Principles of microRNA Regulation Revealed Through Modeling microRNA Expression Quantitative Trait Loci

Stefan Budach,*,[1] Matthias Heinig,†,[2] and Annalisa Marsico*,‡,[2]
*RNA Bioinformatics, Max Planck Institute for Molecular Genetics, 14195 Berlin, Germany, †Institute of Computational Biology, Helmholtz Zentrum München, 85764 Neuherberg, Germany, and ‡High Throughput Genomics, Department of Mathematics and Computer Science, Freie Universität Berlin, 14195 Berlin, Germany
ORCID ID: 0000-0002-5612-1720 (M.H.)

ABSTRACT Extensive work has been dedicated to study mechanisms of microRNA-mediated gene regulation. However, the transcriptional regulation of microRNAs themselves is far less well understood, due to difficulties determining the transcription start sites of transient primary transcripts. This challenge can be addressed using expression quantitative trait loci (eQTLs) whose regulatory effects represent a natural source of perturbation of *cis*-regulatory elements. Here we used previously published *cis*-microRNA-eQTL data for the human GM12878 cell line, promoter predictions, and other functional annotations to determine the relationship between functional elements and microRNA regulation. We built a logistic regression model that classifies microRNA/SNP pairs into eQTLs or non-eQTLs with 85% accuracy; shows microRNA-eQTL enrichment for microRNA precursors, promoters, enhancers, and transcription factor binding sites; and depletion for repressed chromatin. Interestingly, although there is a large overlap between microRNA eQTLs and messenger RNA eQTLs of host genes, 74% of these shared eQTLs affect microRNA and host expression independently. Considering microRNA-only eQTLs we find a significant enrichment for intronic promoters, validating the existence of alternative promoters for intragenic microRNAs. Finally, in line with the GM12878 cell line derived from B cells, we find genome-wide association (GWA) variants associated to blood-related traits more likely to be microRNA eQTLs than random GWA and non-GWA variants, aiding the interpretation of GWA results.

KEYWORDS microRNA; eQTL; promoter; variation; regulation

SMALL noncoding RNAs known as microRNAs (miRNAs) are playing an important role in the post-transcriptional regulation of gene expression. Only in 2000 was it discovered that the sequence of the *let-7* family of miRNAs is conserved among multiple species, attracting great attention to these small RNAs which had previously been ignored (Pasquinelli *et al.* 2000). Their biogenesis and function in metazoans have been extensively researched over recent years (Pasquinelli *et al.* 2005; Ha and Kim 2014) and the majority of the human protein-coding genes are regulated by miRNAs; ~60% of them possess at least one known conserved miRNA-binding site (Friedman *et al.* 2009). Due to their importance and abundance, miRNAs have been increasingly linked to medical conditions (Sayed and Abdellatif 2011; Im and Kenny 2012; Lujambio and Lowe 2012).

miRNAs are located within different genomic contexts (intragenic or intergenic regions), either possessing their own promoter or sharing the promoter of a host gene (Ozsolak *et al.* 2008). They are transcribed by RNA polymerase II into a primary-miRNA (pri-miRNA) of up to several kilobases in length (Corcoran *et al.* 2009) and the following processing by Drosha and DGCR8 results in a structured stem-loop precursor-miRNA (pre-miRNA) of length ~75 nt, which is in turn cleaved by Dicer into an miRNA duplex. The ~22-bp-long duplex is loaded onto the RNA-induced silencing complex (RISC) and one strand of the duplex is released

doi: 10.1534/genetics.116.187153
Manuscript received January 15, 2016; accepted for publication May 27, 2016; published Early Online June 1, 2016.
Available freely online through the author-supported open access option.
Supplemental material is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.187153/-/DC1.
[1] Corresponding author: RNA Bioinformatics, Max Planck Institute for Molecular Genetics, Ihnestraße 63–73, 14195 Berlin, Germany. E-mail: budach@molgen.mpg.de
[2] These authors contributed equally to this work.

(passenger strand). The remaining strand leads RISC to a target messenger RNA (mRNA), resulting in the degradation or repression of that mRNA (Pasquinelli *et al.* 2005; Ha and Kim 2014). According to miRBase nomenclature, the strands originating from the 5′ and 3′ arms of the pre-miRNA are named mature 5p and 3p miRNA, respectively (Kozomara and Griffiths-Jones 2014).

Despite great progress in understanding the biological role of miRNAs in regulating gene expression, our understanding of how miRNAs themselves are regulated, both at transcriptional and post-transcriptional level, is still developing. Recent research focuses increasingly on how miRNA expression is controlled, *e.g.*, which regulatory elements cause tissue-specific expression and deregulation in pathological conditions. Among others, the short length of miRNAs and the rapid cleavage by Drosha present technical issues complicating their experimental analysis. For this reason, computational methods are increasingly used to predict miRNA-related annotations, such as promoters, from both sequencing data and primary sequences (Marsico *et al.* 2013; Georgakilas *et al.* 2014).

Most *cis*-regulatory elements are encoded in the DNA sequence either by sequence motifs or higher-order patterns. Sequence variants affecting these regions are thus expected to change expression patterns of the associated gene and can be viewed as naturally occurring perturbations of *cis*-regulatory elements. These variants can affect expression of many types of genes including protein-coding genes and miRNAs. Such changes in expression manifest themselves between individuals, populations, and distinct tissue-specific phenotypes (Lappalainen *et al.* 2013; GTEx Consortium 2015). Expression quantitative trait locus (eQTL) studies enable the systematic detection of genomic loci associated with the expression levels of transcripts (Jansen and Nap 2001; Gilad *et al.* 2008; Lappalainen *et al.* 2013; GTEx Consortium 2015). Previously, eQTLs were used to gain better insight into the biology of human gene expression regulation for protein-coding genes by incorporating them with genomic regulatory annotations (Lee *et al.* 2009; Gaffney *et al.* 2012; Battle *et al.* 2014).

Concerning miRNAs, the mechanisms of miRNA eQTLs (sequence variants associated with miRNA expression levels) are less known due to the limited availability of data sets which map eQTLs to miRNA expression across different tissues. Two studies report a limited number of miRNA eQTLs in adipose tissue (Parts *et al.* 2012) and dendritic cells (Siddle *et al.* 2014), respectively. Gamazon *et al.* (2013) provide a map of *trans*-only miRNA eQTLs in liver and Huan *et al.* (2015) compile a genome-wide map of miRNA eQTLs in whole blood. This final study provides a comprehensive mapping of miRNA eQTLs, however, it is inappropriate for the study of cell type-specific regulatory elements as whole blood constitutes a mixture of many cell populations.

Here we chose the data set from Lappalainen *et al.* (2013), since it comprises a large number of *cis*-miRNA eQTLs, the full genotype information as part of The 1000 Genomes Project Consortium (2012), and its tissue specificity (B-lymphoblastoid cell line GM12878). This cell line has been profiled extensively by the ENCODE project (ENCODE Project Consortium 2012) and epigenetic and genomic annotations are publicly available. We combined the miRNA eQTLs and regulatory annotations with a methodology for miRNA promoter prediction, previously developed in our group (Marsico *et al.* 2013), to address how genetic variation affects miRNA expression. Therefore, we trained a logistic regression model to classify miRNA/SNP pairs into eQTLs and non-eQTLs based on the overlap of the SNPs with a range of genomic features, such as miRNA gene structure, miRNA promoters, and epigenome mapping.

The final model selected by the Akaike information criterion (AIC) achieves 85% accuracy on an independent test set. miRNA eQTLs were enriched for regions of the miRNA precursor, tissue-specific miRNA promoters, enhancers, and transcription factor binding sites (TFBS). Conversely, odds of miRNA eQTLs were decreased when an insulator was between SNP and miRNA, or the SNP was in a region of repressed chromatin. miRNA eQTLs were also enriched for eQTLs of host genes and a substantial fraction of eQTLs was shared between intragenic miRNAs and their hosts. We found, however, that the majority of shared eQTLs affected miRNA and host expression differently. miRNA-only eQTLs were enriched for miRNA promoters, mainly intronic promoters, suggesting a decoupling of host and miRNA expression that is modulated by genetic variation. Finally, we applied our model to predict miRNA eQTLs for SNPs that were identified in genome-wide association studies (GWAS). We found that the predicted probabilities of being an miRNA eQTL are significantly higher for variants associated to blood-related traits compared to random genome-wide association (GWA) and non-GWA variants. This is in line with the GM12878 cell line derived from B cells.

Materials and Methods

eQTL and SNP data

The set of *cis*-eQTLs originated from Lappalainen *et al.* (2013). They performed mRNA (462 individuals) and small-RNA sequencing (seq) (452 individuals) on GM12878 cell line samples from the European (EUR) and Yoruba (YRI) populations and determined both mRNA eQTLs and miRNA eQTLs. The set of non-miRNA eQTLs was defined as all remaining SNPs in the region ±500 kb around pre-miRNAs (The 1000 Genomes Project Consortium 2012, hg19, phase 1, version 3). All SNPs are filtered for a minor allele frequency >5%.

Data for general cis-regulatory elements

The following regulatory genomic annotations from ENCODE (ENCODE Project Consortium 2012) for the GM12878 cell line were used: Chromatin immunoprecipitation followed by sequencing peak data for 76 TFBS (http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=wgEncodeRegTfbsClusteredV3), DNase-seq peak data for DNase I hypersensitive sites (DHSs) (GSM816665) and the ChromHMM chromatin states (GSM936082). ChromHMM uses a hidden Markov model to annotate genomic regions according to combinations of chromatin modifications. This results in 15 different possible states associated to, for example, enhancer regions, promoter regions, insulators, and regions in a repressed state.

Data for miRNA promoter predictions

Two sets of putative miRNA promoters were predicted. In both cases we used PROmiRNA (Marsico *et al.* 2013), a method previously developed in our lab and based on a semisupervised machine-learning approach trained on deep cap analysis gene expression (CAGE) data and promoter sequence features. The algorithm considers upstream loci of pre-miRNAs enriched in CAGE signals as putative promoters and incorporates CpG content, conservation, TATA box affinity, and mature miRNA proximity to score real promoters *vs.* background transcription. PROmiRNA provides multiple predictions per miRNA that correspond to different promoters which are potentially active in different tissues and conditions.

For the first set [feature *promoter (unspecific)* in our model] we used the predictions reported in the PROmiRNA article, derived from applying PROmiRNA to the human genome using the FANTOM4 data on >33 nonredundant tissues (Kawaji *et al.* 2009). These predictions are complemented by RNA-seq-based predictions from the microTSS software for hESC and IMR90 cell lines, as reported in the supplementary tables accompanying Georgakilas *et al.* (2014). These data do not contain the GM12878 cell line and promoter predictions in this case may not correspond to promoters active in this cell line, but represent all possible alternative promoters. To restrict predictions to active promoters in the GM12878 cell line [feature *promoter (specific)* in our model] PROmiRNA was retrained on ENCODE CAGE data (ENCODE Project Consortium 2012) for the GM12878 cell line (two replicates), keeping only promoters found in both replicates. Predictions were further filtered for promoters overlapping at least 1 bp with DHS peaks (GSM816665). Predictions for both sets, on average 20–30-bp long due to the CAGE peaks, were extended by 100 nt in both directions to cover a region that putatively represents the core promoter. miRNA promoters of intergenic miRNAs are defined as intergenic promoters, promoters overlapping the ±100-bp region around transcription start sites (TSSs) of miRNA host genes are defined as host gene promoters, and promoters located inside introns of miRNA host genes are defined as intronic promoters.

