

Article

The Impact of Population Variation in the Analysis of microRNA Target Sites

Mohab Helmy[†], Andrea Hatlen and Antonio Marco^{*ID}

School of Biological Sciences, University of Essex, Colchester CO4 3SQ, UK; mhaelb@essex.ac.uk (M.H.); a.hatlen@essex.ac.uk (A.H.)

* Correspondence: amarco.bio@gmail.com; Tel.: +44-1206-873339

† Current address: Faculty of Life Sciences & Medicine, King's College London, London SE1 9RT, UK.

Received: 22 March 2019; Accepted: 20 June 2019; Published: 22 June 2019



Abstract: The impact of population variation in the analysis of regulatory interactions is an underdeveloped area. MicroRNA target recognition occurs via pairwise complementarity. Consequently, a number of computational prediction tools have been developed to identify potential target sites that can be further validated experimentally. However, as microRNA target predictions are done mostly considering a reference genome sequence, target sites showing variation among populations are neglected. Here, we studied the variation at microRNA target sites in human populations and quantified their impact in microRNA target prediction. We found that African populations carry a significant number of potential microRNA target sites that are not detectable in the current human reference genome sequence. Some of these targets are conserved in primates and only lost in Out-of-Africa populations. Indeed, we identified experimentally validated microRNA/transcript interactions that are not detected in standard microRNA target prediction programs, yet they have segregating target alleles abundant in non-European populations. In conclusion, we show that ignoring population diversity may leave out regulatory elements essential to understand disease and gene expression, particularly neglecting populations of African origin.

Keywords: miRNAs; human populations; gene regulation; microRNA target prediction; evolution

1. Introduction

The study of gene regulatory iterations is at the heart of biomedical research. Many disease-causing mutations affect regulatory motifs that eventually lead to a miss-regulation of fine-tuned biological processes. In the last years, microRNAs have emerged as an important type of regulatory molecules involved in virtually all biological networks [1,2]. Both the deletion and overexpression of microRNAs have been associated to human diseases (see for instance the compilation by [3]). Notably, the role of microRNAs in cancer is a particularly active research field [4,5]. Unlike other regulatory molecules (such as transcription factors or RNA binding proteins), the interaction between regulator and target is mediated by pairwise nucleotide complementary [1]. Specifically, a microRNA is partially complementary to motifs in RNA transcripts and, after binding, translational repression or even RNA degradation mechanisms are triggered [1,6]. Indeed, changes at microRNA target sites are now associated to major human diseases including cancer, neurodegeneration and metabolic disorders (see [7,8] and references therein), among others. The study of microRNA target sites is becoming more important as we start to understand the complexities behind microRNA regulatory networks.

From a computational point of view, the microRNA targeting mechanisms have a clear advantage: targets can be predicted by scanning genomic sequences for potential complementary sites [9–11]. However, microRNA target prediction algorithms have, in general, high rates of false positives [12,13]. Consequently, most research on microRNA targets complements a computational sieve of potential

target sites plus experimental validations, usually by luciferase assays or similar (e.g., [14]). Therefore, the first step in microRNA target analysis is the prediction of binding sites from a primary sequence. In biomedical research, this primary sequence is often, if not always, the human reference genome sequence (currently hg38). However, a reference genome does not take into account existing variation within human populations and, therefore, any inference from such a sequence may be biased. Populations of African origin show a higher nucleotide diversity than other populations, probably as a consequence of a recent origin of Eurasian populations from a small group of African migrants (reviewed in [15]). There exist several projects aimed to account for all this nucleotide variability in human populations (i.e., [16]). Nevertheless, microRNA target prediction programs rely on reference genome sequences (as other regulatory interaction programs do), where this diversity is not truly represented.

Variation at gene regulatory sites, particularly at transcription factor binding sites, has been studied in the past (e.g., [17,18]). In microRNA target sites, some studies have shown that, as expected, functional target sites are often conserved in populations [19,20]. On the other hand, the creation of novel, potentially harmful, microRNA target sites, is selected against in populations [21,22]. In summary, the literature clearly shows evidence that nucleotide variation occurs at microRNA target sites. The question is, how much of this variation is neglected in the human reference genome sequence, and therefore not accounted for in standard microRNA target prediction programs? In other words, what is the impact of nucleotide variation among humans in microRNA target prediction research? This is the question we are tackling in this paper.

