Nucleic Acids Research Advance Access published March 26, 2015

doi: 10.1093/nar/gkv256

SubmiRine: assessing variants in microRNA targets using clinical genomic data sets

Evan K. Maxwell[1,2], Joshua D. Campbell[2,3], Avrum Spira[2,3] and Andreas D. Baxevanis[1,*]

[1]Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA, [2]Bioinformatics Program, Boston University, Boston, MA 02215, USA and [3]Division of Computational Biomedicine, Department of Medicine, Boston University School of Medicine, Boston, MA 02215, USA

Received September 05, 2014; Revised February 25, 2015; Accepted March 13, 2015

ABSTRACT

MicroRNAs (miRNAs) regulate gene expression by binding to partially complementary sequences on target mRNA transcripts, thereby causing their degradation, deadenylation, or inhibiting their translation. Genomic variants can alter miRNA regulation by modifying miRNA target sites, and multiple human disease phenotypes have been linked to such miRNA target site variants (miR-TSVs). However, systematic genome-wide identification of functional miR-TSVs is difficult due to high false positive rates; functional miRNA recognition sequences can be as short as six nucleotides, with the human genome encoding thousands of miRNAs. Furthermore, while large-scale clinical genomic data sets are becoming increasingly commonplace, existing miR-TSV prediction methods are not designed to analyze these data. Here, we present an open-source tool called SubmiRine that is designed to perform efficient miR-TSV prediction systematically on variants identified in novel clinical genomic data sets. Most importantly, SubmiRine allows for the prioritization of predicted miR-TSVs according to their relative probability of being functional. We present the results of SubmiRine using integrated clinical genomic data from a large-scale cohort study on chronic obstructive pulmonary disease (COPD), making a number of high-scoring, novel miR-TSV predictions. We also demonstrate SubmiRine's ability to predict and prioritize known miR-TSVs that have undergone experimental validation in previous studies.

INTRODUCTION

MicroRNAs (miRNAs) are small RNA molecules of about 22 nucleotides in length that are processed from hairpin-loop structures formed mostly by RNA polymerase II transcripts in the nucleus. The animal miRNA biogenesis pathway is a subset of the larger RNA interference (RNAi) pathway, in which the RNAi-induced silencing complex (RISC) transports miRNAs to mRNA recognition sequences ('target sites'). Upon binding, the miRNA down-regulates the target gene's expression, predominately via mRNA destabilization and decay (1).

The miRNA regulatory mechanism was discovered in the nematode *Caenorhabditis elegans* in 1993, with the identification of small RNAs encoded by *lin-4* regulating the gene *lin-14* through binding in the 3′UTR (2). Since then, the miRNA pathway has been identified in every animal except for those in two of the earliest evolving lineages—the ctenophores and placozoans (3,4)—and many miRNAs (and their target sites) are conserved across species. In humans, the number of identified miRNAs is constantly increasing as sequencing technologies improve, with current inventories in the thousands (5,6). Many roles have been found for miRNAs in the context of the study and treatment of human disease. These include their use as disease biomarkers, as potential therapeutic molecules, and as drivers of genetic disease through mutation (7–11). Genomic variants can alter miRNA functionality through mutation of the primary miRNA's sequence, the miRNA processing machinery (e.g. Dicer, Drosha and Argonaute), or the miRNA's target sites. Mutations in the processing machinery or primary sequence can have severe downstream effects (10–12), whereas mutations that occur within miRNA binding sites likely have more subtle, localized effects, manifesting as relatively moderate deregulation of gene expression. Thus, the mechanistic effects of variants in the miRNA processing machinery and primary sequences are relatively easier to predict than variants in miRNA target sites. Furthermore, mapping genomic variants onto functional miRNA target sites is significantly more difficult than mapping to primary sequences (and the genes encoding miRNA-processing proteins), as determining functional loci for miRNA target sites is not trivial. The significance of variants in non-coding regions of the genome

*To whom correspondence should be addressed. Tel: +1 301 496 8570; Fax: +1 301 480 2634; Email: andy@mail.nih.gov

(13) and the role of gene expression in driving human disease phenotypes (14) suggest that variants in miRNA target sites are important to human disease susceptibility and progression. It has been shown that miRNA binding sites are under selection (15,16), providing further evidence that disrupting their recognition sequences can have significant phenotypic effects. Numerous miRNA target site variants (miR-TSVs) have been identified and linked to human diseases (see Supplemental Table S2), mostly through candidate gene approaches. However, with the increasing number of GWAS hits being identified and with the advancements being made to technologies for whole genome-scale analysis in clinical applications, it seems likely that many more miR-TSVs will be uncovered. This points to an urgent need for methods that can be used to systematically predict miR-TSVs in a genome-wide fashion.

Despite the fact that miRNAs were discovered over ten years ago, methods to confidently identify and predict miRNA target sites are plagued by high false positive rates. This is due in large part to our limited knowledge of how miRNA regulation is directed. For example, miRNAs have generally been understood to bind the 3′UTR of their targets at sites containing perfect Watson-Crick complementarity to the miRNA seed region—nucleotides 2–7 (and possibly 8) from the 5′ end of the miRNA. Additionally, having an adenine on the 3′UTR across from the first nucleotide of the miRNA is thought to enhance the accessibility of RISC (1,17). Using these criteria, a six-to-eight nucleotide sequence is generally sufficient for miRNA target recognition. Thus, candidate miRNA target sites occur quite frequently, yet very few of these are likely to be functional. The increasingly large number of known human miRNAs further exacerbates this issue, as it increases the number of unique 6–8mer sequences that match putative miRNA target sites. In addition, miRNA binding sites that contradict the requirement for perfect seed pairing have been identified (18), and recent high-throughput screens suggest that such non-canonical binding sites (and binding sites outside of the 3′UTR) are potentially more common than has been appreciated (19,20). It is not yet clear whether these non-canonical binding sites function to the same degree (or even by the same mechanism) as canonical binding sites, and bioinformatic methods have not yet been developed to account for all of their activities.

Nonetheless, many methods to predict functional versus non-functional miRNA target sites have been developed, and their accuracy has been demonstrated by comparing their predictions against experimental data (21–26). These methods utilize features such as site conservation, neighborhood sequence context, and thermodynamic properties to distinguish functional sites. As a consequence of this complexity, running any one target prediction program genome-wide can be a laborious process. Therefore, the predictions made by these methods are generally run once on the reference genome, with the results then made available in an online database. While having this kind of public resource is helpful for common tasks such as identifying known miRNA target sites, it does not address our ability to analyze the effect of sequence variants on miRNA functionality. A few of the aforementioned target prediction methods provide source code to run custom sequences locally (21,22,24,26), but analyzing the effects of sequence variation on miRNA binding genome-wide is beyond the scope of these methods. Historically, clinically associated miR-TSVs have been identified by a candidate gene (or SNP) approach, limiting the scale of the search enough to make this analysis feasible. Thus, the issues with using these target prediction methods for analyzing miR-TSVs are multifold: (1) running on a custom data set is non-trivial; (2) running on a custom data set is time consuming; and (3) the methods were not designed to run on multiple alleles, so comparing allelic differences must be done manually. While tools have been developed for analyzing damaging variants in coding regions of the genome, such tools for analyzing lower-impact variants outside of coding regions have not been extensively developed [see (27) for a recent review].

Recently, a few methods have been developed for analyzing miR-TSVs across the genome. These methods utilize target prediction methods (as described above) to score the alleles of variants independently and identify miR-TSVs through changes in the associated target scores. However, similarly to the target prediction tools, they generally present only the results of publicly reported variants and miRNAs in online databases or supplemental tables (28–34). The restricted set of variants and miRNA sequences they can analyze limits their usefulness and applicability to the large-scale clinical genomic data sets that are more frequently becoming available. Furthermore, these methods are designed to be run *independently* of experimental data (i.e. miRNA expression and gene expression). Therefore, these methods produce many miR-TSV predictions that are likely false positives; these methods also do not provide a mechanism to prioritize those that are most likely to be functional. In this work, we present an efficient open-source software tool called SubmiRine that has been designed specifically to address these issues, providing a powerful method for systematically analyzing miR-TSVs genome-wide that is especially suited for use in a clinical research context. SubmiRine performs miR-TSV prediction *de novo*, allowing for the analysis of novel variants and miRNAs, integration of miRNA and gene expression data, and prioritization of miR-TSV predictions by relative significance with respect to a data set-specific background model.

