doi:10.1093/nar/gkr414

Inferring causative variants in microRNA target sites

Laurent F. Thomas[1,2,*], Takaya Saito[1] and Pål Sætrom[1,2,3,*]

[1]Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, N-7489 Trondheim, Norway, [2]Interagon AS, Laboratoriesenteret, NO-7006 Trondheim and [3]Department of Computer and Information Science, Norwegian University of Science and Technology, N-7489 Trondheim, Norway

Received December 21, 2010; Revised May 5, 2011; Accepted May 9, 2011

ABSTRACT

MicroRNAs (miRNAs) regulate genes post transcription by pairing with messenger RNA (mRNA). Variants such as single nucleotide polymorphisms (SNPs) in miRNA regulatory regions might result in altered protein levels and disease. Genome-wide association studies (GWAS) aim at identifying genomic regions that contain variants associated with disease, but lack tools for finding causative variants. We present a computational tool that can help identifying SNPs associated with diseases, by focusing on SNPs affecting miRNA-regulation of genes. The tool predicts the effects of SNPs in miRNA target sites and uses linkage disequilibrium to map these miRNA-related variants to SNPs of interest in GWAS. We compared our predicted SNP effects in miRNA target sites with measured SNP effects from allelic imbalance sequencing. Our predictions fit measured effects better than effects based on differences in free energy or differences of TargetScan context scores. We also used our tool to analyse data from published breast cancer and Parkinson's disease GWAS and significant trait-associated SNPs from the NHGRI GWAS Catalog. A database of predicted SNP effects is available at http://www.bigr.medisin.ntnu.no/mirsnpscore/. The database is based on haplotype data from the CEU HapMap population and miRNAs from miRBase 16.0.

INTRODUCTION

MicroRNAs (miRNAs) are small non-coding single stranded RNAs of about 22 nucleotides length that regulate genes post transcription by partially pairing with 3′-untranslated regions (3′-UTR) of messenger RNA (mRNA) (1). Watson–Crick pairing to nucleotides 2–7 of the 5′-end of microRNAs (seed sites) is known to be important in mRNA targeting. Specifically, miRNAs require almost perfect complementarity at seed sites for binding and reducing the protein levels of targets (2). However, mRNA sites with perfect complementarity to the seed nucleotides are not necessarily functional (3) and those with imperfect seed complementarity can also be functional (2). Consequently, considering seed sites alone gives many false positive miRNA target sites. Predictions can be improved, however, by using information about the target sites' context, such as their position within the 3′-UTR (4) and the distance to neighbouring sites (5), as such context is critical for target site functionality and efficacy.

Genome-wide association studies (GWAS) can identify genomic regions that contain genomic alterations, such as single nucleotide polymorphisms (SNPs), associated with common disease (6). The biological effects of identified alterations are usually not known, however, as few of the functional variants that show association in GWAS change the amino acid sequence. Moreover, a sizeable proportion is thought to reside in regulatory regions, since several associated regions found in GWAS lack known genes (7). Variants in regulatory regions can, for example, result in altered protein levels, so identifying and understanding their effects can improve diagnostics and treatments for diseases (8). Specifically, SNPs in regulatory elements such as miRNA target sites can affect phenotype (9) and have been associated with increased cancer risk (10) and other diseases (11). The increased use of GWAS to study genetic factors in common disease necessitates a tool that can identify and interpret effects of regulatory variants.

Several research groups have tried to look at regulatory variant effects. Bao *et al.* (12) looked for SNPs in putative conserved miRNA target sites [from the target site prediction tool TargetScan (13)], and integrated such SNP sites with phenotype (physiological and behavioural traits

*To whom correspondence should be addressed. Fax: +47 72571463; Email: laurent.thomas@ntnu.no
Correspondence may also be addressed to Pål Sætrom. Tel: +4798203874; Email: pal.satrom@ntnu.no

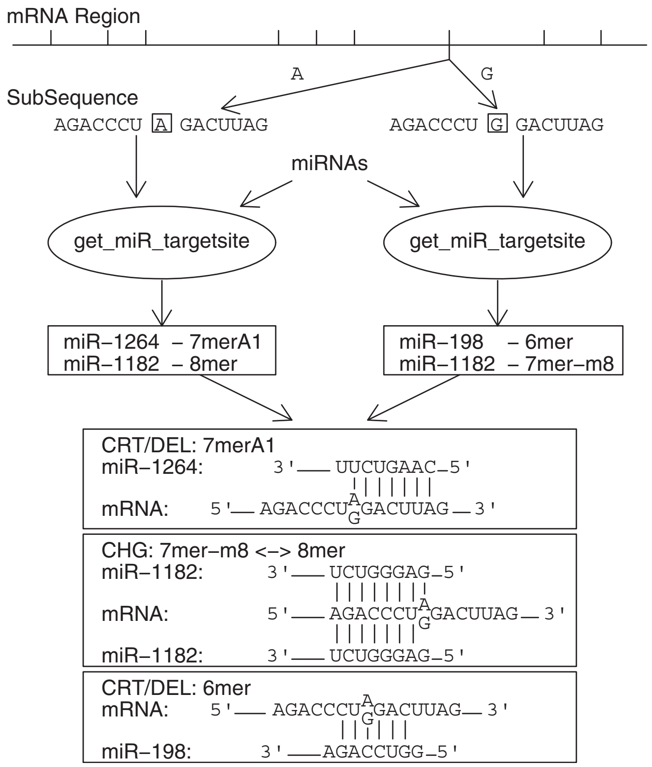

Figure 1. Identifying SNPs in miRNA target sites. The illustration shows an mRNA region that contains SNPs represented by small vertical lines. The considered SNP has two alleles: A and G. We make one subsequence for each allele by using the flanking regions of the SNP (7 nucleotides on each side). Given miRNA seed motifs (nucleotides 2–8 from the 5′-end of miRNA sequences), we look for target sites in each allele sequence and then compare results to characterise the effect of the SNP (create/delete (CRT/DEL) target sites, or change (CHG) site type).

of mice as quantitative trait loci) and expression data (of mice and human transcripts) into a database. However, the studied phenotypes only concern physiology of mice instead of human diseases. Georges *et al.* (14) also made a database with SNPs in putative miRNA target sites [regulatory motifs identified in (15) and predicted sites from (13)], but Georges *et al.* (14) did not map their site SNPs to phenotypes, except for one SNP in sheep. Barenboim *et al.* (16) developed an online tool that finds SNPs in microRNA target sites on the fly. The tool takes haplotype into account, but is limited to one single gene and six SNPs per run and does not quantify SNP effects. Nicoloso *et al.* (17) used the miRanda tool (18) to identify breast cancer-associated SNPs that disrupt miRNA target sites. The authors filtered SNPs based on minimum free energy (MFE) and tested the remaining ones in a case-control study.

