (GSE16579), human normal liver (GSE21279), as well as cross-species tests on mouse brain, embryo, testes, ovary, and whole newborns (GSE20384). Because microRNA expression can vary from tissue to tissue, all programs were evaluated against the expressed miRs for that data set with at least one read aligned. miRWoods recalled on average 10% more annotated miRs, and obtained greater F-scores except in the case of mouse embryo where the F-sore was 0.624 for miRWoods compared to 0.626 for miRDeep2. Higher F-scores were obtained for all sets when miRWoods was compared with miREAP.

Remarkably, miRWoods performed better on cross-species tests on mouse data compared to tests on human data (Figure 2.8), providing justification for its application to other mammalian genomes when trained on human. Typically, miRWoods has a greater number of false positives and fewer false negatives than miRDeep2 when compared to miRBase annotations (Figure 2.9a).

We tuned thresholds for expression level, proportion of negative samples, and decision values threshold on a separate dataset from what RFs were trained on (see Methods, Figure 2.10). A summary of the data sets and values resulting from the tuning experiment is provided in Table 2.6.

The decision value threshold that has been tuned to optimize the F-score for the identification of valid loci correlates well with decreased expression in Dicer knockdown MCF-7 cells (Figure 2.9b, Figure 2.11a-b Fig). On average, the novel predictions of miRWoods show a greater decrease in the cytoplasm of Dicer knockdowns compared to novel predictions from miRDeep2 and miREAP (Figure 2.9c) and on par in total cell content (Figure 2.11b). We calculated p-values for each of these comparisons using two-sample t-tests and found that novel predictions in cytoplasm from miRWoods and miREAP had a significant reduction in Dicer-knockdown expression compared to miRBase. Novel predictions in total cell content for all programs showed a significant reduction in Dicer-knockdown expression compared to miRBase (Table 2.7). Similarly, empirical cumulative distribution functions (ECDFs) of the fold change in Dicer knockdowns compared to wild type show a greater proportion of novel predictions highly depleted in Dicer knockdowns (Figure 2.11c,d). Examples of novel predictions found to be reduced in expression in Dicer mutants include hsa-Novel35, hsa-Novel28, hsa-Novel23, hsa-Novel65, has-Novel92, and hsa-Novel99 (Figure 2.12).

One advantage of miRWoods over the other methods is that it prints a score for each genomic locus evaluated, whether or not it is predicted to be a microRNA. Therefore, the output is amenable to creating precision-recall (PR) curves [42], such as Figure 2.9d and Figure 2.9f. The area under the PR curve (AURPC) evaluates the performance of the prediction, and has the advantage over Receiver Operator Characteristic (ROC) curves [43] of not being overwhelmed by the large number of true negatives associated with genome-wide microRNA prediction. We present PR curves for predictions in mouse, with an average AUPRC of 70.3. Comparisons with miRDeep2 show that miRWoods has greater false positives, but fewer false negatives (Table 2.5, Figure 2.9e-g, Figure 2.13). Comparisons with miREAP show that miRWoods has a lower false positive rate, and a higher F-score on average (Table 2.5, Figure 2.14). Overall, miRWoods shows equal or greater F-score than both miRDeep2 and miReap for all data sets (Table 2.5).

Many of the “false positive” microRNA predictions are actually novel predictions of valid miRs. Despite how complete the human microRNA annotation is, we were able to identify 682 potential novel loci in the human data sets. We found that many of our novel predictions, despite being unannotated, had homology to known miRs in other species. In some cases, miRWoods identified more instances of known miR families. For example, there are 72 known precursors from the mir-548 family in the human genome annotated by miRBase. miRWoods was able to identify an additional 34 novel members of the mir-548 family (Figure 2.15), suggesting this family could be larger than previously thought.

2.3.5 Novel microRNA predictions in the feline genome

We next sought to predict microRNA loci in species with limited microRNA annotations, including the feline and bovine genomes. We ran miRWoods on small RNA samples isolated from muscle and skin tissue for 3 different cats. Currently, there are two studies of feline microRNAs that we are aware of. In one study, Sun *et al.* did an analysis with miREAP in the context of the mink enteritis virus (MEV) [44]. In a more recent study, Laganà *et al.* identified feline microRNAs with miRDeep2 in a multi-tissue cohort [45]. miRWoods identified 495 microRNA loci, with 293 of them having significant homology to microRNA precursors from miRBase. Among the miRWoods predictions, 198 overlapped with the microRNA found in Sun *et al.*, and 213 overlapped with microRNAs found in Laganà *et al.*, and 215 were newly discovered (Figure 2.16a).

Expression of three novel microRNAs in feline skin and muscle were examined by qPCR and normalized expression relative to 2 control miRs with low variability across our tissue samples, miR-25 and miR-191 (Figure 2.16b,d, Figure 2.17). These examples included a novel member of the miR-133 family, with enriched expression in muscle that was validated by qPCR (Figure 2.16b) and a predicted structure that strongly matches expectations for microRNAs (Figure 2.16c). We also identified a novel miR with no homology to known miRs, with a statistically significant tissue-specific enrichment based on a voom analysis [46], including some more abundant in muscle (Figure 2.16b-e). In addition, we validated two predicted miRs previously described by Laganà *et al* that we determined to be significantly differentially expressed. As predicted, fca-mir-1-1 was more abundant in muscle whereas fca-mir-205 was abundant in skin (Figure 2.17). Overall, our analysis of the expression of our predicted microRNAs identified 71 differentially expressed miRs using a voom FDR of 0.05, with 33 enriched in muscle, and 38 enriched in skin tissue.

Several known and novel let-7 family precursors were found within clusters including multiple let-7 miRs. For example, we found a cluster on chromosome D4 containing fca-let-7f and two novel let-7 miRs denoted fca-let-7-Novel2 and fca-let-7-Novel3 (Figure 2.18a). The predicted novel miRs (Figure 2.18b-c) have predicted secondary structures with similar bulges and/or internal loops observed in other let-7 family members including fca-let7f (Figure 2.18d). A phylogenetic tree of known and novel let-7 miRs shows comparable sequence similarity, although not necessarily correlated with proximity of the genomic loci (Figure 2.18e).

Feline microRNAs were found within 51 clusters, 28 overlapped with the 31 previously described [45]**.** miRWoods identified a novel precursor (fca-Novel45) near chr-X-38640 and chr-X-38642, which were two previously identified novel microRNA that may be associated with testis development and physiology [45]. Two additional feline-specific miRs (fca-Novel10 and fca-

Novel13) where found on a cluster within the ARHGEF10L gene. miRWoods also identified two mir-30 homologs near fca-mir-30c-1 within an intron on the NFYC gene.

2.3.6 Novel microRNA predictions in the bovine genome

For the bovine genome, there are 811 known microRNA precursors producing 881 mature microRNA annotations, compared to 1187 precursors for mouse and 1881 for human, which generate 2045 and 2813 mature products, respectively.

We used miRWoods to predict bovine microRNA loci using small RNA-seq samples from 10 bovine tissues including corium from the hoof (corium feet), dental pulp, oral papillae, penis, retina, iris, optic nerve, brain stem, bone marrow, and submandibular lymph node. We selected tissues that were highly diverse and whose microRNA profiles had not been examined before. Our pipeline identified a set of 810 predicted microRNA loci. Among these, 409 were already in the miRBase R21 *Bos taurus* annotations, 91 had homology to microRNA annotations in cow and other species, and 310 were novel predictions with no known homology.

Overall, miRWoods identified 401 novel bovine microRNAs. In addition, clustering of microRNA loci revealed 76 clusters, including 63 known and 13 novel clusters. Two bovine-specific half-mirtrons, (bta-Novel68 and bta-Novel71), were found within the *PLD2* gene. A bovine specific mirtron (bta-Novel210) and another half-mirtron (bta-Novel212) were found within the *MCAM* gene. Another bovine specific half-mirtron (bta-Novel208) was found on a cluster with bta-mir-140 on the *WWP2* gene.

Figure 2.19a shows an Euler diagram comparing miRBase annotations to miRWoods predictions for bovine samples. To test the validity of the novel predictions, we performed RT-qPCR on available samples, and normalized expression relative to 5 control miRs with low variability across our tissue samples. After normalization, expression levels for control miR-7 are compared with RT-qPCR (Figure 2.19b). Strong correspondence between small RNA-seq and RT-qPCR are observed for 2 of the tested microRNAs (Figure 2.19c,d), suggesting that the mature product was detectable with both methods in the tissue it was expressed. Expression was observed for all tested novel bovine miRs using RT-qPCR, validating the expression of these predicted mature products (Figure 2.19e).

2.3.7 Novel Predictions in the Bovine miR-2284 Precursor Family

The miR-2284 family has previously been found to be expressed in tissues relevant to the immune system but gene targets are currently unknown [47]. Within the mir-2284 family, miRWoods predicted 29 known and 68 homologous precursors. Of the 68 homologous precursors, only 35 fit the criteria of having the same seed region as other miRs. Removing the seed requirement, 33 additional mir-2284 family precursors were identified. Unique reads were found in 51.5% of homologous precursors and 37.00% of already annotated precursors. Hierarchical clustering was performed on mir-2284 family microRNAs based on their normalized expression profiles and a heat map was generated (Figure 2.20a). Despite having a shared homology, expression of the microRNAs in the mir-2284 family are highly diverse in the tissues assayed, but show greatest expression in submandibular lymph node (SLN). Interestingly, this is consistent with prior studies of this family that observe greatest expression in bovine immune cells [48] given recent studies of the immunosuppressive properties of SLNs [49]. A phylogenetic tree was created to show all annotated and newly predicted miRs in the mir-2284 family (Figure 2.20b).

Read abundances showed a tendency for mir-2284 and mir-2285 precursors to favor a single (opposite) side of the precursor. The abundance of each microRNA for the mir-2284 and mir-2285 precursors within the mir-2284 precursor family was examined (Figure 2.20c) For annotated microRNA precursors, 82.05% of mir-2284 loci had the most abundant read on the 5p-side, and 89.36% of mir-2285 had the most abundant read on the 3p-side. Similarly, for our predicted microRNAs with homology to this family, 73.91% of mir-2284 examples had the most abundant read on the 5p-side, and 88.89% of mir-2285 examples had the most abundant read on the 3p-side.

2.3.8 Discovery of novel miR families

We found that 11 novel predictions in human were within clusters of annotated microRNAs, and 39 novel predictions in new clusters. Of the 9 potential miR families which matched the criteria found in the methods section one contained a snoRNA and was removed. Of the remaining candidate miR families two had miRs which were found across species. We identified a novel miR family with two instances in the human genome; one example was a mirtron in an intron of *LAMA5*, and the other a half-mirtron in an intron of *CHD3* (Figure 2.21a-e). Both of the examples in human were observed to have no expression in Dicer knock-out cells (Figure 2.21a,b). We observed a strong level of similarity in predicted secondary structures of the two examples observed in human (Figure 2.21c,d). We compared these introns across several mammalian species and observed patterns of conservation that suggest an ancestral divergence of these two mirtrons rather than a more recent duplication (Figure 2.21e). We found another novel miR family with an example in both the bovine and feline genomes, but not observed in mouse or human (Figure 2.21f-g). Strikingly, both of these miRs (bta-Novel5 and fca-Novel70) were found within the same intron of *TYK2* in cow and cat genomes (Figure 2.21f,g), and both examples showed nearly identical hairpin precursor sequences (Figure 2.21h,i). We did not observe this miR in human or mouse data sets, and we also observed greater sequence divergence of this intron in human and mouse (Figure 2.21j).

2.4 Discussion

Our study demonstrates that despite a long history of microRNA discovery tools and annotations, there is still room for improvement. Despite the maturity of microRNA annotations for the human genome, our approach was still able to find novel human miRs. We have identified several miRs with annotated positions shifted from the correct location, and that have been resolved with miRWoods.

The inclusion of the duplex-focused method in miRWoods improved hairpin precursor span identification over the other programs. Not only did miRWoods match the miRBase hairpin precursor annotation more often, in some instances miRWoods predictions corrected the miRBase annotation. Splice junction boundaries for the mirtron and half-mirtron examples provide evidence for the validity of the miRWoods duplex method because the optimal precursor span closely corresponds to the splice junctions, as expected given mirtron biogenesis mechanisms [25] despite the fact that these hairpin boundaries were computed without the use of intron annotations. Similarly, Boruta feature-importance analysis showed that the duplex energy was more important than the minimum free energy of the hairpin. These observations support the idea that the thermodynamic stability of intermediate RNA duplex formed by miR and miR* may serve important roles in microRNA function, consistent with previous studies showing this affects efficient loading into Argonaut [50]. We also found that the distance between the miR* sequence and the loop have a greater importance than that of the major product and the loop. Future work is needed to determine the relative importance of stable mature miR:miR* duplex formation compared to stable stem-loop formation in microRNA biogenesis.

We demonstrated in this study that miRWoods is capable of correctly identifying microRNA loci with only one read more than other programs. Although this displays the strength the miRWoods approach, in practice users should seek further evidence to support the validity of any novel miRs only supported by one read.

Predictions from miRWoods consist of 215 potential novel microRNA annotations for cat and 417 novel candidates for cow. These findings support the expectation that these organisms have comparable number of microRNAs to human and mouse but are currently less-well annotated due to greater research focus on human and mouse. Future work could expand miR annotation in feline and bovine further by sequencing other tissues, as well as identifying regulatory targets for miRs in specific tissues.

Finally, our approach is able to identify more examples of known families, suggesting that they are larger than previously thought. While these large families retain sequence similarity at the hairpin-level, they are often the result of seed shifting and mismatches, suggesting a wide range of potential gene targets. Predictions using miRWoods showed an expansion in the number of microRNAs within the mir-548 family in human, and the mir-2284 family in the bovine genome. These families are often defined in terms of homology to the hairpin sequence rather than the seed

[51]. We observed several mutations within the seed region of mir-2284 family miRs that result in the complex phylogeny, and which indicate that a much wider range of genes may be targeted by this family than currently accepted. The fact that we observed miR-2284 family members to be differentially expressed across diverse tissue types supports the idea that this family expanded and sub-functionalized in various tissues. As noted previously, the widespread genomic distribution of the primate-specific mir-548 family supports the hypothesis that it may have been evolutionarily derived from transposable elements [52]. Similarly, mir-2284 family may be more expansive than previously thought, and the observed diversity of sequence and expression supports the hypothesis that this family has shaped ruminant evolution [51].

2.5 Methods

2.5.1 Ethics Statement

All bovine tissues were harvested from animals that were already scheduled to be slaughtered, and collected immediately after slaughter. All slaughter operations were performed under USDA-FSIS supervision in accordance with the Humane Slaughter Act (1978), the Federal Meat Inspection Act (1906), and using a percussive captive bolt stunner. The feline tissue samples were obtained through the biobank at the Carlson College of Veterinary Medicine at Oregon State University. Tissues had been banked for research purposes with owner consent and approval of the institutional animal care and use committee.

2.5.2 Tissue samples small RNA sequencing

We examined small RNA samples collected from 10 bovine tissues including submandibular lymph node (SLN), bone marrow, brain stem, optic nerve, retina and iris of the eye, penis (corpus cavernosum), oral papillae (buccal mucosa), dental pulp, and hoof corium (corium feet) from three Angus steers collected just after slaughter at the Meat Science laboratory at Oregon State University. The feline tissue samples were obtained through the biobank at the Carlson College of Veterinary Medicine at Oregon State University and included normal haired skin and normal skeletal muscle from three male neutered domestic short hair cats aged 10–13 years. Tissues had been banked for research purposes with owner consent and approval of the institutional animal care and use committee. RNA was isolated from tissue by chloroform-isopropanol extraction. RNA quality was analyzed on a Bioanalyzer 2100 Nano chip (Agilent Technologies, Santa Clara, CA), with a minimum acceptable RIN of 7. Small RNA sequencing was performed at the Center for Genome Research and Biocomputing (CGRB) at Oregon State University (OSU). Libraries were prepared using the Illumina TruSeq small RNA sample preparation kit (Illumina, San Diego, CA) for library preparation and size-separation by polyacrylamide gel electrophorese. Library size was determined with the Bioanalyzer 2100 HS-DNA chip and the KAPA biosystem's library quantification kit, and libraries normalized to 2 nM. Multiplexed samples (6/lane) were sequenced with a 50 cycle v3 sequencing kit on an Illumina HiSeq 3000 sequencer.

2.5.3 The miRWoods pipeline

The miRWoods pipeline consists of two random forests with readily interpretable, biochemically-motivated features. The pipeline's two layers correspond to classifiers that recognize different components of the microRNA (Figure 2.1). The first random forest layer predicts likely mature miRNAs products. In this way, the first random forest acts to filter out a large number of loci before precursors are considered, thereby improving accuracy and reducing the overall runtime. The second random forest layer scores the precursors around the predicted mature miRNAs and is used to generate the final set of predictions.

The miRWoods pipeline is perl software largely derived from miRTRAP [21], but with significant improvements on speed and memory efficiency, as well as two random forest layers rather than user-defined thresholds. The pipeline now includes the integration of indexed bam files for faster read processing, and the RNAfold perl module for rapid secondary structure prediction. The processing of sequencing data begins with one or more small RNA-seq FASTQ files. We trim the reads using cutadapt, a software designed to remove the 3′ adapter sequences [53]. We set cutadapt's options to remove sequence from the 3′ end with a PHRED score of 10 or less prior to trimming, allow for a 20% error in the adapter sequence during trimming, and remove sequences less than 17 in length after trimming. After trimming we filter the reads further using a custom script that removes reads with an average PHRED score of 30 or more. Sequencing data is mapped to the genome using bowtie [54]. Bowtie's option are set to return the best mappings that have one or fewer mismatches with a seed length of 18 nucleotides, an error of 50 after the seed length, and occur within 10 or fewer places in the genome. Before sorting and indexing the bam files, we add additional tags according to samtools specifications describing the number of hits (NH-tags) to the genome for each read[55], which is used later in the pipeline to normalize expression.

2.5.4 Mature product random forest

The miRWoods pipeline consists of several data-processing steps. Next, after read alignment, we identify "read regions", which consist of reads that map to overlapping positions in the genome. Each of these read regions are evaluated as putative mature microRNA products based on a number of features calculated from genomic loci and read distributions. Basic features for each read region are computed, such as GC-content, dinucleotide frequencies, Wootton-Federhen sequence complexity [56], and the median length of reads mapped to the locus. Because the function of microRNAs involves the position of seed sequences relative to the 5′ end, the 5′-heterogeneity is computed for each read region as previously described [57]. In addition, we compute the number of reads mapping to positions within a fixed offset from the most abundant product. We also computed the minimum duplex energy between the read stack's most frequent read sequence and the surrounding 70bp region. These and other features are input into our first random forest, called the "mature product random forest" (MPRF), which classifies read regions as mature microRNA products or non-miR loci.

2.5.5 Hairpin precursor span optimization

We found that a major source of error in high-throughput microRNA discovery was the prediction of the span (start and end positions) of genomic location of the hairpin precursor, and therefore we developed a new method of precursor span prediction (Figure 2.6). While most other approaches predict secondary structure of the region surrounding a putative mature product, our approach computes the RNA:RNA duplex energy of the mature products (without the loop). Each putative microRNA product identified by the MPRF is used to compute the optimal duplex energy between the most abundant product and the surrounding 70bp window using RNAduplex [25], as depicted in Figure 2.6a. The region spanning this most abundant read and the optimal duplex subsequence is then used as a putative hairpin precursor sequence. In addition, a second method folds between any two products that are separated by 5 nt or more. Both methods are used and create several secondary structure predictions, all of which are the basis of a putative hairpin precursor to be input to the next random forest. When the hairpins are subsequently evaluated in the next step, overlapping hairpins are dropped and only the predicted hairpin with the highest decision value from that random forest is retained.

2.5.6 Hairpin precursor random forest

The second random forest within miRWoods, called the "hairpin precursor random forest" (HPRF) is used to evaluate the putative hairpin precursors from 71 features, which provide scores based on its sequence, structure, and folding energy. Many of the features for the hairpin phase come from the original miRTRAP software [21].

The features for the HPRF can be categorized as sequence features, structural features, and product-distribution features. Sequence features include dinucleotide frequencies, GC Content, and sequence complexity over the entire precursor sequence. Structural features include the minimum free energy returned by RNAfold [58], and the optimal duplex energy of the most abundant product and hairpin precursor region computed by RNAduplex [59]. The decision value from the MPRF for the most abundant product within putative hairpin precursors is also included as a feature.

Expression levels for a locus L are quantified with adjusted reads per million (ARPM), which are defined by

$$ARPM(L) = \frac{10^6 \sum_{r \epsilon L} 1/n_r}{\sum_{r \epsilon S} 1/n_r}$$

Total read counts, separately computed for the sense and antisense strands of the precursor, are first adjusted, meaning when a read r aligns to n_r locations in the genome, the read contributes a fractional count of $1/n_r$ to each location, essentially uniformly distributing the count to each locus [60]. These calculated values are then normalized for each sample S to parts-per-million. The product-distribution features are computed by first naming read stacks as the products that would be expected in the event of Dicer and Drosha cuts by a previously defined algorithm [21]. A number of features describe the abundance and mapping of reads for each of these products. The unique read fraction describes the proportion of reads mapping only to the locus. Various features, such as the $5'$ heterogeneity, and average hit count were evaluated for the most abundant mature product. For each of the mature products, several features describe the relative frequency of reads for miRs, moRs, loop products, and other products within the precursor. Several other features were created to describe the variance and weighted variance of reads associated with mature products relative to the most frequent cut variant and to the hairpin.

Dicer and Drosha tend to make precise cuts to produce well-defined $5'$ ends of the mature products for proper functionality. Because of this, several features describe the amount of overlap across all products and across each product relative to its surrounding products. In addition, reads within a product would not be expected to be significantly offset from the product on the opposite arm of the hairpin, or relative to any moR products on the same arm. Therefore, features measuring the amount of shift between miR products are included.

A number of features were generated to describe the structure of the predicted hairpin. Two features, base pair density (fraction of paired nucleotides in predicted structure) within the major product, and base pair density within the optimal duplexed region. These features may be different due to bulges being present on one arm of the hairpin but not the other. Features for the part of the fold around the miR products include the sizes of the largest bulge, size of largest internal loop, size difference between the two halves of internal loops, and overhangs on the major miR product, which are defined as the maximum number of unpaired bases on either end of the miR. The dupLoopLength feature measures the largest region of unbound nucleotides on the duplex across from the most abundant miR Product. A dupSize feature is a measure of the size of the region predicted to duplex with the most abundant product. Since the duplex is expected to be around the same size as the miR product this feature may help exclude cases where there are large unpaired stretches on the duplex or most of the major product is unbound to the duplex. A feature called innerLoopGapCount scores the number of occurrence of spans of 3 or more unpaired nucleotides in the loop region (i.e. more than one indicates a multi-branched loop). This feature may help in situations in which there is a multiloop or where the loop structure is uncommon to known miR precursors. Additionally, a feature measuring the size of the hairpin loop is included. We added new features quantifying the size of the largest bulge in the hairpin structure, which is known to affect Dicer specificity [61].

