©2014

LIYANG DIAO

ALL RIGHTS RESERVED

APPLICATIONS OF THE MIXED LINEAR MODEL IN GENOME-WIDE ASSOCIATION STUDIES AND SMALL RNA MOTIF DISCOVERY

by

LIYANG DIAO

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in

Computational Biology and Molecular Biophysics

Written under the direction of

Dr. Kevin C. Chen

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

OCTOBER, 2014

ABSTRACT OF THE DISSERTATION

Applications of the mixed linear model in genome-wide association studies and small RNA motif discovery

By LIYANG DIAO

Dissertation Director:

Kevin C. Chen

If sheer number of papers published is indicative of anything, it suggests that the age of genome-wide association studies, or GWAS, is here to stay. However, in spite of the influx of data, several issues remain, one of which is the presence of confounding factors caused by relatedness within the study sample. This can cause many false positive results. In recent years, the use of mixed linear models to correct for unknown types of relatedness, i.e. "cryptic relatedness", has been very popular. While this model has been shown to be successful in some cases, here we address the feasibility of performing GWAS in a highly structured population such as *Saccharomyces cerevisiae*, and find that the inclusion of fixed local ancestry covariates can sometimes lend a study more power.

Furthermore, we explore the application of mixed linear models in a different type of biological problem of discovering motifs associated with active microRNAs. While there exist several algorithms for miRNA motif discovery, only a few consider background sequence composition of the 3' UTR binding site in addition to seed sequence motif enrichment, which is known to factor into miNRA binding efficacy. The methods that do

account for 3' UTR sequence composition do so by rescoring motif counts based on the background UTR sequence in which it appears. Though computationally efficient, these methods are unable to simultaneously compare both gene expression values and UTR sequence, which our method, named MixMir, is able to do, with favorable results. When compared to the simple linear model, as well as existing motif discovery algorithms, MixMir is able to rank true motifs more highly in multiple data sets. Such computational methods are biologically significant because although it is possible to sequence small RNAs in a sample, their expression may not be perfectly correlated with the size of their effect, which is what we observed.

ACKNOWLEDGMENTS AND DEDICATION

First and foremost, I would like to thank my advisor Dr. Kevin C. Chen for all of the mentorship and advice he provided me with for the last three years. He was responsive to my questions, practical in his answers, and always asked me what my opinion was, pushing me to ask and answer questions of my own. I thank him for his insight and moral support, and for seeking out opportunities to make me a well-rounded as a scientist. Without his guidance I would not be where I am now, nor have the opportunities I have before me.

I would also like to thank the past and present members of the Chen and Xing labs, who were ever friendly and ready with helpful comments and criticisms. I am grateful for your time and attention, even through long and rambling presentations.

I would like to thank Michael Seiler, for supporting me and being a positive influence especially during difficult times.

Finally, I would like to thank my parents, to whom I also dedicate this thesis, for inspiring me on my journey as I followed in their footsteps as best I could.

Chapter 2 of the thesis is based on the published work, Local ancestry corrects for population structure in *Saccharomyces cerevisiae* genome-wide association studies (1). Chapter 3 of the thesis is based on the published work, MixMir: microRNA motif discovery from gene expression data using mixed linear models (2).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1: GENOME-WIDE ASSOCIATION STUDIES

1.1.1: THE ERA OF BIG DATA: A BRIEF HISTORY OF GWAS

Genome-wide association studies (GWAS) have become a primary tool in mapping genetic traits, particularly with the growing quantity of whole-genome data quickly becoming available. Thanks to high-throughput sequencing methods that have made the collection of data faster and more inexpensive than ever before, the exploration of complex genetic traits and architectures is possible. We now find ourselves in a situation where the development of statistical methods, and the computational aspects of applying such methods to huge amounts of data, has become a crux in the analysis pipeline.

The idea of GWAS began with the idea of constructing linkage maps to uncover disease-associated variants (3,4). To understand linkage maps, one must first understand the concept of recombination. Recombination between loci occurs when homologous chromosomes cross over during meiosis, resulting in the exchange of genetic information and increased diversity. A graphic of chromosomal crossover and a brief discussion of linkage blocks and how they can be used in association studies can be found in Figure 1. When recombination frequency is low, then fewer crossovers take place and linkage blocks tend to be large. In this case, linkage mapping can be a viable method of discovering causal genes located nearby mapped loci. However, one of the tradeoffs made with large linkage blocks is a loss in resolution (3), so that if for example the causal variant is far from the mapped locus, and the linkage block is too large, it may be difficult to pinpoint the exact cause of the disease. Thus, successful linkage mapping relies on disease loci being inherited alongside nearby loci in linkage blocks.

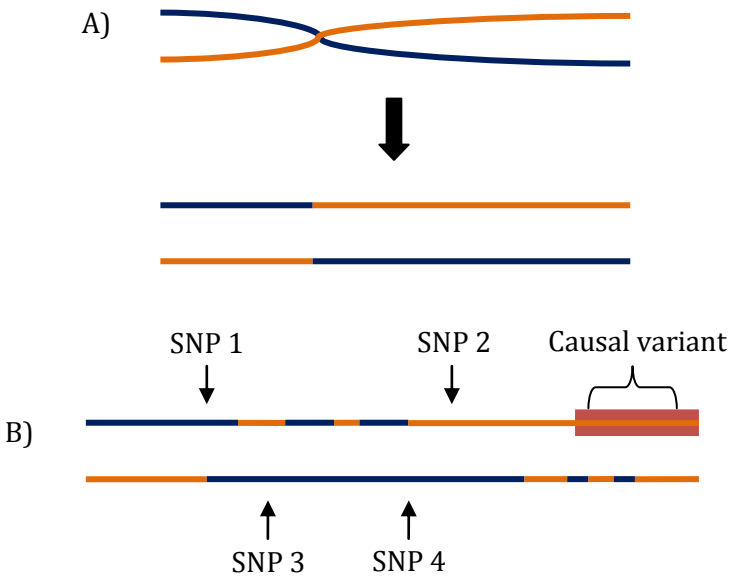


Figure 1

a) Example of chromosomal crossover between the arms of two homologous chromosomes during meiosis. After crossover occurs, part of the genetic information on each arm has been swapped with genetic information from the other. Loci that are inherited together are said to be located on the same linkage block (linkage disequilibrium, or LD, block), and have a low recombination frequency between them. **b)** After several recombinations, we may see some loci are frequently inherited together in what are known as linkage blocks. In this example, SNP 1 and SNP 2 are separated by several recombination events, while SNP 3 and SNP 4 are not, and are located on the same linkage block. If SNP 2 is a causal allele, we may not be able to uncover its association if we have only included SNP 1 in our analysis. However, if SNP 4 is a causal allele, we may be able to do so if SNP 3 is included in the analysis, because they reside on the same linkage block.

In general, linkage analyses rely on disease variants having a large effect size, and indeed they have trouble identifying disease variants and causal loci, particularly in complex human traits (3). One of the differences between linkage analyses and GWAS is that linkage analyses rely on LD blocks arising within a pedigree, while GWAS rely on LD blocks arising at the population level (3). The principles depicted in Figure 1 remain the same. Risch and Merikangas (5) first pointed out that while linkage analyses may suffer

from power to detect common variants of smaller effect, association studies using many more data points, such as single nucleotide polymorphisms (SNPs), may point towards a solution. Undertakings such as the human genome project (6-9) attempted to collect such large-scale data, with the hope that its completion would give scientists the data necessary to improve our understanding of human disease, migration, evolution, and gene therapy, among other things.

While the initial idea was to apply such studies to simple Mendelian traits, it quickly became apparent that the vast majority of phenotypes and genetic architectures are complex, and furthermore that the effect size of any one particular mutation may be small (4). In the presence of other confounding factors such as population structure, researchers soon found that the limitations of GWAS may not be able to be remedied simply by an increase in the amount of data, but also required statistical innovations. Additionally, scientists began to question whether or not the basic principles of GWAS were sound. In a report by Visscher et al., the authors surveyed researchers in the field and specifically address four points of contention: That

1. GWASs are founded on a flawed assumption that genetics plays an important role in the risk to common diseases;
2. GWASs have been disappointing in not explaining more genetic variation in the population;
3. GWASs have not delivered meaningful, biologically relevant knowledge or results of clinical or any other utility; and
4. GWAS results are spurious (3)

Nonetheless, in the past several years some large successes have been reported by the GWAS community.

Among those are the studies in 2005 and 2006 that implicated variants of the CFH gene and HTRA1 promoter with increased risk of developing age-related macular degeneration (AMD). In these studies, Klein et al. and DeWan et al. discovered single nucleotide polymorphisms (SNPs) associated with nonneovascular (dry) and neovascular (wet) AMD, respectively, located in the CFH gene and the promoter of HTRA1 (10,11). In dry AMD, the variant in CFH confers a greater than 7 fold risk for the disease in individuals who are homozygous (10). In wet AMD, the variant confers a 10 fold risk (11).

Other often-cited examples of the successes of GWAS are the studies in 2007 and 2008 that uncovered the association between variants of the BCL11A gene with differing levels of fetal hemoglobin (HbF) expression (12-14). Different levels of HbF are associated with varying degrees of morbidity and mortality in sickle cell disease and β-thalassemia (13,15), which although they are monogenic disorders, show a surprising range of phenotypes, some of which are explained by levels of HbF. These computational results were later experimentally validated by Sankaran et al., who showed that down-regulation of BCL11A expression results in increased HbF expression in primary adult erythroid cells, and who also proposed that directed down-regulation like this could be used to treat patients with sickle cell disease and β-thalassemia (15).

In addition to these well-known examples, many other GWAS studies have identified susceptibility loci in diseases such as Crohn's disease (16), type 2 diabetes (17), and obesity (18). A quick search for the keyword "genome-wide association study" results in nearly 300,000 research papers published since 2005, with nearly a third of those having been published in the last four years, indicating that GWAS continue to be widely practiced and in greater numbers. In spite of this, however, many doubts remain as to whether GWAS have

been biologically useful in most cases, especially when working with complex phenotypes. If the goal is to discovery targets for gene therapy, for example, then first we must first consider that many phenotypes may not be purely genetic, that the environmental component may be large, and that the effect of markers—even truly causal ones—may be limited (19). For the purposes of this thesis, however, we are less concerned with the potential therapeutic benefits of successful GWAS than we are with the possibility of first discovering true and causal markers. In chapter two we address one particular difficulty with GWAS, namely the confounding effects of population structure in the sample, which we discuss below. We focus on the possibility of performing GWAS on *S. cerevisiae*, as a model organism that is highly structured.

1.1.2: POPULATION STRUCTURE AND CRYPTIC RELATEDNESS

Though access to large datasets now lends new studies greater power, several persistent complications remain. One of these is the presence of population structure in the samples being studied, which can lead to both false positives and a reduction in power. This can arise due to shared ancestry among the individuals in the sample, or it can appear more subtly as "cryptic relatedness", relatedness between individuals for reasons unknown to the researcher. In a classic example of the former, if testing for single nucleotide polymorphisms (SNPs) associated with hypertension in a population of African Americans and Caucasians, any SNPs which are found more commonly in African Americans would be more likely to be implicated, as hypertension is known to occur more frequently in that population. In the latter case, large datasets such as those used in case-control studies of complex diseases may find that the affected individuals being analyzed may share genetic markers, either through unknown familial relatedness or simply a shared genetic background which contributes to the propensity for the disease (20). While one can in

theory attempt to remove some of this confounding by carefully carrying out an experiment to avoid such factors, in practice this can be both impractical and costly.

One example of this is a study of human height using more than five thousand individuals from the Framingham Heart Study (FHS), which reported a prediction R^2 of 0.25 using ten-fold cross validation when the prediction was performed on all individuals used in the analysis, reviewed in (19). However, because there are many known familial relationships among individuals in the FHS, the authors performed the cross validation again while restricting the sample to individuals with no known relationship, which reduced the R^2 to 0.15. The authors of (19) performed their own analysis of height in more than seven thousand individuals from the FHS and found a similar R^2 with close relatives removed. However, they whittled down the sample further by restricting to those individuals with pairwise relatedness estimated using SNPs less than a given threshold. They found that as this threshold decreases, so does the R^2 obtained. Thus, the authors concluded, using a genetic definition of relatedness and statistical methods that correct for cryptic relatedness are important, as reported family relationships may not be sufficient to reduce confounding.

One of the early and still popular methods for correcting for population structure possible relatedness between individuals is genomic control (GC) (20). One of the reasons for the development of GC was to address the fact that in case-control studies, individuals with a particular disease are naturally more likely to be related to one another, and thus associations will be found to be stronger than they actually should be (20). The method compares the both the Armitage trend test and the allelic chi-squared test and shows that the test statistics are inflated when compared to the case where the members in a population under study are related. The authors define this inflation factor as the *variance inflation factor*, or VIF. In brief, the inflation factor λ describes the additional variance

present in the expression dataset that can be attributed to the presence of collinearity between variables (such as the presence of highly linked genetic markers). Pritchard and Donnelly (21) reformulate this inflation factor as

$$\lambda = 1 + \frac{RF\sum_k[(f_k - g_k)^2] - 2F}{1 + F},$$

where R is the number of cases and controls, K is the number of subpopulations within the population sampled, f_k and g_k are the fractions of cases and controls drawn from subpopulation k, respectively, and F is Wright's coefficient of inbreeding (21). As we can see here, if the fractions of cases and controls drawn from each subpopulation k is equal, and there is no inbreeding in the population, then the inflation factor is 1, i.e., there is no inflation of the test statistic.

Developed for case-control studies and later extended to quantitative traits (22), GC was shown to be effective at controlling the number of false positives induced by stratification. One of the important assumptions when applying GC is that the population structure affects each locus identically—that is, GC produces a uniform correction over all loci. Among the conditions under which this is approximately true are that the loci cannot have significantly different mutation rates or F (20). Furthermore, it was also shown that in some cases, particularly when the effect of structure is large and a lower threshold for significance is desired, GC can suffer from a lack of power (23-25).

Other methods applied to this problem include principal component analysis (PCA) and structured association. PCA is a commonly used method of dimension reduction that can represent a dataset of many dimensions as a combination of a smaller number of linearly independent variables, where this number can be selected by the user. Typically PCA chooses the top two independent variables (the "principal components"), so that each original data point can be represented in a two-dimensional graph for easy visualization.

While the idea of PCA extends back more than a hundred years or more, it has only recently been applied to the issue of population stratification (23). Here the authors take as their high-dimensional data matrix the matrix of biallelic markers for each individual in their analysis, with individuals in rows. Denote this matrix *M*. The authors then perform PCA on the matrix

$$X = \frac{1}{n} MM',$$

where *n* is the number of markers in the study.

While the first few principal components may not represent any particular subpopulations or ancestors exactly, the authors mention that these terms may be socially defined and as such may not best categorize the individuals (also cited in (21)). This is in contrast to structured association (SA) methods that try to delineate the details of the population structure and correct for it accordingly. The most well-known of these methods is STRUCTURE (26,27), which uses Markov chain monte carlo (MCMC) to estimate the proportion of ancestry of each individual from k ancestral populations, where k is defined by the user and may not be known. In short, STRUCTURE defines each ancestral population by a set of allele frequencies at each marker, jointly estimating these allele frequencies as well as each individual's ancestry. Furthermore, STRUCTURE allows for the presence of admixture. Later versions of STRUCTURE also allowed for the presence of different types linkage disequilibrium between markers and also provided marker-specific estimates of ancestry (27). A more detailed description of STRUCTURE can be found in *2.2.3: Brief description of STRUCTURE and parameters*. After obtaining a genome-wide estimation of ancestry, then, these factors may be incorporated in a GWAS in order to control for false positives due to shared ancestry.

It has now become popular to utilize mixed linear models (MLMs) to address the issue of population structure, and specifically cryptic relatedness. Unlike methods such as

STRUCTURE which explicitly estimate the ancestry of each individual at both the genome level and at each individual marker, and also unlike genomic control methods which provide a single variable for correction across all markers, MLMs consider the similarity in ancestry pairwise between all individuals in the sample using a kinship matrix. The kinship matrix can be constructed in any number of ways, provided it is positive semidefinite. Most typically this matrix is defined as the identity-by-state (IBS) matrix between all pairs of individuals (28). Then, some part of the phenotypic similarity between individuals can be explained by their genotypic similarity.

One of the factors that makes MLMs appealing is that it is not necessary to explicitly define individual ancestry as STRUCTURE does, and as such may be able to better capture cryptic relatedness, which might not be clear when we are forced to assume k ancestral populations. At the same time, MLMS are also more flexible than GC, allowing for different levels of similarity between individuals in the sample. However, the issue of whether MLMs are effective at analyzing highly structured populations such as *S. cerevisiae* is less clear, and this is the topic we focus on in *Chapter 2: Genome-wide association studies in highly structured populations*. A detailed description of the mixed linear model is given in *1.3: Mixed linear models*.

As mentioned above, the variance inflation factor, or VIF, is one way to determine whether or not population structure exists in a sample. Another commonly used visualization for this is by means of a quantile-quantile (QQ; sometimes percentile-percentile, PP) plot. These plots are constructed by plotting the significance values returned by the model along the *y* axis against their expected values along the *x* axis, which should be uniform. Thus, deviation between model and data results in a deviation between the observed versus expected line and the *x*-*y* axis. In cases where stratification causes an inflation of the significance of the markers under study, the QQ plot will present as a

concave curve. The presence of true associated markers will also cause a shift in the curve, but as the number of true markers is typically much smaller than the set of all markers tested, these can be identified as a slight overabundance of markers with low *p*-values, while the remainder of the curve is unaffected.

1.2: MICRORNA MOTIF DISCOVERY

1.2.1: HISTORY, BIOGENESIS, AND MECHANISM

In the years since their discovery, small, non-coding RNAs (ncRNAs) have been found to be pervasive actors in gene regulation (29-32), adding yet another layer of complexity to our understanding of how genes are expressed. Among the most common of these that have been discovered are small interfering RNAs (siRNAs), piwi-interacting RNAs (piRNAs), and microRNAs (miRNAs). These three classes of small RNAs have different modes of biogenesis, different mechanisms, and play different roles in the regulation of gene expression as well as the expression of transposable elements, for example.

siRNAs were the first type of small RNA discovered, in a seminal paper by Fire and Mello in 1998 (33). While the process which later came to be known as RNA interference, or RNAi, was first observed in the 1980s in bacteria, *Drosophila*, mammalian cells, and plants (34), the first major step towards understanding the mechanism came when Fire and Mello found that the introduction of double-stranded RNA (dsRNA) into *C. elegans* was effective in repression of gene expression, and neither the sense nor antisense strand separately. Meanwhile in 1993 the first observation of a miRNA was reported (35), and in the years that have followed, a flurry of studies of small (as well as not-so-small) ncRNAs has uncovered new variations in a host of different organisms.

The functions of these various ncRNAs are as of yet only partially understood. While siRNAs were originally thought to be generated strictly from exogenous agents such as viruses, potentially as a protective measure, several forms of endogenous siRNAs, sometimes called endo-siRNAs (29), have also been discovered, with several classes being present in plants. These endogenous siRNAs are thought to protect the organism against retrotransposons and also potentially to promote heterochromatin formation (29).

piRNAs are one of the most recently discovered class of small ncRNAs, and are thought to primarily protect germ line cells against transposons (29,36). As a testament to the plasticity of the field, a review of small RNAs published in 2008 (37) described both repeat-associated RNAs (rasiRNAs) and 21U-RNAs, which in a review published just a year later were both reclassified as piRNAs (29). A potential new class of tRNA-derived small RNAs have also been reported recently (38-42); however, reports are not consistent on, for example, the RNA fragment length, or even if we should expect a narrow window of fragment lengths as we would with siRNAs, miRNAs, and piRNAs. There is also no known function for these tRNA-derived RNAs; however, at least one such RNA is reportedly essential for cell proliferation (42).