A basic way of detecting SNPs in microRNA target sites (mirSNPs) in a gene g, starts by looking at SNPs lying in a region of interest, such as 3′-UTR, 5′-UTR, coding or promoter region (Figure 1). Here, we will use the 3′-UTR as an example, since SNPs affecting miRNA target sites are more likely to reside in the 3′-UTR (19,20). Let us consider a SNP s in this region of interest. The SNP s has several alleles, usually two, that we want to evaluate for targeting by a microRNA seed motif m. Specifically, for each allele a_i, we determine whether there is a microRNA target site in a sequence als_i consisting of the allele a_i and its flanking sequences. Target sites are detected by using any miRNA target site prediction tool based on sequence search. It is convenient to disregard target sites with mismatches in the seed region and only consider 6-mer, 7-mer and 8-mer seed sites. For each allelic sequence als_i, we get a list l_i of target sites for microRNA m. We can then compare these lists to determine if a target site is created, deleted, or changed between the alleles (Figure 1).

All existing tools use variants of the approach above of evaluating candidate sites individually (Figure 1), but this approach ignores that 3′-UTRs can contain multiple linked SNPs that can affect miRNA targeting by altering site context. Instead, we propose to analyse all the SNPs of the 3′-UTR at the same time, to have a general overview of the SNPs' regulatory effect on the considered mRNA.

In this article, we present a computational tool that can help identifying SNPs causative to diseases, such as cancer. The tool focuses on SNPs that may affect miRNA targeting and thereby cause gene dysregulation. More precisely, the tool predicts the effects of SNPs in miRNA target sites and uses linkage disequilibrium to map those mirSNPs to SNPs of interest in GWAS. We show that the tool's predictions correspond well to the SNP's measured effects on miRNA regulation, and that the predictions correlate better to those effects than do the predictions of other existing tools. We further demonstrate the tool's utility by analysing two published GWAS data sets and specific SNPs reported to affect miRNA targeting.

MATERIALS AND METHODS

The following sections will present a method that uses context-based miRNA target prediction to quantify the effects of SNPs in miRNA target sites (mirSNPs) and uses linkage disequilibrium to map candidate mirSNPs to disease data from GWAS. The tool allows additional filtering of candidate genes and candidate miRNAs. The tool's mapping method is general and can therefore be applied to SNPs independent of the scoring method used.

Data

We used the SNP data from the human haplotype map project [HapMap, (21)]; particularly, SNP data from the CEU population (CEPH - Utah residents with ancestry from northern and western Europe), release 22 for haplotype data, and release 27 for linkage disequilibrium data. We used DNA sequences from the human and mouse genome assemblies hg18 and mm9 (22,23). SNPs and Gene annotations (hg18,mm9) came from UCSC Genome browser (24). MicroRNA sequences came from miRBase, release 13.0 and 16.0 (25). GWAS data were from a breast cancer study from Cancer Genetic Markers of Susceptibility (CGEMS) (26), from a Parkinson disease study (*P*-values from tier 1) (27), and from the NHGRI GWAS catalog (28) (http://www.genome.gov/gwastudies).

MicroRNA regulation score of haplotypes

To analyse all the SNPs of the 3′-UTR at the same time, we use population haplotype data for the 3′-UTR (Figure 2 and Supplementary Figure S1). Specifically, we first use haplotype data to build haplotype sequences hs_i; i.e. 3′-UTR sequences containing the combinations of alleles found in the considered population. Second, for a given miRNA m, we use a miRNA target prediction tool (29) to score each haplotype sequence hs_i. The prediction tool uses a two-step SVM classifier, where one SVM step classifies individual target sites and a subsequent SVM step classifies overall mRNA targeting potential. Features the SVM uses at the first step include seed pairing, 3′ supplementary pairing, the site's AU context and relative position in the 3′-UTR, and distance to neighbouring sites, whereas features at the second step include 3′-UTR length, the number and predicted strength of target sites, and the number of optimally spaced sites in the 3′-UTR (29). As output, the SVM-based prediction tool gives a score such that a high output score indicates that the miRNA m is likely to down-regulate this mRNA. Third, we compare the score-haplotype pairs to find the differences of haplotypes that can explain any differences of SVM scores. From the differences of haplotypes, we can make a list of candidate SNPs and predict their impact on gene regulation.

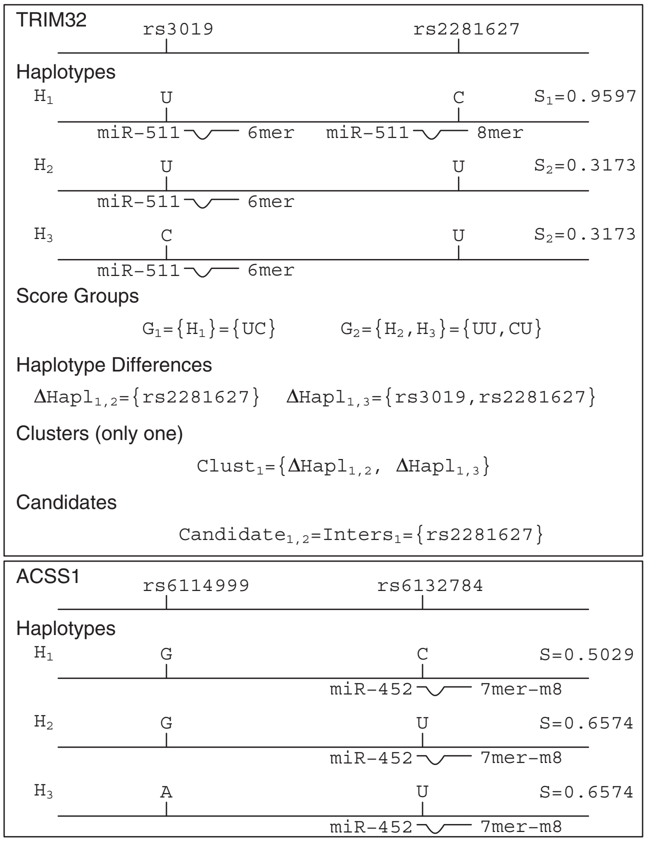

Figure 2. Scoring SNPs in miRNA target sites. rs3019 and rs2281627 are SNPs in the 3′-UTR of *TRIM32*. There are 3 different haplotypes in the CEU population: UC/UU/CU. *TRIM32* is targeted by miR-511, but the U allele of rs2281627 disrupts one seed site, which results in a lower score S_2 for the UU/CU haplotypes. To identify rs2281627 as the effect SNP, first the 3 haplotypes H_1, H_2 and H_3 are grouped by scores into G_1 and G_2. Second, we identify the differences between haplotypes from groups G_1 and G_2; i.e. differences between H_1 and H_2 and between H_1 and H_3. Third, we cluster those haplotype differences, so that the intersection within the cluster is not empty; here, there is only one cluster. Finally, we take the intersection of haplotype differences within this cluster, which gives the SNP rs2281627. Similarly, rs6114999 and rs6132784 lie in the 3′-UTR of *ACSS1*. There are 3 haplotypes: GC/GU/AU. Both SNPs lie outside of any seed sites of miR-452, but rs6132784 lies in a 3′-supplementary site and has a small effect on the scores.