Because microRNA loci tend to cluster together, we incorporated a neighbor count feature, which is a score tallying the number of neighboring hairpins that occur within 1000 nucleotides of the precursor being analyzed. The neighbor count feature counted all small RNA loci, including both miR and non-miR loci, and reduced the number of observed false positives.

2.5.7 Training, tuning and model selection

The miRWoods pipeline requires models for both MPRF and HPRF layers that have been trained on positive examples, which are annotated microRNAs, and negative examples, which are loci containing read regions not overlapping annotated microRNAs. The training data for the MPRF is produced by a script that collects loci based on the overlap of the products with the mature microRNAs in miRBase annotations, with X-fold more negative examples than positives for some input X. The training data for HPRF is created by using hairpins with the best overlap of the known hairpin annotation.

Our strategy for tuning the thresholds of miRWoods focused on three parameters: the decision value $\hat{y}_{HPRF}$ for the hairpin random forest output, the expression level threshold E_{th} in units of ARPM, and the proportion X of negative loci used in stratified sampling. To determine these thresholds, we trained and tested on different small RNA deep sequencing data sets. We selected four large data sets from sequencing read archives (SRA) from diverse tissues and developmental stages. We trained on one of the data sets, which produced optimal RFs. We then applied this to a second data set and computed F-scores for different $\hat{y}_{HPRF}$, E_{th}, and X parameters, and chose the set of parameters that gave the highest F-score.

Our strategy for training and tuning models was to train with one data set, tune on another, and ultimately select final models were chosen based on the highest F-Score when tested on a test set. Two sets of models were trained using either tonsillar B-cell populations from GSE23090 or human cerebellum, heart, kidney, and testis tissue from GSE40499 (Figure 2.10a). The frontal cortex data was excluded from the GSE40499 set to make read counts in tissues more balanced. Each of the two resulting models were tuned using a grid-search for the $\hat{y}_{HPRF}$, E_{th}, and X parameters to optimize F-score when evaluated on either cancer cells from GSE18381 and GSE20592 or stem cells from GSE65706 and GSE62501; therefore, four tuning experiments were performed, corresponding to the four arrows in Figure 2.10a. Afterwards, models tuned using the cancer cell sets were validated using the stem cell sets and vice versa. The model resulting in the highest F-score from the test set was chosen for all remaining tests. Plots of the F-score as a function of each of the tuned parameters are presented in Figure 2.10b-d. In each training experiment the stratified sampling for the product model was set such that the negative set would be equal in size to the positive set. This was to allow as many products as possible to enter the hairpin phase while still filtering out enough that the resulting folds could be generated in reasonable amount of time.

The model with the highest F-score resulted from training on the set of tonsillar B-cell populations (GSE23090) and tuning on human melanoma cells (GSE18381) and human normal and cancerous cervical cells (GSE20592) when validated against stem cell sets (GSE62501,GSE65706). Tuning through a grid search resulted in an optimum decision value of 0.28, an ARPM of 0.11, and a 1:25 ratio of positive to negative training data used in stratified sampling.

2.5.8 Comparisons with other tools

miRWoods was compared with miRDeep2 and miReap in the prediction of microRNA loci from small RNA sequence data in well-annotated genomes. The data used were MCF-7 cell cytoplasmic and total-cell extract from GSE31069, human cancer cell lines from GSE16579, healthy human liver samples from GSE21279, and mouse brain, embryo, newborn, testes, and ovary from GSE20384. For miRDeep2 the FASTQ files were combined and the program was run with the same settings as previously published [35].

We ran miReap with default parameters. FASTQ files were combined into a FASTA file with its reads collapsed. Reads were aligned with bowtie using the same settings used for miRWoods. However, because miRWoods uses quality scores and miReap does not, the allowable error outside of the bowtie alignment seed was changed from 50 to 80 to allow for at least 2 mismatches. Bowtie considers the default value of a mismatch without quality scores present to be 40.

In order to provide a comparison of the three pipelines, a separate set of scripts was used to determine accuracy. For each pipeline being tested a common set of functions was used to score each prediction as a true positive or false positive. We imposed more stringent requirements for true positives than most previous studies that require just overlap with annotated microRNAs. Predicted hairpins where the precursor folded in the wrong direction and only partially overlapped the annotation were named "overlaps" and scored as false positives. Additionally, precursors on the antisense strand of an annotation were named false positives because there is uncertainty whether they are really active as miR precursors.

A set of custom-made scripts was also developed to find homology for novel predictions from each of the three pipelines. Mature products from precursors that did not overlap annotations were searched with BLAST to the database of mature microRNA found in miRBase [62]. Mature products were named homologous if they had the same seed region and an E-value less than 0.05 when compared with a miRNA in the database.

The sensitivity, specificity and F-scores were used to compare each of the three pipelines. The F-score was used to evaluate performance for two reasons. First, different mapping and filtering methods result in variable numbers of precursors being expressed. Because the F-score does not rely on a tally of the number of true negatives, it is better for comparisons. Second, the type of data being analyzed will tend to be very unbalanced with far more non-miRs than miRs, which leads to an inflated accuracy.

2.5.9 Dicer knockdown comparisons

The differences in microRNA expression between wild-type cells and cells in which Dicer had been knocked down were compared across pipelines. Small RNA samples collected from total cell content and cytoplasmic fraction for this test came from the series GSE31069 downloaded from GEO. For each pipeline a set of predictions was generated for both wild-type samples. In each case, the log fold change was computed for each novel prediction comparing the expression of the wild-type cells versus cells in which Dicer had been knocked down. A pseudocount of 0.015 ARPM was used to avoid taking the log of zero.

2.5.10 Validation of bovine and feline microRNA predictions

Novel microRNA predictions were evaluated with homology to known microRNAs from other species and validated by qPCR. We validated the expression of the novel miRs with the highest decision value using qPCR across the tissues we examined.

Feline microRNAs. Feline RNA samples were reverse transcribed with the HiSpec Buffer system of the miScript II RT kit. We performed qPCR in 96 well plates with the ABI StepOnePlus using cDNA generated from 2.5 ng total RNA, miScript Primer assays, and miScript SYBR Green PCR mix combined in 25 µL reaction volumes. Cycling followed manufacturer's instructions. Melt Curve analysis was performed to insure single product generation and the average of all primer efficiencies was 1.8. Of the four potential reference genes selected from feline sequencing data, two, miR-25 and miR-191, were found to be stable across tissues and the average of CT values was used to normalize expression.

Bovine microRNAs. RT was performed using the miScript II RT Kit and qPCR was performed in a HT7900 ABI system in 384-well plate using the Custom miScript Primer Assay and miScript

SYBR Green PCR Kit, following the manufacturer-instructions with a 4-fold dilution of cDNA prior qPCR. We performed normalization using internal control genes (ICGs) or reference genes as indicated by the MIQE guideline [63]. It has been proposed and demonstrated that the use of ICGs for normalization for miRs qPCR provides a more accurate measure of expression than other methods, such as normalization with 5S RNA, U6 snRNA, or total RNA [64]. In order to identify the best ICGs to normalize the novel miRs, we selected predicted miRs with low-variability and similar levels in expression across various tissues, as previously performed [64]. The miRs selected for bovine were miR-7, bta-miR-32, bta-miR494, bta-miR-1388, bta-miR-2431, bta-miR-2483, and bta-miR-6520; Final qPCR data for bovine were analyzed using LinRegPCR to account for efficiency of amplification [65]. Bovine qPCR data from the tested internal control miRs were normalized using geNorm to determine the M- and V-values [66]. bta-miR-7 had a M-value >1.5 and was therefore not used for normalization but rather as a positive control, while the most stable miR pair was miR-494 and miR-6520 (M=0.98). The most stable normalization was obtained by using the 6 most stable miRs with a final V-value of 0.245. The normalization factor was calculated by geNorm as the geometrical mean of the most stable miRs.

2.5.11 Hierarchical clustering

Hierarchical clustering was performed for the expression of known and novel mir-2284/mir-2285 family miRs in bovine. Expression was normalized by computing z-scores, subtracting the mean and dividing by the standard deviation across tissues.

2.5.12 Identification of clusters

Clusters were identified by locating sets of precursors with genomic positions within 10 kbp of each other. Prior to detecting clusters using novel predictions, the set of annotated microRNAs were grouped into clusters. This was done first because if novel microRNAs fell within a cluster of annotated microRNAs it may count as further evidence that that microRNA is real. After the clusters of annotated miRs were identified, novel microRNAs were grouped into new clusters or incorporated into clusters of annotated microRNAs.

2.5.13 Identification of novel miR families

In order to search for novel miR families, the sequences of each novel miR was blasted to a set containing a combination of the novel miRs found using miRWoods and the set of all known miRs from miRBase. Family membership requires a perfectly matching seed sequence, both products were on the same arm for each hairpin, and a BLAST E-value less than or equal to 0.5 for the mature product. In addition, we excluded examples with top hits that are antisense to itself and cases with identical mature sequences to prevent inclusion of loci originating from repetitive regions.

2.6 Tables

Table 2.1: Features used in the mature products random forest (MPRF).

fivePrimeHet	5′-heterogeneity of product reads
medianLength	Median length of product reads
gcContent	GC content of product sequence
aa,ac,ag,at,ca,cc,cg,ct,ga,gc,gg,gt,ta,tc,tg,tt (16 features)	product dinucleotide frequencies
r7,r6,r5,r4,r3,r2,r1,s0,f1,f2,f3,f4,f5,f6,and f7 (15 features)	read abundance 7 nt downstrem to 7 nt upstream product start position
WFC	Wooten-Federhen Complexity of product sequence
Duplex Energy	Minimum free energy of product duplex with surrounding genomic region.

Table 2.2: Features used in the hairpin precursor random forest (HPRF).

Name	Description	Reference
Mfe	minimum free energy of hairpin fold	40, 38*, 22*, 34, 25, 10, 39
Pbp	frequency of paired bases of miR	38, 21*, 37
Urf	fraction of unique reads to total adjusted reads for locus	60, 37
gcContent	GC content of locus sequence	38, 25
totalSenseRPM	Adjusted reads per million (ARPM) in the sense strand	39*, 21*
loopSize	length of the loop in nucleotides.	25, 39
maxBulge	longest bulge appearing in the region of the hairpin spanning the miR and miR*	25*, 39
Tapd	total displacement of sense to anti-sense products	37
Aapd	average displacement of sense to anti-sense products	37
Ahc	average number of hits to the genome for the major product	37
Afh	average 5′-heterogeneity of major product reads	37
sameShift	Amount of offset between products on the same arm	37
bothShift	maximum amount two products are offset on opposite arms	37
Dinucleotide frequencies (16 features)	precursor dinucleotide frequencies	39
maxInteriorLoop	Length of largest interior loop spanning the miR and miR*	39
intLoopSideDiff	Difference in length of of interior loop branches in miR/miR*	39
OPA	Frequency of the most abundant overlapping product	
Duplex Energy	Duplex energy of major product and surrounding region.	
foldDupCmp	Similarity between dot-bracket sequences from RNAduplex and RNAfold	
dupPBP	base pairing density of region duplexing the major product	
dupLoopLength	Length of biggest bulge or interior loop in region duplexing the major product	
APV	The average variance of read counts for distinct reads for all products	
wAPV	The average variance of read counts for distinct reads weighted across products	
ARV	The average variance of start positions for reads on each product	
wARV	The average variance of start positions for reads weighted by product size	
mpLoopDistance	distance of the miR from the loop	
dupLoopDistance	distance of the miR* from the loop	
totalOverlap	The sum of the amounts of overlap between each pair of overlapping reads.	
totalRelativeOverlapAmount	sum of each overlap multiplied by the abundance ratio of the smaller to larger product	
averageOverlapAmount	sum of each overlapping product multiplied by the frequency of reads of the smaller product within the hairpin	
innerLoopGapCount	number of times 3 or more unbound nucleotides appears in the loop region	
totalAntisenseRPM	Adjusted reads per million (ARPM) in the anti-sense strand	
maxUnboundOverhang	The largest length of unpaired nucleotides on either side of the miR	
numOffshoots	number of additional hairpins formed on or across from the miR or miR*	
dupSize	The size of the region duplexed by the miR product	
neighborCount	The number of regions of contiguous read loci within 1000 nucleotides of the precursor	
RFProductAvg	Decision value returned by the random forest in the product phase	
Product counts (8 features)	The fraction of the product relative to the total for the hairpin	
Product overlaps (11 features)	Overlaping lengths for individual products within the locus (e.g. “miRmoR5pOverlap” the overlap between miR and moR on 5′ arm.	

*References with an asterisk use a variant of the described feature.

Table 2.3: Percentage of cases where duplex method produced span used in final prediction.

	All Annotated Precursors			**Annotated Precursors with candidate spans > 1**		
Sample	Duplex-Focused Span Matches Predicted (%)	Duplex-Focused Span Correct (%)	Predicted Span Correct (%)	Duplex-Focused Span Matches Predicted (%)	Duplex-Focused Span Correct (%)	Predicted Span Correct (%)
MCF7 (total)	90.19	96.56	96.82	77.49	97.66	98.25
MCF7 (cytoplasm)	88.55	95.97	96.10	75.41	97.57	97.84
cell lines	85.99	95.25	96.18	78.90	95.91	97.31
Liver	89.43	96.12	95.86	78.24	96.69	96.14

Table 2.4: Percentage of predicted hairpin spans matching miRBase annotation. The method with the highest percent for a particular sample are presented in bold.

	miRWoods		**miRDeep2**		**miReap**	
Library	total	percent (%)	total	percent (%)	Total	percent (%)
human MCF-7 (total cell)	450	**98.901**	318	98.452	430	98.398
human MCF-7 (cytoplasm)	452	98.69	314	**99.054**	428	97.717
human liver	385	**99.483**	318	99.375	413	98.804
human cell lines	736	**99.325**	532	98.519	228	97.854
mouse brain	405	**100**	330	98.214	370	98.143
mouse embryo	486	**99.59**	412	98.329	398	97.073
mouse newborn	419	**99.762**	335	97.384	179	97.283
mouse ovary	282	**100**	243	97.2	237	98.75
mouse testes	293	**100**	269	97.464	260	97.744

Table 2.5: Performance of miRWoods compared to miRDeep2 and miReap. The method associated with the highest F-score for a particular sample are presented in bold.

	miRWoods			miRDeep2			miReap		
Library	precision	recall	F-score	precision	recall	F-score	Precision	recall	F-score
human MCF-7 (total cell)	0.727	0.501	**0.593**	0.839	0.354	0.498	0.42	0.478	0.447
human MCF-7 (cytoplasm)	0.7	0.511	**0.591**	0.86	0.355	0.503	0.476	0.484	0.48
human liver	0.871	0.447	**0.591**	0.898	0.369	0.523	0.446	0.48	0.462
human cell lines	0.627	0.586	**0.606**	0.834	0.424	0.562	0.264	0.182	0.215
mouse brain	0.849	0.569	**0.681**	0.951	0.463	0.623	0.397	0.52	0.45
mouse embryo	0.694	0.567	0.624	0.898	0.481	**0.626**	0.205	0.464	0.285
mouse newborn	0.836	0.559	**0.67**	0.931	0.447	0.604	0.312	0.239	0.271
mouse ovary	0.953	0.603	**0.738**	0.96	0.519	0.674	0.798	0.506	0.62
mouse testes	0.91	0.579	**0.708**	0.944	0.532	0.68	0.324	0.514	0.397

Table 2.6: Tuning Results.

Train Set	Tuning Set	Validation Set	Ratio of negatives in set (1:X)	Decision Value Threshold	ARPM Threshold	F-Score
GSE23090	GSE65706, GSE62501	GSE18381, GSE20592	25.361	0.25	0.13	0.669
GSE23090	GSE18381, GSE20592	GSE65706, GSE62501	25	0.28	0.11	0.711
GSE40499	GSE65706, GSE62501	GSE18381, GSE20592	27	0.37	0.08	0.644
GSE40499	GSE18381, GSE20592	GSE65706, GSE62501	59.382	0.29	0.11	0.662

Table 2.7: Dicer knockdown of novel predictions compared with known annotations. P-values computed from comparing the log-fold change of Dicer knockdowns compared to wild-type using a t-test, for novel predictions from each software and known annotations from miRBase.

MCF-7 Cytoplasmic Fraction				
	miRWoods	mirdeep	miReap	miRBase
miRWoods	-	0.56541159	0.29649206	0.00012122
mirdeep	0.56541159	-	0.89559594	0.09242175
miReap	0.29649206	0.89559594	-	0.00296758
miRBase	0.00012122	0.09242175	0.00296758	-
MCF-7 Total Cell Content				
	miRWoods	mirdeep	miReap	miRBase
miRWoods	-	0.2670974	0.7934435	1.3489E-14
mirdeep	0.2670974	-	0.14420995	4.8891E-05
miReap	0.7934435	0.14420995	-	1.9469E-28
miRBase	1.3489E-14	4.8891E-05	1.9469E-28	-

2.7 Figures

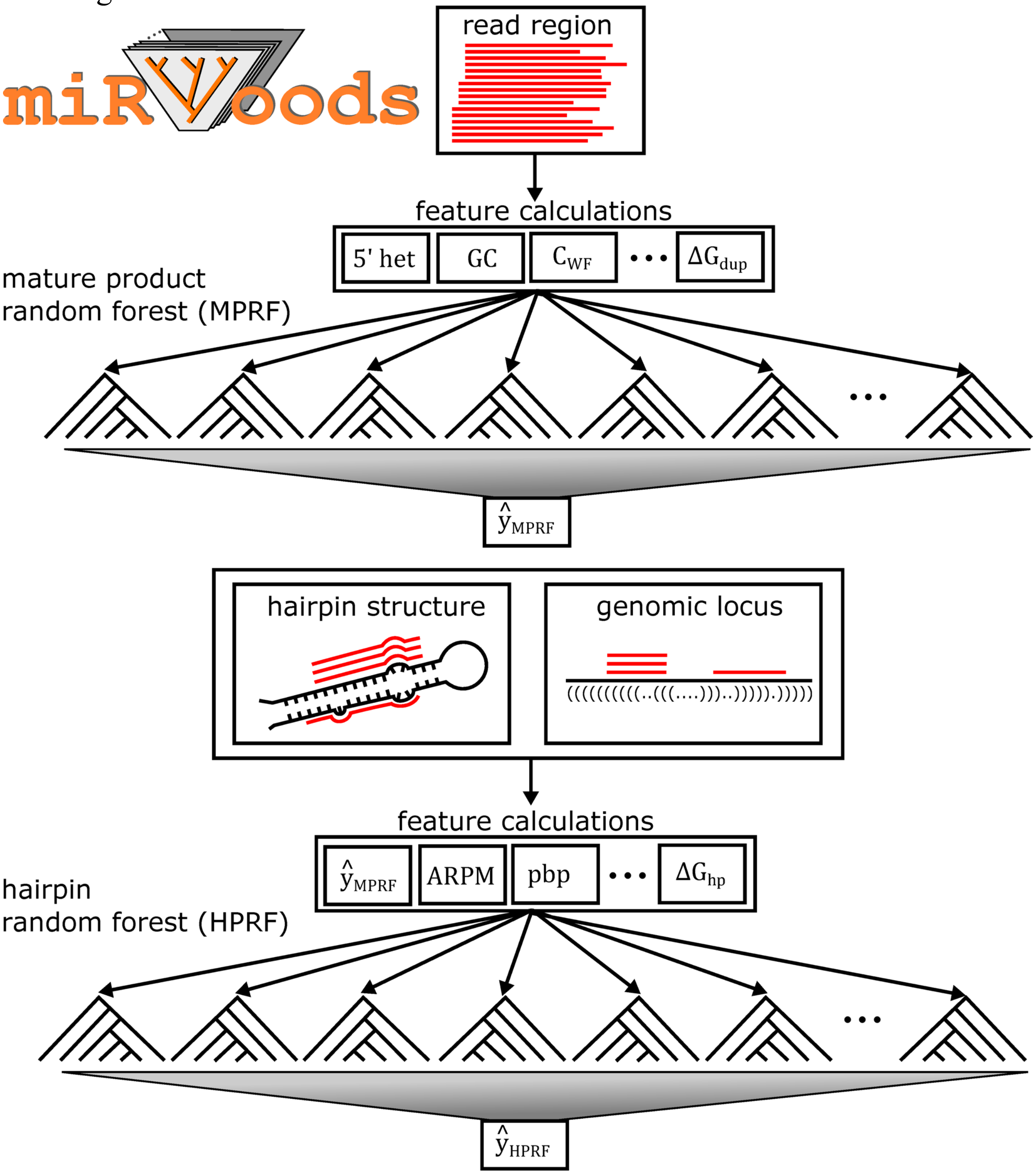

Figure 2.1: Outline of miRWoods Pipeline. After aligning to the genome, overlapping reads are grouped together to form read stacks. Read stacks are scored by the Mature Product Random Forest (MPRF), to predict a set of putative mature microRNAs. Products which meet the minimum threshold score for the MPRF are combined with the surrounding region to form hairpins and each hairpin is folded. Hairpins are scored by the Hairpin Random Forest (HPRF) and a set of final predictions are generated which meet the minimum threshold for the HPRF score.

