EPiC Series in Computing

Volume 83, 2022, Pages 96–107

Proceedings of 14th International Conference
on Bioinformatics and Computational Biology

Variation of miRNA target sites in the Human Genome

Brad V Bellomo, Helen Piontkivska, and Arvind K. Bansal
Kent State University, Kent, OH
bbellomo@kent.edu, opiontki@kent.edu, arvind@cs.kent.edu

Abstract

MicroRNAs (miRNAs) are important regulators of gene expression in humans and many other organisms. Genetic variation in target sites potentially alters this regulation. Better understanding of patterns of nucleotide changes of these sites can provide new insights into human diseases such as cancer, bacterial and viral diseases. Studies of human variation of miRNA binding sites have been done before. However, we focus our study on miRNA-mRNA pairs that are known to be co-expressed. Unlike previous studies, this work considers combinations of several genetic variants, identifies those that are potentially evolutionary conserved, and considers how frequently alleles occur in the human population. New algorithm for matching the target sites has been described. Our findings confirm the putative functional significance of many alleles already reported in medical cancer-related literature, and suggest others not previously reported in literature. These alleles may be worth investigating further.

1 Introduction

MicroRNAs (miRNAs) are approximately 22-nucleotide long noncoding RNAs that bind to target sites on messenger RNA (mRNA), typically in the 3' untranslated region (3'UTR). miRNA-mRNA binding interferes with protein translation and maturation (Lewis, Shih, Jones-Rhoades, Bartel, & Burge, 2003). Consequently, immune system pathways and control pathways that use these proteins are affected. This interference plays an important role in the initiation and progression of cancer (Calin & Croce, 2006), as well as in the adaptive (Zenz, et al., 2009) and innate (Momen-Heravi & Bala, 2018) immune response.

Understanding of the target binding-sites and the resulting interference can help mitigate many diseases, such as cancer (Ryan, Robles, & Harris, 2010), bacterial and viral diseases (Yuan, Burns, Subramanian, & Blekhman, 2018), including COVID-19 (Ahmadi & Moradi, 2021). Elucidating patterns of genetic variation that affect binding sites is important for better understanding of the function of these binding sites, as well as the downstream proteins and pathways regulated by the miRNA.

Previous studies have been limited by considering SNPs individually, focusing on SNPs without considering other variants such as indels (insertions/deletions), not differentiating between rare and common alleles, and not examining whether targets are evolutionarily conserved, including a large number of false positives between mRNA-miRNA pairs that are not co-expressed. The amount of variant data has also grown significantly in recent years.

In this study, we examine how genetic variations affect these target sites. How common or rare are these variants? To what extent do they affect target sites potentially conserved by evolution? How many of these target site variants are new findings versus how many have already been shown to be functionally significant in published biomedical literature? The major contributions of this paper are:

1. A new algorithm has been proposed that analyzes combinations of variants, not just one at a time.
2. We show that human variants can be used to identify important targets similar to how the degree of evolutionary conservation is currently used for such inferences.
3. We use conservation data to show when variation occurs in evolutionarily conserved targets.
4. Our findings overlap with alleles already reported in medical literature validating our approach.
5. Our approach identified many new alleles (not reported previously in biomedical literature) that may be worth investigating further as disease causing candidates.

The paper is organized as follows: Section 2 describes background concepts. Section 3 describes the related work. Section 4 describes the methodology, algorithm and findings. Section 5 discusses the results and limitations. Section 6 describes the future work and concludes the paper.

2 Background

Each miRNA has a seed region, generally the 2nd – 8th nucleotide from the 5’ end. This seed region binds to mRNA with a Watson-Crick complement of the seed region in mRNA. Notably, most such interactions in animal genomes involve an imperfect match (Bartel D. P., 2009). Figure 1 describes a miRNA binding site.

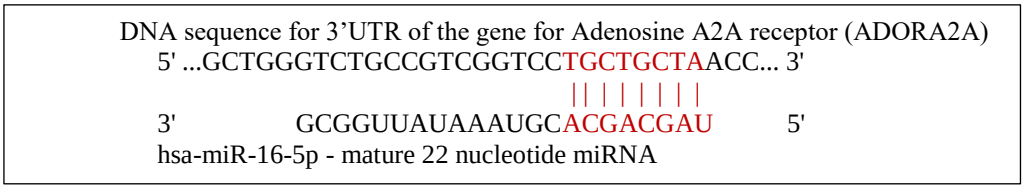

Figure 1. An example miRNA binding-site. The 3’ end of the miRNA aligns with the 5’ end of the mRNA.

