PopTargs: A database for studying population evolutionary genetics of human microRNA target sites

Andrea Hatlen[1], Mohab Helmy[1] and Antonio Marco[1,*]

[1] School of Biological Sciences, University of Essex, Wivenhoe Park, Colchester CO4 3SQ
United Kingdom

[*] To whom correspondence should be addressed

ABSTRACT

There is an increasing interest in the study of polymorphic variants at gene regulatory motifs, including microRNA target sites. Understanding the effects of selective forces at specific microRNA target sites, together with other factors like expression levels or evolutionary conservation, requires the joint study of multiple datasets. We have compiled information from multiple sources and compare it with predicted microRNA target sites to built a comprehensive database for the study of microRNA targets in human populations. PopTargs is a web-based tool that allows the easy extraction of multiple datasets and the joint analyses of them. The user can also compare the allele frequency spectrum between two groups of target sites, and conveniently produce plots. The database can be easily expanded as new data becomes available and the raw database as well as code for creating new custom made databases are available for downloading. We also describe a few illustrative examples. Poptargs is available at http://poptargs.essex.ac.uk

INTRODUCTION

As genome sequencing costs continue to decrease, the interest in population genetics increases. In particular, the analysis of variation at regulatory sites is becoming critical to understand how non-coding sequences emerge and evolve (1). MicroRNAs are important gene regulators that target gene transcripts by partial complementarity (2). The fact that their targets can be predicted from their primary sequence has been exploited to study the potential impact of single-nucleotide polymorphisms at their target sites. Indeed, a number of studies have reported selective pressures at these target sites by investigating the variation in populations (3–6).

Despite the interest in population genetics at microRNA target sites, their analysis imply the combination of various datasets from multiple databases, and handling simultaneously vasts amounts of data, making it cumbersome to research groups without the data analysis expertise or access to high-throughput computing clusters. To facilitate this task we have developed a database which cross-links allele frequencies at predicted microRNA target sites, as well as expression and evolutionary conservation information from other sources, and that permits the analysis of allele frequencies at target sites.

METHODS

Source of data

The human 3'UTRs were downloaded with Biomart (7) and the BiomaRt R package (8) from Ensembl database version 91 (human genome assembly GRCh38), and keeping only 3'UTRs from protein coding transcripts. All mature human microRNAs were downloaded from miRBase version 21 (9). SNPs were also retrieved with BiomaRt, and the allele frequencies were from the 1,000 Genomes Project (10) as compiled in dbSNP build 151 (11) [Ensembl Variation 95]. Genes were classified as 'over-' or 'under-expressed' by tissue according to the Bgee database, version 13.2 (12). MicroRNA tissue expression information was obtained from Meunier et.al (2013) and MiRmine (14). The miRNA data was classified into four groups for analysis, based on their expression in each tissue: 1- zero RPM (reads per million), 2- broad expression (>50RPM), 3- high expression (>500RPM), and 4- specifically expressed in one tissue (highly expressed compared to the other tissues: 1.5 times the interquartile range plus upper quartile across tissues. Target and near-target (one nucleotide difference with a target) sites were found using SeedVicious 1.1 (15), which predicts canonical target sites without filtering out for sequence conservation. Only SNP locations in which one allele was a target and another allele was a near-target were further considered. This important feature allows the study of target sites that are not in the reference genome, but that can be targets in some populations (see Results and Discussion).

Access and Implementation

The database is build in MySQL and it is freely accessible via a dedicated web portal at https://poptargs.essex.ac.uk/. The database provides three main options to explore microRNA target sites. First, users can search (*Search* tab) specific microRNAs or genes, and compare the allele frequencies between two lists of microRNAs or genes. The web form also gives the option to plot the allele frequencies side to side to a fast visual inspection of results. Alternatively, the users may browse the database (*Browse* tab) and select microRNAs with specific expression profiles and/or sequence conservation. This data can be retrieved for all or for specific human populations. Finally, the user has the option to download the whole MySQL database (*Downloads* tab). Researchers can also create their own databases with custom sequences as we also provide the source code and full instructions at https://github.com/ash8/PopTargs