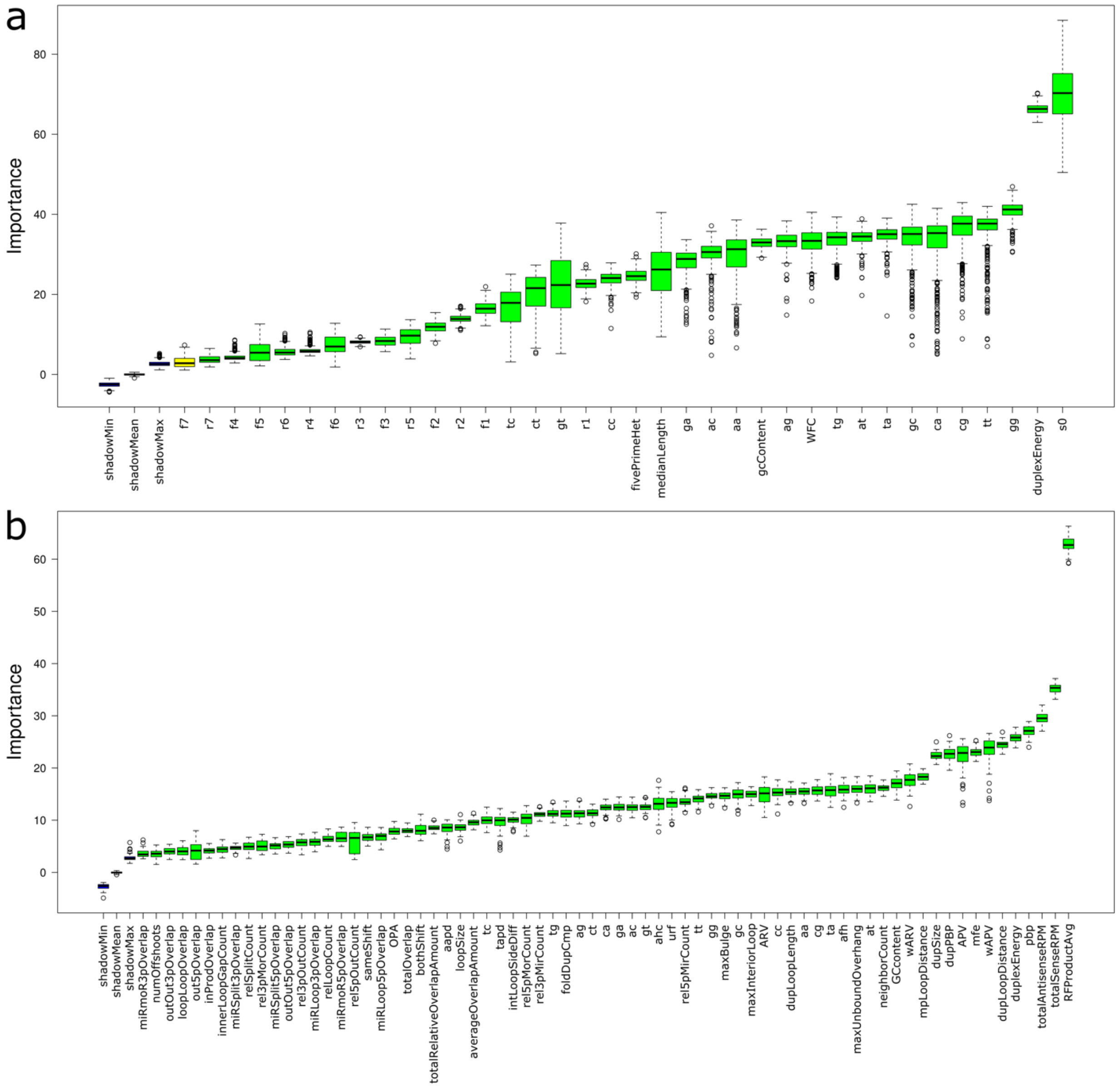

Figure 2.2: Importance of features. **a** The importance of each feature based on the Boruta analysis for the mature product random forest (MPRF) **b** The importance of each feature based on the Boruta analysis for the Hairpin Random Forest (HPRF).

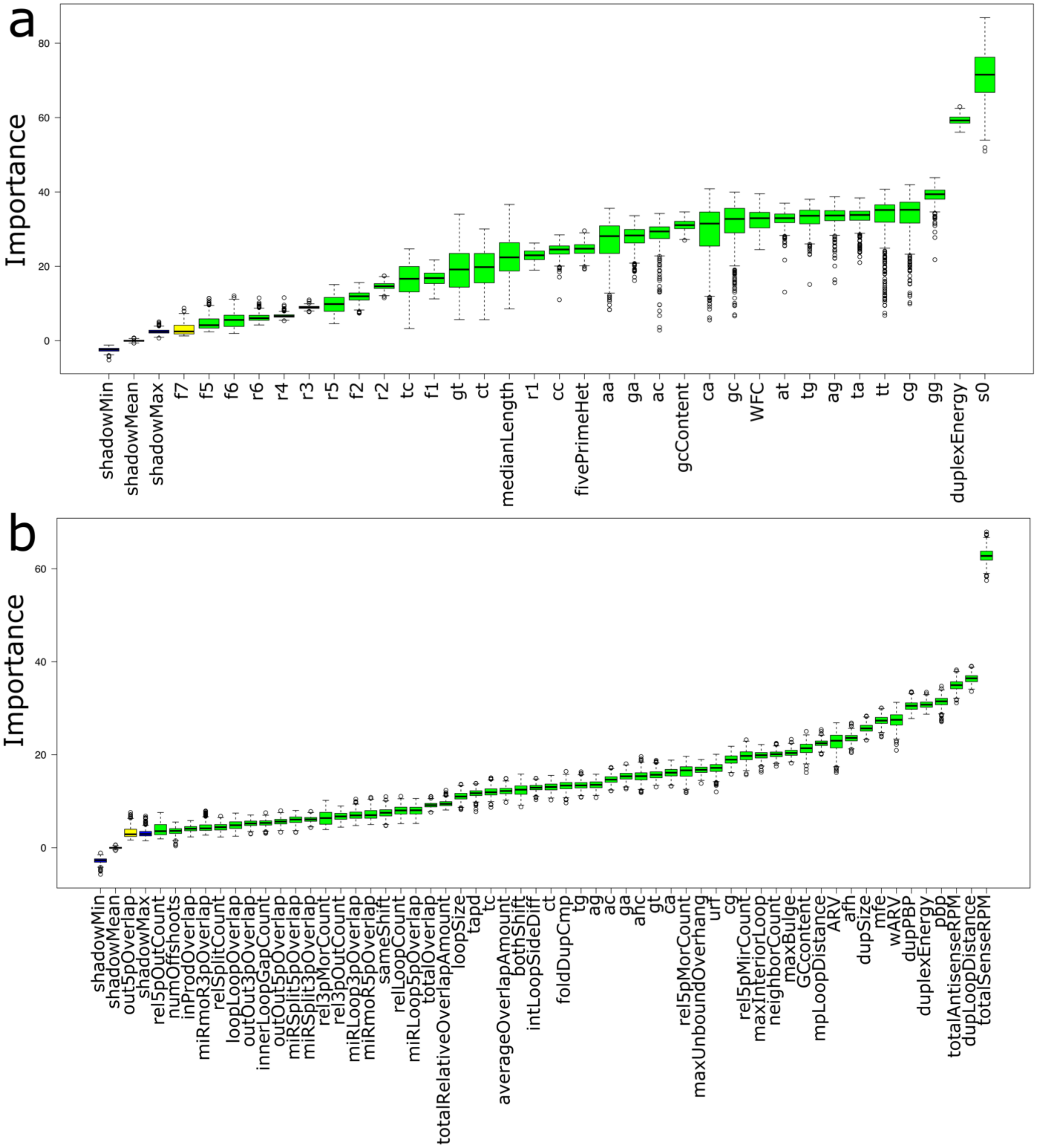

Figure 2.3: Further feature interpretation. Removal of correlated features I. **a** Boruta analysis of feature importance for MRPF with correlated features removed. **b** Boruta analysis for HRPF with correlated features and the MRPF decision value removed.

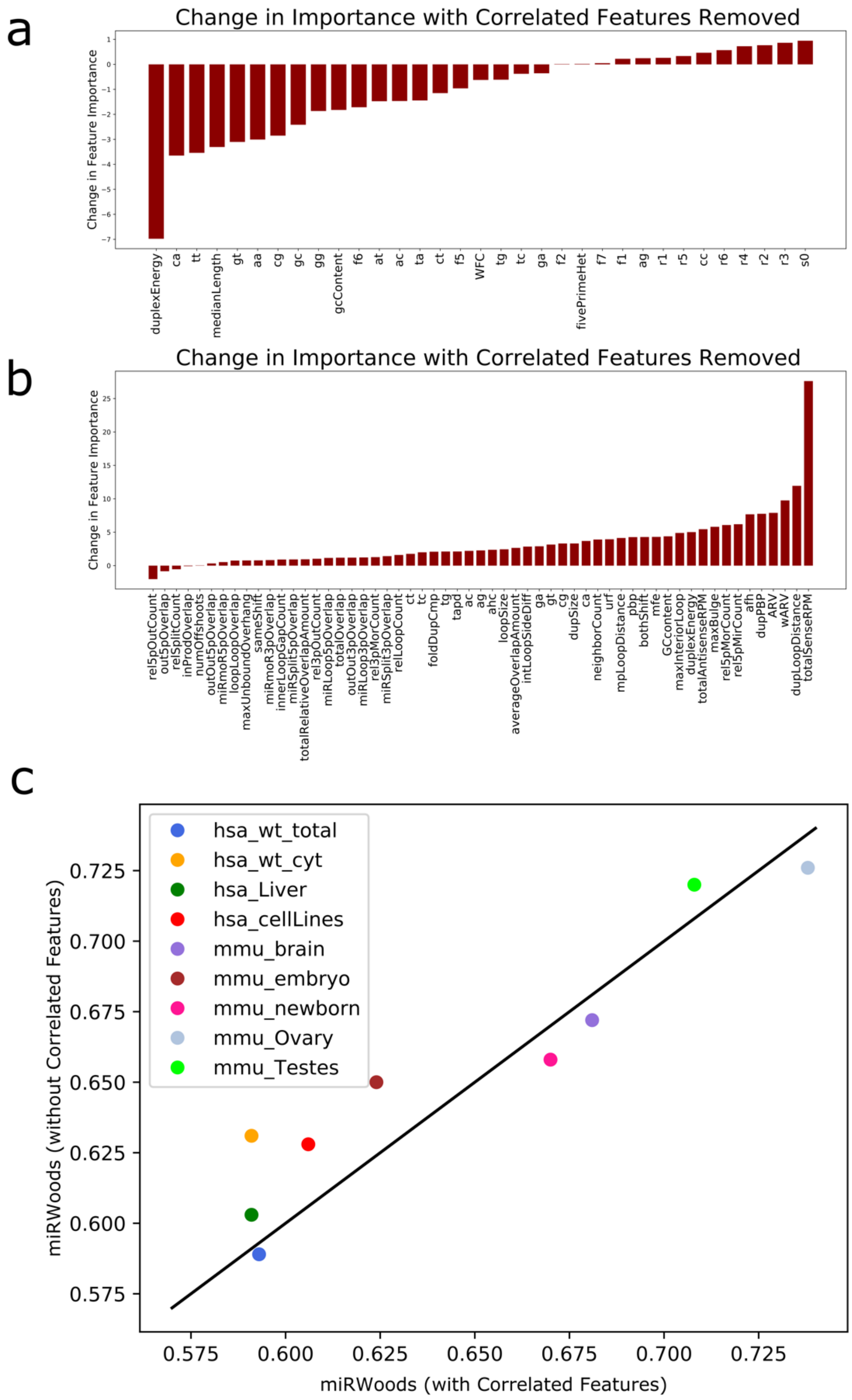

Figure 2.4: Further feature interpretation. Removal of correlated features II. **a** Boruta analysis of feature importance for MRPF with correlated features removed. **b** Boruta analysis for HRPF with correlated features and the MRPF decision value removed.

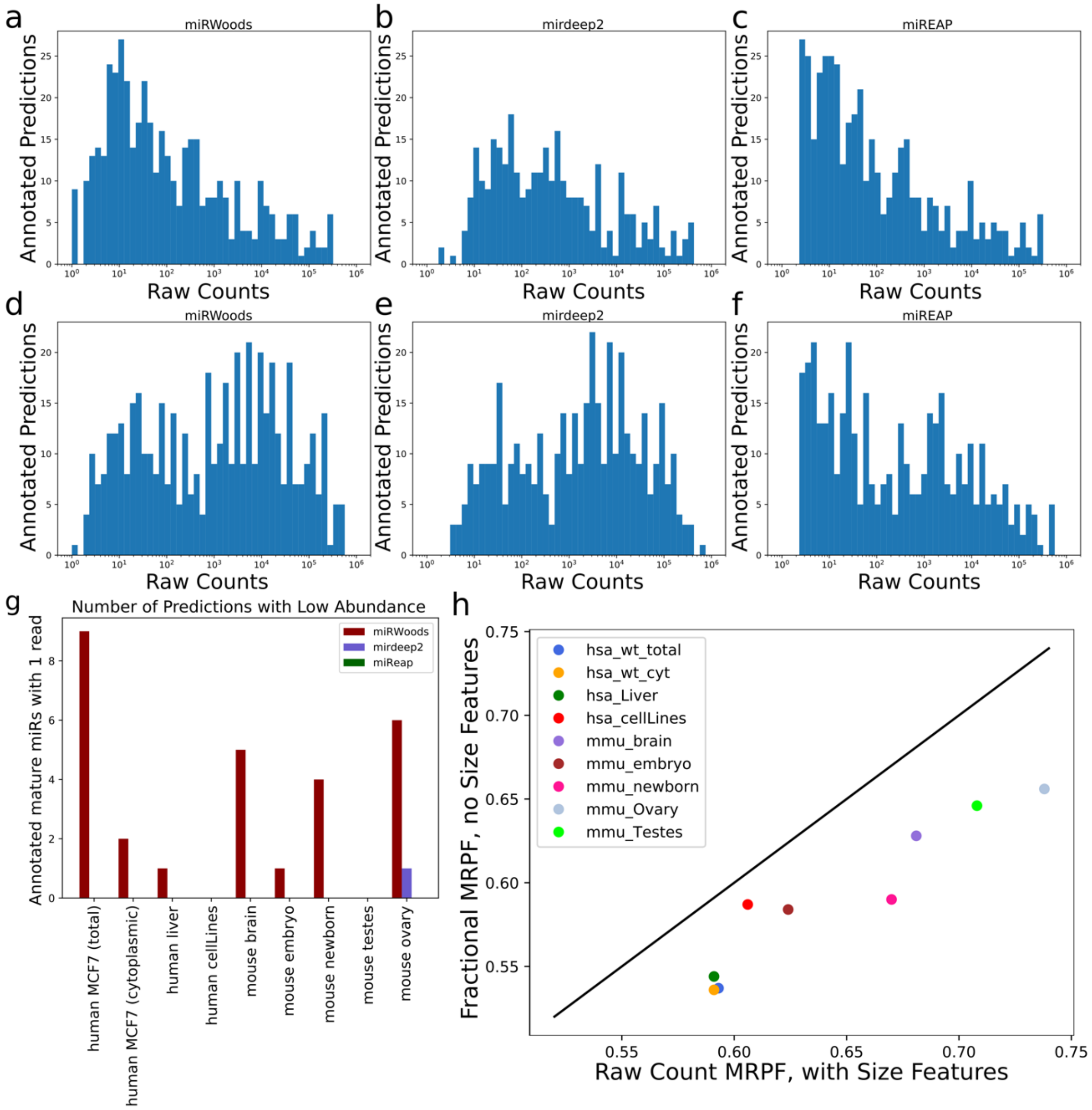

Figure 2.5: Abundance-related features. **a** Distribution of read abundance for correct miRWoods predictions on MCF7 total cell content. **b** distribution of read abundance for correct miRDeep2 predictions on MCF7 total cell content. **c** distribution of read abundance for correct miReap predictions on MCF7 total cell content. **d** Distribution of read abundance for correct miRWoods predictions on mouse embryos. **e** distribution of read abundance for correct miRDeep2 predictions on mouse embryos. **f** distribution of read abundance for correct miReap predictions on mouse embryos. **g** bar plot of correct predictions where the most abundant mature product has one read for samples in human and mouse. **h** F1-score of predictions with size-related features compared to without.

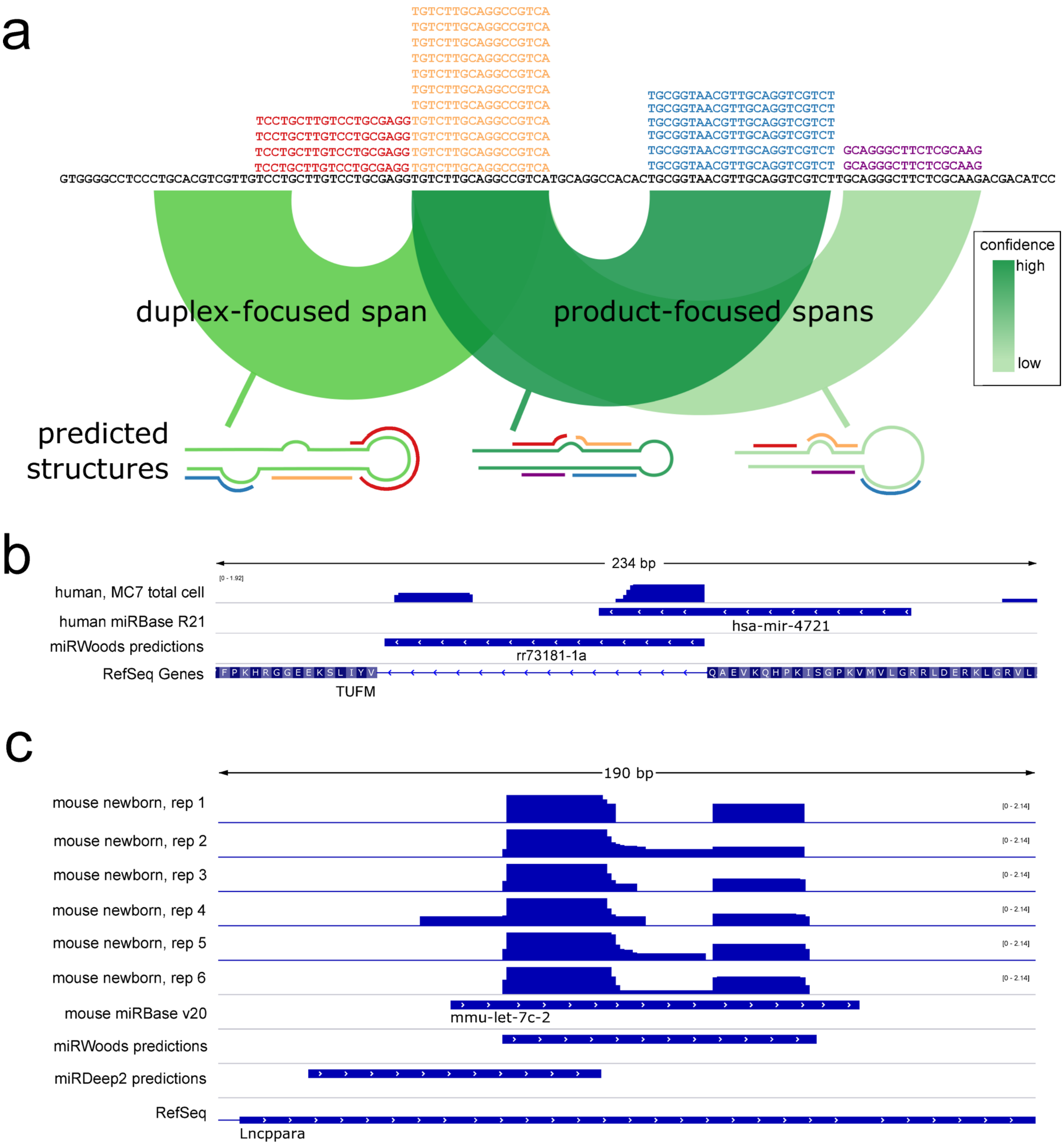

Figure 2.6: Improved Hairpin Precursor Span Identification. **a** miRWoods generates several potential hairpin precursor spans from each product that passes through the MPRF. Duplex-focused spans take the region between the product and the optimal duplex and product-focused spans take the region between the product and other products greater than 4 nt away. Hairpins are selected based on HPRF score. **b** The miRBase annotation for hsa-mir-4721 crosses over an intron boundary. miRWoods corrects the annotation by recognizing a second read stack and produce precursor span that perfectly matches an intron, suggesting mir-4721 is a mirtron. **c** The miRWoods prediction for mmu-let-7c-2 in mouse is consistent with the miRBase annotation, while the best miRDeep2 prediction only partially overlaps with the miRBase annotation.

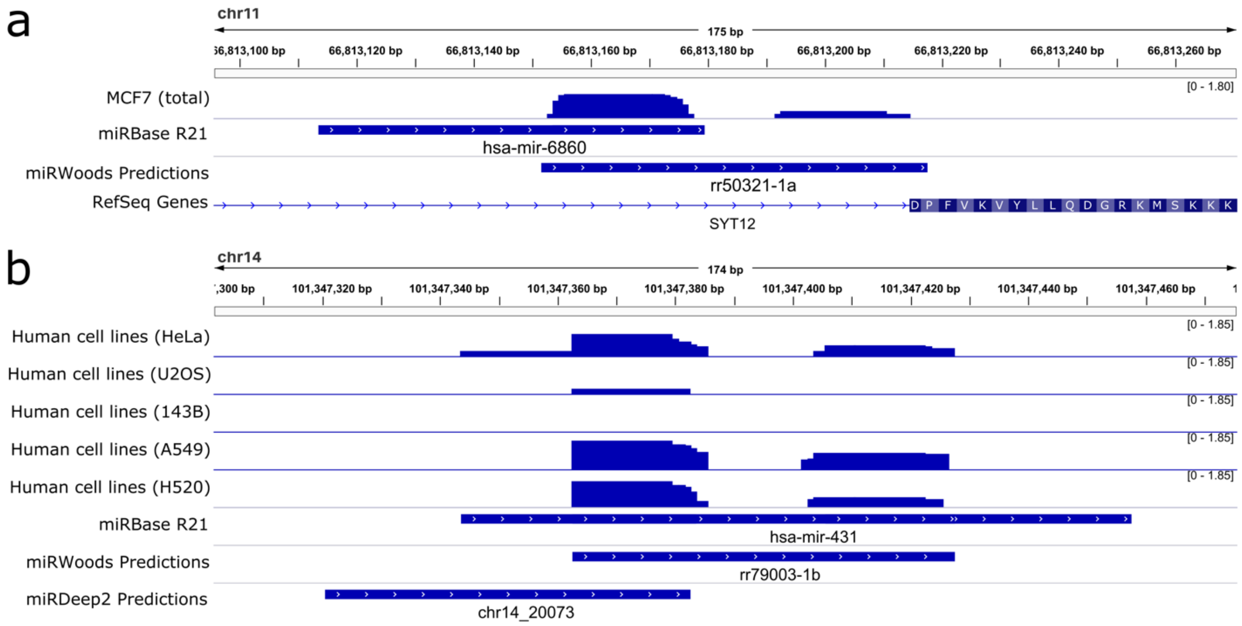

Figure 2.7: Effectiveness of duplex method. **a** RNAseq for hsa-miR-6860 shows miRWoods prediction covering an additional read stack next to the splice junction, which indicates that hsa-miR-6860 may be a half-mirtron. **b** RNAseq for hsa-mir-431 showing predicted folds for miRWoods and miRDeep.

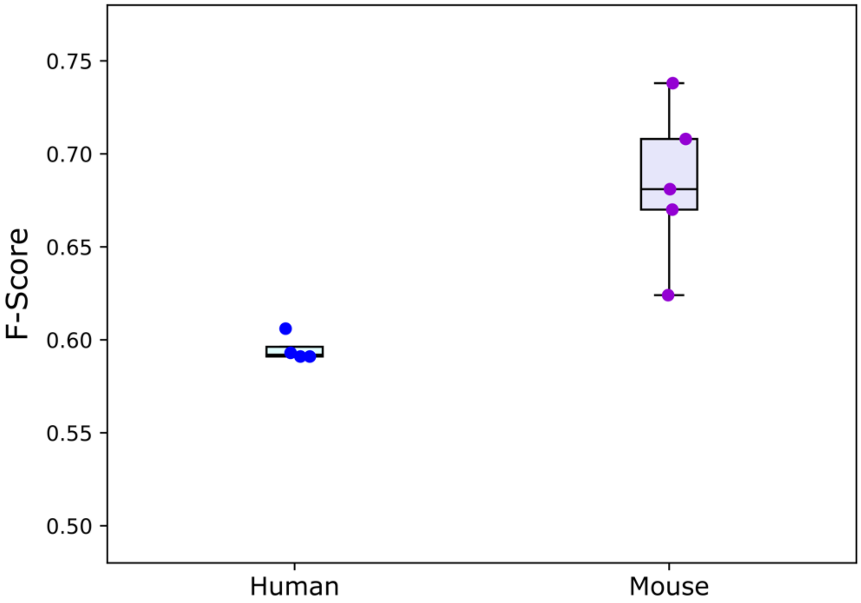

Figure 2.8: Cross-species performance. Comparison between cross-species F1-score and same-species F1-score. All of miRWoods evaluations were tested on a single model trained and tuned on human datasets. The best performance is observed on mouse samples.

