nature genetics



# Natural selection on human microRNA binding sites inferred from SNP data

Kevin Chen[1] & Nikolaus Rajewsky[1,2]

**A fundamental problem in biology is understanding how natural selection has shaped the evolution of gene regulation. Here we use SNP genotype data and techniques from population genetics to study an entire layer of short, *cis*-regulatory sites in the human genome. MicroRNAs (miRNAs) are a class of small noncoding RNAs that post-transcriptionally repress mRNA through *cis*-regulatory sites in 3′ UTRs. We show that negative selection in humans is stronger on computationally predicted conserved miRNA binding sites than on other conserved sequence motifs in 3′ UTRs, thus providing independent support for the target prediction model and explicitly demonstrating the contribution of miRNAs to darwinian fitness. Our techniques extend to nonconserved miRNA binding sites, and we estimate that 30%–50% of these are functional when the mRNA and miRNA are endogenously coexpressed. As we show that polymorphisms in predicted miRNA binding sites are likely to be deleterious, they are candidates for causal variants of human disease. We believe that our approach can be extended to studying other classes of *cis*-regulatory sites.**

Despite longstanding interest in the evolution of human gene regulation, it is only recently that data have become available to study this process genome-wide, with the emergence of SNP genotype data in three human populations from the HapMap[1] and Perlegen[2] projects and multiple sequenced mammalian genomes. Here we focus on miRNAs, which are thought to regulate >30% of human protein-coding genes and have crucial roles in development and metabolism[3,4]. Naturally occurring polymorphisms in miRNA binding sites have been implicated in Tourette's syndrome in humans[5] and muscularity in sheep[6], and a set of SNPs in miRNA-associated motifs has been collected[6].

Several groups have computationally predicted a large class of miRNA binding sites (hereafter referred to as miRNA sites) using a model in which miRNA-mRNA binding is nucleated by an exact Watson-Crick match to the first six to eight bases from the 5′ end of the miRNA (reviewed in ref. 3). Using this approach, we predicted ~22,000 miRNA sites conserved in five mammals (see Methods). Previous analyses using comparative genomics suggested an average false positive rate of ~50% for predicted miRNA sites conserved in these species[3]. However, this estimate is likely to be conservative as it is difficult to construct sets of nonfunctional sequences as controls, and the approach gives no estimate for nonconserved miRNA sites.

Genome-wide SNP genotype data provide a new approach for studying *cis*-regulatory sites. It has recently been shown[7] that SNP density in exonic splicing enhancers is lower than expected, suggesting negative selection, but with the caveat that SNP density is sensitive to mutation rate biases (such as those due to base composition). Other groups used genotype data to demonstrate negative selection, as opposed to a lower mutation rate, on conserved noncoding regions, but they made no explicit connection to *cis*-regulatory sites (see ref. 8 and references therein). These studies examined only the average intensity of selection on an entire class of elements and not the distribution of selective effects within the class.

Here we use SNP density and genotype data from three human populations to study miRNA-mediated gene regulation. We mapped ~25,000 SNPs genotyped in the HapMap and Perlegen projects to human 3′ UTRs, of which 384 mapped to conserved miRNA sites. We found that SNP density in conserved miRNA sites (0.50 SNPs/kb) was lower than in conserved control sites (0.73 SNPs/kb; see Methods) and that SNP density was significantly lower in the region matching the 5′ end of the miRNA than in the rest of the site (**Fig. 1**; $\chi^2$ test, $P = 5 \times 10^{-5}$). Because we used conservation in the 5′ region for target prediction, we verified that the difference in SNP density between the two regions was significantly larger in predicted miRNA sites than in conserved control sites ($\chi^2$ test, $P = 0.0004$). We also observed a slight enrichment for SNPs that preserve GU base pairings in the binding site compared with those in 3′ UTRs overall ($\chi^2$ test, $P = 0.13$), suggesting that GU base pairings may be tolerated in some cases. Because the SNP density analysis supported the target model, we concentrated on SNPs in bases 2–7 of miRNA sites, which are believed to be the most important bases for miRNA target recognition[3].