MATERIALS AND METHODS

Data

SubmiRine was developed and tested using a large clinical genomic data set stemming from the Lung Genome Research Consortium (LGRC; http://www.lung-genomics.org) investigating genetic mechanisms related to chronic lung disease. From these data, we utilized a subset of samples that correspond to lung biopsies classified as either chronic obstructive pulmonary disease (COPD; 116 samples) or control (43 samples); matched genotype, gene expression, and miRNA measurements were available for all of these samples. First, normalized gene expression data from Agilent microarrays were downloaded from GEO (accession GSE47460). Samples were genotyped with the Illumina HumanOmni 2.5M Beadchip, which measures ∼2.5 million SNPs having a minor allele frequency (MAF) of 2.5% or greater, designed around the 1000 Genomes Project

data. Lastly, miRNA expression was measured via small RNA sequencing with Illumina's GAIIx and HiSeq and batch corrected with Combat (35). Trimmed, size-selected reads were mapped to hg19 and miRBase v18 (6), and quantified by $\log_2$ transformation of reads per million (RPM) values normalized within each sample, using a pseudocount of one for each microRNA. Having this type of robust clinical genomic data set in-hand provides a powerful opportunity for looking at miRNA regulatory variants by integrating multiple relevant data points matched to each clinical sample. Specifically, miRNA expression data allows for the identification of both known and novel miRNAs detected in the sample, as well as their isomiRs (i.e. miRNA species processed upstream or downstream of the canonical 5′ locus, producing a miRNA with a shifted seed sequence) (36,37), thereby improving our ability to identify known and novel target sites that are likely to be active *in vivo*. The combination of gene expression and genotype data allows for the identification of variants of clinical interest (through genome wide association), as well as variants that may alter gene regulation, thereby being candidate miR-TSVs. Thus, we have used this rich data set as a representative of the data that can be utilized from modern, large-scale clinical genomic studies with respect to miRNA regulation. We present the results of SubmiRine using this data set as a proof-of-concept of our method, but also to predict novel miR-TSVs that may relate to COPD susceptibility and progression. As the number of genome-wide association studies (GWASs) continues to grow, leading to the identification of more and more human genomic variants associated with disease phenotypes, methods such as SubmiRine will allow for the direct evaluation of possible biological mechanisms responsible for the underlying phenotype.

Preprocessing clinical genomic data for targeted search of microRNA target site variants

Clinical genomic data sets that integrate genotype, gene expression, and miRNA expression information provide an ideal platform for enabling systematic, genome-wide identification of functional miR-TSVs. However, no existing method is designed to utilize these data for performing dynamic searches of miR-TSVs. SubmiRine was designed specifically to harness these data sets for disease context-specific prediction. Figure 1 represents the standard SubmiRine workflow that is used in this manuscript and will be described in the following sections. First, we utilized standard methods for analyzing our COPD clinical genomic data to pre-process variants and miRNAs of clinical interest as input for SubmiRine. Specifically, using genotype, gene expression data, and RefSeq protein annotations, we first identified *cis*-eQTLs localized to 3′UTRs using MatrixEQTL (38). This allowed us to focus on variants that correspond to a gene expression phenotype consistent with the hypothesis of a functional miR-TSV. Second, we used Plink (39) to filter out variants that were not associated with the disease phenotype. This step is optional if no disease association is being tested. Together, we used these filtered variants to generate a single FASTA-formatted file containing all candidate 3′UTR alleles and their relative expression. Next, normalized miRNA expression values were used to identify candidate miRNAs that are present in the sample. A second FASTA formatted sequence file was generated for each identified miRNA, with its mean expression recorded. Notably, this allows for isomiRs to be considered in addition to canonical miRNAs. These two FASTA files are the output of the pre-processing steps for the standard SubmiRine workflow diagrammed in Figure 1A and are the only required input to SubmiRine. Thus, alternative pre-processing procedures can be utilized without affecting SubmiRine's functionality. SubmiRine's use of these two FASTA files as input exemplifies the major advantages SubmiRine exhibits over other miR-TSV prediction tools (Table 1): it allows for analysis of novel and multi-allelic variants (both polymorphisms and indels), analysis of novel miRNAs, and prediction informed by expression data. Additionally, SubmiRine can perform traditional miRNA target site prediction where the input 3′UTR sequences are assumed to be mono-allelic. Sample input files and the source code can be found online at http://research.nhgri.nih.gov/software/SubmiRine.

Prediction of microRNA target site variants

SubmiRine identifies and prioritizes candidate miR-TSVs using a multi-step process. The first step involves the *SubmiRine_Search* module, using the pre-processed clinical data as its input (Figure 1B). First, candidate miRNA target sites are identified on all 3′UTR alleles in preparation for scoring with TargetScan6 context+ scores (21). Context+ scores require that miRNA binding sites be canonical 6mer, 7mer-1a, 7mer-m8 or 8mer sites. Thus, the identification of all candidate binding sites requires searching for all 6–8mer sequences corresponding to the seed sites of the input set of expressed miRNAs. SubmiRine performs this search rapidly against a Burrows-Wheeler transform of the input 3′UTR sequences. Once all candidate miRNA binding sites are identified, each context+ score is computed using the sequence neighborhood surrounding each site.

Multiple studies have demonstrated the importance of miRNA abundance in predicting functional target sites (25,40). Logically, a miRNA must be present and expressed at high enough levels to significantly repress its targets, and individual targets must compete for available miRNAs. Thus, after predicting all context+ scores, SubmiRine utilizes miRNA expression values in order to weight target scores by abundance of the candidate miRNA. This is computed by multiplying the miRNA's normalized expression value by the raw context+ score of each candidate binding site. The raw context+ score is referred to as the 'binary score,' and the miRNA abundance-weighted score is referred to as the 'empirical score.' We retain both the binary and empirical scores, as they each may be meaningful for assessing candidate miR-TSVs. For example, given that miRNA abundance values can vary significantly between miRNA species, the most highly expressed miRNAs (e.g. let-7) can generate extremely high empirical scores for very low-scoring candidate targets. Thus, while the empirical score has been shown to reduce false positive rates (25), the unweighted binary score is also important to consider.

