The evolution of microRNA in primates

Jennifer C. McCreight

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Willie J. Swanson, Chair

Christine Queitsch

Walter L. Ruzzo

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

The evolution of microRNA in primates

Jennifer C. McCreight

Chair of the Supervisory Committee:
Professor Willie J. Swanson
Genome Sciences

MicroRNA play an important role in post-transcriptional regulation of most transcripts in the human genome, but their evolution within humans and across the primate lineage is largely uncharacterized. A particular miRNA can have one to thousands of messenger RNA targets, establishing the potential for a small change in sequence or overall miRNA structure to have profound phenotypic effects. However, the majority of non-human primate miRNA is predicted solely by homology to the human genome and lacks experimental validation. In the present study, we sequenced thirteen species representing a wide range of the primate phylogeny. Hundreds of miRNA were validated, and the number of species with experimentally validated miRNA was tripled. These species include a sister taxon to humans (bonobo) and basal primates (aye-aye, mouse lemur, galago). Consistent with previous studies, we found the seed region and mature

miRNA to be highly conserved across primates, with overall structural conservation of the pre-miRNA hairpin. However, there were a number of interesting exceptions, including a seed shift due to structural changes in miR-501 and an increase in the number of miR-320 paralogs throughout primate evolution. We also identified 4521 SNVs within diverse human populations from the 1000 Genomes Project, again finding the seed region and mature miRNA to be most highly conserved, even among common variants. No variants exhibited population substructure and most were very rare, suggesting that purifying selection has been the driving force for human miRNA evolution. The conservation of human miRNA and the enriched regulation of neuronal processes among miRNA targets of non-conserved non-human primates illustrate the importance of investigating the miRNA of more distantly related primate species in order to learn more about human evolution.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

The love and support of many people made this dissertation possible.

Thank you to all members of the Swanson Lab past and present for the years of advice and feedback: Jan Aagaard, Katrina Claw, Joe Gasper, Renee George, Emily Killingbeck, Melody Palmer, and Damien Wilburn. I especially want to thank my advisor, Willie Swanson, for encouraging me to shoot for the moon when picking a project, supporting my independence, flying me around the globe to present my research, and being patient when times were tough.

Thank you to the Department of Genome Sciences for being such a wonderful place for collaborative, cutting edge research. Many people contributed significantly to this project. Preliminary miRNA sequences from Renee George's exome sequencing project inspired this research. The Thomas Lab provided helpful feedback over many shared lab meetings, and Sean Schneider was essential in many bioinformatics analyses. Sequencing was performed in collaboration with the lab of Debbie Nickerson, with special thanks to Adam Gordon for programming assistance and comments, Chris Frazar for library preparation help, and Tristan Shaffer for quality control on raw reads. My committee members Josh Akey, Debbie Nickerson, Christine Quietsch, and Walter L. Ruzzo provided thoughtful guidance. The GS IT staff is greatly appreciated for dealing with my frequent inquires. And finally, thank you to Brian Giebel for being the guardian angel of GS grad students.

Thank you to those who believed in me enough to fund me. I was supported by a NIH/NHGRI Genome Training Grant HG000035 and a NSF Graduate Research Fellowship. This research was funded by a UW Royalty Research Fund Grant and NIH grant HD076862.

Thank you to all of the scientific mentors I've been lucky enough to have throughout my life. Thank you, Keith Koszut, for introducing a curious 7th grader to genetics and evolution in such an amazing way that it became a lifelong obsession, and for encouraging me to join Science Olympiad when I lacked the confidence. Thank you to my undergraduate research advisors at Purdue who all encouraged me to pursue a PhD: Joe Busch, for training me in experimental work and being the best grad student mentor a wide-eyed undergrad could ask for; Andrew J. DeWoody, for the constant support and intimidating (but in retrospect, extremely educational) lab meetings; and Peter Waser, for the once in a lifetime fieldwork experience (I forgive you for not mentioning the tarantulas and rattlesnakes). Thank you to my human genetics professor Anna Berkovitz for encouraging me to go into the field, being a strong female role model, and leading me to Genome Sciences by exposing me to the research and activism of Mary Claire King.

Thank you to those of you who kept me sane over the years. My fellow grad students, for your comradery and commiseration. My friends near and far and virtual, for much needed laughs and distractions. My therapist (someone most grad students probably secretly thank) Stephen, for keeping it all in perspective and encouraging me not to give up. And last but not least, my cat Pixel, for endless amounts of therapeutic cuddles.

Thank you, thank you, thank you to Sean Schneider. There is no way to summarize how much you mean to me. You were like a second advisor to me, even though you were busy tackling your own PhD. This project wouldn't have been completed without your constant feedback and assistance throughout every step of the process and your skills as a master blaster. But most of all, thank you for being the love of my life, my best friend, and partner in crime. Thank you for believing in me when I didn't believe in myself, and for being by my side through

all the good and the bad. The end of our PhDs always seemed like a distant dream, but we made it – together.

And finally, thank you to my entire family for their infinite love and support, and for encouraging me to follow my dreams even if that meant moving thousands of miles away. Daddy, thank you for encouraging your little girl to believe she could achieve anything she worked hard at, raising me with a skeptical mind, and teaching me my very first Punnett square. Mommy, thank you for nurturing my creativity and sense of wonder about the world, and for always putting me first. Even when I was visiting you in the hospital, you were more concerned with me not giving up my higher education than your own wellbeing. I desperately wish you were here to see me finish, but I know there was never any doubt in your mind that I would succeed. That unwaivering love will be with me forever.

DEDICATION

In loving memory of my mother, Elena Katagas McCreight (1953-2014), and my Papou, Peter Katagas (1922-2016). *Que sera, sera.*

Chapter 1. INTRODUCTION

Comparative genomics is an indispensable tool for studying the evolutionary history of any organism. Humans are no exception: people are perpetually fascinated with the molecular variation that differentiates us from other primates. Studies comparing protein coding sequence data has uncovered rapid evolution between primates in many key areas, including immunity, sensory perception, reproduction, and keratinization (George et al. 2009; The Chimpanzee Sequencing and Analysis Consortium 2004; Goode et al. 2010). However, results from genomic scale analyses continue to reinforce that much, if not most, phenotypic differences between species result from changes in gene expression and not amino acid divergence (King and Wilson 1975; Enard 2002; Lee et al. 2007; Goode et al. 2010; Pai et al. 2011). While there are many layers of gene regulation that exist between DNA sequence data and expressed proteins, emphasis is often placed on the mechanisms that regulate transcription. There has been much work characterizing the coevolution of transcription factors and their DNA binding elements (Yang et al. 2011). However, less is known concerning the evolution of post-transcriptional regulatory elements. In addition to RNA binding proteins, microRNAs (miRNA) are an important class of post-transcriptional trans-acting factors that regulate mRNA stability and rates of translation (Chen and Rajewsky 2007). Despite their critical importance in seemingly every biological process (cell proliferation, differentiation, metabolism, apoptosis) (He and Hannon 2004), the role of miRNA in primate evolution has yet to be thoroughly examined. In this manuscript we provide an in-depth characterization of miRNA identification and evolution across multiple primate lineages.

1.1 MIRNA BIOCHEMISTRY

MiRNAs are short, noncoding, single-stranded RNAs important for post-transcriptional regulation in eukaryotes. MiRNAs are a relatively new addition to our understanding of genetics: the first miRNA was discovered in *Caenorhabditis elegans* in 1993, but the widespread effects of miRNAs were not fully recognized until the early 2000s (Berezikov 2011). Since then, discoveries in the miRNA field have expanded our understanding of genetics, illustrating the complexity of regulatory networks and the interplay between sequence and structure. Phylogenetic studies have shown that miRNAs have been present throughout the evolution of metazoans and that increased number and expression of miRNAs are positively associated with structural and organismal complexity (Berezikov 2011; Lee et al. 2007). Non-conserved miRNAs can be an indicator of adaptation in the genome of an organism, leading to novel phenotypes and a number of diseases, including heart disease, schizophrenia, and numerous types of cancers (Lee et al. 2007; He and Hannon 2004; Li and Kowdley 2012).

The intricate process of miRNA biogenesis (Figure 1) plays a crucial role in generating diverse phenotypes in organisms, as an alteration to any step may have profound downstream effects. MiRNA genes are transcribed from the genome, resulting in a primary miRNA transcript that may include a single miRNA or a cluster of miRNAs (Berezikov 2011). Regions of a primary miRNA form hairpin structures that are recognized by the endonuclease drosha, which cleaves the double-stranded stem region of the hairpin to produce an approximately 83 nucleotide (nt) precursor miRNA (pre-miRNA) (Fang et al. 2013). After being exported to the cytoplasm, pre-miRNA are further processed by a second endonuclease, dicer, which cleaves off the loop region of the hairpin to produce an approximately 22 nt double stranded RNA duplex that contains the mature miRNA and its complement (termed the star strand, or miRNA*). The

strand with the less thermodynamically stable 5’ end becomes the mature sequence and is loaded into the RNA-induced silencing complex (RISC), while the star sequence is degraded. Occasionally a pre-miRNA has both of its miRNA and miRNA* strands lead to mature sequences. The mature miRNA base-pairs with complementary sequence within the 3’ untranslated region (UTR) of messenger RNA (mRNA). This process guides RISC to specific transcripts, resulting in down-regulation of the targets through degradation of the transcripts or inhibition of translation (Nilsen 2007). Although there are varying degrees of complementarity between a miRNA and its mRNA target, binding is most highly dependent on positions 1 through 8 of the 5’ end of the mature miRNA, known as the seed region (Berezikov 2011). Although 75% of downregulated mRNA have canonical seed sites in their 3’ UTR, the seed region is not always sufficient for causing downregulation (Grimson et al. 2007). The 3’ end of the mature miRNA can also have an effect: positions 13 – 16 are highly conserved, and their proper complementary base pairing to a mRNA target is associated with downregulation (Grimson et al. 2007).

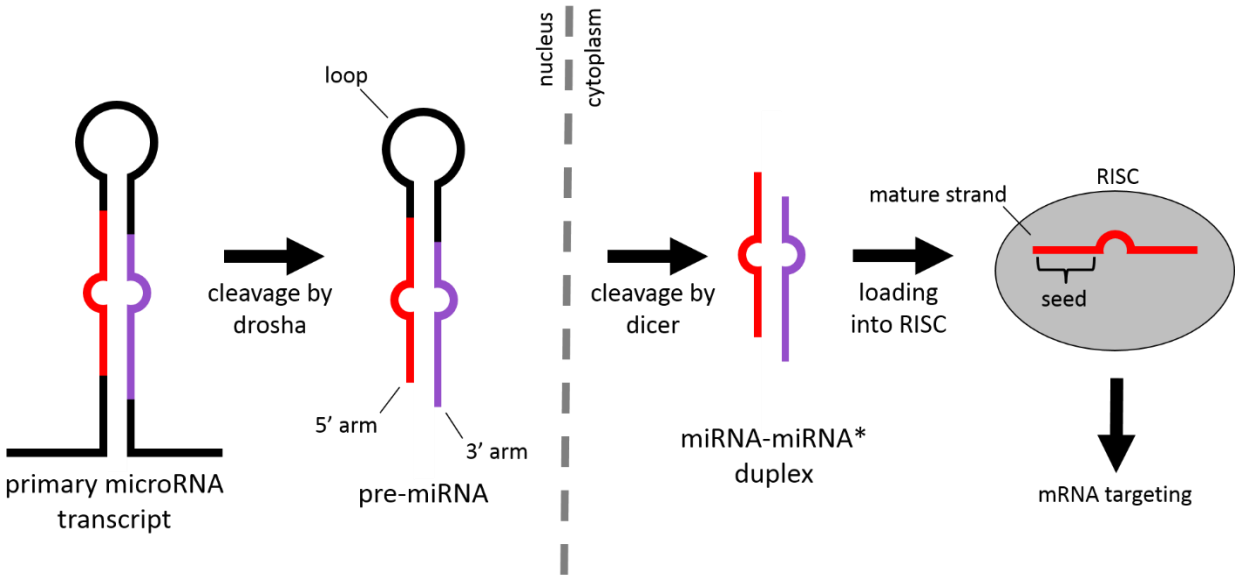

Figure 1. miRNA biogenesis. miRNA genes are transcribed from the genome, resulting in a primary miRNA transcript. Regions of the primary miRNA form a hairpin structure that is recognized by the endonuclease drosha, which cleaves the double-stranded stem region of the hairpin to create a pre-miRNA of ~83 nt in length. The pre-miRNA is exported to the cytoplasm where it is further processed by dicer, which cleaves off the loop region of the hairpin. This results in an approximately 22 to 23 nt double-stranded RNA called the miRNA-miRNA* duplex. The mature miRNA strand is loaded into the RNA-induced silencing complex (RISC), where its 8 nt seed region complementarily base pairs with messenger RNA targets, leading to their downregulation.

About 60% of all human transcripts contain known or predicted miRNA target recognition sites (Friedman et al. 2009). A single type of miRNA can have one to thousands of targets, which establishes the potential for small changes in miRNA sequence to have profound phenotypic effects: each miRNA may result in varying degrees of phenotypic plasticity for different cell types, which have different target mRNAs to act upon. Additionally, many miRNA

have multiple paralogs throughout the genome. Gene duplication followed by mutation in one copy is a common avenue for the evolution of novel functions: by maintaining more than one gene copy for a given miRNA, purifying selection to preserve function is often relaxed for one of the paralogs, allowing for mutational acquisition, differentiation, neofunctionalization, and subfunctionalization (Conant and Wolfe 2008). Most miRNA are highly conserved across species and show higher rates of purifying selection than protein-coding regions of the genome, suggesting that variation found in miRNA sequence may play a vital role in the evolution of metazoans (Hausser and Zavolan 2014; Pang et al. 2006; Altuvia et al. 2005; Berezikov et al. 2005).

1.2 MIRNA STRUCTURE

While many studies focus on changes in miRNA expression or variants in the seed region, changes in pre-miRNA secondary structure can also dramatically affect downstream function through several different mechanisms. In general, variants in the stems of pre-miRNAs that decrease overall structural stability of the hairpin reduce the production of mature miRNA (Gong et al. 2012). If the pre-miRNA has a secondary structure that is very divergent from the standard hairpin, the ability of drosha to recognize and process the pre-miRNA may be reduced or completely eliminated. Small changes in sequence may have drastic effects: a variant in the mature sequence of miRNA-125a blocks the processing of primary miRNA to pre-miRNA, resulting in complete loss of function (Duan et al. 2007). However, sequence divergence does not always imply structural divergence, as compensatory mutations often help conserve a pre-miRNA's hairpin structure.

A mutation in the primary miRNA sequence could also result in a different but stable hairpin structure. This could alter the location of drosha cleavage sites during pre-miRNA

biogenesis (Han et al. 2006), and in turn shift the cleavage sites of dicer, resulting in a different mature miRNA and thus different seed region. Sun et al. identified such a variant with an altered cleavage site and seed region shift (Sun et al. 2009). Previous studies have only investigated a limited number of pre-miRNA variants, and the effects of most are still unknown.

Chapter 2. PRIMATE MIRNA EVOLUTION

A major roadblock to studying miRNA across primates is the lack of experimentally verified miRNA in non-human primates. Only 13 of the ~300 known primate species (Perelman et al. 2011) have any entries in miRBase (Table 1, Figure 2) (Kozomara and Griffiths-Jones 2011). The number of characterized human pre-miRNAs (n=1881) is still more than twice as large as that of chimpanzee (n=655), the most well studied non-human primate. The majority of these miRNA are predicted based only on homology to the human genome; only four species (chimpanzee, gorilla, orangutan, and rhesus macaque) have sequences that are experimentally validated through RNAseq or other expression analyses. While homology is a useful tool for identifying orthologs, it may result in an overrepresentation of conserved sequences with respect to humans, missing sequence diversity in more distantly related primate species. Homology alone cannot identify whether a sequence is actually expressed or forms a stable hairpin that successfully completes processing by drosha and dicer. Even in cases where a mature miRNA is produced, one cannot determine the exact boundaries of the mature sequence (and thus seed region) without expression data (Pritchard et al. 2012). The current best mature predictive software (Mature Bayes) only comes within 1 nt of the true mature sequence 49% of the time, which can result in an incorrect seed region and thus target repertoire (Gkirtzou et al. 2010).

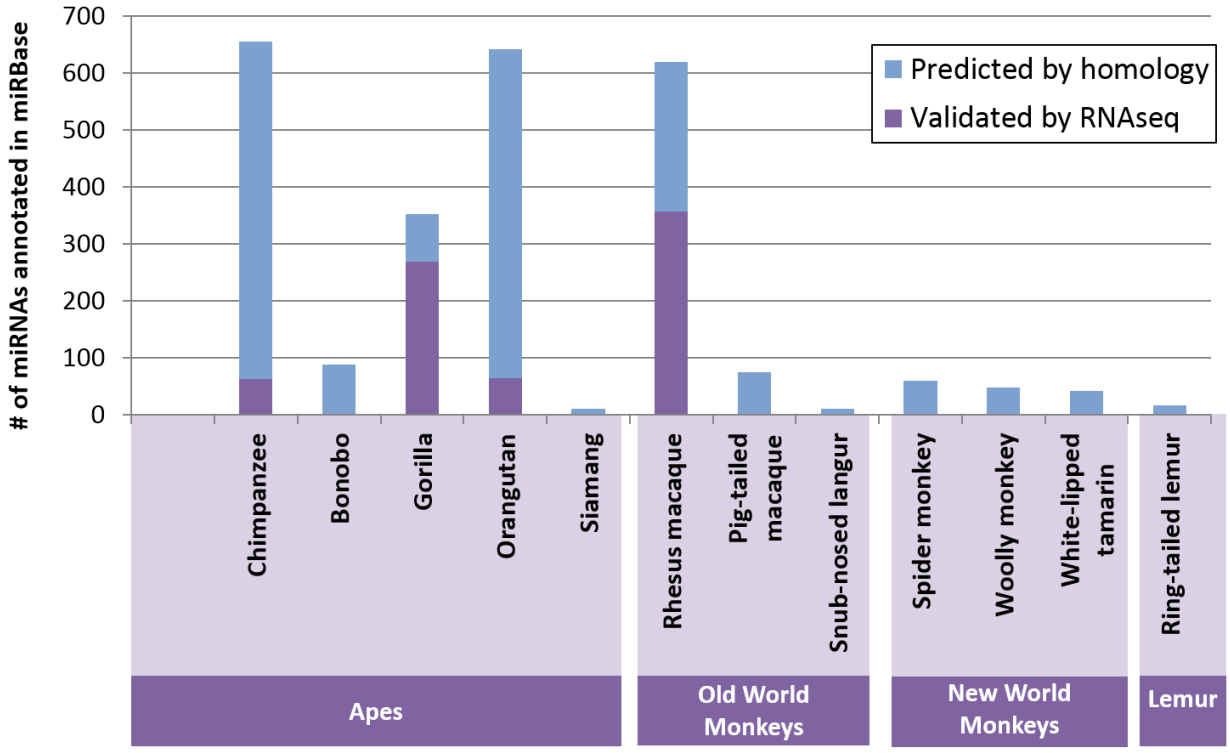

Figure 2. Non-human primate miRNA is poorly characterized. Only 12 of the ~300 known primate species have any entries in miRBase (release 21). The majority of these miRNA are predicted based only on homology (shown in blue); only four species (chimpanzee, gorilla, orangutan, and rhesus macaque) have sequences that are experimentally validated through RNAseq or other expression analyses (shown in purple). The number of characterized human pre-miRNAs (n=1881) is still more than twice as large as that of chimpanzee (n=655).

Despite this lack of validated miRNAs, some prior studies have compared differences in miRNA across primates. Berezikov et al. used high throughput sequencing technology to discover miRNAs in the brains of human fetuses and chimpanzee adults, identifying hundreds of miRNAs specific to primates with dozens not conserved between humans and chimpanzees (Berezikov et al. 2006). Hu et al. recently discovered several miRNA that were differentially

expressed in human and chimpanzee brains, and that this differential expression resulted in downregulation of several neuronal genes (Hu et al. 2011). Zhang et al. discovered an X-linked miRNA cluster that was rapidly evolving in primates, and these miRNA had increased expression during male sexual maturation (Zhang et al. 2007).

However, techniques investigating only the expression level of miRNA would miss any phenotypic differences caused by changes in miRNA target specificity. Target specificity could change due to sequence differences in the seed region, or sequence differences that change the secondary structure of pre-miRNA and thus alter its downstream processing. In this study, we focus on nucleotide variation found within the mature and pre-miRNA sequences rather than expression levels. Non-conserved miRNA provide insights into primate evolutionary history, including what differentiates humans from other primates. In order to investigate how conserved or divergent miRNA are within primates, we sequenced miRNA from thirteen species, greatly expanding the number of experimentally validated non-human primate miRNA and our knowledge of their evolution.

2.1 RESULTS

2.1.1 *miRNA discovery*

In order to better characterize patterns of miRNA evolution across primates, small RNAseq was performed on fibroblast cells cultured from 13 divergent primate species (Figure 3, see Methods). Our study drastically expanded our knowledge of primate miRNA, tripling the number of primate species with experimentally validated miRNA in miRBase (from 4 to 13), and adding dozens to hundreds of miRNA per species (Figure 4, Table 2). This includes the only experimentally validated miRNA sequences available for bonobo, one of human's closest evolutionary relatives; and for the first time, experimentally validated miRNA sequences are

available for New World monkeys, lemurs, and a galago. For non-human primate miRNAs that to date were computationally predicted by homology alone, 27% (211/766) were sequenced in at least one of our primate species, with the majority of these sequences (86%) being represented by at least two species (Figure 5).

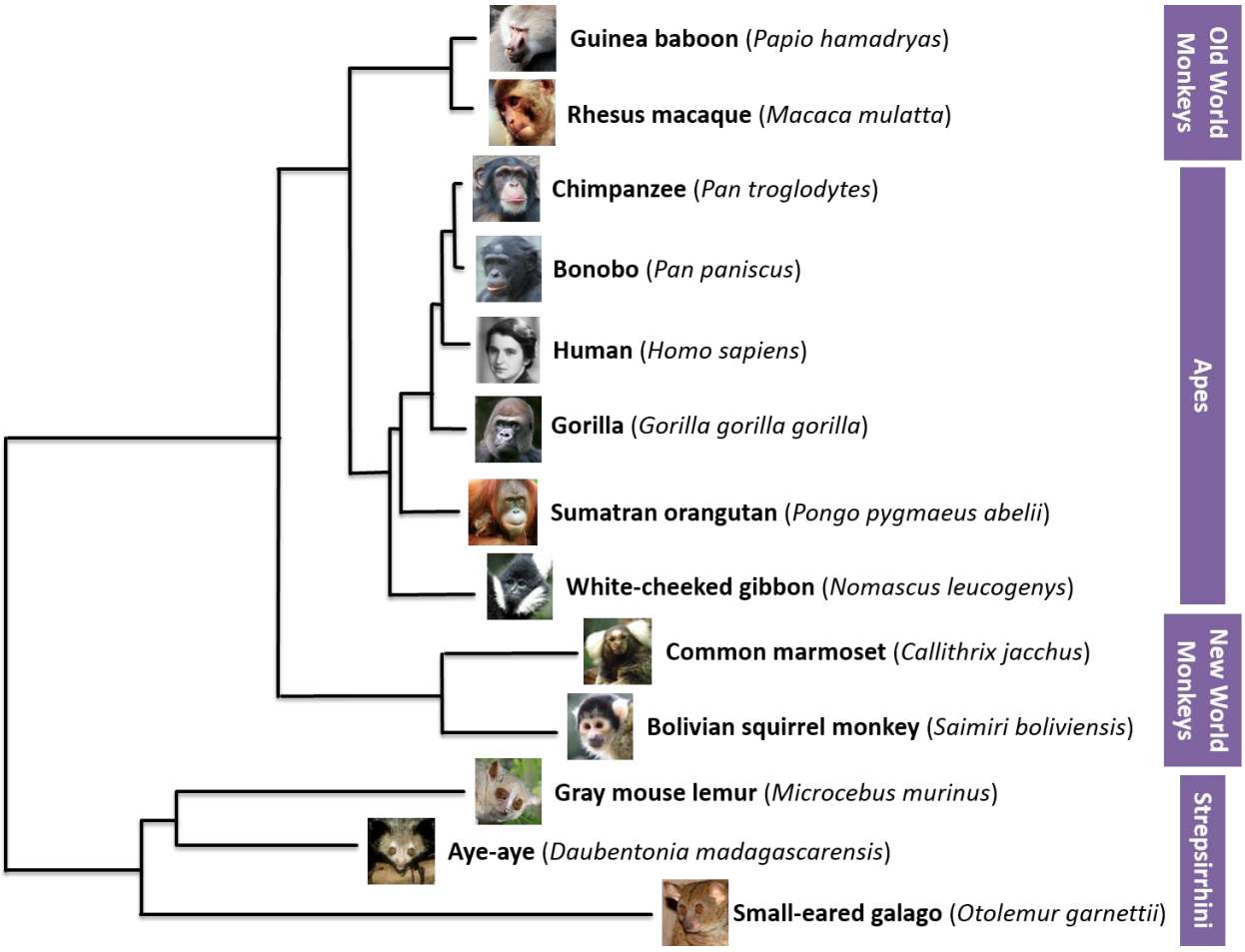

Figure 3. Phylogeny of primate genome assemblies included in our study (adapted from Perelman et al. 2011). We selected species that had both a sequenced genome, and fibroblast cell culture available through Coriell Cell Repositories.