2. Results and Discussion

We compiled biallelic SNPs at predicted microRNA target sites (see Methods) for which one of the alleles is a target site and the other is not a target site. These are the target/near-target pairs as defined before [11,21]. Considering that populations of African origin host more variation, including ancestral alleles, than other populations, we hypothesize that ancestral target sites that are not present in the reference genome sequence were lost in Out-of-Africa populations. Thus, we first consider target sites whose allele in the reference genome sequence is not a target, yet the ancestral allele was a target site. The distribution of these alleles frequencies in five major groups (African, European, South Asian, Native American and East Asian) indicates that non-African populations tend to have the non-target allele whilst African populations show a wider variation, and a paucity of non-targets compared to the other populations (Figure 1, top). Additionally, we consider target sites in the reference genome sequence that were not ancestral to human populations and, as expected, we found that African populations have lower frequencies of the target allele (Figure 1, bottom). These results suggest that a significant loss and gain of microRNA target sites happened in Out-of-African populations.

To illustrate the impact of variation at target sites we selected SNPs with a very high degree of population differentiation ($F_{st} > 0.7$; see Methods) that are not present in the human genome reference sequence. We found 26 target sites with such a degree of variation across populations (Table 1) for highly expressed microRNAs. Many of these target sites are lost in Out-of-Africa populations, suggesting that losses rather than the gains of target sites are more likely to happen during differentiation between populations (see discussion in [22]).

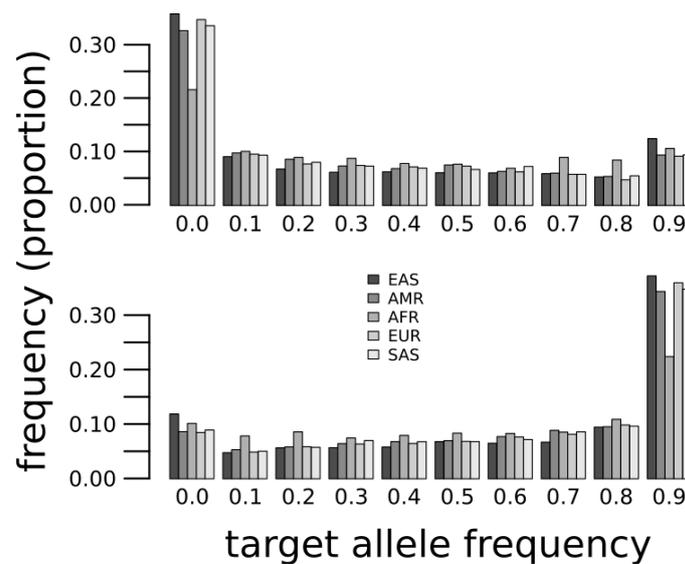


Figure 1. Target allele frequencies at microRNA target sites where the target is ancestral and not in the reference genome sequence (**top**) or the target is derived and present in the reference genome sequence (**bottom**). Super-populations are African (AFR), Mixed American (AMR), East Asian (EAS), European (EUR) and South Asian (SAS).

Table 1. MicroRNA target sites, not in the reference genome sequence, with a very high differentiation among human populations.

MicroRNA	Gene	SNP ^a	EAS ^b	AMR ^b	AFR ^b	EUR ^b	SAS ^b	Fst
miR-519a-5p ¹	MTAP	rs7868374	0.0079	0.0476	0.6838	0.0050	0.1176	0.8239
let-7a-5p ²	MTAP	rs7875199	0.0536	0.0576	0.7958	0.0070	0.1288	0.7997
miR-337-3p ³	ATP1A1	rs1885802	0.0427	0.1268	0.8109	0.0338	0.0450	0.7914
miR-185-3p ⁴	SLC5A10	rs1624825	0.9970	0.5159	0.9667	0.2028	0.7495	0.7859
miR-1180-3p	MYEF2	rs2470102	0.7530	0.3386	0.9251	0.0060	0.2720	0.7726
miR-151a-5p ⁵	SCN2B	rs624328	0.9405	0.8905	0.2519	0.9513	0.9673	0.7674
miR-4732-5p	AP5M1	rs1889720	0.9335	0.9107	0.1551	0.9433	0.9182	0.7637
miR-584-5p	CYB5R4	rs6903739	0.0000	0.0749	0.7738	0.0477	0.0061	0.7514
miR-618	CYB5R4	rs9449733	1.0000	0.9251	0.2247	0.9523	0.9939	0.7397
miR-18a-3p	ANKRD65	rs904589	0.1131	0.1945	0.8222	0.0964	0.1881	0.7332
miR-329-3p ⁶	CEP162	rs6901546	0.0139	0.1628	0.7670	0.0706	0.0900	0.7326
miR-150-5p ⁷	HM13	rs6059873	1.0000	0.8156	0.1725	0.8429	0.9622	0.7298
miR-513a-3p ⁸	TCERG1	rs3822506	0.7183	0.1167	0.0325	0.0915	0.2198	0.7224
miR-128-1-5p ⁹	BCL7C	rs11864054	0.9117	0.5447	0.0242	0.3817	0.1595	0.7144
miR-192-5p ¹⁰	C12orf65	rs1533703	0.9980	0.7262	0.1490	0.7694	0.7945	0.7074