The haplotype score comparison works as follows. First we group haplotypes H_i by scores, since we are interested in score differences:

$$G_s = \{H_i \in H \mid Score(H_i) = s\}.$$

Second, we look at the difference of haplotypes between groups, to identify which SNPs differ between two score groups: $\forall (G_m, G_n)$, $m \neq n$, $\forall H_i \in G_m$, $\forall H_j \in G_n$,

$$\Delta Haplo_{ij} = \{snp \mid H_i(snp) \neq H_j(snp)\}.$$

Third, we cluster the $\Delta Haplo$ SNP sets, to handle particular cases such as two SNPs in one target site (Supplementary Figure S2). Specifically, we cluster $\Delta Haplo$ sets such that in each cluster, the intersection of all the $\Delta Haplo_{ij}$ of the cluster is not empty:

$$Clust_k = \left\{\Delta Haplo_{ij} \middle| \bigcap \Delta Haplo_{ij} \neq \emptyset\right\}.$$

Fourth, we take the intersection of the $\Delta Haplo$ SNP sets in each cluster, to identify which SNP is responsible for the score difference in each cluster:

$$Inters_k = \bigcap Clust_k = \bigcap_{\Delta Haplo_{ij} \in Clust_k} \Delta Haplo_{ij}.$$

Finally, we merge all the clusters to create a list of SNPs responsible for the score difference for the clusters:

$$Candidate_{mn} = \bigcup_k Inters_k.$$

$Candidate_{mn}$ are candidate SNPs that might explain the difference between the scores m and n.

Normalization of target site scores

The miRNA target site prediction tool (29) predicts both the targeting potential of individual candidate sites and the total regulatory potential of candidate 3′-UTRs; i.e. if a gene's 3′-UTR sequence contains one or more candidate miRNA target sites, the tool scores the miRNA's regulatory effect on the target gene. However, the tool does not score mRNAs without target site candidates. Consequently, to score and compare scores for sequences with and without candidate sites, we needed to create a normalized score. The desired distribution should be mainly uniform, because the difference between two transformed scores should reflect a difference in percentiles in the original distribution. Since we only get scores for sequences with target sites, we had to find a way to score sequences that do not have target sites and to compare sequences with and without target sites. Our solution consisted of normalizing the scores in the interval [0, 1]. As there are more sequences without target sites than with target sites, we normalized scores so that the codomain of the normalization has an exponential distribution in [0, 0.01] and a uniform distribution in [0.01, 1], according to the following probability density function:

$$df(y) = \begin{cases} \lambda e^{-\alpha\lambda y} & y \in [0, \tau] \\ \frac{P_{Unif}}{1-\tau} & y \in [\tau, 1]. \end{cases}$$

Here, τ is the threshold that separates the two distributions in the codomain. To jointly score sequences with and without target sites, we considered sequences with only one target site as an intermediate. Since we needed to put the worst target site scores in the exponential part, we used the score distribution of mRNAs that have only one target site, which is a 6-mer. Specifically, we used the fifth percentile of the 6-mer distribution to define the threshold T: $P(X_{6m} < T) = 0.05$. This threshold then separated the exponential distribution from the uniform distribution in the domain of the normalization morphism. As a result, the exponential part contained scores for sequences that have no target site (TS) (including those with mismatch target sites) or canonical target sites with a score lower than T. The proportion of scores that will be in the uniform part is $P_{Unif} = P[X \geq T]P_{TS}$, where P_{TS} is the probability of having a target site and $P[X \geq T]$ is the proportion of scores greater than T. The proportion of scores in the exponential part is $P_{Exp} = 1 - P_{Unif}$. The parameter $\lambda = -\frac{1}{\alpha\tau}\log(1 - \alpha P_{Exp})$ makes the cumulative distribution of the exponential part fit P_{Exp}. The parameter $\alpha \in \left]0, \frac{1}{P_{Exp}}\right[$ makes the two distributions continuous in τ and minimizes

$$f(\alpha) = \left(-\frac{1 - \alpha P_{Exp}}{\alpha\tau}\log(1 - \alpha P_{Exp}) - \frac{P_{Unif}}{1-\tau}\right)^2.$$

We chose τ = 0.01 as a trade-off between τ being so small that all the scores from the exponential part had the same tendency, and being so large that we could find the α that minimized $f(\alpha)$.

Mapping candidate SNPs to disease

We can map candidate mirSNPs to disease by filtering on genes that are dysregulated in a given disease, filtering on miRNAs that are dysregulated in a given disease, and filtering on disease-associated SNPs from the same genomic region as the candidate. As filtering on genes or miRNAs simply involves focusing on subsets of the UTRs or miRNAs, we detail the filtering on disease-associated SNPs.

Association studies can show association of marker SNPs with a disease, but not necessarily association of a causal SNP with the disease. Consequently, if we want to know whether a candidate mirSNP may be causal, we first have to map it to associated marker SNPs.

Mapping candidate SNPs to association studies consists in looking for GWAS top ranking SNPs that have been inherited together with our candidate SNPs; i.e. looking for candidate SNPs that have alleles that correlate with alleles of associated marker SNPs. This can be achieved by computing inheritance blocks.

Inheritance blocks are DNA regions with highly correlated alleles. Consequently, by knowing the alleles of one SNP of the block one can predict the alleles at another SNP of the block. This measure of inheritance is called linkage disequilibrium (LD). Given a candidate SNP, we can compute its inheritance block, according to HapMap data. The block is an area of strong linkage disequilibrium and shows SNPs that have high correlation between themselves and with the candidate SNP.

We can define a block as a set of successive SNPs:

$$Block = \{s_l, \ldots, s_r\},$$

where s_l and s_r are the left and right bound SNPs of the block.

A block spine is a set of LD values:

$$Spine = \{D'_{lj}\} \cup \{D'_{ir}\},$$

such that $l < j \leq r$ and $l < i < r$ and where D'_{xy} is the linkage disequilibrium between the SNPs s_x and s_y. In short, the spine consists of the borders of the block (the two borders of the triangle block).

A solid spine is a spine where a relative amount α of the spine's LD values is below a threshold T. For example, we can use α = 10% and T = 0.8, to detect blocks with strong LD.