In vertebrates, miRNAs occupy a significant portion of genomes (1-2% of regulating non-coding nucleotides). Hence, it is postulated that they play a major role in mRNA regulation and protein maturation (Lewis, Shih, Jones-Rhoades, Bartel, & Burge, 2003). Every developmental stage of every cell-type can have a distinct miRNA expression profile, therefore allowing for fine-tuning of transcriptomes (Bartel D. P., 2004). For example, through miRNA-mRNA binding in ribosomes that interferes with the translation and maturation of the protein and mostly down-regulates the protein translation. Seventy-five miRNA coding genes have been conserved since the ancestor of humans and bony fish. The lack of changes in these genes show these genes are important, and mutations are deleterious (Bartel D. P., 2018). Conserved miRNAs have conserved targets that are also important, however non-conserved targets are estimated to outnumber conserved targets ten to one (Farh, et al., 2005).

Approximately every 500-1000 nucleotides, the human genome has a variation in a single nucleotide (SNP). Polymorphisms that result in insertion or deletion of one or more nucleotides also occur, but less commonly. Genetic variations in the seed regions of target sites of miRNA can alter whether miRNA binds, and to what extent, the mRNA, thus, influencing protein expression levels (Sun, et al., 2009).

Bindings between miRNA and mRNA binding-sites are classified as *canonical* or *non-canonical*. In a canonical binding-site, the nucleotides in a binding subsequence of miRNA are contiguous and fully match for the positions 2 to 7. In the non-canonical binding, there are two or more binding subsequences in miRNA that bind to the target-site; gaps are possible in the alignment. Canonical binding-sites can be further classified into three types: 1) full match; 2) M8 canonical binding-site; 3) A1 canonical binding-site. An M8 canonical binding-site has the same Watson-Crick pairing as a full canonical site, but lacks the adenosine across from the first nucleotide from the 5' end of the mature miRNA. Regardless of whether this nucleotide pair is a Watson-Crick match, the site is classified as an M8 canonical target.

The Needleman-Wunsch algorithm is a dynamic programming $O(n^3)$ algorithm used to align two genomic sequences and derive insertions and deletions (indels) based on user-defined gap-penalty for mismatching nucleotides.

3 Related Works

Genome-wide studies of human variation affecting miRNA targets have been done before (Gong, et al., 2012). Of 30 million SNPs (Single Nucleotide Polymorphisms) in dbSNP build 132 (dbSNP), 225,759 were in the 3' UTR of a gene. Variants of each SNP were run through both TargetScan (Agarwal, Bell, Nam, & Bartel, 2015) and miRanda (Enright, et al., 2003), which identified 58,977 SNPs that could disrupt 90,784 target locations. 59,810 SNPs could create 91,711 new targets. 20,779 SNPs have potential to create and to remove a target, for a total of 89,008 SNPs potentially affecting target-sites.

An update by the same authors (Gong, et al., 2015) included 53 million SNPs, of which 566,176 SNPs were in the 3' UTRs, of which 236,241 SNPs had the potential to delete targets; 263,596 had the potential to add targets; 135,546 SNPs could potentially both add and delete targets. 64% of all SNPs in the 3' UTR were predicted to affect miRNA targets. This update also included data from Diana Tarbase (Karagkouni, et al., 2018), starBase (Li, Liu, Zhou, Qu, & Yang, 2014), miRecords (Xiao, et al., 2009), miRTarBase (Huang, et al., 2020) and miR2diseases (Jiang, et al., 2009) along with data from genome-wide association studies (Hindorff, et al., 2009).

MirSNP (Liu, et al., 2012) is a database of SNPs that affect miRNA target sites. It uses miRanda algorithm (Betel, Wilson, Gabow, Marks, & Sander, 2008) restricted to exact seven nucleotide matches in the 3'UTRs, and includes optional filtering by minor allele frequency. It included 513,249 SNPs from dbSNP build 135 in 42,733 mRNA sequences and includes linkage disequilibrium for nearby SNPs to infer associations from genome-wide association studies. It does not restrict mRNA-miRNA pairs based on expression data.

PolymiRTS (Bhattacharya, Ziebarth, & Cui, 2014) includes indels as well as SNPs in 3'UTRs from dbSNP build 137. It matches variant locations against data from TargetScan as well as from CLASH experimental data showing miRNA-mRNA pairs and binding site location. It includes biological pathway data to show the possible biological impact of changes to target sites. 22979 SNPs and 1047 indels were identified as creating, disrupting or changing target sites.

In contrast to previous studies, this work limits false positives by only working with canonical seed matches in 3'UTRs for well curated co-expressed miRNA-mRNA pairs. It considers combinations of SNPs and indels that can affect the same target site. Data from TargetScan is loaded for comparison as

well as alignments with mouse and chimpanzee genomes to show which sites are evolutionarily conserved.

4 Methods

TargetScan (Agarwal, Bell, Nam, & Bartel, 2015) is a widely used computational tool. However it requires a large sequence of the mRNA transcript to score predictions. To study combinations of genetic variants, it is computationally infeasible to run TargetScan against all possible transcripts for an mRNA. Hence, this study was limited to just full, M8 and A1 canonical matches in the seed region. This approach fails to capture other features of miRNA target site identification, such as location in the 3'UTR, target site accessibility and AU content (Grimson, et al., 2007). However, the seed match region is by far the best prediction of target-site efficacy (Agarwal, Bell, Nam, & Bartel, 2015).