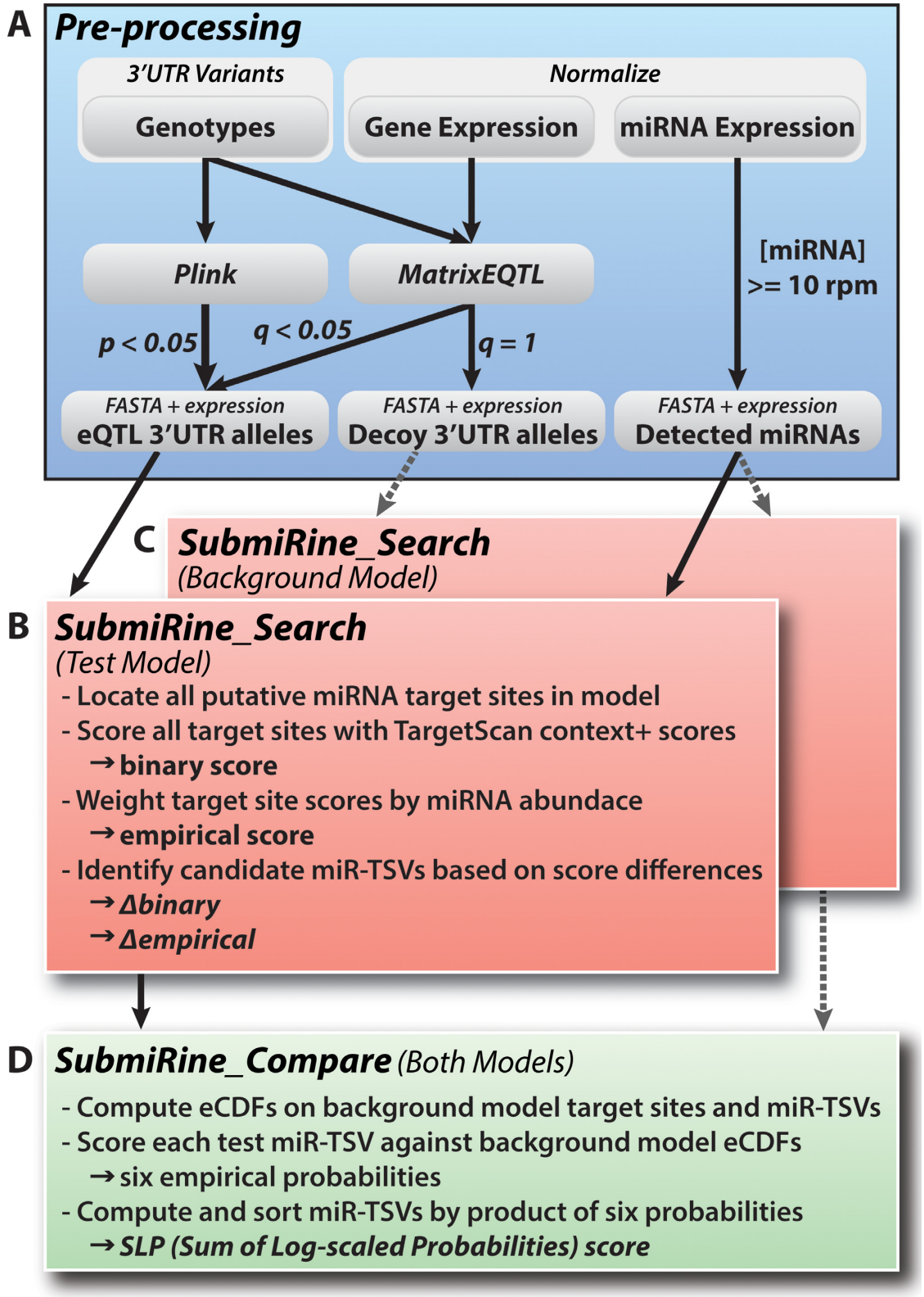

Figure 1. The standard SubmiRine workflow. (**A**) Pre-processing steps used to select candidate variants (the 'Test Model') and decoy variants (the 'Background Model') for miR-TSV prediction. The *SubmiRine_Search* module is then run independently on (**B**) the Test Model and (**C**) the Background Model. *SubmiRine_Search* takes two FASTA files as input, one for 3′UTRs and one for miRNAs, and each FASTA record contains representative expression values. *SubmiRine_Search* outputs the scored set of candidate miR-TSVs identified in the input model. (**D**) The *SubmiRine_Compare* module is then used to prioritize the miR-TSVs from the Test Model by comparing them to the decoy miR-TSVs from the Background Model. *SubmiRine_Compare* computes the SLP score (Sum of Log-scaled Probabilities), representing the joint, empirical probability of the scores computed for each candidate miR-TSV.

Table 1. Comparison of miR-TSV prediction methods

Variant types	SubmiRine	mrSNP	PolymiRTS3	MirSNP	MicroSNiPer	mirsnpscore	Patrocles[a]
Single Nucleotide Polymorphisms (SNPs)—publicly reported	✓	✓	✓	✓	✓	✓	✓
Single Nucleotide Polymorphisms (SNPs)—novel	✓	✓			✓		
Insertions/deletions (INDELs)—publicly reported	✓		✓	✓			
Insertions/deletions (INDELs)—novel	✓						
Combinations of variants in phase/haplotypes	✓				✓	✓	
Variants with >2 alleles	✓						
Non-3′UTR variants							
Testable limit (maximum no. of variants per run)[b]	None	None	None	None	6	1	NA
MicroRNA sequences							
Publicly reported (miRBase)	✓	✓	✓	✓	✓	✓	✓
isomiRs	✓						
Novel miRNAs	✓						

For each miR-TSV prediction method (columns), the types of variants, variant relationships, and microRNA sequences that can be analyzed (rows) are indicated by check marks. Aside from SubmiRine, the only methods that can handle novel variants—to our knowledge—are mrSNP and MicroSNiPer. However, these can only handle novel bi-allelic SNPs and score them independently of other variants that may occur in the region.

[a]The Patrocles online database is non-functional at the time of this writing. Its stated abilities are based upon the description provided in the associated manuscript (28).

[b]The testable limit refers to, for a hypothetical list of variants of interest to the user, how many variants the public version of each tool allows the user to test at once. 'None' implies no limit exists. Regarding mirsnpscore, runs can be based upon a single SNP, a single gene or a single miRNA. Note that tools with a limit could be downloaded and run (or queried) locally to bypass any limitation.

After all candidate target sites have been identified and scored, SubmiRine compares all target sites across the set of alleles for a particular 3′UTR and identifies miRNA target sites whose score differs in at least one allele. If a target site does not exist on a particular allele, its score is considered to be zero. This score difference allows for identification of variants that create or destroy miRNA binding sites altogether, as well as variants that merely alter the predicted binding strength of a particular miRNA between alleles. Collectively, these variant-miRNA pairs represent the set of candidate miR-TSVs. Finally, because the input 3′UTR alleles were derived from *cis*-eQTLs and therefore have allele-specific expression values, SubmiRine ignores all candidate miR-TSVs that have scoring differences inconsistent with the direction of change in gene expression. Following this comparison, SubmiRine produces one output file containing all of the raw target site predictions (with both binary and empirical scores), and a second output file reporting only candidate miR-TSVs with target site scoring differences between alleles. miR-TSVs are reported by their change in binary and empirical score ($\Delta binary$ and $\Delta empirical$), and the scores of the strongest target site (i.e. the most negative) in the allele group are also reported. Additionally, two similar files are produced where all target sites for a particular miRNA that occur in the same UTR are summed, representing a miRNA-wise ('mir-wise') view of target prediction opposed to the site-specific ('site-wise') view described above [see (21,23,24,41) for more detail]. In this work, we focus on the site-wise predictions because we do not have a large density of SNPs per 3′UTR, but the mir-wise predictions are produced for cases where they may provide more global information relative to site-wise predictions.

Defining a background model with decoy variants

The list of candidate miR-TSVs identified in the *SubmiRine_Search* module can be quite large, especially for sets of 3′UTRs and miRNAs generated from genome-wide scans. Thus, determining which miR-TSVs are the most likely to be functional (and have the strongest impact) requires prioritizing the list of predicted miR-TSVs. SubmiRine prioritizes the predicted miR-TSVs by comparing them to a background model consisting of decoy 3′UTR variants. Here, decoy variants are defined as genomic variants that occur in a 3′UTR but do not correlate with allele-specific gene expression. Because SubmiRine was designed to run on custom clinical genomic data sets, the decoy variant set can be generated alongside the clinically relevant set during standard pre-processing (Figure 1A).

During pre-processing, we identified local 3′UTR *cis*-eQTLs (i.e. variants in 3′UTRs whose genotype is correlated with expression of the underlying gene) to select variants that may be functional miR-TSVs. MatrixEQTL models the genotype-gene expression interaction as a quantitative trait using a linear model (or, optionally, a variation-based model), and assigns a p-value to each variant based upon its probability of being a true eQTL. While likely eQTLs are selected on the lower tail of this p-value distribution where $p \approx 0$ (we used a cutoff of FDR = 0.05), unlikely eQTLs are conversely predicted at larger values of p, with the most confident negative predictions occurring on the upper tail of the distribution where $p \approx 1$. These correspond to variants that do not fit a linear model relative to the gene expression quantitative trait, or fit a linear model with a slope ≈ 0. Thus, our background model contains a set of 3′UTR variants that lie on the upper tail of the p-value distribution produced by MatrixEQTL, representing true decoy variants. While comparisons of variant sets at different p-value intervals indicate that most variants outside of the lower tail of the eQTL p-value distribution can be used for building a background model, our selection of variants specifically from the upper tail does a slightly better job of selecting for miR-TSVs with low-scoring Δbinary and Δempirical metrics. More importantly, we confirmed that

almost any background model based on observed variants was preferable to a background model based on simulated sequence variants, which have no experimental evidence of being non-functional miR-TSVs and are not representative of true negative examples (see the supplemental text and Supplemental Figures S2–S6 for more detail). Using the 3′UTRs with the selected decoy variants (i.e. the background model), *SubmiRine_Search* is run a second time in the same fashion as for the clinically relevant set to generate decoy miR-TSVs (Figure 1C). These decoy miR-TSVs can be used to determine, for each miRNA, how often particular target site scores (and variant-driven scoring changes) occur in a non-functional, background model.

Prioritizing predictions by microRNA regulation-altering potential

Qualitatively, to prioritize predicted miR-TSVs for further clinical study, each candidate miR-TSV should be assessed for the repressive strength of the created (or destroyed) binding site, the magnitude of the variant's effect on miRNA binding, the availability (abundance) of the miRNA, and the relative significance of these metrics compared to others in the genome. SubmiRine utilizes the background model to make these assessments quantitatively in the *SubmiRine_Compare* module (Figure 1D). Specifically, for each predicted miR-TSV from the clinically relevant set (Figure 1B), SubmiRine compares the binary score, empirical score, *Δbinary* score and *Δempirical* score to the distribution of corresponding metrics observed in the decoy (non-functional) miR-TSVs (Figure 1C). Using these four scoring metrics, *SubmiRine_Compare* computes a total of six empirical probabilities reflecting how common each metric for a candidate miR-TSV is expected to be in non-functional data. The six probabilities of this form are defined in Supplemental Table S1. In practice, a series of empirical cumulative distribution functions (eCDFs) are built with the background model to compute these probabilities directly. To prevent probabilities of 0 and 1, each eCDF contains two extra pseudocount values (one above the observed maximum and one below the observed minimum) such that empirical probabilities follow Laplace's Rule of Succession. Note that two of the six empirical probabilities are computed relative to a background set containing only the subset of decoy miR-TSVs that correspond to the miRNA associated with the candidate miR-TSV being tested; this allows us to model competition among sites of a particular miRNA implicitly.