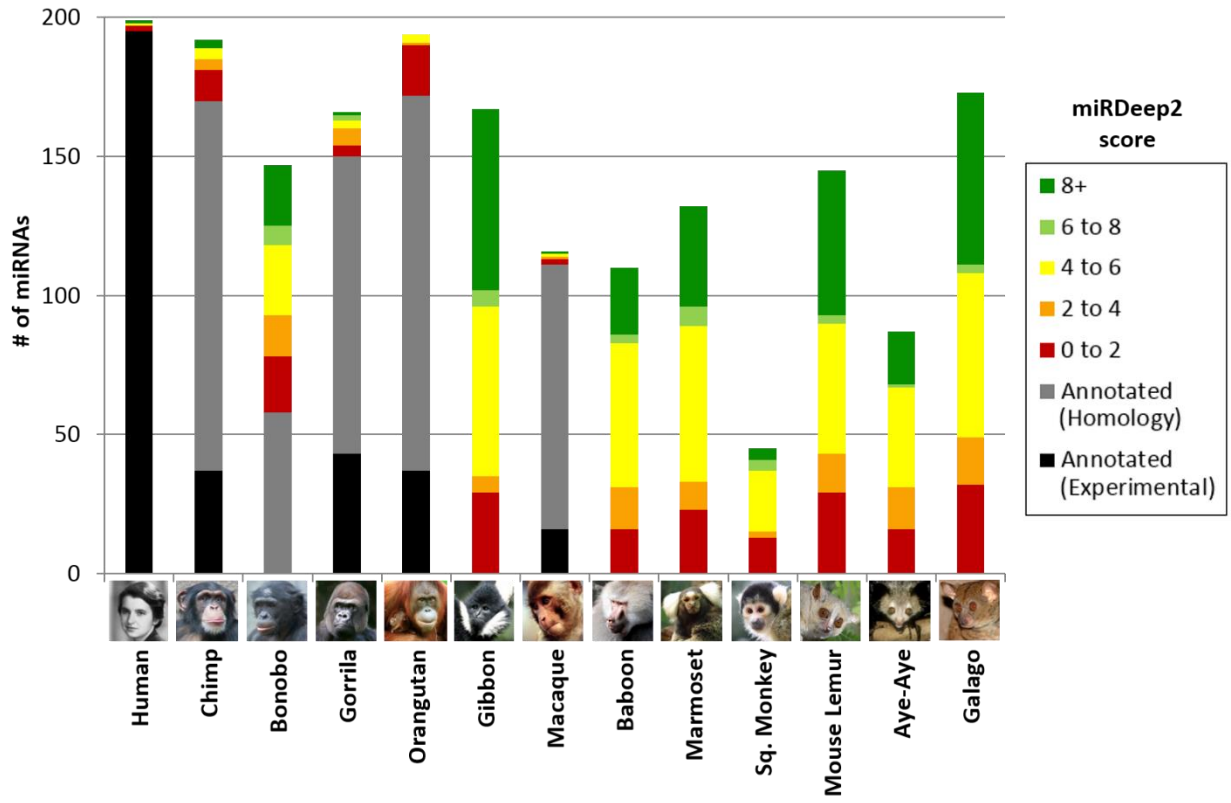

Figure 4. miRDeep2 results by score. MiRDeep2 scores range from -10 to 10, with a higher number corresponding to increased likelihood that a miRNA is genuine. A cut-off of 0 was used to be included in this study. miRNA already annotated in miRBase are represented in black and gray: black represents miRNA with experimental validation, and gray represents miRNA previously predicted solely by homology to the human genome that have now been validated in this study. Novel miRNA are shown in a color corresponding to their miRDeep2 score; this score is partially determined by the availability of any previously annotated miRNA, which would inherently result in lower scores for our primates with no information in miRBase.

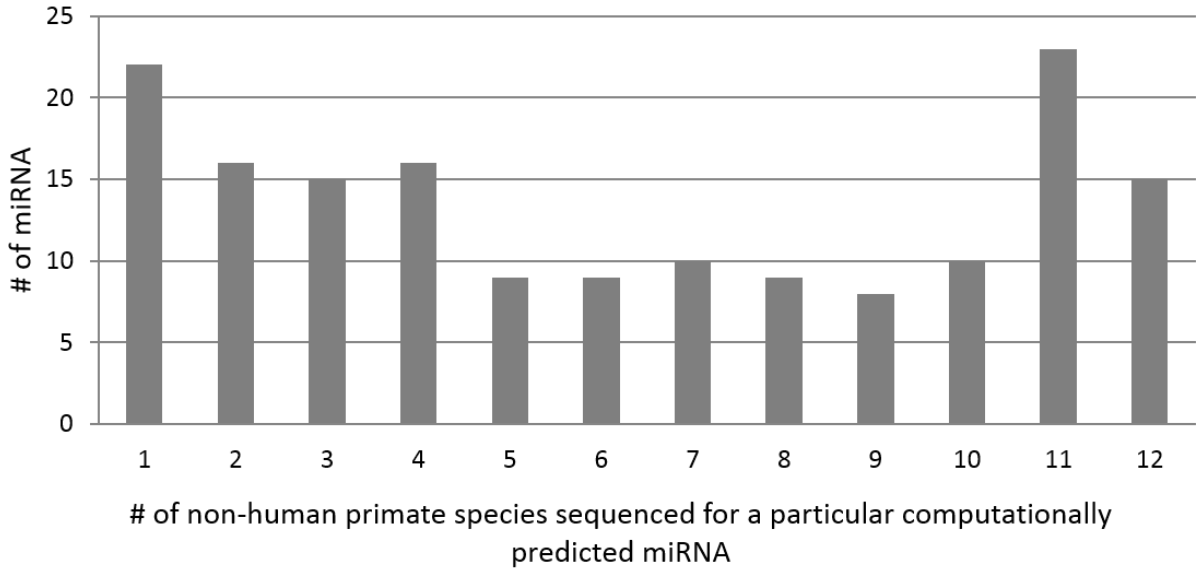

Figure 5. Distribution of the number of non-human primate species from our dataset sequenced for a particular miRNA that was previously computational predicted by homology alone. 140/163 (86%) have experimental support from at least two primates. Because of the difficulty distinguishing between paralogs with identical mature sequences, only the paralog with the most coverage from a family of miRNA is shown in this chart.

2.1.2 *Primate miRNA evolution*

Sequenced miRNAs were computationally clustered into groups of homologs that had at least 70% identity within the mature region (see Methods). Homology groups containing paralogs were further subdivided into their individual miRNA orthologs, resulting in 188 particular miRNA ortholog groups with representation in at least three primate species. As expected based on previous studies (Pang et al. 2006), primate miRNA appears to be highly conserved, with 173 of 188 miRNA ortholog groups (~92%) showing no variation within the mature region across primates. Of the 15 miRNA ortholog groups that contained variation within the mature region, none of these variants occurred within the seed region. This is consistent with

previous studies that show the seed region to be the most highly conserved region of miRNA and the most important determinant of target recognition (Figure 6) (Grimson et al. 2007). Only one variant was found within positions 13-16, the second most conserved region of miRNA that is sometimes involved in 3' complementary base pairing during target recognition. Most variation was observed in basal primate species: 14/21 variant sequences were from the basal Strepsirrhini suborder, and 5/21 were from New World monkeys (Table 3). This is concordant with the hominoid slowdown hypothesis, which shows that rates of nucleotide substitution in primates decrease as generation time increases (Li and Tanimura 1987).

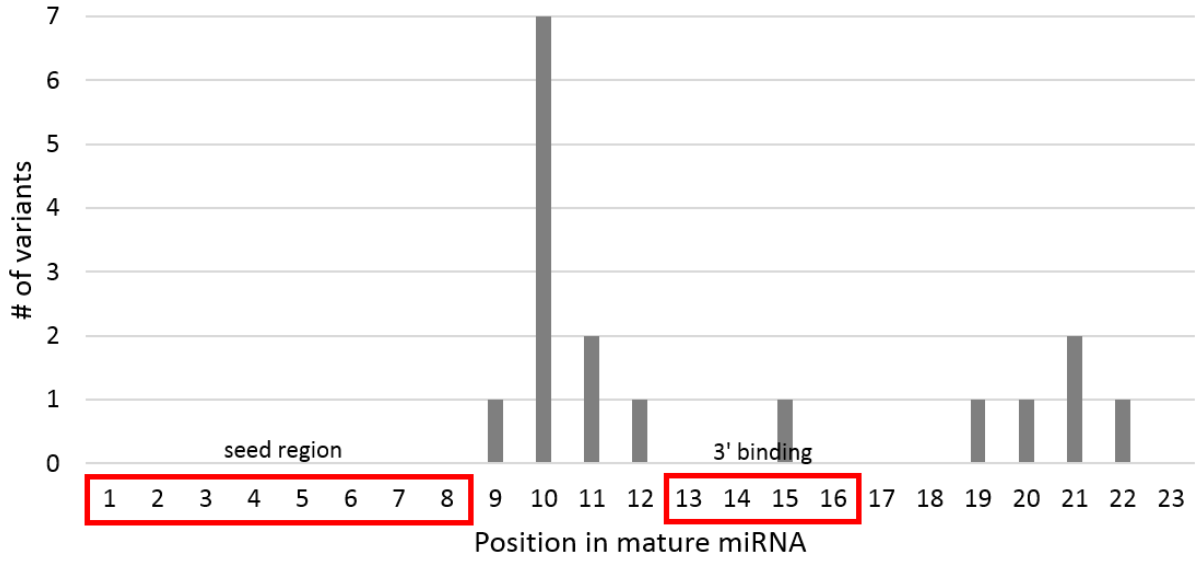

Figure 6. Location of variants within mature miRNA across the thirteen primate species sequenced in this study. The 5' end of the mature miRNA has an 8 nt "seed region" in positions 1 through 8 that complementary base-pairs with the 3' untranslated region (UTR) of messenger RNA (mRNA). The 3' end of the mature miRNA can also have an effect: positions 13 – 16 are highly conserved, and their proper complementary base pairing to a mRNA target is associated with downregulation (Grimson et al. 2007). As expected, the vast majority of the variants

sequenced in our study appear in positions with relaxed evolutionary constraints (positions 9-12 and 17-23).

2.1.3 *Structural analysis*

We analyzed the thermodynamic stability and structural conservation of any miRNA with at least 5 species in its alignment (n = 152, see Methods). Our pre-miRNA structures are thermodynamically stable as measured by z-score (where more negative values indicate stability), with most of the analyzed miRNAs (120/152) having a z-score that indicates very conserved structures ($z < -3.0$) (Figure 7). Structural stability is further evidence that a miRNA sequence is genuine (Bonnet et al. 2004). Structural conservation as measured by the Structural Conservation Index (SCI), where an SCI of 1 indicates complete structural conservation, generally decreases as sequence divergence increases (Figure 8), but this correlation is weak ($R^2 = .1719$). This is to be expected, as SCI only approximately captures true structure conservation, but a weak correlation is also concordant with the properties of miRNA: a miRNA with low sequence identity may still be structurally conserved due to compensatory mutations, or a miRNA with high sequence identity may have one variant that results in drastic (and perhaps functionally significant) structural differences.

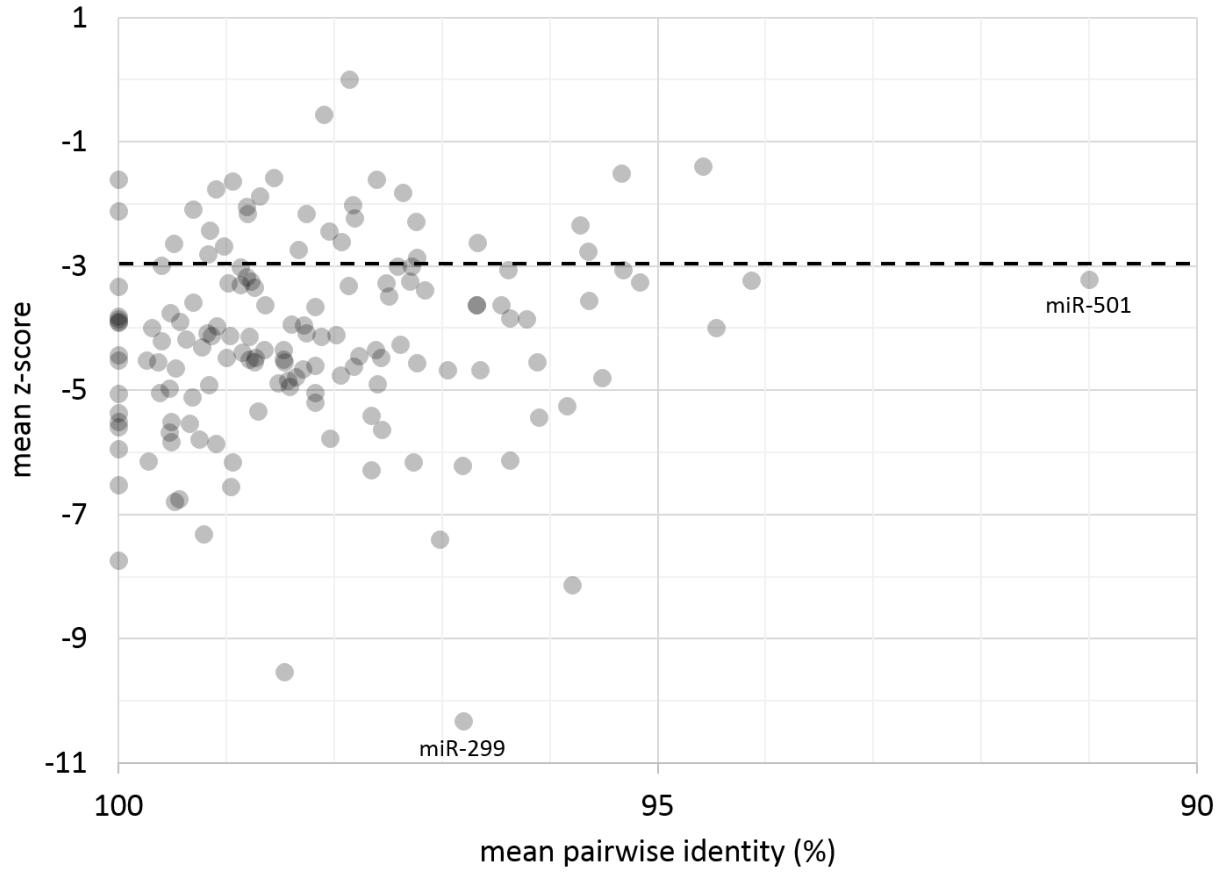

Figure 7. Mean pairwise sequence identity compared to the z-score, where a more negative z-score indicates increased structural stability. Scores below -3 (represented by the dotted black line) generally indicate very stable structures.

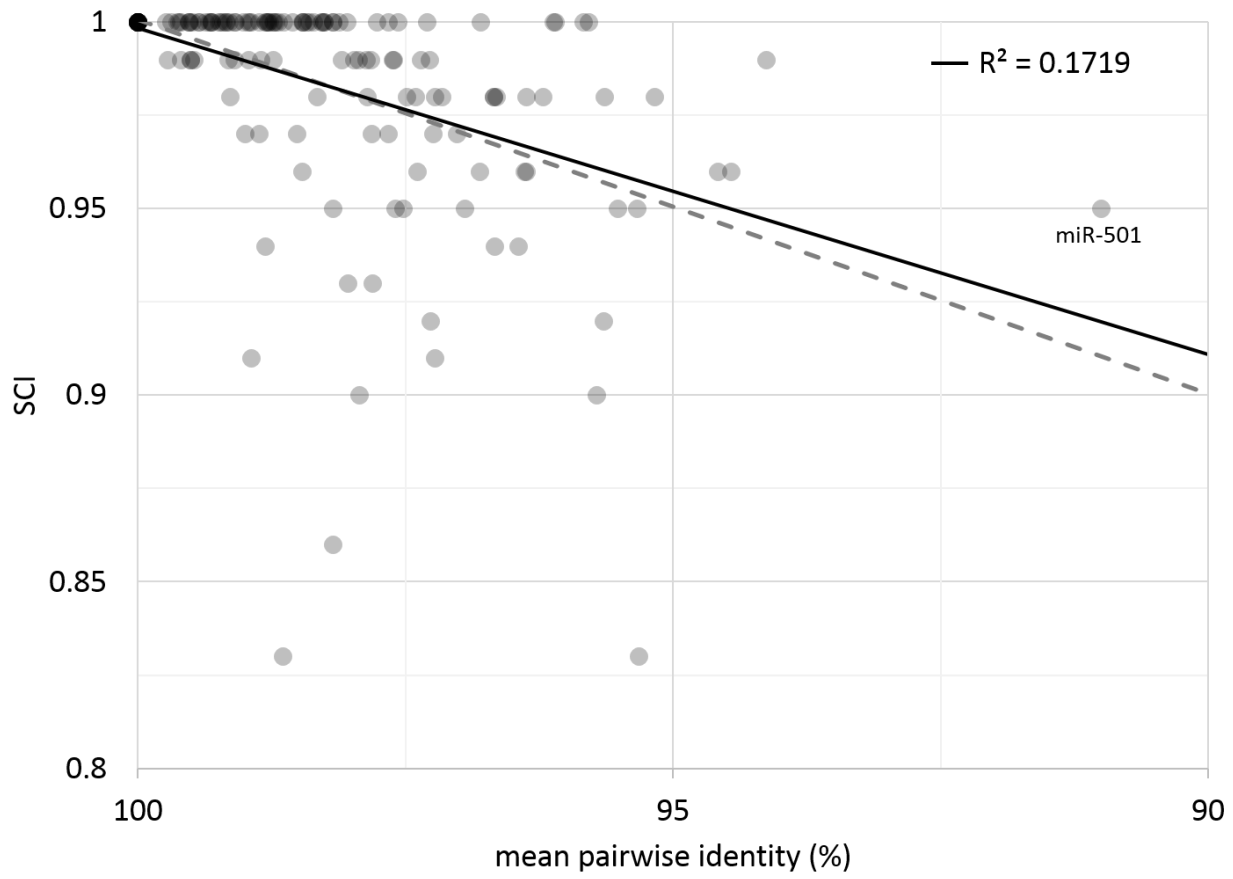

Figure 8. Mean pairwise sequence identity compared to the Structure Conservation Index (SCI). In general, an SCI near or above the mean pairwise identity indicates structural conservation (dotted gray line). The black line is the linear regression for our data ($R^2 = 0.1719$).

2.1.4 *miR-2355-3p*

MiR-2355-5p was the only alignment to contain a variant within the conserved 3' binding location (positions 13 – 16) of the mature miRNA. Homologous sequences from additional primate species and a mouse (*Mus musculus*) outgroup extracted from the UCSC Genome Browser reveal an interesting evolutionary event: the closest sister taxa to humans (chimpanzee, bonobo, and gorilla) have variant T15C in the mature sequence, while humans have seemingly reverted to the ancestral T (Figure 9) (Speir et al. 2016). This reversion is conserved across

humans in dbSNP (Sherry et al. 2001). This specific transition, as well as other mutations throughout the pre-miRNA, do not seem to alter the secondary structure of the hairpin. This is the first time RNAseq has confirmed the expression of this miRNA in a non-human primate (chimpanzee, bonobo, gorilla, and orangutan); the only other species confirmed to express this miRNA is cow (*Bos taurus*) (Glazov et al. 2009), suggesting it is likely to be expressed in other primates as well. Human miR-2355-5p has been shown to be expressed in embryonic stem cells (Hansen et al. 2010), neural stem cells (Goff et al. 2009), and throughout the female reproductive tract (Witten et al. 2010; Creighton et al. 2010). The specific targets and function of miR-2355-5p are currently unknown.

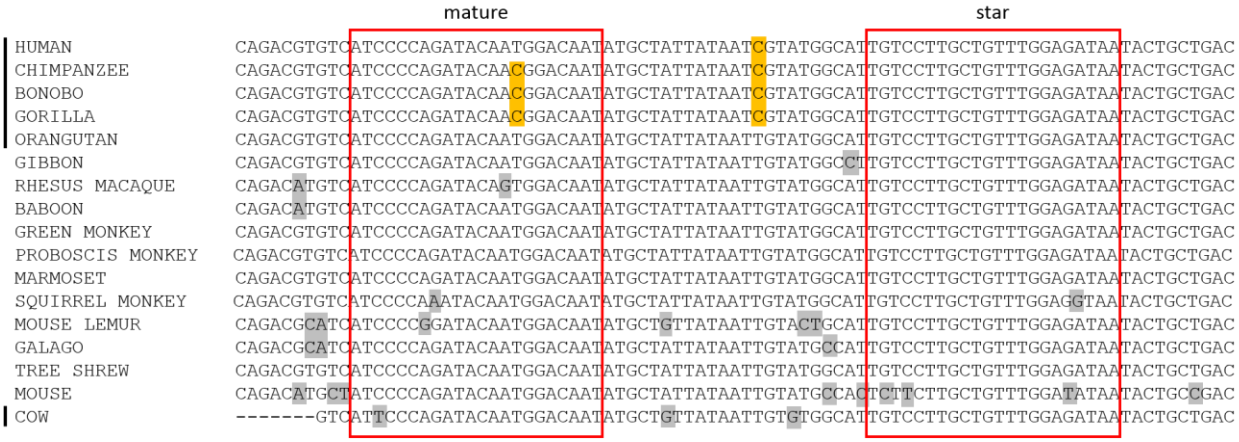

Figure 9. Alignment of miR-2355 homologous sequences. The black line indicates species that have experimental support for the transcription of miR-2355, either in miRBase (human, cow) or from this study (chimpanzee, bonobo, gorilla, and orangutan). The red boxes outline the mature and star sequences within the miRNA. Variants only found among the great apes are highlighted in yellow, while all other variants are marked in grey. Humans have experienced a reversion at position 15 of the mature miRNA, restoring that nucleotide to its ancestral state.

2.1.5 *miR-299-3p*

miR-299-3p was completely identical across all primates except for a single change (C10T) in humans. The secondary structure of miR-299 was entirely conserved (SCI = 1) and was the most thermodynamically stable hairpin of all the miRNA in this study (z = -10.33). Previous research identified this change based on sequences from human, chimpanzee, and macaque, and we show that the ancestral sequence is shared across all primates except human. Initial research found that miR-299-3p has human-specific expression, with preferential expression in neurons; although targets of this miRNA were enriched for neuronal function and axon guidance, there was no difference in target specificity between the human and chimpanzee versions (Hu et al. 2015). A more recent study confirmed that miR-299-3p has targets enriched for neuronal function; however, contrary to the previous study, changes in target repertoire and expression levels between humans and non-human primates were identified (Gallego et al. 2016). These conflicting results are likely due to tissue specificity: Gallego et al. found that miR-299-3p was highly expressed only in cerebellum, whereas Hu et al.'s analyses were performed in neuroblastoma cell lines. These expression and target differences illustrate how a change outside of the seed region can still have profound effects on miRNA function.

2.1.6 *miR-501-3p*

Of all of the alignments, miR-501 had the lowest mean pairwise identity (91%) across primates, similar to the average pairwise identity between most human and mouse pre-miRNA (>90%) (Pang et al. 2006). However, miR-501 still appeared to be structurally conserved due to a number of compensatory mutations (SCI = 0.95, covariance contribution = -0.24). The mature sequence miR-501-3p also contains a variant just outside the seed region at position 9 in the

basal Strepsirrhini suborder that introduces an additional bulge in the hairpin. This variant, as well as variants outside of the mature sequence, likely alters the overall secondary structure of the hairpin, resulting in the mature sequence being shifted downstream by 1 nt in apes relative to Strepsirrhines (Figure 10). This change would alter the dicer cleavage position and shift the seed region by one nt, likely changing the target repertoire of this particular miRNA.

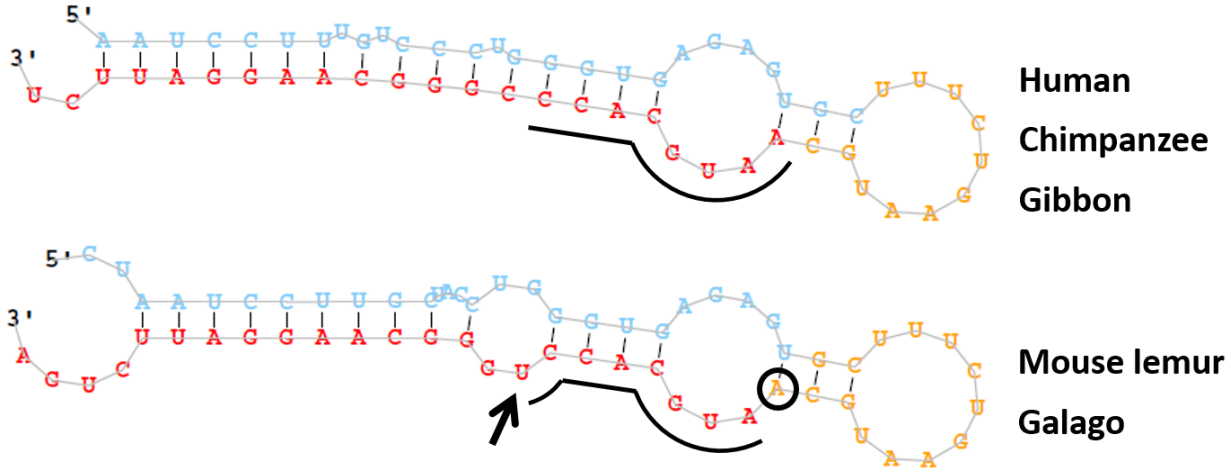

Figure 10. Predicted structure of miR-501 by miRDeep2. Red indicates the mature miR-501-3p sequence supported by reads, yellow the predicted loop, and blue the predicted star sequence. In mouse lemur and galago, the mature sequence contains a variant (arrow) immediately following the seed region (underlined); this as well as variants outside of the mature sequence appear to alter the overall secondary structure of the hairpin, resulting in the mature sequence and thus the seed region being shifted downstream by 1 nt (circled in black).

Previous research has shown that miR-501-3p localizes to dendrites and plays a key role in NMDA-induced dendritic spine remodeling, which is thought to be the "structural basis of information storage in the brain for cognitive functions such as learning and memory" (Hu et al. 2015). Suppression of *GluA1* expression is necessary for long term maintenance of NMDA-induced spine modifications: experimental assays showed that miR-501-3p targets the transcript

of *GluA1*, and their expression are inversely correlated during postnatal brain development. NDMA stimulation increased the expression of the primary miR-501 transcript, and still increased mature miR-501-3p levels even when a transcription inhibitor was present, suggesting that miR-501-3p undergoes post-transcriptional regulation (Hu et al. 2015). This regulation may be controlled by the structure of miR-501, as hairpin structural stability increases the production of the mature miRNA (Gong et al. 2012). The sampled apes in our study have increased complementary base pairing throughout the hairpin, lacking the mid-mature bulge found in Strepsirrhines. This structural difference and its resulting seed shift may indicate an important moment in primate brain evolution.

2.1.7 *miR-320*

One of our largest homology groups was composed of the miRNA-320 paralogs (Figure 11). In humans, the miR-320 family consists of one copy of miR-320a (chr8), two copies of miR-320b (chr1), two copies of miR-320c (chr18), two copies of miR-320d (chr13 & chrX), and a single copy of 320e (chr19). Gene duplication allows novel functions to evolve, as one paralog is maintained by purifying selection for its previous function while other copies are allowed to acquire mutations and neofunctionalize and/or subfunctionalize. Although our data showed miR-320a to be present across the entire primate lineage, we only identified RNAseq reads for miR-320b and miR-320c in apes and Old World monkeys, matching the copy number found in humans (we did not sequence any copies of miR-320d or miR-320e from any species, likely because they are not expressed in this particular cell type). This pattern of only being found in the most derived species is unlikely to occur by random chance, and may indicate that these additional paralogs do not exist in these genomes.