As sequencing technologies become cheaper and more efficient, our ability to characterize in detail the small RNA landscape across and within organisms will undoubtedly uncover exciting new regulatory functions and complexities. However, the scope of this thesis will focus on miRNAs, one of the very well studied classes of small ncRNAs, which we will discuss further here.

miRNAs are approximately 22 nucleotides long, and can be differentiated from other small RNAs such as siRNAs by the fact that they arise from their primary transcripts, small hairpin structures called pri-miRNAs (see Figure 2). Though not well characterized, the pri-miRNA is transcribed from miRNA genes primarily by RNA polymerase II (43), though

transcription has also been reported by Pol III (44). These primary transcripts are then processed in the nucleus into so-called precursor miRNAs, or pre-miRNAs, by the Drosha RNase III endonuclease and cofactor DGCR8, which cleaves the bottom of the hairpin structure (30,45). The pre-miRNAs are exported into the cytoplasm, where the final processing step is performed by the RNase III enzyme Dicer (30), which cleaves away the hairpin portion of the pre-miRNA, leaving a duplex consisting of the mature miRNA strand and its complement, often denoted miRNA* and called the "star strand" (46), which is typically degraded in the cytoplasm, though functional miRNA* activity has been reported (47). Similar to siRNAs, the mature miRNA strand is loaded into the Argonaute protein of an RNA-induced silencing complex, or RISC (48) and then guided to its target, where targeting is often defined by some level of sequence complementarity of the miRNA to the mRNA, though this mechanism is considerably more complex in animals than in plants (see *1.2.2: miRNA target prediction and motif discovery*). In the case of perfect complementarity, Argonaute-catalyzed mRNA cleavage typically occurs; otherwise some form of translational repression occurs instead (48).

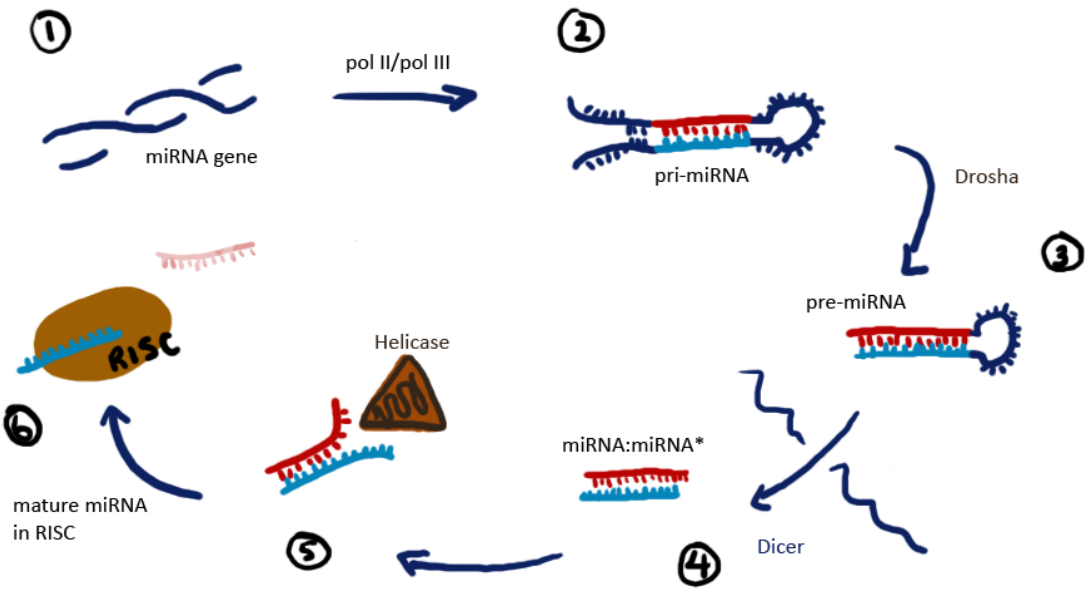


Figure 2

miRNA biogenesis begins with the miRNA gene being transcribed by RNA polymerase II or RNA polymerase III, into a hairpin structure called the pri-miRNA. The pri-miRNA is processed by Drosha into what is called the pre-miRNA, which is exported into the cytoplasm, where the hairpin is removed by Dicer and the miRNA:miRNA* duplex is unzipped by a helicase to separate the mature miRNA and the star strand, which usually is quickly degraded. The mature strand is then loaded into the RNA silencing complex and is guided to its target.

It is estimated that more than 60% of all human protein coding genes are regulated by miRNAs (32,49). They have been implicated in essential processes such as cell development, hematopoiesis, immune function, and differentiation (50-52), as well as in many diseases, such as autoimmune diseases (53) and several types of cancer (54-57). miRNA genes as well as their targets have been found to be widely conserved among mammals (31,32,49), which further supports the importance of the role they play in fine tuning gene regulation. Their expression is not only highly tissue- and cell- specific (30), but their expression also varies temporally (46), so characterizing which miRNAs are present and playing an active role is a complex endeavor, and furthermore high expression

of a particular miRNA species does not necessarily indicate a large effect size, since competition for mRNAs or even argonaute can affect miRNA efficacy (58).

1.2.2: miRNA TARGET PREDICTION AND MOTIF DISCOVERY

Given the biological significance of miRNAs, being able to predict a particular miRNA's target genes, or being able to determine which miRNAs are active in a particular cell type, are of great interest.

As mentioned in the previous section, the biological mechanism by which miRNAs act is partially understood, insofar as the mature miRNA strand is directed to its target by the RISC. However, the sequence specific determinants of binding are still somewhat unclear. Predicting the targets of plant miRNAs is more straightforward, as sequences that demonstrate extensive complementarity have been shown to largely be true targets (48). However, this is not always the case for metazoa, and many computational algorithms have been developed to address the nuances of this issue. Together they suggest three things: First, that perfect complementarity to the seed region of the miRNA (nts 2-7 at the 5' end) is required; second, that conserved pairing to this region is enough for predicting targets better than random; and last, that highly conserved miRNAs have many conserved targets (48). In general, pairing at the 6 nt seed region alone is not strong enough to discover true targets, but the three so-called canonical sites are. These canonical sites consist of a seed site match with an additional A paired to the first nucleotide of the miRNA (from the 5' end), called a 7mer-A1, or simply A1, site; a seed site match with an additional Watson-Crick match across nucleotide 8 of the miRNA, called a 7mer-m8 site; and a seed site match with both A1 and nt 8 matches, called an 8mer site (48).

Occasionally there is also compensatory pairing at the 3' end of the miRNA, for example if there is imperfect pairing in the seed region, but such examples are rare.

Perhaps the most well known instance of imperfect pairing in the seed region is a "G-bulge" site, which occurs when a G nucleotide appears in the mRNA between nucleotides 5 and 6 of the mature miRNA. This type of non-canonical binding was shown to be common in miR-124 binding sites in mouse brain, occurring in at least 15% of all Ago-miRNA interactions (59).

While the importance of stringent Watson-Crick binding to the seed sequence is certain, there are other factors which can affect miRNA binding efficacy. For example, binding sites that are flanked by regions of high AU content and that are located away from the centers of long UTRs and stop codons are more likely to be bound by miRNAs (48). Additionally, when a miRNA binds near another, their combined effect can be greater than the sum of their individual effects (48).

Other factors that have been explored in miRNA:mRNA target pair prediction are the expression levels of the argonaute proteins, the miRNAs themselves, the expression levels of potentially competing miRNAs, and the expression levels of mRNAs competing for miRNA binding. Stanhope et al. (58) incorporated all of these into a multifactor linear model aimed at delineating true pairs of miRNAs and their target mRNAs across various tissue samples. In the following, *i* denotes the index of the tissue sample, and *j* indicates the index of the miRNA:mRNA target pair.

$$
\begin{aligned}
mRNA_i^j = {} & \beta_0^j + \beta_1^j Ago2_i miRNA_i^j + \beta_2^j Ago2_i + \beta_3^j miRNA_i^j \\
& + \beta_4^j Ago134_i + \beta_5^j Ago134_i miRNA_i^j + \beta_6^j miRNA_i^{-j} \\
& + \beta_7^j Ago2_i miRNA_i^{-j} + \beta_8^j Ago134_i miRNA_i^{-j} \\
& + \beta_9^j mRNA_i^{-j} + \varepsilon_i^j
\end{aligned}
$$

Here $mRNA_i^j$ and $miRNA_i^j$ are the expression levels of the mRNA and miRNA, respectively, in the *j*th target pair tested, in the *i*th tissue sample; Ago2 and Ago134 represent the expression levels of Argonaute 2 and Argonautes 1, 3, and 4; and $mRNA_i^{-j}$ and $miRNA_i^{-j}$ are the expression levels of the all mRNAs and miRNAs, respectively, not in target pair *j*. The authors found that after performing model selection based on the AIC, model fit increased and the so-called AIC-optimized submodel was able to identify many more target pairs than the "marginal model", i.e. a simple correlation of miRNA and mRNA expression levels without systems biology components (58).

In this manuscript we focus on the problem of miRNA motif discovery, which attempts to predict active miRNAs within a particular cell type and/or tissue instead of attempting to discovery active miRNA and mRNA target pairs. In modeling, then, we assume that miRNA expression levels are unknown, or that those most highly expressed miRNAs are not necessarily the ones with the largest effects. Existing methods instead use thousands of mRNA or protein expression measurements coupled with sequence data to predict potential seed motifs or canonical binding motifs that may correspond to the active miRNAs. A more in-depth discussion of current methods in miRNA motif discovery can be found in *3.3.1: Current methods in miRNA motif discovery*.

1.3: MIXED LINEAR MODELS

1.3.1: HISTORY AND SIGNIFICANCE OF THE MIXED LINEAR MODEL

Mixed linear models have been applied extensively in genome-wide association studies (GWAS) for the purpose of correcting for population stratification. The seed of the idea was planted by R. A. Fisher in his major 1918 work, *The correlation between relatives*

on the supposition of Mendelian inheritance, where the author was concerned with the amount of phenotypic variation attributable to genetic relatedness between individuals, and how to separate this from the variation attributable to other, such as environmental, effects (60). In today's terminology, this is equivalent to the separation of variance components, and determining the amount of heritability of a trait. At the time, however, even the idea of "variance" was not established as it is today, and in fact Fisher was the first to coin this term in the selfsame paper (60). Perhaps surprisingly to us now, there was some hesitation on the part of the reviewers (Karl Pearson of the Pearson correlation and Reginald Punnett of the Punnett square) to accept the paper (61)—not with the mathematics of the paper, but primarily with Fisher's assumption that there may be many genes underlying a phenotype. As Pearson concluded in his review, "Whether the paper be published or not should depend ... [on] the probability that Mendelians will accept ... a multiplicity of independent units not exhibiting dominance or coupling" (61). This assumption of complex phenotypes is an idea that we now take for granted.

Since then the theory of mixed linear models (also linear mixed models) has been well developed and widely applied. Its usefulness is readily appreciated: In the linear model, one of the assumptions is that the model's residuals are independent. However, in many cases this assumption is violated. For example, when the individuals being studied are related, their response variables are expected to be correlated as well. Mixed linear models are popular in analyses of longitudinal data as well, since repeated measurements from a single individual are expected to be related. In these cases we would like to distinguish between what is called a *fixed* effect and a *random* effect.

The easiest way to understand the difference between the two is via example: Suppose we were a pharmaceuticals company interested in studying the effect of a drug. We distribute the drug to several hospitals, who each administer it to several of their

patients. While we are interested only in the effect of the drug, we will likely see positive correlations among the responses of patients attending a single hospital, in what is known as a "batch effect". In this toy example, the effect of the administering hospital on the response is the batch effect, which is not of interest but which we must tease out of the drug effect. Depending on the situation, the hospital effect may be considered a fixed or a random effect: If the scope of our study extends only to a particular subset of hospitals, and we are interested in the effect of each of those hospitals specifically, then what we have is a fixed effect. If, on the other hand, we randomly selected a handful of hospitals from across the nation, with no interest in the results of any particular hospital, then we have a random effect. Another way to define the difference between a fixed and a random effect is that a fixed effect in a model encompasses all possible levels of that factor—for example, the drug factor in this study would be a fixed effect, since we are studying all levels of the drug (either it is administered or it is not). However, the hospital effect as we have described it does not encompass all possible levels, since we are randomly selecting a few hospitals out of thousands of potential hospitals across the nation. This is where the terminology *random* effect comes from: In effect, the effect attributable to a particular hospital is *randomly* selected from a probability distribution.

1.3.2: OVERVIEW OF THE MIXED LINEAR MODEL

Here we give a brief description of the mixed linear model, taken primarily from papers developed by members of the GWAS community, Searle's book *Variance components*, and also Jiang's *Linear and Generalized Linear Mixed Models and Their Applications* (25,28,62,63), which the reader can reference for further detail. We first discuss briefly the linear model with respect to GWAS, and then expand the model to include random effects, which produces the mixed linear model.

In the case of a simple linear model in which we are testing for the effect of a single SNP, with no other covariates, the model is formulated as

$$y_i = \mu + x_{ij}\beta_j + \varepsilon_i,$$

where y_i represents the phenotype of individual *i*, x_{ij} is a binary variable representing either the presence or absence of SNP *j* in individual *i*, β_j is the phenotypic effect of SNP *j* on individual *i*, and ε_i is an error term associated with individual *i*. The variance depends solely on ε_i, such that

$$E(y_i) = \mu$$

$$E(\varepsilon_i) = 0$$

$$Var(y_i) = Var(\varepsilon_i) = \sigma_e^2$$

$$Cov(\varepsilon_i, \varepsilon_j) = 0 \text{ for } i \neq j$$

In this case, x_{ij} is called a *fixed* effect. Fixed effects are constants, and represent all levels of a factor of interest. Here, x_{ij} is either 0 or 1, representing all possible levels of the SNP *factor* (either present or absent).

For a mixed linear model, the equation is written similarly, with the addition of a *random* effect, α_i:

$$y_i = \mu + x_{ij}\beta_j + \alpha_i + \varepsilon_i,$$

While α_i is presented no differently than a fixed effect, the way in which it is interpreted is different. Namely, the random effect is not an effect that we have any particular interest in. As in the example we presented in the previous example with a drug trial in randomly selected hospitals, the hospital effect is not of interest to us—it is the performance of the drug. This is not reflected in the equations above for readability, but is implicit.

The mixed linear model is subject to the following additional constraints:

$$E(\alpha_i) = 0$$

$$Var(\alpha_i) = \sigma_\alpha^2$$

$$Cov(\alpha_i, \varepsilon_j) = 0 \text{ for all } i, j$$

From this we can derive

$$Var(y_i) = \sigma_\alpha^2 + \sigma_e^2,$$

which are called the *variance components* of the model.

In general, the mixed linear model can be formulated as

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e},$$

where in GWA studies,

$\boldsymbol{y}$: $n \times 1$ vector of phenotypes

$\boldsymbol{X}$: $n \times q$ matrix of genotypes

$\boldsymbol{\beta}$: $q \times 1$ vector of the coefficients of the fixed effects

$\boldsymbol{Z}$: $n \times t$ incidence matrix of individual membership in populations

$\boldsymbol{u}$: vector of random effects

Here,

$$Var(\boldsymbol{e}) = \sigma_e^2 \boldsymbol{I}$$

$$Var(\boldsymbol{u}) = \sigma_\alpha^2 \boldsymbol{K},$$

and thus,

$$Var(\boldsymbol{y}) = \sigma_\alpha^2 \boldsymbol{ZKZ'} + \sigma_e^2 \boldsymbol{I}.$$

The notable addition in these equations are the matrices *Z* and *K*. *Z* is simply an incidence matrix which allows the relevant random effects to be assigned to the relevant individuals. In the drug example, patients would be assigned to clinics via the *Z* matrix.

Generally speaking, *K* constrains the covariance structure of the individuals in the analysis, which is not defined by the symmetrical error ε_i alone, which can be defined in an arbitrary way, as long as the matrix is positive semidefinite, i.e., the eigenvalues of the matrix are strictly positive, which is important in solving for the parameters. In GWAS, this

matrix is typically calculated as an identity-by-descent (IBD) or identity-by-state (IBS) matrix. In an IBS matrix, a nucleotide or sequence segment is considered identical between two individuals if they are simply the same. In an IBD matrix, there is the additional condition for identity that the nucleotide or sequence segment is inherited without recombination from a common ancestor.

Kang et al. (25,28) cite their own observations and several other studies that show that use of various types of kinship matrices such as IBS and IBD result in little estimation differences, and in some cases IBS is able to outperform more complex matrices, though they note that this may change in studies where the individuals are more recently related. Due to the nature of GWAS, it is oftentimes impossible to determine exact ancestries of the samples, and thus an IBS matrix is usually applied.

Mixed linear models may be solved using a restricted maximum likelihood estimation, where the variance components σ_α^2 and σ_e^2 are estimated while the fixed effects are integrated out. This involves several expensive operations, including taking the eigendecomposition and inverse of the kinship matrix, which in cases of large GWA studies, can be extremely unwieldy. Thus, scientists in the GWAS community have been at the forefront in creating software to handle large datasets quickly and accurately. These software are discussed in section 2.2.6: *Current methods in GWAS*.

Mixed linear models are typically solved either by maximum likelihood (ML) or restricted maximum likelihood (REML) methods. The algorithms we investigated in chapters 2 and 3 typically allow users to select either ML or REML and in each case we select REML. The issue with the maximum likelihood estimators (MLE) of the variance components is that they can be biased; furthermore, when we are primarily interested in the estimation of the variance components, it may be beneficial to bypass the estimation of all parameters in favor of one which eliminates nuisance parameters, such as the fixed

effects (63,64). The REML first estimates θ, the vector of all variance components, after which estimation of the fixed effects β can be estimated by ML, with $\theta = \hat{\theta}$, with $\hat{\theta}$ the REML estimator. Later, Bayesian derivations of the REML estimator showed that it could be derived as the marginal likelihood of the model with the fixed effect parameters integrated out (63).

In the chapters that follow we consider the efficacy of mixed linear models in correcting for population and cryptic relatedness in a highly structured population *S. cerevisiae*. We consider the benefits of including additional fixed effect factors to a mixed linear model and find that in some cases, the additional factors increase the power of our model. Then, we apply a mixed linear model to a new type of problem, of computationally discovering small RNA motifs, and show that reinterpreting the parameters of the MLM in GWAS leads to improved miRNA prediction, by correcting for background sequence composition.

CHAPTER 2: GENOME-WIDE ASSOCIATION STUDIES IN HIGHLY STRUCTURED POPULATIONS

2.1: INTRODUCTION

As covered briefly in the introduction, one of the primary issues with GWAS today is the presence of different types of population structure in experimental samples. In this chapter we investigate the effectiveness of using GWAS to map complex traits in highly structured populations such as *Saccharomyces cerevisiae*. *S. cerevisiae* is a model organism for several reasons, some of which are: Firstly, it is relatively easy to acquire and study; secondly, due to its small genome size, it is relatively cheap to sequence; and thirdly, it is widely used in human consumption. The different strains of *S. cerevisiae* are commonly used in beer, wine, and sake fermentation, as well as bread baking. Other strains are associated with food spoilage, natural fermentation, and on the fruits of ripe plants (65).

We demonstrate that while the population structure of *S. cerevisiae* is well-defined, association studies still benefit from applying mixed linear models for the purposes of correcting for cryptic relatedness between individual strains. Furthermore, we demonstrate that the addition of a local ancestry variable can reduce the deviance between expected and observed values even more.

2.2: DATA AND METHODS

2.2.1: S. CEREVISIAE SEQUENCE AND EXPRESSION DATA

We obtained 38 whole-genome sequences of *S. cerevisiae* from Liti et al. (65). Each sequence belongs to one of the five known populations of the yeast, labeled European,

Malaysian, North American, West African, and Sake. These strains were from varying sources, including the lab, baking, fermentation, and sake (65). In total we found 150,077 SNPs over 16 chromosomes, excluding triallelic sites. After applying a sequential SNP selection procedure to eliminate SNPs that are tightly linked (see 2.2.5: SNP selection procedures), we retain 3,723 SNPs.

In addition to genomic data, we obtained quantitative phenotype expression data for most of the same strains for 201 traits from Warringer et al. (66). The 201 traits correspond to three types of measurements in each of 67 different environments. The three measurements types are growth rate, adaptation, and efficiency, where growth rate is how quickly proliferation occurs, adaptation is proliferation lag, and efficiency is how population density change (66). There was both sequence and expression data for 35 strains of *S. cerevisiae*.