The block detection method (Figure 3) is called Solid Spine by Expansion and is an adaptation of the Solid Spine algorithm developed within the Haploview software (30). This expansion algorithm uses a candidate SNP as input. It starts the expansion from this SNP and then tries to expand the block successively in the downstream and upstream directions. An expansion occurs if the spine of the expanded block fits a rule depending on α and T. This algorithm needs an area of high LD to expand, which ensures that the algorithm returns few false positive blocks. The expansion can start on the left side as well as on the right side and the two directions can give different results. As we are interested in finding all SNPs that reside in blocks that have high LD with of the input SNP, we consider both resulting blocks.

Given a block of SNPs identified by the Solid Spine by Expansion algorithm above, we then extract GWAS top ranking SNPs from the block, to identify if the candidate SNP is correlated with any associated SNPs. We consider a SNP to be top-ranking when its rank is less than a given threshold.

We define three scores to assess the level of LD of the block defined by the candidate SNP and a top ranking SNP. The spine score is the mean of all LD values of the spine between the SNPs s_x and s_y:

$$Sc_{spine} = \frac{1}{2(y-x)-1}\left(\sum_{j=x+1}^{y} D'_{xj} + \sum_{i=x+1}^{y-1} D'_{iy}\right).$$

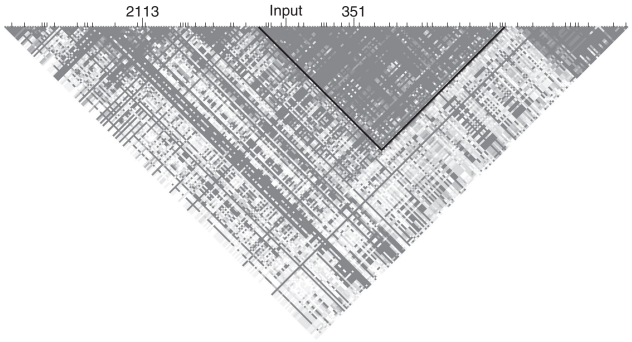

Figure 3. Example of a linkage disequilibrium block. Given an input SNP, we compute its linkage disequilibrium block (delimited by dark lines), and then look for top ranking SNPs in the block (here a SNP ranking as 351).

The triangle score is the mean of all LD values of the inner triangle between the SNPs s_x and s_y:

$$Sc_{triangle} = \frac{2}{(y-x)(y-x+1)}\left(\sum_{i=x+1}^{y-2}\sum_{j=i+1}^{y-1} D'_{ij}\right).$$

A block score is the sum of the spine score and the triangle score:

$$Sc_{block} = Sc_{spine} + Sc_{triangle}.$$

RESULTS

We first use data from allelic imbalance sequencing (31) to test our SNP scoring method and to compare our method with existing ones. Then we use two different GWAS data sets to evaluate the mapping method. Finally, we show that the method can find known altered miRNA targets associated with disease.

Scoring method predicts effects of mirSNPs

Kim and Bartel (31) used allelic imbalance sequencing to measure for three miRNAs, *in vivo* miRNA-directed repression at polymorphic target sites in mice. They provide allelic ratios (target versus non-target allele) $AR = \frac{|target\ allele|}{|non\ target\ allele|}$ for 65 SNPs in 3′-UTRs that create or disrupt miRNA target sites in tissues expressing (AR_E) and not expressing (AR_{NE}) the considered miRNA. We used 47 of these SNPs (those that have both allelic ratios AR_E and AR_{NE}) to test our method. For each of these 47 SNPs, we computed miRNA regulation scores for the target allele S_T and non-target allele S_{NT}. We compared the difference of our scores between the two alleles $\Delta S = S_T - S_{NT}$ with the difference of logarithms of allelic ratios $\Delta AR = \log_2(AR_{NE}) - \log_2(AR_E)$ (Figure 4) and found a clear and significant correlation (Pearson's correlation *P*-value 0.0025, Spearman's rank correlation *P*-value 0.00019).

In comparison, using MFE given by RNAhybrid 2.1 (32) to predict SNP effects gave insignificant correlations, whereas using TargetScan 5.0 context scores (13) (computed without taking conservation into account) gave significant but lower correlation (Table 1). Furthermore, our normalization method could improve the correlation based on TargetScan scores.

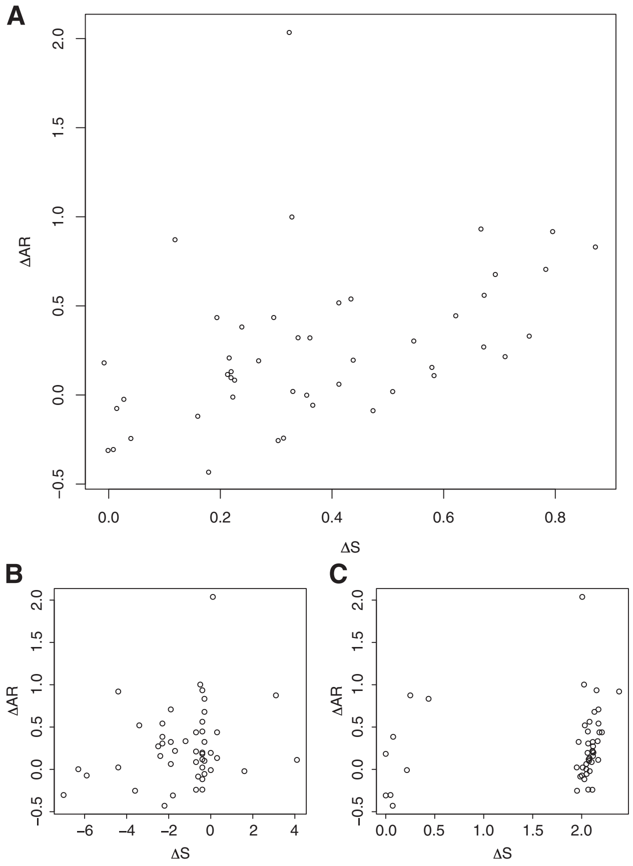

Figure 4. Predicted SNP effects correspond with observed effects. Correlation between the measured allelic ratio ΔAR and (**A**) the difference of our predicted allelic scores ΔS (with transformation), (**B**) MFE differences, and (**C**) TargetScan score differences (without transformation, but where the minimum TargetScan value represents the score for sequences without predicted target sites). See Table 1 for correlations and *P*-values.

This result suggests that our scoring method for SNP effects fits data from allelic imbalance sequencing better than TargetScan context scores (13) or changes in MFE [for example, used in (17)]. Our method therefore appears to be the best choice for predicting effects of SNPs in microRNA target sites.