Both miRNA and mRNA must be expressed within the same cell and at the same time for any interaction to occur. To ensure co-expression, miRTarBase (Huang, et al., 2020), which is a well curated source of 13,404 experimentally validated miRNA and Target interaction, was used. The comparison was limited to 5,257 interactions between 2,016 genes and 677 miRNAs reported in Homo sapiens genome.

For each of these genes, Ensembl release 104 (Howe, et al., 2021) was used to obtain coordinates for the protein coding 3' UTRs. Over 1 billion reported variants in dbSNP build 155 (dbSNP) were filtered to identify 2,006,524 genetic variants in 3' UTR, including SNPs and indels. RNA sequences for the miRNAs were taken from the miRBase (Griffiths-Jones, Grocock, van Dongen, Bateman, & Enright, 2006).

ALFA (Phan, et al., 2020) global population totals (SAMN10492705) was used to estimate the frequency of an allele. 1,118,471 of derived variants, or just over 55%, had ALFA frequency data available. Alternate Alleles with no ALFA frequency data, or ALFA frequency data of less than one percent of the population, were designated as rare alleles. Rare alleles were also evaluated. However combinations involving multiple rare alleles were not evaluated to save computational time and eliminate false positives. Reference alleles were always included, and even when rare, did not limit combinations with other alleles. Due to the availability of only population-level data, and not haplotype data from individuals, it was not certain that any two variants, even common alleles, can occur together.

hg19 (human genome chromosome-loci 19) coordinates were downloaded for TargetScan's (Agarwal, Bell, Nam, & Bartel, 2015) default conserved target prediction, and mapped to hg38 using LiftOver (Haeussler, et al., 2019). miRNA names and families were manually mapped between the names in TargetScan and miRTarBase (Huang, et al., 2020). While this data contains 126,698 predicted targets, only 2823 occurred within the mRNA – miRNA pairs in miRTarBase.

4.1 Target Frame Algorithm

Our algorithm identifies binding-sites affected by combinations of variants. A fixed length nucleotide window was positioned at each coordinate in the reference genome in the 3' UTR. A window size of eight nucleotides is required to identify canonical miRNA binding sites. At each position, all combinations of SNPs and other variants that could affect nucleotides in the window are loaded into a list. This includes variants with coordinates before the start of the window, such as if the nucleotide at the reference coordinate is part of a delete or replacement of greater than one-nucleotide length.

The number of combinations can be very large, even for a small window size. For example, the 3' UTR of the gene TRIM13 has 199 variants in dbSNP build 155 affecting a single window location of length eight. 197 of these variants have two different alleles, one has three alleles, and one has 20 alleles. This gives around 10^{61} combinations. Since some of these variants delete the same portion of the reference sequence, some of these combinations cannot occur together. However, it is not obvious

which combinations are valid. Due to the large number of potential combinations, it is infeasible to check the validity of each combination and calculate all possibilities for the eight-nucleotide window.

To build the list of possible nucleotides for each window, an empty eight-nucleotide "Frame object" is used along with a list of variants that can affect the window, sorted by the starting coordinate. Each variant at the current position is applied by creating a copy of the "Frame object" for each possible allele by filling in nucleotides from either the reference or alternate alleles, advancing the reference genome coordinates, and tracking the variants in a list for the "Frame object". Multiple variants can start at the same reference position. The reference sequence is used to fill data for locations that are not part of variant allele(s). The process is repeated until eight-nucleotides of each "Frame object" are filled, ignoring additional variants.

An exception is alternate alleles longer than the window-size. Additional nucleotides are lost when substituting alleles into the reference sequence. Separate frames are created to capture these potential targets.

Once all possible eight-nucleotide combinations are calculated for each reference position in the 3'UTR, all combinations are evaluated to see if they contained canonical miRNA targets. If no combination contains canonical targets, no data is stored for that position.

Reference Genome: TGAGATGTGCAT	Variants: Position 3: SNP Position 4: Deletion of 3 nucleotides Position 7: Insertion of 4 nucleotides	
	Frame Data	Reference Genome:
no variants at the first or second position:	*T G* ▒ ▒ ▒ ▒ ▒ ▒	AGATGTGCAT
a SNP at the third position A/T:	T G *A* ▒ ▒ ▒ ▒ ▒ T G *T* ▒ ▒ ▒ ▒ ▒	
a potential Deletion of GAT:	T G A ▒ ▒ ▒ ▒ ▒ ▒ T G A ▒ ▒ ▒ ▒ ▒ T G T ▒ ▒ ▒ ▒ ▒ T G T ▒ ▒ ▒ ▒ ▒	GATGTGCAT GTGCAT GATGTGCAT GTGCAT
no variants at the next 3 positions:	T G A *G A T* ▒ ▒ T G A ▒ ▒ ▒ ▒ ▒ T G T *G A T* ▒ ▒ T G T ▒ ▒ ▒ ▒ ▒	GTGCAT
an optional Insertion of TTTT:	T G A G A T *T T* T G A G A T ▒ ▒ T G A T *T T T T* T G A ▒ ▒ ▒ ▒ ▒ T G T G A T *T T* T G T G A T ▒ ▒ T G T T *T T T T* T G T ▒ ▒ ▒ ▒ ▒	
no more variants, fill remaining positions from reference genome:	T G A G A T T T T G A G A T *G T* T G A T T T T T T G A *G T G C A* T G T G A T T T T G T G A T *G T* T G T T T T T T T G T *G T G C A*	

Figure 2: Example of variation algorithm. Variants are applied in the order of the column on the left, completed frame data is in the middle column and the reference genome is on the right.

