bioRxiv preprint doi: https://doi.org/10.1101/420646; this version posted September 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Pervasive selection against microRNA target sites in human populations

Andrea Hatlen and Antonio Marco*

School of Biological Sciences, University of Essex, Colchester, United Kingdom

Short title: Selection against microRNA target sites

* To whom correspondence should be addressed

School of Biological Sciences

University of Essex

Wivenhoe Park, Colchester CO4 3SQ

United Kingdom

Email: amarco.bio@gmail.com

Telephone: +44 (0) 120 687 3339

bioRxiv preprint doi: https://doi.org/10.1101/420646; this version posted September 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

ABSTRACT

MicroRNA target sites are often conserved among related species. On the other hand, purifying selection against novel microRNA target sites can also exist. However, the relative impact of conservation versus selection against target sites is still unknown. We investigated these processes in human populations by focusing on polymorphic sites in which one allele is a target site and the other is not. Target allele frequencies were significantly lower than expected at these sites. The analysis of derived allele frequencies revealed that, when the non-target allele is ancestral, the proportion of non-target sites is higher than expected by chance. Conversely, when the target allele is ancestral, the proportion of non-target alleles is also significantly higher than expected. These analyses reveal a selective pressure against microRNA target alleles, which is more effective than selection to conserve target sites. Additionally, microRNA target sites show relatively low levels of population differentiation (Fst). However, when we analyse separately target sites in which the target allele is ancestral in the population, the proportion of SNPs with high Fst significantly increases. These findings support a scenario in which population differentiation (and possible local adaptation) is much more likely in target sites that are lost than in the gain of new target sites. Taking all the results together, we conclude that there is evidence of pervasive selection against microRNA target sites in human populations. The overall impact across untranslated regions is not negligible and should be taken into account when studying the evolution of genomic sequences.

Keywords: Purifying selection; miRNA; human populations; target sites.

bioRxiv preprint doi: https://doi.org/10.1101/420646; this version posted September 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

INTRODUCTION

MicroRNAs are small endogenous RNAs that can regulate virtually any type of biological process. In humans they were not discovered until this century [1] and there are now about 1900 human microRNA precursors annotated in miRBase [2], although less than 700 are classified with high confidence. Soon after microRNAs were found in multiple animal species [3–5], the first target prediction tools became available [6–8]. Only in the last few years have these developments permitted the evolutionary analysis of target sites [9–13] revealing that many microRNA target sites are highly conserved among species. On the other hand, whilst some microRNA families have been conserved for millions of years, their targets appear to differ between species (see for instance [14]). Indeed, evidence from vertebrates suggests that gains and losses of target sites may be more important than changes in the microRNAs themselves during the evolution of microRNA-based gene regulation, as microRNAs are usually highly conserved (see Discussion in [15]).

Several studies have found that gene transcripts are depleted of target sites for co-expressed microRNAs [10–12]. In particular long 3' UTR might accumulate microRNA target sites by random mutation, yet they actually have a lower frequency than expected by chance, suggesting that there has been selection against these sequences [10]. These missing sites have been called 'anti-targets' [16,10]. Interestingly, target sites for the same microRNAs tend to be conserved in transcripts expressed in neighboring tissues [11]. These studies have shown that selection against microRNA target sites can be inferred from comparisons among distantly related species. However, the relative impact of selection against microRNA target sites in human populations is not known.

Analysis of human populations has suggested purifying selection was particularly strong at microRNA target sites, even in non-conserved sites [17,18]. It was also detected negative selection against gaining microRNA target sites in Yoruban populations, but the pattern was not detected in other populations [17]. In a study of *Drosophila* populations we found evidence of selection against

bioRxiv preprint doi: https://doi.org/10.1101/420646; this version posted September 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

microRNA target sites [19]. Specifically, we found selection against target sites of maternal microRNAs in maternally deposited transcripts. More recently is has been shown that this effect is particularly strong for the *mir-309* cluster, whose microRNAs are abundant in the egg and almost absent in the zygote [20]. Characterising this type of selection in humans would reveal to which extent it shapes our genomes. However, the strength and prevalence of selection against target sites is human populations is still unknown. Here we investigate polymorphisms at human microRNA target sites and quantify the impact of selection.

bioRxiv preprint doi: https://doi.org/10.1101/420646; this version posted September 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

RESULTS

Bias towards microRNA non-target alleles in human populations

In order to investigate the selective pressures on microRNA target sites in human populations, we first mapped human single-nucleotide polymorphisms (SNPs) to putative canonical microRNA target sites such that one allele is a target site and the alternative allele is not a target site (see Methods for details). The non-target allele in this pair is called a 'near-target' [21]. We compiled 709,854 polymorphic target sites in 42,221 gene transcripts (15,026 genes) for 2,584 microRNAs. We first compared the allele frequency distribution of target sites for broadly expressed microRNAs with a background distribution obtained by conducting the same analysis on the reverse complement sequences of 3'UTRs (see Methods). The distribution is biased to the near-target allele: there are more non-target sequences than background at low population frequencies, and fewer at high frequencies (one-tailed Mann-Whitney test, $p \ll 0.001$; Figure 1A). Selection against the target sites might be expected in cases where the microRNA/transcript pairs are co-expressed and therefore more likely to interact. We therefore compared these pairs with those that are co-expressed and found a significant difference (Figure 1B; $p=0.029$).