ANALYSIS OF GWAS DATA

To generate a list of candidate SNPs involved in miRNA-based regulation, we computed differences of scores for all 3′-UTR haplotypes for all coding genes (UCSC RefSeq Genes hg18) and all miRNAs (from miRBase 13.0). Specifically, we analysed mRNAs that had more than 1 haplotype in their 3′-UTR (12 808 of the 26 963 coding transcripts) according to the CEU population from HapMap. Of the 12 808*698 = 89 39 984 mRNA/miRNA pairs, 396 851 had at least one haplotype score that differed from the other haplotype scores of the

Table 1. Correlations between the measured allelic ratio ΔAR and predicted SNP effects from several methods

Method	Pearson's corr.		Spearman's corr.	
	coeff.	*P*-value	coeff.	*P*-value
SVM (raw scores)	0.383	0.0079	0.507	0.00033
SVM (w/ transformation)	0.431	0.0025	0.524	0.00019
SVM (w/ transf, w/o 1 outlier)	0.562	$4.8*10^{-5}$	0.548	0.00010
MFE (no helix constraint)	0.223	0.1324	0.177	0.2345
MFE (helix constraint 2–7)	0.124	0.405	0.084	0.5736
TargetScan (raw scores)	0.168	0.2582	0.394	0.0062
TargetScan (w/ transformation)	0.299	0.0409	0.413	0.0039

same mRNA/miRNA pair. As explained in the methods, the haplotype score distribution has an exponential and a uniform part. Consequently, differences of scores also have a distribution with an exponential part, describing small differences in miRNA targeting. We used a threshold of 0.15 to filter out the exponential part. Of the 396 851 mRNA/miRNA pairs (which correspond to 401 983 ΔS values, as several mRNAs had several haplotype score differences), 55 707 pairs (60 751 ΔS values) had at least one $\Delta S > 0.15$. We selected the SNPs that generated a difference in score $\Delta S > 0.15$ as candidate SNPs (18 325 SNPs).

To further analyse the candidate mirSNPs, we mapped the mirSNPs to the breast cancer GWAS from CGEMS, as described in the methods. One would usually choose a high T threshold as parameter for the mapping method to identify blocks with high LD. We chose $T = 0$, however, to have data with low LD to analyse the block score variation in relation to the SNP and GWAS scores, as the block scores quantify the link between the candidate mirSNPs and the GWAS SNPs. We computed block scores for each pair of candidate SNP and top ranking SNP detected by the mapping method.

Top-ranking SNPs are likely in strong LD with their causative SNP. Consequently, we would expect that if mirSNPs are a significant factor behind the top-ranking CGEMS SNPs, high ΔS scores would be enriched among the highest scoring blocks. Since a candidate SNP can have several corresponding ΔS due to several miRNAs and transcripts, we assigned to each SNP its maximum ΔS value: ΔS_M. To test whether an increase in block score threshold between top-ranking SNPs and candidate SNPs causes any shift in the ΔS_M distribution, we computed the probability density of ΔS_M for different subsets of SNPs. These subsets were defined by a block score greater than a threshold, starting from all block scores and gradually reducing to only the best ones.

Figure 5 shows for SNPs mapped to the 2112 top-ranking CGEMS SNPs, the distributions of ΔS_M (from 0.15 to 1) for several subsets of SNPs based on different block score thresholds. The distributions show a shift of the main peak at $\Delta S = 0.33$ to $\Delta S = 0.53$ as the block score threshold increases. This shift is consistent with mirSNPs being significant causative factors behind the top-ranking CGEMS SNPs.

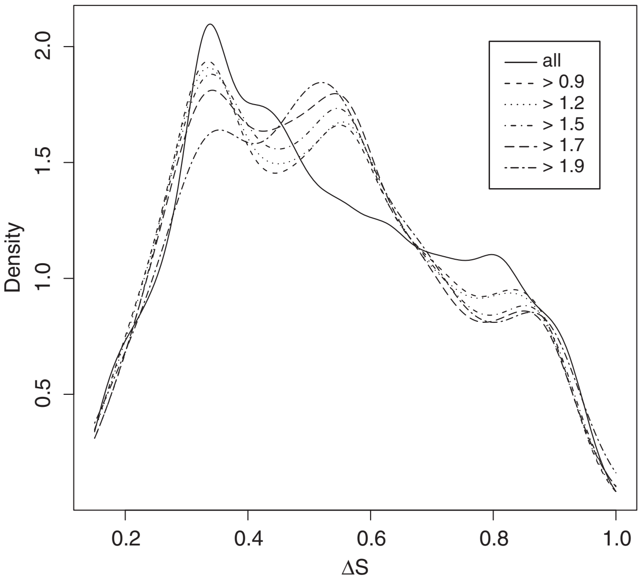

Figure 5. Distribution of mirSNP scores ΔS_M for SNPs mapped to high-ranking SNPs from the CGEMS breast cancer GWAS. ΔS_M is the maximum difference of scores for each SNP, where the scores are normalized scores from the SVM. Each curve shows the distribution for SNPs that have a block score greater than a given threshold. 'All' refers to ΔS_M of all SNPs. '>0.9' refers to ΔS_M of SNPs that have a block score >0.9 with one of the 2112 top-ranking CGEMS SNPs. The peak at 0.33 is decreasing as the block score threshold increases, whereas the peak at 0.53 is increasing with the block score threshold.

We would also expect that the shift will be less pronounced if we consider more candidate SNPs (by using a higher rank threshold on GWAS SNPs), as these SNPs will likely have a higher proportion of false positives. We therefore looked at different top-ranking thresholds to check that as the top-ranking threshold increases, the shift occurs later and later in terms of block score threshold. Figure 6A–D show 3D plots for top-ranking thresholds 528, 1056, 2112, and 4224. As in Figure 5, the plots show a shift of the main peak at $\Delta S = 0.33$ to $\Delta S = 0.53$ as the block score threshold increases.

The lower part of the plots shows all ΔS_M for all block scores—the background distribution of ΔS_M scores without taking LD into account. Increasing the block score threshold removes mirSNPs that are not linked to breast cancer-associated GWAS marker SNPs, thereby increasing the proportion of candidate mirSNPs that are associated with breast cancer. The shift in ΔS_M towards the right for high block score thresholds therefore shows that mirSNPs associated with breast cancer have a stronger effect on miRNA targeting than have the background of all mirSNPs.