4.2 Evolutionary Conservation Data

To measure whether miRNA target sites we conserved evolutionarily, data from Pan troglodytes, the Chimpanzee, was used (BioProject PRJNA13184). For each of the 2,016 genes in the human genome assembly, Ensembl 104 (Howe, et al., 2021) human-chimp orthologs were downloaded and associated with 3'UTR coordinates along with the Pan tro 3.0 assembly GCF_000001515.7, originally reported in (Mikkelsen, et al., 2005). While Mus musculus (mouse) is more distant human relative than the chimpanzee, its genome is better annotated in Ensembl (Howe, et al., 2021), facilitating the identification of conserved sites between mouse and human that could not be identified for chimpanzee. Mus musculus (mouse) assembly GCF_000001635.27, GRCm39 3' UTR coordinates and human-mouse orthologs for the same 2,016 set of human genes were downloaded. 104 of the human genes did not have chimp orthologs and 46 did not have mouse orthologs.

Needleman-Wunsch algorithm was used to optimally align each human 3' UTR with its chimpanzee and mouse ortholog. The scoring system used assigned +2 for each match, -1 for each mismatch and -2 for each gap.

Since a single gene can have multiple 3' UTR annotations, each human 3' UTR was aligned to every chimp and mouse 3' UTR for the orthologous gene. The scaled score for each pair was the score of the optimal alignment divided by the average length of both UTRs in the pair. Only the best alignment was used based on a scaled score. The scaled score for each pair was the score of the optimal alignment divided by the average length of both UTRs in the pair.

5 Results and Discussion

We identified 14,438 different potential canonical targets, but a majority were only present when rare alleles were included. 3072 targets required at least one non-reference allele that was not tested in the ALFA study (Phan, et al., 2020). 2685 required an allele with ALFA population statistics, but no individuals in the ALFA data were found to have the allele. 687 further targets required an allele in less than one percent of the population. Of the remaining 7994 targets, 7933 were present in more than half the population. 6260 canonical targets were always present in the population, including 5985 that always had the same canonical target type.

Of the 2823 targets in the TargetScan data (Agarwal, Bell, Nam, & Bartel, 2015) within the miRNA-mRNA pairs in this study, we identified 2701. Of 3289 target sites always present in our population 1949 were missing in the TargetScan data. For 1048 of these target sites, we could not identify conserved sites in either the chimpanzee or mouse. Evolutionary conservation of target site sequences is an important criterion to determine if predicted target sites are biologically important, as mutation is likely to overwrite non-functional sites.

In our 3'UTR data, the human reference genome has 5,552,947 seven-nucleotide sequences. Only 387,260 of these, just under 7.5%, are not subject to variation, including rare alleles. 5,163,716, or just under 93% are not subject to variations in one percent or more of the population. If variations occurred at random locations, of our 7933 target sites in more than half the population, 553 would be unaffected by variations, and 7377 would be unaffected by variations in more than one percent of the population. Our actual results show 3526 unaffected by variations and 7843 present in more than ninety-nine percent of the population. Based on this, looking at SNPs and other variants in the same way as conservation is an additional criterion to identify target sites which will become more useful as the number of documented human variations increases.

In total, 37,999 of 2,006,524 variants in 3'UTRs, or just under two percent, had the potential to create, remove or change a target site, including rare alleles, a much smaller portion than the 64 percent found previously (Gong, et al., 2015). This is due to our focus on the seed region of canonical matches

only. While some noncanonical sites are biologically functional, most are not. Restricting our search to only known miRNA-mRNA pairs that are co-expressed also reduced the number of results.

Of these 7993 targets found in more than half the population, 3363 had a conserved canonical target in the chimpanzee reference genome, and 1486 had a conserved target in the mouse genome, including 657 we could not find in the chimpanzee genome. 4042 of the 7993 had either a mouse or a chimpanzee site. Of the 892 found as conserved in both the chimpanzee and mouse, 780 were the same canonical site type.

Including combination analysis, 26853 different alleles had potential to disrupt at least one of these 4713 targets. 16070 alleles remained after including variants without ALFA Frequency Estimates. Only 188 remained after including variants in more than once percent of the population. 10 were found in the default TargetScan data. 56 more, shown in Table 1, involve only a single variant.