To further investigate the selective forces operating at microRNA targets sites we computed the derived allele frequency (DAF, also known as the unfolded site frequency spectrum). We initially considered those SNPs for which the derived allele (the non-conserved allele) is the target allele (Figure 1C). As expected the overall DAF distribution is L-shaped. However, the distribution of target sites for broadly expressed microRNAs is significantly skewed to the near-target allele compared to the background distribution, indicating selective pressure against the newly arisen target allele ($p=0.007$). Strikingly, when we explored SNPs for which the derived allele is the near-target allele, the distribution is also skewed towards the near-target allele (Figure 1D; $p=0.022$). In summary, the distribution of allele frequencies shows evidence of selection against microRNA target sites in human populations.

bioRxiv preprint doi: https://doi.org/10.1101/420646; this version posted September 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The effect of microRNA expression levels and evolutionary conservation

We next considered the potential impact of microRNA conservation. On the one hand, evolutionarily conserved microRNAs may have a weaker effect on selection against microRNA targets, as partly deleterious target alleles may have been cleared from the population. On the other hand, evolutionarily conserved microRNAs tend to be highly expressed, and therefore it is expected that the selective pressure to avoid targets for such microRNAs should be stronger. In other words, expression and conservation are not independent to each other. We therefore included both factors in the analysis: the level of expression of the microRNA and a measure of the phylogenetic conservation of the microRNA sequence.

We analysed cases where the derived allele is a target site (as in Figure 1C); testing whether frequency spectra were different across different levels of microRNA conservation (human-primate specific, conserved in mammals, and conserved in animals) and different levels of expression (low, mid and high as described in Methods). For all 81 possible pairwise comparisons we performed a one-tailed Mann-Whitney test and displayed the resulting p-values in a heatmap in Figure 2. The light grey tones indicate lower frequencies of the target allele, suggesting selection against the derived microRNA target sites. This graph allows a quick visual exploration of the results from many tests. First, it can be seen that lower-left colors are lighter than top-right. That is, for conserved microRNAs, the target allele frequencies are smaller than for less conserved microRNAs. At the same time, within each conservation level (squares in the diagonal in Figure 2) the colors are lighter in the top-right sections, indicating that for higher expressed microRNAs the target allele frequencies are lower than for less expressed microRNAs. In other words, selective avoidance against target sites is higher for conserved and higher for highly expressed microRNAs.

bioRxiv preprint doi: https://doi.org/10.1101/420646; this version posted September 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

To evaluate the joint effect of conservation and expression, we performed a two-way non-parametric Sheirer-Ray-Hare test ([22], pp.445). The effect of microRNA expression level in target allele frequencies was significant (p=0.00001) whilst the effect of conservation and the interaction between conservation and expression was not significant (p=0.20100 and p=0.79595 respectively). This analysis therefore supports one of the two trends suggested by inspection of Figure 2: the depressed frequency of target sites is significantly stronger for more highly expressed microRNAs, but there is insufficient evidence to establish a difference between microRNAs that have been strongly conserved among species and the others.

Population differentiation at target sites

If a microRNA target site is under selective constraints, we should expect differentiation among populations at these sites to be relatively low. To investigate this prediction, we grouped SNPs at target sites for broadly expressed microRNAs according to their Fst and compared the relative frequency of these SNPs compared to the background (see Methods). In Figure 3A we observed that, overall, SNPs at microRNA target sites are depleted for high Fst values. This depression of Fst at microRNA target sites was not detected in the analysis of another group [23] (see Discussion). We further explored the distribution of Fst separately depending on the inferred state of the ancestral allele (target versus near-target). The most striking observation was that, incontrast to the general trend, when the conserved microRNA target site is segregating in human populations (i.e. the target allele was ancestral), there can be considerable variation in more frequency among human populations than in the background loci: in some populations the non-target allele is more frequent than the ancestral allele. The same result appears when we expanded the dataset to microRNAs that are expressed in specific tissues (Figure 3C, red line). In Table 1 we show SNPs in microRNA target sites with a Fst greater than 0.6. In most cases, the ancestral target is lost in populations outside Africa. Interestingly, for HM13 (Minor histocompatibility antigen H13) two target sites for miR-150-5p are lost simultaneously (these sites are about 1kb away to each other).

bioRxiv preprint doi: https://doi.org/10.1101/420646; this version posted September 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

On the other hand, in cases where the ancestral allele was the near-target, there is a deficit of high Fst SNPs compared to the background sites(Figure 3B-C, blue lines). In these data, we did not find cases where the derived target allele varied atypically among populations, nor cases where the derived allele reached a very high frequency in a subset of human populations.