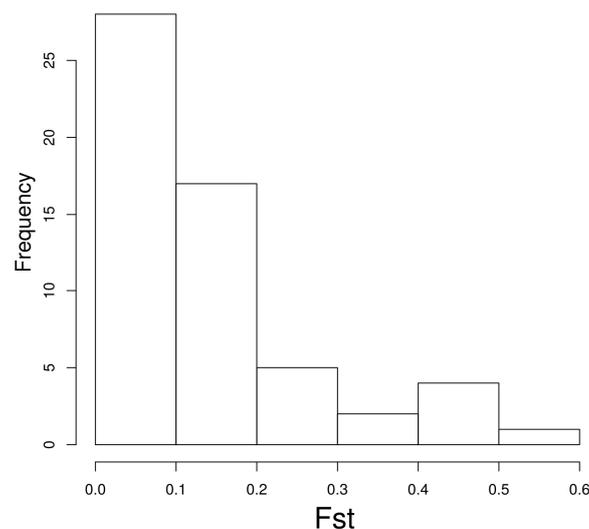
^a SNP names are dbSNP accession numbers. ^b Allele frequencies refer to the target allele frequency in each of the five super-population groups. Superscript in microRNA names indicates other mature sequences with the same target site: ¹ miR-520c-5p, miR-519b-5p, miR-526a-5p, miR-519c-5p, miR-518f-5p, miR-523-5p, miR-518d-5p, miR-522-5p, miR-518e-5p; ² let-7e-5p, let-7c-5p, let-7f-5p, let-7d-5p, let-7b-5p, miR-98-5p, let-7i-5p, let-7g-5p, miR-1294; ³ miR-202-5p; ⁴ miR-128-1-5p; ⁵ miR-561-5p, miR-3605-5p, miR-24-3p, miR-151b; ⁶ miR-362-3p; ⁷ miR-296-3p, miR-4446-3p, miR-23b-5p, miR-3158-3p; ⁸ miR-513c-3p; ⁹ miR-491-5p, miR-1296-5p; ¹⁰ miR-215-5p.

We observed that *MTAP* have lost two target sites for two highly expressed microRNAs in non-African populations. These two target sites were the ancestral form in human populations and conserved as targets in other primates (Figure 2A). Genome-wide association studies have linked variants at *MTAP* with the abundance of naevi (moles) and also with melanoma incidence [23,24], although the association between *MTAP* and melanoma seems to be rather complex [23]. We speculate that the loss of two target sites at *MTAP* transcripts could be associated with an elevated expression of the gene in Out-of-Africa populations, perhaps by a relaxation on the selective pressures that maintained the target sites under purifying selection in African populations, where light exposure is

Table 2. Target sites identified in high-throughput experiments with high population differentiation.

MicroRNA	Gene	SNP ^a	EAS ^b	AMR ^b	AFR ^b	EUR ^b	SAS ^b	Fst
miR-1307-3p	VSTM4	rs4240499	0.2361	0.4971	0.9017	0.3976	0.5031	0.5352
miR-512-3p	EIF2B2	rs4556	0.8393	0.6441	0.1884	0.5388	0.5194	0.4327
miR-197-3p	DBT	rs6701655	0.9296	0.8905	0.4168	0.8648	0.8978	0.4186
miR-17-5p	MRPS10	rs3199638	0.7937	0.6369	0.2148	0.6372	0.7679	0.3946
miR-326	SLC27A4	rs7048106	0.3185	0.2176	0.7617	0.2386	0.3476	0.3782
miR-24-3p	CASP10	rs13432040	0.000	0.0259	0.2587	0.000	0.000	0.2943
miR-342-3p	LRPAP1	rs3468	0.5089	0.2507	0.1611	0.3489	0.2669	0.238

^a SNP names are dbSNP accession numbers. ^b Allele frequencies refer to the target allele frequency in each of the five super-population groups.