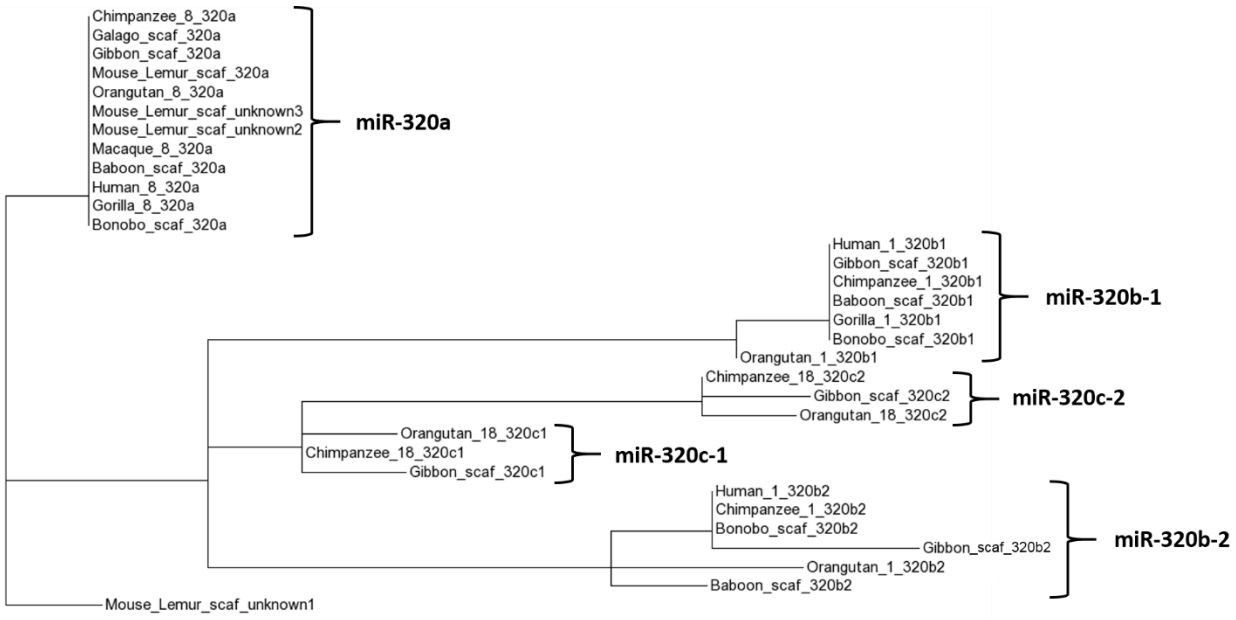

Figure 11. Phylogenetic tree of the predicted pre-miRNAs of the miR-320 family, based on experimentally determined mature sequences. Paralogs miR-320b and miR-320c are only expressed in apes and Old World monkeys, lacking representation from New World monkeys and Strepsirrhines.

miR-320a is mammalian-specific, and its paralogs have been identified in previous studies (experimentally or computationally) only in primates: 320b in gorilla, 320b and 320c in macaque, and all paralogs in chimpanzee and orangutan (Kozomara and Griffiths-Jones 2011; Dannemann et al. 2012; Brameier 2010; Baev et al. 2009). We confirmed the presence or absence of each miR-320 paralog using blastn (see Methods, Table 4). 320a was present in all primates, concordant with our RNAseq data. 320b1, 320c1, 320c2 and 320d2 were present in all primates except for the basal Stepsirrhini suborder. 320b2, 320d1, and 320e were present only in apes and Old World monkeys, with 320e having an additional duplication in orangutan. Alignments clearly indicated whether or not a sequence was present; for example, the pre-miRNA of miR-320b1 is absent in Strepsirrhines despite the conservation of flanking sequence,

illustrating the insertion event that took place sometime after the Strepsirrhini suborder split from the rest of the primate lineage (Figure 12). Paralogs found in a genome but missing in our sequencing data may be identified in future RNAseq efforts; however, it is also possible that despite being in the genome, they are not expressed, are expressed only in different cell types, or are not successfully processed into mature miRNA. Our data indicates that the miR-320 family has undergone multiple gene duplications throughout primate evolution, and suggests that only more derived species successfully express mature forms of miR-320b and miR-320c, but more extensive investigation of expression in other species is needed to expand on these conclusions.

Figure 12. Alignment of miR-320b1 homologous sequences. The yellow box denotes the pre-miRNA sequence, red outlines the mature sequence, and variants with respect to humans are marked in grey. miR-320b1 is found in all apes, Old World monkeys, and New World monkeys (only human and marmoset are shown for simplicity). The entire pre-miRNA sequence is clearly absent in Strepsirrhines (aye-aye and galago), despite conservation of flanking sequence, demonstrating an insertion event that took place after the Strepsirrhini suborder split from the rest of the primate lineage.

Numerous studies have illustrated the role of the miR-320 family as a regulator of neural development. The family is highly expressed in rat neurons, with enrichment specifically in axons (Natera-Naranjo et al. 2010), and is also targeted by REST, a transcription factor that silences neuronal genes in non-neuronal tissue and is essential in neuronal differentiation as well as the maintenance of neural stem cells (Otto et al. 2007; Gao et al. 2012). Transfection experiments confirm that miR-320b inhibits expression of neuron-related mRNA targets, and *in situ* hybridizations to investigate histological expression patterns show miR-320b co-localized with neurons in both human and macaques (Somel et al. 2011). Increased levels of miR-320b have also been shown to increase neurite length, further suggesting that increased copies of this miRNA may play a role in neuronal development (White and Gifford 2012). The miR-320 family is also frequently dysregulated in neurological disorders: miR-320 was downregulated in blood of schizophrenia patients (Vachev et al. 2016) and the striatum of forebrains of Huntington Disease patients (Martí et al. 2010), and upregulated in the cortex of patients with sporadic Alzheimer's disease (Hébert et al. 2008) and mouse brains undergoing prion-induced neurodegeneration (Saba et al. 2008). In both control and Huntington Disease forebrains, miR-320b and 320c have such high rates of post-transcriptional nucleotide substitutions compared to their primary transcripts that researchers have suggested these edited sequences be considered the reference miRNA in brain tissue; additionally, these two paralogs were the only miRNA with post-transcriptional substitutions at several positions across the mature sequence (Martí et al. 2010). In addition to its role in neuronal development, miR-320c appears to have a wide range of functions, including regulation of chondrocytes in cartilage (Ukai et al. 2012), differentiation of skeletal stem cells (Hamam et al. 2014), inhibition of proliferation, migration, and invasion in bladder cancer (Wang et al. 2014), and induction of resistance to the chemotherapeutic agent

gemcitabine in pancreatic cancer (Iwagami et al. 2013). Given the number of duplications this miRNA family has undergone throughout primate evolution and the enrichment of neuronal functions among its paralogs, we speculate that the miR-320 family played a role in primate brain evolution.

2.2 METHODS

2.2.1 *Primate samples*

In order to represent a broad span of the primate phylogeny, we selected thirteen primate species that had sequenced genomes and fibroblast cell cultures available through Coriell Cell Repositories (Figure 3, Table 5):

- **Apes:** human (*Homo sapiens*), chimpanzee (*Pan troglodytes*), bonobo (*Pan paniscus*), western lowland gorilla (*Gorilla gorilla gorilla*), Sumatran orangutan (*Pongo pygmaeus abelii*), northern white-cheeked gibbon (*Nomascus leucogenys*)
- **Old World monkeys:** Indian rhesus macaque (*Macaca mulatta*), Hamadryas baboon (*Papio hamadryas*)
- **New World monkeys:** Common marmoset (*Callithrix jacchus*), Bolivian squirrel monkey (*Saimiri boliviensis*)
- **Strepsirrhines:** Gray mouse lemur (*Microcebus murinus*), Aye-aye (*Daubentonia madagascarensis*), Small-eared galago (*Otolemur garnetti*)

Marmoset is the only species where the genome of a closely related species (*C. jacchus*) was used as the reference for the cell line available (*C. geoffroyi*).

2.2.2 *RNA sequencing*

Mature miRNA from fibroblast cells was extracted using the Qiagen miRNeasy Mini Kit following the manufacturer's protocols (Catalog #217004, Qiagen, Valencia, CA). We prepared and barcoded samples using Illumina's TruSeq Small RNA Library Preparation Kit (Catalog #RS-200-0012, Illumina). Barcoded samples were multiplexed for 25bp paired-end sequencing on a single lane of an Illumina MiSeq. The Picard tool ExtractIlluminaBarcodes separated raw Illumina reads by barcode, and IlluminaBasecallsToFastq output the results in fastq format. Fastqc was run on the original fastq, with FastqTOSam importing the fastqs into bam format. Bam files were processed with Picard MarkIlluminaAdapters tool. The adapter-masked fastqs were put into PEAR with no minimum overlap size in order to merge the reads. Fastqc was run on merged, trimmed fastq and compared to the original fastqc to confirm that adapters were trimmed.

2.2.3 *miRNA identification*

miRDeep2 was used to predict novel and identify previously annotated miRNA (Friedländer et al. 2008). Merged, trimmed reads were mapped to their respective genomes using the miRDeep2 mapper.pl module with the following parameters: -c -j -l 18 -m -p -s -t –v. MiRDeep2 was executed with default parameters. When making novel miRNA predictions, miRDeep2's algorithm accounts for already known miRNAs of the species being analyzed and of any related species. We retrieved a list of known miRNAs from miRBase (release 21) for any of our primates that were in the database (*H. sapiens*, *P. troglodytes*, *P. paniscus*, *G. gorilla*, *P. pygmaeus*, *M. mulatta*), and used all known metazoan miRNAs as our "related species" reference. MiRDeep2 assigned a score from -10 to 10 to each miRNA, with a higher number corresponding to increased likelihood that the putative miRNA is functional. This score is

partially determined by the availability of any known miRNA, which would inherently result in lower scores for our primates with no information in miRBase. Because of this, we chose a relaxed score cut-off of 0 and minimum read depth of 3 to include a miRNA in our analyses, with the expectation that false positives would be removed during paralog clustering and alignment.

2.2.4 *Confirmation of predicted miRNAs*

We compiled a list of 776 different miRNAs from miRBase that lacked experimental validation in non-human primates prior to this study. Blastn (NCBI blast 2.2.21) was then used to find a conservative match (100% identity over at least 18 nt) of these sequences within our experimentally validated miRNA. If a match was found in at least one of our primate species, that particular miRNA was counted as now having experimental validation in a non-human primate. When determining whether a given primate genome contained a particular predicted miRNA, paralogs were collapsed into a single group and the highest score was taken, as it is difficult to distinguish between paralogs that have identical or nearly identical mature sequences.

2.2.5 *Homolog clustering*

To determine which miRNA in our data set were homologous (either as orthologs or paralogs), our mature miRNA sequences were clustered by an all-versus-all search using blastn (NCBI blast 2.2.21) at a permissive e-value (1E-02). Additional e-value cut-offs were evaluated, but given the short nature of the search queries (~22 nt mature miRNAs), it was assessed that a more permissive value was necessary to find correct matches for sequences of this length. The results were filtered for matches between sequences that had 70% identity over at least 18 nt. The sequences that remained were then clustered into groups by a custom Python script, with a

sequence being added to a group if it shared at least 70% identity to any other sequence in that group. In this way, every sequence found in the search was placed into a group or was identified as not having known homologs within our dataset.

2.2.6 *Phylogenetic analysis*

The 100 largest homology groups were selected for phylogenetic analysis, with two trees generated per group: one based on our experimentally validated mature sequence, and another based on the excised sequences as predicted by miRDeep2. These excised sequences contain the actual pre-miRNA plus ~20 nt of flanking sequence on either side. Specifically, miRDeep2 searches for the highest local stack of mature reads and excises it twice, once with 20 nt upstream and 70 nt downstream flanking sequence, and once with 70 nt upstream and 20 nt downstream flanking sequence. This is in order to determine if the mature reads occur on the 5' or 3' part of the hairpin, with miRDeep2 attempting to fold both excised sequences into stable hairpins. Thus, any excised sequence that is confirmed to include a pre-miRNA will have exactly 20 nt flanking on one side, and ~20 nt flanking the other (exact length of this flank can vary depending on the length of the mature, star, and loop sequences). This flanking sequence adds robustness to our alignments of already very short sequences. Sequences were aligned using the Fast Statistical Alignment Algorithm (FSA) (Bradley et al. 2009). Trees were then generated using RAxML with the following parameters: -f a -m GTRGAMMA -p 12345 -x 12345 -# 1000 -s -n (Stamatakis 2014). Mature and excised miRNA alignments were visualized with the Max Plank Institute's Bionformatics Toolkit (Biegert et al. 2006) and trees were visualized with Phylodendron (http://iubio.bio.indiana.edu/treeapp/). Trees were visually examined for evidence of potential adaptation, such as an excess of paralogs found only in a particular subgroup of primates, or basal primates whose sequences represent an intermediate step between paralogs.

These large paralog trees were then subdivided into groups that contained only one particular miRNA (in at least three species), based on visual assessment of both mature and excised miRNA alignments and trees. In nearly all cases, miRNAs were clustered into obvious ortholog groups. In rare cases where it was difficult to determine where a particular sequence belonged, miRNA sequences in the trees were searched within the miRBase database for the closest match. These searches clearly labeled known paralogs, confirming that our self-blast clustering worked as intended. After subdivision, each particular miRNA was realigned using FSA (Bradley et al. 2009) and trees reconstructed with RAxML (Stamatakis 2014). Each alignment was then searched for sequence variants within the mature region of a particular miRNA, as these changes are likely to have phenotypic consequences.

2.2.7 *Structural analysis*

The exact sequence of the pre-miRNA hairpin was retrieved from the excised sequences based on the folding predictions of miRDeep2. Pre-miRNA alignments were analyzed by the Vienna RNA Package's RNAz program (Lorenz et al. 2011), which predicts the secondary structure of noncoding RNA and calculates different measures of structural conservation. Thermodynamic stability of a particular secondary structure is indicated by the z-score, which is the number of standard deviations between the minimum free energy (MFE) of a sequence compared to the MFE of random sequences of the same length and base composition; RNAz circumvents this computationally intensive step by using support vector regression to estimate mean MFE and standard deviation (Gruber et al. 2009). Lower z-scores imply greater thermodynamic stability, and scores below -3 generally indicate very stable structures that are unlikely to arise by random chance. The Structure Conservation Index (SCI) is the most accurate measure of structural conservation currently available (Gruber et al. 2008). SCI compares the

average MFE of individual sequences in an alignment to a consensus MFE of that alignment. This consensus MFE is weighted by a "covariance contribution," which gives a bonus to compensatory and consistent mutations that conserve structure, and a penalty to inconsistent mutations; a negative covariance contribution indicates more compensatory mutations. A SCI close to 1 indicates structural conservation, but SCIs cannot necessarily be compared since they depend on the number of sequences in an alignment and its mean pairwise identity. In general, SCIs near or above the mean pairwise identity of the alignment indicate good candidates for conservation (Washietl 2011).

2.2.8 *Paralog confirmation*

Genomic coordinates of the human pre-miRNA for each member of the miR-320 family were obtained from miRBase, and were then used to extract pre-miRNA sequences plus 1000 nt of flanking sequence on either side from the UCSC Genome Browser. These ~2080 nt sequences were then searched against our thirteen primate genomes using blastn at an e-value of 1E-10, and were filtered for matches with a minimum of 70% identity over at least 300nt. Repetitive elements in the flanking regions that were found in thousands of locations within an individual genome were removed. For each species, the match with the highest blast score that overlapped with the pre-miRNA was identified as the paralog. Some species had no overlapping matches, but did have unique matches in the flanking regions; for these, the sequence containing the hypothetical location of the pre-miRNA (100 nt upstream and 200 nt downstream from the pre-miRNA start) was extracted with a custom perl script, aligned with FSA (Bradley et al. 2009), and visualized with the Max Plank Institute's Bionformatics Toolkit (Biegert et al. 2006) in order to investigate conservation of the pre-miRNA.

Table 1. Current primate submissions in miRBase release 21 (June 2014).

Species	**pre-miRNA**	**mature miRNA**
Human (*Homo sapiens*)	1881	2588
Chimpanzee (*Pan troglodytes*)	655	587
Bonobo (*Pan paniscus*)	88	83
Gorilla (*Gorilla gorilla*)	352	357
Orangutan (*Pongo pygmaeus*)	642	660
Siamang (*Symphalangus syndactylus*)	11	10
Rhesus macaque (*Macaca mulatta*)	619	914
Southern pig-tailed macaque (*Macaca nemestrina*)	74	70
Black snub-nosed langur (*Pygathrix bieti*)	11	9
Black-handed spider monkey (*Ateles geoffroyi*)	60	54
Common woolly monkeys (*Lagothrix lagotricha*)	48	45
White-lipped tamarin (*Saguinus labiatus*)	42	40
Ring-tailed lemur (*Lemur catta*)	16	15

Table 2. Summary of raw RNAseq reads and miRNA identified.

Species	**Raw reads**	**Previously annotated**	**Predicted novel**
Human	1091160	195	4
Chimpanzee	977409	170	22
Bonobo	944548	58	89
Gorilla	1296922	150	16
Orangutan	1189396	172	22
Gibbon	1385649	0	167
Macaque	750310	111	5
Baboon	543490	0	110
Marmoset	457234	0	132
Squirrel monkey	95190	0	45
Mouse lemur	1170417	0	145
Aye-aye	777262	0	87
Galago	1004329	0	173

Table 3. Summary of all variants found within the mature region of a miRNA across at least 3 primates.

microRNA	Variant	Position	Species with variant
miR-26b-5p	T > C	11	galago
miR-501-3p	C > T	9	mouse lemur, galago
miR-28-5p	G > A	12	squirrel monkey
miR-28-5p	G > A	10	galago
miR-34b-5p	C > A	10	mouse lemur, galago
miR-193b-5p	T > A	10	mouse lemur, aye-aye, galago
miR-532-5p	C > T	20	galago
miR-151b-3p	A > G	10	mouse lemur
miR-151b-3p	G > A	11	squirrel monkey
miR-328	T > C	22	mouse lemur
miR-299-3p	C > T	10	human
miR-224-5p	G > A	19	squirrel monkey
miR-195-5p	A > T	10	galago
miR-450b-5p	A > T	10	squirrel monkey
miR-2355-5p	C > T	15	orangutan
miR-374a-5p	T > C	21	galago
miR-539-5p	T > C	21	squirrel monkey

Table 4. Blastn results for the miR-320 family.

species|chrom	% identity	# of matching sites	# of non-matching sites	# of indels	query match start	query match end	database match start	database match end	e-value	blast score
miR-320a Present in all clades										
ggor|8	98.13	2091	25	5	1	2082	22194157	22192072	1.00E-200	3814
hsap|8	100	2082	0	0	1	2082	22103556	22101475	1.00E-200	4127
nleu|GL397351	95.22	2091	61	6	1	2082	495703	493643	1.00E-200	3370
ppan|scf1120388623448	97.52	2098	22	5	1	2082	12112241	12114324	1.00E-200	3757
ppyg|8	95.99	2096	63	8	1	2082	21814846	21812758	1.00E-200	3451
ptro|8	97.86	2103	18	5	1	2082	18444283	18442187	1.00E-200	3816
mmul|8	93.13	1820	90	10	8	1823	22230888	22229100	1.00E-200	2587
pham|scaffold11	93.32	1826	95	8	1	1823	224187	225988	1.00E-200	2627
cjac|13	88.37	1831	169	20	8	1823	15967065	15968866	1.00E-200	1830
sbol|JH378195	88.97	1759	166	13	1	1752	1610280	1608543	1.00E-200	1875
dmad|AGTM010437412[1]	87.83	337	36	4	421	754	3	337	3.00E-83	313
mmur|contig_154725	84.95	844	115	7	421	1259	4410	5246	4.00E-175	620
ogar|scaffold_73	83.88	757	115	5	363	1114	6845856	6846610	6.00E-138	498
[1]This contig is only 348 nt long, and thus sequence overlapping the pre-miRNA could not be retrieved due to quality of the genome. However, given that there is a partial match and we sequenced this miRNA in our dataset, we are confident it exists.										
miR-320b1 Present in Apes, OWM, NWM										
hsap|1	100	2079	0	0	1	2079	117213371	117215449	1.00E-200	4121
ptro|1	97.41	2087	14	6	1	2079	120063806	120061752	1.00E-200	3729
nleu|GL397313	95.57	1601	58	9	483	2079	12090969	12092560	1.00E-200	2547
ppan|scf1120388622982	98.23	1578	15	5	503	2079	33059	34624	1.00E-200	2882
ggor|1	97.78	1529	21	6	559	2079	119974413	119975936	1.00E-200	2728
ppyg|1	95.12	1168	37	7	913	2079	111486941	111485793	1.00E-200	1834
pham|scaffold15	93.14	1341	73	7	519	1850	62509	61179	1.00E-200	1897
mmul|1	92.25	1510	87	10	567	2070	120348622	120350107	1.00E-200	2026
cjac|7	86.02	1180	120	17	667	1834	152794200	152795346	1.00E-200	952
sbol|JH378161[2]	85.25	990	105	17	1048	2025	12528888	12529848	1.00E-200	718
[2]The mature miRNA is conserved in sbol, but it has a completely different 5' hairpin and upstream sequence, possibly due to a large insertion or deletion										

miR-320b2										
Present in Apes + OWM only										
hsap|1	100	2138	0	0	1	2138	224445843	224443706	1.00E-200	4238
ppan|scf1120388623324	98.04	2138	21	4	1	2138	1527109	1524993	1.00E-200	3907
ptro|1	97.94	2139	26	3	1	2138	203426022	203423901	1.00E-200	3897
ggor|1	97.95	1706	25	3	12	1717	205101547	205099852	1.00E-200	3094
ppyg|1	94.98	2133	73	10	12	2138	25488652	25490756	1.00E-200	3348
nleu|GL397374	94.6	2147	84	11	1	2138	509965	507842	1.00E-200	3291
mmul|1	91.74	2154	123	16	1	2138	146748815	146750929	1.00E-200	2809
pham|scaffold5714	92.91	988	49	9	1	983	90675	91646	1.00E-200	1356
pham|scaffold5714	93.07	750	33	4	968	1711	91942	92678	1.00E-200	1072
pham|scaffold5714	90.48	420	37	2	1721	2138	92738	93156	1.00E-139	502
miR-320c1										
Present in Apes, OWM, NWM										
ggor|18	97.7	1869	25	8	1	1865	18663103	18664957	1.00E-200	3320
hsap|18	100	2088	0	0	1	2088	19262471	19264558	1.00E-200	4139
nleu|GL397285	96.18	2095	60	9	3	2088	15937766	15935683	1.00E-200	3469
ppan|scf1120388623344	98.33	2090	20	5	1	2088	1336038	1338114	1.00E-200	3846
ppyg|18	96.79	2088	56	6	3	2088	33388949	33391027	1.00E-200	3570
ptro|18	98.99	2089	16	3	1	2088	17343979	17346063	1.00E-200	3955
mmul|18	93.39	1211	59	7	3	1205	15594341	15595538	1.00E-200	1739
mmul|18	87.65	672	51	5	1417	2088	15595760	15596399	1.00E-200	688
pham|scaffold25104	92.13	2096	117	13	3	2088	26500	28557	1.00E-200	2813
cjac|13	90.91	473	38	2	16	486	58878115	58878584	3.00E-165	587
cjac|13	88.8	1330	89	16	771	2088	58879157	58880438	1.00E-200	1415
The upstream flanking sequence for sbol matches, but does not match the actual pre-miRNA, and long string of Ns follows - ambiguous if present or not because of poor genome quality.										
miR-320c2										
Present in Apes, OWM, NWM										
ggor|18	98.86	613	5	1	866	1476	21382528	21383140	1.00E-200	1154
hsap|18	100	2050	0	0	1	2050	21900650	21902699	1.00E-200	4064
nleu|GL397285	94.9	588	19	3	1	588	13160576	13160000	1.00E-200	920
nleu|GL397285	94.95	1525	45	8	536	2050	13160008	13158506	1.00E-200	2397
ppan|scf1120388623512	97.71	2054	20	8	1	2050	22578880	22576850	1.00E-200	3673
ppyg|18	96.45	2056	60	9	1	2050	35988130	35990178	1.00E-200	3433
ptro|18	98	2054	20	4	1	2050	20023440	20025476	1.00E-200	3749
mmul|18	91.62	1455	103	8	1	1442	18262699	18264147	1.00E-200	1875
mmul|18	94.18	447	25	1	1605	2050	18264293	18264739	1.00E-200	672
pham|scaffold4919	91	2067	153	13	1	2050	48740	50790	1.00E-200	2559
cjac|13	87.66	689	74	4	676	1363	61722113	61722791	1.00E-200	674
sbol|JH378116	83.44	610	71	11	84	670	21676342	21675740	9.00E-97	359
sbol|JH378116	89.17	739	68	5	676	1412	21675640	21674912	1.00E-200	805

miR-320d1										
Present in Apes + OWM										
ggor|13	98.23	1977	25	5	1	1970	22825187	22823214	1.00E-200	3612
hsap|13	100	2048	0	0	1	2048	41303011	41300964	1.00E-200	4060
nleu|GL397327	94.62	1635	70	7	1	1622	1435486	1433857	1.00E-200	2510
ppan|scf1120388623014	98.56	1391	12	4	1	1385	1349533	1350921	1.00E-200	2575
ppyg|13	93.69	1869	88	8	1	1844	40657404	40655541	1.00E-200	2750
ptro|13_GL392075_random	98.65	1847	15	4	1	1839	1782036	1780192	1.00E-200	3443
mmul|17	91.21	1411	86	10	1	1392	19975014	19973623	1.00E-200	1790
pham|scaffold30627	91.17	1416	76	7	1	1392	14557	13167	1.00E-200	1844
miR-320d2										
Present in Apes, OWM, NWM										
ggor|X	97.37	2051	47	5	1	2048	138981785	138979739	1.00E-200	3622
hsap|X	100	2048	0	0	1	2048	140009384	140007337	1.00E-200	4060
nleu|GL397463	95.36	2071	57	8	1	2048	210560	208506	1.00E-200	3342
ppan|scf1120388622949	98.88	2050	16	4	1	2048	130696	128652	1.00E-200	3856
ppyg|X	96.1	2077	52	6	1	2048	140668810	140666734	1.00E-200	3463
ptro|X	98.83	2050	20	3	1	2048	141737651	141735604	1.00E-200	3852
mmul|X	93.7	746	39	3	1	746	140284397	140283660	1.00E-200	1092
mmul|X	94.78	1341	58	6	718	2048	140283657	140282319	1.00E-200	2068
pham|scaffold8723	94.11	1274	71	4	1	1273	47601	48871	1.00E-200	1899
pham|scaffold8723	95.75	729	27	3	1322	2048	48922	49648	1.00E-200	1178
cjac|X	87.2	1148	121	11	1	1140	128747924	128746795	1.00E-200	1053
cjac|X	89.04	858	70	8	1136	1975	128746497	128745646	1.00E-200	924
sbol|JH378238	87.56	1125	123	8	18	1131	450418	451536	1.00E-200	1074
sbol|JH378238	91.03	936	61	5	1136	2048	451844	452779	1.00E-200	1185
miR-320e										
Present in Apes + OWM only										
ggor|19	96.43	1316	32	3	1	1310	44087413	44086107	1.00E-200	2256
ggor|19	97.74	665	11	2	1351	2013	44086062	44085400	1.00E-200	1187
hsap|19	100	2053	0	0	1	2053	47213602	47211550	1.00E-200	4070
ppan|scf1120388623374	98.08	1251	19	4	777	2023	524071	525320	1.00E-200	2260
ppyg|19	95.18	1079	50	1	157	1233	48077543	48076465	1.00E-200	1721
ppyg|19[3]	93.29	1460	61	6	572	2022	48079368	48077937	1.00E-200	2131
ptro|19	98.49	1858	24	2	1	1857	51645333	51643479	1.00E-200	3449
mmul|19	89.69	1213	106	6	1	1203	52731713	52730510	1.00E-200	1392
mmul|19	89	309	23	1	1239	1547	52730486	52730189	6.00E-96	355
pham|scaffold13982	89.62	1387	122	7	1	1374	52648	54025	1.00E-200	1582

[3]Appears to have an additional paralog upstream in orangutan (ppyg). Both copies are 100% identical.