Once these six empirical probabilities have been computed against the background model, *SubmiRine_Compare* computes the product of all six probabilities to produce a single, combined metric by which all predictions can be prioritized. We call this metric the SLP (Sum of Log-scaled Probabilities) score, computed as a natural log. Thus, we consider each of the probabilities to be equally weighted, as the scoring scheme is unsupervised. To improve this metric would require a large number of validated predictions to distinguish true from false positives, yet no such data set currently exists. However, we have found that the unweighted product performs well with the set of known miR-TSVs we were able to test, suggesting all six empirical probabilities are meaningful and have predictive value (see Results).

Implementation details

In order to accommodate large-scale, clinical genomic data sets, SubmiRine was designed to be an open-source program that can run efficiently on a single processor with sufficient RAM. Comparatively, every existing target site and miR-TSV prediction method is primarily presented as a precomputed database and are not designed for runs with custom sequence data in the context of a high-throughput analysis pipeline (21,23–25,28–33,42). To our knowledge, several target prediction methods do provide open-source code that can be installed and run locally (21,22,24,26), but no such miR-TSV prediction tool currently exists. Genome-wide scans using the open-source target prediction tools mentioned above are inefficient in the context of an analysis pipeline, as they require scanning the genome once for every input miRNA to identify candidate target sites. Additionally, many require computationally intensive steps within the scoring framework, including miRNA-to-target sequence alignments (21,22,24,26), secondary structure predictions (22,24,26), and analysis of target site conservation across species (21,26). To perform target prediction efficiently in SubmiRine, we utilized TargetScan6 context+ scores for a few reasons. First, context+ scores are computed independently of target site conservation, making them well-suited for scoring variant effects in miRNA targets, but also avoiding the need to align input 3′UTR sequences across species. Second, context+ scores are based on candidate target sites that follow the rules of canonical seed pairing (i.e. 6mer, 7mer-1a, 7mer-m8 or 8mer target sites). This allows miRNA binding site candidates to be identified via simple string searches, where the number of search strings is linear with respect to the size of the input miRNA set. Thus, SubmiRine indexes the full set of 3′UTR sequences with a Burrows-Wheeler transform, allowing all candidate target sites to be identified rapidly without scanning the genome once for every miRNA. If necessary, future versions of SubmiRine could incorporate certain non-canonical seed sites (such as bulge sites (20)) by extending the set of possible seed sequences for each miRNA. SubmiRine utilizes the open-source 'py-burrows-wheeler' implementation of the Burrows–Wheeler algorithm (http://code.google.com/p/py-burrows-wheeler/). Third, TargetScan6 context+ scores do not require *ad hoc* RNA secondary structure predictions, which are computationally expensive—especially when analyzing a large set of miRNAs and 3′UTRs. Although secondary structure predictions are not performed, the 'AU content' metric that contributes to TargetScan6 context+ scores has been shown to be highly correlated with free energy estimates of target site accessibility (i.e. ΔG-open) (22,26). Also, the seed-pairing stability (SPS) metric pre-computed for TargetScan 6 context+ scores reflects the free energy estimate of miRNA seed/target binding. Therefore, our use of TargetScan 6 context+ scores does consider secondary structure information to some extent, but without the caveat of having to predict binding structures on the fly. TargetScan6 context+ scores do require minimal miRNA-to-target se-

quence alignment in order to assess 3′ supplemental pairing contributions, and this step remains the most intensive part of SubmiRine, consuming over half of the runtime. Together, these implementation details greatly improve scalability for genome-wide target prediction.

SubmiRine utilizes a Python implementation of TargetScan developed within the framework of the miRmap tool (26), with modifications to account for a more recent version of TargetScan (version 6) that includes the seed pairing stability and miRNA target abundance scoring features (21). While miRmap has external library dependencies for certain scoring features, we do not utilize these libraries for TargetScan scoring, and therefore do not require such dependencies. The core scoring algorithm in the SubmiRine framework is the *SubmiRine_Search* module written in Python, which is designed to run on either a clinically relevant set of 3′UTRs or a background model of decoy 3′UTR variants. In most cases (including the application described in this manuscript), it is desirable to compare the output of the clinically relevant set against the background model and compute the empirical probabilities used for miR-TSV prioritization. This is performed with the *SubmiRine_Compare* module—an independent R script—that accepts two output files from the *SubmiRine_Search* module as input (i.e. the output from the clinically relevant model and the background model). Sample use cases are provided with the distribution of the SubmiRine source code online at http://research.nhgri.nih.gov/software/SubmiRine.

Performance

As described above, the majority of existing target prediction and miR-TSV prediction tools are not open-source, so we could not compare the efficiency of SubmiRine to all of these methods. However, we were able to compare the miRNA target site search and scoring steps of SubmiRine to the equivalent steps of TargetScan6 (context+ scores only) (21), miRanda (24), and PITA (22). As a benchmark data set, we used all 3′UTR sequences in Ensembl that have corresponding RefSeq protein identifiers and generated two alleles per 3′UTR by adding a single, simulated SNP or single-nucleotide deletion on each UTR. This included 18 605 3′UTRs that are found genome-wide, resulting in 37 210 3′UTR alleles. In total, this data set comprised roughly 47.62 Mb of sequence. As the miRNA input set, we utilized all mature sequences for human miRNAs from miRBase (6), representing 2042 miRNA species at the time of this writing. We anticipated that this benchmark data set would represent the largest input data set that might be used, although it is significantly larger than what we generated from the COPD clinical genomic data set. Comparatively, with the COPD data, our test and background models contain 8.08 Mb and 336.6 kb of 3′UTR sequence, respectively, and 418 miRNAs were identified with expression above 10 RPM (see above).

In order to make a fair comparison between different methods, we searched only for canonical seven to eight nucleotide seed sites with each method, allowing no mismatches or GU-wobbles (-strict option in miRanda). As SubmiRine performs additional functions beyond target site identification and scoring, we only considered runtime through these steps, which are restricted to the *SubmiRine_Search* module. Specifically, this includes the computation of the Burrows-Wheeler Transform, candidate target site searching, and scoring with context+ scores. Notably, PITA was not scalable to this large of a data set, so we calculated its runtime on a small sample of the input 3′UTR sequences (all miRNAs) and extrapolated its runtime accordingly. SubmiRine processed the benchmark data set in 2.35 h, representing a >6-fold increase in speed relative to the next fastest comparison tool (TargetScan6, at 14.16 h), and speed improvements of roughly 11-fold and 400-fold to miRanda (26.26 h) and PITA (1023.59 estimated hours), respectively. These speed-ups are not surprising given the computationally intensive secondary structure and sequence alignment steps performed by miRanda and PITA, but the improved speed when compared to TargetScan6 is purely a result of faster candidate target site identification via the use of a Burrows–Wheeler transform in SubmiRine.

Because SubmiRine has to compare target sites to identify miR-TSVs, each candidate target site must be stored in memory; thus, the memory footprint can be quite large, scaling linearly with the number of targets identified. Thus, we can compute a rough estimate of the memory (RAM) requirements *a priori* by estimating the expected number of target sites, which is a function of the amount of UTR sequence and the size of the input miRNA set being analyzed (i.e. an estimate of the proportion of k-mers that will match a miRNA seed site, extrapolated according to the amount of UTR sequence). Realistically, the true number of target sites identified will also be dependent on the redundancy of the input miRNA set and the non-randomness inherent to UTR sequence data, and additional memory overhead is required outside of storing target sites. Nonetheless, our runs on the benchmark and COPD data sets suggest that 3–5 kb of RAM per expected target site is required for larger data sets, where the number of expected target sites, $E[|T|]$, can be estimated by the equation:

$$E[|T|] = \frac{M}{4^s}[D - U(S-1)]$$

where M is the number of miRNAs, S is the minimum seed site length, D is the amount of DNA sequence (in bases) and U is the number of UTR records. Peak memory usage was 25.4 Gb for the benchmark data set, 2.4 Gb for the COPD background model, and 263.4 Mb for the COPD eQTL test model. Thus, for general clinical genomic data use cases, a standard desktop or laptop should have sufficient memory, but larger machines may be required for unfiltered, genome wide searches.