Many of these 56 variants have been reported in biomedical literature, such as rs56109847 (Kapeller, et al., 2008), reported as a risk factor for irritable bowel syndrome, due to miRNA targeting. Rs4742098 is a known target linked to lung cancer (Du, et al., 2017). The T allele of rs1042538 disrupts a known miRNA target and is linked to lower breast cancer (Zheng, et al., 2011). This variant may also be an important part of human evolution. The T allele shows evidence of recent positive selection (Saunders, Liang, & Li, 2007), because it increases IQGAP1 in the brain, is linked to better cognitive performance, and this target is otherwise conserved in other primates (Yang, et al., 2014).

miRNA	Gene	Target	Conserved	RsID	Allele	Ref	%
155-5p	AGTR1	M8	None	5186	A	A	72
29a-3p	AKT2	M8	M8	41275748	A	A	90
122-5p	ATP1A2	A1	A1	78930771	C	C	97
127-5p	BLVRB	Full	Full	45460395	GA	GA	87
214-3p	CADM1	Full	None	45460594	G	G	98
138-5p	CD274	Full	Full	4742098	A	A	78
570-3p	CD274	M8	None	4143815	G	G	74
30a-3p	CDK6	Full	None	2285332	G	G	71
221-3p	CDKN1B	A1	None	34329	C	G	66
210-5p	CFB	M8	None	4151672	C	C	96
125b-5p	CSNK2A1	M8	None	157807	A	A	60
409-3p	CTNND1	M8	M8	145411496	A	A	99
143-3p	CYP2C9	Full	None	9332242	C	C	90
501-5p	DKK1	A1	None	74711339	A	A	97
125a-5p	EIF4EBP1	M8	None	1451277699	G	G	86
433-3p	FGF20	M8	None	12720208	G	G	92
152-3p	HLA-G	Full	None	1063320	G	C	50
129-5p	HMGB1	A1	A1	1042538	T	T	13
510-5p	HTR3E	Full	Full	56109847	G	G	97
603	IGF1	Full	None	6218	A	A	99
625-5p	IGF2BP1	A1	None	6504593	T	T	51
130b-3p	IRF1	A1	None	11955514	T	T	64
194-5p	KDM5B	M8	None	7541392	G	G	95
141-3p	KLF12	M8	None	7993625	C	T	71
141-3p	KLF5	M8	M8	41286082	T	T	98
410-3p	MDM2	A1	None	7956669	A	A	94
199a-3p	MET	M8	M8	1621	A	G	66
384	NTRK3	A1	A1 Mouse	62019226	A	A	96
30e-5p	P4HA1	M8	M8	76981939	A	A	68
182-5p	PDCD4	A1	None	2146544	G	G	96
93-5p	PDCD4	A1	None	56875008	T	T	99

363-3p	PDPN	M8	M8	1061615	T	T	70
372-3p	PHLPP2	M8	M8	1058747	C	C	68
373-3p	PIK3CA	A1	A1	3976507	C	C	79
509-3-5p	PODXL	M8	M8	113382424	A	A	93
133b	PPP2R2D	M8	None	7894	G	G	73
23a-3p	PPP2R5E	M8	None	3742624	A	A	90
124-3p	PRRX1	M8	None	6701640	A	C	82
146a-5p	ROCK1	M8	None	11390155	A	A	88
33a-5p	SATB2	M8	M8	150702506	TGTGTC	TGTGTC	98
96-5p	SCARB1	A1	None	838880	T	C	66
186-5p	SETD2	M8	None	79752114	G	G	96
1260b	SFRP1	M8	None	115846935	T	T	98
206	SMAD2	A1	None	1792663	A	A	50
376a-5p	SNX19	A1	None	10160281	T	T	55
222-3p	SOD2	M8	None	4555948	G	G	73
29b-3p	TCL1A	Full	None	111247694	A	A	99
663a	TGFB1	M8	M8	900617204	C	C	98
663a	TGFB1	M8	M8	1048889853	C	C	98
4273	TOMM20	Full	None	7930	T	T	93
9-5p	UHRF1	M8	M8	1343194864	A	A	85
539-3p	USP13	M8	None	3732993	G	G	84
429	ZEB1	M8	None	77992406	G	G	96

Table 1: Variants disrupting Target Sites found in more than half the population

5.1 Target Sites in less than Half the Population

Five minor alleles with ALFA frequency above 1% of the population support a new target site not present with the more common allele. Of these, the T allele of rs1057233 is strongly associated with Lupus (Hikami, et al., 2011) due to increased levels of SPI1 mRNA. This increase was previously reported (Shen, Liang, Tang, De Vries, & Tak, 2012) as being due to the possible disruption of a target site for hsa-miR-569. rs12010722 alleles CT and TT were found to be unfavorable in ovarian cancer survival (Liang, et al., 2010).