Table 5. Coriell Catalog IDs of the fibroblast cell cultures used in RNA extraction.

Species name	**Common name**	**Coriell Catalog ID**
Homo sapiens	Human	GM03651
Pan troglodytes	Chimpanzee	S003647
Pan paniscus	Bonobo	PR00051
Gorilla gorilla	Lowland gorilla	PR00053
Pongo pygmaeus	Sumatran orangutan	PR01110
Nomascus leucogenys	White-cheeked gibbon	PR00712
Macaca mulatta	Rhesus macaque	PR00418
Papio hamadryas	Guinea baboon	PR00559
Callithrix geoffroyi	White-fronted marmoset	PR00789
Saimiri boliviensis	Bolivian squirrel monkey	PR00474
Microcebus murinus	Gray mouse lemur	PR00275
Daubentonia madagascariensis	Aye-aye	PR00228
Otolemur garnettii	Small-eared galago	PR00048

Chapter 3. HUMAN MIRNA VARIATION

Given the high levels of conservation of miRNA sequence across primate evolution, variation found within human populations may result in disease and could represent recent evolutionary events. Previous studies have investigated miRNA diversity within humans, but their ability to discover variants was dependent on the availability of SNV and miRNA data at the time (Figure 13). The first study to survey human miRNA diversity was in 2005; they sequenced 173 pre-miRNAs from 96 Japanese individuals and found 10 variants, none in the seed region (Iwai and Naraba 2005). The search was expanded in 2006 to all known SNVs in dbSNP within 474 pre-miRNAs; 65 SNVs were identified, three of which were in the seed region (Saunders et al. 2007). A number of papers continued to extend the current state of known human miRNA variation, and unsurprisingly find more variation as more individuals from diverse populations have been sequenced (Duan et al. 2009, Hiard et al. 2010, Bhartiya et al. 2011, Carbonell et al. 2012, Gong et al. 2012, Liu et al. 2012, Zorc et al. 2012).

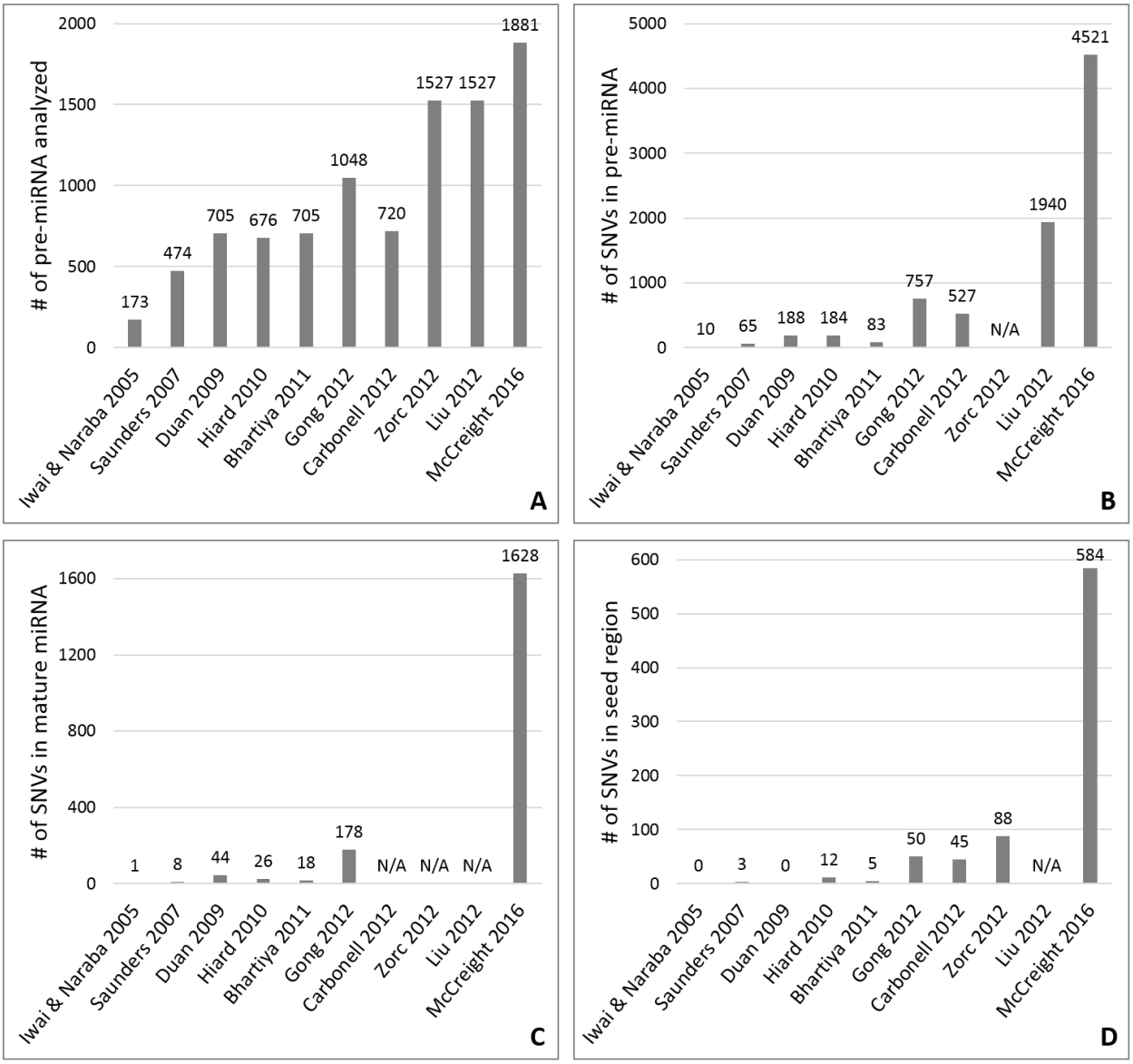

Figure 13. Past surveys of human miRNA variation compared to the current study. A. Total number of pre-miRNA analyzed. B. Number of SNVs found anywhere within the pre-miRNA. C. Number of variants found anywhere within the mature miRNA. D. Number of variants found within the eight nucleotide seed region.

Broad analyses of miRNA variation in human populations has stalled since 2012, despite the rapid expansion of large human genomic datasets in recent years. The number of known

human pre-miRNAs in miRBase has increased from 1,527 (release 18) to 1,881 (release 21) (Kozomara & Griffiths-Jones 2011). Likewise, the number of human SNVs in dbSNP has tripled, growing from 178,140,935 (build 135) to 545,361,347 (build 147) (Sherry et al. 2001). The 1000 Genomes Project has also been updated since last studied (Carbonell et al. 2012); Phase 1 included 38.2 million SNVs identified in 1092 individuals, while Phase 3 now includes 84.4 million SNVs from 2504 individuals (1000 Genomes Project Consortium 2015). Unlike many studies that only include European populations, the 1000 Genomes Project includes 26 populations from around the world that represent the most broad, comprehensive genome-wide scan of human variation to date. Here we summarize the current state of human variation in the 1000 Genomes Project Phase 3 release.

3.1 RESULTS

3.1.1 *Summary of human variants*

A total of 4521 SNVs were identified in 1578 pre-miRNAs; the remaining 303 pre-miRNA were invariant (Figure 14). Most of these variants are rare: 89.4% (4041/4521) of the SNVs have a minor allele frequency (MAF) of less than 1%, with 74.7% (3020/4041) of these rare variants being singletons. Only 5.9% (267/4521) of SNVs had a MAF greater than 5%. Of the 4521 SNVs identified, 2893 were located within the hairpin but outside of the mature sequence, 1044 occurred within the mature sequence but outside of the seed region, and 584 occurred within the seed; Table 6 lists all non-singleton seed region SNVs in Hardy Weinberg Equilibrium (HWE). The approximately 5:2:1 ratio of variants is the same regardless of MAF ($X^2 = 2.78$, df = 6, p = 0.8359). This suggests a narrow range of selection coefficients of purifying selection on different regions of the miRNA hairpin, concordant with the established knowledge that the seed region is most highly conserved, followed by the mature region.

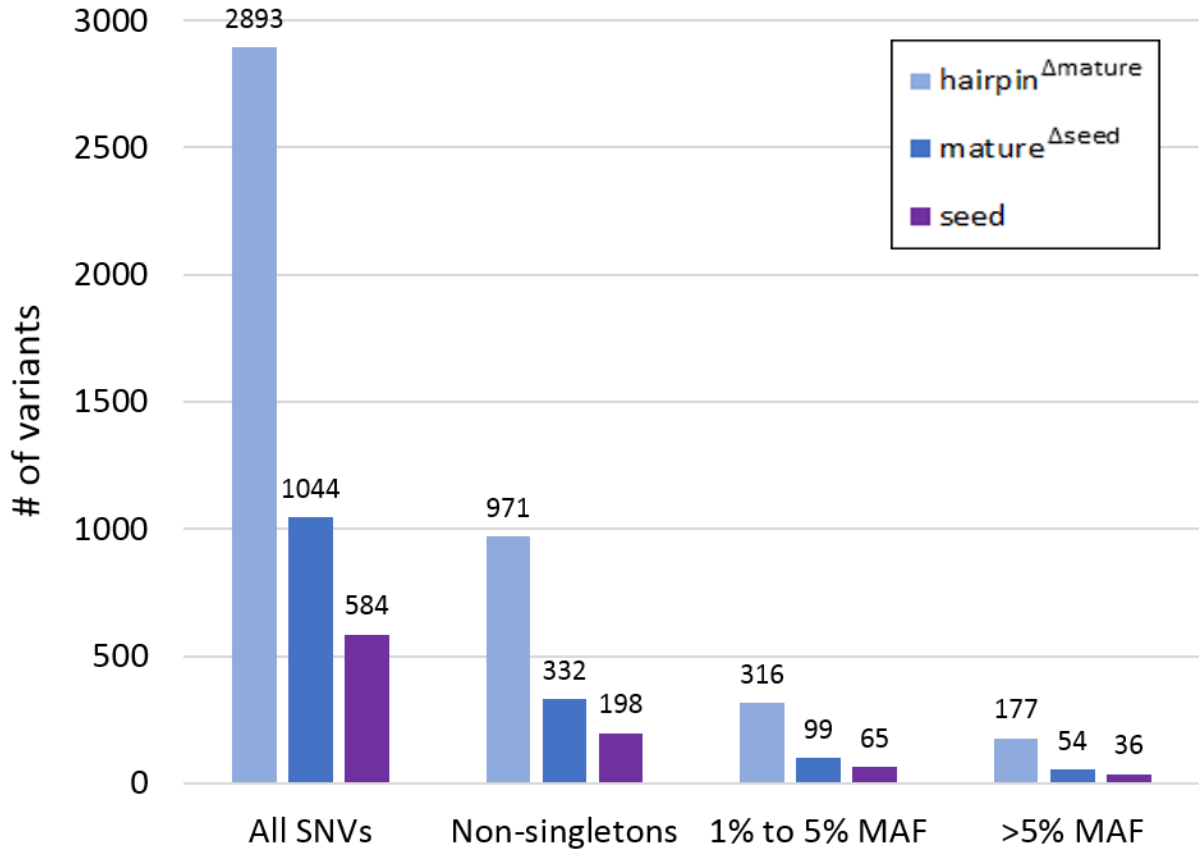

Figure 14. Location of SNVs within a pre-miRNA for different MAF cutoffs. Hairpin$^{\Delta mature}$ represents variants found in the stem and loop of the pre-miRNA, excluding the mature sequence. Mature$^{\Delta seed}$ represents variants found within the mature sequence but outside of the seed region.

Indels (insertions or deletions) were also identified: 181 indels (79 insertions, 102 deletions) were found in 166 miRNA. 142 indels remain when singletons are omitted, 23 of which were in the seed region (Table 7). Indels occurring closer to the beginning of the seed region would have drastic effects on target repertoire, as an indel would alter the whole downstream seed region instead of just a single nucleotide. Most indels are small (Figure 15),

likely because larger events are more disruptive to the overall proper folding and processing of a miRNA hairpin. Large indels tend to occur near the ends of a miRNA, where they would be less deleterious. Furthermore, all insertions larger than 5 nt exactly match the downstream sequence of the human reference genome, suggesting that these "variants" are in fact errors and highlighting the importance of generating sequence information directly from small RNAs, rather than homology from whole genome sequencing.

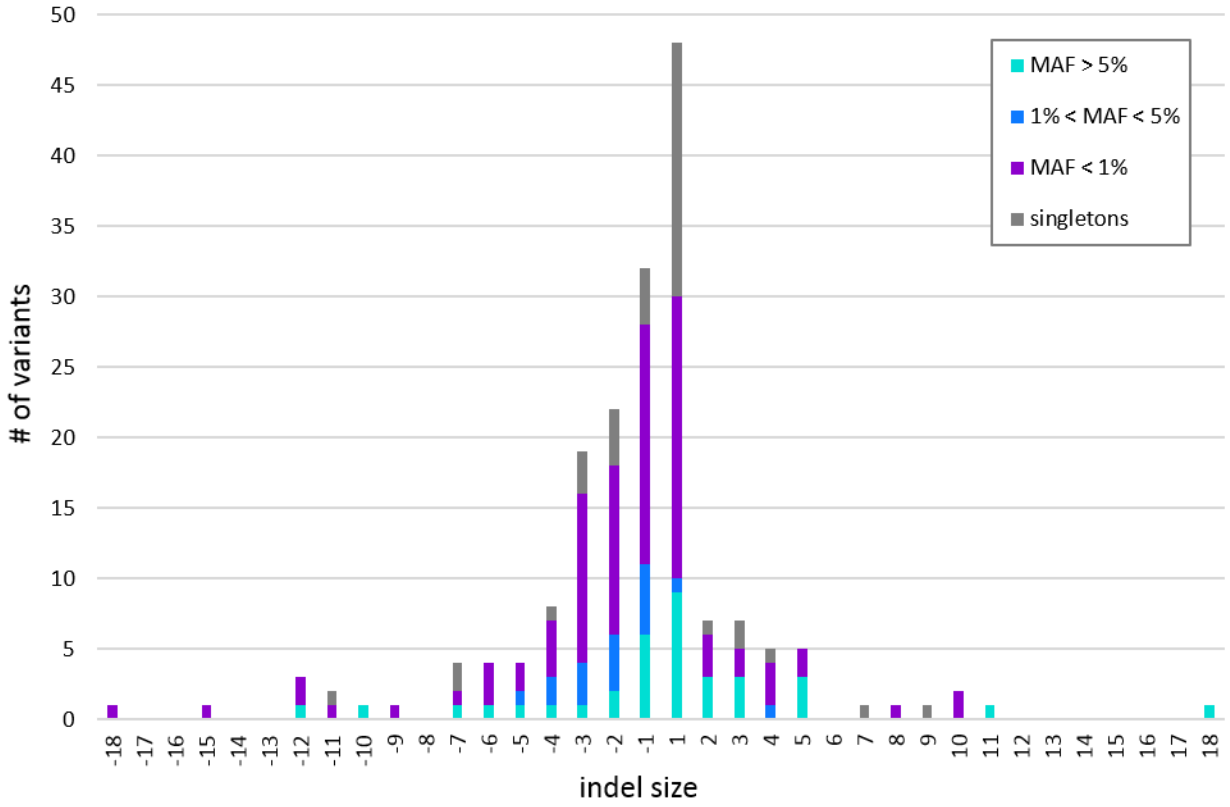

Figure 15. Number of indels identified plotted by size of the insertion or deletion.

The location of SNVs within the mature sequence for a given MAF cutoff is illustrated in Figure 16. Positions 1 through 5 of the seed region have less variation, as expected given their importance in complementarily base pairing with mRNA targets, but positions 6 through 8 appear to be more variable. Positions 9 through 11 appear as conserved as most of the seed

region, highlighting the importance centered sites may play in target recognition (Shin et al. 2010, Martin et al. 2014). While 3' binding at positions 13-16 is sometimes known to occur, this region does not appear to be highly conserved within the 1000 Genomes Project dataset, regardless of minor allele frequency. Compared to the distribution of primate variants across the mature sequence from Chapter 2 (Figure 6), human variants are more likely to occur in typically conserved regions. Variation between distantly related primate species has undergone purifying selection for tens of millions of years; conversely, intra-population human variation represents rare and possibly deleterious variants that have not yet been selected against, and thus variants are more likely to be found in typically conserved regions of miRNA.

Figure 16. Location of variants within mature miRNA of humans from the 1000 Genomes Project. The yellow line represents all SNVs (no insertions or deletions), and the other lines represent progressive filtering steps: removal of singletons (red), removal of variants with less than 1% minor allele frequency (purple), and removal of variants with less than 5% minor allele frequency. The 5' end of the mature miRNA has an 8 nt "seed region" in positions 1 through 8

that complementary base-pairs with the 3' untranslated region (UTR) of messenger RNA (mRNA). Positions 13 – 16 are typically the second most conserved region, as some miRNAs undergo 3' binding.

For each SNV, F_{ST} was calculated for each pairwise combination of super populations in the 1000 Genomes Project: African (AFR), Admixed American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS). F_{ST} is a measure of differentiation between subpopulations and ranges from 0 to 1, where smaller values indicate similar allele frequencies between populations (Holsinger & Weir 2009). Most F_{ST} values were near zero, with 0.0782 as the largest F_{ST} (Figure 17). Given that the average F_{ST} between human subpopulations is approximately 0.10 to 0.13 (Bhatia et al.), human miRNA variants do not appear to have any population substructure. These extremely small F_{ST} values, taken into account alongside the lack of enrichment of seed region SNVs among variants with a MAF > 5%, indicate that little positive selection has occurred in miRNA among humans since the split from chimpanzees.

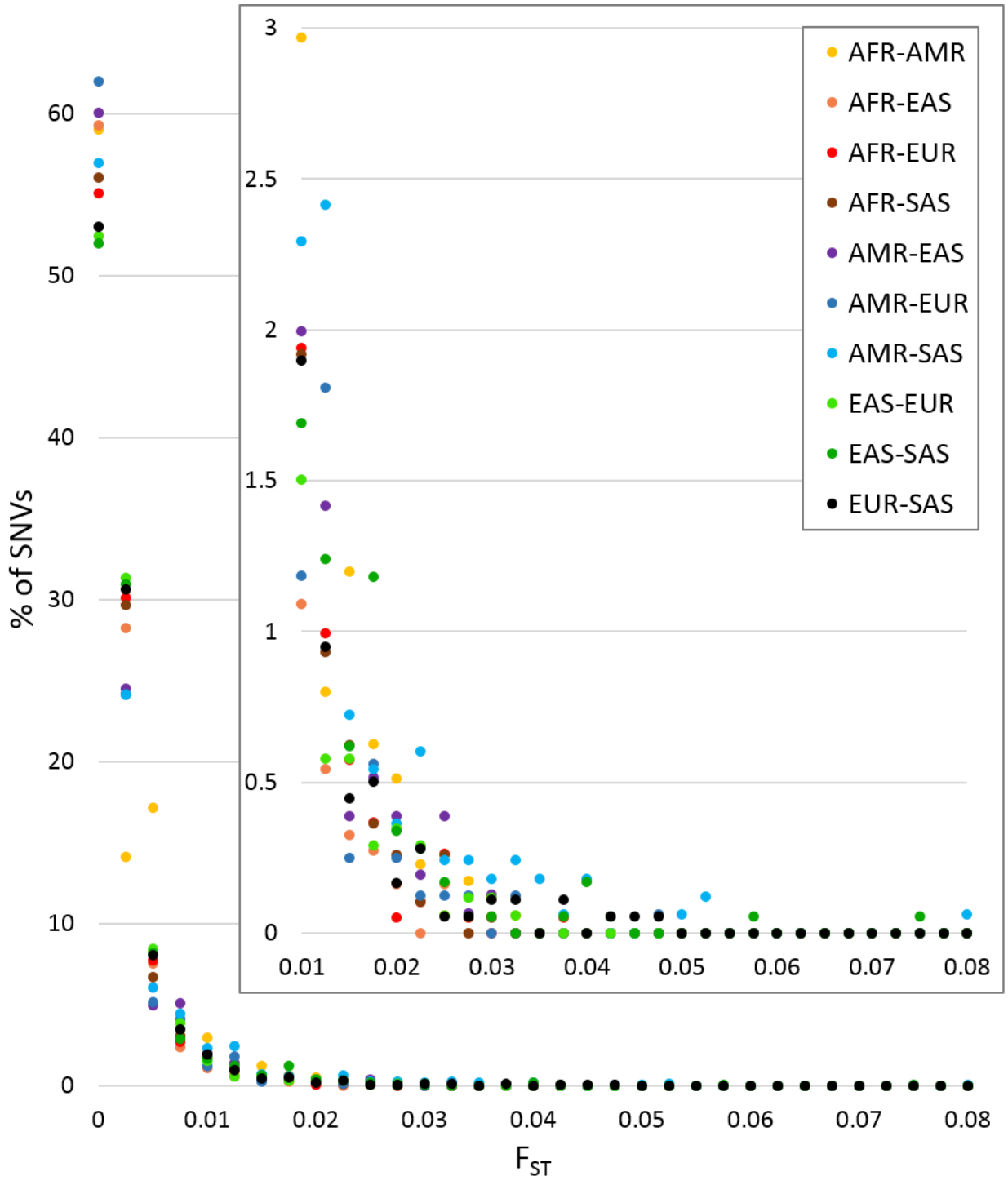

Figure 17. Distribution of F_{ST} values for human miRNA SNVs. The insert zooms in on SNVs with F_{ST}s > 0.01 for clarity.

3.1.2 *Variants deviating from Hardy-Weinberg Equilibrium*

Variant sites were assessed for HWE. 244 SNVs were identified as significantly deviating from expected HWE ($p < 0.05$); 27 of these sites were located within the seed region (Table 8). Deviation from HWE is suggestive of selection, but may also be caused by genetic drift or misaligned paralogs that have not yet been identified. In either of these scenarios, a change within the seed region would result in a change in the mRNA target repertoire. Targets were predicted for each reference and alternate allele, and gene ontology (GO) annotations enriched for each allele were identified (Table 9). Similarity scores range from 0 to 1, where 1 represents complete similarity of enriched GO terms.

3.1.3 *Variants involved in disease*

PhenomiR, a manually curated database of miRNA disease association studies, was used to identify disease associations for three subsets of miRNA seed region variants: SNVs with a MAF of 5% or more, SNVs with significant divergence from HWE, and indels (Ruepp et al. 2012). Few variants were associated with diseases, most of which were cancer (Table 10). However, this does not necessarily indicate that the majority of variants are benign, but rather that they have not yet been assessed: despite PhenomiR being the most recent miRNA disease association database available, it was last updated in 2011. Due to their downstream effects, indel variants are particularly likely to alter the target repertoire of a miRNA, and thus are promising candidates for future functional validation.

3.1.4 *Comparison to non-human primates*

We searched our subset of human SNVs excluding singletons for variants within the four miRNA families important in primate evolution discussed in Chapter 2. miR-299 and miR-2355 had no such variation. Although the miR-320 family had five SNVs, each variant matched the reference allele for another member of the 320 family, making it difficult to distinguish between a true polymorphism and a misalignment. miR-501 had two SNVs, both within the mature sequence (A2G and G10A); although the seed region variant is relatively rare (MAF = 0.0059), it does result in a change in target repertoire (Table 11).

3.2 METHODS

We retrieved a list of human mature and pre-miRNAs coordinates from miRBase (release 21 for GRCh38). Variant call files were downloaded for Phase 3 of the 1000 Genomes Project, and variants within miRNA were analyzed with vcftools (Danecek et al. 2011) and wrapped within a custom Python script. Deviations from HWE and F_{ST} (exact test) were calculated with vcftools. Copy number variants were removed and filtering by MAF was executed in Excel. Indels were checked against the human reference genome using the UCSC Genome Browser. miRmut2GO was used to predict targets via TargetScan for each reference and alternate allele and identify which GO annotations were enriched for each set of targets ($p < 0.05$) (Bhattacharya and Cui 2015). Disease associations were identified with PhenomiR (Ruepp et al. 2012).