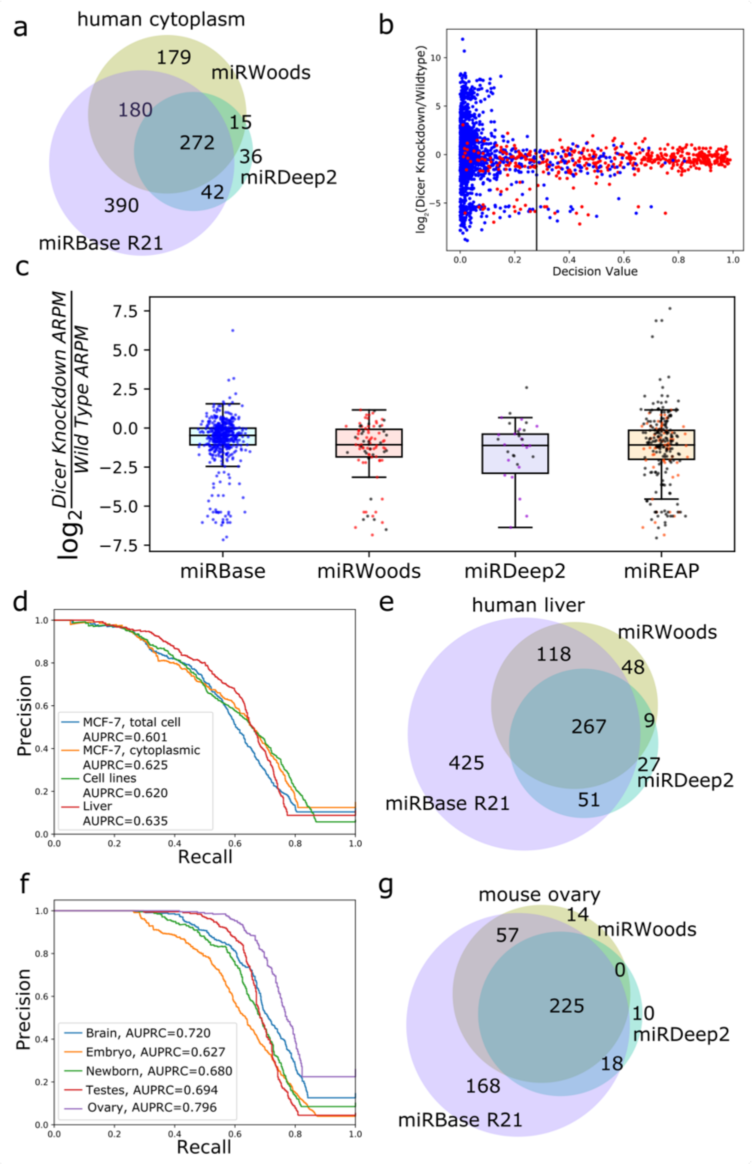

Figure 2.9: Evaluation of miRWoods performance. **a** Euler diagrams comparing predictions from miRWoods and miRDeep with annotations from miRBase for human MCF-7 cytoplasmic extract **b** A scatterplot comparing the miRWoods decision value to the log fold change in Dicer knockdown cells compared to wild-type cells. **c** Scatter-boxplot comparing the log fold change for Dicer knockout to wild type for unprocessed read regions, miRBase annotations, and predictions from miRWoods, miRDeep, and miReap for MCF-7 (cytoplasmic fraction). Black dots indicate predictions that are unique to this method. **d** Precision-recall (PR) Curve and AUPRC of miRWoods predictions for human including MCF-7 (total cell content), MCF-7 (cytoplasmic fraction), cell lines, and liver. **e** Euler Diagrams comparing predictions from miRWoods and miRDeep with annotations from miRBase for human liver. **f** Precision Recall Curve and AUPRC of miRWoods predictions for mouse tissues including brain, embryo, newborn, testes, and ovaries sets. **g** Euler Diagrams comparing predictions from miRWoods and miRDeep2 with annotations from miRBase for mouse ovary.

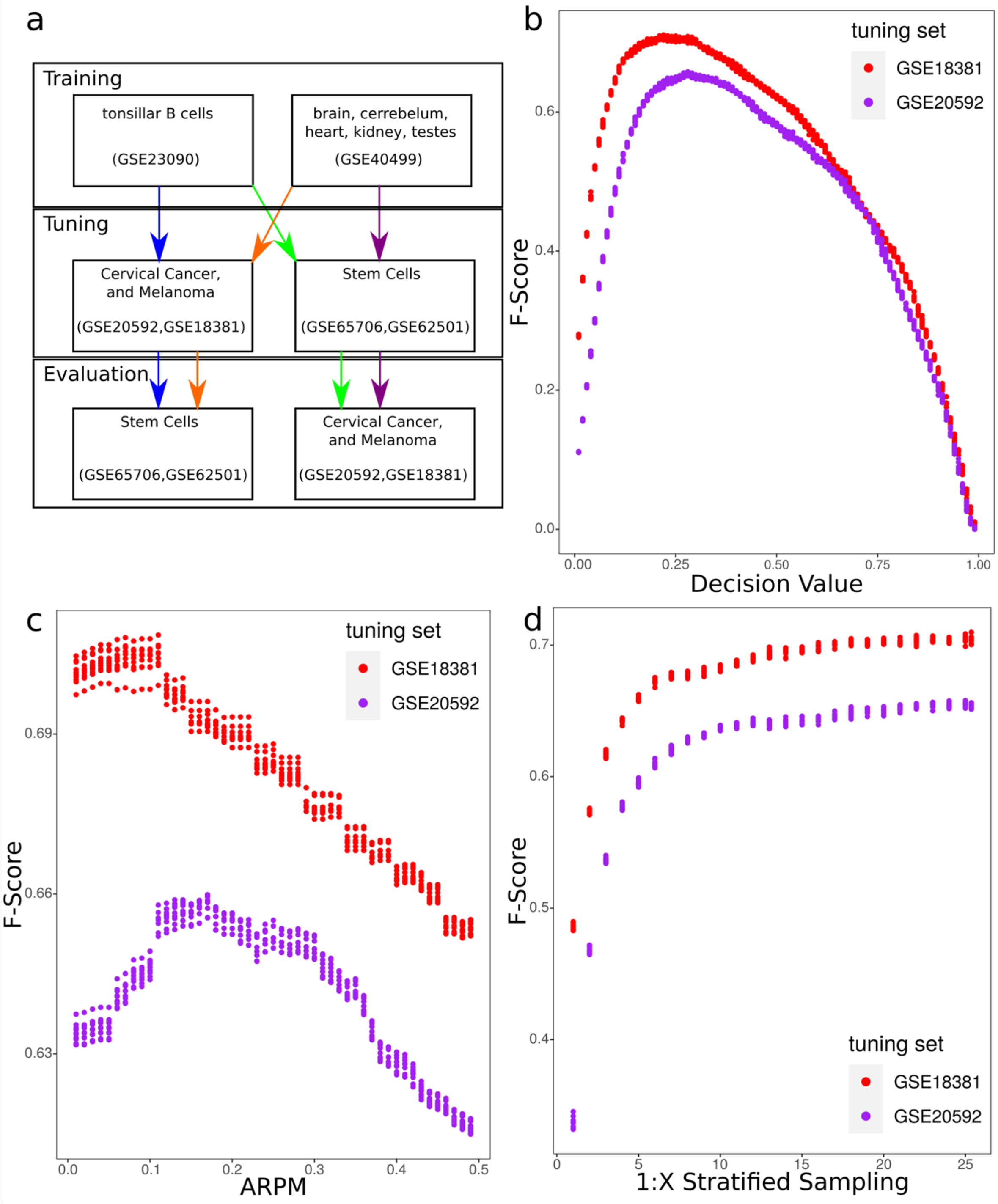

Figure 2.10: Tuning miRWoods. **a** Analysis pipelines and corresponding data sets used for training, tuning, and evaluation correspond to the paths of the arrows. **b** Plot of F1-score versus decision value threshold used in tuning the decision value threshold. **c** Plot of F1-score versus ARPM threshold used in tuning the ARPM threshold. **d** Plot of f1-score versus X in 1:X stratified sampling used to tune the amount of negative (non-miR) loci used in training the HRPF.

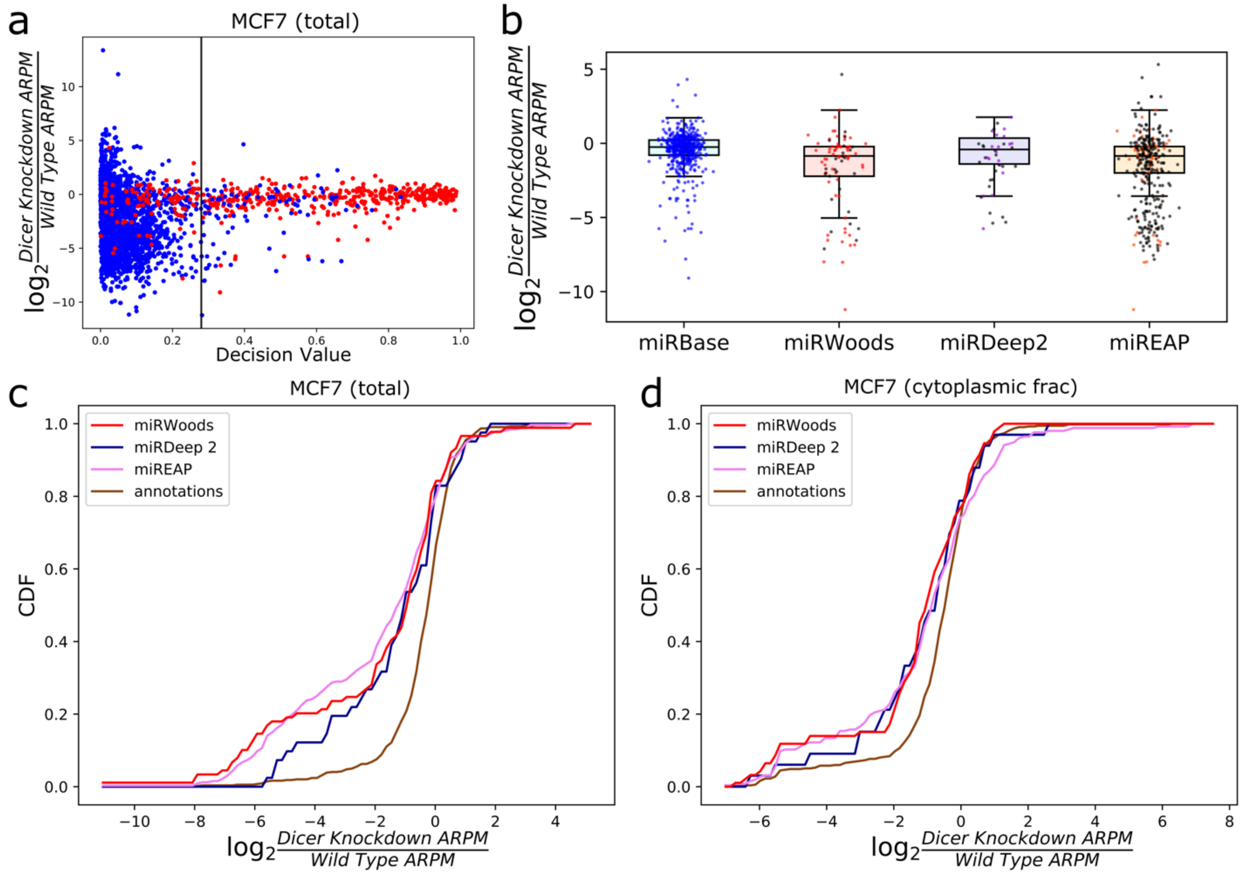

Figure 2.11: Dicer Knockdown Data. **a** Scatter plot for hairpins in the MCF7 (total) set, plotting log fold change of Dicer knockdown vs wildtype against the miRWoods decision value for annotated (red) and novel (blue) hairpins. The vertical line in the plot represents the decision value cut-off with all miRWoods predicted precursors to the right of it. **b** Box plot showing the log fold change of Dicer knockdown vs wildtype of annotated precursors within miRBase and novel precursors is predicted by each software for the MCF7 (total) set. **(c-d)** CDF's for **c** MCF7 (Total) and **d** MCF7 (cytoplasmic) log fold change of Dicer knockdown vs wildtype for novel precursors.

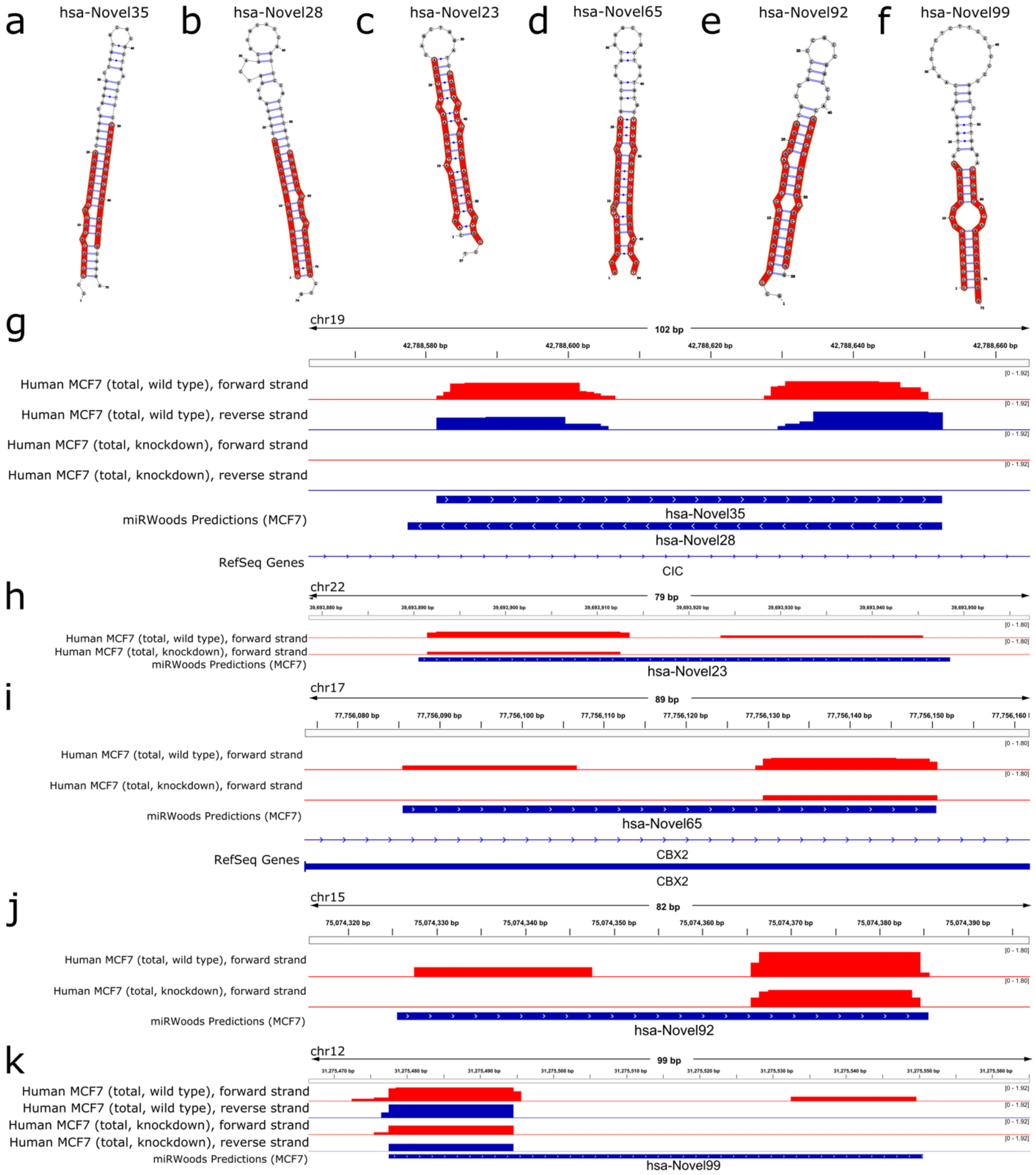

Figure 2.12: Individual Examples of Dicer Knockdown. Predicted secondary structures for **a** hsa-Novel35, **b** hsa-Novel28, **c** hsa-Novel23, **d** hsa-Nove65, **e** hsa-Novel92, and **f** hsa-Novel99. **(g-k)** RNAseq for **g** hsa-Novel35, hsa-Novel28, **h** hsa-Novel23, **i** hsa-Novel65, **j** hsa-Novel92, and **k** hsa-Novel99.

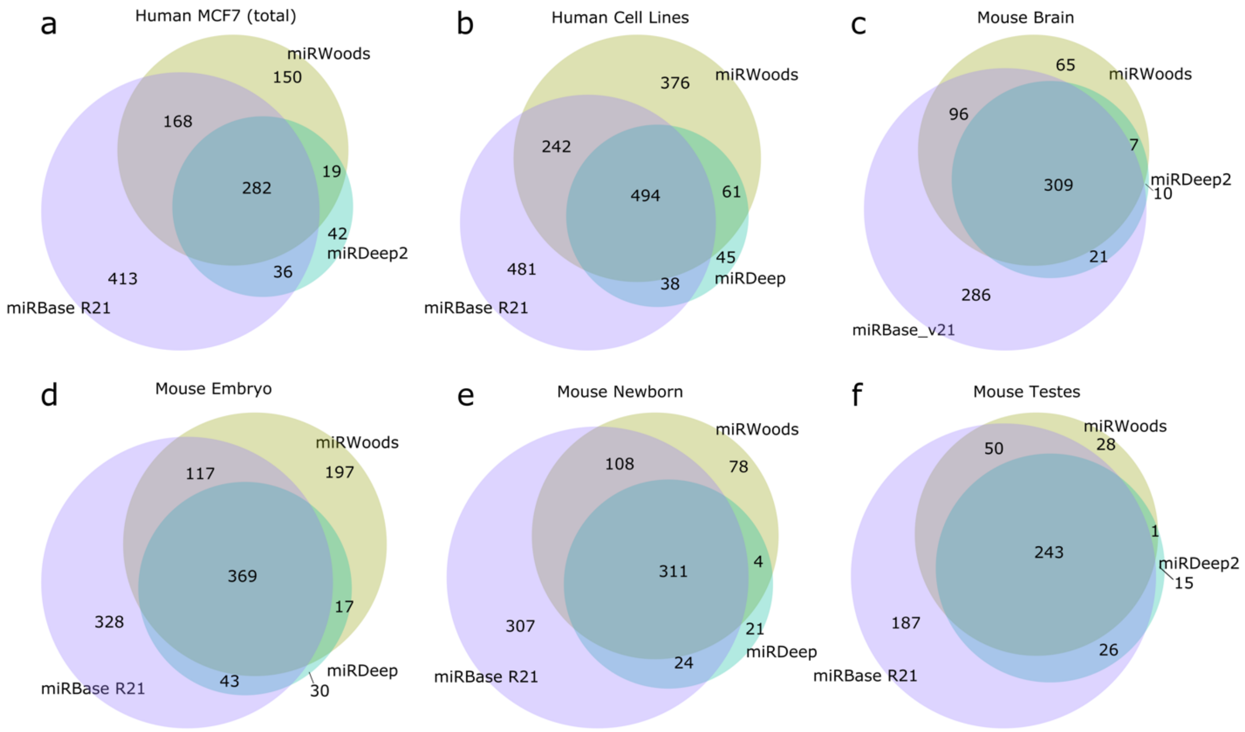

Figure 2.13: Additional Euler Plots comparing miRWoods, miRDeep, and miRBase. Euler plots for **a** Human MCF7 (total), **b** Human cell lines, **c** Mouse brain, **d** Mouse embryo, **e** Mouse newborn, and **f** Mouse testes sets.

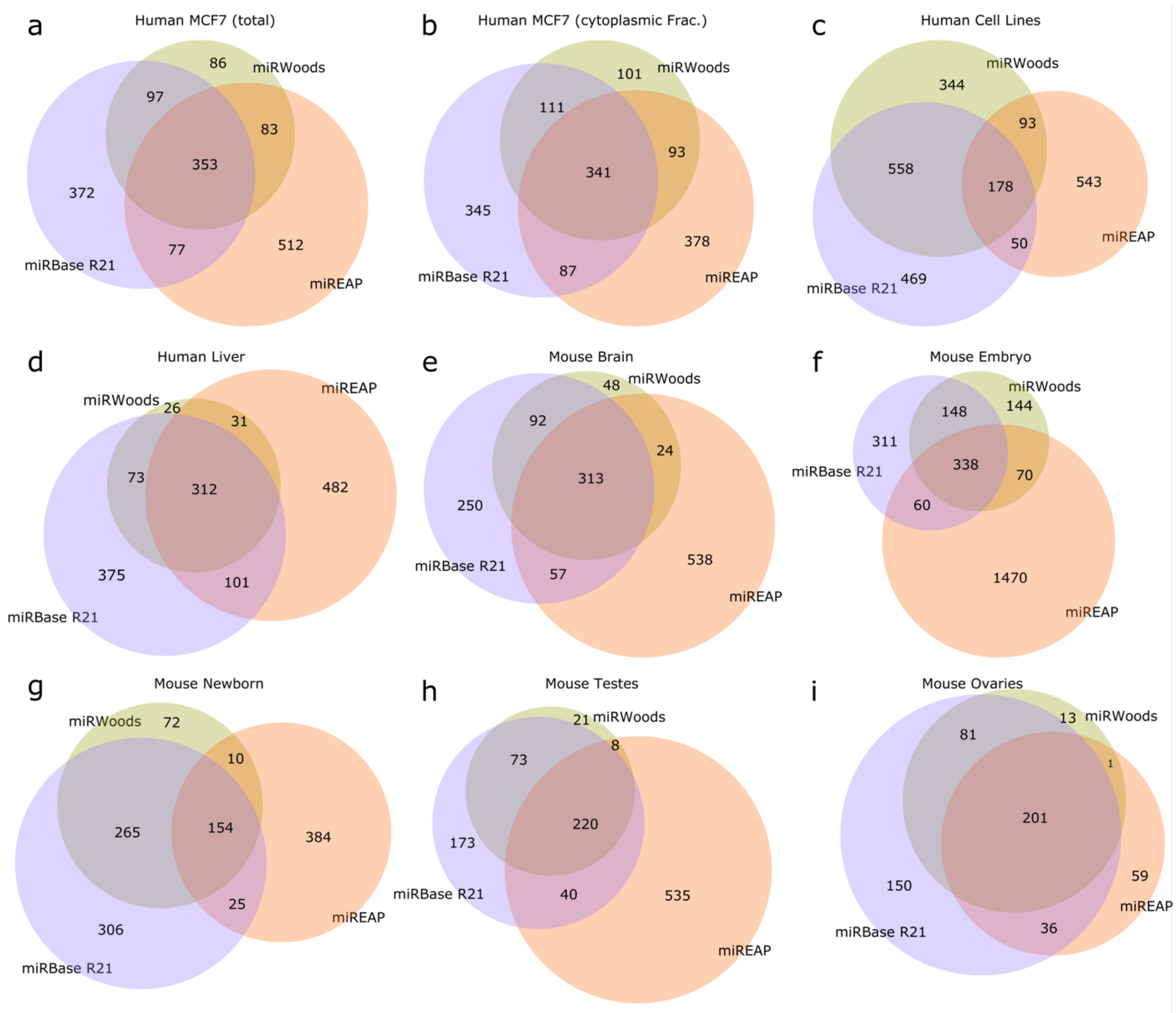

Figure 2.14: Additional Euler Plots comparing miRWoods, miReap, and miRBase. Euler plots for **a** Human MCF7 (total), **b** Human MCF7 (cytoplasmic), **c** Human cell lines, **d** Human liver, **e** Mouse brain, **f** Mouse embryo, **g** Mouse newborn, **h** Mouse testes and, **i** Mouse ovaries sets.