Co-evolution of microRNA and target sites: a potential case study

We found one case in which a microRNA has a SNP within its seed sequence (the region that determines the targeting property of the microRNA), which shows some evidence of population differentiation (rs7210937; Fst = 0.3314). In this case, the Fst value between African and European populations is remarkably high (Fst = 0.6129; Nei's estimate [24,25]). In European populations 92.5% of sequenced individuals present the ancestral form of miR-1269b, whilst in African population, the derived version is more frequent (59.8%). As a shift in the seed sequence may have an impact on the evolution of 3'UTRs, we further studied target sites whose ancestral form is a target for the derived miR-1269 microRNA (786 in total). Then we compared the frequency of the target site allele between European and African populations with the Wilcoxon non-parametric test for paired samples. We found that in African populations the frequency of target alleles is lower than in European populations at these sites ($p<<0.001$) whilst for the ancestral miR-1269b we did not find any significant difference. These results suggest that a shift in the allele frequencies affecting the seed sequence of a microRNA can have an effect on the allele frequencies at the novel target sites, specifically toward the non-target allele.

bioRxiv preprint doi: https://doi.org/10.1101/420646; this version posted September 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

DISCUSSION

The study of allele frequencies has been extensively used to detect selective pressures in human populations [17,26,27]. Here we show that the patterns of allele frequencies at 3'UTRs show evidence of selection against most microRNA target sites. First, the allele frequencies at target sites are biased towards the non-target allele. Second, this trend is also seen in the subset of cases where the derived allele is a target sequence. These effects are strongest in the cases where the corresponding microRNAs are highly expressed, suggesting that interaction between the microRNA and the target is a key is the source of selection against target sequences. The microRNAs that have been conserved over longer periods of vertebrate evolution did not impose detectably greater selection against their target sequences, once the effect of expression levels had been taken into account.

The most popular microRNA target prediction programs rely on target site conservation to reduce the number of false positives [28] and/or do not provide a stand-alone version to run on custom datasets [29]. Therefore, we used a naive microRNA target prediction method that reports canonical targets and near-target sites [21]. That allowed us to study pairs of alleles segregating at target sites without any other constraint. On the other hand, we would expect a high number of false positive in target predictions (reviewed in [30]). Remarkably, we found a significant pattern of selection against microRNA target sites. This reinforces our initial hypothesis and suggest that, if we would be able to restrict the analysis to bona fide target sites, the signal might be stronger. One possibility is to evaluate experimentally validated target sites. However, these experiments are based on reference genomes, so segregating target sites whose target allele is not in the reference genome will be lost from the analysis. The way forward may be to perform high-thoughtput microRNA target experiments, like HITS-CLIP [31], in cells derived from different populations. The continuous drop in the costs of sequencing and high-throughput experiments may allow this in the near future. Indeed, high-throughput experimental evaluation of segregating alleles at regulatory

bioRxiv preprint doi: https://doi.org/10.1101/420646; this version posted September 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

motifs (transcription factor binding sites, RNA binding sites, etcetera) is a promising area of research which will help us to move from a typological (reference genome) to a population view of gene regulation.

Another way to study the effect of selection in populations is to evaluate the population differentiation [23,32]. We found that, at microRNA target sites in general, populations tend to be less differentiated than at control background sites, suggesting that selection acts against an increase in their frequencies. This result was different to that reported by Li et al. [23]. This may be explained by our finding that a subset of target sites do have relatively high Fst values. We found this trend at those sites at which the ancestral allele was a microRNA target site and the derived allele is a non-target. The loss of a microRNA target (as in the examples reported in a previous work [23]) may be relatively frequent. It follows that the non-target allele might be neutral or even advantageous in some of these cases. It is noticeable that in the examples in which the derived allele reaches a high frequency, that occurs in the non-African populations, which is the pattern that would be expected if a neutral derived allele spread by genetic drift during founder events. Loss of target sites could also have advantageous effects, though the complex interactions that occur in regulatory networks, for example, it has been proposed that in the human lineage the loss of microRNA target sites contributed to an increase in the expression levels of some genes [33]. Our work suggest that this loss of targets may be continuing now in human populations. Selection in favour of new target sites appears to be rarer: we found a strong signal of purifying selection against novel microRNA target sites.

Selection against deleterious mutations has been extensively studied in population genetics (reviewed in [34]). For instance, strong purifying selection produces a phenomenon called background selection, in which loci linked to the selected site experience a reduction in their effective population size [35]. That is, purifying selection reduces the influence of selection at

bioRxiv preprint doi: https://doi.org/10.1101/420646; this version posted September 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

linked sites. For weakly selected sites, a similar process has been described: weak selection Hill-Robertson interference (wsHR [36,37]). Under wsHR, multiple alleles are under a weak selective force, very close together so that recombination is small or negligible between sites, interfering with other selective pressures in the area. We believe that this is the case for the selection against microRNA target sites here described: weak selection against multiple target/near-target sites will shape the evolutionary landscape of the entire untranslated region.