RESULTS AND DISCUSSION

The basic search function of PopTargs is the 'Search' form. Users can provide list of mature microRNAs names and gene names (Ensembl unique IDs) and retrieve a table with SNPs variants found at target sites among the queried genes/microRNAs. As we considered near-targets (see above) during the database assembly the user will also find target sites that are not in the reference genome yet one of the alleles is associated to a target site. This feature can be exploited to detect putative target sites not present in the current reference genome sequence (see discussion at the end of this section). The table provides the population frequencies of the target allele, and also reports which allele is ancestral to human populations. Lists of microRNAs of interest can be obtained from miRBase (9) but also from curated databases that may allow the filtering of microRNAs based on evolutionary conservation or other features (e.g. mirGeneDB (16)). The possibility of providing lists of both microRNAs and genes helps to narrow down the targets of interest when a specific subset of experimentally validated interactions (for instance, from TarBase (17) or miRTarBase (18)) is to be explored. The database also allows the possibility of plotting allele frequencies for the queried microRNA/gene interactions. In this case, one can plot the allele frequencies at target sites and compare it with the allele frequencies of either an alternative list of microRNAs or an alternative list of genes. This is particularly handy when visually exploring large amount of data (see below).

To explore variation at target sites in pre-computed lists, the 'Browse' form allows to study microRNAs with different levels of expression, expression breath, evolutionary conservation and even sub-population structure. For instance, we recently reported that in human populations there is detectable selection against microRNA target sites (6). We can explore some specific cases with PopTargs. If we use the *Browse* option we can compare target sites for microRNAs highly expressed in testis (for instance) versus microRNAs not detected in testis. PopTargs will produce an allele frequency and a derived allele frequency plots, showing that the frequency of the target allele is significantly lower for the targets of highly expressed microRNAs (Figure 1). This result suggests that when a target site for a testis microRNA randomly appears in a testis expressed gene, there will be selective pressures to remove this allele from the population.

We can download a full table with the results, which will contain allele and derive allele frequencies, but also the target allele frequencies for different human populations and the estimated Fst (19). From the results produced we can detect unique 12 segregating target:non-target allele pairs for microRNAs highly expressed in testis (Table 1), that have a high degree of population differentiation (Fst>0.5). For instance, transcripts from *CYB5R4*, a gene associated to oxidative stress protection in sperm production (20), has a conserved target site for miR-151-5p (highly expressed in testes) but this target site is not detected in the reference genome. Indeed, the loss of the ancestral target site happened in European populations whilst other human groups mostly maintain the target allele (dbSNP entry rs6912739, Table 1). This result illustrates how population dynamics can be used to detect target sites that are not in the reference genome and, therefore, escape most target prediction programs (Helmy, Hatlen, Marco, under review).

We provided all scripts used to generate the original database and full documentation such that interested users can generate their own database. As the number of available genome sequences increases, this feature can be of use to those interested in expanding the current database.

ACKNOWLEDGMENTS

We thank Stuart Newman for his invaluable help in setting up the server to host our database and web portal.

FUNDING

This work was supported by the Wellcome Trust [grant number 200585/Z/16/Z].

REFERENCES

1. Abecasis, G. R., Auton, A., Brooks, L. D., et al. (2012) An integrated map of genetic variation from 1,092 human genomes, *Nature*, **491**, 56–65.
2. Bartel, D. P. (2009) MicroRNAs: Target Recognition and Regulatory Functions, *Cell*, **136**, 215–233.
3. Chen, K. and Rajewsky, N. (2006) Natural selection on human microRNA binding sites inferred from SNP data, *Nat. Genet.*, **38**, 1452–1456.
4. Saunders, M. A., Liang, H. and Li, W.-H. (2007) Human polymorphism at microRNAs and microRNA target sites, *Proc. Natl. Acad. Sci. U. S. A.*, **104**, 3300–3305.
5. Marco, A. (2015) Selection Against Maternal microRNA Target Sites in Maternal

Transcripts, *G3 GenesGenomesGenetics*, g3.115.019497.