Data for miRNA gene structure

MiRBase 20 was used for coordinates of pre-miRNA and mature 5p/3p miRNAs (Kozomara and Griffiths-Jones 2014). Due to changes in this newer miRBase version, we only used data for 638 autosomal mature miRNAs (478 pre-miRNAs) and not the complete set analyzed in Lappalainen *et al.* (2013) (644 mature miRNAs, miRBase 18).

Feature encoding

We developed a logistic regression model to classify SNPs as miRNA eQTLs or non-miRNA eQTLs according to their location with respect to several genomic features. All annotation-based model features were encoded as 1 or 0, indicating whether a SNP overlaps with an annotation. The following annotations were used as features: binding sites for 76 transcription factors; DHSs; all 15 ChromHMM states; two sets of miRNA promoter predictions; and multiple separate parts of the miRNA primary transcript, namely the mature 5p miRNA, the mature 3p miRNA, the hairpin loop between the mature sequences, and the 22-bp-long regions upstream/downstream of the Drosha 5′/3′ cutting points (see Figure 1 for a visual overview).

In addition to simple overlap-based features, we included (all features encoded as 1 or 0): an *insulator in between* feature (indicating whether a ChromHMM insulator state is located between SNP and miRNA), an *intragenic* feature (indicating whether the miRNA is intragenic), and an *mRNA-eQTL* feature (indicating whether the SNP is an eQTL for the host gene). Finally, we included the *distance* between SNP and miRNA, encoded as follows:

$$1 - \frac{|distance(SNP, miRNA)|}{500\ \text{kb}} \tag{1}$$

According to this formula, a higher distance results in a smaller encoding with a minimum value of 0 for SNPs located 500 kb far away. SNPs located within the pre-miRNA were assigned a default value of 1. The distance feature is needed because the number of eQTLs decreases exponentially with distance from the miRNA. As regulatory annotations are enriched at different distances, the enrichment for miRNA eQTLs in those regions may be due to a "position" effect, rather than to the actual function of the regulatory element (see Figure S1, A and B). The distance feature, encoded linearly in this formula, corresponds to an overall uniform-distance distribution of SNPs in the model (miRNA eQTLs and non-miRNA eQTLs together). To better capture the exponential decay of miRNA eQTLs with distance, we also considered mapping the distances to the quantiles of the empirical distance distribution of mRNA eQTLs. However, since the final model performance and feature importance did not differ notably, we opted for the simpler linear encoding.

Model building

We considered the set of SNPs located in the region ±500 kb around the pre-miRNA start and end positions for our model. This resulted in a data set of 2,002,126 miRNA/SNP pairs for 638 mature miRNAs, from now on referred to as "observations" of the model. From all observations, 4785 were miRNA eQTLs and the remaining 1,997,341 were non-miRNA eQTLs. The 4785 miRNA eQTLs are associated with 58 mature miRNAs and each mature miRNA has on average 83 eQTLs (median 36). The probability P_i of an SNP i to be an miRNA eQTL, given J features F_{ij} ($j = 1 \ldots J$) was modeled by logistic regression:

$$P_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 F_{i1} + \beta_2 F_{i2} + \ldots + \beta_J F_{iJ})}} \tag{2}$$

In this equation we also control for the effect of SNP proximity to the annotated pre-miRNA, encoded in the distance feature F_{i1} as described above.

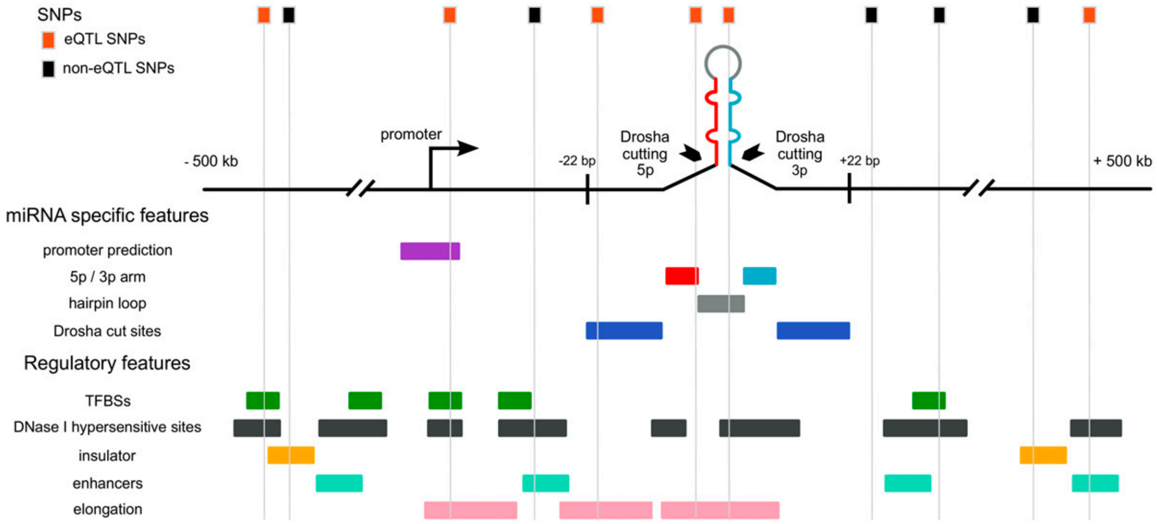


Figure 1 Schematic outline of the classification model. A schematic representation of the ±500-kb genomic region around an miRNA is shown. The miRNA primary transcript is depicted in the middle, and the location of the predicted promoter, the 5p/3p arms of the pre-miRNA stem-loop (highlighted in red/blue), the terminal loop between 5p and 3p and the Drosha cutting points are marked. miRNA eQTLs and non-miRNA eQTLs are shown at the top in orange and black, respectively. For the model building, the overlap of each SNP with the miRNA-specific and the *cis*-regulatory features is shown (vertical lines). For simplicity, not all *cis*-regulatory features used in the full model are shown.

Transcription factor selection

To reduce the number of features, all informative transcription factors were merged into one general *TFBS* feature. To determine the important factors, we applied logistic regression for each factor separately on the complete set of observations; while including the distance as a second feature (as in Equation 2). These and all following logistic regression models use the R generalized linear model (glm) function. All transcription factors exhibiting a significant positive model coefficient ($P < 0.05$) were merged into the TFBS feature.

Feature selection and interpretation

General feature selection was performed using the TFBS feature and all other described features. The same procedure as described above was applied (logistic regression on the complete data using one feature at a time + distance). Significant features were then used to perform further model selection. Considering that many features are correlated as the annotations are partially overlapping or mutually exclusive, we also used the estimated regression coefficients for interpretation (see Figure S1C).

Model selection and testing

To select a final model, logistic regression was performed with different combinations of significant features determined in the previous step. Given the strong class imbalance between miRNA-eQTL observations and non-miRNA-eQTL observations, we sampled balanced subsets of the two classes for model selection and testing. For model selection, we randomly sampled three quarters of the 4785 miRNA-eQTL observations and the same number (3588) of non-miRNA-eQTL observations. The sampling was repeated 50 times. We selected the feature combination which minimized the AIC. For each combination of features we determined the mean AIC value (computed by the R-glm function). The performance of the final model with regard to accuracy and precision was measured using observations that were not part of the training data used during model selection. Again, 50 balanced subsets were sampled comprising the remaining quarter miRNA-eQTL observations and an equal amount of non-miRNA-eQTL observations.

miRNA expression analysis

To visualize the small RNA-seq read coverage shown in Figure 5, raw data from the Geuvadis project (Lappalainen *et al.* 2013) was remapped, as alignment files are provided for mRNA-seq data, but not for small RNA-seq data. Fastq files corresponding to small RNA-seq data from the GM12878 cell line carrying the genotype of interest were downloaded from ArrayExpress (http://www.ebi.ac.uk/arrayexpress/; accessions E-GEUV-1, E-GEUV-2, and E-GEUV-3). Reads <18 nt long were discarded and 3′ adapters clipped using a custom script. The remaining reads were mapped to the hg19 assembly of the human genome with the following command: *bowtie -f -v 1 -a –best –strata*. Raw read counts corresponding to both 5p and 3p mature miRNAs were computed using the quantifier.pl module of miRDeep2 (Friedländer *et al.* 2012). For visualization purposes, read counts were scaled to a total read count equal to the median library size to resolve differences in miRNA expression levels across samples.

miRNA–host gene independence analysis

We also reanalyzed the original gene expression and genotype data from the Geuvadis project to determine whether overlapping eQTLs of host genes and miRNA genes are really shared or independent associations. We downloaded the expression data for mRNAs as well as the genotype data and eQTL results and scaled the RNA-seq read count data as described in the previous paragraph. In addition, we mapped the counts on the quantiles of a standard normal distribution for both mRNAs and miRNAs (as described in Lappalainen *et al.* 2013). We analyzed all triplets of miRNAs, host genes, and SNPs; where the SNP was an eQTL for either the miRNA or the host on gene or exon level in at least one of the two populations (EUR, YRI). For each miRNA, we tested both the 5p and the 3p arm when available. To test for miRNA eQTLs independent of the host gene, we compared the two nested linear models,

$$\begin{aligned}H_1 &: miRNA \sim mRNA + SNP\\ H_0 &: miRNA \sim mRNA,\end{aligned}$$

using a likelihood-ratio test. In addition, we computed the miRNA–SNP and host–SNP associations as well as the correlation of host gene and miRNA. We called miRNA eQTLs independent when the miRNA eQTL was genome-wide significant and the independent eQTL test was significant (false discovery rate < 0.05).

Data availability

Supplemental Material, File S1 contains the promoter predictions used. File S2 contains several data files, including the results of the miRNA–host gene independence analysis. It also contains an R script used to build the model and to create the plots in this article, and instructions to reproduce the calculations. The final full model matrix can be downloaded from https://github.molgen.mpg.de/budach/miRNA_eQTL.

Results

A classification model for miRNA eQTLs

To comprehensively understand the genomic context of genetic variants driving miRNA expression, we developed a classification model to predict the probability of a certain SNP being an miRNA eQTL based on the SNP's location with respect to functional genomic annotations (from now on called "features" of the model). Note that in contrast to previous work (Lee *et al.* 2009; Gaffney *et al.* 2012; Battle *et al.* 2014) we did not aim to resolve the most-likely causal variants underlying the eQTL, but rather built the first interpretable model of miRNA eQTLs, as explained in more detail in the *Discussion*. The eQTL data used to build the model originates from the Geuvadis project (Lappalainen *et al.* 2013). They performed mRNA- and small RNA-seq on hundreds of GM12878 cell line samples from The 1000 Genomes Project (2012) phase 1 data and identified *cis*-eQTLs of protein-coding and miRNA genes in the EUR and YRI populations. For classification we used logistic regression to retrieve model parameters that are easily interpretable in terms of log-odds ratios, which indicate enrichment or depletion of miRNA eQTLs with respect to each feature. Model features include different nonoverlapping parts of the miRNA primary transcript, promoter predictions, TFBS, and other *cis*-regulatory annotations for both open and repressed chromatin regions. Intragenic miRNAs are often coexpressed with their host genes and as they may be regulated differently than intergenic miRNAs, we included additional information into the model, such as whether an miRNA is intragenic and whether a SNP is an eQTL of the host gene. A schematic representation of our logistic regression model is shown in Figure 1 (see *Materials and Methods* for a detailed description of eQTL data and features).