**Figure 3.** Fst values for polymorphic microRNA target sites experimentally evaluated with PASSPORT-seq (see main text).

Last, we investigated polymorphic target sites identified by the popular software TargetScan. We downloaded the whole conserved and non-conserved target datasets and identified polymorphic target sites also present in our database PopTargets. We compared the target allele frequencies between populations and concluded that, for both conserved and nonconserved targets, the target allele was more common in European than in African populations (Table 3). Indeed, for nonconserved targets, the differences are larger.

Table 3. Wilcoxon signed-rank (paired) test between African and European target allele frequencies in TargetScan predicted microRNA target sites.

TargetScan Dataset	Sample Size (pairs)	<i>p</i>	Shift Median ¹	Shift 95% CI ²
conserved	9245	$<2.2 \times 10^{-16}$	7.976×10^{-4}	$(7.833-7.878) 10^{-4}$
nonconserved	1041868	$<2.2 \times 10^{-16}$	1.174×10^{-3}	$(1.151-1.210) 10^{-3}$

¹ Estimated shift median for the difference: European minus African target allele frequency. ² Confidence Interval.

In conclusion, here, we showed that the study of variation in human populations reveals the presence of microRNA target sites that are not detected with current methodological frameworks. We suggest that variation at target sites must be taken into account in biomedical studies as some populations (mostly of African origin) may be neglected. Further research is needed to quantify the impact of population variation in the function of microRNA target sites.

3. Methods

Polymorphic sites at predicted microRNA target sites were retrieved from the PopTargs database (version 2; Hatlen, Helmy and Marco, under review; <https://poptargs.essex.ac.uk>). Only target sites for highly expressed microRNAs were considered. Population frequencies, Fst values and ancestral alleles were also retrieved in PopTargs, as provided in [30]. We only consider segregating alleles with a frequency between 0.01 and 0.99. Sequence alignment of primate sequences was extracted using the UCSC table browser [31]. The geographical location (latitude) of the genome samples sequenced in the 1000 genomes project was obtained from the sample descriptions at <https://www.coriell.org/1/NHGRI/Collections/1000-Genomes-Collections/1000-Genomes-Project>. When the specific location was ambiguous, we used the location of the capital city of the country of origin. For multiple collection points we computed the centroid middle point. Recently migrated populations were discarded (CHD, CEU, ASW, ACB, GIH, STU and ITU as described at IGSR: The International Genome Sample Resource (<http://www.internationalgenome.org/category/population/>)) were discarded from the analysis in Figure 2B. We extracted microRNA targets from the miRTarBase database [32], which were detected only in high-throughput experiments (PAR-CLIP and/or HITS-CLIP).

Author Contributions: Conceptualization, A.M.; methodology, A.H. and A.M.; formal analysis, M.H. and A.M.; investigation, M.H., A.H. and A.M.; data curation, M.H., A.H. and A.M.; writing—original draft preparation, A.M.; writing—review and editing, M.H. and A.M.; supervision, A.M.; project administration, A.M.; funding acquisition, A.M.

Funding: This research was funded by Wellcome Trust, grant number 200585/Z/16/Z.