6. Hatlen, A. and Marco, A. (2018) Pervasive selection against microRNA target sites in human populations, *bioRxiv*, 420646.
7. Kinsella, R. J., Kähäri, A., Haider, S., et al. (2011) Ensembl BioMarts: a hub for data retrieval across taxonomic space, *Database J. Biol. Databases Curation*, **2011**, bar030.
8. Durinck, S., Spellman, P. T., Birney, E., et al. (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt, *Nat. Protoc.*, **4**, 1184.
9. Kozomara, A. and Griffiths-Jones, S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data, *Nucleic Acids Res.*, **42**, D68-73.
10. 1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., et al. (2010) A map of human genome variation from population-scale sequencing, *Nature*, **467**, 1061–1073.
11. Sherry, S. T., Ward, M. H., Kholodov, M., et al. (2001) dbSNP: the NCBI database of genetic variation, *Nucleic Acids Res.*, **29**, 308–311.
12. Bastian, F., Parmentier, G., Roux, J., et al. In *Data Integration in the Life Sciences*; Lecture Notes in Computer Science; Springer, Berlin, Heidelberg, 2008; pp. 124–131.
13. Meunier, J., Lemoine, F., Soumillon, M., et al. (2013) Birth and expression evolution of mammalian microRNA genes, *Genome Res.*, **23**, 34–45.
14. Panwar, B., Omenn, G. S. and Guan, Y. (2017) miRmine: a database of human miRNA expression profiles, *Bioinformatics*, **33**, 1554–1560.
15. Marco, A. (2018) SeedVicious: Analysis of microRNA target and near-target sites, *PLOS ONE*, **13**, e0195532.
16. Fromm, B., Domanska, D., Hackenberg, M., et al. (2018) MirGeneDB2.0: the curated microRNA Gene Database, *bioRxiv*, 258749.

17. Vlachos, I. S., Paraskevopoulou, M. D., Karagkouni, D., et al. (2015) DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions, *Nucleic Acids Res.*, **43**, D153–D159.

18. Chou, C.-H., Shrestha, S., Yang, C.-D., et al. (2018) miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions, *Nucleic Acids Res.*, **46**, D296–D302.

19. Pybus, M., Dall'Olio, G. M., Luisi, P., et al. (2014) 1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans., *Nucleic Acids Res.*, **42**, D903-9.

20. H.m., Y., Kumar, S., Dubey, P. P., et al. (2017) Profiling of sperm gene transcripts in crossbred (Bos taurus x Bos indicus) bulls, *Anim. Reprod. Sci.*, **177**, 25–34.

TABLES

Table 1. Target sites for testis-expressed microRNAs with a high degree of population differentiation.