RESULTS

COPD

SubmiRine was run on our COPD clinical genomic data set following the pipeline illustrated in Figure 1. The preprocessing steps produced a total of 93 3′UTRs containing 127 SNPs in the target model and 1608 3′UTRs containing 2106 decoy SNPs in the background model. Concurrently, we applied a threshold to include only miRNAs that

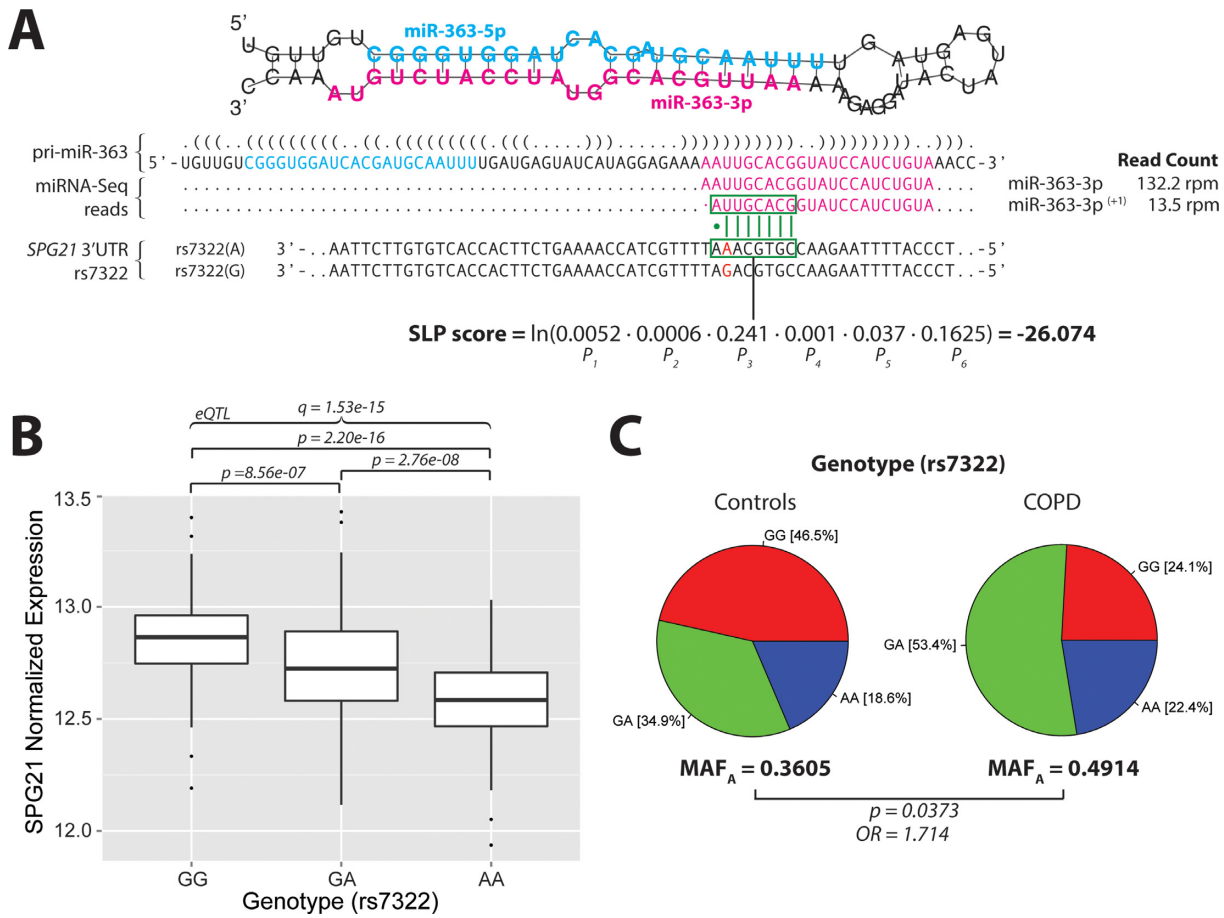

Figure 2. Highest-scoring miR-TSV prediction on the COPD data set. (**A**) The primary miR-363 locus and secondary structure are displayed, along with the quantified miRNA-Seq reads mapping to this region. Also shown is the locus on the 3′UTR of *SPG21* containing the rs7322 SNP. Highlighted in green is the predicted 8mer miRNA target site predicted on the rs7322(A) allele only for the miR-363-3p$^{(+1)}$ isomiR. Based on the SLP score computed by SubmiRine, this represents the highest-scoring miR-TSV prediction the COPD data set (see Table 2). (**B**) The rs7322 *cis*-eQTL relationship with *SPG21*. Consistent with the predicted miR-TSV in (A), the rs7322(A) allele is associated with lower *SPG21* expression. (**C**) The rs7322(A) allele is also moderately associated with the COPD phenotype, being more frequent in cases than controls.

were expressed at a level of 10 RPM or higher in at least one of the samples in the cohort, and the mean expression was recorded across the cohort as the representative expression value. This produced a total of 418 miRNAs. Using these pre-processed data, SubmiRine produced 467 putative miR-TSVs that were consistent with the expression changes reported in the eQTL models (out of 1074 miR-TSVs total), with SLP scores ranging from −0.875 to −26.074. The top 12, non-redundant predictions are presented in Table 2, with the top prediction further detailed in Figure 2.

The highest scoring predictions presented in Table 2 represent a broad range of potential miR-TSVs. For example, SNPs that are predicted to alter the sequence neighborhood of very strong target sites for highly expressed miRNAs represent miR-TSVs that have a relatively indirect effect on miRNA binding. In contrast, other high scoring predictions directly create or abolish a strong miRNA seed site, but do so for moderately expressed miRNAs. It is unclear if either of these characteristics should be more heavily weighted than the other or if they are equally important. Additionally, we highlight the fact that two of the top miR-TSV predictions are not based on the canonical mature miRNA, but rather on an isomiR.

To further demonstrate the results obtained from SubmiRine, Figure 2 summarizes the data behind the highest scoring miR-TSV prediction: the creation of a miR-363-3p isomiR binding site by the A/G SNP rs7322 on the 3′UTR of *SPG21*. Figure 2A displays the secondary structure of the primary miR-363 locus (pri-miR-363), along with the quantified mature miRNA reads mapped to this region. Note that the canonical mature miR-363-3p exhibits the largest read population of 132.2 reads per million (rpm) and that the 5′ shifted isomiR miR-363-3p$^{(+1)}$ is also moderately expressed at 13.5 rpm. Next, the two alleles of rs7322 are displayed within the *SPG21* 3′UTR, and the only predicted miRNA target site occurring between these miRNA and 3′UTR combinations is highlighted: rs7322(A) and miR-363-3p$^{(+1)}$. Thus, the canonical form of miR-363-3p is not predicted to target this locus on *SPG21*, but the miR-363-3p$^{(+1)}$ isomiR is predicted to bind the rs7322(A) allele (binary score = −0.3714) and not the rs7322(G) allele (binary score = 0). The six empirical probabilities and SLP score computed by SubmiRine for this miR-TSV are shown. P_1

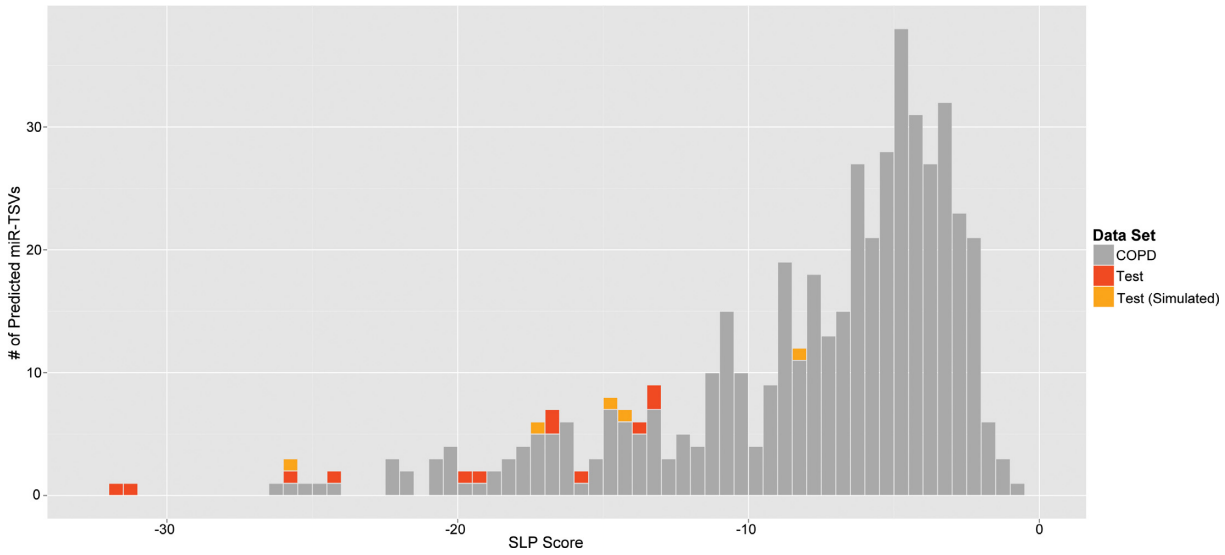

Figure 3. SubmiRine SLP score distribution of known miR-TSVs and predicted COPD-related miR-TSVs. Histogram showing the number of candidate miR-TSVs predicted at different SLP score cutoffs from the COPD data set relative to the test set of known miR-TSVs identified in previous studies. The test set (see Supplemental Table S2) is divided into subsets representing known miR-TSVs whose miRNA was identified in our lung samples ('Test'), and known miR-TSVs whose miRNA was not identified (or identified below our RPM threshold) in our lung samples, resulting in its expression being simulated at 100 RPM ['Test (Simulated)'].