Feature selection identifies relevant annotations

We first sought to filter out unimportant features that do not help in predicting the response. We also quantified the enrichment or depletion for miRNA eQTLs relative to each model feature, while controlling for the SNP-to-miRNA distance. Therefore we performed logistic regression with each model feature separately, while including the distance as a second feature. In Figure 2 the resulting coefficients (β values of the regression) are shown. The coefficients in a logistic regression represent the log-odds ratios between the odds of miRNA eQTLs *vs.* non-miRNA eQTLs in a certain feature compared to the odds in background regions. A positive log-odds ratio implies that a feature is likely to increase the chance of being an miRNA-eQTL SNP. A total of 9 out of 27 features do not possess a significant log-odds ratio and hence were excluded from the further analysis ($P > 0.05$). Among these are seven ChromHMM states (2_Weak_Promoter, 5_Strong_Enhancer, 6_Weak_Enhancer, 8_Insulator, 11_Weak_Txn, 14_Repetitive, and 15_Repetitive), the DHSs, and a set of promoter predictions not specific to the cell line.

Final model selection yields 85% accuracy

Combinations of significant features determined above were tested to minimize the AIC and to select a final model. The AIC estimates the goodness of fit of a model to its observations and adds a penalty for the number of features; thus it penalizes the model complexity. A lower AIC indicates a better model quality. On this basis, the combination of all significant features provides the best fit to the data (Figure 3A). In particular, it is noticeably superior to a distance-only model. Adding single features to the distance-only model improves the AIC. Adding the mRNA-eQTL feature yields the largest improvement. Correlations between features do not affect the performance of the model, but as they do alter the estimated log-odds ratios, the subsequent biological interpretation is based on the coefficients obtained during the feature selection (Figure 2, for correlations see Figure S1C).The final full model performance (Figure 3, B and C) was measured on observations not used during the model training. The logistic regression predicts the probability of being an miRNA eQTL for each observation. The maximum accuracy of the predictions amounts to 85% for a probability cutoff of 0.5 and is stable for different samples (Figure 3B). For

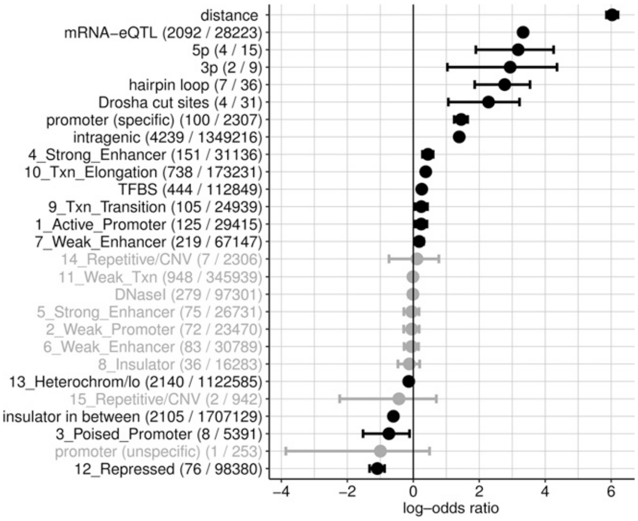


Figure 2 Feature selection and log-odds ratios. For each feature a separate logistic regression was performed, always including the distance as a second feature. Insignificant log-odds ratios (natural logarithm) are shaded ($P > 0.05$). Error bars illustrate 95% C.I. Numbers in parentheses show the amount of miRNA eQTLs/total amount of SNPs for each feature. Feature names starting with a number and an underscore represent the original ChromHMM identifiers.

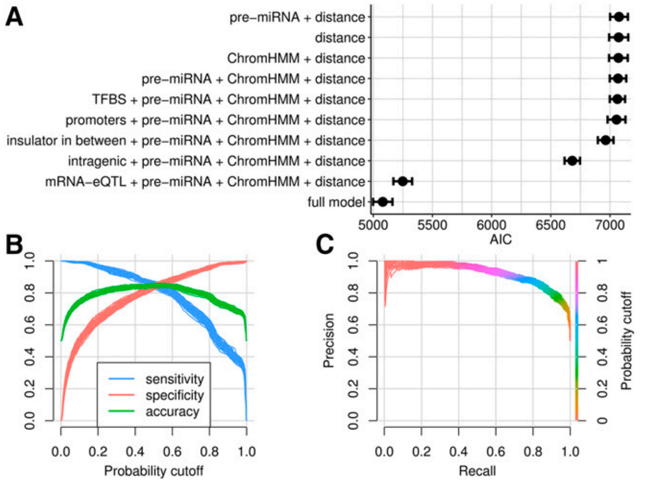


Figure 3 Model selection and performance. (A) AIC of different feature combinations. "pre-miRNA" represents several features: mature 5p and 3p miRNA, hairpin loop, and the flanking regions of the Drosha cutting points. "ChromHMM" represents all remaining significant ChromHMM features (see Figure 2). A low AIC indicates a better model quality. The dots are mean values, error bars depict the SDs obtained from repeated random sampling of the data (see *Materials and Methods*). (B) Accuracy, sensitivity, and specificity of the full model. In this case, the number of miRNA eQTLs and non-miRNA eQTLs is balanced and for a probability cutoff of 0.5 the model achieves an accuracy of 85%, which is stable for different samples. (C) Precision/recall plot colored by the probability cutoff. The preferred cutoff of 0.5 leads to both a high precision and a high recall (sensitivity) of ~85% each.

the same cutoff, precision and recall are 85% as well, and the corresponding false discovery rate is 15% (Figure 3C). A model containing only the two most-important features, distance and mRNA eQTL (Figure 2 and Figure 3A), achieves near full model accuracy of 83%. Since mRNA eQTLs are already enriched for regulatory regions (Lappalainen *et al.* 2013), this is expected and accordingly the full model without the mRNA-eQTL feature achieves 79% accuracy (model not shown).

miRNA eQTLs are enriched for TFBSs related to the transcription process and immune functions

Our model shows miRNA eQTLs are enriched for regulatory elements such as active enhancers, active promoters, regions of transcriptional elongation, and TFBS. Conversely, miRNA-eQTL SNPs are depleted for heterochromatic and repressed regions, and for poised promoters. The existence of an insulator element between SNP and miRNA also makes the probability of being an miRNA-eQTL SNP less likely. As miRNA eQTLs were enriched for TFBSs, we analyzed the individual factors in more detail and performed logistic regression with each transcription factor separately, always including the distance as a second feature (see Figure S1D). The ranking of these factors according to their *P*-values and log-odds ratios (Table 1) shows highly ranked factors which are crucial in general transcription processes or in immune response. Factors with a low *P*-value tend to have more binding events and therefore a high number of overlapping SNPs, whereas those with a high log-odds ratio usually have a lower number of overlapping SNPs, but a higher portion of miRNA eQTLs. Among the top-ranked factors we found transcriptional regulators, such as *CHD1* [chromatin-remodeling factor regulating the RNA polymerase II transcription and known to recruit several complexes to the H3K4me3 chromatin mark (Sims *et al.* 2007)]; *IKZF1* [associated with chromatin remodeling (Payne and Dovat 2011)]; *BRCA1* [interacts with RNA polymerase II holoenzyme (Scully *et al.* 1997)]; and factors related to immune-specific functions, such as *JUND* [also known to interact with *BRCA1* (Hu and Li 2002)], *CEBPB* (Ramji and Foka 2002), *MEF2A* (McKinsey *et al.* 2002), and *PAX5* (Schebesta *et al.* 2007).

Enrichment of miRNA eQTLs around the miRNA precursor

All features representing the individual parts of the miRNA stem-loop show significant miRNA-eQTL enrichment. These regions play important roles in the miRNA biogenesis; hairpin loops are important for the binding of the Dicer–TAR-binding protein complex (Zeng 2006) and the flanking regions are critical for the detection of cutting sites by Drosha. Studies have shown that sequence determinants of Drosha processing are located in a region ~20 nt downstream/upstream of the cutting sites (Auyeung *et al.* 2013; Conrad *et al.* 2014). When looking at the data we found that miRNA eQTLs occur up to position 22 downstream/upstream of cutting points and the next ones do not occur until position 59, supporting previous studies (only non-miRNA eQTLs were located between these positions).

Cell line-specific promoter predictions are enriched for miRNA eQTLs

miRNA promoters were predicted using the PROmiRNA software (Marsico *et al.* 2013), which originally used CAGE data from all available tissues in the FANTOM4 database. Our promoter (unspecific) model feature is based on those original predictions, covering all the FANTOM4 data, complemented by the predictions of the microTSS software (Georgakilas *et al.* 2014). The promoter (unspecific) predictions do not include our cell line of interest. This feature was not significant during the feature selection ($P \approx 0.32$, $\beta \approx -1.00$), highlighting the importance of locating active promoters in the cell line of interest. To do this, we retrained PROmiRNA on ENCODE CAGE data for the GM12878 cell line (ENCODE Project Consortium 2012). The resulting promoter (specific) model feature was significant ($P \approx 6.26\text{e}{-44}$, $\beta \approx 1.45$), indicating that functional miRNA eQTLs are enriched for cell type-specific, active miRNA promoters.

Intronic promoters initiate miRNA transcription independent from the host gene

The majority (313 or 65%) of miRNA precursors from our data were intragenic, located within exons, introns, or untranslated regions of host genes; and 165 (35%) were intergenic. Thus we tested whether intragenic miRNAs are regulated by their host gene promoters or by their own independent promoters. In the GM12878 cell line we predicted at least one promoter for 312 miRNA precursors out of 478 (a total of 1501 miRNA promoters). Intriguingly, only 475 (32%) were coincident with the host gene promoter whereas the majority showed independent promoters located in intergenic regions (451 or 30%) and within introns of host genes (575 or 38%). A total of 100 out of the 4785 miRNA/SNP pairs categorized as miRNA eQTLs overlapped predicted promoters. The majority of these were intronic miRNA promoters distinct from host gene promoters (49 out of 100), followed by intergenic promoters (32 out of 100) and host gene promoters (only 19 out of 100). Compared to host gene promoters, miRNA eQTLs show a significant enrichment for intronic promoters ($P = 2.93\text{e}{-05}$, Fisher's exact test; see Figure 4). Intronic promoters are thought to be alternative miRNA promoters driving intragenic miRNA transcription independently from their host genes in a tissue-specific manner (Monteys *et al.* 2010). Although miRNA promoter prediction algorithms report a large number of intronic miRNA promoters, the existence and activity of these promoters has been validated in few experimental studies and large-scale experimental validations are currently missing. Here we showed that 49 miRNA eQTLs overlap predicted intronic promoters in GM12878, indicating that these promoters have a potential function in this cell line and that genetic variation in such promoter elements likely contributes to a decoupling of miRNA expression from host gene expression.