Acknowledgments: The authors acknowledge the use of the High Performance Computing Facility (Ceres) and its associated support services at the University of Essex in the completion of this work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bartel, D.P. MicroRNAs: Target Recognition and Regulatory Functions. *Cell* **2009**, *136*, 215–233. [[CrossRef](#)] [[PubMed](#)]
2. Axtell, M.J.; Westholm, J.O.; Lai, E.C. Vive la différence: Biogenesis and evolution of microRNAs in plants and animals. *Genome Biol.* **2011**, *12*, 221. [[CrossRef](#)] [[PubMed](#)]
3. Jiang, Q.; Wang, Y.; Hao, Y.; Juan, L.; Teng, M.; Zhang, X.; Li, M.; Wang, G.; Liu, Y. miR2Disease: A manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res.* **2009**, *37*, D98–D104. [[CrossRef](#)] [[PubMed](#)]
4. Di Leva, G.; Garofalo, M.; Croce, C.M. microRNAs in cancer. *Annu. Rev. Pathol.* **2014**, *9*, 287–314. [[CrossRef](#)] [[PubMed](#)]
5. Croce, C.M. Causes and consequences of microRNA dysregulation in cancer. *Nat Rev Genet* **2009**, *10*, 704–714. [[CrossRef](#)] [[PubMed](#)]
6. Ghildiyal, M.; Zamore, P.D. Small silencing RNAs: An expanding universe. *Nat. Rev. Genet.* **2009**, *10*, 94–108. [[CrossRef](#)]
7. Bhattacharya, A.; Ziebarth, J.D.; Cui, Y. PolymiRTS Database 3.0: Linking polymorphisms in microRNAs and their target sites with human diseases and biological pathways. *Nucleic Acids Res.* **2014**, *42*, D86–D91. [[CrossRef](#)] [[PubMed](#)]
8. Hiard, S.; Charlier, C.; Coppieters, W.; Georges, M.; Baurain, D. Patrocles: A database of polymorphic miRNA-mediated gene regulation in vertebrates. *Nucleic Acids Res.* **2010**, *38*, D640–D651. [[CrossRef](#)]
9. Agarwal, V.; Bell, G.W.; Nam, J.-W.; Bartel, D.P. Predicting effective microRNA target sites in mammalian mRNAs. *eLife* **2015**, *4*, e05005. [[CrossRef](#)]
10. Paraskevopoulou, M.D.; Georgakilas, G.; Kostoulas, N.; Vlachos, I.S.; Vergoulis, T.; Reczko, M.; Filippidis, C.; Dalamagas, T.; Hatzigeorgiou, A.G. DIANA-microT web server v5.0: Service integration into miRNA functional analysis workflows. *Nucleic Acids Res.* **2013**, *41*, W169–W173. [[CrossRef](#)]