Table 2. Top COPD-related miR-TSV predictions by SubmiRine

Gene	SNP	miRNA	[miRNA]	SNP Effect	SLP Score
SPG21	rs7322	miR-363-3p$^{(+1)}$	13.5 rpm	8mer site creation	−26.074
ETFDH	rs17843966	miR-100-5p	3977.1 rpm	Neighborhood alteration	−25.513
DIP2A	rs2839340	miR-125a-5p	3373.8 rpm	Neighborhood alteration	−25.171
ASPRV1	kgpl663794	miR-501-3p	31.0 rpm	8mer site creation	−22.132
MRPL20	kgp3485594	miR-4677–3p	14.4 rpm	8mer site deletion	−20.719
ASPRV1	rs3087933	miR-16–5p	5650.2 rpm	6mer/7mer-1a alteration	−20.538
GTF3C4	kgp6017974	miR-30a-3p$^{(+1)}$	247.8 rpm	8mer site creation	−20.413
DIP2A	rs2839340	let-7a-5p	49792.1 rpm	Neighborhood alteration	−20.255
ZNF419	kgp3109439	miR-136-5p	8.8 rpm	8mer site deletion	−19.700
HDAC7	kgpl0188643	miR-143-3p	146844.8 rpm	Neighborhood alteration	−19.080
SCAMP4	rs8730	miR-151a-5p	1351.6 rpm	7mer-m8 site creation	−18.997
PHLPP2	kgp7275857	miR-182-5p	1570.0 rpm	7mer-m8 site deletion	−18.004

The gene, SNP, and miRNA for each of the top 12 non-redundant miR-TSVs predicted by SubmiRine on the COPD data set are displayed. Each miRNA's mean expression is reported in reads per million (rpm), along with the predicted effect of each miR-TSV and SLP score. The full list of COPD miR-TSV predictions is provided in Supplemental Table S3. Please note that site creation and deletion are conceptually equivalent. In this table, the designation is based on the effect of the minor allele as determined in the genome-wide association analysis.

shows that, relative to all target sites in the background model, this binary score is 'stronger' than over 99.4% of the decoy target sites. However, when considering only other potential target sites for miR-363-3p$^{(+1)}$, P_2 demonstrates that this site is stronger than 99.9% of the decoy target sites. Thus, P_2 demonstrates that this particular site is predicted to be highly competitive for available miR-363-3p$^{(+1)}$. Because this isomiR is not particularly highly expressed, the empirical score (P_3) is not quite as significant as the binary score (P_1). Probabilities $P_4 - P_6$ show that the *Δbinary* and *Δempirical* scores correspond to similar levels of significance as those described above. Figure 2B and C demonstrate the relevance of rs7322 to the COPD clinical phenotype. Figure 2B demonstrates that the genotype at rs7322 is associated with differential *SPG21* expression levels ($p = 1.53 \times 10^{-15}$), with the rs7322(A) allele corresponding to lower expression, consistent with the miR-TSV prediction. We note that the expression of miR-363-3p$^{(+1)}$ does not significantly differ among the rs7322 genotype subpopulations (data not shown), rejecting the possibility that the observed rs7322-*SPG21* eQTL is an artifact of varying miR-363-3p$^{(+1)}$ levels. Finally, Figure 2C shows that the rs7322(A) allele is more frequent in the COPD cases than in the controls ($p = 0.0373$), suggesting this miR-TSV prediction could be associated with COPD susceptibility. Together, these results demonstrate that the top predictions by SubmiRine on the COPD data set appear to be consistent with a potentially functional and clinically relevant miR-TSV and may warrant experimental verification.

Validation of known SNPs affecting miRNA target sites

In order to assess the accuracy of SubmiRine, we tested its ability to predict and prioritize known miR-TSVs alongside the novel predictions made from the COPD clinical

Figure 4. Novel miR-TSV predictions on variants of previously validated miR-TSVs. Using the miRNA expression data from the COPD data set, SubmiRine predicted novel high-scoring candidate miR-TSVs on *MDM4* and *HLA-C* in addition to the experimentally validated miR-TSVs reported in the original studies (see Supplemental Table S2).

data. Currently, many miR-TSVs have been reported in the literature, but only a handful have been experimentally validated to the point of demonstrating functional, allele-specific miRNA regulation *in vivo*. Since the goal of SubmiRine is to both predict miR-TSVs and prioritize them by their probability of being functional, we utilized only known miR-TSVs that have been experimentally validated in our 'test set'. In total, we identified 26 such cases in the literature (see Supplemental Table S2). Using this test set, we evaluated SubmiRine's ability to identify known miR-TSVs as well as to efficiently prioritize them via the SLP score, highlighting miR-TSVs (both known and novel COPD predictions) that are the most likely to be functional.

Despite our relatively strict criteria for selecting cases for the test set, we anticipate that some of the validated miR-TSVs we include could be false positives. Validation of miR-TSVs is often demonstrated *in vitro* through over-expression of the candidate miRNA, yet miRNA expression levels have been shown to affect the ability of target genes to be repressed (25,40). Thus, over-expressing a miRNA beyond normal physiological levels could produce false positive interactions. Furthermore, many validated miR-TSVs were first identified *in silico* using a single candidate gene and the entire set of miRNAs annotated in miRBase (6), which may include many tissue-specific and low-confidence miRNAs not relevant to the clinical phenotype. For example, rs3134615, which lies in the 3′UTR of *MYCL1*, was predicted to have allele-specific expression due to binding of miR-1827 to the G allele only, which is associated with small-cell lung cancer (see Supplemental Table S2). However, miR-1827's expression was not validated *in vivo*, and there is very minimal experimental evidence for miR-1827 presented in miRBase, suggesting it may not even be a real miRNA. Nonetheless, transfection of a miR-1827 construct successfully repressed *MYCL1* in an allele-specific fashion.

Starting with the 26 miR-TSVs in our test set, we filtered out any miR-TSVs where the associated SNP was not included in our input 3′UTR data, as well as miR-TSVs corresponding to non-canonical target sites, which cannot be scored by the TargetScan context+ scores utilized in SubmiRine. In total, these filtering steps removed seven miR-TSVs from the test set, including four cases of non-

canonical seed sites, two SNPs that do not map to 3′UTRs, and one SNP whose record is deprecated in dbSNP and could not be mapped (see Supplemental Table S2). Of the 19 remaining test miR-TSVs, five correspond to miRNAs that were not identified in our lung tissue samples (miR-367, miR-510, miR-513a, miR-1827 and miR-3148), and one was expressed below our cutoff threshold of 10 RPM (miR-433). Thus, for each test miR-TSV, if the miRNA was identified above 10 RPM in our lung tissue samples, we utilized the detected mean expression value; otherwise, we 'simulated' the miRNA's expression by imputing it into our COPD data set at a moderate expression level of 100 RPM. Note that, with the exception of miR-1827, all of the miRNAs for which expression was simulated were shown to be expressed in the respective miR-TSV study, supporting the fact that some miRNAs are tissue-specific and simply may not have been detected in the lung. While these miRNA expression values could be specific to lung tissue, we nonetheless used these values as a proxy for the relative expression of each miRNA *in vivo* so that the test miR-TSVs could be scored alongside the COPD predictions.