As SNP density is sensitive to mutation rate heterogeneity, we corroborated our findings using derived allele frequency (DAF) data (Methods), which is robust to these effects[9]. Given a set of SNPs, an excess of rare derived alleles is a signature of negative selection. Because ascertainment bias with SNP data is well known[10], we avoided methods that compare SNPs with neutral models (such as Tajima's *D*) and instead compared different DAF distributions directly using a

[1]Center for Comparative Functional Genomics, Department of Biology, New York University, New York, New York 10003, USA. [2]Max Delbrück Centrum für Molekulare Medizin, Robert-Rössle-Strasse 10, Berlin-Buch, Germany. Correspondence should be addressed to N.R. (rajewsky@mdc-berlin.de).

Received 6 July; accepted 21 September; published online 29 October 2006; doi:10.1038/ng1910

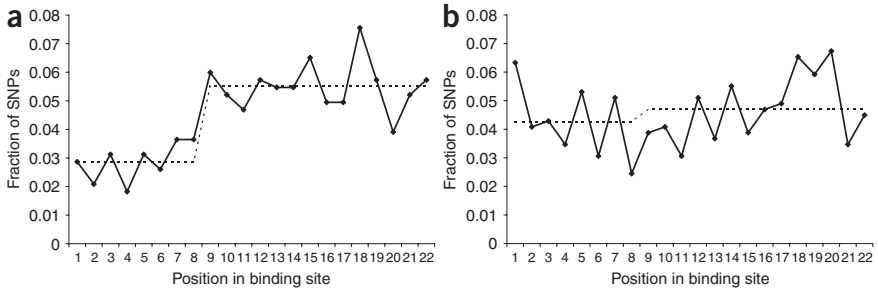


**Figure 1** SNP density in conserved miRNA sites. (**a**,**b**) SNP density in (**a**) predicted conserved miRNA binding sites and (**b**) conserved random controls. The dotted lines track the average SNP density in the region matching the 5′ end of the miRNA (bases 1–8; region 1) and the rest of the binding site (bases 9–22; region 2). The significant difference in SNP density between regions 1 and 2 of conserved miRNA binding sites ($\chi^2$, $P = 5 \times 10^{-5}$), when compared with the much smaller difference between regions 1 and 2 of conserved random controls, supports the target prediction model ($\chi^2$, $P = 0.0004$).

Mann-Whitney test[11]. This approach is still sensitive to ascertainment bias between functional classes (for example, synonymous and nonsynonymous sites), so we focused on SNPs in 3′ UTRs, for which ascertainment bias should be uniform.

In the HapMap data, we found that the DAF distribution of these SNPs was significantly skewed toward rare alleles in all three populations relative to synonymous sites, nonsynonymous sites, 3′ UTRs and conserved 7-mers in 3′ UTRs (*P* values in **Table 1**; **Fig. 2**). In the Perlegen data, we found similar trends (*P* values in **Table 1**; **Supplementary Fig. 1** online), although statistical significance is lower, probably owing to the smaller number of SNPs. Although nonselective forces such as demography, drift and hitchhiking can affect allele frequencies[9], they should affect all functional classes equally. Thus, the DAF analysis is indicative of stronger negative selection acting on predicted conserved miRNA sites than on other major functional classes. This includes other conserved 7-mers in 3′ UTRs (many of which should be functional) which should have the same ascertainment bias as conserved miRNA sites.

Polymorphism data are generally informative only of weak selective effects in recent evolution (~80,000 years in the human lineage), whereas divergence data are potentially informative of stronger selective effects and more distant evolutionary events (~6 million years in the human lineage)[12]. Thus, we used the McDonald-Kreitman test[13] which compares levels of intraspecies polymorphism and interspecies divergence in two classes of sites assumed *a priori* to experience different selective intensities. Weakly deleterious mutations contribute more to polymorphism than to divergence, so an excess of polymorphism is consistent with negative selection[13]. When comparing conserved miRNA sites with either 3′ UTRs or conserved 7-mers in 3′ UTRs, we found more polymorphism than divergence in conserved miRNA sites, suggesting stronger negative selection on conserved miRNA sites, although this result was not statistically significant (**Supplementary Methods** online). This result can be reconciled with the previous analyses by hypothesizing that some miRNA sites are either neutrally or adaptively evolving, thus inflating the amount of divergence and reducing the statistical power to detect negative selection.