As expected, increasing the threshold on top-ranking GWAS SNPs results in the shift occurring later and later on the *y*-axis. Using a higher top-ranking threshold gives a bigger proportion of false positive SNPs, whereas in contrast, a higher block score threshold gives a smaller proportion of false positives. Consequently, to compensate for the additional false positive SNPs that were added when increasing the rank threshold, a higher

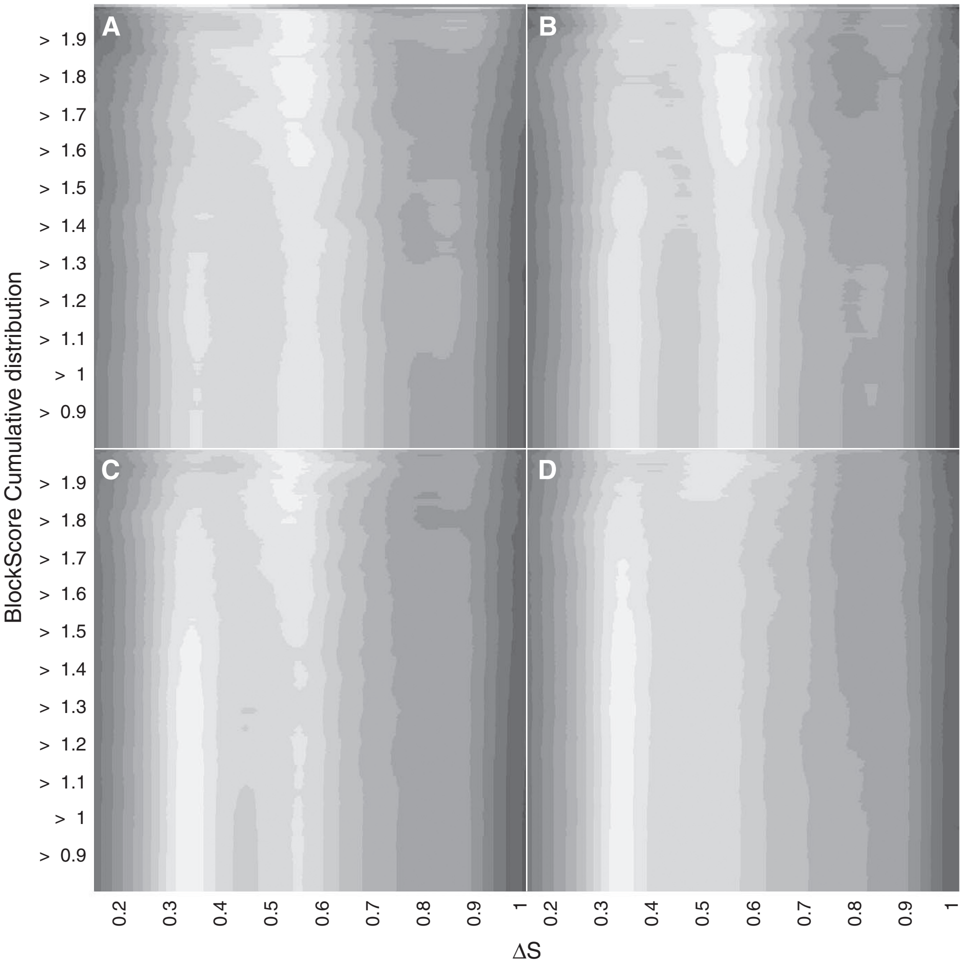

Figure 6. Distributions of ΔS_M for SNPs mapped to different numbers of high-ranking SNPs from the CGEMS breast cancer GWAS. The distributions vary with the number of candidate SNPs and block score thresholds. The graphs show ΔS_M on the x-axis (range [0.15, 1]), complementary cumulative distribution of block scores (from all block scores on the bottom, to gradually filtering to the best block scores on the top) on the y-axis, and density of ΔS_M for a given block score threshold (specifically, the distribution of ΔS_M for SNPs that have a block score > the value on the y-axis) on the z-axis (in grayscale). Dark grey, light grey and white are respectively low, intermediate, and high-density values. Panels (A), (B), (C) and (D) show 3D plots for top-ranking thresholds 528, 1056, 2112 and 4224, respectively. The plots show a shift of the main peak at $\Delta S_M = 0.33$ to $\Delta S_M = 0.53$, as the block score threshold increases.

block score threshold is needed to observe the shift in ΔS. These results indicate a link between high ΔS and high-block score top-ranking SNPs. Furthermore, the analyses give a good overview of how our predicted scores ΔS fit some GWAS data and show that our approach can identify SNPs in regulatory elements that may be causal in disease.

Using TargetScan's context scores (13) computed for all 3′-UTR haplotypes (without considering conservation), gave similar results indicating that the analysis is robust to the choice of prediction method (Supplementary Figures S3 and S4).

We also repeated the analysis on a GWAS for Parkinson's disease. This analysis gave similar results, indicating that the method works with other data sets and diseases (Supplementary Figures S5 and S6). Finally, we analysed the significant trait-associated SNPs from the NHGRI GWAS Catalog (28) and found a similar shift in the ΔS distribution at very high-block scores between miRSNPs and associated SNPs from caucasian-based studies (Supplementary Figure S7; see Supplementary Table S1 for the list of the best-scoring miRSNPs strongly linked to caucasian-based trait-associated SNPs). This result is consistent with us using Hapmap CEU haplotypes and linkage disequilibrium data for the analysis and indicates that miRSNPs explain some of the trait-associations in the NHGRI GWAS Catalog.

Disease-related examples

To further evaluate our methodology, we used it to analyse three miRNA/SNPs involved in breast cancer, asthma and Parkinson's disease.

Saetrom *et al.* (33) found that the SNP rs1434536 lies in the target site of the microRNA miR-125b within the gene *BMPR1b*, and is associated with breast cancer. In that study, we used the disease mapping method presented above to map the candidate SNP rs1434536 to the breast cancer GWAS from CGEMS. We computed the LD block of rs1434536, in which we found 5 SNPs that rank within the 500 best in the association study (ranks 67, 79, 291, 409 and 424) out of 528.000 SNPs; the candidate SNP lay in between the SNPs ranked 67 and 79 (Figure 7). The difference of scores for rs1434536 is 0.39. Saetrom *et al.* (33) verified that the SNP affects miR-125b's regulation of *BMPR1b* and verified the SNP's breast cancer association in an independent cohort.

Tan *et al.* (34) found that the SNP rs1063320 is associated with asthma, depending on the mother's disease status. rs1063320 lies in the 3′-UTR of *HLA-G*, and the authors showed that this SNP affects miR-148a, miR-148b and miR-152 targeting of the *HLA-G* gene. They suggested that this altered miRNA targeting increases the risk of asthma.

With our haplotype scoring method run genome-wide, we found 3 SNPs (rs1063320, rs1610696 and rs1707) in the 3′-UTR of *HLA-G* that can affect 28 miRNAs (data not shown). rs1063320 affects 10 miRNAs (data not shown), and its three largest differences of scores are given by the same three miRNAs reported by Tan *et al.* (34): 0.76, 0.78 and 0.81, respectively for miR-148b, miR-148a and miR-152. The other scores range from 0.33 to 0.55, indicating that the three miRNAs are clear candidates.