2.2.2: ESTIMATION OF GLOBAL AND LOCAL ANCESTRY

We used the program STRUCTURE (27,67) to determine both global and local ancestry of each of our 35 strains of *S. cerevisiae*. For each sequence, global ancestry is interpreted as the percentage of the sequence which originated from a particular ancestral strain, for each of K ancestral strains. The number of ancestral strains is predefined by the user and given to STRUCTURE as a parameter. To determine the most likely number of ancestral strains, we ran a shorter iteration of STRUCTURE (see *2.2.3: Brief description of STRUCTURE and parameters*) for values of K from 3 to 8 (1). For each iteration, STRUCTURE returns a log-likelihood of the probability of the data, $\ln \Pr(X|K)$, given the parameters. Higher (i.e. less negative) log-likelihoods indicate a better fit to the data. Though the selection of K with STRUCTURE via this method is *ad hoc*, we find that K=6 provides the best fit to the data, which is concordant with previous studies (65,68), and so we use this

parameter for an extended analysis of the data in order to obtain the most accurate estimate of the population structure (see long run parameters in *2.2.3: Brief description of STRUCTURE and parameters*). A breakdown of global ancestry for each strain can be seen in Figure 3.

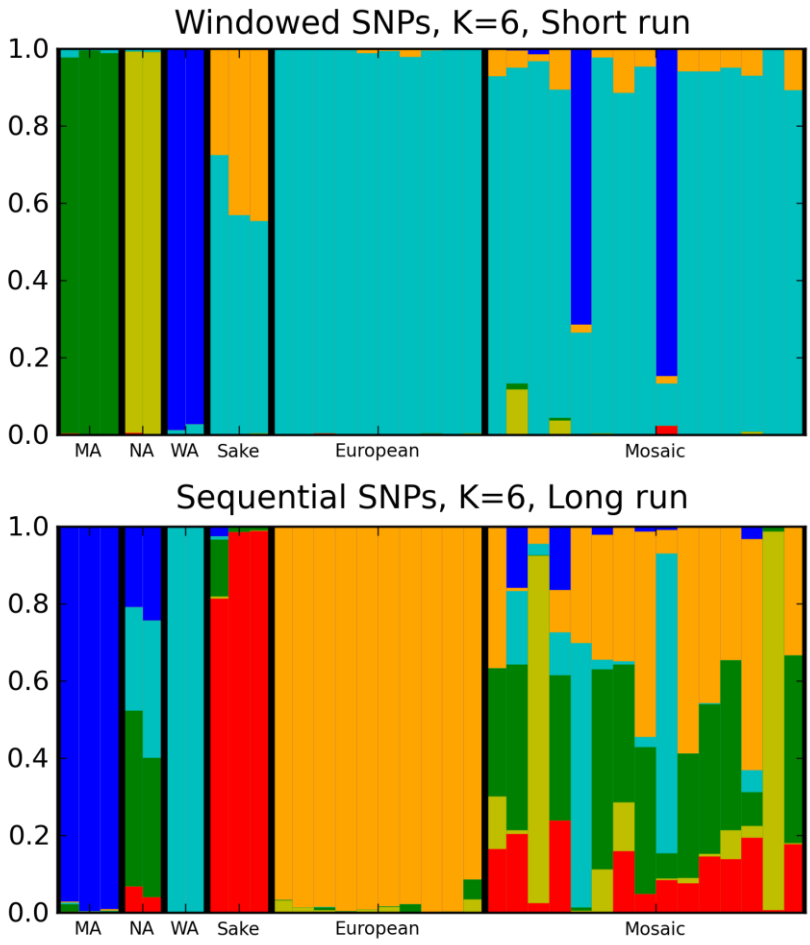


Figure 3

Results of STRUCTURE global ancestry estimates, both for windowed and sequential SNP selection procedures. STRUCTURE estimates are very sensitive to the choice of SNPs used in the analysis—the selection of SNPs relatively uniformly along the genome results in decreased signal in the results. The global ancestry maps seen in the bottom, resulting from sequential SNP selection, is very similar to maps of global ancestry in *S. cerevisiae* obtained previously (65,68).

We found that the Malaysian, West African, Sake, and European strains were primarily descended from a single ancestral population, whereas the North American

strains were a so-called "mosaic" of four different ancestral populations. The remaining strains show heavy mosaicism, and are grouped together and referred to simply as mosaic strains. These estimates of global ancestry are in line with a PCA plot of the same SNPs, which shows five distinct populations as well as the mosaics more spread out (see Figure 5).

Figure 3 also demonstrates that the SNPs used in the STRUCTURE analysis significantly impact the estimates obtained. Because SNPs in background linkage disequilibrium (LD) confound STRUCTURE results, they must first be removed. The way in which these SNPs are removed affects the results to no small degree. The results obtained in Figure 3 are obtained using the sequential SNP selection procedure detailed in *2.2.5: SNP selection procedures*.

In addition to global ancestry estimates, STRUCTURE also produces local ancestry estimates based on a hidden Markov model (27). To verify the STRUCTURE results, we used the program WINPOP, which has improved accuracy for local ancestry estimation when the ancestral populations under analysis are closely related (69). For WINPOP, we needed to input both "ancestral" strains genotypes as well as mosaic strain genotypes in order to learn the local ancestry of the mosaics. For this purpose, we assumed that the populations deemed by STRUCTURE to be primarily unmixed, that is, the Malaysian, West African, and European strains, to represent the "ancestral" genotypes. The North American strains were grouped with the mosaics for this analysis. Further details of the WINPOP algorithm and parameters used can be found in *2.2.4: Brief description of WINPOP and parameters*.

2.2.3: BRIEF DESCRIPTION OF STRUCTURE AND PARAMETERS

The authors of STRUCTURE introduce it as a Bayesian clustering algorithm, which seeks to determine two parameters: the population of origin of the individuals *Z*, and the

allele frequencies of the ancestral populations *P*, given the genotypes of the individuals in a sample *X* (67). In the original paper, the authors assumed Hardy-Weinberg equilibrium within populations and complete linkage equilibrium between loci in each population and were interested in finding *Z* and *P* that maximized the likelihood function $\Pr(X|Z,P)$. In the admixture model, there is an additional parameter *Q*, a vector of length *K* for each individual denoting the proportion of the individual's genome that originated from one of *K* ancestral populations. Let $q^{(i)}$ denote the admixture proportions of individual *i*, $q^{(i)} = \left(q_1^{(i)}, \ldots, q_K^{(i)}\right)$, with $q_j^{(i)}$ denoting the proportion of individual *i*'s genome derived from ancestor *j*. The authors assume a Dirichlet distribution for $q^{(i)}$, i.e. $q^{(i)} \sim \mathfrak{D}(\alpha, \alpha, \ldots, \alpha)$ (67). The parameter α essentially determines the distribution of ancestry proportions: When α is small, an individual's ancestry will be primarily from a single population; when α is large, an individual's ancestry will evenly originate from all *K* populations.

The parameters are solved using an MCMC algorithm and Gibbs sampling, via Algorithm 2 in (67). In short, the parameter *P* and *Q* are first sampled from the distribution $\Pr(P, Q|X, Z^{(m-1)})$. Then, *Z* is estimated from $\Pr(Z|X, P^{(m)}, Q^{(m)})$. Lastly, α is updated using a Metropolis-Hastings step. These three steps are performed iteratively for a user-specified number of steps. The parameter not estimated from the data is *K*, the number of ancestral populations. While the authors give some indications on how to estimate *K* by approximating $\Pr(K|X)$, the solutions are *ad hoc* and we primarily rely on our own prior knowledge of how many distinct populations are on our data set. For example, values of *K* less than 6 in our data fail to show clear separation of the five "clean" populations of *S. cerevisiae* and the mosaic strains.

Updates in the STRUCTURE algorithm (27) allowed it to handle some types of linkage disequilibrium between markers. In our analyses, we allow both admixture to occur and apply the linkage model. Introducing admixture into the model allows for an individual

to belong to multiple ancestral populations, while the linkage model accounts for linkage disequilibrium (LD) between markers that arise due to admixture (27). The latter, termed "admixture linkage disequilibrium", should not be mistaken for "background linkage disequilibrium", or LD present in ancestral populations prior to admixture. STRUCTURE's linkage model accounts for two types of linkage: Admixture LD, mentioned previously, and mixture LD, which the authors termed the linkage expected when large chunks of a genome are inherited together. Essentially, the linkage model allows markers to be inherited in linkage blocks that derive from a single ancestral population, instead of considering each marker separately.

The MCMC procedure in STRUCTURE jointly estimates population allele frequencies for each of the ancestral populations and also the population of origin of each locus in the individuals being analyzed. The number of ancestral populations *K* is predefined by the user. For the purposes of model selection, STRUCTURE returns the estimated log-likelihood of the data given the choice of *K*.

We used two sets of parameters for two separate types of STRUCTURE runs: One was a preliminary run used for SNP selection (see 2.2.5: SNP selection procedures), where all parameters used were default except for PLOIDY = 1; BURNIN = 5000; NUMREPS = 5000; LINKAGE = 1; ADMBURNIN = 2500; and SITEBYSITE = 1 (1). Briefly, these parameters can be described as: PLOIDY: ploidy of the organism; BURNIN: how many iterations of MCMC to run before data collection begins; NUMREPS: number of iterations of MCMC to run after the burn-in period; LINKAGE: indicates usage or not of the linkage model; ADMBURNIN: number of burnin iterations to run with the admixture model when using the linkage model; SITEBYSITE: whether or not to output local ancestry percentages (1 indicates True).

We also ran STRUCTURE for an extended period of time in order to obtain the most accurate estimations of local and global ancestry. The long STRUCTURE runs used for the

rest of the analysis used the same parameters except for BURNIN=50000; NUMREPS=50000; ADMBURNIN=25000.

2.2.4: BRIEF DESCRIPTION OF WINPOP AND PARAMETERS

The primary purpose of STRUCTURE is to provide global estimations of ancestry; however, STRUCTURE also returns per-locus estimations. We use the STRUCTURE estimations of ancestry but verify that these are accurate using WINPOP, a program specifically addressing the issue of determining local ancestry where the ancestral populations are closely related. Our results show that the STRUCTURE and WINPOP results are similar, and as we do not believe the ancestral populations of *S. cerevisiae* to be closely related enough as to warrant a separate estimation of local ancestry using WINPOP, we use the STRUCTURE estimations of local ancestry for consistency in the remainder of our analyses.

WINPOP builds off of a previous method LAMP (70), which uses a windowed method, assigning ancestry per window and casting a vote for markers that fall into more than one window. The shortcomings of LAMP that WINPOP addresses are 1) LAMP assumes no recombinations have occurred within any given window; 2) window length is a fixed number that depends on the number of generations since admixture and recombination rate, which means it remains fixed regardless of how closely or distantly related the populations admixed are (69). The authors demonstrated that by implementing these two fixes, namely by assuming a single recombination per window instead of none, and allowing for variable window size dependent on the genetic similarity between the two admixed populations within the window, the improvement in estimating local ancestry can be substantial. This is particularly true when the admixed populations are very similar, such as admixture between Japanese and Chinese populations (69).

WINPOP estimates per locus ancestry for admixed populations, and unlike STRUCTURE must be given genotypes of the ancestral populations being mixed. Thus, we took the Malaysian, West African, Sake, and European strains as the ancestral populations (as estimated by STRUCTURE), and took the North American and remaining mosaic strains as the admixed populations, to estimate their local ancestries. WINPOP also requires a recombination rate *r* and an estimated number of generations that have passed since admixture *g*. The former we took from a previous study (71) for a value of $r = 3.5 \times 10^{-6}$.

Additionally, we found the results to be very sensitive to the parameter *g*, the estimated number of generations that have occurred since admixture. While the number of generations since admixture is unknown, since *S. cerevisiae* primarily reproduces asexually it can be estimated that the number of outcrossing events since the MCRA of two particular strains is g = 314 (71). STRUCTURE gives a very different estimate of g = 28. Thus, we ran WINPOP for the following range of generations: 5, 10, 15, 28, 157 and 314. The first four values result in very similar patterns of ancestry while the last two values give very noisy results. We found that the local ancestry calls were significantly similar between WINPOP and STRUCTURE (data not shown) so we used the STRUCTURE local ancestry estimates for the remainder of the analysis.

2.2.5: SNP SELECTION PROCEDURES

As described previously, it is important to remove SNPs in high linkage disequilibrium (LD) because STRUCTURE does not model background LD. To do this, we implemented two procedures for selecting an independent set of SNPs: a "sliding window" procedure and a "sequential" procedure to test different methods of SNP selection and how it affects STRUCTURE results.

For the sliding window procedure, we selected a small number of SNPs within each window of *N* consecutive SNPs. The SNPs within a window were selected using a linkage disequilibrium (LD) criterion, where we computed, for each pair of SNPs in the window, the measure of LD defined by D'^2, where we compute D' as

$$D' = \frac{D}{D_{max}} = \frac{Freq(A_1B_1) - Freq(A_1)Freq(B_1)}{D_{max}}$$

$$D_{max} = \begin{cases} \min\left(Freq(A_1)Freq(B_1), Freq(A_2)Freq(B_2)\right), D < 0 \\ \min\left(Freq(A_1)Freq(B_2), Freq(A_2)Freq(B_1)\right), D > 0 \end{cases}$$

If the LD was above a threshold, the SNP with more missing data was removed, with ties broken randomly. We tested window sizes of N = 10 to 100 consecutive SNPs in increments of 10, and D'^2 thresholds of 0.1 to D = 0.9 in increments of 0.1. These parameters resulted in approximately 1700 SNPs to 15,500 SNPs selected out of a total of 150,077 SNPs. Most windows contained one SNP, resulting in a roughly uniform distribution of SNPs across the genome. The window size N was the main determinant of the number of SNPs selected and the threshold D'^2 had a relatively small effect on the choice of SNPs (data not shown). We chose a final window size of N = 30 as the number of SNPs chosen at this window size was approximately the average of all trials (data not shown). Instead of removing the SNP with more missing data, we also implemented several other procedures, including removing a random SNP, choosing the SNP with the higher minor allele frequency (MAF) and choosing the SNP that was more differentiated among the ancestral populations, as previously described (72). However, the set of SNPs selected was essentially unchanged (data not shown).

For our sequential SNP selection procedure, we identified LD blocks along a chromosome and selected one representative SNP per block. Starting with the first SNP in a chromosome (SNP A), we calculated LD as described previously between this SNP and the next SNP (SNP B). If SNP B was in high LD with SNP A where "high LD" is defined by some upper bound on D'^2, the SNP with less missing data was kept, with ties broken randomly. If SNP B was in LD with SNP A but not above a certain threshold (lower bound), the next SNP was considered and no change was made to SNP A. If SNP B was not in LD with SNP A, then we kept SNP A as the defining SNP for the previous LD block and let SNP B begin a new LD block. We varied the lower bound cutoff and the upper bound cutoff from $D'^2 = 0.05$ to 0.95 in increments of 0.05 and found that the upper bound cutoff did not significantly impact the number of SNPs selected, while the lower bound cutoff had a significant impact, with more SNPs selected with a higher lower bound. This trend plateaued at a lower bound of 0.90 (given an upper bound of 0.95), with approximately 5700 SNPs. We chose a D'^2 cutoff for high LD of 0.95 and a D'^2 cutoff for low LD of 0.16, resulting in a total of 3723 SNPs. We compared the distribution of SNPs chosen with a more relaxed lower bound of $D'^2 = 0.5$ to the distribution of SNPs chosen with $D'^2 = 0.16$ and found that both patterns were consistent (data not shown).

We compared the SNPs selected by both of these methods in each chromosome by looking at a few measures: First, we looked at the distribution of SNPs across the chromosome. Second, we compared the SNP distribution with known areas of high and low recombination rate. These two measures give us a sense of how much information is captured by the selected SNPs—we want to select more SNPs in areas of higher recombination and low LD and fewer SNPs in areas of low recombination and high LD, since we are interested in SNPs that can give us a better view of the ancestral landscape.

We obtained regions of exceptionally high and low recombination activity from (73) and show a representative plot of the overlap of these regions with the SNPs selected by the windowed and sequential methods in Figure 4.

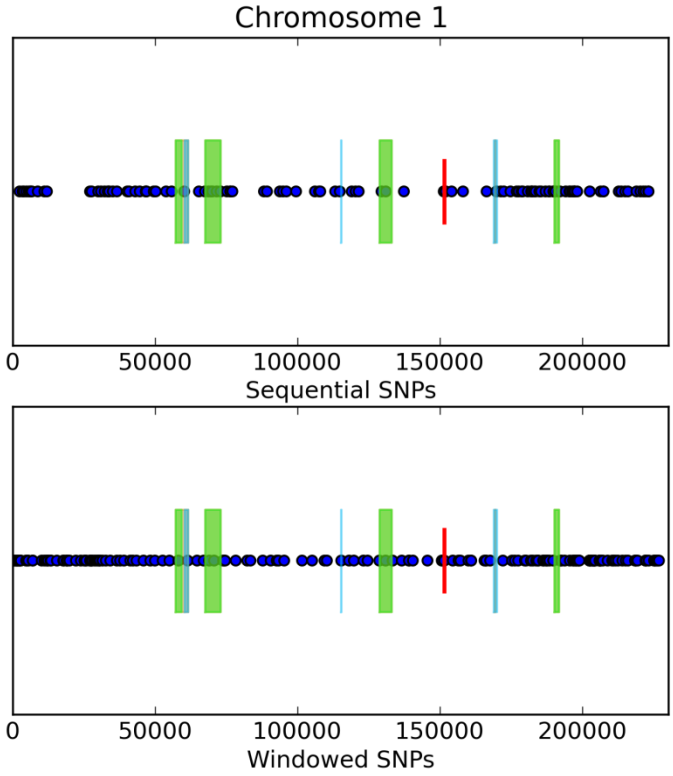


Figure 4

Comparison of sequential and windowed SNPs. SNPs selected by the windowing method (bottom) tend to be more uniform across the genome whereas SNPs selected by the sequential method (top) tend to follow the recombination landscape of the genome more closely. Red lines represent the position of the centromere, bright green bars represent crossover recombination hotspots, cyan rectangles represent non-crossover recombination hotspots, and orange represents overall recombination hotspots, as determined by (73).

We see that the SNPs selected by the sequential method are spread much less uniformly along the chromosome, and that in general we see that blocks of high crossover recombination correspond to denser SNP islands. Furthermore, we see that fewer SNPs

were selected near the centromeres (represented by the red bar), where fewer crossover recombinations occur. These patterns are not observed by SNPs selected via the windowed method, which in general are spread very uniformly regardless of recombination hotspots. Thus, we performed all further analyses with SNPs selected by the sequential method. When we set the LD cutoff to $D'^2 = 0.16$, the sequential procedure resulted in a set of 3723 SNPs, which is similar to the number of SNPs used in the STRUCTURE analysis of (65). To reiterate the admixture result of STRUCTURE and our selection of k=6, we performed principal components analysis using the 3723 SNPs and found that the non-admixed strains found by STRUCTURE cluster together (Figure 5) in five distinct clusters, suggesting that our selection of k was appropriate.