To test this hypothesis, we turned to the Poisson Random Field (PRF) framework[14]. We used a likelihood approach to fit a PRF mixture model with two classes of sites, one with selection coefficient $\leq 0$ and the other with selection coefficient $\geq 0$, to the DAF data. We used formal likelihood ratio tests to show that this model fits the data significantly better ($\chi^2$, 1 degree of freedom, $P < 1.2 \times 10^{-7}$) than one- or three-class mixture models (**Supplementary Methods**). We then used the two-class PRF mixture model to estimate the fraction of functional and nonfunctional miRNA sites. The maximum likelihood fit was achieved for a model in which 85% of the conserved miRNA sites are under negative selection (in other words, are functional) (**Supplementary Methods**), which is a higher estimate than previous comparative genomic analyses[3] but is concordant with experimental evidence from *Drosophila melanogaster*[15].

Next we extended our analysis to nonconserved miRNA sites by predicting ~600,000 miRNA sites genome-wide for all 328 known human miRNAs[16] (Methods). Several groups have suggested that many of nonconserved sites are functional when the mRNA is expressed in the same tissue as the miRNA and that there is widespread selection for mRNAs ('anti-targets') to avoid miRNA sites that would otherwise be detrimental[3,15,17–19]. We investigated both gains and losses of nonconserved miRNA sites genome-wide but did not find any evidence that they are evolving differently than 3′ UTRs overall in all populations when analyzing either the HapMap or Perlegen DAF distributions (one-sided Mann-Whitney, $P > 0.16$) or the McDonald-Kreitman test ($\chi^2$, $P > 0.48$).

However, many predicted nonconserved miRNA sites may not be endogenous targets because the mRNA and miRNA are not expressed in the same spatiotemporal domain. Thus, we focused on a smaller set of ~12,000 nonconserved miRNA sites in mRNAs expressed in the same tissue as the miRNA ('coexpressed' miRNA sites; see Methods). In this set, we found that the DAF analysis was indeed consistent with negative selection against the loss of nonconserved miRNA sites when compared with 3′ UTRs overall in the HapMap data ($P < 0.02$) (**Fig. 3**), and the Perlegen data showed a similar trend that was less significant ($P < 0.08$), probably owing to the smaller number of SNPs (**Supplementary Fig. 2** online). We used the same PRF approach as before to estimate that 30%–50% of

**Table 1 Derived allele frequency analysis of SNPs in conserved miRNA sites**

| Data set | Population | Synonymous | Nonsynonymous | 3′ UTR | 3′ UTR conserved 7-mers | Number of SNPs |
|---|---|---|---|---|---|---|
| HapMap | European | $8.5 \times 10^{-7}$ | 0.0001 | 0.0001 | 0.004 | 48 |
| | Chinese + Japanese | $2.8 \times 10^{-7}$ | 0.0001 | 0.001 | 0.03 | 50 |
| | Yoruban (Nigerian) | $1.7 \times 10^{-5}$ | 0.004 | 0.0005 | 0.03 | 47 |
| Perlegen | European | **0.07** | **0.2** | 0.04 | **0.17** | 35 |
| | Chinese | 0.01 | 0.05 | 0.01 | 0.05 | 35 |
| | African American | 0.003 | 0.04 | 0.002 | 0.04 | 35 |

*P* values in a one-sided Mann-Whitney U test comparing the derived allele frequency (DAF) distribution of SNPs in conserved miRNA sites with the distribution of SNPs in other major functional classes of nucleotides. Boldface indicates $P > 0.05$. The last column refers to the number of SNPs in conserved miRNA sites. The number of SNPs in 3′ UTR conserved 7-mers is >989 in all populations, and the number of SNPs in the synonymous, nonsynonymous and 3′ UTR classes is >3,421 in all populations.