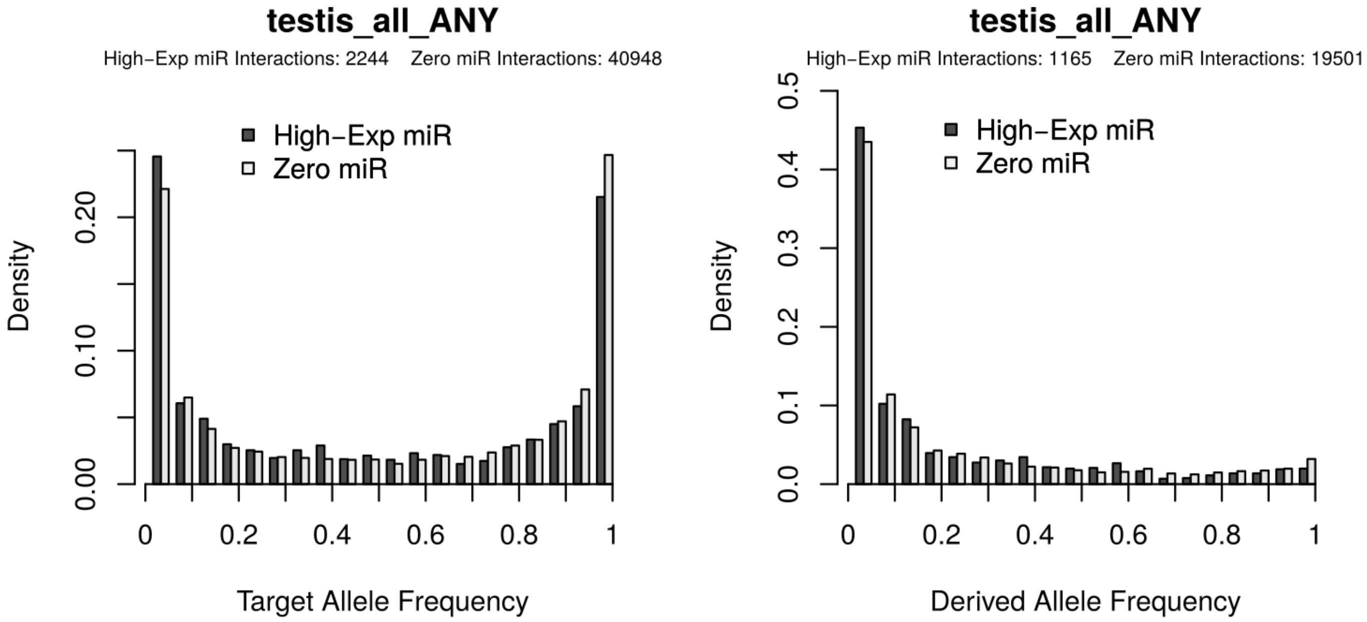
MicroRNA	Gene	SNP	target is ancestral	EAS	AMR	AFR	EUR	SAS	all	Fst
miR-151-5p	CYB5R4	rs6912739	yes	0	0.0677	0.6188	0.0457	0.0061	0.1831	0.6458
miR-22-3p	OPLAH	rs28475718	yes	1	0.9568	0.4365	0.9891	0.9693	0.8380	0.6414
miR-9-5p	C8orf74	rs10107820	no	0.0218	0.0389	0.5696	0.006	0.0133	0.1639	0.6363
miR-197-3p	RGMA	rs4411467	yes	0.3214	0.6412	0.975	0.9622	0.7607	0.7536	0.5957
miR-30-5p	NUP155	rs10473043	yes	0.124	0.219	0.7905	0.1521	0.2219	0.3381	0.5738
miR-181a-2-3p	C8orf74	rs10094968	no	0.0228	0.0389	0.59	0.007	0.0532	0.1775	0.5722
miR-30-5p	CCDC14	rs2700372	no	0.2252	0.6066	0.3525	0.9732	0.7311	0.5615	0.5714
miR-378g	BMP3	rs1495637	no	0.8185	0.7046	0.1089	0.7773	0.5869	0.5624	0.5704
miR-381-3p	PAWR	rs2307220	yes	0.3065	0.2147	0.9115	0.1551	0.1922	0.4009	0.5343
miR-486-5p	ZCCHC14	rs1050863	yes	0.996	0.7478	0.9448	0.4592	0.8967	0.8216	0.5316
miR-103a-3p/107	WDR27	rs3800547	no	0.5972	0.2709	0.0106	0.0646	0.1135	0.1957	0.5114
miR-30-3p	ABRAXAS1	rs17006851	no	1	0.964	0.4221	0.996	1	0.8426	0.506

The target allele frequencies are provided for East Asian (EAS), Mixed American (AMR), African (AFR), European (EUR) and South Asian (SAS) populations, as described in the 1000 genomes project (see Methods).

FIGURE LEGENDS

Figure 1. Allele frequency distributions as generated from the PopTargs web server. The left panel show the target allele frequency distribution for microRNAs highly expressed in testes (grey bars) and for microRNAs whose expression was not detected in testes (white bars). Likewise, the right panel shows the target allele frequency distribution of derived alleles, that is, where the ancestral allele is a non-target. The latter plot is also often called in population genetics the Site Frequency Spectrum.

Figure 1


testis_all_ANY
High−Exp miR Interactions: 2244 Zero miR Interactions: 40948
■ High−Exp miR
□ Zero miR
Density
0.00
0.10
0.20
0
0.2
0.4
0.6
0.8
1
Target Allele Frequency
testis_all_ANY
High−Exp miR Interactions: 1165 Zero miR Interactions: 19501
■ High−Exp miR
□ Zero miR
Density
0.0
0.1
0.2
0.3
0.4
0.5
0
0.2
0.4
0.6
0.8
1
Derived Allele Frequency