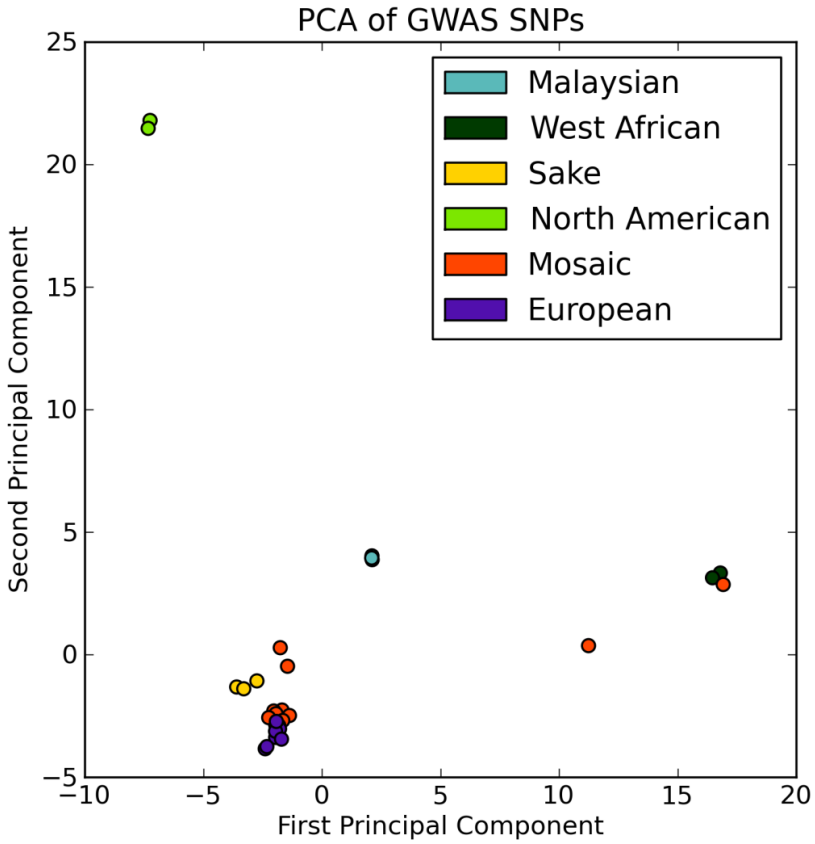


Figure 5

PCA plot of sequentially selected SNPs. Second principal component plotted against the first for each of the strains after performing principal components analysis on the genotype matrices including the sequentially selected SNPs. The results are consistent with the STRUCTURE estimates of global ancestry for each of the six "populations".

2.2.6: CURRENT METHODS IN GWAS

We focused on the presence of global and local ancestry covariates in both linear and mixed linear models. A general linear model (which we refer to as LM) without covariates is our baseline comparison model, as it does not correct for any kind of population structure. In LM models, the estimates of global ancestry can be used as a fixed covariates to correct for population stratification. We also analyze the linear model with local ancestry estimate covariates, as well as the linear model with both global and local ancestry estimate covariates. In addition to the simple linear model, a popular method for population stratification correction is the mixed linear model (MLM) with the kinship matrix as a random effect. We utilize MLM with global, local and both global and local ancestry estimates as covariates in our analyses. At the time of these analyses, there were several existing programs that implemented these methods in an efficient way for genome-scale studies, including TASSEL (74) and EMMAX (28), which both produce approximate results. TASSEL can run both LM and MLM models, while EMMAX can only run MLM. We implemented an LM both with and without covariates in R, and we used both EMMAX and TASSEL's MLM algorithms. A summary of all methods follows:

LM methods, implemented in R:

1. Whole-genome ancestry covariates (estimated by STRUCTURE);
2. Local ancestry covariates (estimated by STRUCTURE);

3. Both whole-genome and local ancestry covariates.

MLM methods:

1. Kinship matrix only (estimated by EMMAX-KIN), implemented in EMMAX and TASSEL;
2. Kinship matrix with whole-genome ancestry fixed covariates, implemented in EMMAX and TASSEL;
3. Kinship matrix with local ancestry fixed covariates, implemented in EMMAX;
4. Kinship matrix with both global and local ancestry fixed covariates, implemented in EMMAX.

A list of these methods and their abbreviations can be found in Table 1.

Abbreviation	Statistical method	Covariates
R-LM	LM	None
R-Q	LM	Q
R-LA	LM	LA
R-LAQ	LM	LA+Q
EMMAX-K	MLM	K
EMMAX-QK	MLM	Q+K
TASSEL-K	MLM	K
TASSEL-QK	MLM	Q+K
EMMAX-KLA	MLM	LA+K
EMMAX-KLAQ	MLM	LA+K+Q

Table 1

Summary of GWAS methods used. The first column contains the abbreviation we use in this manuscript; the second column gives the statistical model; the last column lists the covariates used along with each statistical model.

The two MLM solvers we used initially in these analyses were EMMAX and TASSEL, which are essentially two implementations of the same algorithm. However, we note that we found discrepancies in their performance, with EMMAX clearly outperforming TASSEL in terms of speed and accuracy (in some cases we found that TASSEL would return NANs where EMMAX would not). Therefore we used EMMAX for all of our analyses.

EMMAX is an updated version of the program EMMA (25). Its advantage lies in a smart approximation made under the assumption that most SNPs tested in GWAS have a small effect. Briefly, Kang *et al.* (28) note that one of the issues with GWAS is that while the model being tested is generally represented as

$$y_i = \beta_0 + \sum_{k=1}^{M} \beta_k X_{ik} + e_i$$

where k is the SNP index, so that the summation represents the effect of each of M SNPs on the phenotype of individual i, the model being tested is necessarily

$$y_i = \beta_0 + \beta_k X_{ik} + \eta_{i\bar{k}},$$

i.e., only the effect of a single SNP at a time is being analyzed. From a comparison with the previous equation, we see that error term $\eta_{i\bar{k}}$ here includes the effect of all SNPs in the full model that are not SNP k. In other words, $\eta_{i\bar{k}} = \sum_{l \neq k} \beta_l X_{il} + e_i$. In the case where the individuals in the sample are completely unrelated so that there are no dependencies among the X_{il}s, $\eta_{i\bar{k}}$ can be assumed to be independent and identically distributed, and the model reduces to a simple linear model. However, when individuals exhibit some amount of relatedness, or for example when linkage affects the X_{il}s, this is no longer the case. In these cases, the mixed linear model can be used to estimate the $\eta_{i\bar{k}}$ properly using variance components techniques (28).

The issue with using mixed linear models on such a large dataset is that solving for the variance components is computationally demanding, and if it needs to be solved for each of possibly hundreds of thousands of SNPs, it may be prohibitively so. EMMAX makes the simplifying assumption in this case that

$$\eta_{i\bar{k}} = \sum_{l \neq k} \beta_l X_{il} + e_i = \sum_{k} \beta_l X_{il} + e_i$$

so that the variance components only need to be estimated once, and then reused for all SNPs tested. This assumption essentially states that the effect of each individual SNP is small compared to the sum of effects of all the SNPs, which can be true in many phenotypes analyzed in GWAS.

More recently, other fast and exact methods of solving mixed linear models have been proposed, namely GEMMA and FaST-LMM. However, at the time of this analysis these were not yet available, and so we give an overview of these programs in section *3.3.2: Application of mixed linear models to motif discovery: MixMir*.

2.2.7: COMPUTING THE KINSHIP MATRIX

In GWAS, the relatedness matrix typically defines pairwise genetic relatedness between all individuals in the study, and as such is referred to as the "kinship matrix". We compute this matrix for all MLMs performed using EMMAX-KIN, a script which comes in the EMMAX package. The program PLINK is first used to convert .ped and .map files containing genotype information to .tped and .tmap files, which are readable by EMMAX-KIN. EMMAX-KIN can compute two different types of kinship matrices, "identity-by-descent" (IBD) and "identity-by-state" (IBS). We elected to use the IBS option, as suggested by (75). The difference between IBD and IBS is that in the calculation of an IBS matrix, we are only interested in shared alleles between individuals and not necessarily in the those alleles

which are shared due specifically to descent. The authors Zhao *et al.* found that in their analysis of *Arabidopsis thaliana*, using the IBS matrix removed more false positives than the IBD matrix, and in fact that it alone did almost as well as using both kinship matrix and global ancestry covariates (75).

2.2.8: SIMULATION STUDIES

In addition to our analyses with the true *S. cerevisiae* genotypes and phenotypes, we also performed two sets of simulations, in which we tested the power of each statistical method in recovering "true" SNPs for simulated phenotypes.

In the first set of simulations, we randomly selected as a baseline phenotype the "maltose 2% growth rate" phenotype from Warringer *et al*. (66). Then from the 3723 SNPs used in our analyses, we randomly assigned n of them to be "causal" for the phenotype, such that each strain harboring the major allele of each of the "causal" SNPs received an additive fixed effect of size V to the phenotypic value of the strain. We tested a variety of genetic architectures by allowing n = [3, 10, 100] and V = [1, 3], representing few to many SNPs of small effect; as well as one set of simulations with n = 3 and V = 10 to simulate the case of few SNPs of large effect. After adding in the fixed effects of the "causal" SNPs, we then re-normalized the phenotype in a standard way by subtracting the mean and dividing by the standard deviation. The mean and variance for the original maltose 2% growth rate phenotype were 0.9347 and 0.5781, respectively.

One of our concerns with the various models we tested was potentially having too few samples for sufficient power. Since the *Liti et al.* data contained only 35 strains, we looked to a study by Schacherer *et al.* (68), which contained the sequences for almost twice as many strains of *S. cerevisiae*. To perform a set of simulations comparable to the first and also in line with the rest of our analyses with the Liti et al. (65) data, we applied our

sequential SNP selection procedure to the set of 101,343 SNPs reported by Schacherer *et al.* (68). This resulted in a set of 12,916 SNPs. Again we used STRUCTURE to obtain estimates of global and local ancestry with $K = 6$ ancestral populations, which is consistent with the number of populations observed by Schacherer *et al.* (68), who used $K = 5$ instead. With $K = 6$ we see a similar picture of global ancestry (see Figure 6). As before, we used the IBS kinship matrix obtained by EMMAX-KIN for all mixed linear models.

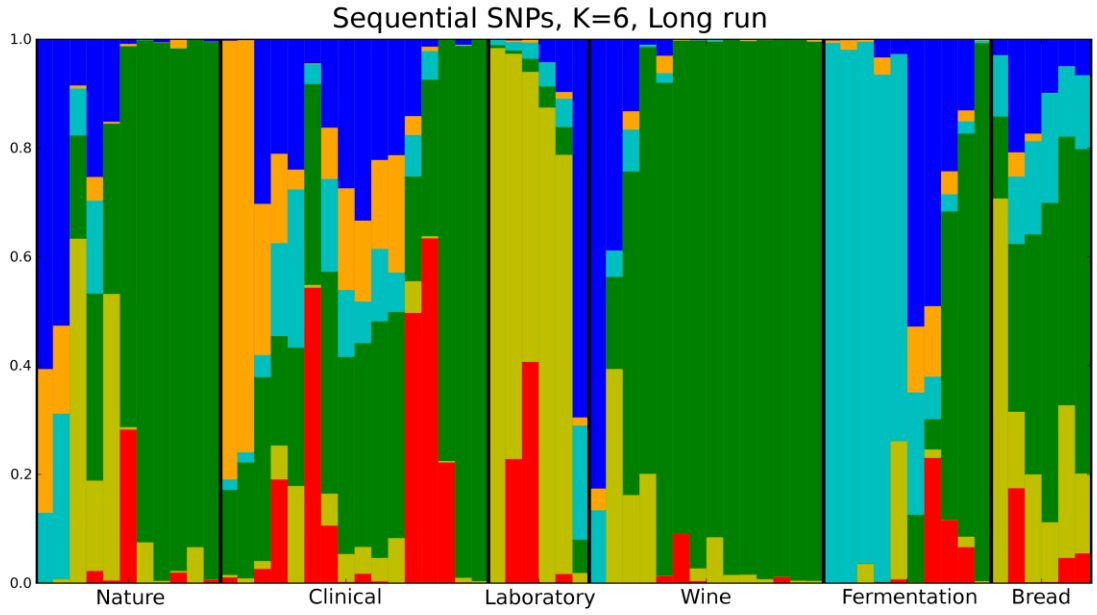


Figure 6

Results of STRUCTURE global ancestry estimates for 63 strains of *S. cerevisiae* from Schacherer et al. Again we use the sequential SNP selection procedure described previously, and assume $K = 6$ ancestral populations. Compared to our original 35 strains from Liti et al. , we find that the population structure is less clear, especially without the characteristic "clean" strains.

Since we did not have phenotypic data for the 63 strains in the larger data set, we simulated phenotypes by first setting the phenotypic value for all strains to 0, and then continued with adding phenotypic effects as in the first set of simulations. i.e., for each strain harboring the major allele of a randomly selected "causal" SNP, we add a fixed

additive effect to the phenotype. As in the first set of simulations, we let n = [3, 10, 100] and V = [1, 3], and additionally ran one set of simulations with n = 3 and V = 10 to simulate the case of few causal SNPs with very large effect.

Each set of simulations consisted of 200 iterations for each of the pairs of possible n and V values, for each of the statistical methods listed in Table 1. Analysis of the results, i.e. comparison of the performance of each of the statistical methods, was performed using receiver operator characteristic (ROC) curves. ROC curves are a standard way of measuring performance by plotting the true positive rate against the false positive rate, graphed as:

$$y = \frac{TP}{TP + FP}$$

$$x = 1 - \frac{TN}{TN + FP}$$

where *TP*, *TN*, and *FP* are the numbers of true positive, true negative, and false positive predictions, respectively. The area under the ROC curve, or the AUROC, can be used to compare ROC curves, where the larger AUROC value indicates better performance. In our simulation studies, the average ranks of the planted "causal" SNPs was taken and plotted as the ranks of true positives.

2.3: RESULTS

2.3.1: PHENOTYPES AND POPULATION STRUCTURE OF 35 STRAINS OF S. CEREVISIAE

We obtained 35 strains of *S. cerevisiae* from Liti *et al.*. For each of these strains we also obtained their phenotypes in different environments, which consisted of three types of quantitative measurements in each of 67 environments. We consider each such

measurement in each environment a "phenotype". With these 201 phenotypes, we investigated the possibility of performing genome-wide association studies (GWAS) in a highly structured population such as *S. cerevisiae*, the issue with such populations being the presence of highly confounding factors resulting from relatedness between individuals. To address this issue, we first needed to determine the structure of the population.

One popular program for estimating population structure was developed by Pritchard *et al.* and aptly named STRUCTURE (27,67,76), which estimates the amount of shared ancestry in a given sample using Markov chain Monte Carlo (MCMC). One issue with STRUCTURE that is not frequently discussed is the way in which SNPs must be pruned prior to use in the program, as certain types of linkage disequilibrium (LD) are not accounted for in the model. As such, we compared different SNP selection procedures and found that a sequential procedure that takes into consideration the LD landscape in the genome produces the most accurate STRUCTURE results (see *2.2.5: SNP selection procedures*). This resulted in a set of 3723 SNPs distributed across 16 chromosomes according to the linkage disequilibrium landscape, with fewer SNPs concentrated in areas of high LD and fewer recombination events (such as near centromeres) and more SNPs in areas of higher recombination, consistent with previous reports of recombination hotspots in *S. cerevisiae* (73).

Running these SNPs through STRUCTURE and setting $K = 6$ ancestral populations gave us a clear picture of the relatedness between individual strains on a global scale. We obtained local ancestry estimates using the same STRUCTURE results, which were verified again with a separate program (see *2.2.2: Estimation of global and local ancestry*).

We performed a preliminary analysis of the 201 phenotypes using a simple linear model. An empirical way of determining the amount of population structure present is by plotting the *p*-values obtained for each SNP against their expected *p*-values in what is

commonly known as a QQ plot. Expected *p*-values under the null hypothesis should follow a null distribution, and in the case of GWAS, if the number of significantly associated SNPs is small compared to the total number of SNPs tested, we should observe in our QQ plots a line mostly following the *x*-*y* axis with some deviation at lower *p*-values, where the observed values would be lower than the expected.

We use as a quantitative measure of the QQ plot deviation the mean squared deviation, or MSD, between the expected and observed *p*-values. Using this as a score for genetic stratification, we determined the top ten most stratified phenotypes to be rapamycin 0.5 mg/ml adaptation; rapamycin 1mg/ml adaptation; pH 3.5 adaptation; LiCl 150 mM efficiency; CuCl2 0.75 mM rate; CuCl2 0.375 mM rate; KCl 1.45 M rate; CoCl2 0.015 mM adaptation; maltose 2% rate; and LiCl 225 mM efficiency.

Additionally, we took the phenotypic variance across all 35 strains for each of the phenotypes measured, and found that the average phenotypic variance of the most stratified phenotypes (0.77) was significantly higher than expected (0.30; P-value 0.0036, 100,000 boostrap replicates). This suggests that conditions with high levels of population structure at the genotype level are also more varied phenotypically. Since we have three types of measurements in each of 67 environments, we also consider which environments show the most genetic stratification across growth adaptation, efficiency, and rate measurements. These results are presented in Table 2.

From these preliminary analyses we concluded that correction for population structure is important in *S. cerevisiae*, and that depending on the phenotype different degrees of stratification may be present and may require different methods of correction.

Environment	**P-value**

Rapamycin	0.0006
CuCl2	0.0024
LiCl	0.026
Kcl	0.084
pH	0.095

Table 2

Environment types that were significantly genetically structured compared to the background. The *p*-value reflects the significance of how much more structured the environment was. We used as a measure of genetic stratification the MSD of the QQ plot.

2.3.2: COMPARISON OF STATISTICAL CORRECTIONS FOR GLOBAL AND LOCAL ANCESTRY

For each of the methods listed in Table 1 and each of the 201 phenotypes described in *2.2.1: S. cerevisiae sequence and expression data*, we performed a GWAS analysis. As described previously, there was some discrepancy between results obtained by TASSEL and by EMMAX, where EMMAX was faster and more accurate than TASSEL. Therefore, we omitted TASSEL from further analyses. For the remaining methods, we drew QQ plots for each phenotype. We present a series of these plots in Figure 7. For each phenotype, we also determine which method performs best using the MSD statistic as before. Across all 201 phenotypes, we see that EMMAX-KLA corrects for the most amount of stratification more than any other method, followed by R-LAQ. These results are presented in Table 3 and Figure 9, where we compare the overall performance using the average MSDSs, and also the variance of the MSDs obtained from each method. We find that the best and worst correction methods both involve the mixed linear model: the best and second best methods are EMMAX-KLA and EMMAX-K, respectively; the worst methods are EMMAX-QK and EMMAX-KLAQ. If the mixed linear models and the kinship matrix are correcting for effects

very similar to that of global ancestry, then we might expect then that adding the global ancestry covariate could overcompensate for stratification and cause the poor performance.