11. Marco, A. SeedVicious: Analysis of microRNA target and near-target sites. *PLoS ONE* **2018**, *13*, e0195532. [[CrossRef](#)] [[PubMed](#)]
12. Alexiou, P.; Maragkakis, M.; Papadopoulos, G.L.; Reczko, M.; Hatzigeorgiou, A.G. Lost in translation: An assessment and perspective for computational microRNA target identification. *Bioinformatics* **2009**, *25*, 3049–3055. [[CrossRef](#)] [[PubMed](#)]
13. Pinzón, N.; Li, B.; Martinez, L.; Sergeeva, A.; Presumey, J.; Apparailly, F.; Seitz, H. microRNA target prediction programs predict many false positives. *Genome Res.* **2017**, *27*, 234–245. [[CrossRef](#)] [[PubMed](#)]
14. Kuhn, D.E.; Martin, M.M.; Feldman, D.S.; Terry, A.V.; Nuovo, G.J.; Elton, T.S. Experimental validation of miRNA targets. *Methods* **2008**, *44*, 47–54. [[CrossRef](#)] [[PubMed](#)]
15. Barbujani, G.; Colonna, V. Human genome diversity: Frequently asked questions. *Trends Genet. Fig.* **2010**, *26*, 285–295. [[CrossRef](#)] [[PubMed](#)]
16. 1000 Genomes Project Consortium; Abecasis, G.R.; Altshuler, D.; Auton, A.; Brooks, L.D.; Durbin, R.M.; Gibbs, R.A.; Hurles, M.E.; McVean, G.A. A map of human genome variation from population-scale sequencing. *Nature* **2010**, *467*, 1061–1073. [[CrossRef](#)] [[PubMed](#)]
17. Garfield, D.; Haygood, R.; Nielsen, W.J.; Wray, G.A. Population genetics of cis-regulatory sequences that operate during embryonic development in the sea urchin *Strongylocentrotus purpuratus*. *Evol. Dev.* **2012**, *14*, 152–167. [[CrossRef](#)]
18. Kasowski, M.; Grubert, F.; Heffelfinger, C.; Hariharan, M.; Asabere, A.; Waszak, S.M.; Habegger, L.; Rozowsky, J.; Shi, M.; Urban, A.E.; et al. Variation in transcription factor binding among humans. *Science* **2010**, *328*, 232–235. [[CrossRef](#)]
19. Chen, K.; Rajewsky, N. Natural selection on human microRNA binding sites inferred from SNP data. *Nat. Genet.* **2006**, *38*, 1452–1456. [[CrossRef](#)]
20. Saunders, M.A.; Liang, H.; Li, W.-H. Human polymorphism at microRNAs and microRNA target sites. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 3300–3305. [[CrossRef](#)]
21. Marco, A. Selection Against Maternal microRNA Target Sites in Maternal Transcripts. *G3 Genesgenomesgenetics* **2015**, *5*, 2199–2207. [[CrossRef](#)] [[PubMed](#)]
22. Hatlen, A.; Marco, A. Pervasive selection against microRNA target sites in human populations. *bioRxiv* **2018**, 420646.
23. Kvaskoff, M.; Whiteman, D.C.; Zhao, Z.Z.; Montgomery, G.W.; Martin, N.G.; Hayward, N.K.; Duffy, D.L. Polymorphisms in naevus-associated genes MTAP, PLA2G6, and IRF4 and the risk of invasive cutaneous melanoma. *Twin Res. Hum. Genet.* **2011**, *14*, 422–432. [[CrossRef](#)] [[PubMed](#)]
24. Bishop, D.T.; Demenais, F.; Iles, M.M.; Harland, M.; Taylor, J.C.; Corda, E.; Randerson-Moor, J.; Aitken, J.F.; Avril, M.-F.; Azizi, E.; et al. Genome-wide association study identifies three loci associated with melanoma risk. *Nat. Genet.* **2009**, *41*, 920–925. [[CrossRef](#)] [[PubMed](#)]
25. Li, J.; Liu, Y.; Xin, X.; Kim, T.S.; Cabeza, E.A.; Ren, J.; Nielsen, R.; Wrana, J.L.; Zhang, Z. Evidence for Positive Selection on a Number of MicroRNA Regulatory Interactions during Recent Human Evolution. *PLoS Genet.* **2012**, *8*, e1002578. [[CrossRef](#)]
26. Chang, C.L.; Cai, J.J.; Huang, S.Y.; Cheng, P.J.; Chueh, H.Y.; Hsu, S.Y.T. Adaptive Human CDKAL1 Variants Underlie Hormonal Response Variations at the Enteroinular Axis. *PLoS ONE* **2014**, *9*, e105410. [[CrossRef](#)]
27. Chang, C.L.; Cai, J.J.; Cheng, P.J.; Chueh, H.Y.; Hsu, S.Y.T. Identification of Metabolic Modifiers That Underlie Phenotypic Variations in Energy-Balance Regulation. *Diabetes* **2011**, *60*, 726–734. [[CrossRef](#)]
28. Duan, Y.; Hu, L.; Liu, B.; Yu, B.; Li, J.; Yan, M.; Yu, Y.; Li, C.; Su, L.; Zhu, Z.; et al. Tumor suppressor miR-24 restrains gastric cancer progression by downregulating RegIV. *Mol. Cancer* **2014**, *13*, 127. [[CrossRef](#)]
29. Ipe, J.; Collins, K.S.; Hao, Y.; Gao, H.; Bhatia, P.; Gaedigk, A.; Liu, Y.; Skaar, T.C. PASSPORT-seq: A Novel High-Throughput Bioassay to Functionally Test Polymorphisms in Micro-RNA Target Sites. *Front. Genet.* **2018**, *9*, 219. [[CrossRef](#)]
30. Pybus, M.; Dall’Olio, G.M.; Luisi, P.; Uzkudun, M.; Carreño-Torres, A.; Pavlidis, P.; Laayouni, H.; Bertranpetit, J.; Engelken, J. 1000 Genomes Selection Browser 1.0: A genome browser dedicated to signatures of natural selection in modern humans. *Nucleic Acids Res.* **2014**, *42*, D903–D909. [[CrossRef](#)]

31. Karolchik, D.; Hinrichs, A.S.; Furey, T.S.; Roskin, K.M.; Sugnet, C.W.; Haussler, D.; Kent, W.J. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **2004**, *32*, D493–D496. [[CrossRef](#)] [[PubMed](#)]
32. Chou, C.-H.; Shrestha, S.; Yang, C.-D.; Chang, N.-W.; Lin, Y.-L.; Liao, K.-W.; Huang, W.-C.; Sun, T.-H.; Tu, S.-J.; Lee, W.-H.; et al. miRTarBase update 2018: A resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res.* **2018**, *46*, D296–D302. [[CrossRef](#)] [[PubMed](#)]



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).