To test the 19 remaining validated miR-TSVs, we added the 3′UTRs of each test miR-TSV into our COPD data set of eQTLs, using dummy expression values to reflect the allele predicted to have lower expression. We then ran SubmiRine on the COPD data merged with the validated miR-TSVs and compared the results side-by-side. Figure 3 and Supplemental Table S2 show that SubmiRine identified all 19 test miR-TSVs, but reported two as false positives (i.e. having a SLP score of zero). Both of these cases (rs12537 and rs3853839) correspond to miR-TSVs where the TargetScan context+ score of the target site is greater than zero, corresponding to a prediction of a non-functional target site. Surprisingly, both of the publications reporting these miR-TSVs used TargetScan to identify the putative target sites, but included the sites despite their positive scores. These cases either represent false negatives from TargetScan or false positives due to non-physiological experimental conditions during validation. In fact, in the case of rs12537, the comparison of expression levels between *MTMR3* alleles was not significant in all cases and did not show differential expression following transfection of miR-181 inhibitors as expected. In the case of rs3853839, the G allele (predicted to disrupt the target site) also shows significant repression as a result of miR-3148 transfection, suggesting binding may not be specific and may be the result of miRNA saturation.

The majority of the 17 test miR-TSVs with a non-zero SLP score were very highly scored by SubmiRine. Two of the 17 scored higher than any predicted miR-TSV from the COPD set, including the miR-148 binding sites on *HLA-C* and *HLA-G*. The miR-148 binding site in *HLA-C* is altered by rs67384697, which encodes a single nucleotide deletion within a haplotype containing multiple other nearby variants. Thus, recovery of this validated SNP demonstrates that SubmiRine can successfully handle indel variants and combinations of variants (see Table 1). Also among the highest scoring predictions are the miR-191 binding site on *MDM4*, the miR-125 binding site on *BMPR1B* and the miR-510 binding site (with simulated expression) on *HTR3E*. Two of the three lowest-scoring test miR-TSVs correspond to SNPs that do not alter the seed region of the predicted target site but still manage to score in the 66th and 85th percentile, respectively, of all miR-TSVs predicted in the COPD data; rs1044129 slightly alters the AU content of the neighborhood surrounding the target site, and rs193302862 increases the 3′ supplemental pairing of the miR-24-5p target site. The other fifteen test miR-TSVs are predicted to alter the target's seed region, with two altering the seed type (e.g. 6mer to a 7mer-m8 site) and the other 13 completely destroying (or creating) a seed site. These results indicate that SubmiRine is able to identify and highly prioritize experimentally validated miR-TSVs. Notably, these results demonstrate that miRNA abundance influences SubmiRine SLP scores without overwhelming them. Validated miR-TSVs that correspond to lowly expressed miRNAs are still recovered at relatively high SLP scores, and simulated miR-TSVs reflect a broad range of SLP scores despite having equivalent levels of miRNA expression (see Figure 3 and Supplemental Figure S6). Furthermore, considering these validated miR-TSVs alongside predicted miR-TSVs from the COPD data helps contextualize the strength of the novel predictions from the clinical genomic data. Additional analyses to demonstrate the predictive value of the SLP score and the underlying empirical probabilities are presented in the supplementary text (see Supplemental Figures S7 and S8).

Using SubmiRine to predict novel miRNA binding site alterations in validated SNPs

In addition to recovering 17 known miR-TSVs, SubmiRine identified a handful of additional high-scoring predictions on the 3′UTRs from test miR-TSVs for uninvestigated miRNAs (Figure 4). First, we predicted that the SNP rs4245739 found in *MDM4* with C allele-specific binding to miR-191 also has an overlapping binding site predicted for miR-887, which has almost as strong of a predicted effect. Second, in *HLA-C*, the deletion encoded by rs67384697 is reported as part of a larger haplotype (43). We investigated additional variants in this region, and predicted that the G>T SNP at position 324 of the aligned UTR (43) creates a very strong binding site for two miRNAs that overlap in seed sequence: miR-146 and miR-589. Interestingly, because SubmiRine uses miRNA-Seq reads to define candidate miRNAs, we also detected a substantial level of an isomiR of miR-146 (miR-146$^{(-1)}$, which is processed one nucleotide upstream on the 5′ end of the canonical miR-146), and predicted that 324G enhances the binding of the miR-146 isomiR to *HLA-C* by altering a 7mer-1a site to an 8-mer site. The original study on *HLA-C* investigated other SNPs in this region and showed experimentally that they did not alter miRNA functions, but this particular SNP at position 324 and the corresponding miRNAs were not reportedly tested. While these novel predictions may be influenced by our use of miRNA expression values from lung tissue, these miRNAs are all reported as relatively highly expressed in miRBase (6) and may be worth investigating further to determine the degree to which they contribute to their respective disease mechanisms.

DISCUSSION

In this work, we present the tool SubmiRine, which we have developed for analyzing miRNA target site variants (miR-TSVs) identified in clinical genomic data sets. SubmiRine is an open-source computational framework written in Python and R that allows researchers the ability to efficiently predict both miRNA target sites and miR-TSVs on a genome-wide scale. Furthermore, it provides a scoring mechanism for prioritizing miR-TSVs that are more likely to be functional. We demonstrated SubmiRine's effectiveness using both a novel COPD clinical genomic data set and a set of known miR-TSVs that have been validated elsewhere. Using the clinical genomic data, we have predicted a number of miR-TSVs that may indeed be functional. This includes novel predictions related to COPD, but also a handful of alternative miR-TSV predictions that may coordinate with known miR-TSVs from our validation set. SubmiRine's scoring scheme is based on empirical probabilities computed relative to a background model of decoy variants. We show that known miR-TSVs score highly relative to novel predictions in COPD, demonstrating that SubmiRine's SLP score has high precision.

To our knowledge, SubmiRine is the first miR-TSV prediction tool developed to analyze clinical genomic data sets in a high-throughput fashion. Existing tools are limited to pre-computed predictions reported in an online database and do not have a mechanism to prioritize predictions relative to expected functional significance. SubmiRine is designed specifically for such clinical contexts and, in addition to providing more informative output, is shown to be much faster than existing target prediction tools. Our underlying use of TargetScan6 context+ scores enables identification of miRNA target sites having canonical seed site sequences (6–8mers), which we show encompass the majority of known miR-TSVs. While a subset of known miR-TSVs from our validation set do not correspond to canonical miRNA seed sites or do not fall in 3′UTRs, we expect that canonical 3′UTR miR-TSVs will tend to have the strongest effects, especially from a genome-wide perspective. Previous reports have shown that non-canonical target sites and target sites outside of the 3′UTR tend to be less repressive (44). However, as the community's understanding of miRNA regulation by these alternative mechanisms improves, SubmiRine's scoring method can be easily adapted.

As genomic data begins to work its way into more clinical settings, the need for high-throughput tools to assess genomic variants of clinical importance is becoming imperative, particularly given how the widespread use of genome-wide association studies (GWASs) to identify variants associated with a given condition has become standard practice. However, the vast majority of GWAS hits are not associated with specific biological mechanisms, greatly limiting their potential use for development of therapeutics. It is hoped that tools such as SubmiRine—ones that can be used in the context of real genomics use cases to assess the effect of a specific kind of variation—can speed up the process of identifying promising targets worthy of experimental verification, with an eye towards downstream translational studies having tangible clinical applicability.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Praveen Sethupathy for his assistance in constructing the test set of experimentally validated miR-TSVs, Richard Myers, Adam Labadorf and Vinay Kartha for supplying the high-quality data sets that were used in the development and testing of SubmiRine; Joan Bailey-Wilson and Claire Simpson for providing motivation for the development of SubmiRine, Derek Gildea for his assistance in the interpretation of next-generation sequencing data for the design of SubmiRine, and Niraj Trivedi for his help in running software used to analyze the output of SubmiRine.

FUNDING

Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health. Funding for open access charge: Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health.
Conflict of interest statement. None declared.