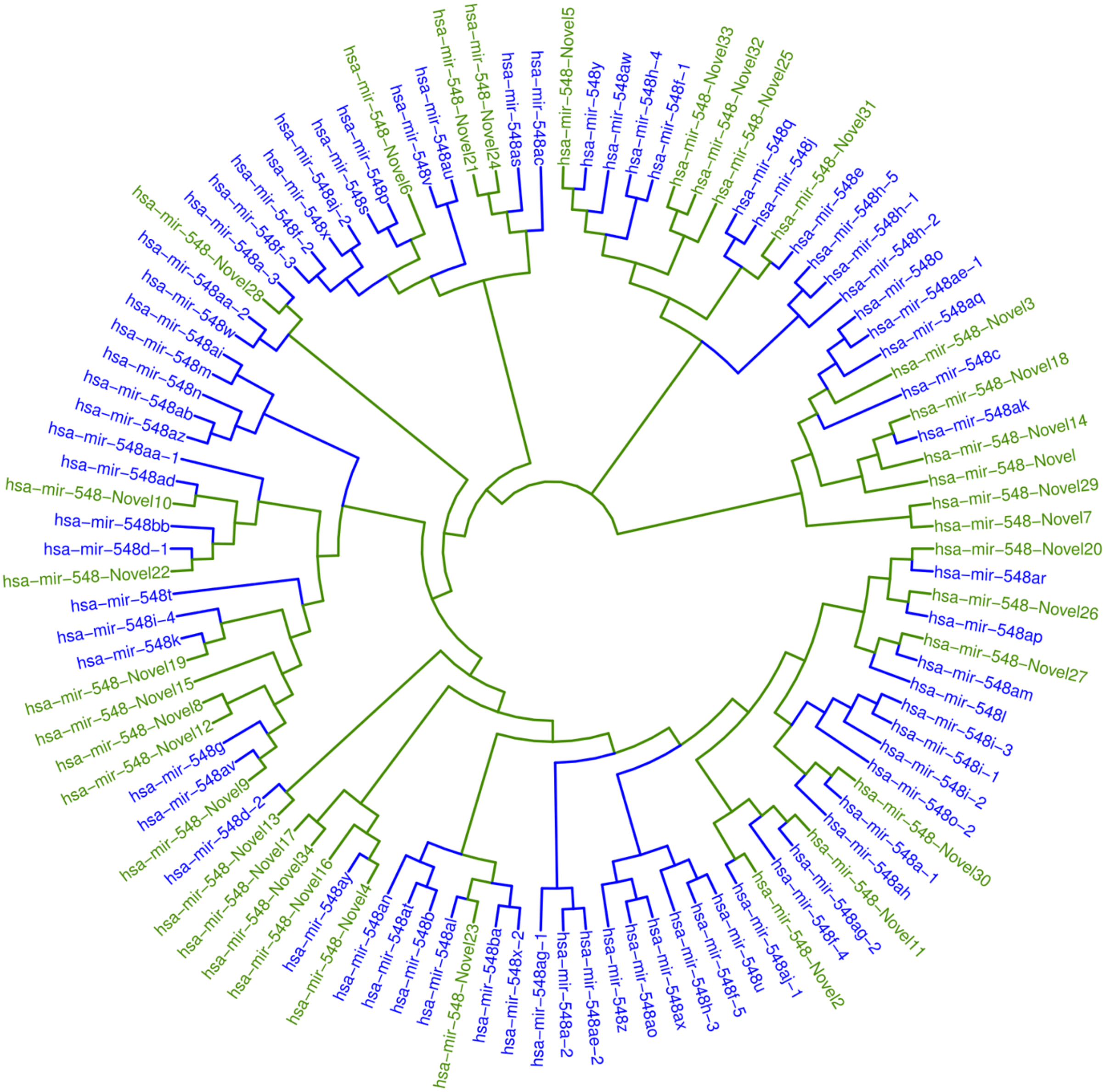

Figure 2.15: mir-548 Phylogenetic tree. Phylogenetic tree showing expansion of the mir-548 precursor family in human. Annotated mir-548 precursors are shown in blue and predicted novel precursors are shown in green.

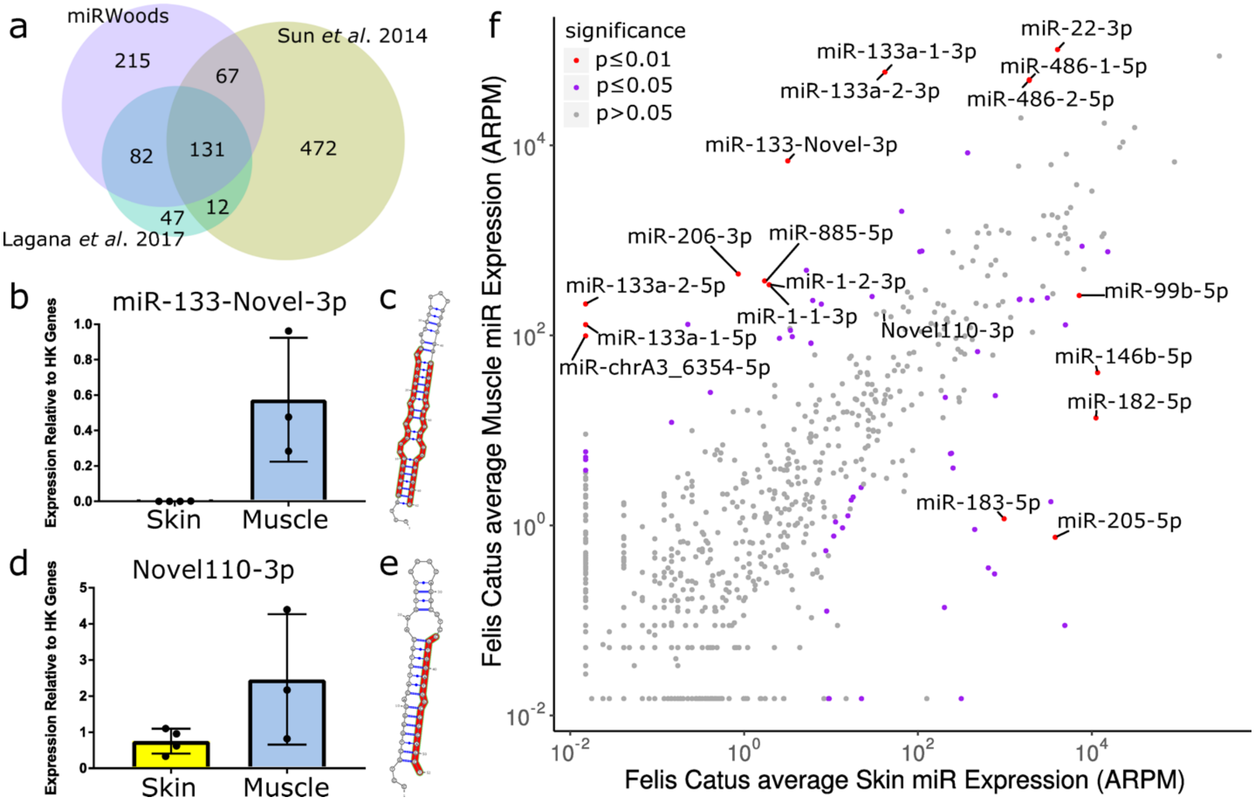

Figure 2.16: miRWoods predictions in the feline genome. **a** Euler diagram of the predictions from miRWoods with predictions from Sun *et al.* (2014) and Lagana *et al.* (2017). **b** The expression in skin and muscle for miR-133-Novel-3p **c** Hairpin for mir-133-Novel precursor. **d** The expression in skin and muscle for Novel110-3p. **e** Hairpin for Novel110 precursor. **f** Scatterplot of average muscle expression vs average skin expression for each mature microRNA.

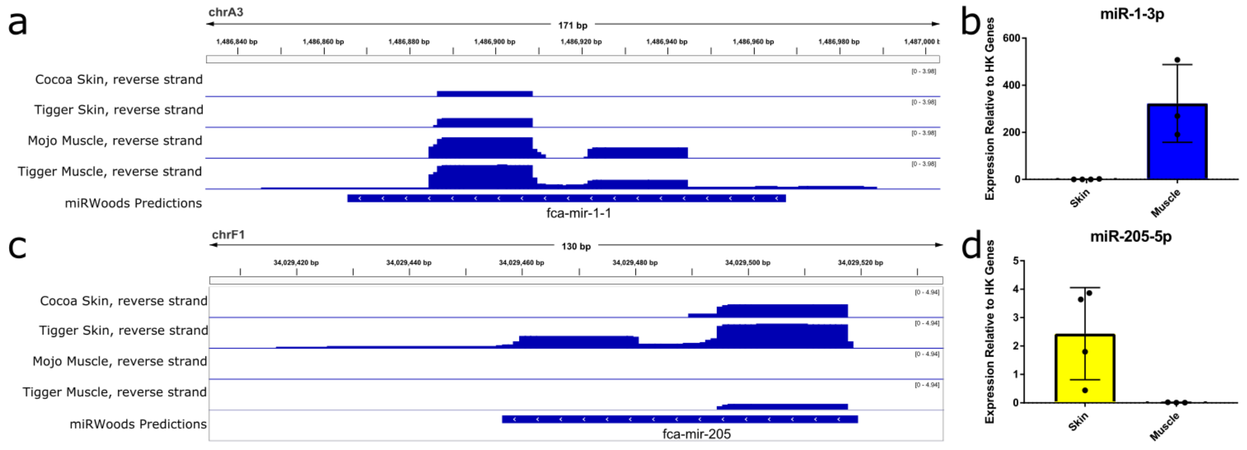

Figure 2.17: Differential Expression analysis of miRs. Expression of fca-mir-1-1 using **a** RNAseq and **b** qPCR validation of differential expression in muscle. **(c-d)** Expression of fca-mir-205 using **c** RNAseq and **d** qPCR validation of differential expression in skin.

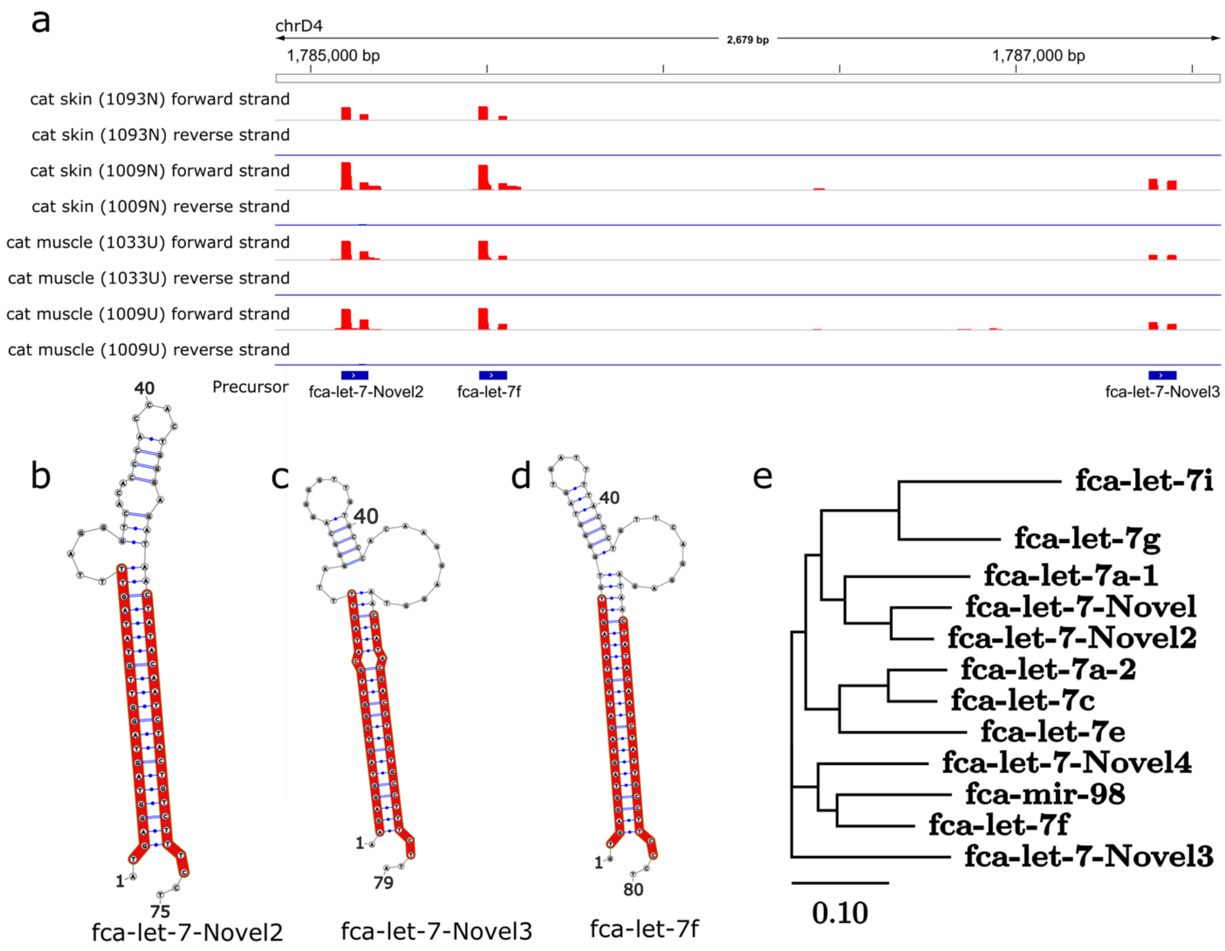

Figure 2.18: Novel let-7 microRNAs in the feline genome. **a** RNA-seq of cluster containing fca-let7-Novel2, fca-let7f, and fca-let7-Novel3 for each skin and muscle sample from *Felis catus*. **b** Hairpin structures for fca-let7-Novel2, **c** fca-let7-Novel3, and **d** fca-let7f. **e** Phylogenetic tree of let-7 miRs including those previously found by Lagana *et al.* (2017).

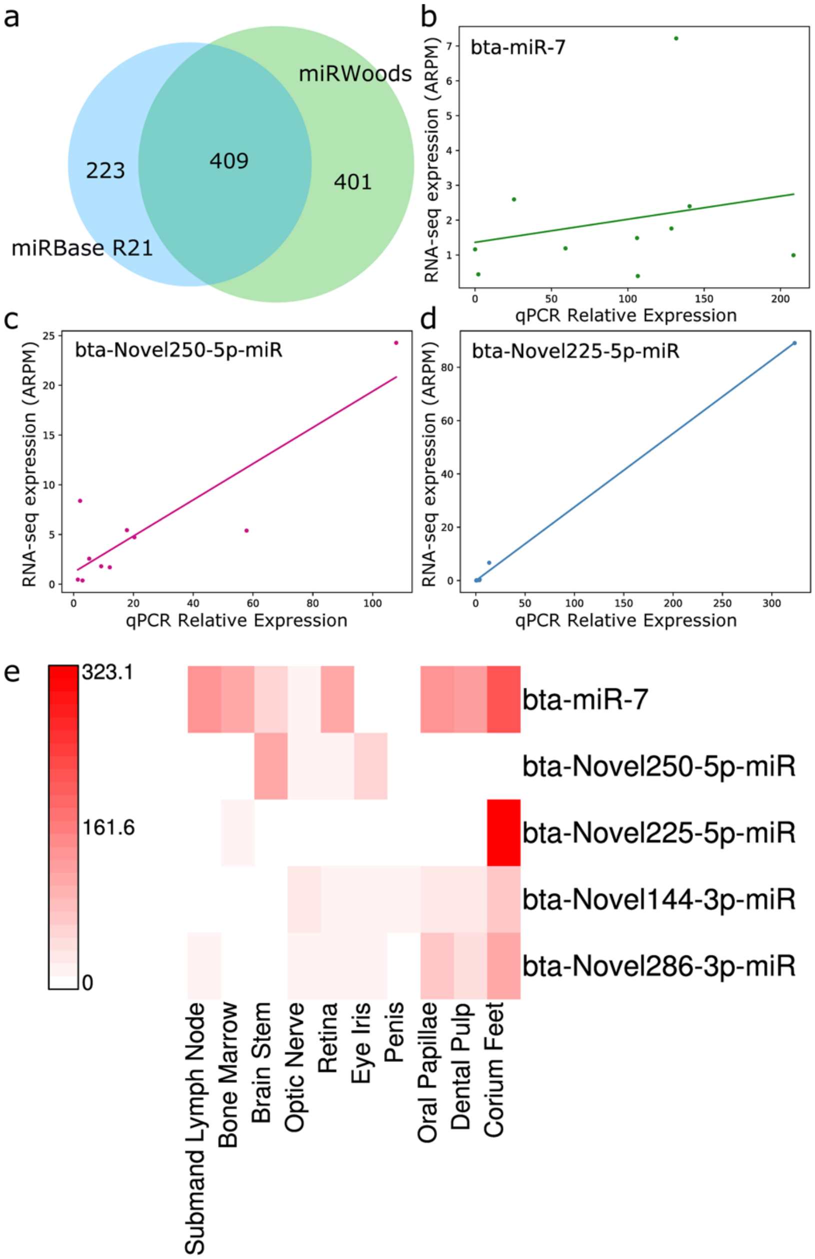

Figure 2.19: Novel microRNA predictions in the bovine genome. **a** Euler diagram comparing miRWoods predictions in the cow genome with miRBase annotations. **b** Scatterplot and best fit line comparing the normalized RT-qPCR expression and RNA-seq for the control miR bta-miR-7. **c** Scatterplot and best fit line comparing the normalized RT-qPCR expression and RNA-seq for a novel predicted miR with enriched expression in brain stem. **d** Scatterplot and best fit line comparing the normalized RT-qPCR expression and RNA-seq for a novel predicted miR with enriched expression in corium feet. **e** Heat map of RT-qPCR expression values over tissues examined.

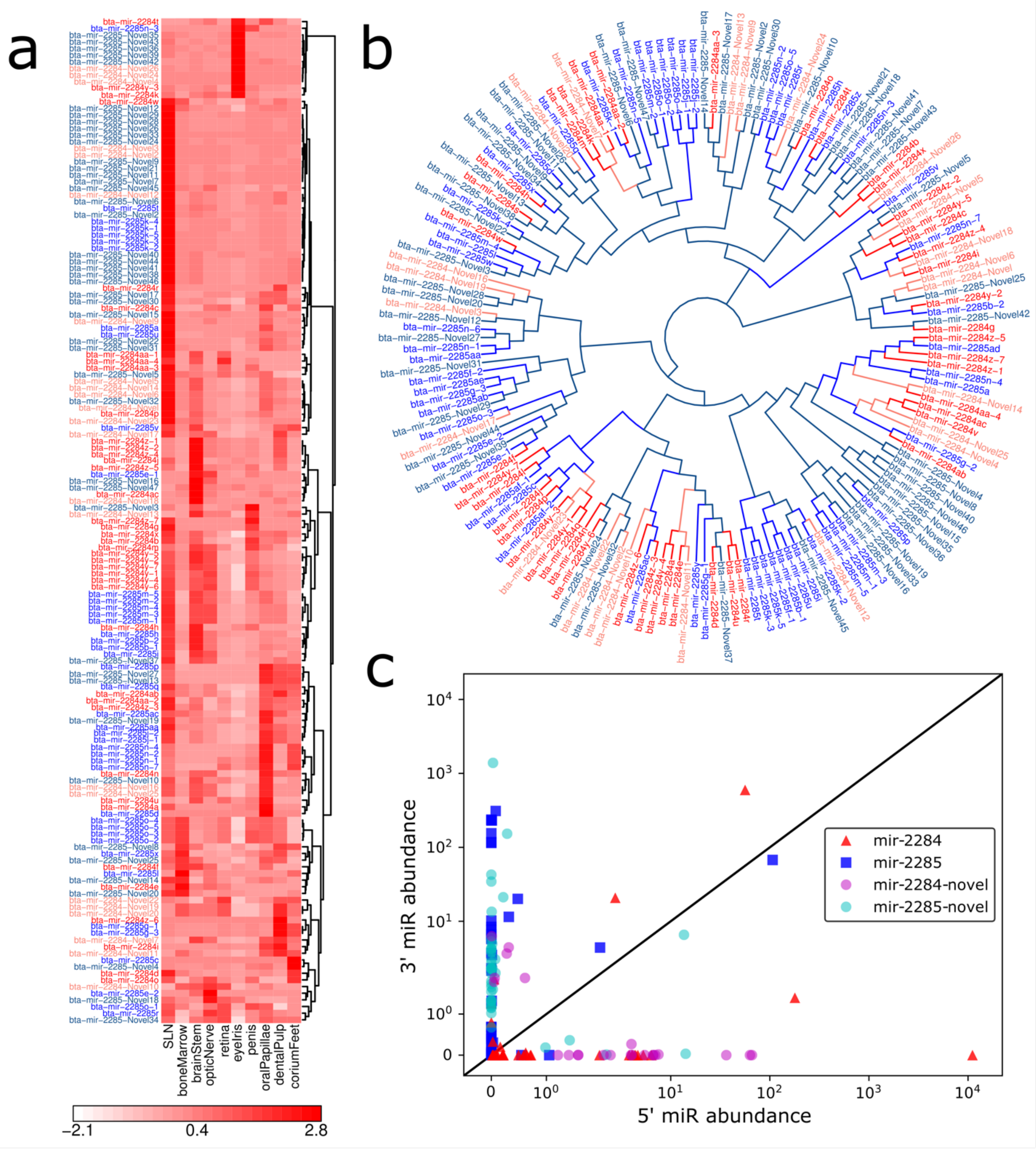

Figure 2.20: mir-2284/mir-2285 family miRs in *Bos taurus*. **a** A heat map for the expression of annotated and novel mir-2284/mir-2285 family miRs. **b** A phylogenetic tree for the bta-2284/bta-2285 family. Variants of bta-mir-2284 appear in red and variants of bta-mir-2285 appear in blue. Colors for novel predictions appear lighter than those for annotated predictions. **c** Abundance of miRs for the 5′ and 3′ sides of the mir-2284/mir-2285 family. The 5′ product tends to show greater expression in the mir-2284 family while the 3′ product shows greater expression in the mir-2285 family.