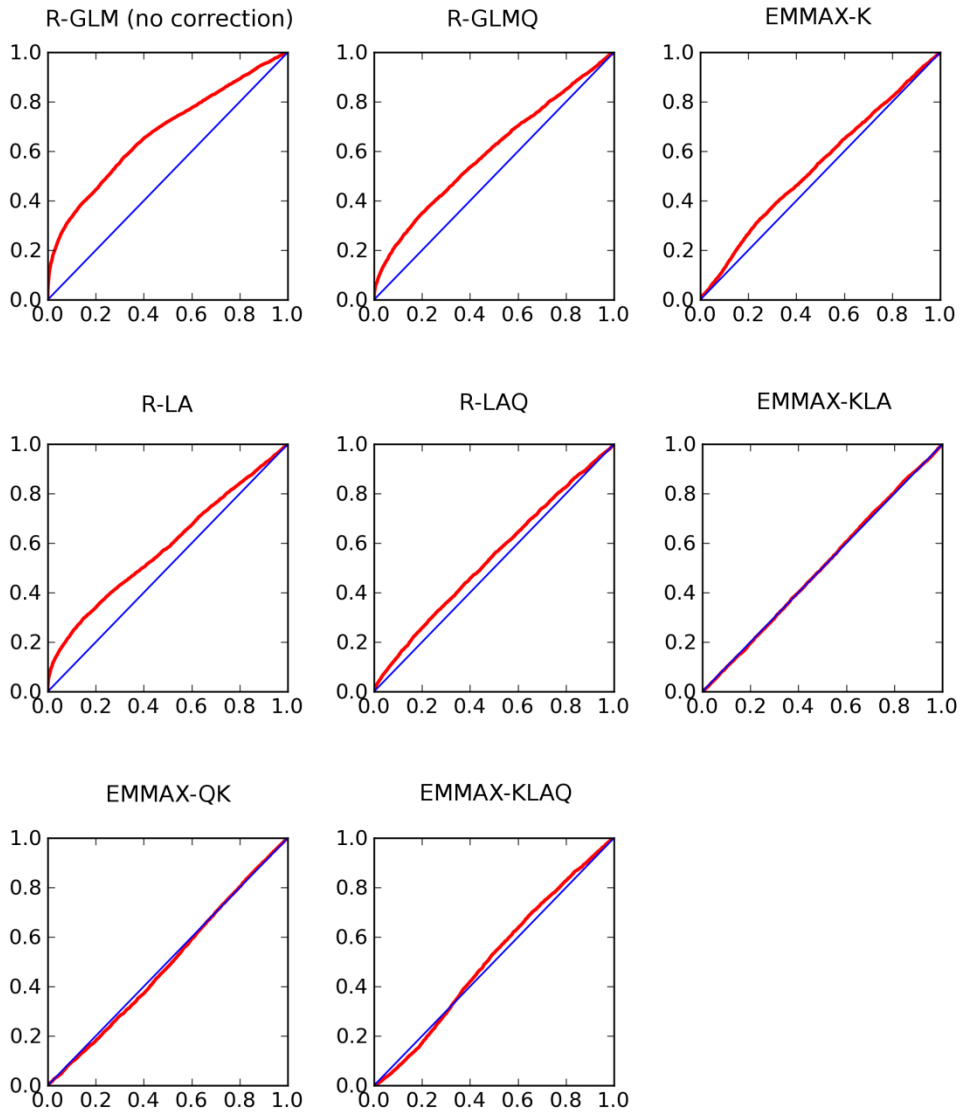


Figure 7

Representative QQ plots for the different statistical methods tested for population structure correction. In each plot, the *y*-axis shows the expected *p*-value and the *x*-axis shows the observed *p*-value. The blue line represents the observed = expected line, and the red line is what we actually observe. The large deviation of the latter from the former in the simple linear regression indicates that there is a significant amount of population structure

present associated with the phenotype. In this particular example, we see that EMMAX-KLA provides the best correction.

We also repeated these tests using the variance inflation factor (VIF) as proposed by (20) instead of the MSD, and found that at least in our case, there was no significant difference between the two, namely that EMMAX-KLA still performed better than the other GWAS methods as determined by the VIF (data not shown). We note that the mean of the test statistics has been proposed as a good estimator for the VIF (77,78), so our MSD statistic has some basis in formal statistical theory.

Statistical method	No. of phenotypes where method performed best
R-LM	11
R-Q	22
R-LA	14
R-LAQ	49
EMMAX-K	36
EMMAX-QK	5
EMMAX-KLA	62
EMMAX-KLAQ	2

Table 3

Number of phenotypes in which each method performed best. Using the MSD as the measure of performance, we ranked the all methods for each phenotype separately, and determined which method performed best. The second column is the number of phenotypes in which the respective method had the lowest MSD. While EMMAX-KLA performs the best more often than any other method, methods such as R-LAQ also perform well. On the other hand, EMMAX-QK and EMMAX-KLAQ performed the worst. This may be due to overcorrection

resulting from combining the mixed linear model and the global ancestry covariates, if they are correcting for similar effects.

Since it appears that different statistical methods may perform best under different circumstances, we selected the best method in each case and created the meta-statistical method which we will refer to as "BEST". We select as significant SNPs the ones which meet a threshold *p*-value of 0.05. While the number of SNPs and therefore the multiple testing burden was reduced by our SNP selection procedure, we still employed two types of correction: One which corrects for family-wise error, or FWER (Holm), and one which corrects for the false discovery rate, or FDR (Benjamini-Hochberg, BH).The number of GWAS loci detected under the Holm correction was small, considering that the total number of SNPs found was 389 by the simple linear model, across all 201 phenotypes. As another example, EMMAX-KLA finds 637 SNPs significant, so that the expected number of associated SNPs per phenotype is just over three. As seen by these numbers, we found a wide range of significant SNPs called by the different statistical methods (see Table 4). This demonstrates the importance of considering different kinds of statistical methods when performing a GWAS.

	Including duplicates		**Unique SNPs**	
Method	**Holm**	**BH**	**Holm**	**BH**
R-LM	389	2662	99	1200
R-Q	249	1024	106	415
R-LA	235	353	31	57
R-LAQ	68	97	14	29
EMMAX-K	431	2154	220	891

EMMAX-QK	191	1261	107	723
EMMAX-KLA	637	1477	65	171
EMMAX-KLAQ	90	144	17	44
BEST	162	359	63	177

Table 4

Number of SNPs found significant using each method across all phenotypes, after two different kinds of multiple testing correction: Family-wise error (FWER, Holm) and false discovery rate (FDR, Benjamini-Hochberg). For both we use a threshold of $p = 0.05$. Multiple testing corrections were made for each phenotype separately, not for the 201 phenotypes combined. For the numbers in the first two columns, we count SNPs which are found significant in more than one phenotype multiple times. For the last two columns, if a SNP is found significant in more than one phenotype, it is counted just once.

We compared the similarities between the statistical methods pairwise by taking the Pearson correlation of the *p*-values produced (Table 5). We notice that while there are generally higher similarities between the MLM-based methods, as expected, there is not a pervasive pattern of similarity given covariates or statistical method. For example, the highest degree of similarity exists between R-LM and EMMAX-K. Also, EMMAX-K and EMMAX-KLA are relatively dissimilar, at least in terms of the *p*-values obtained. This reiterates the importance of testing different statistical methods in GWAS analyses, as method similarity is not obvious.

	R-Q	**R-LA**	**R-LAQ**	**EMMAX-K**	**EMMAX-QK**	**EMMAX-KLA**	**EMMAX-KLAQ**
R-LM	0.1684	0.0850	0.0566	0.7341	0.2024	0.0753	0.0936
R-Q		0.2886	0.0989	0.2540	0.2847	0.1686	0.0557
R-LA			0.2654	0.0977	0.1190	0.2208	0.0849
R-LAQ				0.0619	0.0492	0.1027	0.0738

EMMAX-K	0.338	0.1374	0.1241
EMMAX-QK		0.2493	0.4173
EMMAX-KLA			0.2683

Table 5

Similarities between methods studied as calculated by the Pearson correlation between SNP *p*-values.

2.3.3: BIOLOGICAL AND FUNCTIONAL ANALYSIS OF GWAS SNPS

We performed a biological and functional analysis of the GWAS SNPs identified by the various statistical methods using the meta-statistical "BEST" method, as described in *2.3.2: Comparison of statistical corrections for global and local ancestry*. To investigate the functional significance of the statistically significant SNPs, we examined the fraction of significant SNPs contained in genes. While the *S. cerevisiae* genome is gene-rich, with 63% of all SNPs in our analysis falling in genes, we found that the SNPs called significantly under the FDR correction were enriched in genic regions. 75% of SNPs called significantly under the FDR correction were in genes compared to 63% of all SNPs used in our analysis (*p*-value, 1e-4, calculated by a permutation test).

We also examined the biological functions of the GWAS SNPs using the GO term enrichment program FuncAssociate (79). We found several interesting enriched functions, including biotin biosynthesis for the phenotype pH 3.5 adaptation (Fisher's exact test, *p*-value 0.001). Because of the small number of SNPs, most functions did not reach statistical significance. Nonetheless, among the uncorrected P-values, we observed many suggestive functions, such as glucoside transport for the phenotype, glucose 8% efficiency (Fisher's exact test, uncorrected *p*-value 0.003), and oligosaccharide metabolic process for the phenotype glucose 0.5% rate (Fisher's exact test, uncorrected *p*-value 0.003). Although

these functional results are preliminary, they suggest that more highly powered GWAS in *S. cerevisiae* may be able to elucidate important biochemical pathways.

Next we followed up on several associations previously detected by Cubillos *et al.* (80) and Warringer *et al.* (66). These researchers studied four broad phenotypes and their associated genes: copper tolerance, associated with CUP1/2; NaCl and LiCl tolerance, associated with ENA1/2/5; galactose growth, associated with GAL1/2/3; and maltose growth, associated with MAL31/32/33. To determine if the GWAS methods that we tested discovered the previously published associations, for each condition, gene, and GWAS method, we searched for all SNPs that were nominally significant at a *p*-value threshold of 0.05 in the vicinity of the relevant gene(s).

We computed how many relevant SNPs that each GWAS method found for all four reported associations and the percentage of SNPs found compared to the total number of nominally significant SNPs found by each algorithm (data not shown). The closest SNP to CUP1 and CUP2 (chr08:214751) was found by EMMAX-QK and LM-Q, while the next closest SNP (chr08:221695) was found by EMMAX-KLAQ and LM. For NaCl tolerance, four SNPs that fell within theENA1gene were discovered by several GWAS methods. Similarly for LiCl tolerance, four SNPs were discovered, three of which fell within ENA1, and one of which fell 1743 bp downstream of ENA5 (chr04:525679). The GWAS methods combined also discovered four SNPs associated with the maltose growth environments, all of which were located within MAL31. No significant SNPs were found in or near GAL1/2/3 by any of the GWAS methods only because there were few SNPs in our set near these genes (data not shown). We conclude that the GWAS methods are often able to recover previously known associations, but that different statistical methods may be needed in order to elucidate those associations.

2.3.4: EVOLUTIONARY ANALYSIS OF GWAS SNPS

It is also important to understand the nature of the evolutionary forces acting on SNPs that affect phenotypic variation. To address this issue, we considered the distributions of minor allele frequencies of the 3723 SNPs used in our analysis and all intergenic SNPs and compared them to the distribution of minor allele frequencies (MAF) of the SNPs that were found to be statistically significant by the GWAS methods. We did not attempt to root the SNPs to obtain derived allele frequencies, similar to a previous study (81). Overall, significantly associated SNPs were highly enriched for rare alleles compared to either the 3723 SNPs used in our analysis or all intergenic SNPs (Table 6, Figure 8). These results were robust whether we used an FWER or an FDR multiple testing correction method and whether we considered nonpleiotopic or pleiotropic SNPs. Note that in general GWAS methods have more statistical power for SNPs with higher MAF, so our tests were conservative because they showed that GWAS SNPs were nonetheless enriched in lower MAF. Also, the strain sampling and SNP selection procedures should not bias our result because all sets of SNPs should be equally affected.

Type of SNP	Multiple testing correction	*p*-value vs. GWAS	*p*-value vs. intergenic	Mean MAF
Nonpleiotropic	Holm	1.676e-15	0.0088	0.1329
	FDR	1.251e-17	4.975e-85	0.1227
Pleiotropic	Holm	0.0028	7.692e-07	0.0898
	FDR	1.005e-12	0.0125	0.1228

Table 6

Minor allele frequencies for nonpleiotropic and pleiotropic SNPs found by the BEST method. The BEST method refers to the best method of correction for each condition separately. The *p*-values are from one-sided Wilcoxon

tests. By comparison, the mean MAF for intergenic SNPs was 0.1684, and the mean MAF for GWAS SNPs was 0.1839.

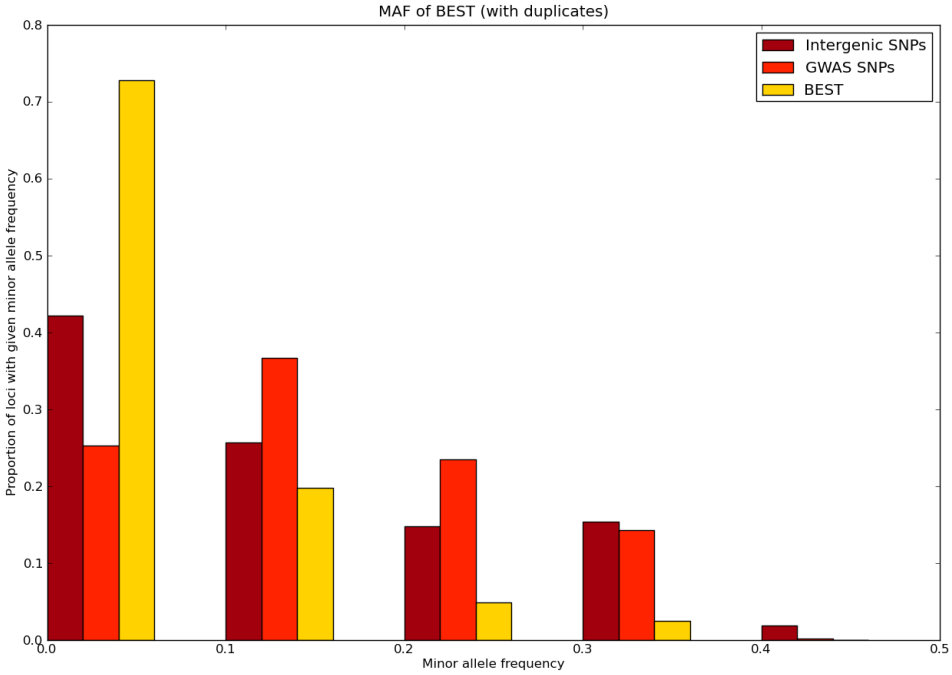


Figure 8

Comparison of minor allele frequencies between intergenic SNPs throughout the genome, SNPs used in our GWAS analysis, and SNPs selected as significant by the BEST meta-statistical method, after correction for multiple testing. We see a significant enrichment for rare alleles in the selected SNPs.

2.3.5: SIMULATION STUDIES REVEAL HIGH VARIANCE ASSOCIATED WITH LOCAL ANCESTRY ESTIMATION

In addition to our GWAS on real data, we performed two sets of simulations in which we fixed a certain number of "true" SNPs and compared the ability of each statistical method to find them under a wide variety of genetic architectures. The first set of simulations was based on an existing phenotype from the Warringer *et al.* data (66), while the second set was based on a simulated phenotype (see *2.2.8: Simulation studies*). We

measured the performance of each method using a receiver operator characteristic (ROC) plot.

Overall, we find that all methods perform better under simple genetic architectures, with the number of SNPs being the primary indicator of performance (see Figure 10). We see very little change in performance when changing the SNP effect size *V* while fixing the number of SNPs *N*. We find that over the average performance of 200 simulation iterations, EMMAX-K performs best, while EMMAX-KLA appears to perform similarly to the linear models with different ancestry covariates. To address why this is the case, we computed the variance of the ranks of the simulations. We observed that the methods based on local ancestry have higher variance than the other methods, which is an effect we also observed in our analysis of the original *S. cerevisiae* data (see Figure 9). While we find here that the mean squared distance of EMMAX-KLA is the lowest of all models tested, its range of values is greater than that of EMMAX. This is presumably because when randomly selecting SNPs to be causal SNPs, some of those were invariably correlated with local ancestry. This is a common issue with mixed linear models in GWAS as well, where inclusion of the candidate marker in the computation of the kinship matrix results in a loss of power of the study due to overfitting of the candidate marker (82). The example is easily extensible to overfitting by including fixed ancestry covariates related to the candidate marker.

By adding a local ancestry covariate, we were inadvertently correcting the effect of the SNP by the GWAS method. This raises an important larger point: until this point we were primarily focused on reducing the number of false positives, by selecting methods which minimize the MSD. However, reducing the false positive rate will necessarily also reduce the power the statistical approach. Nonetheless, we find that our methods have enough power to elucidate meaningful biological associations as previously published in Cubillos *et al.* (80) and Warringer *et al.* (66).

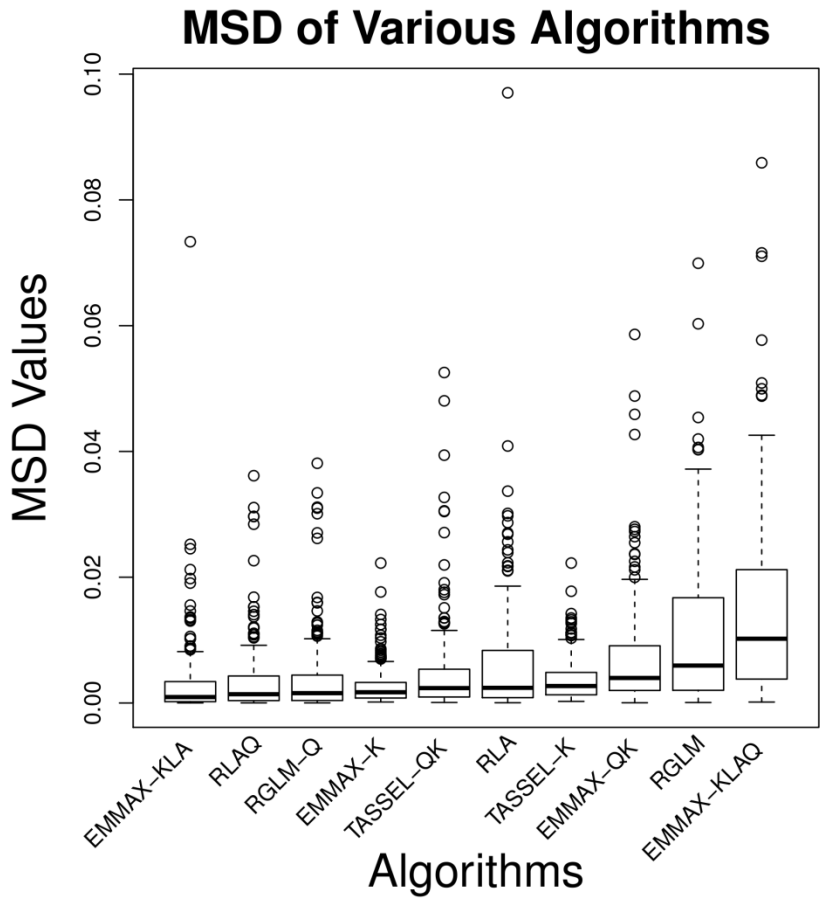


Figure 9

Boxplots of mean squared distance (MSD) across all phenotypes for each statistical method, ordered by lowest average MSD. While EMMAX-KLA has the lowest average MSD, the other methods including a local ancestry covariate show higher variance than methods not including a local ancestry covariate. In particular, if we compare EMMAX-KLA with EMMAX-K, we see that the MSD values from EMMAX-K have lower variance. This may be due to some loci being associated both with the phenotype and also with the population structure, which we also observed in the simulation studies. Further discussion in *2.3.5: Simulation studies reveal high variance associated with local ancestry estimation*.