Figure 5 shows three examples of predicted miRNA promoters and miRNA eQTLs that can be instructive in understanding the biological mechanisms of host-independent miRNA transcription. In all three cases the deepCAGE peaks upstream of the precursor correspond to predicted promoters that harbor miRNA eQTLs. This is indicated by changes in read coverage at the 5p and 3p miRNAs for different SNP genotypes. Figure 5, A and B, shows intronic miRNA promoters for which sequence variants influence miRNA expression, but not host gene expression. These promoters can be cell type-specific (Figure 5B) or present in multiple cell lines (Figure 5A). Figure 5C shows a cell type-specific intergenic and bidirectional predicted promoter that is located upstream of the host gene promoter and for which the overlapping eQTL changes both miRNA and mRNA expression.

74% of shared eQTLs affect miRNA and host expression independently

We found that 18 intragenic mature miRNAs share *cis*-eQTLs with their host genes and thus asked for which functional annotations shared eQTLs and miRNA-only or mRNA-only eQTLs are enriched. While shared eQTLs (2092) were found to be enriched for several TFBS, miRNA-only eQTLs (2147) were significantly enriched for promoter regions, mainly intronic promoters, as well as miRNA-hairpin subregions (one-sided Fisher's exact test, see Table S1). This further strengthens the argument that SNPs in intronic miRNA promoters and in the miRNA hairpin affect miRNA biogenesis and expression independently from the host. Conversely, when looking for enriched features for mRNA-only eQTLs (26131) we find that insulator in between is the only highly significant regulatory feature ($P \approx 9.47\text{e}{-90}$). This indicates that the cell makes use of insulator elements to decouple the expression of intragenic miRNAs from the expression of the corresponding host transcripts. For miRNA eQTLs that overlapped with mRNA eQTLs, we performed a test of independence (see *Materials and Methods*) to assess whether the associations between SNPs and miRNAs remained significant when conditioned on the host gene expression level using miRNA and mRNA expression data. We found that 74% of the miRNA/SNP associations remained significant at a false discovery rate of 5%, if conditioned on the host gene expression level. These results indicate that shared eQTLs can affect both miRNA and host gene expression independently, *i.e.*, expression values of miRNA and host gene are not correlated within each genotype group (see Figure S1, E and F, for examples).

The eQTL model predicts tissue-specific GWAS variants

The National Human Genome Research Institute (NHGRI) GWAS catalog contains SNPs associated to clinical conditions and phenotypic traits (Welter *et al.* 2014). As GWAS SNPs located in regulatory regions are potentially causal for the associated phenotypes, we tested whether that is also the case for miRNA eQTLs. In other words, we investigated if an miRNA-eQTL-specific model built on the B-lymphoblastoid cell line

Table 1 The 10 top-ranked transcription factors

P-value		Log-odds ratio	
MEF2A	**9.45e−15**	**JUND**	**2.05**
PAX5	7.35e−14	BRCA1	1.59
JUND	**7.82e−14**	NR2C2	1.58
IKZF1	**7.86e−14**	**CEBPB**	**1.23**
CEBPB	**2.21e−13**	**IKZF1**	**1.16**
CHD1	**4.10e−13**	**MEF2A**	**1.14**
ATF2	2.24e−10	**CHD1**	**1.04**
YY1	3.75e−10	ZBTB33	1.03
PML	6.53e−10	SIX5	0.99
RELA	1.66e−09	RCOR1	0.94

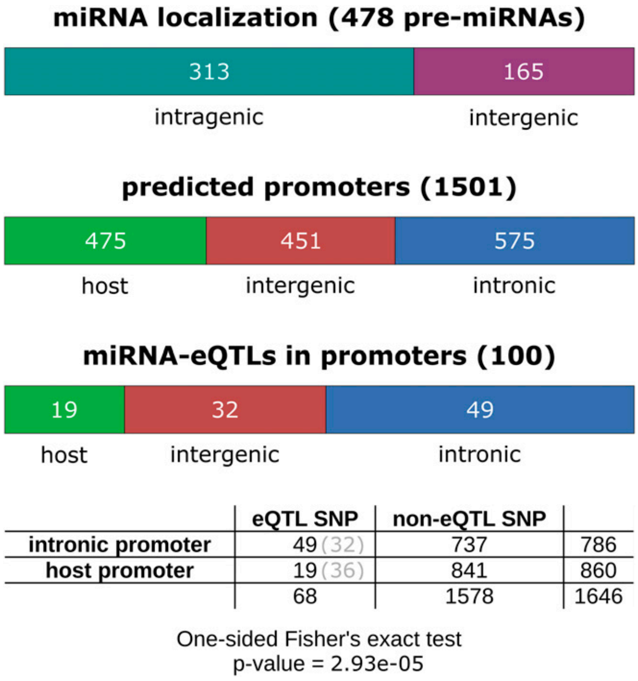
Factors found in both top 10 rankings are shown in boldface type.

GM12878 provides relevant information to interpret GWAS SNPs. The GWAS SNP annotation was retrieved from the NHGRI GWAS catalog and filtered using a *P*-value cutoff < 5e–8. We selected GWAS SNPs related to traits of the blood (GM12878 being a blood cell line) according to the International Classification of Diseases. This includes all GWAS SNPs related to primary diseases of the hematopoietic system, as well as blood tumors, immunological diseases of the blood, and other blood-related phenotypes. For the "blood"-associated GWAS SNPs that were also in our data set used to build the model (108 SNPs out of which 10 were miRNA eQTLs), we predicted the probability of being an miRNA eQTL. We also sampled equal-sized random sets of non-GWAS SNPs and GWAS SNPs which did not belong to the blood category and computed the probabilities. By comparing the cumulative distributions of these probabilities for the three groups, we could show that blood-associated SNPs have a significantly higher probability of behaving as miRNA eQTLs compared to the other two groups (Kolmogorov–Smirnov test, Figure 6). The significantly skewed cumulative distribution illustrates that the annotations of our model carry disease-relevant information and that it can be used to enhance the interpretation of GWAS variants.

Discussion

Despite its relevance, the biology underlying miRNA expression remains poorly understood, mainly due to the difficulty of locating the TSSs for the primary transcripts. Consequently, a large-scale experimental annotation of regulatory elements such as promoters is not yet available. Here we used a genome-wide map of *cis*-miRNA eQTLs in the human cell line GM12878 (Lappalainen *et al.* 2013), data from The 1000 Genomes Project Consortium (2012), and miRNA promoter predictions (Marsico *et al.* 2013) to get an overview of the regulatory landscape of miRNA transcription and to answer the question of how regulatory variation affects miRNA expression.

In contrast to previous work on protein-coding genes (Lee *et al.* 2009; Gaffney *et al.* 2012; Battle *et al.* 2014) we did not aim to resolve the most-likely causal variants underlying the eQTL, but rather we built the first interpretable model to classify miRNA/SNP pairs into miRNA eQTL or non-miRNA eQTL.



	eQTL SNP	non-eQTL SNP	
intronic promoter	49 (32)	737	786
host promoter	19 (36)	841	860
	68	1578	1646

Figure 4 Overview of miRNA and promoter localization. About two thirds of pre-miRNAs in our data set are intragenic, *i.e.*, located within a host gene. About one third of the predicted promoters coincide with a host gene promoter, the majority represents independent intergenic or intragenic alternative promoters. A Fisher's exact test shows that miRNA eQTLs are significantly enriched for predicted intronic promoters as compared to host gene promoters. Shaded numbers in the contingency table are expected values computed by Pearson's Chi-squared test.

Our model achieves a classification accuracy of 85%, but one should keep in mind that eQTL studies give rise to multiple significant SNPs per gene. Of those, many may not be causal, but are in linkage disequilibrium (LD) with true causal SNPs. Nevertheless, we labeled all miRNA eQTLs as positives—including possible noncausal SNPs—to achieve higher statistical power. To study the effect of this possible mislabeling in the absence of a ground truth about causality, we monitored the effect of a reduced portion of causal SNPs on our prediction accuracy by gradually relabeling randomly sampled miRNA eQTLs as non-miRNA eQTLs. Progressively lowering the chance of including true causal miRNA eQTLs among the positives resulted in a gradual decrease of prediction accuracy which converges to a value of 65%. As expected, lowering the portion of true miRNA eQTLs also leads to models where the regulatory features no longer contribute to prediction accuracy (*i.e.*, do not lead to better AIC values, see Figure S1G), while the distance feature still remains important due to the correlation between LD and distance. Using only the single most-significant eQTL per miRNA [58 miRNA eQTLs, as this should be a reasonable approximation of a subset of causal SNPs (Lappalainen *et al.* 2013)] and relabeling the remaining miRNA eQTLs as non-miRNA eQTLs results in an accuracy of 71%, which is better than randomly selecting one miRNA eQTL per miRNA (65%). However, in doing so we also strongly decrease the number of eQTL observations from 4785 to 58,