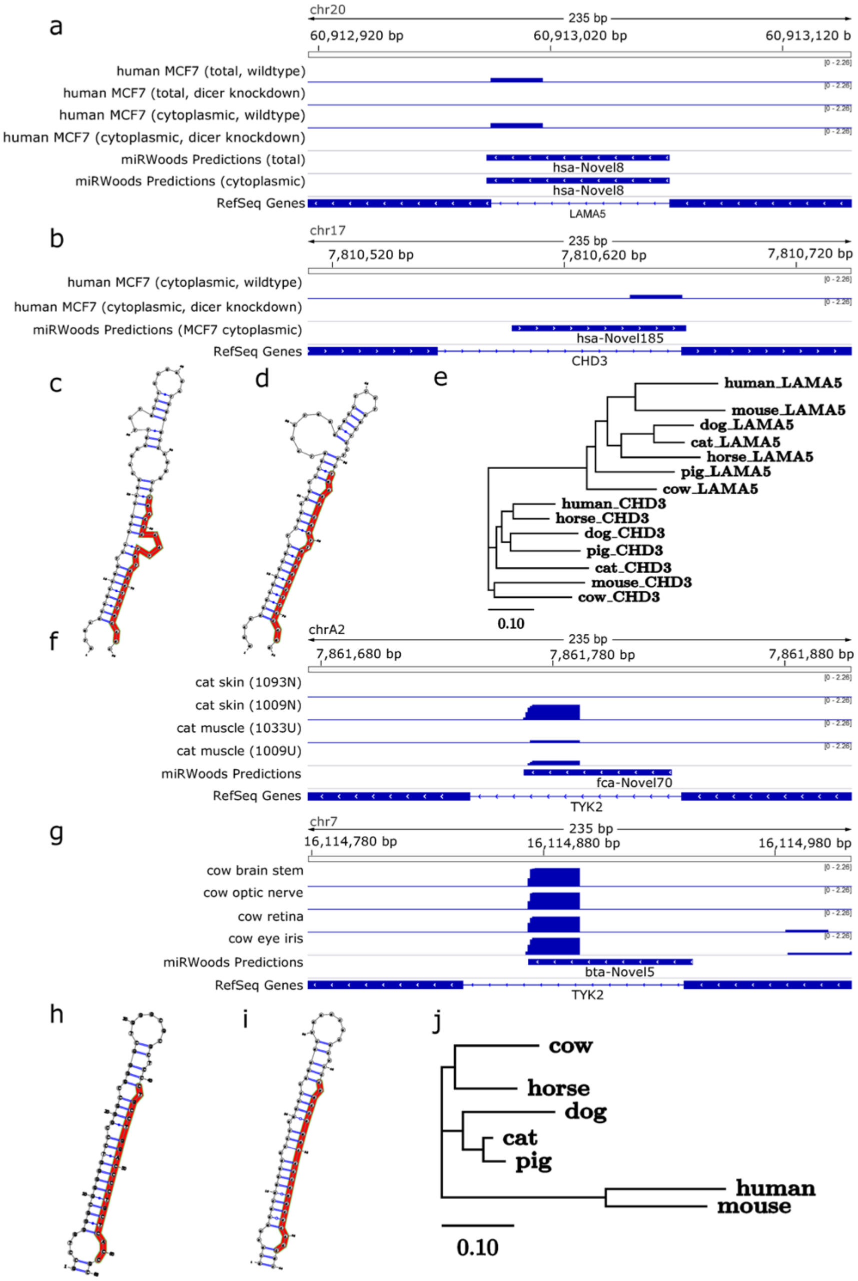

Figure 2.21: Novel microRNA families identified by miRWoods. **a** hsa-novel-8 is a mirtron predicted for both MCF-7 sets where expression was decreased in the Dicer knockdown sets. **b** hsa-Novel-185 is a mirtron predicted within the human cell lines set and the MCF-7 (cytoplasmic fraction) set. It also shows reduced expression in the Dicer knockdown version of the MCF-7 set. **c** The structure of hsa-novel-8. **d** The structure of hsa-Novel-185. **e** Phylogeny comparing the LAMA5 intron and CHD3 intron for several mammals. **f** Novel miR predicted in bovine genome in an intron of TYK2. **g** novel predicted miR in the feline genome in the same intron of TYK2 **h** structure of novel feline miR. **i** structure of novel bovine miR. Eight nucleotides were removed from the 5′ end, and two were added to the 3′ end to match the feline hairpin precursor boundaries. **j** A phylogeny comparing the TYK2 intron in several mammals.

3 Predicting Dicer and Drosha cleavage sites through deep learning.

3.1 Abstract

MicroRNAs are a highly conserved class of small endogenous RNA involved in post-transcriptional gene silencing and have prominent roles in disease and development. Since their discovery, many tools have been developed to identify Novel microRNA. However, few attempts have been made to predict on the individual processes of microRNA biogenesis. So far SVM based methods to predict Dicer cleavage sites have been developed but there's been no attempts to predict cleavage sites for Drosha or to identify cleavage sites using deep learning. Here, we present DeepMirCut, an LSTM based software designed to predict both Dicer and Drosha cleavage sites. We compare models trained on sequence data, sequence and fold data, and sequence and fold data with the labeled loops. Our results show that it still performs reasonably well on just sequence data but achieves the biggest gains when fold data is added. Labeling the loop resulted in the best performance. Experiments on our best model show that DeepMirCut is able predict cuts within closer average proximity than results reported for other methods. Point mutation plots for the sequence showed that A GU pair across the cleavage site on Dicers 3′ arm has a positive effect on DeepMirCut's prediction, while a UG pair has a detrimental effect. Point mutation plots created for the fold and labeled loops were also used to identify several positions where bulges had either positive or negative effects on the score.

3.2 Introduction

MicroRNAs (miRs) are a highly conserved class of small endogenous RNA around 22nt in length. Mature microRNAs modulate a variety of different processes through post-transcriptional gene silencing, which result in either transcript degradation or translational inhibition [1]. MicroRNAs have a wide range of functions including cancer (both tumor-suppressor and oncogenic) [2], development [3], stress response [4], aging [5], and circadian rhythms [6]. Nucleotide positions 2 through 8 on the mature microRNA are called the seed sequence and help direct the sequence-specific activity of the RNA-induced silencing complex (RISC), where it binds to a complementary strand on the 3′- UTR of an mRNA transcript.

Biogenesis of mature microRNAs begins with the transcription of a primary miRNA (pri-miRNA) transcript by RNA Polymerase II [7, 8], or in rare cases RNA Polymerase III [9]. A microprocessor complex associates with the hairpin, whereby the action of the component enzyme Drosha produces a double-stranded cleavage that results in the microRNA precursor (pre-miR) leaving a 2-nt overhang on the 3′ end [10, 11].. Exportin-5 associates with the 3′ overhang and transports the precursor from the nucleolus to the cytoplasm [12]. Here an enzyme known as Dicer removes the loop through an additional double-stranded cleavage. Taken together, the activity of these enzymes result in four distinct cleavage sites (here also called "cut-sites") of the pri-miRNA transcript and therefore result in a double-stranded miR:miR* duplex. Dicer passes the miR:miR* duplex to Argonaut, a core enzyme of RISC, which binds with only one of the strands while the other one is degraded.

While most microRNA tools have been developed for homologous and novel microRNA discovery, some have been developed to learn more about individual processes involved in microRNA biogenesis. PHDCleave is an SVM designed to identify Dicer cut-sites on a microRNA precursor [32]. While PHDCleave has decent sensitivity and specificity on their test set, when the SVM is applied in a sliding window across the entire precursor, the cut-site predictions are on average 3.1 nucleotides offset from the annotation [32]. LBSizeCleave is similar but adds features describing the length of loop and bulge structures [33]. LBSizeCleave performs with greater accuracy than PHDCleave at finding cleavage sites that were within 1nt of the annotated site, but has lower accuracy when more of an offset was allowed [33].

Deep learning approaches overcome the need for feature engineering by learning the features themselves. Several deep learning approaches such as convolutional neural networks (CNNs) [27] and recurrent neural networks (RNNs) [28, 29] have been used for microRNA classification. While these approaches have addressed the limitations of feature engineering, they only predict loci and don't perform cleavage site prediction. RNNs, such as Long Short-Term Memory (LSTM) networks, have been used in natural language processing applications such as named entity recognition [30] and part-of-speech tagging [31], which are similar tasks to cleavage site recognition. Motivated by the challenges of microRNA analysis and the success of deep learning applications for NLP, we present DeepMirCut, an LSTM-based algorithm that predicts Dicer and

Drosha cleavage sites within microRNAs. DeepMirCut predicts the locations of the four cut-sites of Drosha and Dicer from an input RNA sequence.

3.3 Results

For our analysis, we processed microRNA annotations from miRBase and genomic sequences to extract microRNA precursor sequences as well as 300-nt flanking genomic sequence. We refer to the precursor and flanking sequence as an “extended sequence”. Our data processing resulted in a collection of 34,714 extended sequences to be further refined to create datasets for training and testing.

For the current study, we focused on precursors from metazoan species having both mature microRNAs (5′ and 3′) annotated in miRBase, which consists of 11,296 records. Precursor sequences that were within a sequence identity threshold of 0.8 of other sequences were excluded from the set using CD-Hit [67] in order to ensure low similarity between the training, validation, and testing sets. A standard 80:10:10 split was used to produce a training set with 3,923 examples, validation set with 490 examples, and test set with 491 examples. To increase our training examples, we added back sequences that CD-Hit had identified as similar to those in the training set but were below the sequence identity threshold of the validation and testing sets, which increased the training set to 8,491 examples. We compared each sequence of the training set with sequences of the validation and testing sets to verify that an identity threshold of 0.8 was maintained for sequences between sets as demonstrated in Figure 3.1. Random amounts of buffer sequence between 30-nt and 50-nt were added to each of the precursors for the training, validation, and testing sets. An augmented training set with 84,910 examples and an augmented validation set with 4,900 examples was generated by randomly selecting buffer sequences 9 more times for each example (Figure 3.2a).

We trained three different sets of models defined by the type of input data that were used. First, model 1 was trained on only the extended RNA sequence. Second, model 2 was trained on sequence and fold. RNAfold [59] was used to predict the secondary structure of the extend RNA sequence, to provide the dot-bracket [68] sequence for each RNA within each of the train, test, and validation sets. Finally, for model 3, we further annotated the sequence using a modified bpRNA structure array [69] to provide context such as whether each nucleotide was on a bulge, internal loop, or hairpin loop.

DeepMirCut has the option of predicting cut-sites based on inputs with sequences only, sequences and folds, or sequences and labels indicating the context of the fold produced by bpRNA. The architecture includes an embedding dimension, a dropout layer, two bidirectional LSTM layers, and a time distributed dense layer. The time distributed layer outputs a set of 5 values for each nucleotide which represent weights for a Drosha cut on the 5′ arm (DR5), a Drosha cut on the 3′ arm (DR3), a Dicer cut on the 5′ arm (DC5), a Dicer cut on the 3′ arm (DC3), or no cut-site present (O). By default, DeepMirCut labels the position with the maximum weight for DR3, DR5, DC3, and DC5 as a cleavage site, but the O-sites are not labeled (Figure 3.2b.) Labeling is done in this way so that each cut-site will only be labeled once whereas labeling using that maximum weight at each position could result in cut-sites being labeled more than once or not at all. See Figure 3.3 for an example of DeepMirCut predicting cleavage sites for hsa-mir125a.

Hyper-parameters were tuned to identify the best parameter combinations for models trained using each of DeepMirCut's three input options. The top 10 architectures identified through tuning were evaluated with 20 replicates each to identify parameters resulting in the best median F-score (Figure 3.4,Figure 3.5, and Figure 3.6). All models were evaluated using the augmented validation set. The parameter combinations that showed the best performance during tuning are shown in Table 3.1.

Models trained on sequence only, sequence and fold, and sequence and bpRNA data were tested using the optimum parameter combinations for each type of input. Training with sequence and bpRNA-context resulted in the highest median F-score (F-score = 0.348), followed by models trained using the sequence and fold (F-score = 0.345). Surprisingly, DeepMirCut was still able to identify several cleavage sites using only the sequence as input (F-score = 0.183) (Figure 3.7a) We note that while the F-score was used to select models, it can be an unforgiving metric for this kind of task because predictions offset by 1 nt will be counted as a false prediction. Therefore, we further evaluated the mean distance between the predicted and annotated cut-sites for each replicate. Training with sequence and bpRNA-context resulted in the best median average distance (dist = 2.611), followed by models trained using the sequence and fold (dist=2.623), followed by models trained with just the sequence (dist = 4.943) (Figure 3.7b)

The best performing model based on the validation set was re-evaluated against the test set (F-score = 0.358; dist = 2.579). The model identified Dicer cleavage sites better on the 3′ arm (F-score = 0.43; dist = 2.214) than 5′ arm (F-score = 0.30; dist = 2.802) and identified Drosha cleavage sites better on the 5′ arm (F-score = 0.36; dist = 2.902) than the 3′ arm (F-score = 0.34; dist = 2.397). Most of DeepMirCut's predictions fell within one nucleotide of the annotated cleavage sites. Predictions of Dicer's cut along the 3′ arm result in both higher decision values and predictions within a closer proximity than all of the other cleavage sites. (Figure 3.7c-e.) These results indicate that DeepMirCut is better at finding cleavage sites at the 5′ ends of mature microRNA compared to the 3′ sites. A possible reason for this may be that isomirs are more likely occur at the 3′ end of microRNAs [70] making training and testing more difficult.

We performed a point-mutation analysis on the nucleotides surrounding each cut-site in order to assess what DeepMirCut is learning (Figure 3.8) The effect on scores from Dicers 3′ arm was the most pronounced. A GU pair across the cleavage site on Dicers 3′ arm has a positive effect on DeepMirCuts prediction while a UG pair appears to have the opposite effect. Uracil is the top logo character appearing 1 nt downstream from the cleavage site so DeepMirCut is likely to identify the presence of Uracil as an important feature to base its predictions on. (Figure 3.8b.)

A point mutation-analysis was also performed on the fold and contextual information as well (Figure 3.9) A bulge occurring 3 nt upstream has a positive influence on DeepMirCuts ability to identify Dicers cleavage site on the 3′ arm (Figure 3.9d). On the 5′ arm prediction performance improved when a bulge was present 1 nt downstream but not 1nt-2nt upstream from Dicers cleavage site (Figure 3.9b).

We tested the performance of DeepMirCut on precursors with only one annotated microRNA by generating sets with the annotated microRNA either on the 3′ arm or the 5′ arm. Cleavage sites on the arm opposite to the annotated microRNAs were assumed based on read stacks from small RNA Sequencing data (see Methods) and were used to assess DeepMirCut's performance. DeepMirCut identified unannotated Dicer cleavage sites on the 3′ arm (F-score = 0.53; dist = 1.316) than on the 5′ arm (F-score = 0.36; dist = 3.273) and identified unannotated Drosha cleavage sites on the 5′ arm (F-score = 0.32; dist = 5.00) better than the 3′ arm (F-score = 0.11: dist = 2.105)

3.4 Discussion

We report on the training, testing and evaluation of DeepMirCut for the site-labeling of Dicer and Drosha cleavage sites. While other similar prediction programs only predict Dicer cleavage sites on the folded precursor sequences, DeepMirCut predicts both Dicer and Drosha cleavage sites on full-length extended precursor sequences that include flanking sequence of randomly-sampled length. Our experiments with annotations from miRBase show that DeepMirCut predicts cleavage sites with close average proximity, and more closely predicts Dicer sites than results reported for other methods.

Central to microRNA function is the seed sequence, which is necessary for RISC to target specific mRNAs and is defined relative to the 5′ end of the mature microRNA. Consistent with these functional requirements, we observed that DeepMirCut performed better at the identification of cleavage sites that correspond to the 5′ ends of mature microRNAs compared to their 3′ ends. These data support the idea that sequence and structural information are sufficient to locate these cleavage sites.

We trained and evaluated different versions of DeepMirCut that incorporate varying levels of input information, including sequence, sequence and structure, and sequence, structure, and loop information. We found that DeepMirCut can predict moderately well based on sequence alone, suggesting it is not completely relying on structural information about the loop for its predictions. This is consistent with the fact that point-mutation analysis reveals strong changes in score due to perturbations to sequence alone. We further added structural information in the form of predicted secondary structure dot-bracket sequences and bpRNA-computed loop-type labels, which significantly improved the performance. We observed the greatest improvement in prediction performance with the addition of the dot-bracket sequence, which provides information regarding the presence and absence of base pairs. It has been shown biochemically that the hairpin loop position [71] and the locations of bulges and other unpaired nucleotides [72] may help direct the function of Dicer. This fact may explain the modest improvement we observed with the addition of loop-type labels, such as distinguishing bulges from internal loops.

3.5 Methods

3.5.1 Data Preprocessing

All microRNA GFF annotations files were downloaded for miRBase v22.1, and then used to locate precursor sequences within the genomic context for each species. Genome FASTA files were downloaded from various sources including NCBI Assembly and organism-specific genome resources when needed. Precursor sequences were extracted from each genome along with a buffer sequence extending 300nt upstream to 300nt downstream. Cleavage-sites were determined by folding the original precursor sequences found on miRBase and determining which arm each mature microRNA was on. Examples where microRNA overlap the loop were removed to avoid ambiguity in cleavage-sites corresponding to each microRNA. In several cases either the name or location of the miR was inconsistent between the miRBase GFFs and the miRBase FASTA files and in a few cases defunct miRs were present in the miRBase GFFs. In order to improve testability, microRNAs were dropped whenever there was a naming inconsistency between GFF and FASTA files or an inconsistency between the annotation and genomic sequence (70 loci in total). Sequences from *Brassica napus*, *Schistosoma japonicum*, *Schmidtea mediterranea*, and *Triticum aestivum* were excluded from the set because of difficulties in finding versions of the genome that corresponded to locations of each sequence within the miRBase GFF files.

3.5.2 Hyperparameter Tuning

Hyperparameters for three different models were tuned using a training set composed of either training data, sequence and fold data, or sequence an bpRNA data. Hyperopt [73] was used to search for a model producing an optimal F-score with an embedding dropout between 0 and 0.5, an embedding dimension of 32, 64, 96, 128, or 160 units, a first bidirectional LSTM layer with 64, 128, 192, 256, or 320 units, a second bidirectional LSTM layer with 0, 64, 128, 192, 256, or 320 units, a learning rate for the adam optimizer between 0.00001 and 0.1, and an epsilon between 10^{-10} and 10^{-4}. The top 10 models identified by hyperopt were retrained 20 times and a model for each of the three training sets was chosen based on F-score and consistency.

3.5.3 Point Mutation Analysis

In order to determine what DeepMirCut is learning about nucleotide sequences, every possible mutation from -10 nt upstream to -10nt downstream was generated for cleavage sites within the test set. DeepMirCut was run on the mutated and unmutated datasets and decision values were converted into scores using the logit function. The mean difference between scores for mutated nucleotides vs scores for nucleotides in the unmuted set was used to evaluate the effects that mutations would have on the model’s ability to predict cleavage sites.

3.5.4 Identification of Unannotated mature microRNAs

In order to test DeepMirCuts performance on microRNAs with only one annotated mature microRNA, wildtype MCF-7 total cell content (GSE31069) and MCF-7 cell fractions (GSE31069) were downloaded from GEO [41] A script called miRPreprocess from miRWoods [74] was run in order to group read stacks and identify unannotated microRNAs which would be used to assume positions of unknown cleavage sites.

Test sets were grouped into those with annotated mature microRNAs occurring on either the 5′ arm or the 3′ arm. Cases where multiple hairpin precursors had identical sequences were filtered down to one example prior to test time. DeepMirCut predicted cleavage sites on each set and positions identified from read stacks were used to evaluate performance.

3.6 Tables

Table 3.1: Tuned Parameters

Input Type	Sequence Only	Sequence & Fold	Sequence & Context
Embedding layer	96 units	32 units	32 units
Dropout	0.315	0.213	0.247
Bi-LSTM layer 1	320 units	64 units	64 units
Bi-LSTM layer 2	192 units	256 units	320 units
Learning rate	$3.2 * 10^{-3}$	$1.91 * 10^{-3}$	$2.64 * 10^{-3}$
Epsilon (10^{x})	-7.56	-6.79	-6.66

3.7 Figures

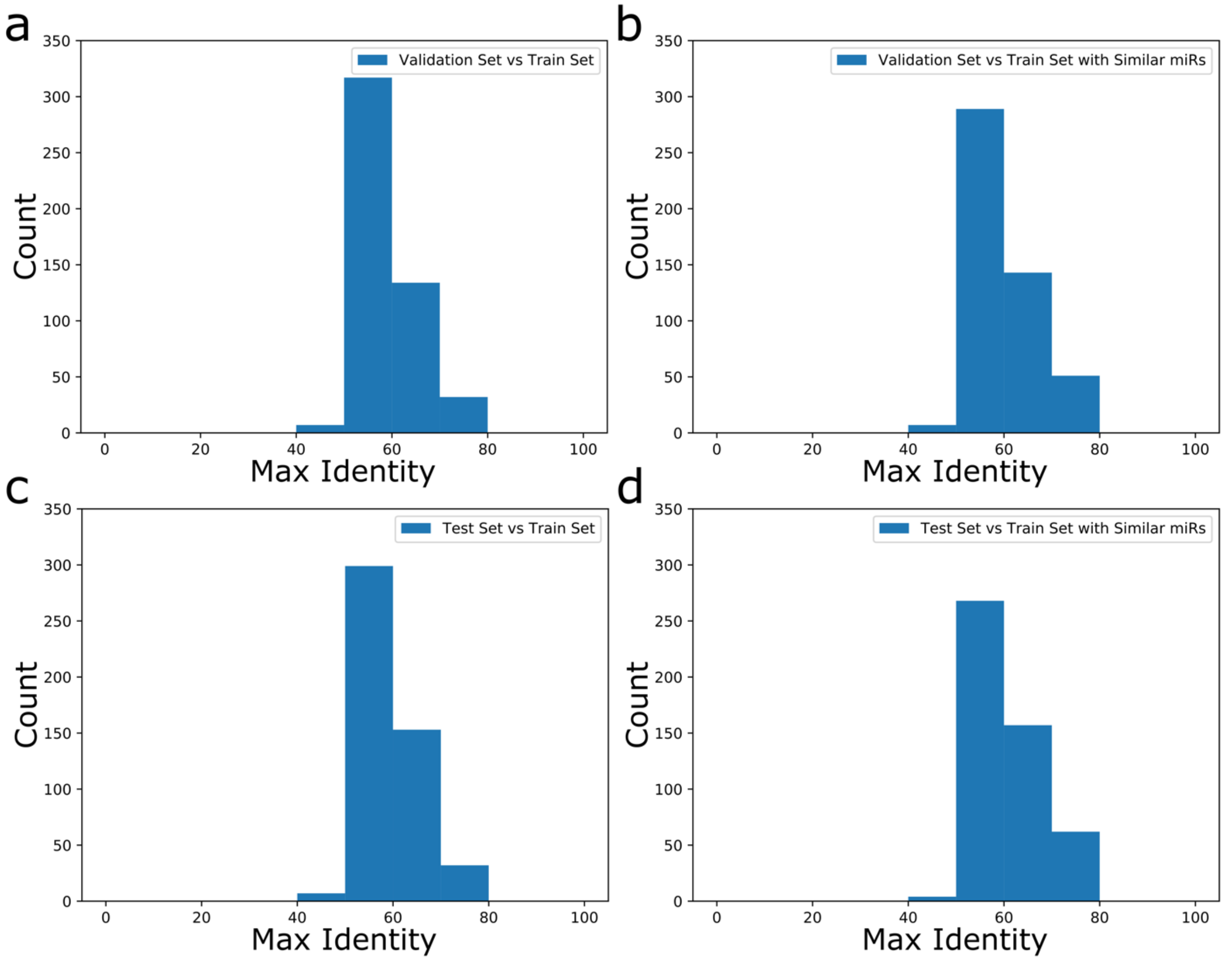

Figure 3.1: Verification of maximum identity threshold between sets. Histograms of maximum global identity comparing sequences of **a** validation set vs train set **b** validation set vs train set with similar sequences **c** test set vs train set **d** test set vs train set with similar sequences.