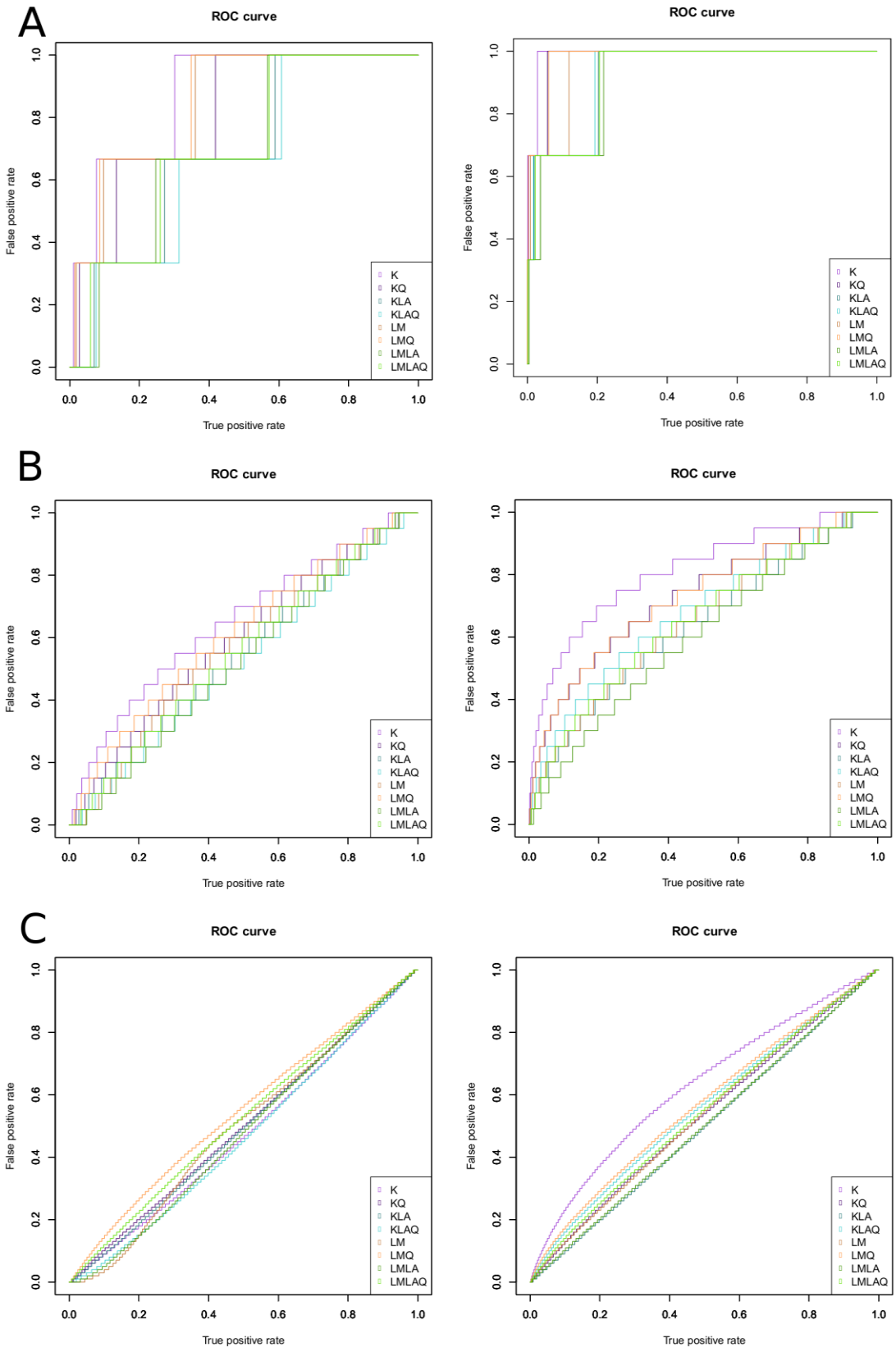

A
B
C
ROC curve
False positive rate
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
K
KQ
KLA
KLAQ
LM
LMQ
LMLA
LMLAQ

Figure 10

Receiver operator characteristic (ROC) curves for two sets of simulation studies. Left columns: First set of simulation studies, based on an existing phenotype. Right columns: Second set of simulation studies, based on a purely simulated phenotype and also using a larger set of genotypes (see *2.2.8: Simulation studies*). ROC curves are drawn for the average performance over 200 iterations. Figures A-C demonstrate the performance of each method under different numbers of simulated causal SNPs. In each figure, we hold the effect of the causal SNP fixed at $V = 3$. The number of causal SNPs in figures A-C are 3, 20, and 100, respectively. Overall, it is the number of SNPs determines the performance of the statistical methods. We also see that the purely simulated phenotype sees better performance, which may be due to two factors: First, the phenotype being less noisy; and second, the larger number of strains. In general, we see that EMMAX-K performs best on average. We explore this issue further in *2.3.5: Simulation studies reveal high variance associated with local ancestry estimation*.

2.4: DISCUSSION

GWAS have proven to be a highly effective way to map the genes underlying complex phenotypic traits in many species. In all applications of GWAS, it is crucial to control for underlying population structure, since it can cause spurious associations. Here we have performed an empirical study of statistical methods for correcting for population structure when performing GWAS in the important model organism, *S. cerevisiae*. Our main results are that GWAS is indeed a feasible approach in *S. cerevisiae* and that it is important to take into account the local ancestry of an *S. cerevisiae* strain when performing GWAS. At a practical level, the EMMAX mixed linear model implementation (28) using an identity-by-state kinship matrix as a random effect and local ancestry inferred by STRUCTURE (27) as a fixed effect performed best in our experiments. Importantly, our work also shows that existing methods for detecting local ancestry, such as STRUCTURE (27) and WINPOP (69), are effective in *S. cerevisiae*, at least for the purposes of GWAS. Nonetheless, the demographic history of *S.cerevisiae* is complex (65,68) and properly modeling it will probably require more specialized statistical methods than the methods designed for the

comparatively simpler cases of recent punctate admixture in human populations, particularly Latinos and African Americans (83).

There are many differences between performing GWAS in *S. cerevisiae* and humans. First, the burden of multiple hypothesis testing correction is much lower in *S. cerevisiae* because it has a much smaller genome size. Our analysis used only 3723 SNPs compared to the 500,000 typically used in human GWAS studies. If a simple Bonferroni-type correction is used, we would expect an *S. cerevisiae* GWAS to be far more powerful than a comparable GWAS in humans. Second, the extent of linkage disequilibrium is much less in *S. cerevisiae*, so GWAS in *S. cerevisiae* is more likely to pinpoint the actual causal variant than in humans, where it is more likely to find an association with a tag SNP. The *S. cerevisiae* genome is also much more gene-rich than the human genome, so each significant SNP is easier to link to a putative causal gene than in the human case. Third, since it is relatively cost effective to fully resequence *S. cerevisiae* genomes, we were able to use whole-genome resequencing data compared to the SNP genotyping chips still typically used in human GWAS studies (although continued decreases in sequencing cost may make whole-genome resequencing for humans feasible at some point in the future). Thus GWAS in *S. cerevisiae* can in principle test causal SNPs for association rather than tag SNPs. It has previously been shown that the power to detect association is much higher when testing the causal SNPs than when testing a tag SNP (84). Fourth, with *S. cerevisiae* it is possible to perform replicate phenotypic measurements to reduce the environmental noise. For all of these reasons, we believe that the power of GWAS in *S. cerevisiae* mitigates the relatively small sample sizes of individuals used in our study. We also note that it is possible to study many environmental conditions in *S. cerevisiae*, such as drug treatments, which would be impossible or unethical to do in humans.

S. cerevisiae is an important model organism for many aspects of molecular biology. Recent work on mapping complex traits in this species using recombinant inbred lines has yielded many important insights (85). In addition to its use as a model organism, *S. cerevisiae* is also an important agricultural species in its own right. Thus we hope that the statistical methods for GWAS investigated here will lead to further advances in our understanding of the genotype–phenotype map in this important species. Our comparisons to previous mapping results in *S. cerevisiae* (66,80) are promising in this regard. A recent study of GWAS in *S. cerevisiae* also found similar results to our study (86). In particular, they showed through simulations on the same set of *S. cereivisae* strains that GWAS in *S. serevisiae* is generally difficult because of the complex population structure but is feasible for Mendelian trait and cis QTL mapping. One difference is that Connelly and Akey (86) stressed the difficulties of GWAS in *S. cerevisiae* whereas we have stressed the relative utility of using local ancestry corrections in *S. cerevisiae* GWAS, while continuing to acknowledge the overall difficulty of using GWAS methods in this species. Nonetheless, our improved GWAS performance on the larger set of *S. cerevisiae* strains from Schacherer et al. (68) suggests that increased sampling and sequencing of strains will improve GWAS results in the future. Such studies will be facilitated by the small size of the *S. cerevisiae* genome (12 Mb), the decreasing cost of DNA sequencing, and the relative tractability of high-throughput phenotyping in yeast (87).

In addition, there are many other studies of GWAS in other model organisms that are similar to our study, including studies in mice (88), Arabidopsis (75), maize and rice (89), tomato (90), dog (91), and *Drosophila melanogaster* (92). Recent admixture is a pervasive phenomenon in many species. For example, there is strong evidence of non-African admixture in the DPGP *D. melanogaster* lines from Africa (J. Pool, unpublished results). GWAS in admixed human populations is also an important current research

problem, and a very interesting goal for the future will be to combine admixture mapping with association mapping (93,94). Thus we believe that our results will also be useful for GWAS analyses in humans and other model systems as well.

CHAPTER 3: MIRNA MOTIF DISCOVERY USING MIXED LINEAR MODELS

3.1: INTRODUCTION

Since their discovery, small non-coding RNAs have proven to be an important and pervasive mechanism for gene regulation. MicroRNAs, or miRNAs, are a special class of these small RNAs that specifically arise from small hairpin structures, and were first observed nearly two decades ago. In 2002, a miRNA database was established, called miRBase, containing just 218 entries. Since that time, the sheer volume of data has increased exponentially every year, and the database now contains sequence information for tens of thousands of mature miRNAs, in 193 species, including *C. elegans*, *D. melanogaster*, human, and mouse (95). In spite of the wealth of data, the exact mechanism by which miRNAs suppress gene expression has not been entirely elucidated, and their purpose is not entirely understood. One key issue now is to determine the potential targets of known miRNAs, which could help us understand the extent and nature of their function; another is de novo discovery of the miRNAs themselves. Here we will see that in fact these two problems are closely related.

Specifically, we propose a new model for small RNA motif discovery employing mixed linear models. As inspiration we use the problem of population stratification in genome-wide association studies. In such studies, one of the primary hurdles is the presence of relatedness among individuals that can lead to inflated test statistics for markers that are in fact not linked to the phenotype being assessed. It has repeatedly been shown that mixed linear models are effective in correcting for this type of confounding factor (28,96), and we make use of this fact by drawing an analogy whereby "SNPs", or

single nucleotide polymorphisms, in a GWAS are represented by potential small RNA motifs in the 3' UTR of a gene, and "population stratification" is translated to be the relatedness between mRNAs—for example, this could be taken to be the sequential relatedness between complete 3' UTR sequences—thus capturing any background effects that could influence the efficacy of a miRNA but that are not readily observable. We first briefly describe our previous work with mixed linear models and GWAS, then follow up with our small RNA motif discovery proposal.

3.2: DATA

3.2.1: MOUSE CD4+ DICER KO EXPRESSION PROFILES

CD4+CD25- Conventional T cell mRNA expression profiles

Mice carrying a floxed Dicer allele in combination with CD4Cre transgene on a mixed C57BL/129 background (97) were maintained under specific pathogen-free conditions. Peripheral CD4+CD25- T cells were sorted on a FACS ARIA (Becton Dickinson) from 6-8 week-old mice and RNA extracted using RNAbee (AMSBio) according to the manufacturer' instructions. 100 nanograms of RNA was used to interrogate the GeneChip Mouse Gene 1.0 ST Array (Affymetrix). We obtained log fold changes in gene expression for 24,601 mRNA transcripts between WT and Dicer KO mouse CD4+ CD25- T cells. These experiments were performed by Antoine Marcais.

Conventional T cell miRNA expression profiles

We obtained two data sets of miRNA expression for CD4+CD25- T cells from independent sources using different technologies.

First, from the same cells from which we obtained our mRNA microarray expression data, we also obtained comparative miRNA expression data between CD4+CD25+ T-cells and CD4+25- T-cells from Cobb et al. (22), who studied the differences in miRNA expression profiles for the two types of cells. The authors performed a miRNA microarray analysis with probes for 173 miRNAs from miRBase. Of these, we take the top 20 differentially expressed to be true, "active" miRNAs, as reported by the authors in Figure 2.

To corroborate these results, we also used miRNA expression data from C57BL/6 mice determined by the nCounter miRNA expression assay kit (Nanostring Technologies), from Sommers et al. (24). The authors validated the Nanostring nCounter expression results with Exiqon microarrays and Taqman qRT-PCR assays. 92 probes corresponding to 86 miRNAs in miRBase were evaluated. Of these, we took the top 21 highly expressed for experimental validation of our predictions, corresponding to the most highly expressed miRNAs presented by the authors in Figure S1.

3.2.2: Mouse adrenal cortex Dicer KO mRNA and miRNA expression profiles

We obtained mRNA and miRNA expression data for Dicer KO adrenal cortex tissue from mouse embryos at stages E15.5 and E16.5 from a study by Krill et al. (98). Sf1-Cre mice were crossed with mice carrying a floxed Dicer allele to produce Sf1-Cre/Dicerlox/lox mice. Embryos were harvested at E15.5 and E16.5 and the adrenals from each were collected. A total of 4 control and 4 Dicer KO biological replicates were obtained for each time point. Affymetrix Mouse 430 v2.0 gene expression arrays were used for hybridization.

Krill et al. also report miRNA expression in adrenal cortex at developmental stages E15.5 and E16.5. ABI miRNA OpenArray was used for miRNA expression analysis. These arrays are able to target 750 miRNAs. We took the top 25 highly expressed miRNAs for E15.5, the top 35 highly expressed miRNAs for E16.5, and the top 15 miRNAs which are

found highly expressed at both time points, in line with Figure 4 and Table 1 in Krill et al. (98).

3.2.3: MOUSE EMBRYONIC STEM CELL DICER KO EXPRESSION PROFILES

The embryonic stem (ES) cells were derived and described in Nesterova et al. (99). ES cell lines were maintained on a feeder layer (mitomycin-inactivated primary mouse embryonic fibroblasts) in Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% fetal calf serum (FCS, Autogen Bioclear), 7% Knockout Serum Replacement (KSR), 2 mM L-glutamine, 1× non-essential amino acids, 50 μM 2-mercaptoethanol, 50 μg/ml penicillin/streptomycin (all from Invitrogen) and LIF-conditioned medium, made in house, at a concentration equivalent to 1000 U/ml. Cells were grown at 37°C in a humid atmosphere with 5% CO2. Affymetrix GeneChip Mouse Gene 1.0 ST Arrays were used to perform the microarray. These experiments were performed by Antoine Marcais.

3.2.4: HELA TRANSFECTION EXPRESSION PROFILES

We obtained two types of data for five miRNA transfection experiments in human HeLa cells: microarray and proteomics from Selbach et al. (100). The authors performed transfections by synthetic miRNAs and mock transfections in human HeLa cells for let-7b, miR-1, miR-155, miR-16, and miR-30a. The amount of protein synthesis was given by the log of the ratio of protein synthesized in the miRNA transfected cells divided by the mock transfection between 8hrs and 32hrs post transfection. Microarray analyses were performed with the Affymetrix Human Genome U133 Plus 2.0 chip. We used the microarray log fold change values taken at both 8hrs and 32hrs post transfection for each miRNA transfection experiment.

In this paper, the authors developed the pSILAC (standing for "pulsed" SILAC) method for measuring the change in protein production between two different samples. Reported proteomics data was obtained using the pSILAC method. We mapped the pSILAC Uniprot protein IDs to Refseq transcript IDs by downloading an ID mapping table from the Uniprot website. For the different transfection experiments, there were slightly different numbers of proteins with expression values, resulting in a range of the number of protein expression data points with corresponding 3' UTR sequences from ~3000 - 3600 across all the transfection experiments.

3.2.5: 3' UTR SEQUENCE DATA FOR MOUSE AND HUMAN

The 3' UTR sequences for mouse and human RefSeq gene mRNAs were both downloaded from the UCSC Genome Browser. In total we downloaded 26,845 sequences for mouse and 40,571 sequences for human, versions mm10 and hg19, respectively (101,102). In the case of transcript variants, we retained only the longest transcript. Furthermore, we removed all UTRs of length 10 or lesser. We are able to associate 17,988 unique mouse UTR sequences to their microarray expression values for the mouse Tconv Dicer KO dataset, and 22,266 unique human UTR sequences to their microarray expression values for each of the Selbach *et al.* miRNA transfection experiments. The number of transcripts able to be associated to the Selbach pSILAC data is ~4k, varying slightly depending on the transfection.

3.2.6: MIRNA MOTIF DATABASE: MIRBASE

We downloaded 1,908 mature mouse miRNA sequences corresponding to ~1200 distinct 6mer seeds and 2,578 mature human miRNA sequences corresponding to ~1500 distinct 6mer seeds from the miRBase database (release 20) (95).

3.3: METHODS

3.3.1: CURRENT METHODS IN miRNA MOTIF DISCOVERY

There are several existing methods for miRNA motif discovery that we compared our methods to. The foremost of these is miReduce (103), based on the Reduce algorithm (104), which is essentially a forward stepwise regression. More recently, two other algorithms were also published, Sylamer (105) and cWords (106), which have both implemented ways to correct for background sequence composition. In all of these methods, we take as the dependent variable the set of log fold changes in expression between two experiments, in particular a Dicer knockout (KO) experiment versus a wild-type. The main idea in each of these algorithms is to correlate motif presence in the 3' UTRs of the genes with their change in expression. We would expect, for example, that in cells in which Dicer has been knocked out, that miRNAs would no longer work properly, thus leading to an overall increase in gene expression. Here we give a brief overview of these three algorithms and how they differ from each other and from MixMir (2).

miReduce

miReduce implement a forward stepwise regression, which adds one factor at a time into a linear model as long as the factor still contributes significantly to the model. The basic model can be written as

$$y_i = \beta_0 + \sum_{j \in M} \beta_j x_{ij}$$

where y_i is the log fold change in expression for gene *i*, β_0 is some baseline change in expression, β_j is the effect of the presence of motif *j*, and x_{ij} is the presence or absence of motif *j* in the 3' UTR of gene *i*. Here we let *M* be the set of significant motifs.

We start by assuming that $M = \{\ \}$, the empty set. Then we iteratively select the motif which minimizes the error when fitted to the model above. If found significant, this motif, m_1, is then added to the set *M*, and the vector of errors from fitting m_1 is subtracted from $\overline{y}$, the vector of all expressions. This procedure is then repeated until no motifs are found significance. The measure of statistical significance is determined using the extreme value distribution, which describes the probability that the largest of *M* samples from a normal distribution (104).

miReduce is a simple and fast method that performs remarkably well in some cases, for example in the miRNA transfection data we analyzed (*3.2.4: HeLa transfection expression profiles*). However, the method is purely based on motif presence/absence, and as described in *1.2.2: miRNA target prediction and motif discovery*, there are other sequence-based factors affecting miRNA binding, in particular the sequence composition around the binding site. To that end, two other methods have been published which address this very issue.

Sylamer

Sylamer (105) was published in 2008 and approaches the problem differently. Instead of using the actual expression values, for example, Sylamer ranks the gene list by change in expression and uses this ranking to select for overexpressed motifs in the top *N* elements of the list.

We ran Sylamer using the web-based implementation Sylarray (107). We did this instead of running the command line version of Sylamer primarily because we noticed that in order for Sylamer to perform on par with the other methods investigated, some preprocessing needed to be performed, namely that sequence "purging" needed to be done using a third party program such as RSAT (108). The "purging" function of RSAT helps to remove redundancies in the sequence, such as repetitive monomers and dimers, which can result in many false positives, and biases towards AU rich motifs.

Simply speaking, the Sylamer algorithm first ranks the genes in a list by decreasing expression value. It then considers as the first "bin" the top N genes, and seeks to determine whether a motif in question is over- or under-expressed in the top N genes compared to the remainder of the list, determined by a hypergeometric distribution. After the first bin, Sylamer then proceeds in increments of N, considering over- and under- representation of a motif in the top kN genes at step k. The significance of each motif in each bin is then log transformed if the motif is overrepresented, and negative log transformed if underrepresented, resulting in a table of m motifs by k bins.

Due to computational constraints, Sylarray can be run either on the set of all words (by default, 6mers, 7mers, and 8mers), or only on a subset of words which correspond to known miRNAs in miRBase. If run on a the set of all words, Sylarray then returns the motif enrichment table above for all motifs satisfying a p-value cutoff of $p < 0.01$. This table can be interpreted visually by simply plotting the enrichment table in the natural way, with log p-values along the y axis for each word, for each bin along the x axis. The interpretation of this enrichment plot is simple if there are a few words which are highly over- or under-expressed, as these will result in a steep incline (similarly, decline) for the plot of a particular motif, either at the start of the plot, indicating a strong effect in the first few bins, or along the entire length of the plot, indicating an effect spanning many genes. The

Sylarray output is in the form of a java file which can draw such a plot and also show the top three over- and under-expressed words and their corresponding miRNAs.

Additionally, the Sylamer algorithm attempts to correct for composition biases by replacing the word counts in a particular bin by their expected word counts, as determined by the composition bias of that particular bin. It does this via a higher order Markov model, conditioning the expected word count of a particular word on shorter words within the bin. The authors note that as the bin size grows, the expected word counts computed in this manner will deviate further and further from the true word counts, so for our analyses we use the default bin size of $N = 200$, which should not see this issue.