Table 6. Seed region SNVs in HWE, excluding singletons.

miRNA	Chr:position	position in seed	SNV	MAF
hsa-mir-1178	chr12:119713688	8	T/C	0.1055
hsa-mir-1178	chr12:119713688	8	T/G	0.00721371
hsa-mir-1181	chr19:10403518	8	G/A	0.000901713
hsa-mir-1203	chr17:48156504	3	G/A	0.000901713
hsa-mir-122	chr18:58451126	3	C/T	0.00360685
hsa-mir-1236	chr6:31956854	7	G/C	0.00360685
hsa-mir-1255b-1	chr4:36426426	1	G/A	0.0856628
hsa-mir-1255b-2	chr1:167998666	2	G/C	0.000901713
hsa-mir-1292	chr20:2652820	2	C/T	0.00901713
hsa-mir-1302-1	chr12:112695096	7	G/A	0.0198377
hsa-mir-1322	chr8:10825388	6	A/G	0.00225428
hsa-mir-1469	chr15:96333267	7	G/C	0.00676285
hsa-mir-1539	chr18:49487407	6	C/T	0.00135257
hsa-mir-1973	chr4:116299753	4	G/A	0.000901713
hsa-mir-199a-1	chr19:10817443	8	A/G	0.00270514
hsa-mir-20b	chrX:133303896	7	C/T	0.00118694
hsa-mir-212	chr17:2050343	8	G/A	0.003156
hsa-mir-2392	chr14:100814555	5	A/G	0.00270514
hsa-mir-2682	chr1:98045291	2	C/T	0.0193868
hsa-mir-3119-1	chr1:170151453	2	C/T	0.000901713
hsa-mir-3126	chr2:69103691	1	T/C	0.000901713
hsa-mir-3147	chr7:57405036	7	G/T	0.000901713
hsa-mir-3150a	chr8:95072962	1	C/T	0.00135257
hsa-mir-3175	chr15:92904413	6	A/G	0.00405771
hsa-mir-3193	chr20:31607192	7	G/A	0.00135257
hsa-mir-3196	chr20:63238789	2	G/A	0.0275023
hsa-mir-328	chr16:67202389	1	C/G	0.000901713
hsa-mir-3615	chr17:74748666	4	C/T	0.00270514
hsa-mir-3620	chr1:228097290	6	C/T	0.108206
hsa-mir-3622a	chr8:27701697	8	G/A	0.297115
hsa-mir-3655	chr5:140647850	7	C/T	0.00856628
hsa-mir-3682	chr2:53849189	3	T/C	0.00586114
hsa-mir-3689b	chr9:134850221	7	T/A	0.0261497
hsa-mir-3689d-1	chr9:134849674	8	C/T	0.00135257
hsa-mir-3689d-2	chr9:134850346	8	C/T	0.00405771
hsa-mir-3689f	chr9:134850800	5	T/C	0.00360685

hsa-mir-3690-1	chrX:1412821	2	C/G	0.00721371
hsa-mir-378i	chr22:41923287	5	C/T	0.00135257
hsa-mir-3916	chr1:247202035	7	T/C	0.00135257
hsa-mir-412	chr14:101065469	5	C/T	0.00135257
hsa-mir-4257	chr1:150551992	6	G/A	0.00405771
hsa-mir-4284	chr7:73711334	8	C/T	0.025248
hsa-mir-4290	chr9:90023467	3	G/A	0.00180343
hsa-mir-4293	chr10:14383222	5	G/C	0.0572588
hsa-mir-4305	chr13:39664119	7	G/T	0.0144274
hsa-mir-4305	chr13:39664120	6	T/C	0.0130748
hsa-mir-4315-1	chr17:45475423	3	C/T	0.00180343
hsa-mir-4322	chr19:10230461	7	G/A	0.00766456
hsa-mir-4423	chr1:85133843	2	T/C	0.00225428
hsa-mir-4433b	chr2:64340782	8	C/A	0.00405771
hsa-mir-4433b	chr2:64340782	8	C/G	0.148332
hsa-mir-4440	chr2:239068848	3	A/G	0.00135257
hsa-mir-4460	chr5:129397131	6	C/A	0.00180343
hsa-mir-4467	chr7:102471476	5	G/A	0.00631199
hsa-mir-4472-2	chr12:116428308	4	C/T	0.000901713
hsa-mir-450a-1	chrX:133674439	6	G/A	0.00118694
hsa-mir-450a-2	chrX:133674612	4	A/G	0.00237389
hsa-mir-4517	chr16:28958592	6	T/C	0.00135257
hsa-mir-4520-2	chr17:6655449	1	C/T	0.58972
hsa-mir-4532	chr20:57895400	2	C/T	0.0225428
hsa-mir-4532	chr20:57895403	5	G/A	0.0306583
hsa-mir-4537	chr14:105859550	4	C/G	0.000901713
hsa-mir-4638	chr5:181222632	1	T/G	0.00360685
hsa-mir-4641	chr6:41598771	8	G/A	0.0162308
hsa-mir-4646	chr6:31701084	8	T/C	0.00225428
hsa-mir-466	chr3:31161729	8	A/G	0.00225428
hsa-mir-4661	chr8:91205534	4	G/T	0.0207394
hsa-mir-4666a	chr1:228462130	6	A/G	0.000901713
hsa-mir-4669	chr9:134379455	7	C/G	0.00135257
hsa-mir-4669	chr9:134379456	8	G/A	0.00405771
hsa-mir-4670	chr9:92528050	2	T/G	0.000901713
hsa-mir-4676	chr10:72721042	5	C/T	0.00135257
hsa-mir-4679-1	chr10:89063350	7	A/T	0.000901713
hsa-mir-4685	chr10:98431352	6	C/T	0.000901713
hsa-mir-4695	chr1:18883265	7	C/T	0.0130748

hsa-mir-4706	chr14:65044700	4	G/A	0.0135257
hsa-mir-4706	chr14:65044703	7	G/T	0.00676285
hsa-mir-4721	chr16:28843941	8	A/C	0.000901713
hsa-mir-4731	chr17:15251649	7	T/A	0.453562
hsa-mir-4737	chr17:60043095	2	A/G	0.0184851
hsa-mir-4738	chr17:75784546	6	G/A	0.00225428
hsa-mir-4741	chr18:22933413	7	G/A	0.000901713
hsa-mir-4741	chr18:22933414	8	T/G	0.000901713
hsa-mir-4743	chr18:48670606	5	C/T	0.00180343
hsa-mir-4747	chr19:4932725	7	C/T	0.000901713
hsa-mir-4749	chr19:49854632	1	C/T	0.003156
hsa-mir-4749	chr19:49854633	2	G/A	0.000901713
hsa-mir-4756	chr20:54068432	8	A/T	0.00135257
hsa-mir-4762	chr22:45760532	1	C/G	0.0171326
hsa-mir-4763	chr22:46113584	1	C/T	0.00135257
hsa-mir-4782	chr2:113721314	6	C/T	0.000901713
hsa-mir-4802	chr4:40502106	2	T/C	0.00450857
hsa-mir-4804	chr5:72878605	6	C/G	0.840848
hsa-mir-499a	chr20:34990448	4	A/G	0.158702
hsa-mir-499a	chr20:34990452	8	C/T	0.00135257
hsa-mir-5001	chr2:232550551	3	C/T	0.00135257
hsa-mir-501	chrX:49774381	2	A/G	0.00593472
hsa-mir-5096	chr4:78820771	8	C/T	0.000901713
hsa-mir-518d	chr19:53734935	7	G/A	0.00180343
hsa-mir-5197	chr5:143679930	8	G/T	0.00270514
hsa-mir-526a-1	chr19:53706273	8	G/A	0.00135257
hsa-mir-548al	chr11:74399308	8	A/G	0.177187
hsa-mir-548f-3	chr5:110513855	8	C/T	0.000901713
hsa-mir-548t	chr4:173268209	5	A/C	0.00450857
hsa-mir-550b-1	chr7:30289828	6	G/A	0.00135257
hsa-mir-551a	chr1:3560728	3	C/T	0.00450857
hsa-mir-552	chr1:34669669	2	A/G	0.00360685
hsa-mir-5585	chr1:32086991	6	T/G	0.00225428
hsa-mir-5586	chr14:59646979	5	C/T	0.00225428
hsa-mir-5589	chr19:10038360	7	G/A	0.00541028
hsa-mir-5692b	chr21:42951004	2	T/C	0.862038
hsa-mir-5739	chr22:28459928	4	G/A	0.000901713
hsa-mir-575	chr4:82753367	4	G/A	0.000901713
hsa-mir-590	chr7:74191216	4	C/T	0.00360685

hsa-mir-593	chr7:128081882	7	C/T	0.0198377
hsa-mir-598	chr8:11035275	5	A/G	0.000901713
hsa-mir-602	chr9:137838436	3	C/T	0.00360685
hsa-mir-604	chr10:29545030	8	C/T	0.00811542
hsa-mir-6078	chr10:3991179	2	C/T	0.000901713
hsa-mir-6083	chr3:124374413	6	A/G	0.000901713
hsa-mir-615	chr12:54033969	3	G/T	0.00135257
hsa-mir-627	chr15:42199650	2	A/C	0.0680794
hsa-mir-642b	chr19:45674997	3	A/C	0.00721371
hsa-mir-6499	chr5:151522138	7	C/T	0.0946799
hsa-mir-6511b-2	chr16:15134093	6	G/A	0.000901713
hsa-mir-6515	chr19:12940522	2	C/T	0.000901713
hsa-mir-6716	chr11:118644051	4	G/A	0.000901713
hsa-mir-6717	chr14:21023372	5	C/T	0.000901713
hsa-mir-6717	chr14:21023373	4	G/A	0.0266005
hsa-mir-6720	chr6:1390341	5	G/A	0.00360685
hsa-mir-6721	chr6:32170102	5	G/A	0.00225428
hsa-mir-6726	chr1:1296122	8	A/G	0.000901713
hsa-mir-6726	chr1:1296127	3	G/A	0.1055
hsa-mir-6729	chr1:12029168	6	G/A	0.0139766
hsa-mir-6736	chr1:145850601	7	C/T	0.00135257
hsa-mir-6736	chr1:145850603	5	C/T	0.000901713
hsa-mir-6742	chr1:228397061	7	C/T	0.00405771
hsa-mir-6743	chr11:209390	8	C/T	0.00135257
hsa-mir-6746	chr11:61878271	3	C/T	0.00541028
hsa-mir-6753	chr11:68044933	2	G/A	0.000901713
hsa-mir-6761	chr12:111799883	2	C/T	0.00270514
hsa-mir-6761	chr12:111799887	6	C/T	0.000901713
hsa-mir-6763	chr12:132582046	6	C/T	0.758341
hsa-mir-6788	chr18:10759603	3	G/A	0.00180343
hsa-mir-6808	chr1:1339702	2	T/C	0.00180343
hsa-mir-6810	chr2:218341922	7	A/G	0.0752931
hsa-mir-6823	chr3:48549975	7	A/C	0.0901713
hsa-mir-6839	chr7:64679160	6	T/C	0.00135257
hsa-mir-6845	chr8:143837807	5	C/T	0.00135257
hsa-mir-6863	chr16:56904332	6	G/A	0.0468891
hsa-mir-6867	chr17:40193646	4	T/C	0.00135257
hsa-mir-6870	chr20:10649689	2	C/A	0.000901713
hsa-mir-6879	chr11:65018557	8	C/T	0.00631199

hsa-mir-6883	chr17:8145059	8	C/A	0.00360685
hsa-mir-6891	chr6:31355243	6	T/C	0.264202
hsa-mir-7108	chr19:2434993	8	G/A	0.000901713
hsa-mir-7151	chr10:67403364	8	G/T	0.00135257
hsa-mir-7153	chr18:11654898	5	T/C	0.000901713
hsa-mir-759	chr13:52810074	2	C/G	0.00270514
hsa-mir-7703	chr14:24143511	1	A/G	0.00135257
hsa-mir-7854	chr16:81533949	8	A/G	0.386384
hsa-mir-8061	chr19:54645323	5	G/T	0.00180343
hsa-mir-8070	chr11:11783210	3	C/T	0.00901713
hsa-mir-8073	chr13:110340972	5	G/T	0.000901713
hsa-mir-8074	chr19:51206956	7	G/A	0.0108206
hsa-mir-8081	chr9:106600997	8	C/T	0.00360685
hsa-mir-8088	chrX:52079770	5	C/T	0.00356083
hsa-mir-933	chr2:175167656	8	T/C	0.00405771
hsa-mir-936	chr10:104048170	3	T/C	0.00631199
hsa-mir-941-3	chr20:63919612	6	G/A	0.00721371

Table 7. Indels found within human pre-miRNA, excluding singletons.

miRNA	Chr:position	Variant location	Reference/Alternate alleles	indel size	MAF
hsa-mir-562	chr2:232172719	seed (7)	GCTGTACCATTTGCACTCC/G	-18	0.00631199
hsa-mir-593	chr7:128081929	hairpin	GTGCTGGGTTTGTCTC/G	-15	0.00135257
hsa-mir-3945	chr4:184851078	mature (11)	TCCTATGCCCTCC/T	-12	0.0843102
hsa-mir-6071	chr2:85783627	hairpin	CAGTAAGCTAGGG/C	-12	0.00450857
hsa-mir-466	chr3:31161769	hairpin	AACACACATATAC/A	-12	0.000901713
hsa-mir-4472-1	chr8:142176402	seed (7)	GGGTGTTGTTTT/G	-11	0.00225428
hsa-mir-548a-3	chr8:104484407	hairpin	CATTGAAAGTA/C	-10	0.0532011
hsa-mir-6786	chr17:81693813	seed (5)	TGGGGCCGGA/T	-9	0.000901713
hsa-mir-3652	chr12:103930452	hairpin	AGGGGTGG/A	-7	0.102344
hsa-mir-641	chr19:40282587	hairpin	TAGAGGAC/T	-7	0.000901713
hsa-mir-548i-2	chr4:9556194	hairpin	TAGAAGG/T	-6	0.348963
hsa-mir-620	chr12:116148604	hairpin	GATATCT/G	-6	0.00766456
hsa-mir-3688-1	chr4:159128882	hairpin	TTGAAAG/T	-6	0.000901713
hsa-mir-3688-2	chr4:159128882	mature (22)	TTGAAAG/T	-6	0.000901713
hsa-mir-548h-4	chr8:27048916	hairpin	TTAAAG/T	-5	0.148783
hsa-mir-920	chr12:24212423	hairpin	AGTTGT/A	-5	0.03156
hsa-mir-3924	chr10:57304555	hairpin	ATTTAT/A	-5	0.00180343
hsa-mir-559	chr2:47377689	hairpin	TTAAAG/T	-5	0.000901713
hsa-mir-548aj-2	chrX:37883167	mature (19)	AAAGT/A	-4	0.144214
hsa-mir-302c	chr4:112648383	seed (6)	CACTT/C	-4	0.012624
hsa-mir-373	chr19:53788735	hairpin	TTGTC/T	-4	0.0103697
hsa-mir-6763	chr12:132582020	mature (19)	GCAGA/G	-4	0.00856628
hsa-mir-4633	chr5:129097725	hairpin	AAATG/A	-4	0.00180343
hsa-mir-8077	chr19:42351128	hairpin	AGGGT/A	-4	0.00135257
hsa-mir-492	chr12:94834422	hairpin	CATCG/C	-4	0.000901713
hsa-mir-550a-3	chr7:29680776	hairpin	TACA/T	-3	0.0716862
hsa-mir-4483	chr10:113777997	hairpin	AAAC/A	-3	0.0153291
hsa-mir-550a-1	chr7:30289837	mature (23)	CTGT/C	-3	0.0117223
hsa-mir-550b-1	chr7:30289837	hairpin	CTGT/C	-3	0.0117223
hsa-mir-6127	chr1:22633256	hairpin	AAAG/A	-3	0.00946799
hsa-mir-1302-7	chr8:141786307	hairpin	ATGT/A	-3	0.00450857
hsa-mir-3607	chr5:86620571	hairpin	ACTC/A	-3	0.00450857
hsa-mir-5692a-1	chr7:97963710	hairpin	ATAT/A	-3	0.00270514
hsa-mir-5692a-2	chr8:12719179	hairpin	ATAT/A	-3	0.00180343
hsa-mir-6864	chr17:4969719	seed (4)	TCAC/T	-3	0.00135257
hsa-mir-4517	chr16:28958590	seed (4)	TATG/T	-3	0.00135257
hsa-mir-4738	chr17:75784588	seed (1)	TAAG/T	-3	0.00135257
hsa-mir-3926-1	chr8:12727263	hairpin	CGCT/C	-3	0.00135257
hsa-mir-3926-2	chr8:12727263	hairpin	CGCT/C	-3	0.00135257
hsa-mir-3185	chr17:48724460	mature (9)	CCTT/C	-3	0.00135257
hsa-mir-6127	chr1:22633266	hairpin	AAGG/A	-3	0.000901713
hsa-mir-3938	chr3:55852545	hairpin	TAA/T	-2	0.176285

hsa-mir-3171	chr14:27633270	hairpin	CTA/C	-2	0.176285
hsa-mir-4640	chr6:30890954	seed (6)	CCT/C	-2	0.0387737
hsa-mir-631	chr15:75353624	hairpin	CCT/C	-2	0.0369702
hsa-mir-516b-2	chr19:53725519	hairpin	CTT/C	-2	0.0184851
hsa-mir-1303	chr5:154685821	hairpin	TTA/T	-2	0.0171326
hsa-mir-3143	chr6:27147673	hairpin	CTT/C	-2	0.00856628
hsa-mir-6133	chr7:133290981	hairpin	CAG/C	-2	0.00631199
hsa-mir-4740	chr17:81400728	mature (10)	CCT/C	-2	0.00450857
hsa-mir-466	chr3:31161779	hairpin	TAC/T	-2	0.00180343
hsa-mir-3175	chr15:92904411	seed (4)	GGA/G	-2	0.00180343
hsa-mir-516a-1	chr19:53756776	mature (21)	TTC/T	-2	0.00135257
hsa-mir-548aa-2	chr17:67471514	hairpin	CTG/C	-2	0.00135257
hsa-mir-548d-2	chr17:67471514	mature (12)	CTG/C	-2	0.00135257
hsa-mir-3917	chr1:25906415	hairpin	ACT/A	-2	0.00135257
hsa-mir-4527	chr18:47380521	mature (19)	CTG/C	-2	0.000901713
hsa-mir-3678	chr17:75406086	mature (10)	ACT/A	-2	0.000901713
hsa-mir-4296	chr10:125032828	hairpin	AAC/A	-2	0.000901713
hsa-mir-4794	chr1:64579919	hairpin	TA/T	-1	0.741659
hsa-mir-548az	chr8:119325212	hairpin	TG/T	-1	0.666817
hsa-mir-4472-2	chr12:116428308	seed (4)	CA/C	-1	0.580703
hsa-mir-1303	chr5:154685822	hairpin	TA/T	-1	0.576195
hsa-mir-3908	chr12:123536488	hairpin	AT/A	-1	0.17899
hsa-mir-1250	chr17:81133217	hairpin	CT/C	-1	0.0527502
hsa-mir-1260a	chr14:77266279	hairpin	CA/C	-1	0.0333634
hsa-mir-6838	chr7:44073397	seed (1)	TC/T	-1	0.0198377
hsa-mir-4300	chr11:81890820	hairpin	GC/G	-1	0.0198377
hsa-mir-4289	chr9:88745840	hairpin	CT/C	-1	0.0153291
hsa-mir-4540	chr9:36864291	hairpin	TC/T	-1	0.0112714
hsa-mir-6766	chr15:89326742	mature (22)	TG/T	-1	0.00991885
hsa-mir-4306	chr13:99643108	hairpin	CT/C	-1	0.00856628
hsa-mir-320e	chr19:46709312	hairpin	TC/T	-1	0.00586114
hsa-mir-4271	chr3:49274155	hairpin	TG/T	-1	0.00360685
hsa-mir-6732	chr1:37480236	seed (2)	AG/A	-1	0.00360685
hsa-mir-7106	chr12:113159114	mature (22)	TG/T	-1	0.00270514
hsa-mir-3622a	chr8:27701724	hairpin	GC/G	-1	0.00270514
hsa-mir-3622b	chr8:27701724	hairpin	GC/G	-1	0.00270514
hsa-mir-3663	chr10:117167728	hairpin	AT/A	-1	0.00270514
hsa-mir-921	chr1:166154748	hairpin	GA/G	-1	0.00225428
hsa-mir-3680-1	chr16:21506081	seed (6)	GC/G	-1	0.00180343
hsa-mir-1303	chr5:154685809	hairpin	AT/A	-1	0.00135257
hsa-mir-4635	chr5:1062912	mature (13)	TC/T	-1	0.00135257
hsa-mir-6082	chr4:171186220	hairpin	TC/T	-1	0.00135257
hsa-mir-887	chr5:15935225	hairpin	TG/T	-1	0.000901713
hsa-mir-6508	chr21:39447049	seed (4)	GC/G	-1	0.000901713
hsa-mir-6085	chr15:62343123	mature (13)	GA/G	-1	0.000901713
hsa-mir-3125	chr2:12737375	hairpin	G/GA	1	0.702885
hsa-mir-5094	chr15:89850717	hairpin	C/CT	1	0.642471

hsa-mir-4797	chr3:197293909	hairpin	G/GA	1	0.62624
hsa-mir-4511	chr15:65719324	hairpin	C/CT	1	0.328224
hsa-mir-8086	chr10:28289300	hairpin	C/CA	1	0.288999
hsa-mir-4737	chr17:60043057	hairpin	G/GC	1	0.116321
hsa-mir-3199-1	chr22:27920603	seed (1)	T/TG	1	0.0946799
hsa-mir-3199-2	chr22:27920603	hairpin	T/TG	1	0.0946799
hsa-mir-1268b	chr17:80098843	mature (13)	T/TG	1	0.0734896
hsa-mir-3940	chr19:6416432	hairpin	G/GT	1	0.018936
hsa-mir-1289-1	chr20:35454060	hairpin	T/TA	1	0.00991885
hsa-mir-570	chr3:195699497	hairpin	T/TC	1	0.00631199
hsa-mir-7156	chr1:77060173	hairpin	T/TG	1	0.00586114
hsa-mir-4466	chr6:156779731	hairpin	T/TC	1	0.00450857
hsa-mir-548t	chr4:173268204	hairpin	C/CA	1	0.00405771
hsa-mir-216b	chr2:56000752	hairpin	T/TC	1	0.003156
hsa-mir-1303	chr5:154685809	hairpin	AT/ATT	1	0.00270514
hsa-mir-525	chr19:53697575	hairpin	G/GA	1	0.00270514
hsa-mir-190b	chr1:154193738	hairpin	G/GA	1	0.00270514
hsa-mir-4440	chr2:239068880	hairpin	A/AG	1	0.00270514
hsa-mir-1225	chr16:2090261	hairpin	T/TC	1	0.00225428
hsa-mir-5008	chr1:227941660	seed (3)	T/TC	1	0.00180343
hsa-mir-3921	chr3:99964342	seed (6)	C/CA	1	0.00180343
hsa-mir-760	chr1:93846840	hairpin	G/GC	1	0.00135257
hsa-mir-320b-2	chr1:224257035	hairpin	C/CT	1	0.00135257
hsa-mir-5582	chr11:46753137	mature (10)	T/TA	1	0.000901713
hsa-mir-132	chr17:2050000	hairpin	C/CG	1	0.000901713
hsa-mir-3658	chr1:165907957	hairpin	A/AT	1	0.000901713
hsa-mir-3938	chr3:55852534	seed (3)	A/AT	1	0.000901713
hsa-mir-4463	chr6:75428468	hairpin	A/AT	1	0.000901713
hsa-mir-4463	chr6:75428430	hairpin	C/CAG	2	0.653742
hsa-mir-943	chr4:1986461	hairpin	C/CCT	2	0.266005
hsa-mir-6087	chrX:108297779	seed (6)	C/CGG	2	0.0510386
hsa-mir-520h	chr19:53742585	mature (20)	A/AGT	2	0.00766456
hsa-mir-4526	chr18:13611157	hairpin	G/GAC	2	0.00135257
hsa-mir-7854	chr16:81533910	hairpin	A/ATC	2	0.00135257
hsa-mir-630	chr15:72587312	hairpin	A/ATTG	3	0.944545
hsa-mir-3131	chr2:219058688	hairpin	G/GAGA	3	0.495942
hsa-mir-3665	chr13:77698098	hairpin	T/TGCC	3	0.200631
hsa-mir-548a-1	chr6:18571856	mature (13)	T/TTAC	3	0.000901713
hsa-mir-4795	chr3:87226232	hairpin	T/TAAC	3	0.000901713
hsa-mir-567	chr3:112112876	hairpin	T/TAAAA	4	0.0374211
hsa-mir-429	chr1:1169026	hairpin	C/CCAGA	4	0.00541028
hsa-mir-6087	chrX:108297779	seed (6)	C/CGGGG	4	0.00237389
hsa-mir-4633	chr5:129097731	hairpin	G/GCATT	4	0.00180343
hsa-mir-4274	chr4:7460097	seed (7)	T/TCCCCA	5	0.864743
hsa-mir-6087	chrX:108297780	seed (7)	G/GGGGGC	5	0.292582
hsa-mir-516b-2	chr19:53725488	hairpin	G/GAAAGA	5	0.12624
hsa-mir-3620	chr1:228097285	seed (1)	G/GTGGGC	5	0.00225734

hsa-mir-6756	chr11:119312951	mature (19)	T/TGGGCA	5	0.000901713
hsa-mir-4284	chr7:73711370	hairpin	G/GGGTAGTTA	8	0.003156
hsa-mir-6727	chr1:1312563	hairpin	A/ACCCTGCCCTG	10	0.00180343
hsa-mir-1227	chr19:2234135	mature (15)	A/ACCGCCTGGCC	10	0.000901713
hsa-mir-6891	chr6:31355262	hairpin	T/TGAAGGGCTCCA	11	0.564472
hsa-mir-7150	chr9:123485617	hairpin	A/ACCGTGTGTGTGTGTGCGC	18	0.185302

Table 8. Seed region SNVs significantly deviating from HWE, excluding singletons.

miRNA	Chr:position	position in seed	SNV	MAF	pHWE
hsa-miR-1269b	chr17:12917329	6	G/C	0.287196	7.73E-07
hsa-miR-3117-3p	chr1:66628488	4	G/A	0.275023	1.05E-02
hsa-miR-3124-5p	chr1:248826385	3	C/T	0.00811542	1.82E-03
hsa-miR-3910	chr9:91636294	3	T/G	0.00180343	2.71E-03
hsa-miR-3928-5p	chr22:31160117	1	A/G	0.669071	8.20E-03
hsa-miR-4467	chr7:102471478	7	C/T	0.0135257	1.52E-02
hsa-miR-4472	chr12:116428309	3	A/C	0.418846	3.96E-07
hsa-miR-4472	chr8:142176399	4	G/C	0.226781	1.21E-09
hsa-miR-4482-5p	chr10:104268396	3	G/A	0.196123	2.19E-02
hsa-miR-4513	chr15:74788737	8	G/A	0.324617	8.83E-14
hsa-miR-4707-3p	chr14:22956973	6	C/A	0.534265	2.28E-22
hsa-miR-4741	chr18:22933411	5	C/T	0.11046	5.71E-03
hsa-miR-4781-3p	chr1:54054127	4	G/A	0.0834085	8.28E-05
hsa-miR-5090	chr7:102465754	3	G/A	0.119026	4.38E-02
hsa-miR-548ad-3p	chr2:35471453	1	G/A	0.0464382	2.69E-02
hsa-miR-548ao-3p	chr8:41271080	7	G/A	0.0139766	1.73E-02
hsa-miR-557	chr1:168375591	8	C/T	0.0329125	5.24E-03
hsa-miR-5589-3p	chr19:10038396	6	A/G	0.0302074	2.40E-03
hsa-miR-585-3p	chr5:169263631	4	C/T	0.0928765	1.95E-03
hsa-miR-662	chr16:770249	7	G/A	0.0333634	9.23E-04
hsa-miR-6777-5p	chr17:17813539	2	G/A	0.00901713	2.85E-03
hsa-miR-6796-3p	chr19:40369893	7	C/G	0.444995	4.15E-04
hsa-miR-6810-5p	chr2:218341923	8	C/T	0.0171326	3.81E-02
hsa-miR-6811-3p	chr2:237510968	1	A/G	0.334536	3.68E-03
hsa-miR-6826-5p	chr3:129272155	5	T/C	0.21055	9.09E-17
hsa-miR-6850-3p	chr8:144791952	3	G/A	0.0279531	1.18E-04
hsa-miR-6886-5p	chr19:11113481	3	C/T	0.0320108	2.28E-02
hsa-miR-938	chr10:29602331	2	C/T	0.146979	3.27E-06