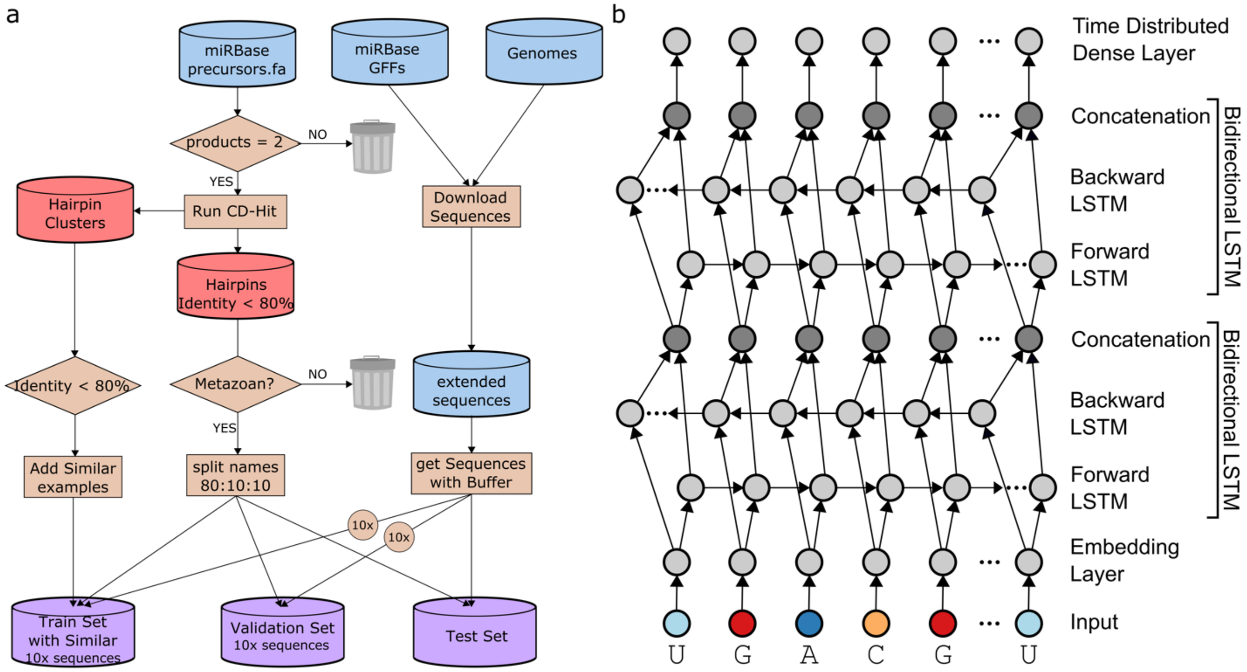

Figure 3.2: Data Processing and Architecture. **a** Flowchart describing the generation of the train, validation, and test sets. **b** Diagram showing DeepMirCut's architecture.

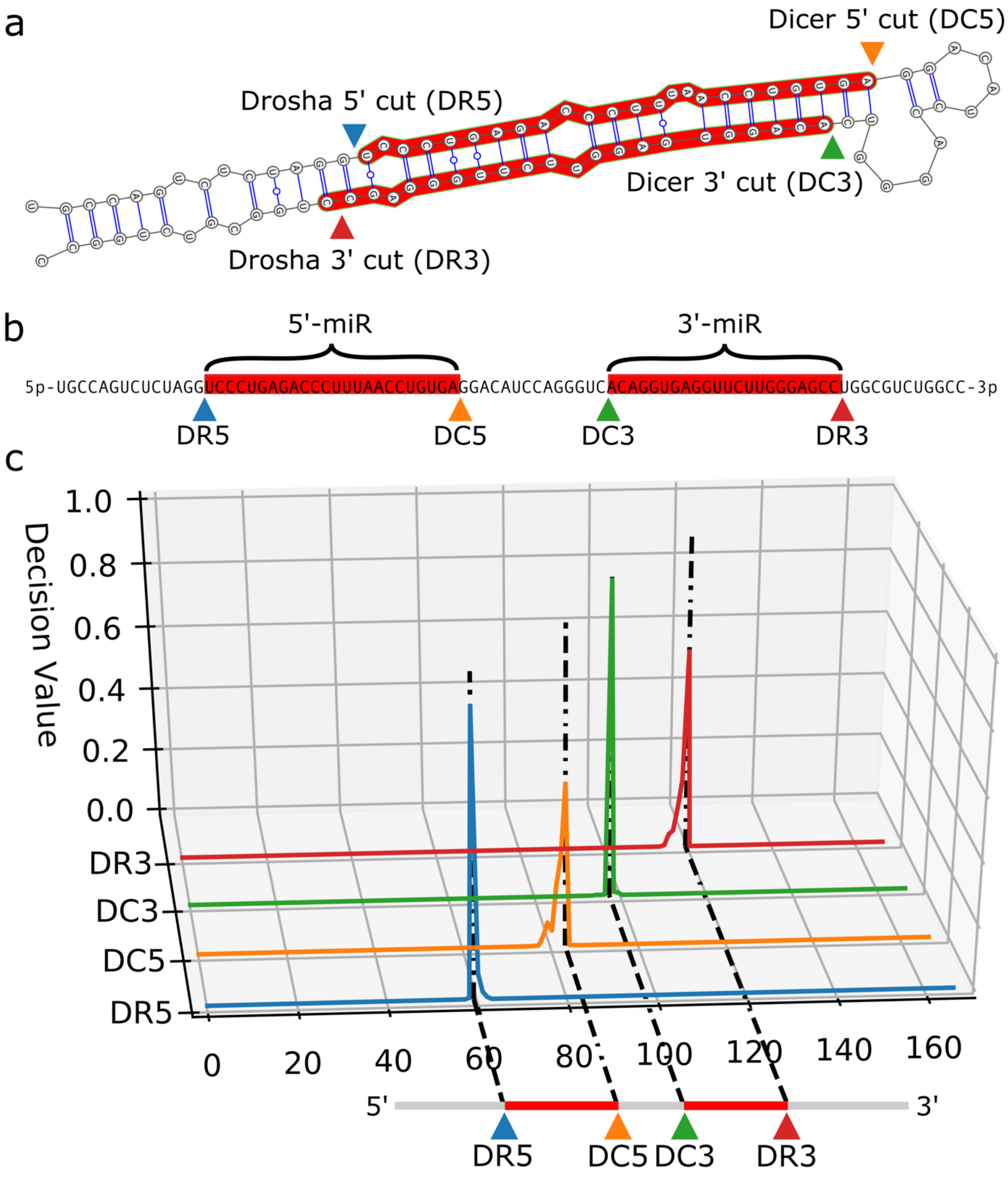

Figure 3.3: Example cleavage site identification for hsa-mir-125a. **a** Dicer and Drosha cleavage sites surrounding the two mature microRNA sequences highlighted in red. **b** An unfolded version of the precursor sequence. **c** Three-dimensional line plot showing decision values for each Dicer and Drosha cleavage site in hsa-mir-125a.

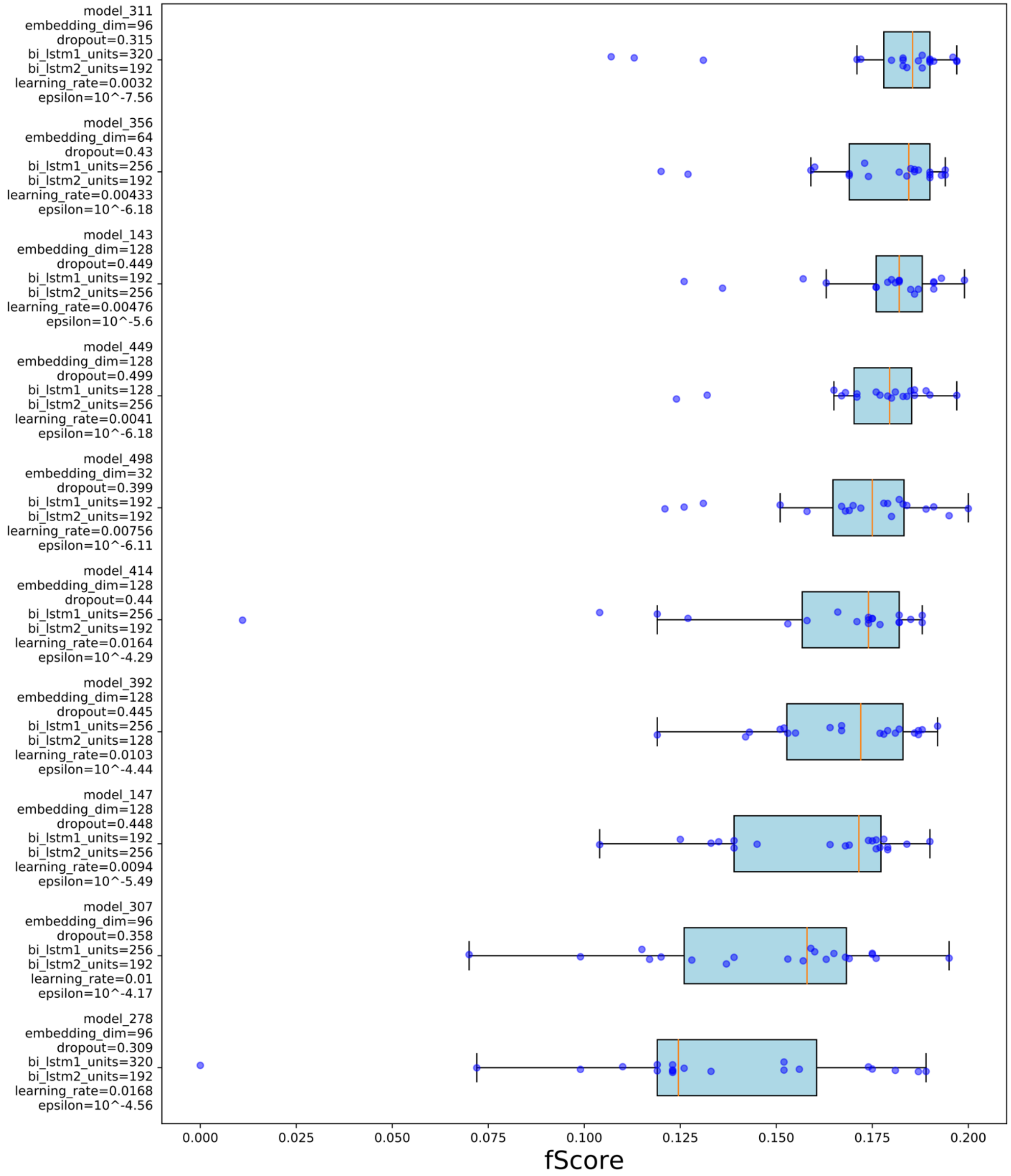

Figure 3.4: Top ten models for models trained with sequence data. Box plot shows F-Score and for each replicate and parameters used in training.

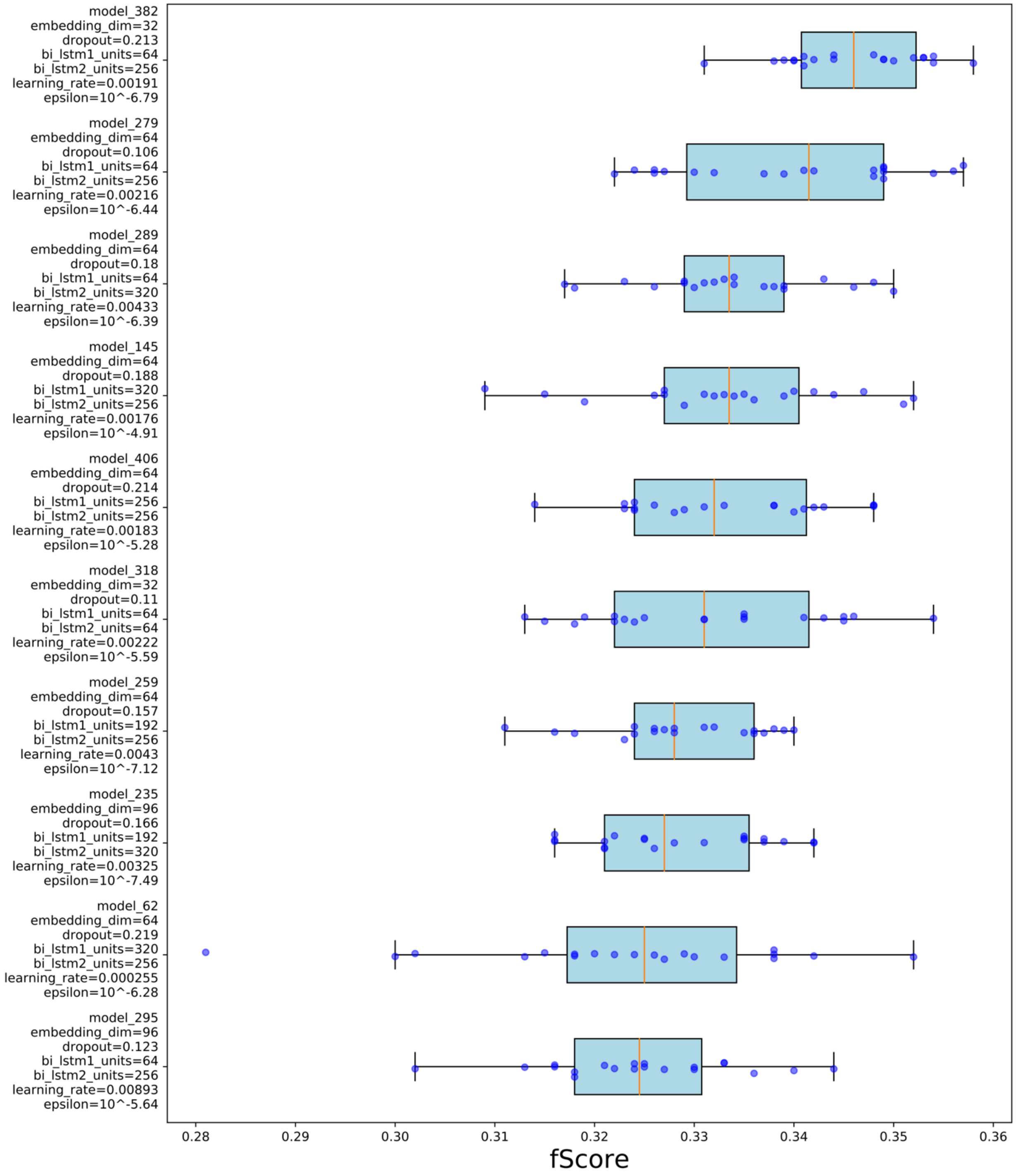

Figure 3.5: Top ten models for models trained with sequence and fold data. Box plot shows F-Score and for each replicate and parameters used in training.

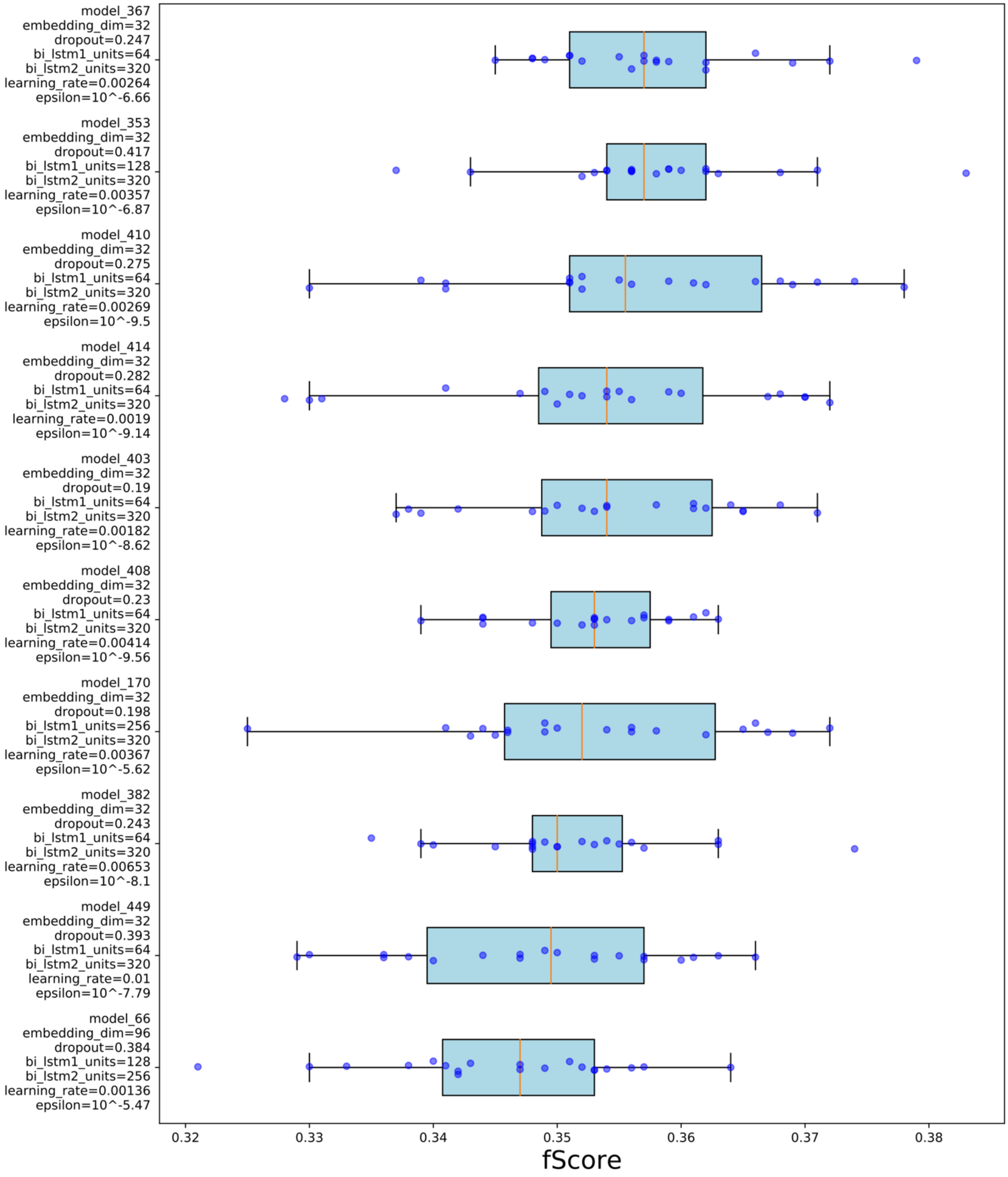

Figure 3.6: Top ten models for models trained with sequence and bpRNA provided contextual information. Box plot shows F-Score and for each replicate and parameters used in training.

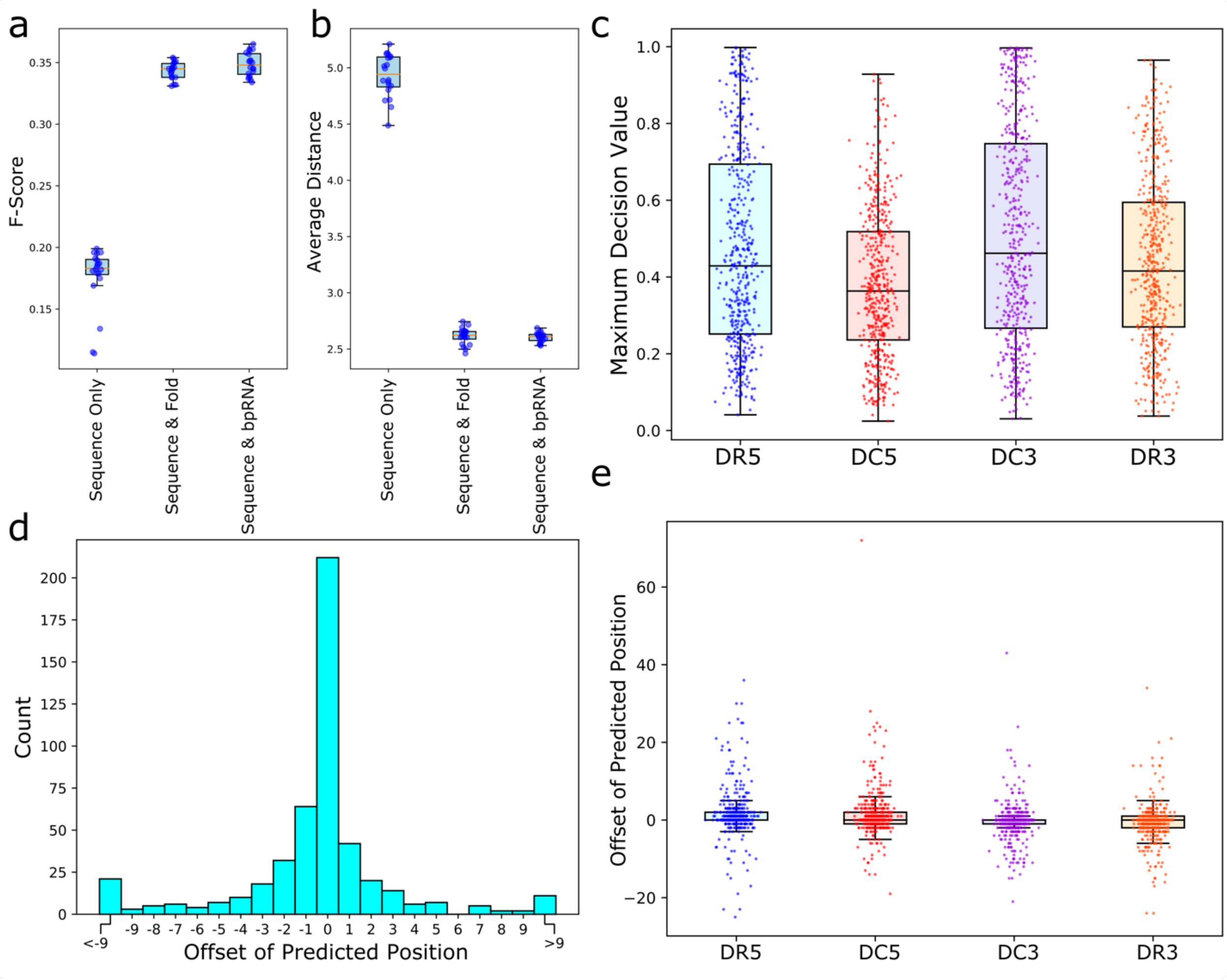

Figure 3.7: Performance with different inputs and test results for chosen model. **a** Boxplots comparing F-scores and **b** average distance of predicted cleavage sites from annotated positions for models generated using optimum parameters for each type of input. **c** Boxplot showing the difference in cleavage site decision values for a model chosen from the set of models trained to use sequence and bpRNA fold context. **d** Differences in position between predicted and annotated cleavage sites for Dicers cut on the 3′ arm and **e** a Boxplot showing differences between all predicted and annotated cleavage sites for the chosen model.