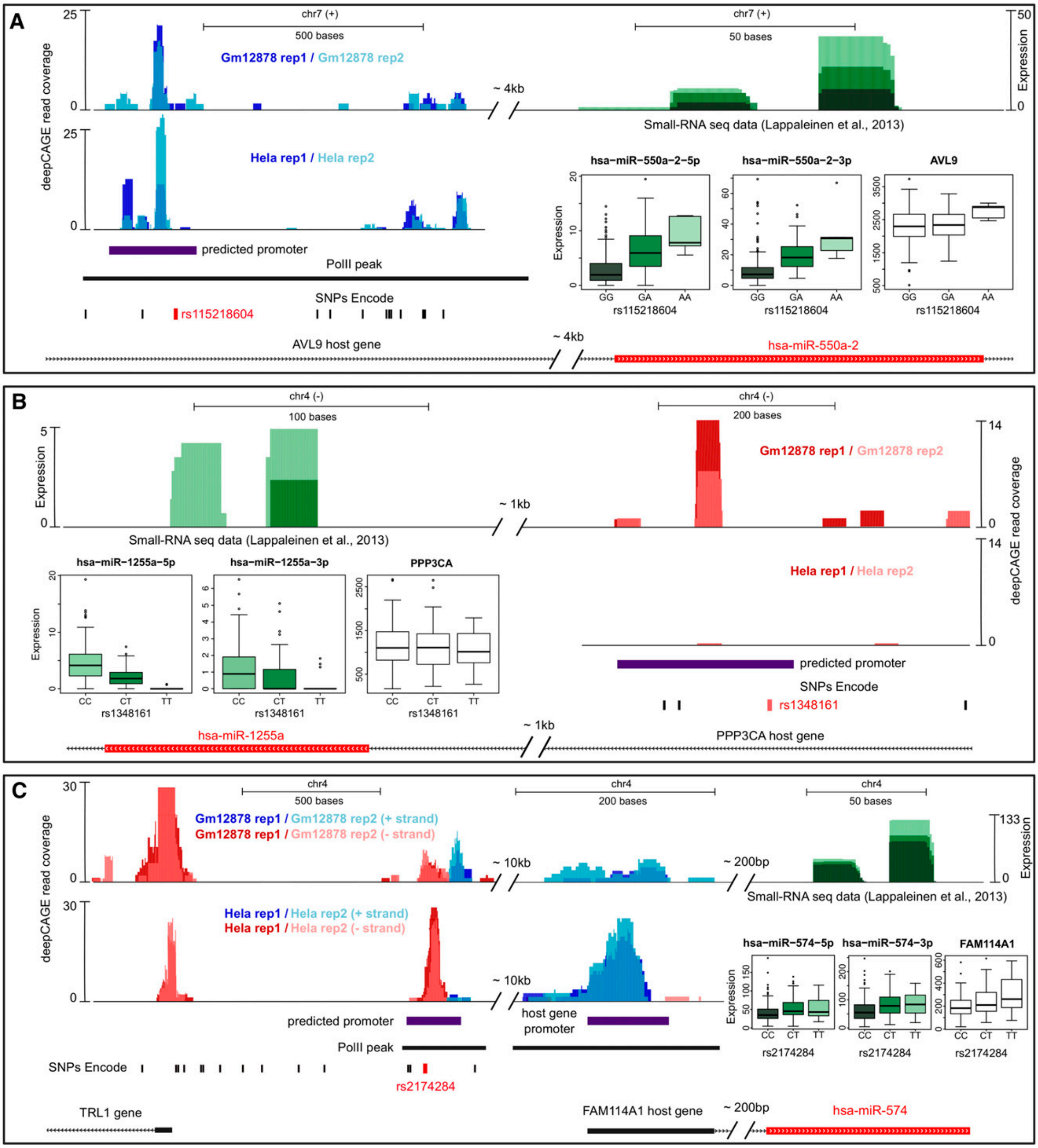


Figure 5 Effect of predicted alternative promoters on miRNA expression. Read coverage from ENCODE CAGE data are shown for GM12878 and HeLa cells for two replicates. Blue signal corresponds to read coverage on the forward (+) strand, while red signal corresponds to read coverage on the reverse strand (−). Normalized read coverage from small RNA-seq data for different genotypes is shown for the 5p and 3p miRNA arms in GM12878 only (green signal and boxplots, see *miRNA expression analysis* in *Materials and Methods*). Violet boxes represent regions around the predicted promoters. The corresponding eQTL SNPs located within the predicted promoters are marked with red. (A) Genome browser view of miR-550a-2 located in a long intron of the AVL9 gene. The miRNA eQTL *rs115218604* is located in the ±100-bp region around the predicted intronic promoter [chr7(+):32767642-32767783]. (B) Genome browser view of miR-1255a located in the first intron of the PPP3CA gene. The miRNA eQTL *rs1348161* is located in the ±100-bp region around the predicted intronic promoter [chr4(−):102252612-102252641]. (C) Genome browser view of miR-574 located in the first intron of the FAM114A1 gene. Two predicted miRNA promoters are highlighted: one corresponding to the host gene promoter and an alternative, and cell type-specific, bidirectional promoter located ∼10 kb upstream of the miRNA precursor [chr4(+):38859404-38859497]. The miRNA eQTL and mRNA eQTL *rs2174284* is located in the ±100-bp region around the alternative promoter. Transcription on the forward strand is specific to the GM12878 cell line, as indicated by the tissue-specific promoter prediction. Transcription on the reverse strand is preserved in both cell lines, corresponding most probably to an alternative promoter for the TRL1 gene.

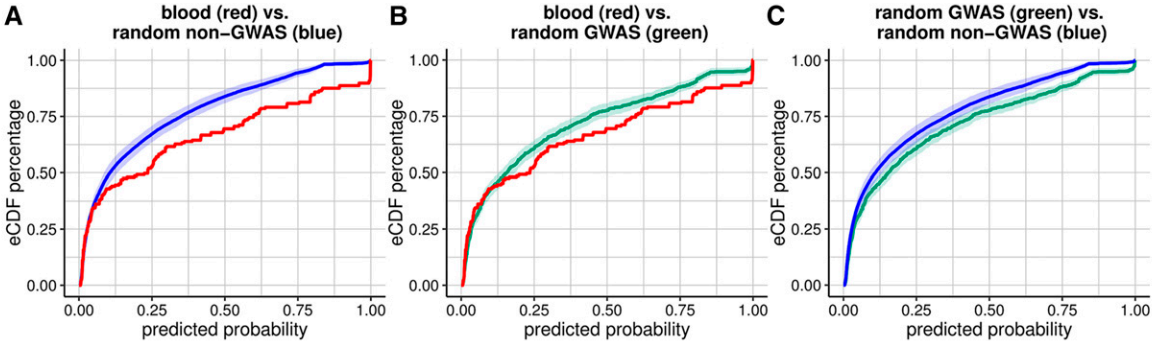


Figure 6 Prediction of GWA variants with the miRNA-eQTL model. Empirical cumulative distribution functions (eCDF) of predicted probabilities for blood GWAS SNPs, random other GWAS SNPs, and random non-GWAS SNPs. For the blood GWAS SNPs all 108 SNPs were always used, for the other two groups 100 random sets were sampled and the mean eCDFs and SDs were plotted. One-sided Kolmogorov–Smirnov tests were applied to compare the eCDFs ($P < 0.05$). (A) In 97 out of 100 cases the eCDF of blood GWAS SNPs was significantly different from the random non-GWAS SNPs eCDF. (B) In 30 out of 100 cases the eCDF of blood GWAS SNPs was significantly different from the random GWAS SNPs eCDF. (C) The eCDF of random GWAS SNPs was not significantly different from the eCDF of random non-GWAS SNPs in any cases.

losing substantial statistical power due to the smaller sample size. Dissecting the interplay of smaller sample size and having a cleaner miRNA eQTL set is not completely possible. Still, seeing that in our case the performance measurements (Figure 3) are stable for random samples suggests that the model offers a high predictive power.

The integrative analysis presented here provides new insights into patterns of *cis*-miRNA eQTLs which affect different steps of miRNA biogenesis. Several notable results emerge from our analysis. We find that both SNPs in the 5p and 3p miRNA arms, as well as around the Drosha processing sites and internal loop, can affect miRNA expression. miRNA regulation takes place at multiple steps, including processing by Drosha and Dicer. Several studies have shown that nucleotide preferences in the flanking regions of the Drosha cutting sites play a role in determining miRNA processing efficiency (Auyeung *et al.* 2013; Conrad *et al.* 2014). Although we do not detect any preferential enrichment of miRNA eQTLs for previously described motifs, we do see that miRNA eQTLs preferentially accumulate within 22 nt upstream and downstream of Drosha cutting sites when compared to more distal regions. This further supports the hypothesis that the sequence and secondary structure of the 20 nt region upstream of the 5p and downstream of the 3p are crucial for Drosha recruitment and function.

Genetic variants that modify chromatin accessibility, promoters, and transcription factor binding are a major mechanism by which genetic variation leads to expression differences for protein-coding genes in humans (Kasowski *et al.* 2010; Degner *et al.* 2012; McVicker *et al.* 2013; del Rosario *et al.* 2015; Waszak *et al.* 2015; Grubert *et al.* 2015). Here we demonstrate that miRNA eQTLs are more likely to overlap activating regulatory elements and less likely to overlap repressive features. In particular, miRNA eQTLs were significantly enriched for predicted promoters in the GM12878 cell line as opposed to unspecific promoters, which do not show enrichment. Our predicted promoters provide a snapshot of the currently active promoters in this cell line and our results indicate that miRNA eQTLs may affect promoter activity in a tissue-specific fashion. Although we cannot quantitatively determine if a certain miRNA eQTL is really GM12878-specific, due to the lack of miRNA-eQTL mapping in other cell types, we suggest that cell type-specific regulatory elements are associated to cell type-specific miRNA eQTLs. In fact, active binding sites for immune-related and/or B cell lineage-specific transcription factors are among the top 10 factors for which miRNA eQTLs are enriched.

The mechanisms of transcriptional regulation of intragenic miRNAs are more complex than for intergenic miRNAs, as intragenic miRNAs may mirror the regulatory mechanisms of their host transcripts and therefore share regulatory elements with their hosts. Here we demonstrate that the majority of intragenic miRNAs is independently regulated by their own promoter regions or *cis*-regulatory elements. Independent alternative promoters for intragenic miRNAs, mainly intronic promoters, have previously been shown to be preferentially tissue-specific, unlike host gene promoters which are preferentially ubiquitously expressed (Marsico *et al.* 2013). We also discovered that miRNA eQTLs show a significant enrichment for mRNA eQTLs, pointing to many cases of coregulation of the intragenic miRNA with its host gene. However, 74% of the shared eQTLs remain significant when conditioned on the corresponding mRNA expression level and we observed a significant enrichment of miRNA-only eQTLs for intronic promoters active in the GM12878 cell line. This supports the hypothesis of cell type-specific intronic miRNA promoters and suggests that alternative promoters of intragenic miRNAs are a rich source of causal genetic variation.

A significant fraction of miRNA eQTLs are located far upstream or downstream of the mature miRNA (up to 500 kb). This suggests that miRNAs distal regulatory elements can also interact with proximal regulatory elements, *e.g.*, via chromatin looping, regulating miRNA expression. The observation that the presence of an insulator between the SNP and miRNA significantly decreases the probability of that SNP to be causal also suggests that insulators may act as a barrier preventing chromatin looping (Grubert *et al.* 2015).

While epigenetic and genomic annotations are available for a variety of cell types and tissues and we can predict miRNA promoters with software such as PROmiRNA (Marsico *et al.* 2013) and microTSS (Georgakilas *et al.* 2014), genome-wide maps of miRNA eQTLs are available for very few cell types. Excluding the Geuvadis project, miRNA-eQTL studies either report a limited number of significant miRNA-eQTL associations (Parts *et al.* 2012; Siddle *et al.* 2014), report only *trans* associations (Gamazon *et al.* 2013), or map miRNA eQTL for a mixture of cells [*e.g.*, blood (Huan *et al.* 2015)]; making the analysis of tissue-specific regulatory elements much harder. In the coming years we also expect that miRNA-eQTL data will be available for diverse cell types and in sufficient sample sizes to train a statistical model, such as the one presented here. By exploiting the richness of regulatory annotations in different tissues, our model, which is now trained to predict miRNA eQTLs *vs.* non-miRNA eQTLs, can be applied easily to other cell lines and trained to discriminate tissue-specific miRNA eQTLs.

Similar to eQTL SNPs which associate a genomic location with transcript expression, GWAS SNPs are associated with phenotypic traits. Most GWAS SNPs are located within noncoding regions of the genome which makes the interpretation of associations difficult. By applying our eQTL model to SNPs associated with diseases/traits in the GWAS catalog, we found that the predicted probabilities for cell line-specific traits are significantly higher than those for other traits and non-GWAS SNPs, indicating that model features can be used to interpret GWAS results. More generally, predictions of our model might be useful to determine whether a cell line is relevant in explaining the effects of a set of GWAS SNPs first by analyzing the distribution of predicted probabilities.

In conclusion, our classifier not only enables for the first time an accurate discrimination between miRNA eQTLs and non-miRNA eQTLs in a cell type-specific fashion, but also helps to identify the most informative functional features that may explain the mechanisms of miRNA regulatory logic improving the interpretation of GWAS.

Acknowledgments

We thank Claudia Truppa for helping with the classification of blood phenotypes from the genome-wide association studies SNPs; and Martin Vingron, Thomas Blasi, Janine Arloth, Alena van Bömmel, and Brian Caffrey for insightful discussion and proofreading of the manuscript. The project was supported by the Freie Universität Berlin within the Excellence Initiative of the German Research Foundation. This work was also supported by funding to M.H. by the Federal Ministry of Education and Research in the project eMed:symAtrial (01ZX1408D).