5.2 Target Sites involving Combinations of SNPs

The M8 target site for hsa-miR-641 in the gene SETD2 depends on two SNPs. It is a canonical target with the reference 'T' allele of rs1051312, which occurs 78% of the population and the alternate 'T' allele of rs3746544, which occurs in 65% of the population. The specific haplotype rs363043 C-rs3746544 T-rs1051312 C alleles has been shown to be protective against Parkinson's disease (Agliardi, et al., 2019). Rs1051312 also influences a target for miR-510. This did not appear in our results as SETD2 - miR-510 was not a miRNA-mRNA pair in the miRTarBase.

The reference T allele of rs2239680 found in 74 percent of the population and the reference T allele of rs17882139 found in 97 percent of the population create a M8 hsa-miR-335-5p target on BIRC5. The alternate 'C' of rs2239680 was found to significantly increase the risk of lung cancer (Zu, et al., 2013) through disruption of this miRNA target. rs17882139 has not yet been reported in literature, and should be investigated as a similar cancer risk.

We found a full target site for hsa-miR-432-5p on RCOR1, also conserved in the chimpanzee. The target is deleted with the alternate allele of rs1169181744 found in 1 percent of the population and can also be disrupted by the alternate 'C' allele of rs1454677527 found in three percent of the population. Neither appears in literature.

The default TargetScan data includes an M8 target for hsa-miR-491-5p on IGF2BP1. This is disrupted by either the alternate 'C' allele of rs1251726793 in four percent of the population or the alternate 'C' allele of rs1360904203 in five percent of the population.

Several other combinations of variants produced target sites in less than half the population. The reference allele C of rs8126, found in 28 percent of the population and the reference allele G of rs1052823, found in 87 percent of the population together create an 'M8' target for hsa-miR-184 on TNFAIP2 that we also found conserved in the chimpanzee. The 'T' allele of rs8126 is associated with cancer, and was investigated together with rs1052823 (Liu, et al., 2011). The 'G' variant of rs1052823, which disrupts the same target, did not reach statistical significance, but was reported independently of rs8126. We suggest investigating if the haplotype combinations of these SNPs were more predictive than rs8126 alone.

We identified many novel combinations not previously reported in biomedical literature. The alternate C allele of rs1410721908, found in 12 percent of our population, and the reference T allele of rs773888708 found in 88 percent of our population, together form an M8 target hsa-miR-625-5p for HMGA1. The alternate A allele of rs74119502, in three percent of the population, together with the reference 'C' allele of rs74121903 in 97% of our population, create an A1 target site for hsa-miR-449a on POU2F1.

5.3 Target Site Variation in the Human Reference Genome Hg38.13

Researchers need to be careful using the human reference genome for miRNA target site identification. For example, using hg38.13 assembly, we found 14,455 non-reference allele that could alter canonical target sites, although only 383 had population frequency above 1%, of which 71 were present in more than half the population.

6 Future Directions

This study focused only on 2,016 genes, just over 3% of human genes. Searching every gene for miRNA targets without evidence of miRNA-mRNA interaction would likely result in far more false positives than correct predictions. However, we intend to broaden our search for additional expression data, both from other sources and future versions of miRTarBase. Most of the potential interactions derived in this study have already been reported in the biomedical literature as the potential causes of serious diseases. This suggests that our computational predications are biologically relevant, and thus, indicate that expanding our study to cover the majority of the human genome may uncover other relevant targets.

This study was limited by the state of software tools to predict functional target sites. Not all sequences that appear to be canonical miRNA target sites are biologically relevant, nor are biologically relevant sites equally important (Grimson, et al., 2007). Many non-canonical sites may be biologically relevant which were ignored in this study. While a more comprehensive scoring system such as the context++ score of TargetScan (Agarwal, Bell, Nam, & Bartel, 2015) would have been more useful than simply matching canonical sequences, no current tool sufficiently predicts the differences between different canonical sites or the binding efficacy of noncanonical sites. In the future, we would like to reassess this study with better target prediction tools.

References

Agarwal, V., Bell, G. W., Nam, J. W., & Bartel, D. P. (2015). Predicting effective microRNA target-sites in mammalian mRNAs. *eLife*. doi:10.7554/eLife.05005

Agliardi, C., Guerini, F. R., Zanzottera, M., Riboldazzi, G., Zangaglia, R., Sturchio, A., . . . Clerici, M. (2019). SNAP25 Gene Polymorphisms Protect Against Parkinson's Disease and Modulate Disease Severity in Patients. *Mol Neurobiol 56*, 4455–63. doi:0.1007/s12035-018-1386-0

Ahmadi, A., & Moradi, S. (2021). In silico analysis suggests the RNAi-enhancing antibiotic enoxacin as a potential inhibitor of SARS-CoV-2 infection. *Sci Rep 11, 10271*. doi:10.1038/s41598-021-89605-6

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell, 116*(2), 281-97. doi:10.1016/S0092-8674(04)00045-5.