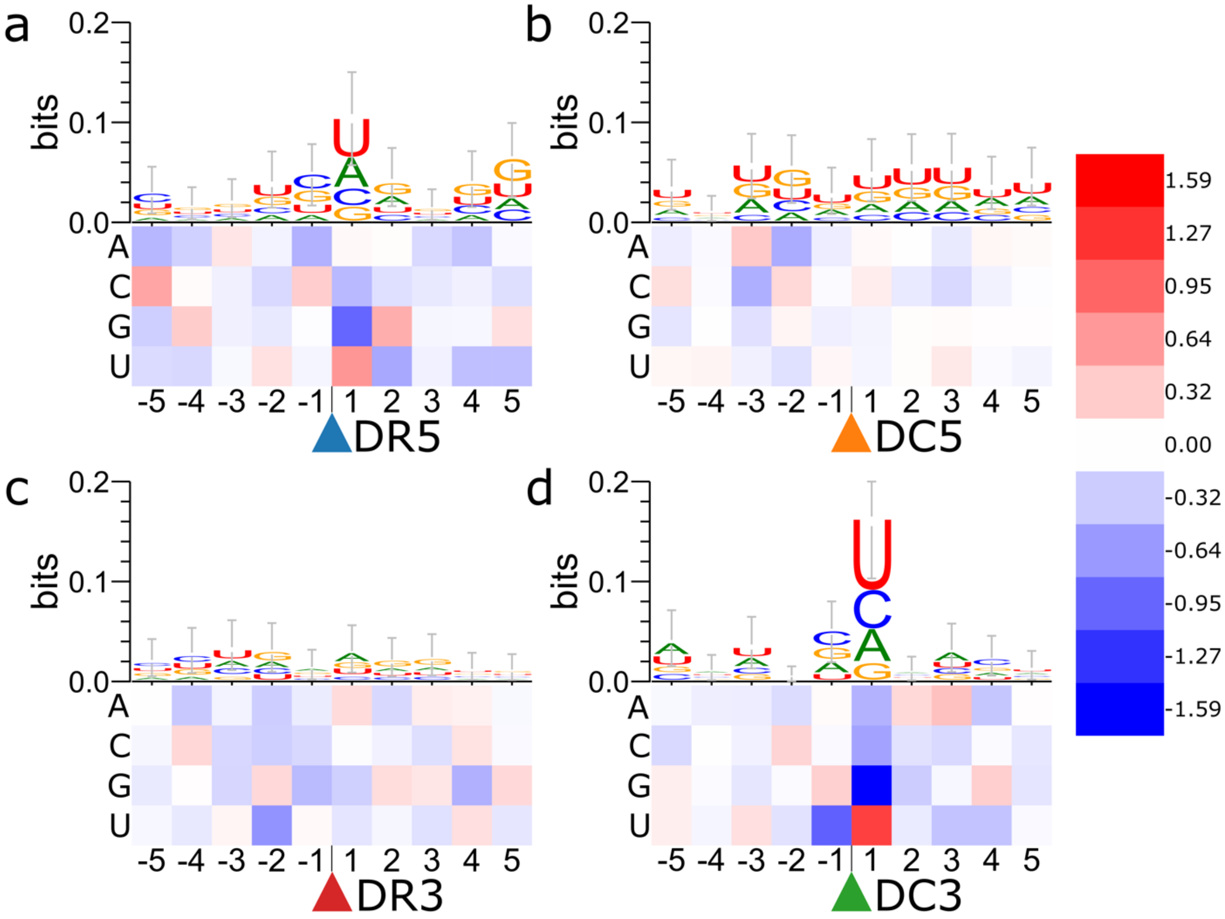

Figure 3.8: Point mutation analysis for nucleotides. Heatmaps showing point mutations for nucleotides surrounding cleavage sites of **a** Dicer on the 5′ arm, **b** Dicer on the 3′ arm, **c** Drosha on the 5′ arm, and **d** Drosha on the 3′ arm.

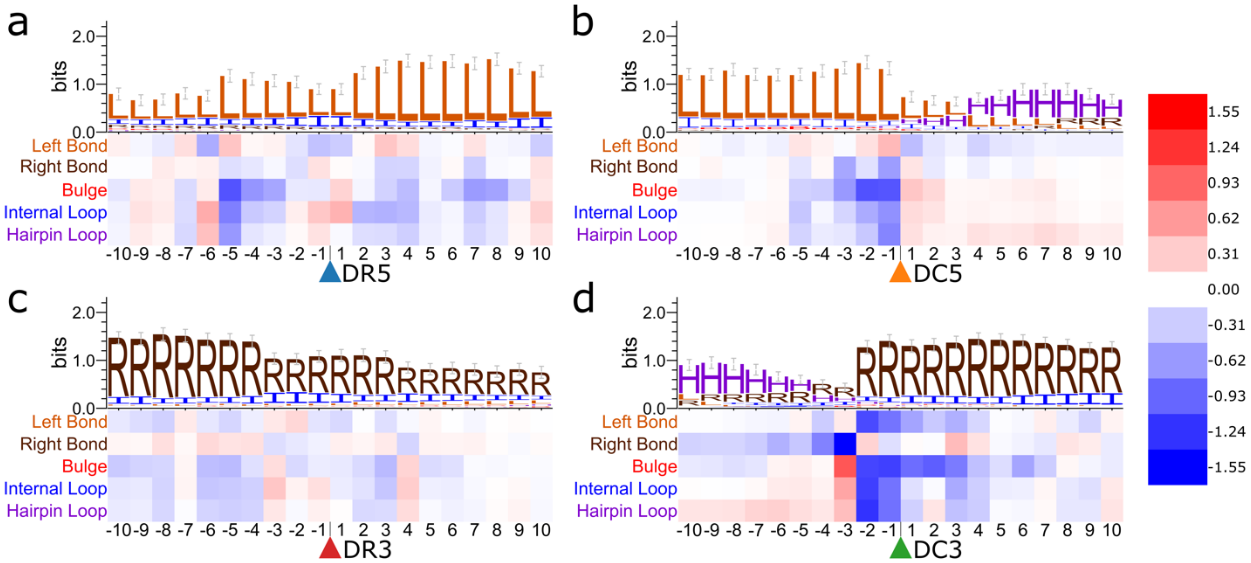

Figure 3.9: Point mutation analysis for folds. Heatmaps showing point mutations for folds and their bpRNA provided loop context surrounding cleavage sites of **a** Dicer on the 5′ arm, **b** Dicer on the 3′ arm, **c** Drosha on the 5′ arm, and **d** Drosha on the 3′ arm.

4 Conclusion

In conclusion, we have presented two software approaches to address several current challenges in microRNA discovery and analysis. First, miRWoods provides a new approach to the detection of valid precursor spans using a duplex-focused method. We also demonstrated that miRWoods was capable of finding microRNA loci with a single read. Second, DeepMirCut provides a deep learning-based method for the prediction of Dicer and Drosha cleavage sites.

One of the core assumptions of molecular biology is that sequence determines structure, which determines function. microRNA biogenesis depends on the proper integration of biological sequence information by the enzymes involved. Our two methods highlight the importance of RNA sequence and secondary structure information in the prediction and characterization of microRNAs. miRWoods uses engineered features based on deep sequencing read abundance and distribution, RNA sequence, and predicted secondary structure, to predict loci. The fact that it predicts loci with only one read, suggests that sequence and structure information can take precedence over read abundance in some cases. DeepMirCut demonstrates that deep learning can infer patterns in input sequence and structural information without any engineered features. Taken together, these observations support the idea that the RNA sequence information that directs microRNA biogenesis is encapsulated by our computational methods.

miRWoods uses dinucleotides as a feature, but without considering context or position. Therefore, one might speculate whether this sequence information is a remnant of evolutionary ancestry or part of purposeful patterns directing their processing and biogenesis. Our deep learning-based method DeepMirCut integrates sequence information in a context-specific way, suggesting that proper context of sequence information is necessary for microRNA function. It is possible that miRWoods learned dinucleotide patterns similar to what DeepMirCut learned, albeit out of context in some situations.

This work has focused primarily on microRNAs from Metazoa. Future work could investigate whether our approaches would be applicable to plant microRNAs. While plant microRNAs have many similarities there are some key differences. For instance, plants have an enzyme called Dicer-like1 (DL1), which performs both of the actions of Dicer and Drosha [75]. Analyzing the effects of point mutations may also provide insight into differences in structural features that DL1 might recognize along either cut.

DeepMirCut and miRWoods were both trained differently and applied to different tasks. DeepMirCut learns patterns in sequential data while miRWoods bases its prediction on features related to dinucleotide content, read distribution, and folding structure. It is possible that future work could combine these two approaches for synergistic effect. For example, the output of DeepMirCut could be used as additional features to miRWoods. The output of DeepMirCut when applied to non-miRs could provide a "signature" that could be detected to improve microRNA the microRNA prediction of miRWoods.

Overall, we have made substantial improvements to the analysis of microRNAs. Our work has already identified errors in current microRNA annotations, and it would be used in the future to further refine our microRNA annotations, ultimately leading to a greater understanding of their biogenesis and processing.

5 Bibliography

1. Liu J. Control of protein synthesis and mRNA degradation by microRNAs. Current opinion in cell biology. 2008;20(2):214-21.

2. Zhang B, Pan X, Cobb GP, Anderson TA. microRNAs as oncogenes and tumor suppressors. Developmental biology. 2007;302(1):1-12.

3. Carrington JC, Ambros V. Role of microRNAs in plant and animal development. Science. 2003;301(5631):336-8.

4. Leung AK, Sharp PA. MicroRNA functions in stress responses. Molecular cell. 2010;40(2):205-15.

5. Smith-Vikos T, Slack FJ. MicroRNAs and their roles in aging. Journal of cell science. 2012;125(1):7-17.

6. Na Y-J, Sung JH, Lee SC, Lee Y-J, Choi YJ, Park W-Y, et al. Comprehensive analysis of microRNA-mRNA co-expression in circadian rhythm. Experimental & molecular medicine. 2009;41(9):638-47.

7. Lee Y, Kim M, Han J, Yeom KH, Lee S, Baek SH, et al. MicroRNA genes are transcribed by RNA polymerase II. The EMBO journal. 2004;23(20):4051-60.

8. Zhou X, Ruan J, Wang G, Zhang W. Characterization and identification of microRNA core promoters in four model species. PLoS computational biology. 2007;3(3).

9. Borchert GM, Lanier W, Davidson BL. RNA polymerase III transcribes human microRNAs. Nature structural & molecular biology. 2006;13(12):1097-101.

10. Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, et al. The nuclear RNase III Drosha initiates microRNA processing. Nature. 2003;425(6956):415-9.

11. Gregory RI, Yan K-p, Amuthan G, Chendrimada T, Doratotaj B, Cooch N, et al. The Microprocessor complex mediates the genesis of microRNAs. Nature. 2004;432(7014):235-40.

12. Yi R, Qin Y, Macara IG, Cullen BR. Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. Genes & development. 2003;17(24):3011-6.

13. Bortolomeazzi M, Gaffo E, Bortoluzzi S. A survey of software tools for microRNA discovery and characterization using RNA-seq. Briefings in bioinformatics. 2019;20(3):918-30.

14. Ambros V, Bartel B, Bartel DP, Burge CB, Carrington JC, Chen X, et al. A uniform system for microRNA annotation. Rna. 2003;9(3):277-9.

15. Mendes ND, Freitas AT, Sagot M-F. Current tools for the identification of miRNA genes and their targets. Nucleic acids research. 2009;37(8):2419-33.

16. Friedländer MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, et al. Discovering microRNAs from deep sequencing data using miRDeep. Nature biotechnology. 2008;26(4):407-15.

17. Yousef M, Showe L, Showe M. A study of microRNAs in silico and in vivo: bioinformatics approaches to microRNA discovery and target identification. The FEBS journal. 2009;276(8):2150-6.

18. Akhtar MM, Micolucci L, Islam MS, Olivieri F, Procopio AD. Bioinformatic tools for microRNA dissection. Nucleic acids research. 2016;44(1):24-44.

19. Lai EC, Tomancak P, Williams RW, Rubin GM. Computational identification of DrosophilamicroRNA genes. Genome biology. 2003;4(7):R42.

20. Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S, Rhoades MW, et al. The microRNAs of Caenorhabditis elegans. Genes & development. 2003;17(8):991-1008.

21. Hendrix D, Levine M, Shi W. miRTRAP, a computational method for the systematic identification of miRNAs from high throughput sequencing data. Genome biology. 2010;11(4):R39.

22. Mathelier A, Carbone A. MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data. Bioinformatics. 2010;26(18):2226-34.

23. Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic acids research. 2012;40(1):37-52.

24. Chen X, Li Q, Wang J, Guo X, Jiang X, Ren Z, et al. Identification and characterization of novel amphioxus microRNAs by Solexa sequencing. Genome biology. 2009;10(7):R78.

25. Hackenberg M, Sturm M, Langenberger D, Falcon-Perez JM, Aransay AM. miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. Nucleic acids research. 2009;37(suppl_2):W68-W76.

26. Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. IEEE transactions on neural networks. 1994;5(2):157-66.

27. Do BT, Golkov V, Gürel GE, Cremers D. Precursor microRNA identification using deep convolutional neural networks. BioRxiv. 2018:414656.

28. Park S, Min S, Choi H, Yoon S. deepMiRGene: Deep neural network based precursor microrna prediction. arXiv preprint arXiv:160500017. 2016.

29. Cao M, Li D, Lin Z, Niu C, Ding C, editors. MiRNN: An Improved Prediction Model of MicroRNA Precursors Using Gated Recurrent Units. International Conference on Intelligent Computing; 2018: Springer.

30. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural architectures for named entity recognition. arXiv preprint arXiv:160301360. 2016.

31. Wang P, Qian Y, Soong FK, He L, Zhao H. Part-of-speech tagging with bidirectional long short-term memory recurrent neural network. arXiv preprint arXiv:151006168. 2015.

32. Ahmed F, Kaundal R, Raghava GP, editors. PHDcleav: a SVM based method for predicting human Dicer cleavage sites using sequence and secondary structure of miRNA precursors. BMC bioinformatics; 2013: BioMed Central.

33. Bao Y, Hayashida M, Akutsu T. LBSizeCleav: improved support vector machine (SVM)-based prediction of Dicer cleavage sites using loop/bulge length. BMC bioinformatics. 2016;17(1):487.

34. Shi W, Hendrix D, Levine M, Haley B. A distinct class of small RNAs arises from pre-miRNA–proximal regions in a simple chordate. Nature structural & molecular biology. 2009;16(2):183.

35. Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic acids research. 2011;40(1):37-52.

36. Yang X, Li L. miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants. Bioinformatics. 2011;27(18):2614-5.

37. An J, Lai J, Lehman ML, Nelson CC. miRDeep*: an integrated application tool for miRNA identification from RNA sequencing data. Nucleic acids research. 2012;41(2):727-37.

38. Sheng Y, Engström PG, Lenhard B. Mammalian microRNA prediction through a support vector machine model of sequence and structure. PloS one. 2007;2(9).

39. Tempel S, Zerath B, Zehraoui F, Tahi F. miRBoost: boosting support vector machines for microRNA precursor classification. RNA. 2015;21(5):775-85.

40. Jiang P, Wu H, Wang W, Ma W, Sun X, Lu Z. MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. Nucleic acids research. 2007;35(suppl_2):W339-W44.

41. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, et al. NCBI GEO: archive for functional genomics data sets—10 years on. Nucleic acids research. 2010;39(suppl_1):D1005-D10.

42. Raghavan V, Bollmann P, Jung GS. A critical investigation of recall and precision as measures of retrieval system performance. ACM Transactions on Information Systems (TOIS). 1989;7(3):205-29.

43. Davis J, Goadrich M, editors. The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning; 2006.

44. Sun J-z, Wang J, Wang S, Yuan D, Li Z, Yi B, et al. MicroRNA miR-320a and miR-140 inhibit mink enteritis virus infection by repression of its receptor, feline transferrin receptor. Virology journal. 2014;11(1):210.

45. Laganà A, Dirksen WP, Supsavhad W, Yilmaz AS, Ozer HG, Feller JD, et al. Discovery and characterization of the feline miRNAome. Scientific reports. 2017;7(1):1-14.

46. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome biology. 2014;15(2):R29.

47. Lawless N, Vegh P, O'Farrelly C, Lynn DJ. The role of microRNAs in bovine infection and immunity. Frontiers in immunology. 2014;5:611.

48. Lawless N, Foroushani AB, McCabe MS, O'Farrelly C, Lynn DJ. Next generation sequencing reveals the expression of a unique miRNA profile in response to a gram-positive bacterial infection. PloS one. 2013;8(3).

49. Sakurai D, Uchida R, Ihara F, Kunii N, Nakagawa T, Chazono H, et al. Immunosuppressive property of submandibular lymph nodes in patients with head and neck tumors: differential distribution of regulatory T cells. BMC research notes. 2018;11(1):479.

50. Gu S, Jin L, Zhang F, Huang Y, Grimm D, Rossi JJ, et al. Thermodynamic stability of small hairpin RNAs highly influences the loading process of different mammalian Argonautes. Proceedings of the National Academy of Sciences. 2011;108(22):9208-13.

51. Bao H, Kommadath A, Sun X, Meng Y, Arantes AS, Plastow GS, et al. Expansion of ruminant-specific microRNAs shapes target gene expression divergence between ruminant and non-ruminant species. BMC genomics. 2013;14(1):609.

52. Piriyapongsa J, Jordan IK. A family of human microRNA genes from miniature inverted-repeat transposable elements. PloS one. 2007;2(2).

53. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet journal. 2011;17(1):10-2.

54. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology. 2009;10(3):R25.

55. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.

56. Wootton JC, Federhen S. [33] Analysis of compositionally biased regions in sequence databases. Methods in enzymology. 266: Elsevier; 1996. p. 554-71.

57. Morin RD, O'Connor MD, Griffith M, Kuchenbauer F, Delaney A, Prabhu A-L, et al. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. Genome research. 2008;18(4):610-21.

58. Zuker M, Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic acids research. 1981;9(1):133-48.

59. Lorenz R, Bernhart SH, Zu Siederdissen CH, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms for molecular biology. 2011;6(1):26.

60. Ruby JG, Stark A, Johnston WK, Kellis M, Bartel DP, Lai EC. Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. Genome research. 2007;17(12):1850-64.

61. Ma H, Wu Y, Niu Q, Zhang J, Jia G, Manjunath N, et al. A sliding-bulge structure at the Dicer processing site of pre-miRNAs regulates alternative Dicer processing to generate 5'-isomiRs. Heliyon. 2016;2(9):e00148.

62. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic acids research. 2013;42(D1):D68-D73.

63. Bustin SA, Beaulieu J-F, Huggett J, Jaggi R, Kibenge FS, Olsvik PA, et al. MIQE precis: Practical implementation of minimum standard guidelines for fluorescence-based quantitative real-time PCR experiments. BioMed Central; 2010.

64. Peltier HJ, Latham GJ. Normalization of microRNA expression levels in quantitative RT-PCR assays: identification of suitable reference RNA targets in normal and cancerous human solid tissues. Rna. 2008;14(5):844-52.

65. Ruijter J, Ramakers C, Hoogaars W, Karlen Y, Bakker O, Van den Hoff M, et al. Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data. Nucleic acids research. 2009;37(6):e45-e.

66. Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, et al. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome biology. 2002;3(7):research0034. 1.

67. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28(23):3150-2.

68. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. Fast folding and comparison of RNA secondary structures. Monatshefte für Chemie/Chemical Monthly. 1994;125(2):167-88.

69. Danaee P, Rouches M, Wiley M, Deng D, Huang L, Hendrix D. bpRNA: large-scale automated annotation and analysis of RNA secondary structure. Nucleic acids research. 2018;46(11):5381-94.

70. Neilsen CT, Goodall GJ, Bracken CP. IsomiRs–the overlooked repertoire in the dynamic microRNAome. Trends in Genetics. 2012;28(11):544-9.

71. Gu S, Jin L, Zhang Y, Huang Y, Zhang F, Valdmanis PN, et al. The loop position of shRNAs and pre-miRNAs is critical for the accuracy of Dicer processing in vivo. Cell. 2012;151(4):900-11.

72. Feng Y, Zhang X, Graves P, Zeng Y. A comprehensive analysis of precursor microRNA cleavage by human Dicer. Rna. 2012;18(11):2083-92.

73. Bergstra J, Yamins D, Cox DD. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. 2013.

74. Bell J, Larson M, Kutzler M, Bionaz M, Löhr CV, Hendrix D. miRWoods: Enhanced precursor detection and stacked random forests for the sensitive detection of microRNAs. PLoS computational biology. 2019;15(10).

75. Kurihara Y, Watanabe Y. Arabidopsis micro-RNA biogenesis through Dicer-like 1 protein functions. Proceedings of the National Academy of Sciences. 2004;101(34):12753-8.