Literature Cited

Auyeung, V. C., I. Ulitsky, S. E. McGeary, and D. P. Bartel, 2013 Beyond secondary structure: primary-sequence determinants license pri-miRNA hairpins for processing. Cell 152: 844–858.

Battle, A., S. Mostafavi, X. Zhu, J. B. Potash, M. M. Weissman *et al.*, 2014 Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res. 24: 14–24.

Conrad, T., A. Marsico, M. Gehre, and U. A. Orom, 2014 Microprocessor activity controls differential miRNA biogenesis In Vivo. Cell Reports 9: 542–554.

Corcoran, D. L., K. V. Pandit, B. Gordon, A. Bhattacharjee, N. Kaminski *et al.*, 2009 Features of mammalian microRNA promoters emerge from polymerase II chromatin immunoprecipitation data. PLoS One 4: e5279.

Degner, J. F., A. A. Pai, R. Pique-Regi, J.-B. Veyrieras, D. J. Gaffney *et al.*, 2012 DNase I sensitivity QTLs are a major determinant of human expression variation. Nature 482: 390–394.

del Rosario, R. C.-H., J. Poschmann, S. L. Rouam, E. Png, C. C. Khor *et al.*, 2015 Sensitive detection of chromatin-altering polymorphisms reveals autoimmune disease mechanisms. Nat. Methods 12: 458–464.

ENCODE Project Consortium, 2012 An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57–74.

Friedländer, M. R., S. D. Mackowiak, N. Li, W. Chen, and N. Rajewsky, 2012 miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res. 40: 37–52.

Friedman, R. C., K. K.-H. Farh, C. B. Burge, and D. P. Bartel, 2009 Most mammalian mRNAs are conserved targets of microRNAs. Genome Res. 19: 92–105.

Gaffney, D. J., J. K. Pritchard, G. E. Crawford, Y. Gilad, A. A. Pai *et al.*, 2012 Dissecting the regulatory architecture of gene expression QTLs. Genome Biol. 13: R7.

Gamazon, E. R., F. Innocenti, R. Wei, L. Wang, M. Zhang *et al.*, 2013 A genome-wide integrative study of microRNAs in human liver. BMC Genomics 14: 395.

Georgakilas, G., I. S. Vlachos, M. D. Paraskevopoulou, P. Yang, Y. Zhang *et al.*, 2014 microTSS: accurate microRNA transcription start site identification reveals a significant number of divergent pri-miRNAs. Nat. Commun. 5: 5700.

Gilad, Y., S. A. Rifkin, and J. K. Pritchard, 2008 Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet. 24: 408–415.

Grubert, F., J. B. Zaugg, M. Kasowski, O. Ursu, D. V. Spacek *et al.*, 2015 Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. Cell 162: 1051–1065.

GTEx Consortium, 2015 The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science 348: 648–660.

Ha, M., and V. N. Kim, 2014 Regulation of microRNA biogenesis. Nat. Rev. Mol. Cell Biol. 15: 509–524.

Hu, Y.-F., and R. Li, 2002 B potentiates function of BRCA1 activation domain 1 (AD1) through a coiled-coil-mediated interaction. Genes Dev. 16: 1509–1517.

Huan, T., J. Rong, C. Liu, X. Zhang, K. Tanriverdi *et al.*, 2015 Genome-wide identification of microRNA expression quantitative trait loci. Nat. Commun. 6: 6601.

Im, H.-I., and P. J. Kenny, 2012 MicroRNAs in neuronal function and dysfunction. Trends Neurosci. 35: 325–334.

Jansen, R. C., and J. P. Nap, 2001 Genetical genomics: the added value from segregation. Trends Genet. 17: 388–391.

Kasowski, M., F. Grubert, C. Heffelfinger, M. Hariharan, A. Asabere *et al.*, 2010 Variation in transcription factor binding among humans. Science 328: 232–235.

Kawaji, H., J. Severin, M. Lizio, A. Waterhouse, S. Katayama *et al.*, 2009 The FANTOM web resource: from mammalian transcriptional landscape to its dynamic regulation. Genome Biol. 10: R40.

Kozomara, A., and S. Griffiths-Jones, 2014 miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. 42: D68–D73.

Lappalainen, T., M. Sammeth, M. R. Friedlander, P. A. C. 't Hoen, J. Monlong *et al.*, 2013 Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501: 506–511.

Lee, S.-I., A. M. Dudley, D. Drubin, P. A. Silver, N. J. Krogan *et al.*, 2009 Learning a prior on regulatory potential from eQTL data. PLoS Genet. 5: e1000358.

Lujambio, A., and S. W. Lowe, 2012 The microcosmos of cancer. Nature 482: 347–355.

Marsico, A., M. R. Huska, J. Lasserre, H. Hu, D. Vucicevic *et al.*, 2013 PROmiRNA: a new miRNA promoter recognition method uncovers the complex regulation of intronic miRNAs. Genome Biol. 14: R84.

McKinsey, T. A., C. L. Zhang, and E. N. Olson, 2002 MEF2: a calcium-dependent regulator of cell division, differentiation and death. Trends Biochem. Sci. 27: 40–47.

McVicker, G., B. van de Geijn, J. F. Degner, C. E. Cain, and N. E. Banovich *et al.*, 2013 Identification of genetic variants that affect histone modifications in human cells. Science 342: 747–749.

Monteys, A. M., R. M. Spengler, J. Wan, L. Tecedor, K. A. Lennox *et al.*, 2010 Structure and activity of putative intronic miRNA promoters. RNA 16: 495–505.

Ozsolak, F., L. L. Poling, Z. Wang, H. Liu, X. S. Liu *et al.*, 2008 Chromatin structure analyses identify miRNA promoters. Genes Dev. 22: 3172–3183.

Parts, L., Å. K. Hedman, S. Keildson, A. J. Knights, and C. Abreu-Goodger *et al.*, 2012 Extent, causes, and consequences of small RNA expression variation in human adipose tissue. PLoS Genet. 8: e1002704.

Pasquinelli, A. E., B. J. Reinhart, F. Slack, M. Q. Martindale, M. I. Kuroda *et al.*, 2000 Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. Nature 408: 86–89.

Pasquinelli, A. E., S. Hunter, and J. Bracht, 2005 MicroRNAs: a developing story. Curr. Opin. Genet. Dev. 15: 200–205.

Payne, K. J., and S. Dovat, 2011 Ikaros and tumor suppression in acute lymphoblastic leukemia. Crit. Rev. Oncog. 16: 3–12.

Ramji, D. P., and P. Foka, 2002 CCAAT/enhancer-binding proteins: structure, function and regulation. Biochem. J. 365: 561–575.

Sayed, D., and M. Abdellatif, 2011 MicroRNAs in development and disease. Physiol. Rev. 91: 827–887.

Schebesta, A., S. McManus, G. Salvagiotto, A. Delogu, G. A. Busslinger *et al.*, 2007 Transcription Factor Pax5 Activates the Chromatin of Key Genes Involved in B Cell Signaling, Adhesion, Migration, and Immune Function. Immunity 27: 49–63.

Scully, R., S. F. Anderson, D. M. Chao, W. Wei, L. Ye *et al.*, 1997 BRCA1 is a component of the RNA polymerase II holoenzyme. Proc. Natl. Acad. Sci. USA 94: 5605–5610.

Siddle, K. J., M. Deschamps, L. Tailleux, Y. Nédélec, J. Pothlichet *et al.*, 2014 A genomic portrait of the genetic architecture and regulatory impact of microRNA expression in response to infection. Genome Res. 24: 850–859.

Sims, R. J., S. Millhouse, C.-F. Chen, B. A. Lewis, H. Erdjument-Bromage *et al.*, 2007 Recognition of trimethylated histone H3 lysine 4 facilitates the recruitment of transcription postinitiation factors and pre-mRNA splicing. Mol. Cell 28: 665–676.

The 1000Genomes Project Consortium, 2012 An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65.

Waszak, S. M., O. Delaneau, A. R. Gschwind, H. Kilpinen, S. K. Raghav *et al.*, 2015 Population Variation and Genetic Control of Modular Chromatin Architecture in Humans. Cell 162: 1039–1050.

Welter, D., J. MacArthur, J. Morales, T. Burdett, P. Hall *et al.*, 2014 The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 42: D1001–D1006.

Zeng, Y., 2006 Principles of micro-RNA production and maturation. Oncogene 25: 6156–6162.

Communicating editor: E. G. Petretto

GENETICS

Supporting Information
www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.187153/-/DC1

Principles of microRNA Regulation Revealed Through Modeling microRNA Expression Quantitative Trait Loci

Stefan Budach, Matthias Heinig, and Annalisa Marsico

DOI: 10.1534/genetics.116.187153

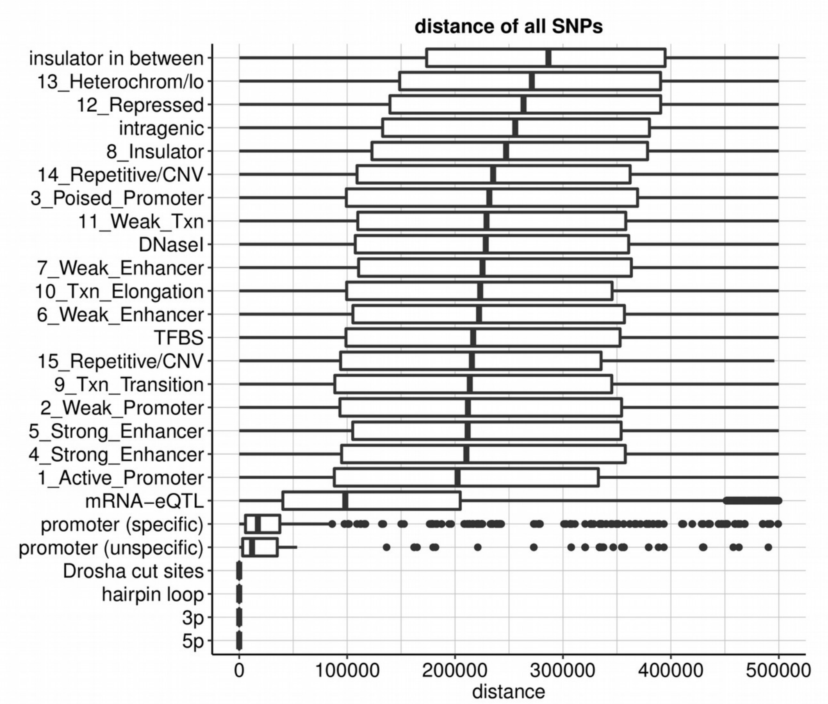


Figure S1A. Distance of all SNPs overlapping a feature with respect to the pre-miRNAs.