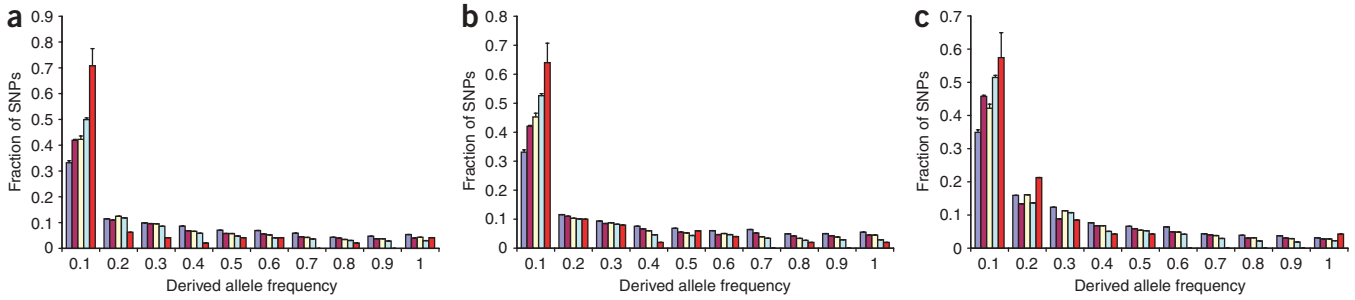


**Figure 2** DAF distributions in conserved miRNA sites suggests stronger negative selection compared with other conserved 7-mers in 3′ UTRs. (**a**–**c**) Derived allele frequency (DAF) distributions of HapMap SNPs in conserved miRNA sites in (**a**) Europeans, (**b**) Asians (Chinese and Japanese) and (**c**) Yorubans. An excess of derived alleles with low frequencies is indicative of weak negative selection. Dark blue: synonymous sites, purple: 3′ UTRs, yellow: conserved 3′ UTR 7-mers, light blue: nonsynonymous sites, red: conserved miRNA sites. Error bars represent s.d. from 1,000 independent bootstrap replicates.



coexpressed sites are functional (**Supplementary Methods**), consistent with previous estimates[17] and other results[18,19].

We also observed greater negative selection against gaining an miRNA site than on 3′ UTRs overall in the Yoruban HapMap population (one-sided Mann-Whitney, $P = 0.027$), although the evidence was weaker or not present in the other populations (**Fig. 3**). Although the general concordance between the HapMap (**Fig. 3**) and Perlegen data (**Supplementary Fig. 2**) for the patterns observed in Europeans, Africans and Asians suggests a true underlying biological cause, with the current data, we conclude that there is evidence for selection against gaining sites ('anti-targets'), but it is weaker than that for selection against losing sites. One possibility is that anti-targets are not as common as true targets. Another possible explanation stems from the difficulty of defining 'nonconservation', as sites could be evolutionarily conserved but not detected as such owing to technical reasons (such as misalignment), biological reasons (such as genome rearrangements), or target model insufficiencies[3]. For example, for 29% of the 'nonconserved' human sites, a motif complementary to bases 2–7 of the miRNA appears somewhere in the orthologous 3′ UTR of each of the other four species. Thus, it is possible that fewer than 30%–50% of true evolutionarily nonconserved sites are functional. This issue should be revisited when better expression and SNP data are available.

Finally, we screened for recent positive selection on miRNA sites (within the last 50,000–75,000 years[12]) by identifying SNPs in conserved miRNA sites whose allele frequencies are highly differentiated in Africans, Asians and Europeans, as measured by the $F_{st}$ parameter (see Methods). Although some variation in $F_{st}$ is expected from nonselective processes, an unusually high $F_{st}$ suggests recent local selection. This analysis does not necessarily pinpoint the causative allele but localizes the putative selective event to a genomic region. We restricted our search to SNPs with global heterozygosity $>0.1$, as these are the only SNPs informative of population differentiation. Out of 25 informative SNPs, we identified a single SNP (rs1054528), located in a conserved binding site of *miR-204* and *miR-211* in the *Map1lc3b* gene, with an $F_{st}$ higher than 99.8% of SNPs in 3′ UTRs with the same heterozygosity. The site is present in 87% of Africans but absent in almost all Europeans and Asians. Post-transcriptional misregulation of this gene has been implicated in giant axonal neuropathy[20] and Fragile X syndrome[21], and the latter has been linked to miRNAs[22]. We note that although positive selection is generally rare in species with small effective population sizes, like humans, it may be easier to detect positive selection on miRNA sites in species with larger effective population sizes, like *Drosophila*.