Table 9. Differences in mRNA targets between the reference and alternative allele for SNVs deviating from HWE.

miRNA	**# of targets (ref)**	**# of targets (alt)**	**Biological processes similarity score**	**Molecular function similarity score**	**Cellular component similarity score**
hsa-miR-1269b	4093	7051	0.476	0.553	0.275
Ref targets functional enrichment			**Alt targets functional enrichment**		
GO:0022008 neurogenesis (p=1.19E-06) GO:0051179 localization (p=2.57E-06) GO:0098589 membrane region (p=4.91E-06) GO:0048518 positive regulation of biological process (p=1.95E-05) GO:0098590 plasma membrane region (p=2.63E-05) GO:0045202 synapse (p=2.73E-05) GO:0023051 regulation of signaling (p=9.39E-05) GO:0010646 regulation of cell communication (p=1.46E-04) GO:2000145 regulation of cell motility (p=2.64E-04) GO:0031175 neuron projection development (p=3.56E-04) GO:0071310 cellular response to organic substance (p=1.40E-03) GO:0009719 response to endogenous stimulus (p=1.58E-03) GO:0030054 cell junction (p=2.02E-03) GO:0007155 cell adhesion (p=2.38E-03) GO:0007610 behavior (p=5.62E-03) GO:0005488 binding (p=8.90E-03) GO:0007166 cell surface receptor signaling pathway (p=1.24E-02) GO:0035556 intracellular signal transduction (p=1.67E-02) GO:0030001 metal ion transport (p=3.46E-02) GO:0022836 gated channel activity (p=3.54E-02) GO:0007156 homophilic cell adhesion via plasma membrane adhesion molecules (p=3.81E-02) GO:0034702 ion channel complex (p=4.72E-02)			GO:0005622 intracellular (p=1.16E-07) GO:0043167 ion binding (p=1.53E-06) GO:0051179 localization (p=4.52E-06) GO:0016482 cytosolic transport (p=8.94E-04) GO:0006643 membrane lipid metabolic process (p=3.30E-03) GO:0003779 actin binding (p=3.52E-03) GO:0007156 homophilic cell adhesion via plasma membrane adhesion molecules (p=3.92E-03) GO:0015629 actin cytoskeleton (p=3.97E-02)		
hsa-miR-3117-3p	3072	4968	0.529	0.435	0.448
Ref targets functional enrichment			**Alt targets functional enrichment**		
GO:0044464 cell part (p=2.16E-05) GO:0043226 organelle (p=3.31E-04) GO:0043167 ion binding (p=8.00E-04) GO:0051179 localization (p=1.12E-03) GO:0065007 biological regulation (p=1.74E-03) GO:0009719 response to endogenous stimulus (p=1.26E-02) GO:0048015 phosphatidylinositol-mediated signaling (p=2.96E-02) GO:1901698 response to nitrogen compound (p=3.15E-02)			GO:0044424 intracellular part (p=2.29E-17) GO:0043231 intracellular membrane-bounded organelle (p=8.80E-11) GO:0051179 localization (p=1.90E-10) GO:0016043 cellular component organization (p=3.41E-09) GO:0005488 binding (p=4.49E-09) GO:0048518 positive regulation of biological process (p=1.37E-07) GO:0019222 regulation of metabolic process (p=2.40E-07) GO:0005654 nucleoplasm (p=1.25E-06) GO:0048856 anatomical structure development (p=1.85E-06) GO:0044707 single-multicellular organism process (p=2.42E-06) GO:0098805 whole membrane (p=3.27E-05) GO:0045202 synapse (p=2.20E-03) GO:0016477 cell migration (p=3.33E-03) GO:0000166 nucleotide binding (p=4.27E-03) GO:0035556 intracellular signal transduction (p=8.19E-03) GO:0007155 cell adhesion (p=9.85E-03) GO:0006397 mRNA processing (p=2.52E-02) GO:0030554 adenyl nucleotide binding (p=2.87E-02) GO:0016773 phosphotransferase activity, alcohol group as acceptor (p=3.19E-02) GO:0016740 transferase activity (p=3.26E-02)		

hsa-miR-3124-5p	198	707	0.5	NA	NA
Ref targets functional enrichment			**Alt targets functional enrichment**		
GO:0030154 cell differentiation (p=2.75E-06) GO:0007275 multicellular organism development (p=5.40E-06) GO:0050789 regulation of biological process (p=2.01E-02) GO:0007409 axonogenesis (p=2.13E-02)			GO:0048731 system development (p=4.31E-07) GO:0007156 homophilic cell adhesion via plasma membrane adhesion molecules (p=3.53E-05) GO:0005886 plasma membrane (p=5.88E-03)		
hsa-miR-3910	7438	876	NA	NA	0.524
Ref targets functional enrichment			**Alt targets functional enrichment**		
GO:0044424 intracellular part (p=1.78E-23) GO:0043167 ion binding (p=6.81E-17) GO:0048856 anatomical structure development (p=9.18E-13) GO:0035556 intracellular signal transduction (p=6.31E-07) GO:0071310 cellular response to organic substance (p=6.71E-05) GO:0016740 transferase activity (p=7.53E-05) GO:0009719 response to endogenous stimulus (p=1.73E-04) GO:0007010 cytoskeleton organization (p=2.48E-04) GO:1901701 cellular response to oxygen-containing compound (p=2.92E-04) GO:0006357 regulation of transcription from RNA polymerase II promoter (p=2.92E-04) GO:0098805 whole membrane (p=3.18E-04) GO:0061028 establishment of endothelial barrier (p=3.28E-04) GO:1901698 response to nitrogen compound (p=1.08E-03) GO:0015629 actin cytoskeleton (p=1.17E-03) GO:0033554 cellular response to stress (p=3.83E-03) GO:0072511 divalent inorganic cation transport (p=6.64E-03) GO:0016023 cytoplasmic, membrane-bounded vesicle (p=1.90E-02) GO:0023061 signal release (p=2.05E-02) GO:0043087 regulation of GTPase activity (p=2.79E-02) GO:0003012 muscle system process (p=2.88E-02) GO:2001257 regulation of cation channel activity (p=4.51E-02)			GO:0043231 intracellular membrane-bounded organelle (p=4.35E-03)		
hsa-miR-4467	1626	1470	0.503	0.526	0.528
Ref targets functional enrichment			**Alt targets functional enrichment**		
GO:0050794 regulation of cellular process (p=5.71E-07) GO:0000981 RNA polymerase II transcription factor activity, sequence-specific DNA binding (p=1.21E-06) GO:0043565 sequence-specific DNA binding (p=1.70E-06) GO:0006357 regulation of transcription from RNA polymerase II promoter (p=7.52E-06) GO:0007399 nervous system development (p=3.99E-05) GO:0044212 transcription regulatory region DNA binding (p=2.96E-03) GO:0005634 nucleus (p=2.10E-02) GO:0031974 membrane-enclosed lumen (p=2.45E-02) GO:0001105 RNA polymerase II transcription coactivator activity (p=2.73E-02) GO:0000989 transcription factor activity, transcription factor binding (p=4.30E-02)			GO:0007399 nervous system development (p=1.51E-08) GO:0005488 binding (p=1.06E-06) GO:0031175 neuron projection development (p=3.11E-05) GO:0043005 neuron projection (p=1.97E-04) GO:0048518 positive regulation of biological process (p=2.10E-04) GO:0048583 regulation of response to stimulus (p=1.29E-03) GO:0051179 localization (p=1.39E-03) GO:0005768 endosome (p=4.44E-03) GO:0040011 locomotion (p=4.60E-03) GO:0023051 regulation of signaling (p=5.71E-03) GO:0070848 response to growth factor (p=8.01E-03) GO:0007167 enzyme linked receptor protein signaling pathway (p=9.18E-03) GO:0030054 cell junction (p=1.14E-02) GO:0044424 intracellular part (p=1.41E-02) GO:0000981 RNA polymerase II transcription factor activity, sequence-specific DNA binding (p=2.62E-02)		

hsa-miR-4472 (A3C)	9663	6054	0.467	0.653	0.682
Ref targets functional enrichment			**Alt targets functional enrichment**		
GO:0044707 single-multicellular organism process (p=3.96E-27) GO:0097458 neuron part (p=6.48E-19) GO:0005488 binding (p=3.19E-18) GO:1903508 positive regulation of nucleic acid-templated transcription (p=4.80E-09) GO:0043565 sequence-specific DNA binding (p=1.23E-06) GO:0030036 actin cytoskeleton organization (p=5.79E-06) GO:0015629 actin cytoskeleton (p=1.18E-05) GO:0071310 cellular response to organic substance (p=1.69E-05) GO:0043547 positive regulation of GTPase activity (p=8.76E-05) GO:0000975 regulatory region DNA binding (p=1.23E-04) GO:1990837 sequence-specific double-stranded DNA binding (p=4.30E-04) GO:0046873 metal ion transmembrane transporter activity (p=1.64E-03) GO:0006468 protein phosphorylation (p=3.20E-03) GO:0030672 synaptic vesicle membrane (p=4.90E-03) GO:0003002 regionalization (p=7.05E-03) GO:0071872 cellular response to epinephrine stimulus (p=1.02E-02) GO:0030325 adrenal gland development (p=1.34E-02) GO:1901700 response to oxygen-containing compound (p=1.50E-02) GO:0016482 cytosolic transport (p=1.96E-02) GO:0003013 circulatory system process (p=2.17E-02) GO:0016773 phosphotransferase activity, alcohol group as acceptor (p=2.32E-02) GO:1990351 transporter complex (p=3.35E-02) GO:0016301 kinase activity (p=3.46E-02) GO:0050890 cognition (p=4.58E-02)			GO:0007399 nervous system development (p=4.98E-28) GO:0043005 neuron projection (p=2.19E-16) GO:0019899 enzyme binding (p=3.41E-11) GO:0006357 regulation of transcription from RNA polymerase II promoter (p=2.17E-09) GO:0098805 whole membrane (p=7.74E-07) GO:0001071 nucleic acid binding transcription factor activity (p=1.03E-06) GO:0000975 regulatory region DNA binding (p=2.71E-06) GO:0000976 transcription regulatory region sequence-specific DNA binding (p=1.13E-05) GO:0030036 actin cytoskeleton organization (p=3.05E-05) GO:0098588 bounding membrane of organelle (p=5.01E-05) GO:0000988 transcription factor activity, protein binding (p=1.13E-04) GO:0015629 actin cytoskeleton (p=1.15E-04) GO:0048489 synaptic vesicle transport (p=1.75E-04) GO:0006468 protein phosphorylation (p=4.10E-04) GO:0034703 cation channel complex (p=6.26E-03) GO:0008021 synaptic vesicle (p=1.02E-02) GO:0001654 eye development (p=1.12E-02) GO:0010863 positive regulation of phospholipase C activity (p=1.35E-02) GO:0016773 phosphotransferase activity, alcohol group as acceptor (p=1.44E-02) GO:0051146 striated muscle cell differentiation (p=1.70E-02) GO:0030001 metal ion transport (p=1.71E-02) GO:1990351 transporter complex (p=1.80E-02)		
hsa-miR-4472 (G4C)	9663	1354	0.287	0.524	0.461
Ref targets functional enrichment			**Alt targets functional enrichment**		
GO:0044707 single-multicellular organism process (p=3.96E-27) GO:0097458 neuron part (p=6.48E-19) GO:0005488 binding (p=3.19E-18) GO:1903508 positive regulation of nucleic acid-templated transcription (p=4.80E-09) GO:0043565 sequence-specific DNA binding (p=1.23E-06) GO:0030036 actin cytoskeleton organization (p=5.79E-06) GO:0015629 actin cytoskeleton (p=1.18E-05) GO:0071310 cellular response to organic substance (p=1.69E-05) GO:0043547 positive regulation of GTPase activity (p=8.76E-05) GO:0000975 regulatory region DNA binding (p=1.23E-04) GO:1990837 sequence-specific double-stranded DNA binding (p=4.30E-04) GO:0046873 metal ion transmembrane transporter activity (p=1.64E-03) GO:0006468 protein phosphorylation (p=3.20E-03) GO:0030672 synaptic vesicle membrane (p=4.90E-03) GO:0003002 regionalization (p=7.05E-03) GO:0071872 cellular response to epinephrine stimulus (p=1.02E-02) GO:0030325 adrenal gland development (p=1.34E-02) GO:1901700 response to oxygen-containing compound (p=1.50E-02) GO:0016482 cytosolic transport (p=1.96E-02) GO:0003013 circulatory system process (p=2.17E-02) GO:0016773 phosphotransferase activity, alcohol group as acceptor (p=2.32E-02) GO:1990351 transporter complex (p=3.35E-02) GO:0016301 kinase activity (p=3.46E-02) GO:0050890 cognition (p=4.58E-02)			GO:0007399 nervous system development (p=4.19E-13) GO:0048522 positive regulation of cellular process (p=5.89E-07) GO:0099537 trans-synaptic signaling (p=1.09E-06) GO:0099536 synaptic signaling (p=1.09E-06) GO:0044459 plasma membrane part (p=7.63E-06) GO:0044456 synapse part (p=1.61E-05) GO:0005509 calcium ion binding (p=2.39E-05) GO:0097458 neuron part (p=2.94E-05) GO:0051179 localization (p=6.37E-05) GO:0043005 neuron projection (p=7.19E-05) GO:0098742 cell-cell adhesion via plasma-membrane adhesion molecules (p=2.59E-04) GO:0031175 neuron projection development (p=4.08E-04) GO:0098805 whole membrane (p=6.14E-04) GO:0044212 transcription regulatory region DNA binding (p=1.28E-03) GO:0097485 neuron projection guidance (p=2.19E-03) GO:0007411 axon guidance (p=2.19E-03) GO:0051899 membrane depolarization (p=4.38E-03) GO:0005515 protein binding (p=5.46E-03) GO:0046903 secretion (p=9.88E-03) GO:0035556 intracellular signal transduction (p=1.15E-02) GO:0000976 transcription regulatory region sequence-specific DNA binding (p=1.67E-02) GO:0036465 synaptic vesicle recycling (p=3.48E-02) GO:0001505 regulation of neurotransmitter levels (p=4.07E-02)		

hsa-miR-4482-5p	4737	4433	0.501	0.309	0.412
Ref targets functional enrichment			**Alt targets functional enrichment**		
GO:0043167 ion binding (p=7.21E-12) GO:0007156 homophilic cell adhesion via plasma membrane adhesion molecules (p=1.08E-08) GO:0005623 cell (p=3.99E-07) GO:0051179 localization (p=1.08E-06) GO:0048731 system development (p=3.09E-05) GO:0006464 cellular protein modification process (p=9.11E-04) GO:0019222 regulation of metabolic process (p=1.35E-03) GO:0048667 cell morphogenesis involved in neuron differentiation (p=1.39E-02) GO:0044260 cellular macromolecule metabolic process (p=1.42E-02) GO:0015085 calcium ion transmembrane transporter activity (p=1.43E-02) GO:0031175 neuron projection development (p=1.55E-02) GO:0048699 generation of neurons (p=1.59E-02) GO:0016477 cell migration (p=3.42E-02) GO:0005261 cation channel activity (p=4.15E-02) GO:0016740 transferase activity (p=4.85E-02)			GO:0044424 intracellular part (p=1.80E-12) GO:0048731 system development (p=8.31E-10) GO:0005515 protein binding (p=9.96E-08) GO:0006464 cellular protein modification process (p=2.47E-05) GO:0007167 enzyme linked receptor protein signaling pathway (p=4.09E-05) GO:0010646 regulation of cell communication (p=6.18E-05) GO:0051056 regulation of small GTPase mediated signal transduction (p=1.01E-04) GO:0001076 transcription factor activity, RNA polymerase II transcription factor binding (p=8.20E-04) GO:0071495 cellular response to endogenous stimulus (p=9.98E-04) GO:0009059 macromolecule biosynthetic process (p=1.77E-03) GO:0071363 cellular response to growth factor stimulus (p=2.89E-03) GO:0043234 protein complex (p=1.19E-02) GO:0048699 generation of neurons (p=1.38E-02) GO:0098588 bounding membrane of organelle (p=1.52E-02) GO:0032869 cellular response to insulin stimulus (p=1.57E-02) GO:1901699 cellular response to nitrogen compound (p=2.42E-02) GO:1903508 positive regulation of nucleic acid-templated transcription (p=3.13E-02) GO:0099537 trans-synaptic signaling (p=4.73E-02)		
hsa-miR-4513	4389	4369	0.525	0.697	0.637
Ref targets functional enrichment			**Alt targets functional enrichment**		
GO:0005622 intracellular (p=1.34E-08) GO:0065007 biological regulation (p=4.56E-06) GO:2001029 regulation of cellular glucuronidation (p=7.31E-06) GO:0052697 xenobiotic glucuronidation (p=7.31E-06) GO:0046872 metal ion binding (p=1.42E-05) GO:0050794 regulation of cellular process (p=6.13E-05) GO:1904223 regulation of glucuronosyltransferase activity (p=6.62E-05) GO:0043229 intracellular organelle (p=2.55E-04) GO:0045202 synapse (p=4.21E-04) GO:0099537 trans-synaptic signaling (p=5.34E-04) GO:0098805 whole membrane (p=5.90E-04) GO:0098588 bounding membrane of organelle (p=1.68E-03) GO:0006464 cellular protein modification process (p=2.82E-03) GO:0044707 single-multicellular organism process (p=8.22E-03) GO:0045922 negative regulation of fatty acid metabolic process (p=1.32E-02) GO:1903506 regulation of nucleic acid-templated transcription (p=4.62E-02)			GO:0046872 metal ion binding (p=3.29E-08) GO:0065007 biological regulation (p=3.05E-06) GO:0006464 cellular protein modification process (p=3.59E-06) GO:0005622 intracellular (p=6.38E-06) GO:0045202 synapse (p=2.42E-04) GO:0007169 transmembrane receptor protein tyrosine kinase signaling pathway (p=7.61E-04) GO:0061564 axon development (p=1.09E-03) GO:0098589 membrane region (p=1.57E-03) GO:0048667 cell morphogenesis involved in neuron differentiation (p=1.62E-03) GO:0044707 single-multicellular organism process (p=2.13E-03) GO:0030054 cell junction (p=3.98E-03) GO:0043226 organelle (p=1.57E-02) GO:0008361 regulation of cell size (p=1.60E-02) GO:0016567 protein ubiquitination (p=2.84E-02) GO:0022008 neurogenesis (p=3.42E-02) GO:0004842 ubiquitin-protein transferase activity (p=4.63E-02)		
hsa-miR-4707-3p	2243	6253	0.272	NA	NA
Ref targets functional enrichment			**Alt targets functional enrichment**		
GO:0007156 homophilic cell adhesion via plasma membrane adhesion molecules (p=1.85E-07) GO:0050793 regulation of developmental process (p=5.38E-03)			GO:0048518 positive regulation of biological process (p=8.03E-21) GO:0030054 cell junction (p=9.92E-11) GO:0019899 enzyme binding (p=7.06E-09) GO:0036211 protein modification process (p=3.08E-06) GO:0006468 protein phosphorylation (p=4.95E-06) GO:0010033 response to organic substance (p=1.04E-04) GO:0006357 regulation of transcription from RNA polymerase II promoter (p=7.94E-04) GO:0008047 enzyme activator activity (p=1.99E-03) GO:0098772 molecular function regulator (p=2.47E-03) GO:0030036 actin cytoskeleton organization (p=6.17E-03) GO:1901700 response to oxygen-containing compound (p=6.44E-03) GO:1990837 sequence-specific double-stranded DNA binding (p=2.77E-02) GO:0060589 nucleoside-triphosphatase regulator activity (p=2.92E-02) GO:0050922 negative regulation of chemotaxis (p=2.93E-02)		

hsa-miR-4741	8202	3445	0.605	0.446	0.177
Ref targets functional enrichment			**Alt targets functional enrichment**		
GO:0048731 system development (p=4.11E-18) GO:0005515 protein binding (p=3.60E-13) GO:0098805 whole membrane (p=1.94E-09) GO:0006357 regulation of transcription from RNA polymerase II promoter (p=8.25E-05) GO:0010033 response to organic substance (p=3.75E-04) GO:0036211 protein modification process (p=5.98E-04) GO:0043565 sequence-specific DNA binding (p=3.22E-03) GO:0098588 bounding membrane of organelle (p=5.31E-03) GO:0046873 metal ion transmembrane transporter activity (p=5.40E-03) GO:0010959 regulation of metal ion transport (p=5.77E-03) GO:0000975 regulatory region DNA binding (p=1.09E-02) GO:0016301 kinase activity (p=1.18E-02) GO:0016773 phosphotransferase activity, alcohol group as acceptor (p=1.34E-02) GO:0001071 nucleic acid binding transcription factor activity (p=1.71E-02) GO:0070838 divalent metal ion transport (p=1.94E-02) GO:0005261 cation channel activity (p=2.22E-02) GO:0006796 phosphate-containing compound metabolic process (p=2.98E-02) GO:0012501 programmed cell death (p=3.38E-02) GO:0016477 cell migration (p=3.58E-02) GO:0022836 gated channel activity (p=3.59E-02) GO:0000988 transcription factor activity, protein binding (p=3.95E-02) GO:0008361 regulation of cell size (p=3.97E-02) GO:1901700 response to oxygen-containing compound (p=4.91E-02)			GO:0048731 system development (p=9.10E-11) GO:0051179 localization (p=4.09E-07) GO:0032879 regulation of localization (p=5.94E-07) GO:0022008 neurogenesis (p=6.69E-07) GO:0036477 somatodendritic compartment (p=1.22E-06) GO:0044459 plasma membrane part (p=2.11E-06) GO:0045202 synapse (p=1.16E-05) GO:0016477 cell migration (p=4.75E-05) GO:0000904 cell morphogenesis involved in differentiation (p=5.22E-05) GO:0043025 neuronal cell body (p=2.46E-04) GO:0007268 synaptic transmission (p=3.81E-04) GO:0000982 transcription factor activity, RNA polymerase II core promoter proximal region sequence-specific binding (p=6.65E-04) GO:0006464 cellular protein modification process (p=1.36E-03) GO:0005515 protein binding (p=1.57E-03) GO:0007610 behavior (p=1.70E-03) GO:0043269 regulation of ion transport (p=3.76E-03) GO:0000978 RNA polymerase II core promoter proximal region sequence-specific DNA binding (p=4.50E-02) GO:0098589 membrane region (p=4.60E-02)		
hsa-miR-4781-3p	5654	4596	0.314	0.78	0.499
Ref targets functional enrichment			**Alt targets functional enrichment**		
GO:0005622 intracellular (p=4.80E-10) GO:0043167 ion binding (p=1.74E-05) GO:0051179 localization (p=3.13E-05) GO:0098805 whole membrane (p=9.84E-04) GO:0043412 macromolecule modification (p=1.72E-03) GO:0050801 ion homeostasis (p=2.86E-02) GO:0019222 regulation of metabolic process (p=4.11E-02)			GO:0043167 ion binding (p=4.88E-07) GO:0044424 intracellular part (p=1.03E-06) GO:0031323 regulation of cellular metabolic process (p=8.45E-06) GO:0044707 single-multicellular organism process (p=1.95E-04) GO:0001071 nucleic acid binding transcription factor activity (p=4.37E-04) GO:0007275 multicellular organism development (p=1.18E-03) GO:0043226 organelle (p=1.23E-03) GO:1903506 regulation of nucleic acid-templated transcription (p=3.32E-02)		
hsa-miR-5090	1959	8761	0.383	NA	0.344
Ref targets functional enrichment			**Alt targets functional enrichment**		
GO:0010646 regulation of cell communication (p=6.17E-07) GO:0023051 regulation of signaling (p=3.18E-06) GO:0007399 nervous system development (p=2.67E-04) GO:0051179 localization (p=3.42E-04) GO:0043005 neuron projection (p=2.75E-03) GO:0048666 neuron development (p=1.74E-02) GO:0044459 plasma membrane part (p=2.61E-02)			GO:0005515 protein binding (p=1.21E-13) GO:0044424 intracellular part (p=2.14E-13) GO:0048522 positive regulation of cellular process (p=2.68E-13) GO:0035556 intracellular signal transduction (p=4.63E-06) GO:0007169 transmembrane receptor protein tyrosine kinase signaling pathway (p=4.63E-04) GO:0000981 RNA polymerase II transcription factor activity, sequence-specific DNA binding (p=5.07E-04) GO:0015629 actin cytoskeleton (p=1.25E-03) GO:0006357 regulation of transcription from RNA polymerase II promoter (p=1.85E-03) GO:0022836 gated channel activity (p=1.87E-03) GO:0030659 cytoplasmic vesicle membrane (p=3.45E-03) GO:0006464 cellular protein modification process (p=3.96E-03) GO:0070382 exocytic vesicle (p=4.34E-03) GO:0016773 phosphotransferase activity, alcohol group as acceptor (p=2.27E-02) GO:0008361 regulation of cell size (p=3.60E-02) GO:0006835 dicarboxylic acid transport (p=3.61E-02)		