Wang *et al.* (35) found that the SNP rs12720208 is associated with Parkinson's disease. rs12720208 lies in the 3′-UTR of *FGF20*. They also showed that this SNP has an effect on miR-433 targeting of *FGF20*. They suggested that this altered targeting increases the risk of Parkinson's disease.

We identified two SNPs (rs1721100 and rs12720208) in the 3′-UTR of *FGF20* that can affect four miRNAs (data not shown). The largest difference of scores for this gene is 0.88 and is given by miR-433 at rs12720208—the same miRNA/SNP pair reported by Wang *et al.* (35). One other miRNA scores 0.44 with rs12720208, whereas SNP rs1721100 scores 0.24 and 0.43 with two miRNAs. Consequently, the pair rs12720208/miR-433 seems to be a clear candidate.

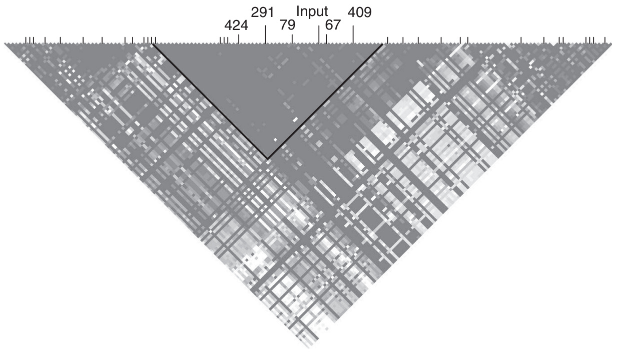

Figure 7. SNP rs1434536 (input) has an LD block (delimited by the dark lines) which contains top ranking SNPs (ranks 67, 79, 291, 409 and 424) from CGEMS's breast cancer GWAS.

DISCUSSION

By evaluating our proposed method on allelic imbalance sequencing data, two different GWAS data sets, and validated mirSNPs, we have demonstrated that our method is useful for identifying potential causative SNPs in miRNA target sites. Specifically, our analyses of the allelic imbalance sequencing data show that our proposed method outperforms existing methods. Although the data set is limited as it contains only 47 SNPs, the data set should be of high quality as it was generated *in vivo* without artificially altering miRNA or target expression (31). Indeed, our results revealed clear differences between the methods. Especially, the method based on changes in predicted miRNA–mRNA hybridization MFE showed poor performance and could not predict the SNPs' effect on miRNA targeting. This result is consistent with overall miRNA–mRNA hybridization in itself being a poor predictor of miRNA targeting and support the model of target site context being essential for miRNA regulation (1).

The basic approach used by many existing tools for detecting SNPs in miRNA target sites looks for SNPs in seed regions of predicted target sites. Seed regions are known to be the most important regions for miRNA targeting efficacy (1). Focusing on seed regions reduces the amount of false positive SNPs predicted to alter miRNA-targeting, but will miss SNPs affecting non-canonical miRNA targeting such as 3′ supplementary sites. This basic method can however be used to filter the mRNA/miRNA pairs that are most likely affected by SNPs. Such filtered SNPs can then subsequently be analysed with our haplotype method.

SNPs outside the seed region can affect miRNA targeting, however, and some existing approaches based on computational RNA–RNA hybridization or thermodynamic calculations consider such SNPs. Our method can also detect SNPs in 3′ supplementary sites, but according to our analyses, such SNPs have a small predicted effect (Supplementary Figure S8). This result is consistent with the observation that conserved 3′ supplementary sites constitute 4.9% of all conserved pairing sites (36). As SNPs affecting seed site pairing have a bigger predicted effect than those affecting other miRNA features, our online database provide allelic sequences for SNPs in target seed sites.

A transcriptome-wide study of interactions between miRNAs and mRNAs estimated that sites with seed mismatches constitute <6.6% of all miRNA target sites (19). By excluding SNPs in mismatch sites, we only miss SNPs that change a mismatch target site (weak) into another mismatch site. Moreover, non-canonical sites appear to have a smaller regulatory effect than canonical target sites have (19). Thus, our method focuses on identifying the SNPs that are most likely to affect and to have the largest effect on miRNA targeting.

Our haplotype scoring method is based on HapMap haplotype data, and only 66% of the SNPs from HapMap have haplotype data. The 34% HapMap SNPs that do not have haplotype data have a very low minimum allele frequency (MAF), usually 0 in the considered hapmap population. Removing low MAF SNPs is an advantage in mapping SNPs to common diseases, resulting in less false positives (false causal SNPs), in a common variant common disease model.

Our haplotype approach also currently only focuses on analysing 3′-UTRs. Although miRNAs can target 5′-UTRs and coding regions, these sites have a limited effect compared to 3′-UTR sites (19,20).

The main advantage of our method compared to existing methods is that we analyse the regulatory effects of all linked genetic variations within regulatory regions, such as 3′-UTRs. Consequently, our method can be used to analyse how SNPs in multiple target sites together contribute to upregulate, downregulate, or compensate each other, through haplotype patterns.

CONCLUSION

We have presented a tool that aims at identifying the causative variation within regions associated with diseases. Specifically, the tool identifies 3′-UTR SNPs that can affect miRNA targeting and predicts the SNPs' effect on miRNA regulation. Our main result is the SNP effect prediction method. The results suggest that the effect predictions are reliable, compare favourably to existing methods, and can be used to filter and identify causative SNPs.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Norwegian Research Council; Interagon AS; the Norwegian Functional Genomics Program (FUGE) of the Norwegian Research Council. Funding for open access charge: Norwegian University of Science and Technology.

Conflict of interest statement. None declared.