In conclusion, we have used human SNP genotype data to show that there is significant negative selection acting on a large class of computationally predicted conserved miRNA target sites and to estimate that ~85% of them are functional. We have also used this approach to estimate that 30%–50% of nonconserved miRNA sites in human 3′ UTRs are functional when expressed in the same tissue as the miRNA. These estimates were derived using the PRF model, which depends crucially on the assumption of independent evolution between SNPs. This assumption may be more reasonable for genotyped SNP data and short *cis*-regulatory sites for many *trans*-acting

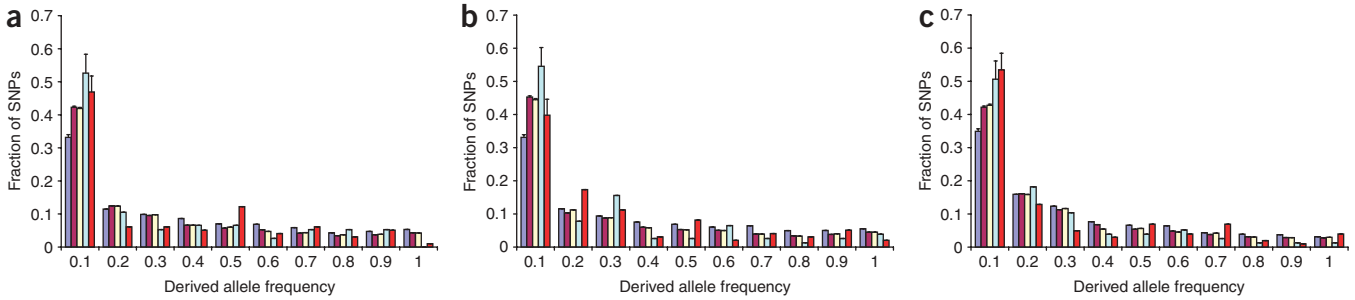


**Figure 3** DAF distributions in nonconserved miRNA sites coexpressed with the miRNAs. (**a**–**c**) DAF distributions of HapMap SNPs in coexpressed miRNA sites in (**a**) Europeans, (**b**) Asians (Chinese and Japanese) and (**c**) Yorubans. An excess of polymorphisms with low frequencies is indicative of weak negative selection. Dark blue: synonymous sites, purple: 3′ UTRs, yellow: gained or lost nonconserved miRNA sites, light blue: coexpressed nonconserved miRNA sites being lost in humans, red: coexpressed nonconserved miRNA sites being gained in humans. Compared with 3′ UTRs, the significant excess of low-frequency derived alleles in lost coexpressed nonconserved sites indicates weak negative selection against losing such miRNA sites in humans. Support for negative selection against gaining such miRNA sites is weaker. Error bars represent s.d. from 1,000 independent bootstrap replicates.

**Table 2 Summary of different classes of miRNA sites used in the analyses**

| Classes of miRNA sites | Description |
|---|---|
| Conserved sites | miRNA sites conserved in human, chimp, mouse, rat and dog for 62 conserved miRNAs |
| Conserved sites lost in humans | miRNA sites conserved in chimp, mouse, rat and dog but lost (that is, not segregating) in humans for 62 conserved miRNAs |
| Nonconserved sites | miRNA sites in human for 328 human miRNAs, minus all conserved miRNA sites or redundant sites |
| Coexpressed sites | Nonconserved sites that fall in mRNAs and are expressed in the same tissue as the miRNA |

All classes of miRNA sites are predicted computationally as described in Methods.



factors dispersed across the genome than for polymorphisms in the coding sequences of one or a few genes, to which the PRF model is traditionally applied.