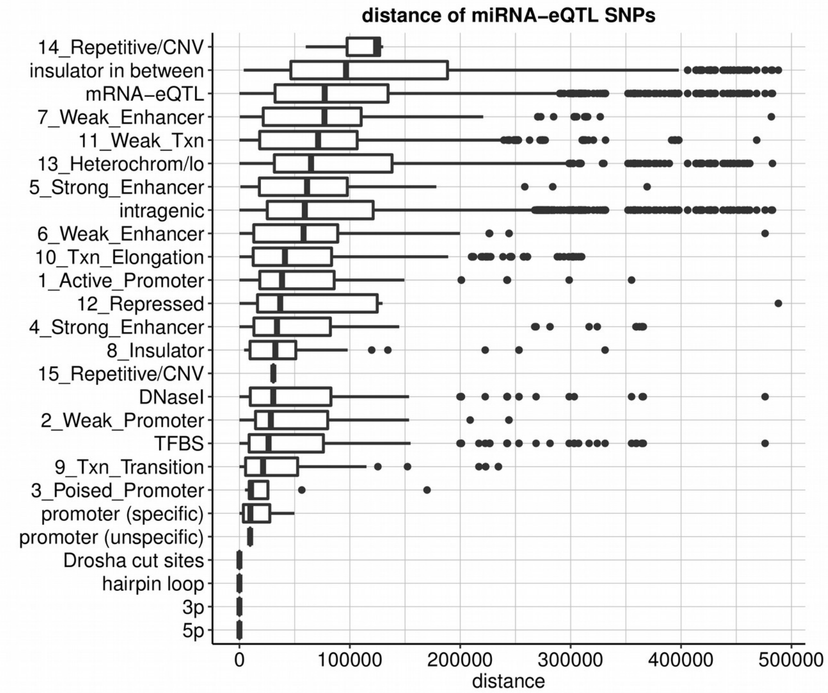


Figure S1B. Distance of miRNA-eQTL SNPs overlapping a feature with respect to the pre-miRNAs.

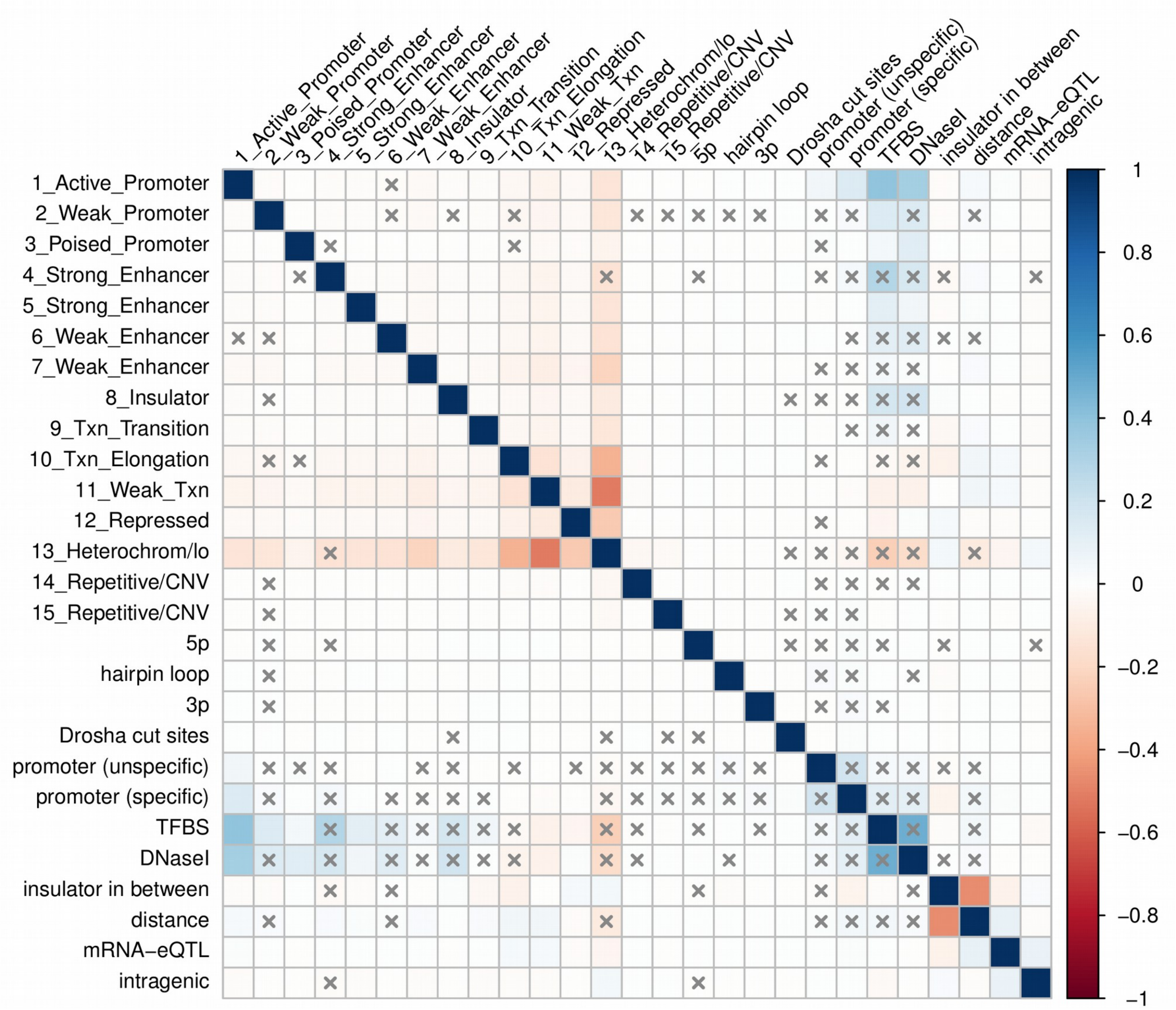


Figure S1C. Pearson's product moment correlation coefficient for all feature pairs. Crossed-out entries are not significant ($p > 0.05$).

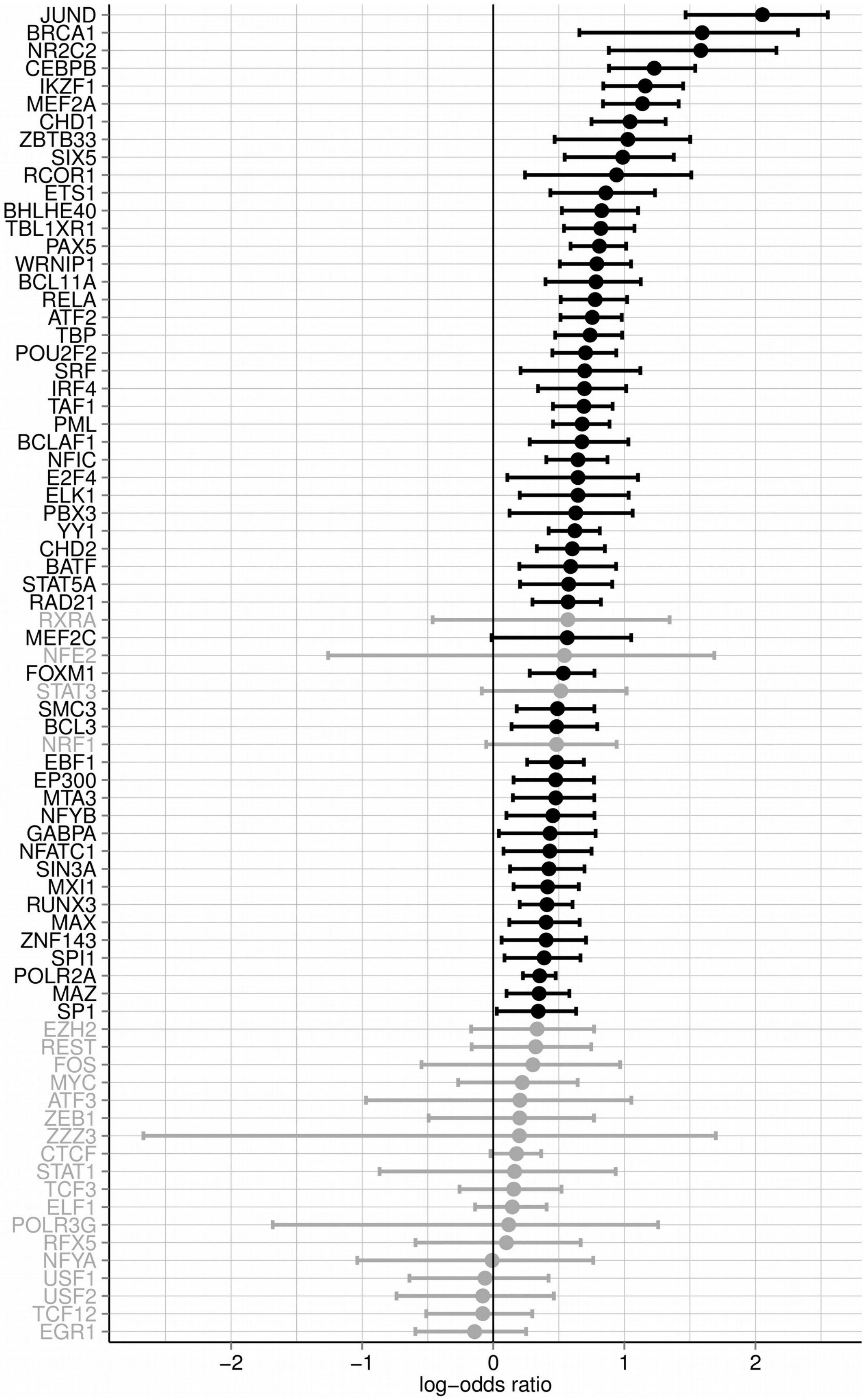


Figure S1D. TFBS selection and log-odds ratios. For each TF a separate logistic regression was performed, always including the distance as a second feature. Insignificant log-odds ratios are shown greyed ($p > 0.05$). Error bars illustrate 95% confidence intervals.