hsa-miR-548ao-3p	2037	6199	0.456	0.497	0.513
Ref targets functional enrichment			**Alt targets functional enrichment**		
GO:0044424 intracellular part (p=4.23E-07) GO:0043231 intracellular membrane-bounded organelle (p=5.38E-06) GO:0008150 biological_process (p=4.20E-05) GO:0060255 regulation of macromolecule metabolic process (p=7.09E-05) GO:0005488 binding (p=2.41E-03) GO:0046328 regulation of JNK cascade (p=2.76E-02) GO:0001071 nucleic acid binding transcription factor activity (p=3.84E-02)			GO:0005622 intracellular (p=1.77E-14) GO:0043167 ion binding (p=4.04E-14) GO:0051179 localization (p=1.39E-09) GO:0043412 macromolecule modification (p=1.18E-07) GO:0006464 cellular protein modification process (p=1.20E-07) GO:0048518 positive regulation of biological process (p=1.80E-06) GO:0016043 cellular component organization (p=4.60E-06) GO:0032502 developmental process (p=5.74E-06) GO:0044707 single-multicellular organism process (p=2.88E-05) GO:0098805 whole membrane (p=1.02E-04) GO:0004842 ubiquitin-protein transferase activity (p=1.82E-04) GO:0048519 negative regulation of biological process (p=3.96E-04) GO:0007156 homophilic cell adhesion via plasma membrane adhesion molecules (p=1.27E-03) GO:0010033 response to organic substance (p=7.24E-03) GO:0098588 bounding membrane of organelle (p=1.01E-02) GO:0010646 regulation of cell communication (p=1.12E-02) GO:0000166 nucleotide binding (p=1.28E-02) GO:0065007 biological regulation (p=1.33E-02) GO:0023051 regulation of signaling (p=1.41E-02) GO:2000112 regulation of cellular macromolecule biosynthetic process (p=2.52E-02) GO:0016477 cell migration (p=2.56E-02) GO:0051172 negative regulation of nitrogen compound metabolic process (p=3.39E-02) GO:0009890 negative regulation of biosynthetic process (p=4.20E-02)		
hsa-miR-557	7895	9552	0.726	0.809	0.813
Ref targets functional enrichment			**Alt targets functional enrichment**		
GO:0019222 regulation of metabolic process (p=3.10E-15) GO:0044424 intracellular part (p=6.91E-14) GO:0005488 binding (p=2.37E-13) GO:0001071 nucleic acid binding transcription factor activity (p=9.28E-09) GO:0035556 intracellular signal transduction (p=3.71E-08) GO:0007167 enzyme linked receptor protein signaling pathway (p=1.22E-06) GO:0070848 response to growth factor (p=2.65E-06) GO:0071310 cellular response to organic substance (p=8.79E-06) GO:0071495 cellular response to endogenous stimulus (p=2.81E-05) GO:0042325 regulation of phosphorylation (p=5.29E-04) GO:0090257 regulation of muscle system process (p=1.59E-03) GO:0007411 axon guidance (p=1.82E-03) GO:0016477 cell migration (p=5.31E-03) GO:0016567 protein ubiquitination (p=8.51E-03) GO:0009952 anterior/posterior pattern specification (p=8.69E-03) GO:0098805 whole membrane (p=1.03E-02) GO:1990234 transferase complex (p=1.51E-02) GO:0038095 Fc-epsilon receptor signaling pathway (p=1.88E-02) GO:0007267 cell-cell signaling (p=2.58E-02) GO:0010863 positive regulation of phospholipase C activity (p=2.66E-02) GO:0003727 single-stranded RNA binding (p=2.93E-02) GO:0030097 hemopoiesis (p=4.17E-02) GO:0002065 columnar/cuboidal epithelial cell differentiation (p=4.31E-02)			GO:0007275 multicellular organism development (p=3.71E-18) GO:0005622 intracellular (p=1.00E-17) GO:0005488 binding (p=8.65E-14) GO:0035556 intracellular signal transduction (p=4.69E-10) GO:0070848 response to growth factor (p=4.36E-08) GO:0007167 enzyme linked receptor protein signaling pathway (p=6.77E-08) GO:0000981 RNA polymerase II transcription factor activity, sequence-specific DNA binding (p=3.02E-07) GO:0071363 cellular response to growth factor stimulus (p=3.27E-07) GO:0009719 response to endogenous stimulus (p=5.14E-07) GO:0006928 movement of cell or subcellular component (p=2.49E-05) GO:0043565 sequence-specific DNA binding (p=1.16E-04) GO:0001067 regulatory region nucleic acid binding (p=4.23E-04) GO:0038095 Fc-epsilon receptor signaling pathway (p=6.54E-04) GO:0099537 trans-synaptic signaling (p=1.21E-03) GO:1990234 transferase complex (p=1.91E-03) GO:0098805 whole membrane (p=2.19E-03) GO:0071375 cellular response to peptide hormone stimulus (p=3.35E-03) GO:1902911 protein kinase complex (p=1.07E-02) GO:0010863 positive regulation of phospholipase C activity (p=1.09E-02) GO:0090257 regulation of muscle system process (p=1.46E-02) GO:0003013 circulatory system process (p=2.58E-02) GO:0030029 actin filament-based process (p=2.89E-02) GO:0003727 single-stranded RNA binding (p=3.06E-02)		

hsa-miR-5589-3p	6301	1894	0.606	0.315	0.421
Ref targets functional enrichment			**Alt targets functional enrichment**		
GO:0065007 biological regulation (p=2.94E-15) GO:0044464 cell part (p=3.08E-12) GO:0005515 protein binding (p=3.32E-08) GO:0071310 cellular response to organic substance (p=1.78E-07) GO:0099537 trans-synaptic signaling (p=1.61E-05) GO:1901699 cellular response to nitrogen compound (p=9.54E-05) GO:0016773 phosphotransferase activity, alcohol group as acceptor (p=1.05E-04) GO:0006366 transcription from RNA polymerase II promoter (p=1.08E-04) GO:0052697 xenobiotic glucuronidation (p=1.22E-04) GO:1904224 negative regulation of glucuronosyltransferase activity (p=8.12E-04) GO:0098772 molecular function regulator (p=1.26E-02) GO:0000981 RNA polymerase II transcription factor activity, sequence-specific DNA binding (p=1.98E-02) GO:0051552 flavone metabolic process (p=3.60E-02)			GO:0052697 xenobiotic glucuronidation (p=6.40E-09) GO:2001030 negative regulation of cellular glucuronidation (p=1.27E-07) GO:1904224 negative regulation of glucuronosyltransferase activity (p=1.27E-07) GO:0065007 biological regulation (p=2.45E-05) GO:0051552 flavone metabolic process (p=4.98E-05) GO:0032940 secretion by cell (p=5.16E-05) GO:0043167 ion binding (p=7.43E-05) GO:0005996 monosaccharide metabolic process (p=1.88E-04) GO:0023061 signal release (p=6.79E-04) GO:0005737 cytoplasm (p=2.55E-03) GO:0019217 regulation of fatty acid metabolic process (p=5.18E-03) GO:0071495 cellular response to endogenous stimulus (p=2.06E-02) GO:0098793 presynapse (p=2.10E-02) GO:0015629 actin cytoskeleton (p=2.62E-02)		
hsa-miR-585-3p	1157	740	0.322	NA	NA
Ref targets functional enrichment			**Alt targets functional enrichment**		
GO:0051179 localization (p=2.92E-02)			GO:0046777 protein autophosphorylation (p=2.23E-03) GO:0006811 ion transport (p=3.34E-03) GO:0005886 plasma membrane (p=3.72E-03) GO:0044463 cell projection part (p=4.00E-02) GO:0055085 transmembrane transport (p=4.62E-02)		
hsa-miR-662	1844	6830	0.429	0.381	0.457
Ref targets functional enrichment			**Alt targets functional enrichment**		
GO:0048518 positive regulation of biological process (p=6.12E-08) GO:0007399 nervous system development (p=3.24E-04) GO:0051641 cellular localization (p=1.07E-03) GO:0031252 cell leading edge (p=3.84E-03) GO:0005515 protein binding (p=4.39E-03) GO:0010033 response to organic substance (p=7.67E-03) GO:0030036 actin cytoskeleton organization (p=1.27E-02) GO:0032880 regulation of protein localization (p=1.30E-02) GO:0030054 cell junction (p=3.31E-02) GO:0098589 membrane region (p=3.34E-02)			GO:0005622 intracellular (p=6.98E-17) GO:0009987 cellular process (p=8.89E-13) GO:0043167 ion binding (p=2.38E-10) GO:0007399 nervous system development (p=1.77E-05) GO:0007167 enzyme linked receptor protein signaling pathway (p=7.94E-04) GO:0007265 Ras protein signal transduction (p=9.44E-04) GO:0016740 transferase activity (p=3.30E-03) GO:0001071 nucleic acid binding transcription factor activity (p=1.15E-02) GO:0030001 metal ion transport (p=2.64E-02) GO:0046873 metal ion transmembrane transporter activity (p=3.24E-02) GO:0048468 cell development (p=4.22E-02)		

hsa-miR-6777-5p	2493	10355	0.565	0.582	0.549
Ref targets functional enrichment			**Alt targets functional enrichment**		
GO:0007399 nervous system development (p=3.98E-15) GO:0048518 positive regulation of biological process (p=2.28E-14) GO:0051179 localization (p=1.86E-10) GO:0023051 regulation of signaling (p=4.87E-10) GO:0098805 whole membrane (p=5.29E-09) GO:0043005 neuron projection (p=2.00E-06) GO:0009966 regulation of signal transduction (p=4.92E-06) GO:0045202 synapse (p=4.12E-05) GO:0007610 behavior (p=5.66E-05) GO:0010033 response to organic substance (p=1.45E-04) GO:0005515 protein binding (p=2.20E-04) GO:0012505 endomembrane system (p=6.76E-04) GO:0050905 neuromuscular process (p=1.95E-03) GO:1901700 response to oxygen-containing compound (p=3.69E-03) GO:0043565 sequence-specific DNA binding (p=3.87E-03) GO:0016477 cell migration (p=4.02E-03) GO:0000981 RNA polymerase II transcription factor activity, sequence-specific DNA binding (p=5.10E-03) GO:0071944 cell periphery (p=8.57E-03) GO:0044424 intracellular part (p=1.27E-02) GO:0016023 cytoplasmic, membrane-bounded vesicle (p=2.30E-02) GO:0008328 ionotropic glutamate receptor complex (p=2.37E-02) GO:0002376 immune system process (p=3.16E-02) GO:0008015 blood circulation (p=4.43E-02) GO:0007264 small GTPase mediated signal transduction (p=4.70E-02)			GO:0048731 system development (p=1.26E-24) GO:0097458 neuron part (p=2.32E-16) GO:0005488 binding (p=3.99E-13) GO:0035556 intracellular signal transduction (p=7.66E-10) GO:0036211 protein modification process (p=3.44E-07) GO:0006357 regulation of transcription from RNA polymerase II promoter (p=1.96E-06) GO:0007166 cell surface receptor signaling pathway (p=2.76E-06) GO:0010033 response to organic substance (p=8.55E-06) GO:0016301 kinase activity (p=2.41E-05) GO:0071310 cellular response to organic substance (p=3.49E-05) GO:0043565 sequence-specific DNA binding (p=4.92E-05) GO:0030659 cytoplasmic vesicle membrane (p=5.94E-05) GO:0001934 positive regulation of protein phosphorylation (p=1.24E-04) GO:0022838 substrate-specific channel activity (p=1.71E-04) GO:1901700 response to oxygen-containing compound (p=2.49E-04) GO:0000975 regulatory region DNA binding (p=3.30E-04) GO:0046873 metal ion transmembrane transporter activity (p=7.42E-04) GO:0097479 synaptic vesicle localization (p=7.58E-04) GO:0000981 RNA polymerase II transcription factor activity, sequence-specific DNA binding (p=1.26E-03) GO:0005057 receptor signaling protein activity (p=1.26E-03) GO:0015085 calcium ion transmembrane transporter activity (p=5.45E-03) GO:0030036 actin cytoskeleton organization (p=1.16E-02) GO:0030073 insulin secretion (p=1.22E-02) GO:0061387 regulation of extent of cell growth (p=2.39E-02) GO:0015629 actin cytoskeleton (p=2.65E-02) GO:0042060 wound healing (p=4.72E-02)		
hsa-miR-6796-3p	4890	5796	0.362	0.78	0.531
Ref targets functional enrichment			**Alt targets functional enrichment**		
GO:0051179 localization (p=3.14E-08) GO:0005488 binding (p=7.06E-08) GO:0044424 intracellular part (p=8.38E-08) GO:0043229 intracellular organelle (p=5.41E-06) GO:0045202 synapse (p=3.46E-05) GO:0036211 protein modification process (p=4.36E-05) GO:0035556 intracellular signal transduction (p=7.85E-05) GO:0099536 synaptic signaling (p=1.32E-03) GO:0032281 AMPA glutamate receptor complex (p=3.93E-03) GO:0006928 movement of cell or subcellular component (p=4.94E-03) GO:1902837 amino acid import into cell (p=1.03E-02) GO:0010468 regulation of gene expression (p=1.86E-02) GO:2000112 regulation of cellular macromolecule biosynthetic process (p=3.07E-02) GO:0030054 cell junction (p=3.34E-02) GO:0000902 cell morphogenesis (p=3.51E-02) GO:0051938 L-glutamate import (p=3.62E-02) GO:0010033 response to organic substance (p=3.67E-02) GO:0006357 regulation of transcription from RNA polymerase II promoter (p=4.54E-02)			GO:0044424 intracellular part (p=2.67E-17) GO:0019222 regulation of metabolic process (p=4.84E-14) GO:0005488 binding (p=8.39E-13) GO:0071310 cellular response to organic substance (p=7.17E-06) GO:0098805 whole membrane (p=1.43E-04) GO:0016740 transferase activity (p=1.49E-04) GO:0005923 bicellular tight junction (p=1.33E-02) GO:0007268 synaptic transmission (p=1.97E-02) GO:0016569 covalent chromatin modification (p=2.32E-02) GO:0036293 response to decreased oxygen levels (p=3.21E-02) GO:0042552 myelination (p=3.52E-02)		

hsa-miR-6810-5p	7641	6956	0.658	0.702	0.654
Ref targets functional enrichment			**Alt targets functional enrichment**		
GO:0007268 synaptic transmission (p=3.22E-19) GO:0098805 whole membrane (p=8.77E-12) GO:0016773 phosphotransferase activity, alcohol group as acceptor (p=2.82E-07) GO:0005488 binding (p=3.27E-07) GO:0046873 metal ion transmembrane transporter activity (p=6.91E-07) GO:0016310 phosphorylation (p=4.82E-06) GO:0033674 positive regulation of kinase activity (p=5.95E-06) GO:0036211 protein modification process (p=1.30E-05) GO:0006357 regulation of transcription from RNA polymerase II promoter (p=6.18E-05) GO:0030036 actin cytoskeleton organization (p=8.42E-05) GO:0071363 cellular response to growth factor stimulus (p=9.38E-05) GO:0034702 ion channel complex (p=4.41E-03) GO:0016482 cytosolic transport (p=4.78E-03) GO:1901700 response to oxygen-containing compound (p=1.29E-02) GO:0017002 activin-activated receptor activity (p=1.44E-02) GO:0005057 receptor signaling protein activity (p=1.46E-02) GO:0043565 sequence-specific DNA binding (p=2.37E-02) GO:0046872 metal ion binding (p=2.46E-02) GO:0000981 RNA polymerase II transcription factor activity, sequence-specific DNA binding (p=2.63E-02) GO:0030659 cytoplasmic vesicle membrane (p=3.68E-02) GO:0048015 phosphatidylinositol-mediated signaling (p=4.23E-02) GO:0006835 dicarboxylic acid transport (p=4.32E-02) GO:0043552 positive regulation of phosphatidylinositol 3-kinase activity (p=4.63E-02)			GO:0048731 system development (p=2.38E-18) GO:0045202 synapse (p=1.42E-12) GO:0005488 binding (p=3.81E-08) GO:0046873 metal ion transmembrane transporter activity (p=6.86E-08) GO:0009719 response to endogenous stimulus (p=7.40E-07) GO:0031328 positive regulation of cellular biosynthetic process (p=7.99E-07) GO:0007166 cell surface receptor signaling pathway (p=1.26E-06) GO:1903508 positive regulation of nucleic acid-templated transcription (p=1.30E-06) GO:0036211 protein modification process (p=1.18E-05) GO:0071310 cellular response to organic substance (p=2.00E-05) GO:0010562 positive regulation of phosphorus metabolic process (p=6.73E-05) GO:1901700 response to oxygen-containing compound (p=1.35E-04) GO:0004672 protein kinase activity (p=3.93E-04) GO:0098588 bounding membrane of organelle (p=5.84E-04) GO:1902531 regulation of intracellular signal transduction (p=7.22E-04) GO:0071417 cellular response to organonitrogen compound (p=1.01E-03) GO:0033674 positive regulation of kinase activity (p=2.00E-03) GO:0034702 ion channel complex (p=5.63E-03) GO:0030659 cytoplasmic vesicle membrane (p=5.71E-03) GO:0019900 kinase binding (p=1.41E-02) GO:0006835 dicarboxylic acid transport (p=2.26E-02) GO:0001071 nucleic acid binding transcription factor activity (p=3.94E-02)		
hsa-miR-6826-5p	3160	6988	0.341	0.608	0.667
Ref targets functional enrichment			**Alt targets functional enrichment**		
GO:0044424 intracellular part (p=3.11E-15) GO:0043167 ion binding (p=6.34E-10) GO:0060255 regulation of macromolecule metabolic process (p=1.27E-09) GO:0044707 single-multicellular organism process (p=6.73E-04) GO:0003677 DNA binding (p=3.31E-03) GO:0070647 protein modification by small protein conjugation or removal (p=1.62E-02) GO:0010720 positive regulation of cell development (p=4.33E-02)			GO:0044424 intracellular part (p=4.10E-26) GO:0019222 regulation of metabolic process (p=1.02E-13) GO:0005515 protein binding (p=3.12E-12) GO:0035556 intracellular signal transduction (p=3.32E-06) GO:0098805 whole membrane (p=4.98E-05) GO:1901214 regulation of neuron death (p=1.88E-04) GO:0016740 transferase activity (p=3.87E-04) GO:0006820 anion transport (p=2.11E-03) GO:0000975 regulatory region DNA binding (p=6.10E-03) GO:0044723 single-organism carbohydrate metabolic process (p=3.04E-02) GO:0051650 establishment of vesicle localization (p=3.61E-02) GO:0007167 enzyme linked receptor protein signaling pathway (p=4.21E-02) GO:0099536 synaptic signaling (p=4.32E-02)		
hsa-miR-6850-3p	1163	4851	0.286	0.238	0.264
Ref targets functional enrichment			**Alt targets functional enrichment**		
GO:0009889 regulation of biosynthetic process (p=6.98E-06) GO:0031326 regulation of cellular biosynthetic process (p=7.25E-06) GO:0043565 sequence-specific DNA binding (p=1.44E-05) GO:0051252 regulation of RNA metabolic process (p=2.75E-05) GO:0000976 transcription regulatory region sequence-specific DNA binding (p=6.66E-05) GO:0007399 nervous system development (p=2.26E-04) GO:0048522 positive regulation of cellular process (p=3.60E-03) GO:0048523 negative regulation of cellular process (p=1.01E-02) GO:0045202 synapse (p=1.88E-02)			GO:0051179 localization (p=3.46E-09) GO:0044459 plasma membrane part (p=2.83E-08) GO:0005515 protein binding (p=1.85E-07) GO:0007268 synaptic transmission (p=1.84E-03) GO:0030036 actin cytoskeleton organization (p=5.33E-03) GO:0014069 postsynaptic density (p=5.73E-03) GO:0036211 protein modification process (p=8.42E-03) GO:0035556 intracellular signal transduction (p=1.61E-02) GO:0046873 metal ion transmembrane transporter activity (p=4.42E-02)		

hsa-miR-6886-5p	1793	10138	0.374	0.659	0.626
Ref targets functional enrichment			**Alt targets functional enrichment**		
GO:0048731 system development (p=3.67E-09) GO:0065007 biological regulation (p=6.90E-09) GO:0051179 localization (p=1.32E-05) GO:0005488 binding (p=1.16E-04) GO:0010646 regulation of cell communication (p=3.57E-04) GO:0043565 sequence-specific DNA binding (p=4.37E-04) GO:0035556 intracellular signal transduction (p=7.72E-04) GO:0000981 RNA polymerase II transcription factor activity, sequence-specific DNA binding (p=1.17E-03) GO:0023051 regulation of signaling (p=1.62E-03) GO:0098805 whole membrane (p=2.99E-03) GO:0051172 negative regulation of nitrogen compound metabolic process (p=5.79E-03) GO:0006366 transcription from RNA polymerase II promoter (p=8.43E-03) GO:0042995 cell projection (p=8.71E-03) GO:0003690 double-stranded DNA binding (p=1.02E-02) GO:0097458 neuron part (p=1.16E-02) GO:0031327 negative regulation of cellular biosynthetic process (p=1.21E-02) GO:0000989 transcription factor activity, transcription factor binding (p=3.17E-02) GO:0010468 regulation of gene expression (p=3.52E-02)			GO:0007275 multicellular organism development (p=3.51E-20) GO:0044424 intracellular part (p=4.81E-19) GO:0005488 binding (p=2.04E-17) GO:0098805 whole membrane (p=7.65E-08) GO:0006464 cellular protein modification process (p=2.98E-07) GO:0016740 transferase activity (p=2.59E-05) GO:0006357 regulation of transcription from RNA polymerase II promoter (p=4.38E-05) GO:0010033 response to organic substance (p=6.83E-05) GO:0043087 regulation of GTPase activity (p=9.44E-04) GO:0052697 xenobiotic glucuronidation (p=2.82E-03) GO:0030036 actin cytoskeleton organization (p=5.55E-03) GO:0005654 nucleoplasm (p=1.05E-02) GO:1901700 response to oxygen-containing compound (p=1.08E-02) GO:2001030 negative regulation of cellular glucuronidation (p=1.31E-02) GO:1904223 regulation of glucuronosyltransferase activity (p=1.31E-02) GO:0001071 nucleic acid binding transcription factor activity (p=1.62E-02) GO:0015629 actin cytoskeleton (p=2.28E-02) GO:0000975 regulatory region DNA binding (p=3.68E-02) GO:0003014 renal system process (p=4.45E-02) GO:0042060 wound healing (p=4.56E-02)		
hsa-miR-938	4396	3036	0.622	0.349	0.759
Ref targets functional enrichment			**Alt targets functional enrichment**		
GO:0048731 system development (p=5.03E-09) GO:0048518 positive regulation of biological process (p=2.13E-08) GO:0005737 cytoplasm (p=8.31E-07) GO:0043167 ion binding (p=5.66E-06) GO:0051179 localization (p=1.76E-05) GO:0010646 regulation of cell communication (p=2.00E-04) GO:0034765 regulation of ion transmembrane transport (p=8.23E-03) GO:0007156 homophilic cell adhesion via plasma membrane adhesion molecules (p=9.97E-03) GO:0005244 voltage-gated ion channel activity (p=1.43E-02) GO:0046873 metal ion transmembrane transporter activity (p=1.58E-02) GO:0009719 response to endogenous stimulus (p=4.65E-02) GO:0070838 divalent metal ion transport (p=5.00E-02)			GO:0044424 intracellular part (p=1.68E-05) GO:0043231 intracellular membrane-bounded organelle (p=1.02E-04) GO:0010646 regulation of cell communication (p=2.41E-04) GO:0019222 regulation of metabolic process (p=3.83E-04) GO:0023051 regulation of signaling (p=1.43E-03) GO:0044260 cellular macromolecule metabolic process (p=2.62E-03) GO:0048518 positive regulation of biological process (p=4.51E-03) GO:0071705 nitrogen compound transport (p=6.04E-03) GO:0016482 cytosolic transport (p=7.30E-03) GO:0071702 organic substance transport (p=1.67E-02) GO:0051179 localization (p=3.87E-02) GO:0048812 neuron projection morphogenesis (p=4.31E-02) GO:1901363 heterocyclic compound binding (p=4.96E-02)		

Table 10. List of variants with known disease associations.

miRNA (subgroup)	Disease	Disease class	Tissue/ Cell line	Regulation	Pubmed ID	Accession ID
hsa-mir-105-2 (indel)	Breast cancer	Cancer	MCF-7 cell	down	16192569	MI0000112
	Breast cancer	Cancer	T-47D cell	down	16192569	MI0000112
	Breast cancer	Cancer	MDA-MB-231 cell	down	16192569	MI0000112
	Breast cancer	Cancer	SK-BR-3 cell	up	16192569	MI0000112
	Breast cancer	Cancer	MDA-MB-361 cell	up	16192569	MI0000112
	Hematological	Hematological	K-562 cell	up	16192569	MI0000112
	Lung cancer	Cancer	A-549 cell	down	16192569	MI0000112
	Lung cancer	Cancer	lung cancer cell line	down	16192569	MI0000112
	Pancreatic cancer	Cancer	PANC-1 cell	down	16192569	MI0000112
	Prostate cancer	Cancer	PC-3 cell	down	16192569	MI0000112
	Prostate cancer	Cancer	Tsu-Pr1 cell	down	16192569	MI0000112
	Prostate cancer	Cancer	PPC-1 cell	down	16192569	MI0000112
	Prostate cancer	Cancer	LNCaP cell	down	16192569	MI0000112
	Prostate cancer	Cancer	DU-145 cell	down	16192569	MI0000112
	Squamous cell carcinoma, head and neck	Cancer	squamous cell carcinoma cell line	down	16192569	MI0000112
	Squamous cell carcinoma, head and neck	Cancer	squamous cell carcinoma cell line	down	16192569	MI0000112
hsa-mir-302c (indel)	Breast cancer	Cancer	SK-BR-3 cell	down	16192569	MI0000773
	Breast cancer	Cancer	MDA-MB-231 cell	down	16192569	MI0000773
	Breast cancer	Cancer	MCF-7 cell	down	16192569	MI0000773
	Breast cancer	Cancer	MDA-MB-361 cell	up	16192569	MI0000773
	Breast cancer	Cancer	T-47D cell	down	16192569	MI0000773
	Breast cancer	Cancer	breast epithelium	down	16754881	MI0000773
	Cancer	Cancer	thyroid gland	up	18270258	MI0000773
	Cancer	Cancer	JURKAT cell	up	16934749	MI0000773

	Cardiomyopathy, dilated	Cardiovascular	left ventricle	down	17606841	MI0000773
	Hematological	Hematological	K-562 cell	up	16192569	MI0000773
	Hodgkin lymphoma	Cancer	lymph node	up	18089852	MI0000773
	Leukemia, acute myelogenous	Cancer	HL-60 cell	up	16934749	MI0000773
	Leukemia, megakaryoblastic, with or without Down syndrome	Cancer	CMK cell	up	16934749	MI0000773
	Lung cancer	Cancer	lung cancer cell line	down	16192569	MI0000773
	Lung cancer	Cancer	A-549 cell	down	16192569	MI0000773
	Melanoma and neural system tumor syndrome	Cancer	cell culture	down	16754881	MI0000773
	Melanoma and neural system tumor syndrome	Cancer	melanocyte	down	18379589	MI0000773
	Miyoshi myopathy	Muscular	muscle	down	17942673	MI0000773
	Non-Hodgkin lymphoma, somatic	Cancer	U-937 cell	up	16934749	MI0000773
	Ovarian cancer	Cancer	ovary	down	17875710	MI0000773
	Ovarian cancer	Cancer	ovary	up	17875710	MI0000773
	Ovarian cancer	Cancer	ovary	down	18458333	MI0000773
	Ovarian cancer	Cancer	ovary	down	16754881	MI0000773
	Ovarian cancer	Cancer	ovary	down	17875710	MI0000773
	Pancreatic cancer	Cancer	PANC-1 cell	down	16192569	MI0000773
	Prostate cancer	Cancer	PC-3 cell	up	17616669	MI0000773
	Prostate cancer	Cancer	PC-3 cell	down	16192569	MI0000773
	Prostate cancer	Cancer	Tsu-Pr1 cell	down	16192569	MI0000773
	Prostate cancer	Cancer	PPC-1 cell	down	16192569	MI0000773
	Prostate cancer	Cancer	LNCaP cell	down	16192569	MI0000773
	Prostate cancer	Cancer	prostate gland	up	17616669	MI0000773
	Prostate cancer	Cancer	DU-145 cell	down	16192569	MI0000773
	Squamous cell carcinoma, head and neck	Cancer	squamous cell carcinoma cell line	down	16192569	MI0000773
	Squamous cell carcinoma, head and neck	Cancer	squamous cell carcinoma cell line	down	16192569	MI0000773