Furthermore, as we have shown that mutations in either conserved or coexpressed miRNA sites are often deleterious, they are good candidates for causal variants in disease mapping studies. All SNPs in miRNA sites of various classes (**Table 2**) and the associated miRNA sites are available in the **Supplementary Note** and **Supplementary Tables 1–3** online, and SNP data and miRNA target predictions are available from the UCSC genome browser[23]. We also note that our conservative definition of a polymorphic site excludes other mutations that affect miRNA regulation, including SNPs in other regions of the site[5] and SNPs in miRNA genes themselves[24,25], which should be interesting for future investigations.

Finally, the approach taken in this study of analyzing genome-wide polymorphism in *cis* elements is not specific to miRNAs, SNPs or humans but rather has the potential to be more widely used in future investigations of other classes of *cis* elements and polymorphisms and other species.

## METHODS

**Alignment data.** We downloaded human RefSeq transcript annotations, human consensus CDS (CCDS) annotations and 17-way vertebrate multiZ alignments from the University of California, Santa Cruz (UCSC) browser[23], discarding transcripts that did not map uniquely to the human genome (release hg17). We assembled the multiZ alignments into contiguous multiple alignments, as described previously in ref. 26, and discarded all species except for the human, chimp, rhesus macaque, mouse, rat and dog. We removed insertions or deletions in the alignment caused by other species not considered in the analysis.

**SNP data.** We downloaded human SNP data (dbSNP build 125) from the UCSC genome browser[23] and discarded all insertion and deletion polymorphisms, SNPs with more than two alleles, SNPs monomorphic (that is, having only one allele) in all populations and SNPs that did not map uniquely to the human genome (release hg17). We downloaded genotype data from the International HapMap Project[1] (release 20) and Perlegen Sciences[2]. HapMap SNPs were genotyped in 90 Yoruban individuals in Ibadan, Nigeria; 44 Japanese in Tokyo; 45 Han Chinese in Beijing and 90 CEPH individuals (Utah residents with ancestry from northern and western Europe). Following common practice, we combined the Japanese and Chinese populations into one population. Perlegen SNPs were genotyped in 23 African Americans, 24 Han Chinese and 24 European Americans. For each of the three groups (Africans, Europeans and Asians), we were able to map ~21,000 HapMap SNPs and ~14,000 Perlegen SNPs to 3′ UTRs, for a total of ~25,000 genotyped SNPs in 3′ UTRs.

**Expression data.** We compiled miRNA expression data comprising 31 miRNAs and 45 miRNA-tissue assignments (**Supplementary Table 4** online) from ref. 17 and data described in ref. 27. We obtained mRNA expression data from ref. 28 and processed it as previously described in ref. 27. An miRNA and an mRNA are defined as 'coexpressed' if the mRNA is expressed at a level >1.2 times the median expression level in the tissue of miRNA expression. The trends reported in the main text are robust for cutoffs in the range 1.0–1.3. At lower cutoffs, the signal is not discernible above the noise, and at higher cutoffs, there is too little data for statistical analysis (data not shown).

**Conserved miRNA binding site predictions.** For the SNP density analysis, we use the term 'miRNA site' to refer to the entire 22-nt site in the 3′ UTR complementary to the miRNA. For the remainder of the analysis, we use the term 'miRNA site' to refer a 6-mer in a 3′ UTR with exact Watson-Crick complementarity to bases 2–7 from the 5′ end of the mature miRNA. We predicted miRNA targets by searching for such 6-mers exactly conserved in human, chimpanzee, mouse, rat and dog, requiring an additional match to either base 1 or 8 of the mature miRNA in each species (not necessarily the same base in all species). This method is similar to the core PicTar algorithm[26], but it ignores the relatively small number of predicted imperfect binding sites. In total, we predicted ~22,000 unique conserved miRNA sites using a set of 62 confidently defined vertebrate-conserved miRNAs[26]. For SNPs in binding sites, we predicted a binding site if either allele of the SNP caused a binding site.

Similarly, we define a conserved 7-mer to be a stretch of seven bases in the human genome that is exactly conserved in the same five species. Overlapping conserved 7-mers are allowed, and, again, for SNPs in conserved 7-mers, both alleles of the SNP were considered. This definition is conservative because it probably includes sites complementary to miRNAs not in our set.