Most analyses of genomic sequence assume that the protein coding sequences are the most important sites of selection. Our results suggests that 3' UTRs should be included as well. New target sites will emerge at a significant rate because many mutations can potentially introduce a new site for one of the many microRNAs. More specifically, there are about 2,000 microRNA families described in TargetScan (see Methods), defined by 7-nucleotide seed sequences. Assuming that 3'UTRs are composed of non-overlapping 7-mers (a simplifying yet conservative assumption) the expected number of near-target sites per kilobase (kb) is about 75. With a mutation rate of $2.5\ 10^{-5}$ per kb [38] and a total length of the genome that encode 3'UTRs of 34Mb, it can be shown that there will be on average one novel microRNA target site on a 3'UTR per genome per generation. That is, one potential deleterious microRNA target site per person per generation.

It is expected that other regulatory motifs influence the evolution of 3'UTRs. For instance, Savisaar et al [39] have described selection against RNA-binding motifs. The selective avoidance of transcription factor binding sites [40,41] and of mRNA/ncRNA regulatory interactions in bacteria [42] have been also described. It is likely that on top of all the selective forces that are usually taken into account, there is a layer of selection against weakly deleterious regulatory motifs that will be influencing the evolution of the genome. In conclusion, selection against microRNA target sites in prevalent in human populations, and it may constrain other selective forces in post-transcriptional regulatory regions.

bioRxiv preprint doi: https://doi.org/10.1101/420646; this version posted September 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

MATERIALS AND METHODS

MicroRNA target and near-target sites were predicted with seedVicious (v.1.1 [21]]) against 3'UTR as annotated in Ensembl version 91 [43] for the human genome assembly hg38. Single nucleotide polymorphisms for the 1000 Genomes project [44] were retrieved from dbSNP (build 137) [45] and mapped to our target predictions. Ensembl sequences and polymorphism data were downloaded using the BiomaRt R package [46]. When plotting allele frequencies (Figure 1) we only considered segregating alleles in which the minor allele is present in at least 1% of the sampled population, but when using all segregating alleles we obtained comparable results. The ancestral allele status was obtained from dbSNP. To compute the background (randomly expected) allele distributions we repeated the process but finding targets in the reverse complement strand of the 3'UTR, to control for sequence length and composition. Expression information for microRNAs was obtained from Meunier et al. [47] and from miRMine [48], and for gene transcripts from the Bgee database (version 13.2, [49]), considering the following tissues: lung, blood, placenta, liver, heart, brain, kidney and testis. MicroRNAs with more than 50 RPM (reads per million) across all tissues were considered broadly expressed. In Figure 2, 'high', 'mid', and 'low' refer to microRNA with more than 500 RPM, between 500 and 50 RPM, and less than 50RPM in any tissue respectively. In Figure 3C, 'broadly and non-broadly expressed' microRNAs were defined as microRNAs with at least 50 RPKM on any tissue. MicroRNAs were grouped into evolutionary conservation categories depending on the species spread of the seed family in TargetScan 7.2 [28]. Fst values were retrieved from the 1,000 Genomes Selection Browser 1.0 [50]. In figure 3 the fold enrichment was computed as the logarithm of the ratio between the proportion of SNPs at target sites and the proportion of SNPs at a background site for each Fst bin. All statistical analyses were done with R (v. 3.4.3, [51]). All processed datasets and online tools to compute the tests here reported are available at our dedicated web server PopTargs (Hatlen and Marco, under review).

bioRxiv preprint doi: https://doi.org/10.1101/420646; this version posted September 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

ACKNOWLEDGEMENTS

We thank Richard Nichols for critical reading of the manuscript and useful comments. This work was supported by the Wellcome Trust [grant number 200585/Z/16/Z].

bioRxiv preprint doi: https://doi.org/10.1101/420646; this version posted September 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

REFERENCES

1. Pasquinelli AE, Reinhart BJ, Slack F, Martindale MQ, Kuroda MI, Maller B, et al. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. Nature. 2000;408: 86–89. doi:10.1038/35040556

2. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. 2014;42: D68-73. doi:10.1093/nar/gkt1181

3. Lee RC, Ambros V. An extensive class of small RNAs in Caenorhabditis elegans. Science. 2001;294: 862–864. doi:10.1126/science.1065329

4. Lau NC, Lim LP, Weinstein EG, Bartel DP. An Abundant Class of Tiny RNAs with Probable Regulatory Roles in Caenorhabditis elegans. Science. 2001;294: 858–862. doi:10.1126/science.1065062

5. Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T. Identification of Novel Genes Coding for Small Expressed RNAs. Science. 2001;294: 853–858. doi:10.1126/science.1064921

6. Stark A, Brennecke J, Russell RB, Cohen SM. Identification of Drosophila MicroRNA targets. PLoS Biol. 2003;1: E60. doi:10.1371/journal.pbio.0000060

7. Enright A, John B, Gaul U, Tuschl T, Sander C, Marks D. MicroRNA targets in Drosophila. Genome Biol. 2003;5: R1. doi:10.1186/gb-2003-5-1-r1

8. Lewis BP, Shih I, Jones-Rhoades MW, Bartel DP, Burge CB. Prediction of mammalian microRNA targets. Cell. 2003;115: 787–798.