Bartel, D. P. (2009). MicroRNAs: Target Recognition and Regulatory Functions. *Cell, 136(2)*, 215-233. doi:10.1016/j.cell.2009.01.002

Bartel, D. P. (2018). Metazoan MicroRNAs. *Cell, 173*(1), 20-51. doi:10.1016/j.cell.2018.03.006

Betel, D., Wilson, M., Gabow, A., Marks, D. S., & Sander, C. (2008). The microRNA. org resource: targets and expression. *Nucleic acids research 36.suppl_1*, D149-D153. doi:10.1093/nar/gkm995

Bhattacharya, A., Ziebarth, J., & Cui, Y. (2014). PolymiRTS Database 3.0: linking polymorphisms in microRNAs and their target sites with human diseases and biological pathways. *Nucleic Acids Research, Volume 42, Issue D1*, D86–D91. doi:10.1093/nar/gkt1028

Calin, G., & Croce, C. (2006). MicroRNA signatures in human cancers. *Nat Rev Cancer, 6*, 857-66. doi:10.1038/nrc1997

dbSNP. (retrieved 2021, Nov). *Database of Single Nucleotide Polymorphisms (dbSNP)*. Retrieved from National Center for Biotechnology Information: http://www.ncbi.nlm.nih.gov/SNP/

Du, W., Zhu, J., Chen, Y., Zeng, Y., Shen, D., Zhang, N., . . . Huang, J. A. (2017). Variant SNPs at the microRNA complementary site in the B7-H1 3'-untranslated region increase the risk of non-small cell lung cancer. *Mol Med Rep*, 2682-90. doi:10.3892/mmr.2017.6902

Enright, A., John, B., Gaul, U., Tuschl, T., Sander, C., & Marks, D. (2003). MicroRNA targets in Drosophila. *Genome biology, 4*(11), 1-27.

Farh, K. K., Grimson, A., Jan, C., Lewis, B. P., Johnston, W. K., Lim, L. P., . . . Bartel, D. P. (2005). The Widespread Impact of Mammalian MicroRNAs on mRNA Repression and Evolution. *Science, 410(5755)*, 1817-21. doi:10.1126/science.1121158

Gong, J., Liu, C., Liu, W., Wu, Y., Ma, Z., Chen, H., & Guo, A. Y. (2015). An Update of miRNASNP Database for Better SNP Selection by GWAS Data, miRNA Expression and Online Tools. *Database*. doi::10.1093/database/bav029

Gong, J., Tong, Y., Zhang, H. M., Wang, K., Hu, T., Shan, G., . . . Guo, A. Y. (2012). Genome Wide Identification of SNPs in microRNA Genes and the SNP Effects on microRNA Target Binding and Biogenesis. *Human Mutation, 33(1)*, 254-263.

Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A., & Enright, A. J. (2006). miRBase: microRNA Sequences, Targets and Gene Nomenclature. *Nucleic Acids Research, 34, Database issue*, D140-D144. doi:10.1093/nar/gkj112

Grimson, A., Farh, K. K., Johnston, W. K., Garrett-Engele, P., Lim, L. P., & Bartel, D. P. (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell*, 91-105. doi:10.1016/j.molcel.2007.06.017

Haeussler, M., Zweig, A. S., Tyner, C., Speir, M. L., Rosenbloom, K. R., Raney, B. J., . . . Gibson, D. (2019). The UCSC Genome Browser database: 2019 update. *Nucleic Acids Research, 47*(D1), D853-8. doi:10.1093/nar/gky1095

Hikami, K., Kawasaki, A., Ito, I., Koga, M., Ito, S., Hayashi, T., . . . Hashimoto, H. (2011). Association of a functional polymorphism in the 3′-untranslated region of SPI1 with systemic lupus erythematosus. *Arthritis Rheum, 63*(3), 755–763. doi:10.1002/art.30188

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences, 106*(23), 9362-7.

Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., . . . Billis, K. (2021). Ensembl 2021. *Nucleic Acids Research, 49, Database issue*, D884-D891. doi:10.1093/nar/gkaa942

Huang, H. Y., Lin, Y. C., Li, J., Huang, K. Y., Shrestha, S., Hong, H. C., . . . Xu, J. T. (2020). miRTarBase 2020: Updates to the Experimentally Validated microRNA–target Interaction Database. *Nucleic Acids Research, 48(D1)*, D148-D154. doi:10.1093/nar/gkz896

Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., . . . Liu, Y. (2009). miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res., 37(suppl_1)*, D98-D104. doi:10.1093/nar/gkn714

Kapeller, J., Houghton, L. A., Mönnikes, H., Walstab, J., Möller, D., Bönisch, H., . . . Gassler, N. (2008). First evidence for an association of a functional variant in the microRNA-510 target site of the serotonin receptor-type 3E gene with diarrhea predominant irritable bowel syndrome. *Hum Mol Genet*, 2967-77. doi: 10.1093/hmg/ddn195

Karagkouni, D., Paraskevopoulou, M. D., Chatzopoulos, S., Vlachos, I. S., Tastsoglou, S., Kanellos, I., . . . Vergoulis, T. (2018). DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA-gene interactions. *Nucleic Acids Res., 46*, D239-D245.