REFERENCES

1. Bartel,D.P. (2009) MicroRNAs: target recognition and regulatory Functions. *Cell*, **136**, 215–233.
2. Brennecke,J., Stark,A., Russell,R. and Cohen,S. (2005) Principles of MicroRNA-target recognition. *PLoS Biol.*, **3**, 404–418.
3. Didiano,D. and Hobert,O. (2006) Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions. *Nat. Struct. Mol. Biol.*, **13**, 849–851.
4. Gaidatzis,D., van Nimwegen,E., Hausser,J. and Zavolan,M. (2007) Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics*, **8**, 248.
5. Saetrom,P., Heale,B.S.E., Snove,O. Jr, Aagaard,L., Alluin,J. and Rossi,J.J. (2007) Distance constraints between microRNA target sites dictate efficacy and cooperativity. *Nucleic Acids Res.*, **35**, 2333–2342.
6. Hirschhorn,J. and Daly,M. (2005) Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.*, **6**, 95–108.
7. Donnelly,P. (2008) Progress and challenges in genome-wide association studies in humans. *Nature*, **456**, 728–731.
8. Mishra,P.J. and Bertino,J.R. (2009) MicroRNA polymorphisms: the future of pharmacogenomics, molecular epidemiology and individualized medicine. *Pharmacogenomics*, **10**, 399–416.
9. Borel,C. and Antonarakis,S.E. (2008) Functional genetic variation of human miRNAs and phenotypic consequences. *Mamm. Genome*, **19**, 503–509.
10. Landi,D., Gemignani,F., Naccarati,A., Pardini,B., Vodicka,P., Vodickova,L., Novotny,J., Foersti,A., Hemminki,K. and Canzian,F. (2008) Polymorphisms within micro-RNA-binding sites and risk of sporadic colorectal cancer. *Carcinogenesis*, **29**, 579–584.
11. Sethupathy,P. and Collins,F.S. (2008) MicroRNA target site polymorphisms and human disease. *Trends Genet.*, **24**, 489–497.
12. Bao,L., Zhou,M., Wu,L., Lu,L., Goldowitz,D., Williams,R.W. and Cui,Y. (2007) PolymiRTS Database: linking polymorphisms in microRNA target sites with complex traits. *Nucleic Acids Res.*, **35**, D51–D54.
13. Lewis,B., Burge,C. and Bartel,D. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.
14. Georges,M., Clop,A., Marcq,F., Takeda,H., Pirottin,D., Hiard,S., Tordoir,X., Caiment,F., Meish,F., Bibe,B. *et al.* (2006) Polymorphic microRNA-target interactions: a novel source of phenotypic variation. *Cold Spring Harb. Symp. Quant. Biol.*, **71**, 343–350.
15. Xie,X., Lu,J., Kulbokas,E., Golub,T., Mootha,V., Lindblad-Toh,K., Lander,E. and Kellis,M. (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature*, **434**, 338–345.
16. Barenboim,M., Zoltick,B.J., Guo,Y. and Weinberger,D.R. (2010) MicroSNiPer: a web tool for prediction of SNP effects on putative microRNA targets. *Hum. Mutat.*, **31**, 1223–1232.
17. Nicoloso,M.S., Sun,H., Spizzo,R., Kim,H., Wickramasinghe,P., Shimizu,M., Wojcik,S.E., Ferdin,J., Kunej,T., Xiao,L. *et al.* (2010) Single-nucleotide polymorphisms inside microRNA target sites influence tumor susceptibility. *Cancer Res.*, **70**, 2789–2798.
18. Miranda,K.C., Huynh,T., Tay,Y., Ang,Y.-S., Tam,W.-L., Thomson,A.M., Lim,B. and Rigoutsos,I. (2006) A pattern-based method for the identification of microRNA binding sites and their corresponding heteroduplexes. *Cell*, **126**, 1203–1217.
19. Hafner,M., Landthaler,M., Burger,L., Khorshid,M., Hausser,J., Berninger,P., Rothballer,A., Ascano,M. Jr, Jungkamp,A.-C., Munschauer,M. *et al.* (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, **141**, 129–141.
20. Grimson,A., Farh,K.K.-H., Johnston,W.K., Garrett-Engele,P., Lim,L.P. and Bartel,D.P. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell*, **27**, 91–105.
21. Int HapMap Consortium. In Frazer,K.A., Ballinger,D.G., Cox,D.R., Hinds,D.A., Stuve,L.L., Gibbs,R.A., Belmont,J.W., Boudreau,A., Hardenbol,P. *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.
22. Int Human Genome Sequencing Conso. In Lander,E., Linton,L., Birren,B., Nusbaum,C., Zody,M., Baldwin,J., Devon,K., Dewar,K., Doyle,M. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
23. Mouse Genome Sequencing Consor. In Waterston,R., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J., Agarwal,P., Agarwala,R., Ainscough,R., Alexandersson,M. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
24. Kuhn,R.M., Karolchik,D., Zweig,A.S., Wang,T., Smith,K.E., Rosenbloom,K.R., Rhead,B., Raney,B.J., Pohl,A., Pheasant,M. *et al.* (2009) The UCSC genome browser database: update 2009. *Nucleic Acids Res.*, **37**, D755–D761.
25. Griffiths-Jones,S., Saini,H.K., van Dongen,S. and Enright,A.J. (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res.*, **36**, D154–D158.
26. Hunter,D.J., Kraft,P., Jacobs,K.B., Cox,D.G., Yeager,M., Hankinson,S.E., Wacholder,S., Wang,Z., Welch,R., Hutchinson,A. *et al.* (2007) A genome-wide association study identifies alleles in

FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nature Genet.*, **39**, 870–874.

27. Maraganore,D., de Andrade,M., Lesnick,T., Strain,K., Farrer,M., Rocca,W., Pant,P., Frazer,K., Cox,D. and Ballinger,D. (2005) High-resolution whole-genome association study of Parkinson disease. *Am. J. Hum. Genet.*, **77**, 685–693.
28. Hindorff,L.A., Sethupathy,P., Junkins,H.A., Ramos,E.M., Mehta,J.P., Collins,F.S. and Manolio,T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.
29. Saito,T. and Saetrom,P. (2010) A two-step site and mRNA-level model for predicting microRNA targets. *BMC Bioinformatics*, **11**, 612.
30. Barrett,J., Fry,B., Maller,J. and Daly,M. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.
31. Kim,J. and Bartel,D.P. (2009) Allelic imbalance sequencing reveals that single-nucleotide polymorphisms frequently alter microRNA-directed repression. *Nat. Biotechnol.*, **27**, 472–477.
32. Rehmsmeier,M., Steffen,P., Hochsmann,M. and Giegerich,R. (2004) Fast and effective prediction of microRNA/target duplexes. *RNA-Publ. RNA Soc.*, **10**, 1507–1517.
33. Saetrom,P., Biesinger,J., Li,S.M., Smith,D., Thomas,L.F., Majzoub,K., Rivas,G.E., Alluin,J., Rossi,J.J., Krontiris,T.G. *et al.* (2009) A risk variant in an miR-125b binding site in BMPR1B is associated with breast cancer pathogenesis. *Cancer Res.*, **69**, 7459–7465.
34. Tan,Z., Randall,G., Fan,J., Camoretti-Mercado,B., Brockman-Schneider,R., Pan,L., Solway,J., Gern,J.E., Lemanske,R.F. Jr and Nicolae,D. (2007) Allele-specific targeting of microRNAs to HLA-G and risk of asthma. *Am. J. Hum. Genet.*, **81**, 829–834.
35. Wang,G., van der Walt,J.M., Mayhew,G., Li,Y.-J., Zuechner,S., Scott,W.K., Martin,E.R. and Vance,J.M. (2008) Variation in the miRNA-433 binding site of FGF20 confers risk for Parkinson disease by overexpression of alpha-synuclein. *Am. J. Hum. Genet.*, **82**, 283–289.
36. Friedman,R.C., Farh,K.K.-H., Burge,C.B. and Bartel,D.P. (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.*, **19**, 92–105.