9. Grün D, Wang Y-L, Langenberger D, Gunsalus KC, Rajewsky N. microRNA Target Predictions across Seven Drosophila Species and Comparison to Mammalian Targets. PLOS Comput Biol. 2005;1: e13. doi:10.1371/journal.pcbi.0010013

10. Farh KK-H, Grimson A, Jan C, Lewis BP, Johnston WK, Lim LP, et al. The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. Science. 2005;310: 1817–1821. doi:10.1126/science.1121158

11. Stark A, Brennecke J, Bushati N, Russell RB, Cohen SM. Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. Cell. 2005;123: 1133–1146. doi:10.1016/j.cell.2005.11.023

12. Sood P, Krek A, Zavolan M, Macino G, Rajewsky N. Cell-type-specific signatures of microRNAs on target mRNA expression. Proc Natl Acad Sci U S A. 2006;103: 2746–2751. doi:10.1073/pnas.0511045103

13. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell. 2005;120: 15–20. doi:10.1016/j.cell.2004.12.035

14. Hui JHL, Marco A, Hunt S, Melling J, Griffiths-Jones S, Ronshaugen M. Structure, evolution and function of the bi-directionally transcribed iab-4/iab-8 microRNA locus in arthropods. Nucleic Acids Res. 2013;41: 3352–3361. doi:10.1093/nar/gks1445

15. Marco A. No evidence of functional co-adaptation between clustered microRNAs. bioRxiv. 2018; 274811. doi:10.1101/274811

bioRxiv preprint doi: https://doi.org/10.1101/420646; this version posted September 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

16. Bartel DP, Chen C-Z. Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs. Nat Rev Genet. 2004;5: 396–400. doi:10.1038/nrg1328

17. Chen K, Rajewsky N. Natural selection on human microRNA binding sites inferred from SNP data. Nat Genet. 2006;38: 1452–1456. doi:10.1038/ng1910

18. Saunders MA, Liang H, Li W-H. Human polymorphism at microRNAs and microRNA target sites. Proc Natl Acad Sci U S A. 2007;104: 3300–3305. doi:10.1073/pnas.0611347104

19. Marco A. Selection Against Maternal microRNA Target Sites in Maternal Transcripts. G3 GenesGenomesGenetics. 2015; g3.115.019497. doi:10.1534/g3.115.019497

20. Zhou L, Lim MYT, Kaur P, Saj A, Bortolamiol-Becet D, Gopal V, et al. Importance of miRNA stability and alternative primary miRNA isoforms in gene regulation during Drosophila development. In: eLife [Internet]. 19 Jul 2018 [cited 9 Aug 2018]. doi:10.7554/eLife.38389

21. Marco A. SeedVicious: Analysis of microRNA target and near-target sites. PLOS ONE. 2018;13: e0195532. doi:10.1371/journal.pone.0195532

22. Sokal RR, Rohlf FJ. Biometry: The Principles and Practice of Statistics in Biological Research. Freeman; 1995.

23. Li J, Liu Y, Xin X, Kim TS, Cabeza EA, Ren J, et al. Evidence for Positive Selection on a Number of MicroRNA Regulatory Interactions during Recent Human Evolution. PLoS Genet. 2012;8: e1002578. doi:10.1371/journal.pgen.1002578

24. Nei M. Definition and estimation of fixation indices. Evol Int J Org Evol. 1986;40: 643–645. doi:10.1111/j.1558-5646.1986.tb00516.x

25. Bhatia G, Patterson N, Sankararaman S, Price AL. Estimating and interpreting FST: The impact of rare variants. Genome Res. 2013;23: 1514–1521. doi:10.1101/gr.154831.113

26. Enard D, Messer PW, Petrov DA. Genome-wide signals of positive selection in human evolution. Genome Res. 2014;24: 885–895. doi:10.1101/gr.164822.113

27. Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. Natural selection has driven population differentiation in modern humans. Nat Genet. 2008;40: 340–345. doi:10.1038/ng.78

28. Agarwal V, Bell GW, Nam J-W, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. eLife. 2015;4. doi:10.7554/eLife.05005

29. Paraskevopoulou MD, Georgakilas G, Kostoulas N, Vlachos IS, Vergoulis T, Reczko M, et al. DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. Nucleic Acids Res. 2013;41: W169–W173. doi:10.1093/nar/gkt393