We run Sylarray allowing sequence purging, using all words and not just those corresponding to miRNAs, and with a bin size of $N = 200$, for all of our analyses. When we create our final ranking of motifs using the motif enrichment table, we take as the p-value for each motif the lowest p-value across all bins. This recapitulates the results returned by the Sylarray java-based graphing program, and also produces the expected p-value distribution for all motifs with $p < 0.01$.

cWords

The most recent of the algorithms we compared against was cWords, published in 2013 (106). Like Sylamer, it takes as input a ranked list of genes according to degree of expression change. However, the way in which cWords determines the significance of a particular word somewhat differs from Sylamer: Given G ranked genes in an analysis and a particular motif m, cWords computes a probability p_i for $1 \leq i \leq G$, for the probability of observing the motif m in gene i. This probability is based on a Markov model like before, which determines the probability of observing a particular word conditioned on the probability of observing shorter words of length k. Specifically,

$$P_k(W) = \mu(w_1 \cdots w_k) \prod_{i=1}^{l_w - k} \pi(w_{i+k} | w_i \cdots w_{i+k-2}, w_{i+k-1}),$$

where $P_k(W)$ is the probability of observing word *W* conditioned on shorter words of length *k*, $\mu(w_1 \cdots w_k)$ is frequency of the first *k* letters of the word *W*, and l_w is the length of *W*. Then the probability of seeing a word appear *m* times or more is given by a binomial distribution:

$$P(q \geq m | n, p) = \sum_{i=m}^{n} \binom{n}{i} p^i (1-p)^{n-i},$$

where $p = P_k(W)$ from above.

The algorithm computes such a *p*-value for each word, for all genes. Thus, for each word we obtain a list of *p*-values as described above, ranked in the same order as the genes. The authors then compute what they call a "running sum" of the log transformed *p*-values, determining the significance of its deviation from random by comparing it to the expected distribution of the maximum running sum. Additional details can be found in (106).

3.3.2: Application of mixed linear models to motif discovery: MixMir

Similar to cWords and Sylamer, we are interested in how correcting for background sequence composition can improve miRNA motif predictions. Unlike the previous methods, however, MixMir employs a pairwise method of correction, which considers similarities in expression change between two transcripts as well as similarities between their 3' UTR *k*mer composition.

To do this, we borrow the mixed linear model (MLM) as it has been applied extensively in GWAS. In that model, we had the following equation:

$$y_i = \mu + x_{ij}\beta_j + \alpha_i + \varepsilon_i,$$

with y_i the phenotype of individual and x_{ij} representing either the presence or absence of SNP *j* in individual *i*, and α_i is a random effect. In GWAS, this random effect factors into the model when we look at the decomposition of the variance of y_i into its components.

In the motif discovery problem, the independent variable is the presence of a particular motif in a sequence instead of a SNP. In our specific application, we are interested in discovering miRNAs, and so x_{ij} represents the presence of miRNA motif *j* in gene *i*, and y_i represents the change in gene expression between two data sets, for example a Dicer knockout and a wild-type cell. Further, we are considering as the random effect the sequence similarities between two 3' UTRs, which may affect miRNA binding efficacy. For example, we know that a high AU content in the region flanking the miRNA binding motif increases miRNA binding efficacy. However, other unknown specific sequence composition rules may affect also affect miRNA binding, and we wish to correct for these even without knowing what they are. The parallel which we wish to draw in this case is the measure of kinship in GWAS and *k*mer content similarity in motif discovery: That is, we define a measure of similarity between 3' UTRs which parallels the definition of the identity-by-state (IBS) kinship matrix we saw previously.

GWAS	**miRNA**	**Model term**
Strain "phenotype"	mRNA expression	y_i
Presence of SNPs	Presence of miRNA motifs	x_{ij}
IBS kinship matrix	3' UTR similarity	K

Table 7

Comparison of components of mixed linear model between GWAS and MixMir.

To this end, we let $K_{ij} = cor(kmer_i, kmer_j)$, where the $kmer_i$ is the vector of *k*mer counts for UTR *i* scaled by UTR length, and *cor* is the Pearson correlation of the two vectors. Thus, *K* represents a measure of similarity, or "kinship", between two transcripts.

We used both GEMMA v0.98 (62) and FaST-LMM v2.07 (109) to solve the MLM described above. These two algorithms were developed after EMMAX and differ primarily in the fact that they are "exact" and do not make the simplifying assumptions that EMMAX does, namely that the effect of each SNP/motif is small. In the case of miRNAs, this assumption is not likely to be valid, so we opted to use an exact method. We began by using GEMMA, as it was the most recent version of an exact MLM solver made tractable for large data sets as those used in GWAS. However, we ran into some issues with GEMMA with some of our analyses, as in some cases it would fail to estimate coefficients and *p*-values. Thus, we provide a short description of both FaST-LMM and GEMMA here.

FaST-LMM is an “exact” method, meaning it yields exact test statistics, unlike EMMAX, which makes the assumption that the variance parameters for all SNPs are the same (see *2.2.6: Current methods in GWAS*). FaST-LMM implements two main innovations: First, the authors note that the restricted maximum likelihood (REML) of a mixed linear model can be written in such a way as to depend on only a single variable (δ); and second, that the data (SNPs, phenotypes, covariates) can be transformed to be uncorrelated using the spectral decomposition of the relationship matrix, so that after transformation only a linear regression is necessary to obtain identical results (109). Thus, FaST-LMM avoids the problem of having to re-estimate the variance parameters for every SNP, while still avoiding having to assume that these parameters are the same across all SNPs (as EMMAX does).

Though theoretically their computational complexities are the same (62), GEMMA uses the Newton-Raphson optimization method, whereas FaST-LMM uses Brent's algorithm (62). In their own analyses, Zhou and Stephens found GEMMA to be much faster than FaST-

LMM on a smaller data set (~700 individuals, 12x speedup), while only twice as fast on a larger data set (~4600 individuals) (62). In our analyses, on the other hand, we found that FaST-LMM outperformed GEMMA, and also that GEMMA often produced *NaN*s, which may have to do with the implementation of the algorithm rather than the algorithm itself. In all cases where GEMMA and FaST-LMM produced sensical results, the results were by and large the same (in terms of p-values and estimation of the size of the fixed effect). Our sample sizes were, however, much larger than those tested in the GEMMA paper, with approximately $n = 18{,}000$ genes ("individuals") and $m = 4096$ ("markers"), so it is possible that at larger values of *n*, GEMMA is not as efficient.

Since both GEMMA and FaST-LMM were built for GWAS, they accept only categorical inputs, and so we compare MixMir against a linear model with binary presence/absence variables. This categorical linear model, which we call "LM Bin", is simply the expression levels of the miRNAs regressed against the presence or absence of a particular motif.

3.3.3: PROPER ESTIMATION AND USAGE OF P-VALUES

Throughout the following analyses, we assume that the p-values returned by the mixed linear model and cWords are properly estimated, as we use these p-values in part to perform a degree of model selection. To verify that this is the case, we performed 20 randomizations of the Tconv sequence and expression data, and ran both MixMir and cWords on the randomized data for k = 2..6. For MixMir, the p-values obtained from the randomized data exactly follow the diagonal line expected under the null hypothesis, suggesting that our p-values are indeed properly estimated.

However, we did not find this to be the case for cWords in the Tconv data in particular, where the p-values were highly skewed towards being very large, i.e. insignificant (see Supplementary Figure 1). Further, we find in *3.4.1*: Parameter testing for

MixMir and cWords on the Tconv data that for the true data, the p-values are very skewed towards being very small, i.e. significant (Supplementary Figure 2). These two factors combined suggest that p-values found by cWords are not in fact properly estimated, at least for the mouse Dicer KO Tconv data. We also performed this test for the ES cell data; however, in that case we found that randomized data resulted in an expected distribution of p-values for both MixMir and cWords, so the effect appears to be dataset dependent.

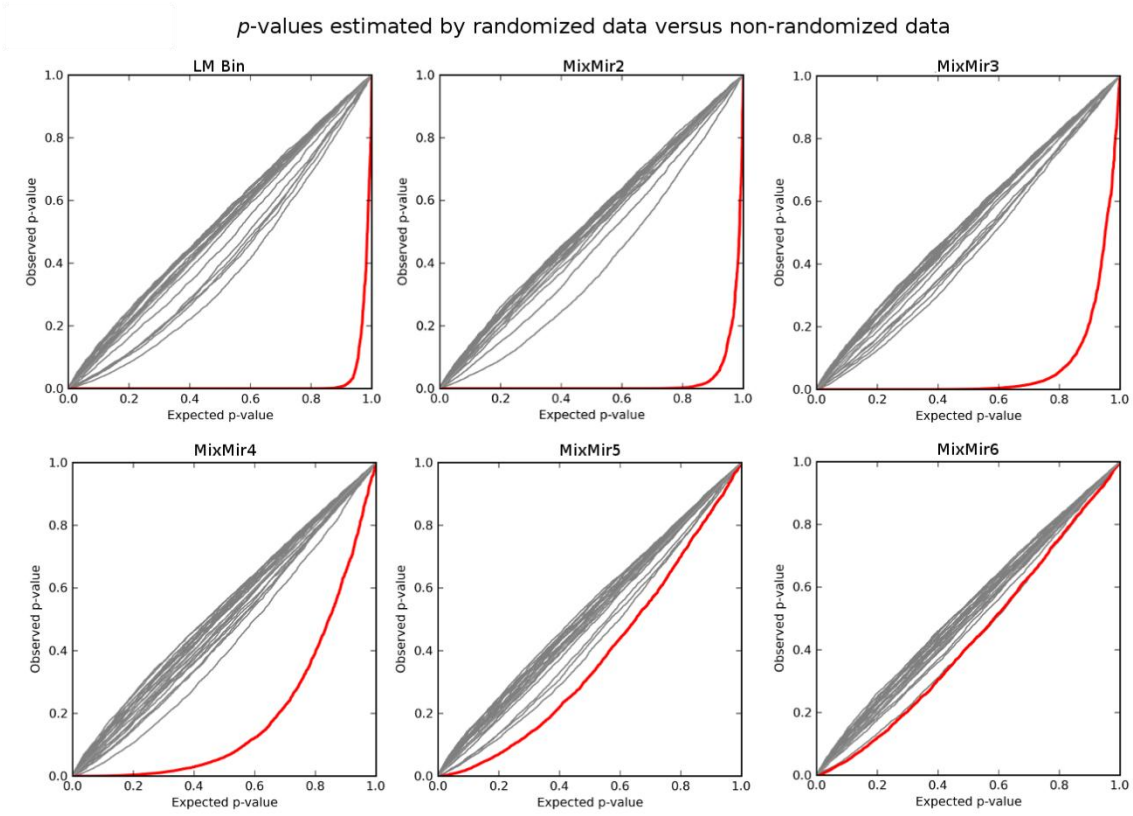


Figure 11

Plots of observed versus expected *p*-values obtained by the linear model and MixMir for different values of *k*, for randomized (grey lines) and non-randomized (red lines) data. From these plots we see that the observed *p*-values estimated with randomized data closely follow the expected, whereas the observed *p*-values with non-randomized data results in a large number that are much smaller than expected, suggesting a large number of false positives. Notably, this is not the pattern observed in cWords, which at least for the Dicer KO Tconv data produces much higher *p*-values than expected given randomized data (i.e. close to 1).

3.4: RESULTS

3.4.1: PARAMETER TESTING FOR MIXMIR AND CWORDS ON THE TCONV DATA

MixMir uses a linear model of miRNA targeting, similar to previous models such as miReduce, but adds a similarity matrix that corrects for background sequence composition (see *3.3.2: Application of mixed linear models to motif discovery: MixMir*). This similarity matrix is calculated by taking the correlation of a vector of *k*mer counts pairwise between 3' UTRs. Since this similarity matrix can be strongly affected by our choice of *k*, we initially tested MixMir with different similarity matrices computed using $k = 2..6$. For values of *k* greater than 6, the similarity matrices became inaccurate due to the limited total amount of 3' UTR sequence in the genome and the running time of the implementation was slow, so we did not consider higher values of k further.

We performed model selection by comparing PP plots as we used in *Chapter 2: Genome-wide association studies in highly structured populations*. The results of these plots, compared alongside the same plot for the linear model, are displayed in Figure 12. As before, the observed *p*-values should follow a uniform distribution under the null hypothesis, and a large deviation is suggestive of the presence of many false positives.

We found that as we increased *k*, the observed p-values obtained from MixMir approached the expected p-values. Furthermore, we found that as *k* increased, MixMir became less similar to the baseline linear model, where similarity is defined by the Pearson correlation of the *p*-values of the motifs tested. These patterns were similar if we computed the Pearson correlation of motif ranks instead of motif *p*-values (Table 8). Combined, these results suggest that the MixMir model which performed the most correction of the *p*-values was MixMir6. Note that this analysis implicitly assumes that there are relatively few highly active miRNAs in any particular cell type compared to the total number of possible miRNA

sequences (in this case 4096 hexamers), an assumption we believe to be generally true biologically (110). Therefore, we selected MixMir6 to represent the mixed linear model results in comparison with the other methods in our analysis.

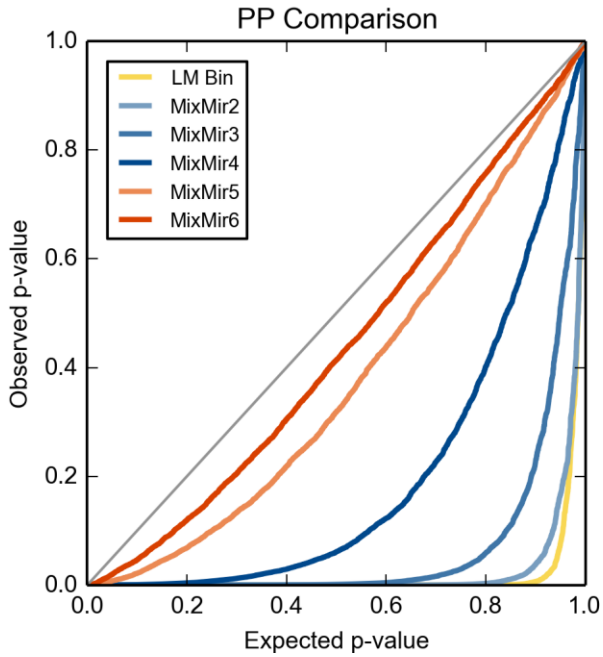


Figure 12

Percentile-percentile plot comparing MixMir with $k = 2 .. 6$. Expected p-values are found on the x-axis; observed p-values are found on the y-axis. The yellow line is the PP plot for the linear model, which displays an extremely skewed distribution towards many very low p-values. As we increase the length of background words used for correction, i.e. k, we see the observed p-values approach the expected p-values.

Method	LM Bin	MixMir2	MixMir3	MixMir4	MixMir5
MixMir2	0.8876				
MixMir3	0.8101	0.9677			
MixMir4	0.6336	0.8397	0.9363		
MixMir5	0.3407	0.5356	0.6517	0.8012	
MixMir6	0.2293	0.3624	0.4428	0.5643	0.8832

Table 8

Comparison of MixMir result similarities with the simple linear model (LM Bin). Pairwise Pearson correlation of all motif ranks. We saw that the degree of rank similarity between the LM and MixMir results varied directly with the length of the *k*mer used to construct the relationship matrix.

The plots above assume that the *p*-values returned by the mixed linear model are properly estimated. To verify this, we performed 20 randomizations of the sequence and expression data, and ran MixMir on the randomized data for $k = 2..6$. *p*-values obtained from the randomized data exactly followed the diagonal line expected under the null hypothesis line, suggesting that our *p*-values are indeed properly estimated.

We initially tested $k = 2$ to 6 for cWords to perform model selection as we did with MixMir, which we refer to as cWords2 to cWords6. The authors recommended setting $k = 1$ to 3 for cWords, presumably because of the limited amount of sequence in 3' UTRs (*3.3.1: Current methods in miRNA motif discovery*). The resulting PP plots showed that there was a significant discrepancy between observed and expected *p*-values, similar to the simple linear models, suggesting a relatively high false positive rate for cWords on this data set, or perhaps that *p*-values are incorrectly estimated. To first test this possibility, we also performed 20 randomizations of the data as above for use with cWords. We found that the randomized data resulted in observed *p*-values much higher (i.e. closer to 1) than their expected values, which suggested that perhaps in this data set, *p*-values are not properly estimated (Supplementary Figure 1).

We observed that little improvement was gained by using any *k*mer background correction as judged by PP plots (Supplementary Figure 2), so we did not use this as a criterion for model selection. We tested the similarity of the results of the different cWords models with the results of the linear model and found that $k = 2$ gave results least similar to

those of the linear model, where similarity is defined by the ranks of the motifs tested (Table 9).

Additionally, when we examined the prediction performance on the T conv data set, we saw a large drop in performance for *k*=5 and *k*=6, with many fewer matches to miRBase miRNAs and T conv cell highly expressed miRNAs than for *k*=2, 3, and 4. This is entirely consistent with the authors' recommendation and our observation above that there is insufficient 3' UTR sequence data to train higher orders of the Markov model. Thus, for further analyses, we retained just cWords2 as representative of the algorithm.

Method	LM Bin	cWords2	cWords3	cWords4	cWords5
cWords2	0.7978				
cWords3	0.8611	0.9335			
cWords4	0.9287	0.8682	0.9312		
cWords5	0.9616	0.8476	0.9020	0.9610	
cWords6	0.9616	0.8475	0.9019	0.9609	1.000

Table 9

Comparison of cWords result similarities against simple linear model. Pairwise Pearson correlation of all motif ranks. Here we saw an opposite effect of what we observed with MixMir (Table 8): as we increase *k* in cWords, the results became closer to those of the linear models, with cWords2 and cWords3 producing the most different results. cWords5 and cWords6 were nearly identical in motif ranking.

3.4.2: MixMir outperforms current methods in discovering miRNA motifs found in miRBase in the Tconv data

To compare MixMir against the previous motif discovery methods, we tested a total of five models: the simple linear model based on motif presence/absence (LM Bin), which we take as our baseline method, miReduce, cWords2, and Sylamer, and MixMir6. For the linear models, all possible motifs were ranked by p-value; for miReduce, we set the p-value cutoff to be 0.5, resulting in 57 motifs returned (see Methods for a discussion of this choice of p-value cutoff). Sylarray returned 885 words with p-value < 0.01, so these were ranked according to p-value. The motifs in the cWords results are ranked according to a Z-score (106), which we found were not consistent with the p-value, so we retained the original Z-score ranking, which produced better results.

We compared the significant hexamer motifs found by each method to miRNAs in miRBase (*3.2.6*: miRNA motif database: miRBase). We performed two matching procedures to the miRNAs. First, in our stringent matching criterion, we considered a hexamer a match to a particular miRNA only if it matches the seed sequence of a mature miRNA. Second, in our relaxed matching criterion, we allowed the hexamers to match to any of three positions starting at nucleotides 1, 2, or 3 from the 5' end of the mature miRNA. We included offset match positions 1 and 3 in order to include all possible types of marginal binding site matches (48), including the potential for extensive complementarity through nts 1-8. This relaxed criterion also allows for shifts in the discovered motifs, which are common in practical applications of motif-finding algorithms to biological data. In general we expect to see more false positives when including matches to offset seed sequences, so for all comparisons we considered both the results from the stringent and the relaxed matching criterion.

As described in *1.2.2: miRNA target prediction and motif discovery*, there are additional types of seed sequence matching, not limited to exact seed only and offset seed sequences. Namely, another common type of site is the A1 site, in which there is an

additional A paired to the first nucleotide of the 5' end of the miRNA as well as matching to nts 2-6 of the mature miRNA (48). However, we found that including the A1 site resulted in many matches to low-confidence miRNAs (e.g. poorly conserved and having very low expression) for the categorical linear model, while influencing the results of other models very little. Therefore we chose to omit the A1 site matches from our analyses (data not shown).