**Nonconserved miRNA binding site predictions.** We predicted ~626,000 total miRNA sites in the human genome for all 328 miRNAs in Rfam 8 with no conservation filter[16]. We defined a set of 'nonconserved sites' by taking this set and removing all conserved miRNA sites. We also remove nonconserved miRNA sites for a particular miRNA that fall into the same 3′ UTR as a conserved miRNA site for the same miRNA, as selective pressure on these redundant sites is not expected to be the same as that on nonredundant sites. For example, if a 3′ UTR contains a conserved site for let-7 and nonconserved sites for let-7 and miR-1, we remove both let-7 sites for the nonconserved analysis but retain the miR-1 site. After we filtered by expression, 4,764 mRNAs remained. All conserved and coexpressed miRNA sites are given as **Supplementary Tables 1** and **2**. Instructions for uploading these data into the UCSC Genome Browser are given in the **Supplementary Note**.

**SNP density analysis.** For the SNP density analysis, we combined all SNPs genotyped by either HapMap or Perlegen in each of three groups: Asians (HapMap Chinese and Japanese and Perlegen Chinese), Africans (HapMap Yoruban and Perlegen African American) and Europeans (HapMap European and Perlegen European American). For the random controls in **Figure 1b**, we took all 62 conserved miRNAs and shifted each one by 6, 8 and 10 bases, and we used the resulting sequences with the above miRNA site prediction algorithm to predict conserved control binding sites, averaging the results of the three trials.

**McDonald-Kreitman analysis.** A full presentation and discussion of this analysis is provided in the **Supplementary Methods**.

**Derived allele frequency (DAF) analysis.** For each SNP, we defined the 'ancestral allele' as the human allele equal to the chimpanzee allele at the aligned base and the 'derived allele' as the other human allele. If the chimpanzee allele did not match either human allele, we used the rhesus macaque allele as the outgroup, and if this also did not match one of the human alleles, the SNP was discarded for the DAF analysis. We were able to define a derived allele for >98% of genotyped SNPs in 3′ UTRs. For miRNA sites containing a SNP, we inferred a gain of a site if the derived allele created a binding site and a loss if the ancestral allele created a binding site. The error bars in **Figures 2** and **3** represent s.d. from 1,000 independent bootstrap replicates.

**Poisson random field analysis.** A full presentation and discussion of this analysis is provided in the **Supplementary Methods**.

***F*st analysis.** The inbreeding coefficient, $F_{st}$, is commonly used as a measure of the differentiation of allele frequencies between different populations[9]. It can be interpreted as the probability that two alleles are identical by descent. We estimated $F_{st}$ only for SNPs with genotype data in all three groups: Africans, Asians and Europeans. To correct for small sample sizes, we estimated $F_{st}$ using the method of ref. 29. For SNPs genotyped by both projects, we combined all the data into the three groups (Europeans, Asians and Africans) because the allele frequencies in both projects are very similar (>10,449 SNPs jointly genotyped, Pearson's $R > 0.97$ in all three groups). As $F_{st}$ is positively correlated with predicted heterozygosity (that is, $2pq$ under Hardy-Weinberg equilibrium), we binned the SNPs in increments of 0.05 in predicted heterozygosity following ref. 30 and computed the empirical rank for the SNPs in each bin separately.

**URLs.** International HapMap Project, http://www.hapmap.org; Perlegen Sciences, http://genome.perlegen.com.

*Note: Supplementary information is available on the Nature Genetics website.*

**ACKNOWLEDGMENTS**

We thank P. Andolfatto, R. Borowsky, E. Halperin, N. Hübner and M. Siegal for helpful discussions. We also thank E. van Nimwegen and R. Nielsen for critical readings of a preliminary version of the manuscript. This research was supported in part by the Howard Hughes Medical Institute grant through the Undergraduate Biological Sciences Education Program to New York University.