30. Alexiou P, Maragkakis M, Papadopoulos GL, Reczko M, Hatzigeorgiou AG. Lost in translation: an assessment and perspective for computational microRNA target identification. Bioinformatics. 2009;25: 3049–3055. doi:10.1093/bioinformatics/btp565

31. Chi SW, Zang JB, Mele A, Darnell RB. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. Nature. 2009;460: 479–486. doi:10.1038/nature08170

bioRxiv preprint doi: https://doi.org/10.1101/420646; this version posted September 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

32. Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, Absher D, et al. The Role of Geography in Human Adaptation. PLOS Genet. 2009;5: e1000500. doi:10.1371/journal.pgen.1000500

33. Gardner PP, Vinther J. Mutation of miRNA target sequences during human evolution. Trends Genet TIG. 2008;24: 262–265. doi:10.1016/j.tig.2008.03.009

34. Charlesworth B. The Effects of Deleterious Mutations on Evolution at Linked Sites. Genetics. 2012;190: 5–22. doi:10.1534/genetics.111.134288

35. Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics. 1993;134: 1289–1303.

36. McVean GAT, Charlesworth B. The Effects of Hill-Robertson Interference Between Weakly Selected Mutations on Patterns of Molecular Evolution and Variation. Genetics. 2000;155: 929–944.

37. Hill WG, Robertson A. The effect of linkage on limits to artificial selection. Genet Res. 1966;8: 269–294.

38. Nachman MW, Crowell SL. Estimate of the Mutation Rate per Nucleotide in Humans. Genetics. 2000;156: 297–304.

39. Savisaar R, Hurst LD. Both Maintenance and Avoidance of RNA-Binding Protein Interactions Constrain Coding Sequence Evolution. Mol Biol Evol. 2017;34: 1110–1126. doi:10.1093/molbev/msx061

40. Babbitt GA. Relaxed selection against accidental binding of transcription factors with conserved chromatin contexts. Gene. 2010;466: 43–48. doi:10.1016/j.gene.2010.07.002

41. Hahn M, Stajich J, Wray G. The effects of selection against spurious transcription factor binding sites. Mol Biol Evol. 2003;20: 906, 901.

42. Umu SU, Poole AM, Dobson RC, Gardner PP. Avoidance of stochastic RNA interactions can be harnessed to control protein expression levels in bacteria and archaea. In: eLife [Internet]. 20 Sep 2016 [cited 6 Aug 2018]. doi:10.7554/eLife.13479

43. Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, et al. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database J Biol Databases Curation. 2011;2011: bar030. doi:10.1093/database/bar030

44. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526: 68–74. doi:10.1038/nature15393

45. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29: 308–311.

46. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc. 2009;4: 1184. doi:10.1038/nprot.2009.97

47. Meunier J, Lemoine F, Soumillon M, Liechti A, Weier M, Guschanski K, et al. Birth and expression evolution of mammalian microRNA genes. Genome Res. 2013;23: 34–45. doi:10.1101/gr.140269.112

bioRxiv preprint doi: https://doi.org/10.1101/420646; this version posted September 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

48. Panwar B, Omenn GS, Guan Y. miRmine: a database of human miRNA expression profiles. Bioinformatics. 2017;33: 1554–1560. doi:10.1093/bioinformatics/btx019

49. Bastian F, Parmentier G, Roux J, Moretti S, Laudet V, Robinson-Rechavi M. Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species. Data Integration in the Life Sciences. Springer, Berlin, Heidelberg; 2008. pp. 124–131. doi:10.1007/978-3-540-69828-9_12

50. Pybus M, Dall'Olio GM, Luisi P, Uzkudun M, Carreño-Torres A, Pavlidis P, et al. 1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans. Nucleic Acids Res. 2014;42: D903-9. doi:10.1093/nar/gkt1188

51. R Development Core Team. {R: A language and environment for statistical computing}. Vienna, Austria: R Foundation for Statistical Computing; 2009.

bioRxiv preprint doi: https://doi.org/10.1101/420646; this version posted September 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

TABLES

Table 1. SNPs at microRNA target sites (ancestral) with high Fst values.

			Target Allele Frequencies					
miRname*	**Gene Name**	**SNP**	**EAS**	**AMR**	**AFR**	**EUR**	**SAS**	**Fst**
let-7e-5p	MTAP	rs7875199	0.054	0.058	0.796	0.007	0.129	0.800
miR-150-5p	HM13	rs6059873	0.000	0.184	0.828	0.157	0.038	0.730
miR-374a-3p	SLC23A2	rs6052937	0.110	0.190	0.903	0.152	0.233	0.709
miR-192-5p	C12orf65	rs1533703	0.998	0.726	0.149	0.769	0.795	0.707
miR-151a-3p	ABCB9	rs4534620	0.002	0.053	0.635	0.002	0.006	0.671
miR-150-5p	HM13	rs6059869	0.000	0.182	0.800	0.157	0.038	0.637
miR-146a-5p	RGMA	rs59585671	0.239	0.657	0.924	0.996	0.764	0.636
miR-146a-5p	PKDREJ	rs12158566	1.000	0.937	0.421	0.963	0.964	0.621
miR-181a-5p	CXXC4	rs2298730	0.145	0.056	0.800	0.022	0.069	0.611