Lewis, B. P., Shih, I. H., Jones-Rhoades, M. W., Bartel, D. P., & Burge, C. B. (2003, Dec). Prediction of Mammalian MicroRNA Targets. *Cell, 115(7)*, 787-98.

Li, J. H., Liu, S., Zhou, H., Qu, L. H., & Yang, J. H. (2014). "starBase v2.0: decoding miRNA–ceRNA, miRNA–ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Research, 42*, D92-D97. doi:10.1093/nar/gkt1248

Liang, D., Meyer, L., Chang, D. W., Lin, J., Pu, X., Ye, Y., . . . Lu, K. (2010). Genetic Variants in MicroRNA Biosynthesis Pathways and Binding Sites Modify Ovarian Cancer Risk, Survival, and Treatment Response. *Cancer Res*, 9765-76. doi:10.1158/0008-5472.CAN-10-0130

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.

Liu, Z., Wei, S., Ma, H., Zhao, M., Myers, J. N., Weber, R. S., . . . Wei, Q. (2011). A functional variant at the miR-184 binding site in TNFAIP2 and risk of squamous cell carcinoma of the head and neck. *Carcinogenesis*, 1668-74. doi:10.1093/carcin/bgr209

Mikkelsen, T., Hillier, L., Eichler, E., Zody, M., Jaffe, D., Yang, S. P., . . . Archidiacono, N. (2005). Initial Sequence of the Chimpanzee Genome and Comparison with the Human Genome. *Nature, 437(7055)*, 69–87. doi:10.1038/nature04072

Momen-Heravi, F., & Bala, S. (2018). miRNA regulation of innate immunity. *J Leukoc. Biol, 103*, 1205-17. doi:10.1002/JLB.3MIR1117-459R

Phan, L., Jin, Y., Zhang, H., Qiang, W., Shekhtman, E., Shao, D., . . . Wang, Z. Y. (2020, March 10). *ALFA: Allele Frequency Aggregator*. Retrieved from National Center for Biotechnology Information, U.S. National Library of Medicine: www.ncbi.nlm.nih.gov/snp/docs/gsr/alfa/

Ryan, B. M., Robles, A. I., & Harris, C. C. (2010, June). Genetic Variation in microRNA Networks: The Implications for Cancer Research. *Nat. Rev. Cancer, 10(6)*, 389–402. doi:10.1038/nrc2867

Saunders, M. A., Liang, H., & Li, W. H. (2007). Human polymorphism at microRNAs and microRNA target sites. *Proc Natl Acad Sci*, 3300-5. doi:10.1073/pnas.0611347104

Shen, N., Liang, D., Tang, Y., De Vries, N., & Tak, P. P. (2012). MicroRNAs—novel regulators of systemic lupus erythematosus pathogenesis. *Nature Reviews Rheumatology*, *8*(12), 701-9.

Sun, G., Yan, J., Noltner, K., Feng, J., Li, H., Sarkis, D. A., . . . Rossi, J. J. (2009). SNPs in human miRNA genes affect biogenesis and function. *RNA*, *15*, 1640-51.

Xiao, F., Zuo, Z., Cai, G., Kang, S., Gao, X., & Li, T. (2009). miRecords: an integrated resource for microRNA–target interactions. *Nucleic Acids Research, 37(suppl_1),*, D105–10. doi:10.1093/nar/gkn851

Yang, L., Zhang, R., Li, M., Wu, X., Wang, J., Huang, L., . . . Su, B. (2014). A functional MiR-124 binding-site polymorphism in IQGAP1 affects human cognitive performance. *PLoS One*, *9*(9). doi:10.1371/journal.pone.0107065

Yuan, C., Burns, M. B., Subramanian, S., & Blekhman, R. (2018, May). Interaction between host microRNAs and the gut microbiota in colorectal cancer. *Systems. 3(3)*, e00205–17.

Zenz, T., Haebe, S., Denzel, T., Mohr, J., Winkler, D., Bühler, A., . . . Hallek, M. (2009). Detailed analysis of p53 pathway defects in fludarabine-refractory chronic lymphocytic leukemia (CLL): dissecting the contribution of 17p deletion, TP53 mutation, p53-p21 dysfunction, and miR34a in a prospective clinical trial. *Blood, 97(114)*, 2589–97.

Zheng, H., Song, F., Zhang, L., Yang, D., Ji, P., Wang, Y., . . . Zhang, W. (2011). Genetic variants at the miR-124 binding site on the cytoskeleton-organizing IQGAP1 gene confer differential predisposition to breast cancer. *Int J Oncol 38*, 1153-61. doi:10.3892/ijo.2011.940

Zu, Y., Ban, J., Xia, Z., Wang, J., Cai, Y., Ping, W., & Sun, W. (2013). Genetic variation in a miR-335 binding site in BIRC5 alters susceptibility to lung cancer in Chinese Han populations. *Biochem Biophys Res Commun.*, 529-34. doi:10.1016/j.bbrc.2012.12.001