**COMPETING INTERESTS STATEMENT**

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/



1. International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
2. Hinds, D.A. *et al.* Whole-genome patterns of common DNA variation in three human populations. *Science* **307**, 1072–1079 (2005).
3. Rajewsky, N. microRNA target predictions in animals. *Nat. Genet.* **38**, s8–13 (2006).
4. Bartel, D.P. MicroRNAs: genomics, biogenesis, mechanism and function. *Cell* **116**, 281–297 (2004).
5. Abelson, J.F. *et al.* Sequence variants in SLITRK1 are associated with Tourette's syndrome. *Science* **310**, 317–320 (2005).
6. Clop, A. *et al.* A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. *Nat. Genet.* **38**, 813–818 (2006).
7. Fairbrother, W.G., Holste, D., Burge, C.B. & Sharp, P.A. Single nucleotide polymorphism-based validation of exonic splicing enhancers. *PLoS Biol.* **9**, E268 (2004).
8. Drake, J.A. *et al.* Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nat. Genet.* **38**, 223–227 (2006).
9. Hartl, D.L. *A Primer of Population Genetics* (Sinauer, Sunderland, Massachusetts, 2000).
10. Clark, A.G., Hubisz, M.J., Bustamante, C.D., Williamson, S.H. & Nielsen, R. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res.* **15**, 1496–1502 (2005).
11. Akashi, H. Inferring the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination. *Genetics* **151**, 221–238 (1999).
12. Sabeti, P.C. *et al.* Positive natural selection in the human lineage. *Science* **312**, 1614–1620 (2006).
13. McDonald, J.H. & Kreitman, M. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* **351**, 652–654 (1991).
14. Sawyer, S.A. & Hartl, D.L. Population genetics of polymorphism and divergence. *Genetics* **132**, 1161–1176 (1992).
15. Stark, A., Brennecke, J., Bushati, N., Russell, R.B. & Cohen, S.M. Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3′UTR evolution. *Cell* **123**, 1133–1146 (2005).
16. Griffiths-Jones, S. *et al.* Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **33**, D121–D124 (2005).
17. Farh, K.K. *et al.* The widespread impact of mammalian microRNAs on mRNA repression and evolution. *Science* **310**, 1817–1821 (2005).
18. Krutzfeldt, J. *et al.* Silencing of microRNAs *in vivo* with 'antagomirs'. *Nature* **438**, 685–689 (2005).
19. Giraldez, A.J. *et al.* Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. *Science* **312**, 75–79 (2006).
20. Allen, E. *et al.* Gigaxonin-controlled degradation of MAP1B light chain is critical to neuronal survival. *Nature* **438**, 224–228 (2005).
21. Lu, R. *et al.* The fragile X protein controls microtubule-associated protein 1B translation and microtubule stability in brain neuron development. *Proc. Natl. Acad. Sci. USA* **101**, 15201–15206 (2004).
22. Jin, P., Alisch, R.S. & Warren, S.T. RNA and microRNAs in fragile X mental retardation. *Nat. Cell Biol.* **6**, 1048–1053 (2004).
23. Karolchik, D. *et al.* The UCSC Genome Browser Database. *Nucleic Acids Res.* **31**, 51–54 (2003).
24. Gottwein, E., Cai, X. & Cullen, B.R. A novel assay for viral microRNA function identifies a single nucleotide polymorphism that affects Drosha processing. *J. Virol.* **80**, 5321–5326 (2006).
25. Iwai, N. & Naraba, H. Polymorphisms in human pre-miRNAs. *Biochem. Biophys. Res. Commun.* **331**, 1439–1444 (2005).
26. Krek, A. *et al.* Combinatorial microRNA target predictions. *Nat. Genet.* **37**, 495–500 (2005).
27. Sood, P., Krek, A., Zavolan, M., Macino, G. & Rajewsky, N. Cell-type-specific signatures of microRNAs on target mRNA expression. *Proc. Natl. Acad. Sci. USA* **103**, 2746–2751 (2006).
28. Su, A.I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA* **101**, 6062–6067 (2004).
29. Weir, B.S. & Cockerham, C.C. Estimating *F*-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370 (1984).
30. Rockman, M.V., Hahn, M.W., Soranzo, N., Zimprich, F. & Goldstein, D.B. Ancient and recent positive selection transformed opioid *cis*-regulation in humans. *PLoS Biol.* **3**, e387 (2005).