* Representative microRNA from the seed family

bioRxiv preprint doi: https://doi.org/10.1101/420646; this version posted September 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

FIGURE LEGENDS

Figure 1. Allele frequencies at microRNA target sites. (A) Target allele frequency distribution at microRNA target sites compared to a background distribution (see Methods). (B) Target allele frequency distribution at microRNA target sites in co-expressed transcripts compared to a target sites in anti-expressed transcripts. (C) Derive allele frequency distribution of microRNA target sites, where the ancestral allele is a target site, compared to a background distribution. (D) As in D but for sites in which the ancestral allele was a near-target.

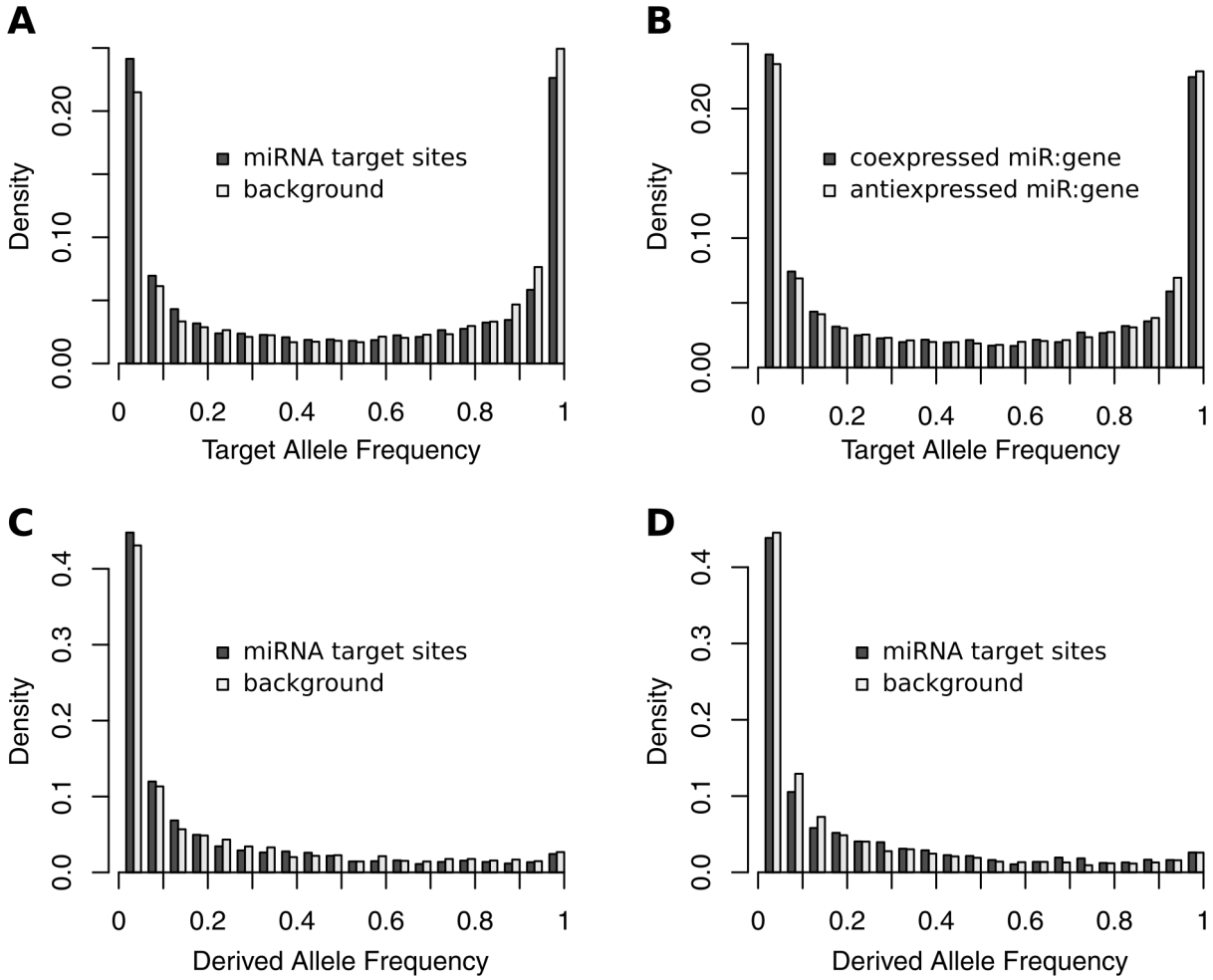
Figure 2. The effect of microRNA conservation and expression level on the frequency of target alleles. Gray tones represent the p-values from a one-tailed Mann-Whitney test for differences between the target allele frequencies of two datasets where the target allele is ancestral. The diagonal represents the self-comparisons (and therefore they have a value of 1) and the bottom-left triangular matrix are the p-values of a test with alternative hypothesis 'y-axis dataset has lower derived allele frequencies than the x-axis dataset'. The top-right matrix reverses the ordering in the test.