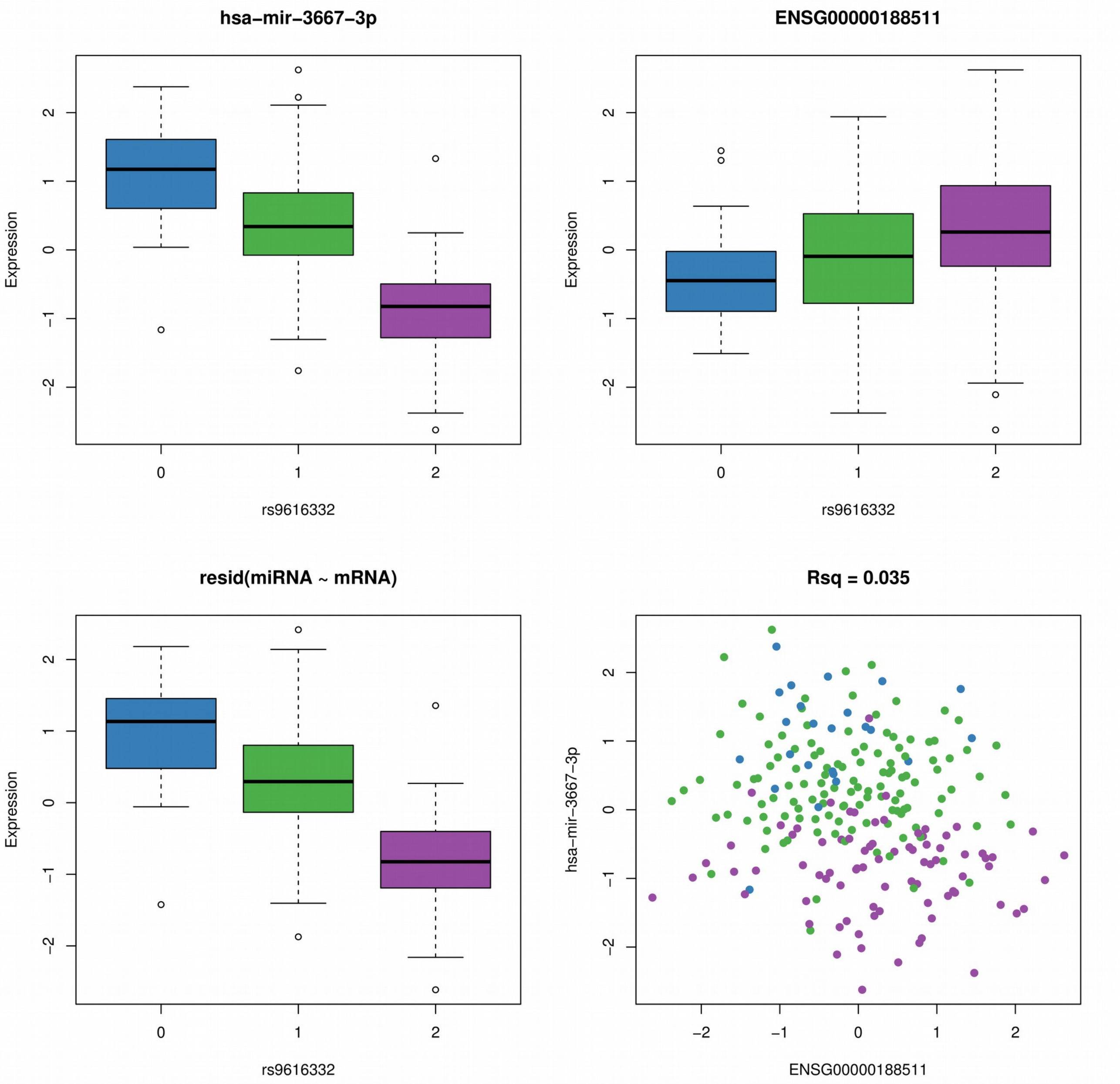


Figure S1E. Example of an independent miRNA and mRNA eQTL. Boxplots and scatterplot depict the miRNA and host gene expression (scaled and processed as described in the section 'MiRNA – Host Gene Independence Analysis' in Methods) for the three different genotypes of the respective eQTL.

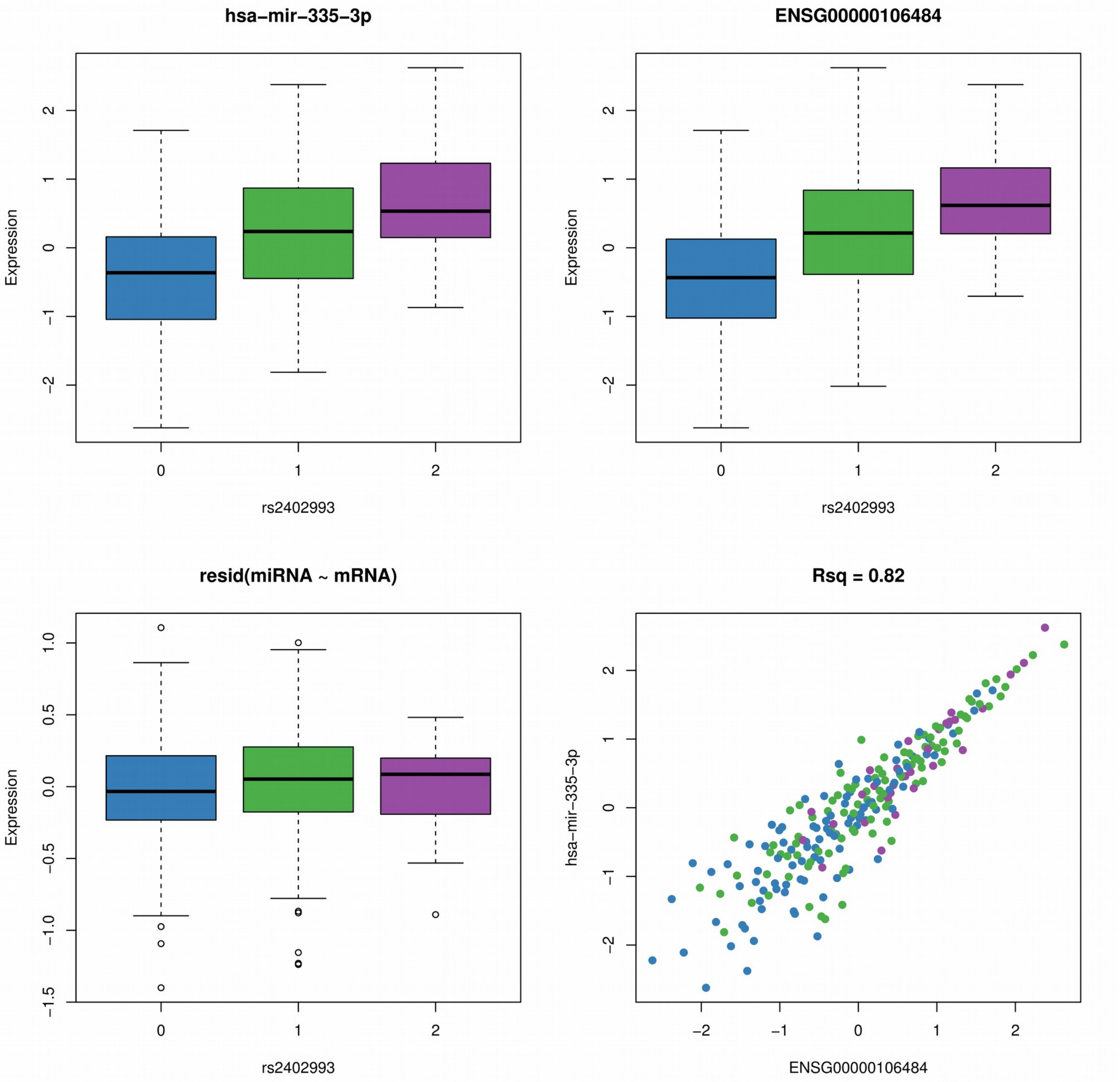


Figure S1F. Example of a shared miRNA and mRNA eQTL. Boxplots and scatterplot depict the miRNA and host gene expression (scaled and processed as described in the section 'MiRNA – Host Gene Independence Analysis' in Methods) for the three different genotypes of the respective eQTL.

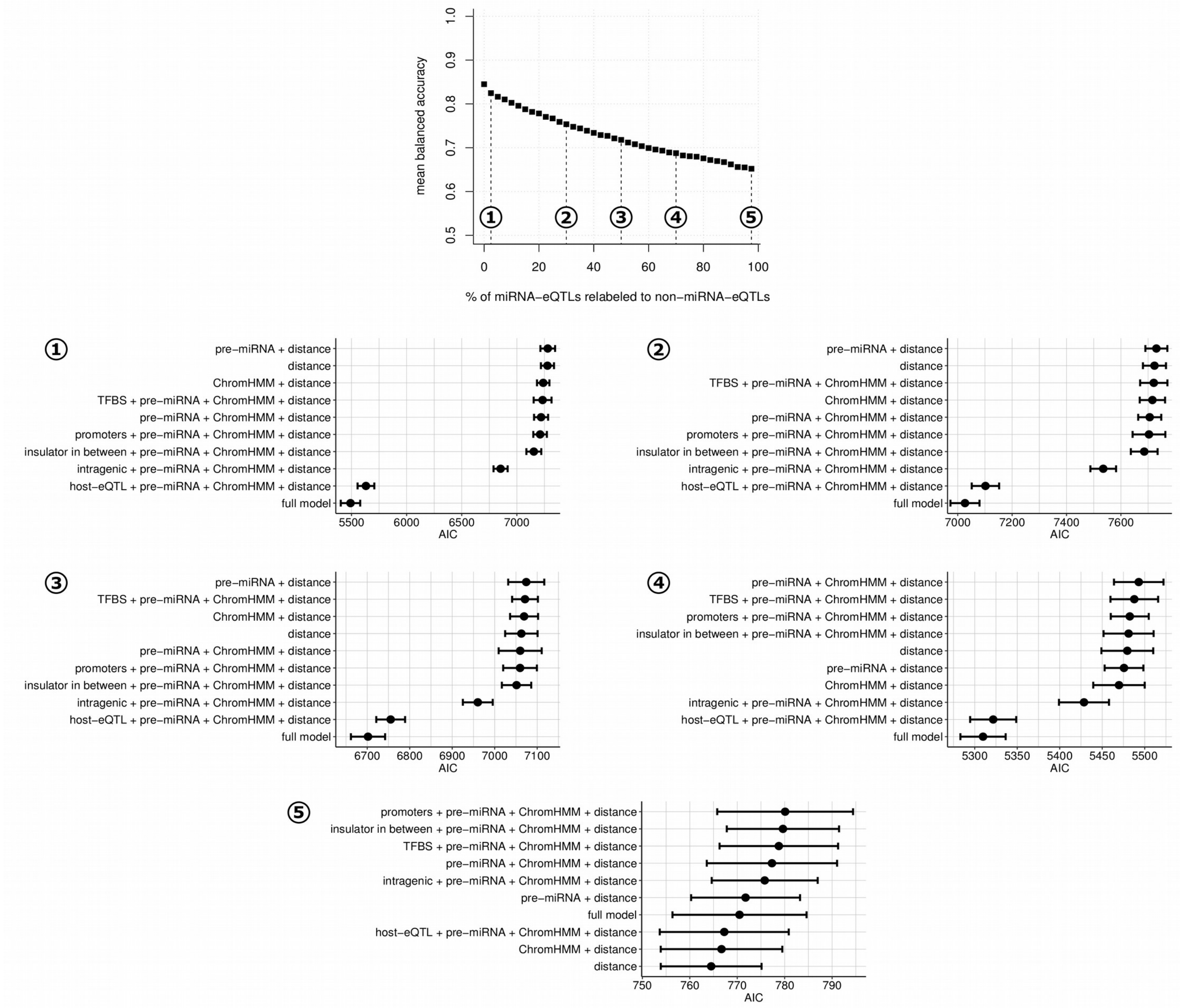


Figure S1G. Effect of randomly relabeling miRNA-eQTLs as non-miRNA-eQTLs to address the fact that not all miRNA-eQTLs used in the model are causal SNPs. The top panel shows the mean balanced accuracy (y-axis) for 25 random samples when relabeling a fraction of the eQTLs (x-axis). The dashed lines marked by circled numbers indicate fractions for which a model selection according to the AIC was performed. Results of the model selection are shown below marked by the corresponding circled numbers. These plots are analogous to Figure 3A in the main text.

Table S1. The numbers are p-values from one-sided Fisher's exact tests representing feature enrichment of different eQTL sets. (.xlsx, 12 KB)

Available for download as a .xlsx file at
www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.187153/-/DC1/TableS1.xlsx

File S1. This file contains the promoter predictions used. (.tar.gz, 99 KB)

Available for download as a .tar.gz file at
www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.187153/-/DC1/FileS1.tar.gz

File S2. This file contains several data files, including the results of the miRNA - host gene independence analysis. It also contains an R script used to build the model and to create the plots in this paper, and instructions to reproduce the calculations. (.tar.gz, 353 KB)

Available for download as a .tar.gz file at
www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.187153/-/DC1/FileS2.tar.gz