REFERENCES

1. Bartel,D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, **136**, 215–233.
2. Lee,R.C., Feinbaum,R.L. and Ambros,V. (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, **75**, 843–854.
3. Maxwell,E.K., Ryan,J.F., Schnitzler,C.E., Browne,W.E. and Baxevanis,A.D. (2012) MicroRNAs and essential components of the microRNA processing machinery are not encoded in the genome of the ctenophore Mnemiopsis leidyi. *BMC Genomics*, **13**, 714.
4. Grimson,A., Srivastava,M., Fahey,B., Woodcroft,B.J., Chiang,H.R., King,N., Degnan,B.M., Rokhsar,D.S. and Bartel,D.P. (2008) Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature*, **455**, 1193–1197.
5. Friedländer,M.R., Lizano,E., Houben,A.J., Bezdan,D., Báñez-Coronel,M., Kudla,G., Mateu-Huertas,E., Kagerbauer,B., González,J., Chen,K.C. *et al.* (2014) Evidence for the biogenesis of more than 1,000 novel human microRNAs. *Genome Biol.*, **15**, R57.
6. Griffiths-Jones,S., Grocock,R.J., van Dongen,S., Bateman,A. and Enright,A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34**, D140–D144.
7. Pereira,D.M., Rodrigues,P.M., Borralho,P.M. and Rodrigues,C.M.P. (2012) Delivering the promise of miRNA cancer therapeutics. *Drug Discov. Today*, doi:10.1016/j.drudis.2012.10.002.
8. Soifer,H.S., Rossi,J.J. and Sætrom,P. (2007) MicroRNAs in disease and potential therapeutic applications. *Mol. Ther.*, **15**, 2070–2079.
9. Davidson,B.L. and McCray,P.B. (2011) Current prospects for RNA interference-based therapies. *Nat. Rev. Genet.*, **12**, 329–340.
10. Ryan,B.M., Robles,A.I. and Harris,C.C. (2010) Genetic variation in microRNA networks: the implications for cancer research. *Nat. Rev. Cancer*, **10**, 389–402.
11. Merritt,W.M., Lin,Y.G., Han,L.Y., Kamat,A.A., Spannuth,W.A., Schmandt,R., Urbauer,D., Pennacchio,L.A., Cheng,J.-F., Nick,A.M. *et al.* (2008) Dicer, Drosha, and outcomes in patients with ovarian cancer. *N. Engl. J. Med.*, **359**, 2641–2650.
12. Jazdzewski,K., Murray,E.L., Franssila,K., Jarzab,B., Schoenberg,D.R. and de La Chapelle,A. (2008) Common SNP in pre-miR-146a decreases mature miR expression and predisposes to papillary thyroid carcinoma. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 7269–7274.

13. Ward,L.D. and Kellis,M. (2012) Interpreting noncoding genetic variation in complex traits and human disease. *Nat. Biotechnol.*, **30**, 1095–1106.
14. Emilsson,V., Thorleifsson,G., Zhang,B., Leonardson,A.S., Zink,F., Zhu,J., Carlson,S., Helgason,A., Walters,G.B., Gunnarsdottir,S. *et al.* (2008) Genetics of gene expression and its effect on disease. *Nature*, **452**, 423–428.
15. Yu,Z., Li,Z., Jolicoeur,N., Zhang,L., Fortin,Y., Wang,E., Wu,M. and Shen,S.-H. (2007) Aberrant allele frequencies of the SNPs located in microRNA target sites are potentially associated with human cancers. *Nucleic Acids Res.*, **35**, 4535–4541.
16. Chen,K. and Rajewsky,N. (2006) Natural selection on human microRNA binding sites inferred from SNP data. *Nat. Genet.*, **38**, 1452–1456.
17. Lewis,B.P., Burge,C.B. and Bartel,D.P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.
18. 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.
19. Chi,S.W., Zang,J.B., Mele,A. and Darnell,R.B. (2009) Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, **460**, 479–486.
20. Helwak,A., Kudla,G., Dudnakova,T. and Tollervey,D. (2013) Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell*, **153**, 654–665.
21. Garcia,D.M., Baek,D., Shin,C., Bell,G.W., Grimson,A. and Bartel,D.P. (2011) Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. *Nat. Struct. Mol. Biol.*, **18**, 1139–1146.
22. Kertesz,M., Iovino,N., Unnerstall,U., Gaul,U. and Segal,E. (2007) The role of site accessibility in microRNA target recognition. *Nat. Genet.*, **39**, 1278–1284.
23. Krek,A., Grün,D., Poy,M.N., Wolf,R., Rosenberg,L., Epstein,E.J., MacMenamin,P., da Piedade,I., Gunsalus,K.C., Stoffel,M. *et al.* (2005) Combinatorial microRNA target predictions. *Nat. Genet.*, **37**, 495–500.
24. Betel,D., Koppal,A., Agius,P., Sander,C. and Leslie,C. (2010) Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol.*, **11**, R90.
25. Coronnello,C., Hartmaier,R., Arora,A., Huleihel,L., Pandit,K.V., Bais,A.S., Butterworth,M., Kaminski,N., Stormo,G.D. and Oesterreich,S. (2012) Novel modeling of combinatorial miRNA targeting identifies SNP with potential role in bone density. *PLoS Comput. Biol.*, **8**, e1002830.
26. Vejnar,C.E. and Zdobnov,E.M. (2012) MiRmap: comprehensive prediction of microRNA target repression strength. *Nucleic Acids Res.*, **40**, 11673–11683.
27. Fernald,G.H., Capriotti,E., Daneshjou,R., Karczewski,K.J. and Altman,R.B. (2011) Bioinformatics challenges for personalized medicine. *Bioinformatics*, **27**, 1741–1748.
28. Hiard,S., Charlier,C., Coppieters,W., Georges,M. and Baurain,D. (2010) Patrocles: a database of polymorphic miRNA-mediated gene regulation in vertebrates. *Nucleic Acids Res.*, **38**, D640–D651.
29. Deveci,M., Catalyürek,U.V. and Toland,A.E. (2014) mrSNP: software to detect SNP effects on microRNA binding. *BMC Bioinformatics*, **15**, 73.
30. Ziebarth,J.D., Bhattacharya,A., Chen,A. and Cui,Y. (2012) PolymiRTS Database 2.0: linking polymorphisms in microRNA target sites with human diseases and complex traits. *Nucleic Acids Res.*, **40**, D216–D221.
31. 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.
32. Liu,C., Zhang,F., Li,T., Lu,M., Wang,L., Yue,W. and Zhang,D. (2012) MirSNP, a database of polymorphisms altering miRNA target sites, identifies miRNA-related SNPs in GWAS SNPs and eQTLs. *BMC Genomics*, **13**, 661.
33. Thomas,L.F., Saito,T. and Sætrom,P. (2011) Inferring causative variants in microRNA target sites. *Nucleic Acids Res.*, **39**, e109.
34. Bulik-Sullivan,B., Selitsky,S. and Sethupathy,P. (2013) Prioritization of genetic variants in the microRNA regulome as functional candidates in genome-wide association studies. *Hum. Mutat.*, **34**, 1049–1056.
35. Johnson,W.E., Li,C. and Rabinovic,A. (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, **8**, 118–127.
36. Cloonan,N., Wani,S., Xu,Q., Gu,J., Lea,K., Heater,S., Barbacioru,C., Steptoe,A.L., Martin,H.C., Nourbakhsh,E. *et al.* (2011) MicroRNAs and their isomiRs function cooperatively to target common biological pathways. *Genome Biol.*, **12**, R126.
37. Morin,R.D., O'Connor,M.D., Griffith,M., Kuchenbauer,F., Delaney,A., Prabhu,A.-L., Zhao,Y., McDonald,H., Zeng,T., Hirst,M. *et al.* (2008) Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res.*, **18**, 610–621.
38. Shabalin,A.A. (2012) Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, **28**, 1353–1358.
39. Purcell,S., Neale,B., Todd-Brown,K., Thomas,L., Ferreira,M.A.R., Bender,D., Maller,J., Sklar,P., de Bakker,P.I.W., Daly,M.J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.
40. Mullokandov,G., Baccarini,A., Ruzo,A., Jayaprakash,A.D., Tung,N., Israelow,B., Evans,M.J., Sachidanandam,R. and Brown,B.D. (2012) High-throughput assessment of microRNA activity and function using microRNA sensor and decoy libraries. *Nat. Methods*, **9**, 840–846.
41. Sumazin,P., Yang,X., Chiu,H.-S., Chung,W.-J., Iyer,A., Llobet-Navas,D., Rajbhandari,P., Bansal,M., Guarnieri,P., Silva,J. *et al.* (2011) An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. *Cell*, **147**, 370–381.
42. Bruno,A.E., Li,L., Kalabus,J.L., Pan,Y., Yu,A. and Hu,Z. (2012) miRdSNP: a database of disease-associated SNPs and microRNA target sites on 3′UTRs of human genes. *BMC Genomics*, **13**, 44.
43. Kulkarni,S., Savan,R., Qi,Y., Gao,X., Yuki,Y., Bass,S.E., Martin,M.P., Hunt,P., Deeks,S.G., Telenti,A. *et al.* (2011) Differential microRNA regulation of HLA-C expression and its association with HIV control. *Nature*, **472**, 495–498.
44. Hafner,M., Landthaler,M., Burger,L., Khorshid,M., Hausser,J., Berninger,P., Rothballer,A., Ascano,M., 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.