Figure 3. Fst enrichment in polymorphic microRNA target sites. Comparisons of the Fst values for SNPs within microRNA target sites with control sites (background sites, see Methods). For each range of Fst values, the proportion sites in that range was calculated for the microRNA target sites (P_t) and for the background sites (P_b). The value plotted is the log(P_t/ P_b). (A) All targets sites for broadly expressed microRNAs. (B) Target sites where the target allele is ancestral (red) and targets sites where the target site is not ancestral (blue). (C) As in (B) but for broadly and non-broadly expressed microRNAs.

bioRxiv preprint doi: https://doi.org/10.1101/420646; this version posted September 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

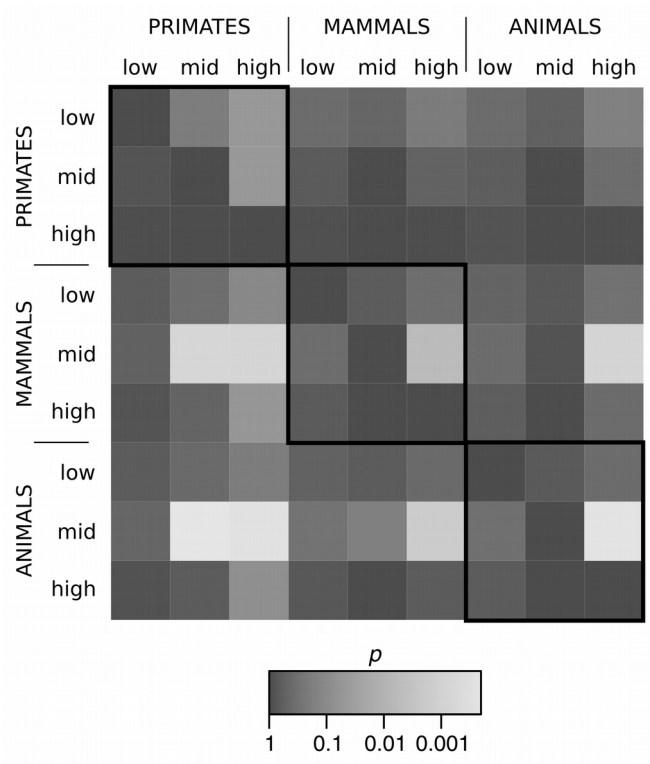
FIGURES

Figure 1



bioRxiv preprint doi: https://doi.org/10.1101/420646; this version posted September 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 2



bioRxiv preprint doi: https://doi.org/10.1101/420646; this version posted September 18, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

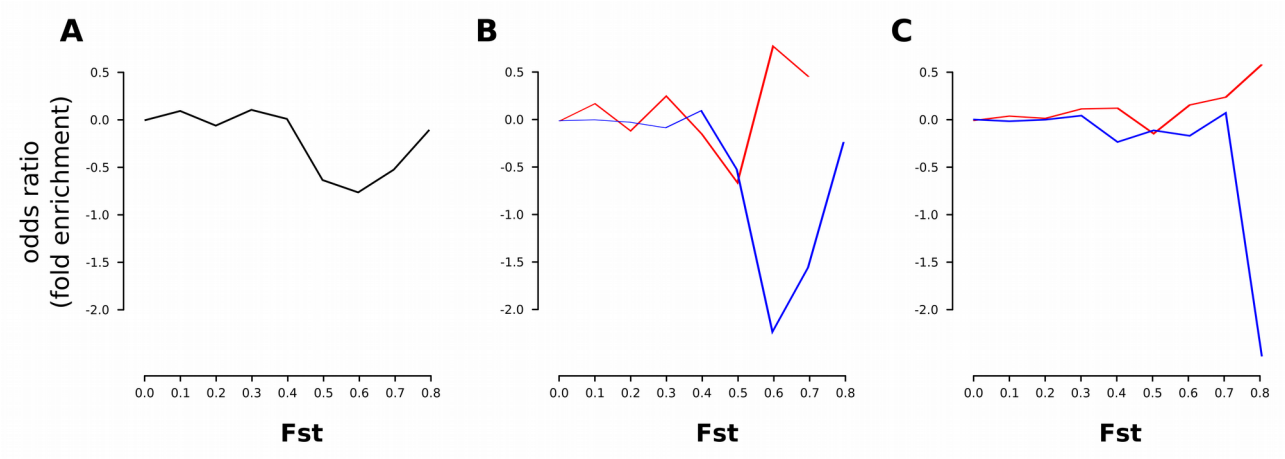



Figure 3