	Squamous cell carcinoma, head and neck	Cancer	HSC-3 cell	down	18381414	MI0000773
hsa-miR-557 (HWE)	Systemic lupus erythematosus (SLE)	Connective tissue	renal cortex	down	18998140	MI0003563
hsa-mir-627 (MAF>5%)	Ovarian cancer	Cancer	ovary	down	18560586	MI0003641
hsa-miR-662 (HWE)	Breast cancer	Cancer	MCF-7 cell	up	20543867	MI0003670
	Ovarian cancer	Cancer	ovary	down	18560586	MI0003670
	Ovarian cancer	Cancer	OVCA-420 cell	down	18560586	MI0003670
	Systemic lupus erythematosus (SLE)	Connective tissue	renal cortex	up	18998140	MI0003670

Table 11. Summary of the SNV within the miR-501 seed region.

miRNA	**Chr:position**		**position in seed**	**SNV**	**MAF**
hsa-mir-501	ChrX:49774381		2	A/G	0.00593472
	# of targets (ref)	**# of targets (alt)**	**Biological processes similarity score**	**Molecular function similarity score**	**Cellular component similarity score**
	3748	3931	0.497	0.358	0.4
Ref targets functional enrichment			**Alt targets functional enrichment**		
GO:0043167 ion binding (p=1.61E-09) GO:0043005 neuron projection (p=1.96E-06) GO:0008150 biological_process (p=4.09E-05) GO:0007399 nervous system development (p=6.11E-05) GO:0007156 homophilic cell adhesion via plasma membrane adhesion molecules (p=1.80E-04) GO:0010468 regulation of gene expression (p=4.54E-03) GO:0044238 primary metabolic process (p=6.06E-03) GO:0004842 ubiquitin-protein transferase activity (p=1.85E-02) GO:0097659 nucleic acid-templated transcription (p=3.09E-02) GO:0050885 neuromuscular process controlling balance (p=3.31E-02) GO:0031344 regulation of cell projection organization (p=3.67E-02)			GO:0048518 positive regulation of biological process (p=6.36E-08) GO:0051179 localization (p=3.83E-07) GO:0050794 regulation of cellular process (p=1.26E-06) GO:0023051 regulation of signaling (p=3.83E-06) GO:0010646 regulation of cell communication (p=3.98E-06) GO:0005622 intracellular (p=1.13E-05) GO:0007275 multicellular organism development (p=2.82E-05) GO:0060255 regulation of macromolecule metabolic process (p=6.52E-05) GO:0016020 membrane (p=1.59E-04) GO:0006366 transcription from RNA polymerase II promoter (p=2.96E-03) GO:0019899 enzyme binding (p=3.54E-03) GO:0071363 cellular response to growth factor stimulus (p=8.55E-03) GO:0045859 regulation of protein kinase activity (p=1.47E-02)		

Chapter 4. CONCLUSIONS

In recent decades, our understanding of functional genomics has been facilitated by technological advances in high throughput transcriptomic and proteomic methods. These tools are also rapidly expanding our understanding of miRNA, a relatively new class of trans-acting regulators of gene expression that are quickly being recognized as potential sources of phenotypic variation or as biomarkers for disease. Reliable sequence information across a diverse range of species is required for researchers to comprehensively study miRNA evolution.

In this study, we have greatly expanded the number of experimentally validated non-human primate miRNAs, especially in more divergent sister taxa. Inclusion of New World monkeys, lemurs, and a galago in this study made it possible to identify more ancient evolutionary events that may have shaped primate evolution, such as the mature sequence shift we identified in miR-501-3p and the duplications of the miR-320 family. We have also demonstrated that more than a fourth of all computationally predicted primate miRNAs were found within our data (despite having only sequenced one cell type), lending confidence to prediction by homology as a method of miRNA discovery. However, our results also illustrate the importance of validating mature miRNA through RNAseq: mature miRNA sequence shifts caused by changes in secondary structure cannot be reliably determined by homology alone, and require sequencing reads to determine their boundaries.

Because this study only sequenced miRNA from a single cell type (cultured fibroblasts), we likely captured only a subset of the miRNAs expressed within each species. However, it is notable that fibroblasts and neurons are both derived from the ectoderm (Chang and Hemmati-Brivanlou 1998), and conversion of fibroblasts to neurons is relatively easy (Vierbuchen et al. 2010); additionally, fibroblasts are frequently used as a model when studying brain disorders

because of their neuron-like signal transduction pathways (Manier et al. 2000; Garbett et al. 2015). This relationship between fibroblasts and neurons possibly explains the abundance of neuronally-expressed miRNA identified in this study. Nonetheless, even a single cell type was sufficient for finding a number of interesting evolutionary events.

We also identified 4521 SNVs across a broad representation of human populations from the 1000 Genomes Project, confirming that the seed region and mature miRNA are most highly conserved, even among common variants. The abundance of rare variants and lack of population substructure suggests that purifying selection has been the driving force for human miRNA evolution since humans migrated out of Africa. Most variants are likely either neutral and propagated through genetic drift, or deleterious and not yet removed by purifying selection. Although human miRNA variation may give important insight into disease (as demonstrated by their common usage as biomarkers), these results suggest it may be less useful in identifying phenotypic differences between human populations.

More sequencing efforts across a diverse range of cell types and stages of development are likely to reveal additional insights, especially in cell types already known to have undergone significant phenotypic changes (i.e., neuronal tissue). Our study also highlights the importance of the inclusion of more distantly related primate species, as most important evolutionary events in miRNA are likely to be ancient. Additionally, more research is needed to confirm the actual biological targets of miRNA of interest described in this study, as target prediction software is not accurate given the complicated binding interactions of miRNA: TargetScan and miranda, two of the best predictive software currently available, both have false positive rates of ~25% (Mazière and Enright 2007). Even target identification is starting to benefit from high throughput technology, such as expression profiling after miRNA knockdown or overexpression (Thomas et

al. 2010), as well as UV cross-linking miRNA-mRNA duplexes to RISC to be pulled out via immunoprecipitation (Hausser and Zavolan 2014). We hope that this study serves as a foundation for future research into the evolution of miRNA and gene regulation in primates.

BIBLIOGRAPHY

1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* **526**: 68-74.

Altuvia Y, Landgraf P, Lithwick G, Elefant N, Pfeffer S, Aravin A, Brownstein MJ, Tuschl T, Margalit H. 2005. Clustering and conservation patterns of human microRNAs. *Nucleic Acids Res* **33**: 2697-2706.

Baev V, Daskalova E, Minkov I. 2009. Computational identification of novel microRNA homologs in the chimpanzee genome. *Comput Biol Chem*. **33**: 62-70. doi: 10.1016/j.compbiolchem.2008.07.024

Berezikov E, Guryev V, van de Belt J, Wienholds E, Plasterk RH, Cuppen E. 2005. Phylogenetic shadowing and computation identification of human microRNA genes. *Cell* **120**: 21-24.

Berezikov E, Thuemmler F, van Laake LW, Kondova I, Bontrop R, Cuppen E, Plasterk RH. 2006. Diversity of microRNAs in human and chimpanzee brain. *Nature Genetics*. **38**: 1375-1377.

Berezikov E. 2011. Evolution of microRNA diversity and regulation in animals. *Nat Rev Gen* **12**: 846-860.

Bhartiya D, Laddha SV, Mukhopadhyay A, Scaria V. 2011. miRvar: A comprehensive database for genomic variations in microRNAs. *Hum Mutat* **32**: E2226-45. doi: 10.1002/humu.21482

Bhatia G, Patterson N, Sankararaman S, Price AL. 2013. Estimating and interpreting F_{ST}: the impact of rare variants. *Genome Res* **23**: 1514-1521

Bhattacharya A and Cui Y. 2015. miR2GO: comparative functional analysis for microRNAs. *Bioinformatics* **31**: 2403-2405.

Biegert A, Mayer C, Remmert M, Söding J, Lupas A. 2006. The MPI Bioinformatics Toolkit for protein sequence analysis. *Nucleic Acids Res* **34**: W335-339. doi: 10.1093/nar/gkl217

Bonnet E, Wuyts J, Rouzé P, Van de Peer Y. 2004. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics* **20**: 2911-2917. doi: 10.1093/bioinformatics/bth374

Bradley RK, Roberts A, Smoot M, Juvekar S, Do J, Dewey C, Holmes I, Pachter L. 2009. Fast statistical alignment. *PLoS Comput Biol* **5**:e1000392. doi: 10.1371/journal.pcbi.1000392

Brameier M. 2010. Genome-wide comparative analysis of microRNAs in three non-human primates. *BMC Res Notes* **3**: 64. doi: 10.1186/1756-0500-3-64

Carbonell J, Alloza E, Arce P, Borrego S, Santoyo J, Ruiz-Ferrer M, Medina I, Jiménez-Almazán J, Méndez-Vidal C, González-Del Pozo M, Vela A, Bhattacharya SS, Antiñolo G, Dopazo J. 2012. A map of human microRNA variation uncovers unexpectedly high levels of variability. *Genome Med* **4**: 62. doi: 10.1186/gm363

Chen K and Rajewsky N. 2007. The evolution of gene regulation by transcription factors and microRNAs. *Nat Rev Genet* **8**: 93-103.

The Chimpanzee Sequencing and Analysis Consortium. 2004. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**: 69-87.

Chang C and Hemmati-Brivanlou A. 1998. Cell fate determination in embryonic ectoderm. J *Neurobiol* **36**: 128-151.

Conant GC and Wolfe KH. 2008. Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet* **9**: 938-950. doi: 10.1038/nrg2482

Creighton CJ, Benham AL, Zhu H, Khan MF, Reid JG, Nagaraja AK, Fountain MD, Dziadek O, Han D, Ma L, et al. 2010. Discovery of novel microRNAs in female reproductive tract using next generation sequencing. *PLoS One* **5**: e9637. doi: 10.1371/journal.pone.0009637

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group. 2011. The variant call format and VCFtools. *Bioinformatics* **27**: 2156-2158

Dannemann M, Nickel B, Lizano E, Burbano HA, Kelso J. 2012. Annotation of primate miRNAs by high throughput sequencing of small RNA libraries. BMC Genomics 13: 116. doi: 10.1186/1471-2164-13-116

Duan R, Pak C, Jin P. 2007. Single nucleotide polymorphism associated with mature miR-125a alters the processing of pri-miRNA. *Hum Mol Genet* **16**: 1124-1131.

Enard W, Khaitovich P, Klose J, Zöllner S, Heissig F, Giavalisco P, Nieselt-Struwe K, Muchmore E, Varki A, Ravid R, et al. 2002. Intra- and interspecific variation in primate gene expression patterns. *Science* **296**: 340–343

Fang Z, Du R, Edwards A, Flemington EK, Zhang K. 2013. The sequence structures of human microRNA molecules and their implications. *PLOS One* **8**: e54215. doi: 10.1371/journal.pone.0054215

Friedländer MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, Rajewsky N. 2008. Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotechnol* **26**: 407-415.

Friedman RC, Farh KK, Burge CB, Bartel DP. 2009. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* **19**: 92-105.

Gallego A, Melé M, Balcells I, García-Ramallo E, Torruella-Loran I, Fernández-Bellon H, Abelló T, Kondova I, Bontrop R, Hvilsom C, Navarro A, Marquès-Bonet T, Espinosa-Parrilla Y. 2016. Functional Implications of Human-Specific Changes in Great Ape microRNAs. *PLoS One* **11**: e0154194. doi: 10.1371/journal.pone.0154194

Gao Z, Ding P, Hsieh J. 2012. Profiling of REST-dependent microRNAs reveals dynamic modes of expression. *Front Neurosci* **6**: 67. doi: 10.3389/fnins.2012.00067

Garbett KA, Vereczkei A, Kálmán S, Brown JA, Taylor WD, Faludi G, Korade Ž5, Shelton RC, Mirnics K. 2015. Coordinated messenger RNA/microRNA changes in fibroblasts of patients with major depression. *Biol Psychiatry* **77**: 256-265. doi: 10.1016/j.biopsych.2014.05.015

George RD, McVicker G, Diederich R, Ng SB, MacKenzie AP, Swanson WJ, Shendure J, Thomas JH. 2011. Trans genomic capture and sequencing of primate exomes reveals new targets of positive selection. *Genome Res* **21**: 1686-1694.

Gkirtzou K, Tsamardinos I, Tsakalides P, Poirazi P. 2010. MatureBayes: A probabilistic algorithm for identifying the mature miRNA within novel precursors. *PLoS One* **5**: e11843. doi: 10.1371/journal.pone.0011843.

Glazov EA, Kongsuwan K, Assavalapsakul W, Horwood PF, Mitter N, Mahony TJ. 2009. Repertoire of bovine miRNA and miRNA-like small regulatory RNAs expressed upon viral infection. *PLoS One* **4**: e6349. doi: 10.1371/journal.pone.0006349

Goff LA, Davila J, Swerdel MR, Moore JC, Cohen RI, Wu H, Sun YE, Hart RP. 2009. Ago2 immunoprecipitation identifies predicted microRNAs in human embryonic stem cells and neural precursors. *PLoS One* **4**: e7192. doi: 10.1371/journal.pone.0007192

Gong J, Tong Y, Zhang HM, Wang K, Hu T, Shan G, Sun J, Guo AY. 2012. Genome-wide identification of SNPs in microRNA genes and the SNP effects on microRNA target binding and biogenesis. *Hum Mutat* **33**: 254-263.

Goode DL, Cooper GM, Schmutz J, Dickson M, Gonzales E, Tsai M, Karra K, Davydov E, Batzoglou S, Myers RM, et al. 2010. Evolutionary constraint facilitates interpretation of genetic variation in resequenced human genomes. *Genome Res* **20**: 301-310.

Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP. 2007. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell*. **27**: 91-105. doi:10.1016/j.molcel.2007.06.017.

Gruber AR, Bernhart SH, Hofacker IL, Washietl S. 2008. Strategies for measuring evolutionary conservation of RNA secondary structures. *BMC Bioinformatics*. **9**: 122. doi: 10.1186/1471-2105-9-122

Gruber AR, Findeiß S, Washietl S, Hofacker IL, Stadler PF. 2009. RNAz 2.0: Improved noncoding RNA detection. *Pac Symp Biocomput* 69-79. doi: 10.1142/9789814295291_0009

Hamam D, Ali D, Vishnubalaji R, Hamam R, Al-Nbaheen M, Chen L, Kassem M, Aldahmash A, Alajez NM. 2014. microRNA-320/RUNX2 axis regulates adipocytic differentiation of human mesenchymal (skeletal) stem cells. *Cell Death Dis* **5**: e1499. doi:10.1038/cddis.2014.462

Han J, Lee Y, Yeom KH, Nam JW, Heo I, Rhee JK, Sohn SY, Cho Y, Zhang BT, Kim VN. 2006. Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell*. **125**: 887-901.

Hansen TB, Bramsen JB, Kjems J. 2010. Re-inspection of small RNA sequence datasets reveal several novel human miRNA genes. *PLoS One* **5**: e10961. doi: 10.1371/journal.pone.0010961

Hausser J and Zavolan M. 2014. Identification and consequences of miRNA–target interactions — beyond repression of gene expression. *Nat Rev Genet* **15**: 599–612. doi:10.1038/nrg3765

He L and Hannon GJ. 2004. MicroRNAs: small RNAs with a big role in gene regulation. *Nat Rev Genet* **5**: 522-531.

Hébert SS, Horré K, Nicolaï L, Papadopoulou AS, Mandemakers W, Silahtaroglu AN, Kauppinen S, Delacourte A, De Strooper B. 2008. Loss of microRNA cluster miR-29a/b-1 in sporadic Alzheimer's disease correlates with increased BACE1/β-secretase expression. *Proc Natl Acad Sci USA* **105**: 6415–6420. doi: 10.1073/pnas.0710263105

Hiard S, Charlier C, Coppieters W, Georges M, Baurain D. 2010. Patrocles: a database of polymorphic miRNA-mediated gene regulation in vertebrates. *Nucleic Acids Res* **38**: D640-51. doi: 10.1093/nar/gkp926

Holsinger KE and Weir BS. 2009. Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. *Nat Rev Genet* **10**: 639-650.

Hu HY, Guo S, Xi J, Yan Z, Fu N, Zhang X, Menzel C, Liang H, Yang H, Zhao M, et al. 2011. MicroRNA expression and regulation in human, chimpanzee, and macaque brains. *PLoS Genet* **7**: e1002327. doi: 10.1371/journal.pgen.1002327

Hu Z, Zhao J, Hu T, Luo Y, Zhu J, Li Z. 2015. miR-501-3p mediates the activity-dependent regulation of the expression of AMPA receptor subunit GluA1. *J Cell Biol* **208**: 949-959. doi: 10.1083/jcb.201404092.

Iwai N and Naraba H. 2005. Polymorphisms in human pre-miRNAs. *Biochem Biophys Res Comm* **331**: 1439-1444.

Iwagami Y, Eguchi H, Nagano H, Akita H, Hama N, Wada H, Kawamoto K, Kobayashi S, Tomokuni A, Tomimaru Y, et al. 2013. miR-320c regulates gemcitabine-resistance in pancreatic cancer via SMARCC1. *Br J Cancer* **109**: 502-511. doi:10.1038/bjc.2013.320

King MC and Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. *Science* **188**: 107-116.

Kozomara A and Griffiths-Jones S. 2011. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res* **39**: 152-157.

Lee C, Risom T, Strauss WM. 2007. Evolutionary conservation of microRNA regulatory circuits: An examination of microRNA gene complexity and conserved microRNA-target interactions through metazoan phylogeny. *DNA Cell Biol* **26**: 209-218.

Li WH and Tanimura M. 1987. The molecular clock runs more slowly in man than in apes and monkeys. *Nature* **326**: 93-9.

Li Y and Kowdley KV. 2012. MicroRNAs in common human diseases. *Genomics Proteomics Bioinformatics* 10: 246-253. doi: 10.1016/j.gpb.2012.07.005

Liu C, Zhang F, Li T, Lu M, Wang L, Yue W, Zhang D. 2012. MiRSNP, a database of polymorphisms altering miRNA target sites, identifies miRNA-related SNPs in GWAS SNPs and eQTLs. *BMC Genomics* **13**: 661. doi: 10.1186/1471-2164-13-661

Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA Package 2.0. *Algorithms Mol Biol.* **6**: 26. doi: 10.1186/1748-7188-6-26

Manier DH, Shelton RC, Ellis TC, Peterson CS, Eiring A, Sulser F. 2000. Human fibroblasts as a relevant model to study signal transduction in affective disorders. *J Affect Disord* **65**: 51-58.

Martí E, Pantano L, Bañez-Coronel M, Llorens F, Miñones-Moyano E, Porta S, Sumoy L, Ferrer I, Estivill X. 2010. A myriad of miRNA variants in control and Huntington's disease brain regions detected by massively parallel sequencing. *Nucleic Acids Res* **38**: 7219 – 7235. doi: 10.1093/nar/gkq575

Martin HC, Wani S, Steptoe AL, Krishnan K, Nones K, Nourbakhsh E, Vlassov A, Grimmond SM, Cloonan N. 2014. Imperfect centered miRNA binding sites are common and can mediate repression of target mRNAs. *Genome Biol* **15**: R51. doi: 10.1186/gb-2014-15-3-r51

Mazière P and Enright AJ. 2007. Prediction of microRNA targets. *Drug Discov Today* **12**: 452-458.

Natera-Naranjo O, Aschrafi A, Gioio AE, Kaplan BB. 2010. Identification and quantitative analyses of microRNAs located in the distal axons of sympathetic neurons. *RNA* **16**: 1516-1529 doi: 10.1261/rna.1833310

Nilsen TW. 2007. Mechanisms of microRNA-mediated gene regulation in animal cells. *Trends Genet* **23**: 243–249

Otto SJ, McCorkle SR, Hover J, Conaco C, Han JJ, Impey S, Yochum GS, Dunn JJ, Goodman RH, Mandel G. 2007. A new binding motif for the transcriptional repressor REST uncovers large gene networks devoted to neuronal functions. *J Neurosci* **27**: 6729–6739.

Pai AA, Bell JT, Marioni JC, Pritchard JK, Gilad Y. 2011. A genome-wide study of DNA methylation patterns and gene expression levels in multiple human and chimpanzee tissues. *PLoS Genet* **7**: e1001316. doi: 10.1371/journal.pgen.1001316

Pang KC, Frith MC, Mattick JS. 2006. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet* **22**: 1-5. doi:10.1016/j.tig.2005.10.003

Perelman P, Johnson WE, Roos C, Seuánez HN, Horvath JE, Moreira MA, Kessing B, Pontius J, Roelke M, Rumpler Y, et al. 2011. A molecular phylogeny of living primates. *PLoS Genet* **7**: e1001342. doi: 10.1371/journal.pgen.1001342

Pritchard CC, Cheng HH, Tewari M. 2012. MicroRNA profiling: approaches and considerations. *Nat Rev Genet* **13**: 358-369.

Ruepp A, Kowarsch A, Theis F. 2012. PhenomiR: microRNAs in human diseases and biological processes. *Methods Mol Biol* **822**: 249-260.

Saba R, Goodman CD, Huzarewich RL, Robertson C, Booth SA. 2008. A miRNA signature of prion induced neurodegeneration. *PLOS One* **3**: e3652. doi: 10.1371/journal.pone.0003652

Saunders MA, Liang H, Li WH. 2007. Human polymorphisms at microRNAs and microRNA target sites. *Proc Nat Acad of Sci* **104**: 3300-3305.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**: 308-311.

Shin C, Nam JW, Farh KK, Chiang HR, Shkumatava A, Bartel DP. 2010. Expanding the microRNA targeting code: functional sites with centered pairing. *Mol Cell* **38**: 789-802.

Somel M, Liu X, Tang L, Yan Z, Hu H, Guo S, Jiang X, Zhang X, Xu G, Xie G, et al. 2011. MicroRNA-driven developmental remodeling in the brain distinguishes humans from other primates. *PLoS Biol* **9**: e1001214. doi: 10.1371/journal.pbio.1001214

Speir ML, Zweig AS, Rosenbloom KR, Raney BJ, Paten B, Nejad P, Lee BT, Learned K, Karolchik D, Hinrichs AS, et al. 2016. The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res* **44**: D717-D725. doi: 10.1093/nar/gkv1275

Stamatakis A. 2014. RAxML Version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312-1313. doi: 10.1093/bioinformatics/btu033

Sun G, Yan J, Noltner K, Feng J, Li H, Sarkis DA, Sommer SS, Rossi JJ. 2009. SNPs in human miRNA genes affect biogenesis and function. *RNA*. **15**: 1640-1651.

Thomas M, Lieberman J, Lal A. 2010. Desperately seeking microRNA targets. *Nat Struct Mol Biol* **17**: 1169-1174.

Ukai T, Sato M, Akutsu H, Umezawa A, Mochida J. 2012. MicroRNA-199a-3p, microRNA-193b, and microRNA-320c are correlated to aging and regulate human cartilage metabolism. *J Orthop Res* **30**: 1915–1922. doi: 10.1002/jor.22157

Vachev TI, Popov NT, Stoyanova VK, Ivanov HY, Minchev DS. 2016. Down regulation of MIR-320 gene family members in the peripheral blood of schizophrenia patients. *Int J Curr Microbiol App Sci* **5**: 221-230. doi: 10.20546/ijcmas.2016.501.020

Vierbuchen T, Ostermeier A, Pang ZP, Kokubu Y, Südhof TC, Wernig M. 2010. Direct conversion of fibroblasts to functional neurons by defined factors. *Nature* **463**: 1035-1041.

Wang X, Wu J, Lin Y, Zhu Y, Xu X, Xu X, Liang Z, Li S, Hu Z, Zheng X, et al. 2014. MicroRNA-320c inhibits tumorous behaviors of bladder cancer by targeting Cyclin-dependent kinase 6. *J Exp Clin Cancer* Res **33**:69. doi: 10.1186/s13046-014-0069-6

Washietl S. 2011. RNAz 2.1 user manual. https://www.tbi.univie.ac.at/~wash/RNAz/manual.pdf

White RE and Gifford RG. 2012. MicroRNA-320 induces neurite outgrowth by targeting ARPP-19. *Neuroreport* **23**: 590-595. doi: 10.1097/WNR.0b013e3283540394.

Witten D, Tibshirani R, Gu SG, Fire A, Lui WO. 2010. Ultra-high throughput sequencing-based small RNA discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls. *BMC Biol* **8**: 58. doi: 10.1186/1741-7007-8-58

Yang S, Yalamanchili HK, Li X, Yao KM, Sham PC, Zhang MQ, Wang J. 2011. Correlated evolution of transcription factors and their binding sites. *Bioinformatics* **27**: 2972-2978. doi: 10.1093/bioinformatics/btr503

Zhang R, Peng Y, Wang W, Su B. 2007. Rapid evolution of an X-linked microRNA cluster in primates. *Genome Res* **17**: 612-617.

Zorc M , Skok DJ, Godnic I, Calin GA, Horvat S, Jiang Z, Dovc P, Kunej T. 2012. Catalog of microRNA seed polymorphisms in vertebrates. *PLoS One* **7**: e30737. doi: 10.1371/journal.pone.0030737

VITA

Jennifer Christina McCreight was born and raised in Northwest Indiana and graduated from Munster High School in 2006. She attended Purdue University in West Lafayette, Indiana where she double majored in Genetics and Ecology, Evolution, and Environmental Biology, and received her Bachelor of Science in Biology in 2010. While at Purdue, she spent three years in the lab of Andrew J. DeWoody studying kangaroo rat population genetics and mating behavior. In 2010 she began her graduate research at the University of Washington, receiving her Doctorate of Philosophy from the Department of Genome Sciences in 2016. In her free time, Jennifer enjoys writing, painting, strategy games, traveling, and cuddling with her cat.