We present results for the two matching criteria using truncated receiver operating characteristic (ROC) curves and analyze the results by computing an area-under-the-curve (AUC) value for each curve (Figure 13). Briefly, we constructed the ROC curves by taking the top 20 and 50 ranked motifs of each method, with true positives taken to be matches to any miRNA in miRBase (see Supplementary Note for details). These truncated ROC curves are exactly a close-up of the bottom left hand corner of an ROC curve over all possible results. We chose to truncate the full ROC curve, which is typically constructed over all possible 6-mer motifs, both because the methods did not return the same number of predictions and most importantly because we believe that focusing attention on only the top motifs is a more biologically meaningful comparison since only a few motifs are likely to be biologically relevant (i.e. only a small fraction of all possible miRNAs in the database are actually expressed in a cell (110)). It is important both that truncating the ROC curve does not change the ranking of the methods and that we believe our results are robust in that the ROC curves for MixMir dominate the other curves over essentially the entire range of sensitivity settings (Figure 13). We caution that the truncated AUC value should not be interpreted as a typical AUC with a baseline value of 0.5 for a random method. Instead, we plot in our curves a baseline expected value for a random predictor given the number of motifs being plotted and the number of possible true positives, to which the other AUC values may be compared.

We present both truncated ROC curves for relaxed motif matching as well as for stringent motif matching (Figure 13). We found that the AUC values for the simple linear model was low and it found fewer miRNAs than miReduce, Sylamer, cWords2, and MixMir6. cWords2, Sylamer, and miReduce were comparable in performance in both window sizes of N = 20 and N = 50 foremost motifs. All four of those methods performed worse than MixMir6, which was more accurate over almost the entire range of sensitivity values. This effect was more noticeable when we used the strict motif matching criterion to position 2 only in the top 50 motifs. These results suggest that MixMir more accurately identifies motifs corresponding to the exact miRNA seed region. The top 50 motifs and whether they match to miRNAs in miRBase are given in Supplementary Table 1.

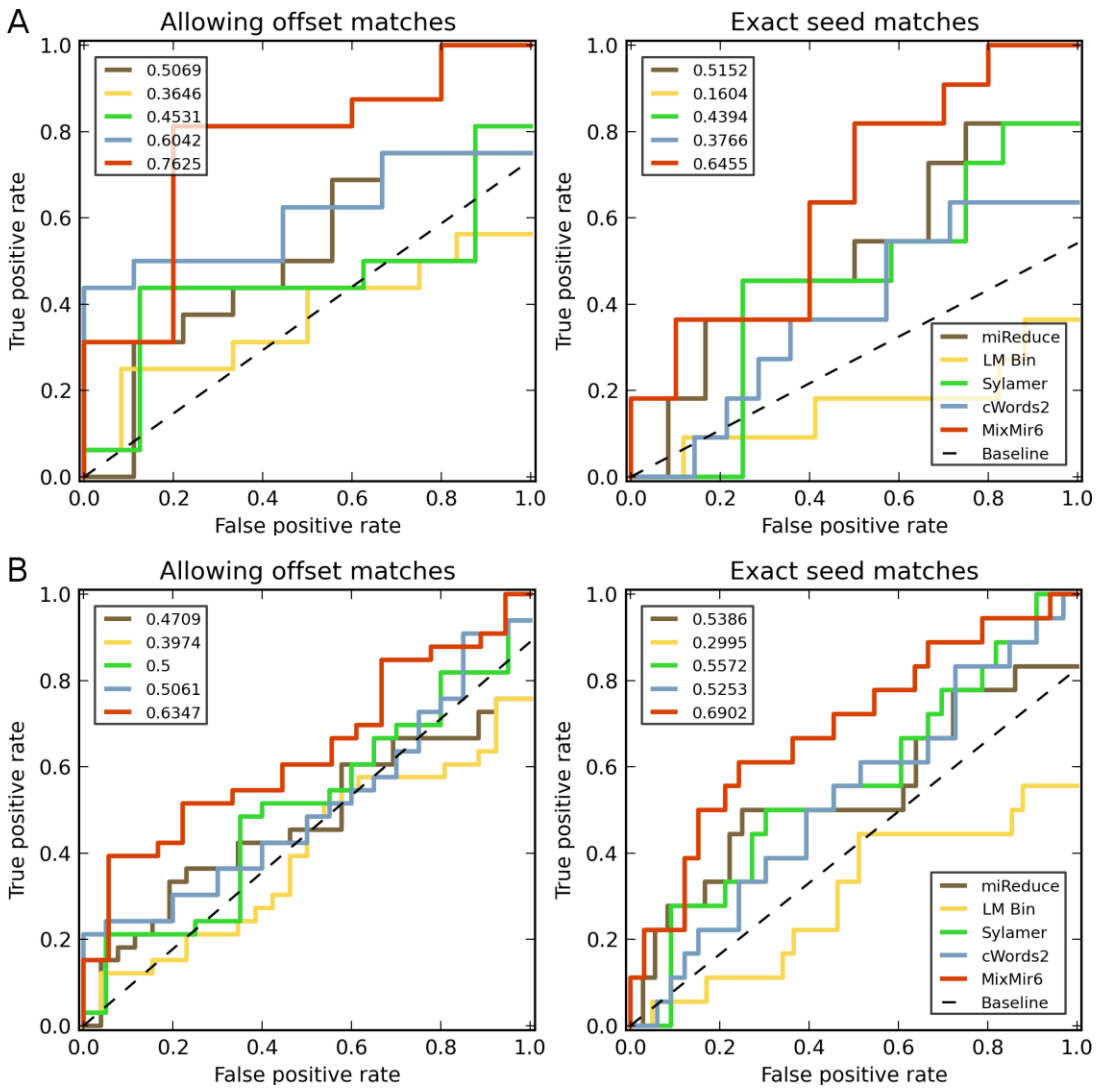


Figure 13

Truncated ROC curves for A) the most highly ranked 20 motifs by each method and B) the most highly ranked 50 motifs by each method, where true positives are taken to be motifs matching to miRNAs in miRBase.

3.4.3: EXPERIMENTAL VALIDATION

Using miRBase as a standard for "true" miRNAs acting in our data is still likely to result in false positives, however, since the database consists of experimentally verified and computationally predicted miRNAs, not all of which may be active in all cell types

(111). Thus, comparing the predicted motifs to all miRNAs in miRBase, while informative, may not be the most biologically meaningful representation of their performance. We therefore further validated our results using miRNA expression levels in CD4+ T cells determined in two independent experiments by Cobb et al. (112) and Sommers et al. (113). These two experiments were conducted using different technologies, the latter measuring miRNA expression using the nCounter system (Nanostring Technologies). The Cobb et al. data compared miRNA expression profiles between CD4+CD25- and CD4+CD25+ T-cells. This experiment has the benefit of being performed in the same laboratory and on the same wildtype T conv cells from which we obtained the mRNA microarray data used in our analysis.

Several, but not all, of the most highly-expressed miRNAs in each of the two data sets overlapped. Notably, the let-7 family (consisting of let-7b, let-7c, and let-7d), miR-30b, miR-26b, miR-142-3p, and miR-15a are among the miRNAs found to be expressed in T conv cells in both data sets. This is consistent with differences between the studies, including the particular labs, quantification technologies and the comparison between two cell types in the case of the Cobb et al. data.

We considered the top 20 highly expressed miRNAs reported by Cobb et al. and the top 21 miRNAs reported by Sommers et al. (*3.2.1: Mouse CD4+ Dicer KO expression profiles*). These results can be found in Supplementary Table 1. In general, we found that while there was clearly a significant overlap between highly expressed and active miRNAs, relatively few of the highly expressed miRNAs were also found to be active by the methods we tested. For example, of the top ten motifs returned, MixMir identified three exact seed sequences corresponding to highly expressed miRNAs. miReduce also performed well, but not as accurately (Supplementary Table 1 and Supplementary Table 2). cWords found more highly expressed miRNAs than the simple linear model, but they are ranked further down

the list than either miReduce or MixMir6. Of the discovered highly-expressed miRNAs, miR-142 (both 3p and 5p) is particularly interesting as it has previously been found to be highly expressed in T conv cells and it plays a significant biologically role in regulating cAMP (114). miR-142-3p was found in the Cobb et al. data and was discovered by both miReduce and MixMir6. These results suggest that miRNA motif finding algorithms can play a significant role in identifying the most biologically active miRNAs in a sample and that simply measuring miRNA expression levels is insufficient to do so.

Overall, MixMir ranked true motifs higher than other methods, while the simple linear model and cWords found fewer matches to miRNAs expressed in this cell type (Supplementary Table 1 and Supplementary Table 2). These results are consistent with our previous analysis of the ROC curves on the full miRBase miRNA data set. These results suggest that MixMir tends to rank true miRNAs higher than other motif-finding methods, an important consideration for experimental groups that might only have the resources to validate a few top candidate miRNAs. It also shows that we were able to discover biologically meaningful results in our mouse Dicer-knockout T conv data set.

3.4.4: Analysis of AU bias and positive effects

It is known that there is often an AU bias in computationally discovered motifs when using microarray data (115). The AU content in the 3' UTRs used in our analyses was 55.9%, while the average AU content in the miRNA seed sequences from miRBase was 48.8%. However, the motifs discovered by the simple linear model had very high average AU content, suggesting that their high false positive rate was partially due to discovering elements representing the AU-rich background sequence (Table 10).

Method	% AU in motif
LM Bin	88.67
miReduce	43.33
Sylamer	58.33
cWords2	81.67
MixMir6	53.67

Table 10

AU content of motifs discovered by the different methods. Simple linear models and cWords2 returned motifs with very high AU content. Both MixMir and miReduce had substantially lower average AU content, closer to the background 3' UTR base composition.

MixMir6 motifs had average AU content similar to that in the background 3' UTR sequence, suggesting that the correlation matrix component of MixMir successfully corrected for the AU bias. Consistent with this idea, as we altered the correlation matrix used in MixMir from k = 2 to k = 6, we observed a linear decrease in the average AU content of motifs as k increases (Table 11). Sylamer showed a similar degree of correction. The miReduce results had an even lower average AU content than the background 3' UTRs. cWords, on the other hand, had motif AU composition similar to that of the simple linear model, which was very high and was not significantly changed by altering the value of k (Table 11). Taken together these results showed that the simple linear model suffered from high AU bias, but this bias was corrected by miReduce, Sylamer, and MixMir. Although miReduce does not have an explicit correction for 3' UTR base composition, it likely implicitly performs this correction by finding a motif highly correlated with background composition and then finding the residuals with respect to that motif to identify the remaining motifs. We observed this phenomenon in our data in practice, where miReduce

often found an AU-rich motif as the most significant motif. As described in *3.3.1: Current methods in miRNA motif discovery*, Sylamer likely removes the AU rich motifs as a separate preprocessing step, unlike the other methods.

Method	% AU in motif
MixMir2	75.0
MixMir3	71.0
MixMir4	66.0
MixMir5	53.67
cWords3	85.67
cWords4	84.33
cWords5	86.33
cWords6	86.67

Table 11

Percentage of A and U nucleotides in the top 50 motifs returned. As k decreases from 6 to 5, we see a decrease in the percentage of AUs.

3.4.5: MIXMIR CORRECTS FOR 3' UTR LENGTH

We expect the coefficient of the fixed effect (i.e. the motif effect) to be positive if the motif represents the seed sequence of an active miRNA since miRNAs almost always downregulate their targets and a positive effect corresponds to higher expression in the Dicer KO. To test this, we looked at the number of motifs with a positive effect in each method, both overall and also compared to all motifs with a significant p-value ($p < 0.01$). We find that this was overwhelmingly true across all motifs, particularly the simple linear

model and cWords. Sylamer returned the lowest number of positive-effect motifs in those which are significant, at only 58.64%, while MixMir showed the best enrichment for positive-effect motifs in those which are found significant, compared to the number positive over all motifs tested (Table 12).

Method	Number of significant motifs	Percent of significant positive coefficients	Percent positive coefficients overall
LM Bin	3726	90.97%	99.34%
Sylamer	885	58.64%	NA
cWords2	3744	99.97%	98.63%
MixMir6	121	96.70%	67.94%

Table 12

Percentage of significant motifs that have positive coefficients in the four models examined. The number of significant motifs in the first column is determined by a cutoff of $p < 0.01$. The percentage of motifs from the first column which are positive (i.e., the percentage of significant coefficients which are positive) is given in the second column. The third column is the percentage of all motifs which have positive coefficients, not limited to those which have been found to be significant.

We reasoned that the overall very high enrichment of positive effects across all motifs in the simple linear model might be an artifact due to the inherent relationship between 3' UTR length and motif count, because longer sequences have a higher probability of containing any given motif, simply by chance. Thus an mRNA that is repressed due to a miRNA motif would also induce a similar correlation for all other motifs found in that 3' UTR. To test this hypothesis, we included 3' UTR length as a covariate to test how it would affect the direction of the miRNA effect. A full discussion of this the 3' UTR length effect can be found in *Appendix B.2*: Analysis of the effects of adding a 3' UTR length covariate. Briefly,

the 3' UTR length covariate strongly shifted the *p*-values of motifs found by the simple linear models, which resulted in the PP plots for the simple linear models being significantly less skewed. These results suggest that an additional reason for the higher performance of MixMir compared to the simple linear models is that MixMir implicitly corrects for 3' UTR length using the relatedness matrix. After correcting for 3' UTR length, we found that the percentage of positive effects across motifs remained high but not artificially high. This is consistent with our biological intuition that while most significant motifs should have positive effects, some significant motifs will appear to have negative effects due to the indirect effects that are not captured by our steady-state microarray expression measurements. In any case, since we found that the additional length covariate did not change the rankings of the top 50 motifs in any of the linear methods, we did not use it for the comparisons between methods presented above.

3.4.6: Application to miRNA transfection datasets

All methods perform well for most miRNA transfections but MixMir performs the best for let-7b

In addition to testing MixMir on our mouse Dicer-knockout T conv data, we also tested our algorithm on miRNA transfection data from human cell lines, to demonstrate that our results are not particular to the mouse microarray data set. We tested both microarray and quantitative protein expression data obtained from Selbach et al. (100) (see *3.2.4: HeLa transfection expression profiles*), and compared our results to those obtained from the same data using miReduce, cWords, Sylamer, and the simple linear model (Table 13 and Table 14). This data extends our analysis to a very different technology, from microarrays to pSILAC, and from mouse to human.

	let-7b	miR-1	miR-155	miR-16	miR-30a
MixMir	**1**[2], **2**[1], **3**[3]	**1**[2], **4**[3], **11**[1]	**1**[2], **2**[3], **6**[1]	**1**[2], **2**[3], **3**[1]	**1**[2]
miReduce	**1**[2]	**1**[2]	**1**[2]	**1**[2]	**1**[2]
Sylamer	**1**[2]	**1**[2], **4**[3]	**1**[2], **12**[3]	**1**[2], **11**[3]	**4**[2]
cWords	**1**[2], **2**[1], **3**[3]	**1**[2], **2**[3], **3**[1]	**1**[2], **2**[3]	**1**[2], **2**[3],	**1**[2], **8**[3], **16**[1]
LM Bin	**NA**	**3**[2]	**1**[2]	**NA**	**NA**

Table 13

Results for five different HeLa cell miRNA transfections, derived from proteomics expression data. Red, boldfaced numbers indicate the rank of the motif found; numbers in square brackets indicate the type of match, with 2 indicating an exact seed match, and 1 and 3 indicating offset matches. Overall, we find that most methods are able to correctly identify the seed sequence of the transfected miRNA as the first motif.

	let-7b	miR-1	miR-155	miR-16	miR-30a
MixMir	**1**[3], **2**[2], **5**[1]	**1**[2], **2**[3], **3**[1]	**1**[2], **3**[3]	**1**[2], **2**[3], **17**[1]	**1**[2], **17**[3]
miReduce	**6**[2]	**1**[2], **10**[3]	**1**[2]	**3**[2]	**1**[2]
Sylamer	**19**[2]	**1**[2], **2**[3], **3**[1]	**1**[2]	**2**[2], **4**[3]	**1**[2], **4**[1], **10**[3]
cWords	**NA**	**1**[2], **2**[3], **3**[1]	**1**[2], **2**[3]	**1**[2], **2**[3], **19**[1]	**1**[2], **3**[3], **5**[1]
LM Bin	**NA**	**1**[2], **2**[3], **3**[1]	**1**[2], **3**[3]	**11**[2]	**1**[2], **3**[3], **5**[1]

Table 14

Results for the same five experiments as depicted in Table 13, but using microarray expression data instead of proteomics data. Again we find that in general all of the statistical methods perform quite well, with the exception of the let-7b transfection experiment. However, MixMir is still able to identify an offset seed sequence as the highest ranked motif, and the exact seed sequence as the second highest ranked motif.

We found that nearly all methods were able to find the exact seed sequence for nearly all the of the quantitative proteomics data sets, with the exception being that Sylamer ranked the seed sequence of miR-30a fourth rather than first. This is an expected

result because unlike the Dicer-knockout scenario where many microRNAs were perturbed, the transfection experiment perturbs one microRNA very strongly and therefore is expected to produce much less noisy expression data. Here our analysis demonstrates that the performance of MixMir extends from the complicated T conv data set considered earlier to other simpler data sets as well.

In addition, we found that MixMir was able find the exact seed sequence or an offset seed sequence (in the case of let-7b) of the transfected miRNA as precisely the most significant motif for each of the microarray experiments at 32hrs post , while in several cases the other methods had difficulty doing so. In particular, the other statistical methods had difficulty identifying both seed and offset matches in the let-7b experiment. No other method was able to identify the seed or any offset matches in the let-7b experiment, with miReduce ranking the seed sixth, and Sylamer and cWords performing very poorly. We found that MixMir was able to find many offset seed matches—all 3 offset seed sequences were found generally within the top 10 motifs. Additionally, we found motifs further downstream of the miRNA seed sequence for let-7b (rank 17, miRNA nts 12-17), miR-155 (rank 5, nts 4-9), and miR-16 (rank 16, nts 9-14), which may be suggestive of noncanonical binding in these miRNAs (116,117). The center of miR-16 has also been suggested to be involved in binding to AU-rich elements (118) although this result has been challenged (119). Since miReduce is a useful tool in experimental labs for validating that a transfection experiment actually worked and MixMir improves on the other methods slightly for several experiments, this is an additional practical use of MixMir as well.

MixMir predicts the miR-290 cluster as biologically most significant in mouse embryonic stem cells

Next we analyzed new unpublished microarray data from mouse Dicer knockout embryonic stem cells (*3.2.3: Mouse embryonic stem cell Dicer KO expression profiles*). We found that all methods implicated the exact seed of the miR-290 cluster (AAGTGC) as the top motif, except Sylamer which ranked it second (

Table 15). It is known that the miR-290 cluster, consisting of miR-290 to miR-295, has very high activity in mouse embryonic stem (ES) cells, to the extent that replacing only this microRNA cluster can rescue most of the Dicer KO phenotype (110). Thus our results are consistent with our results for the single microRNA transfection experiments that on relatively simple experiments where only a single microRNA dominates the microRNA transcriptome of the cell, many methods are generally able to find the correct motif. However, the motif analyses extend to offset seeds and non-canonical miRNA targeting as well. MixMir was able to identify both offset seeds for the miR-290 cluster in the top ten predictions, while the other programs found either 1 or 0 of the offset seeds. We did not observe any obvious non-canonical miRNA seeds among the MixMir motifs, a point we discuss further in section 3.5: Discussion.

MixMir6	**1**[2]**, 6**[1], **7**[3]
miReduce	**1**[2]
Sylamer	**2**[2]
cWords2	**1**[2], **6**[3], **15**[1]
LM Bin	**1**[2]

Table 15

Rank of the exact seed and offset seeds of the miR-290 cluster of miRNAs for each of the methods tested for microarray data obtained from comparing Dicer knockout and WT embryonic stem cells.

MixMir identifies highly expressed miRNAs in mouse Dicer knockout adrenal cortex samples

We further tested MixMir on a published set of adrenal cortex Dicer KO experiments, performed by Krill et al. (98). The authors found that while mouse embryos with Dicer KO adrenal cortex cells developed normally up to E14.5, at E18.5 they experienced total adrenal cortex failure. In all they found 16 miRNAs that were down-regulated in the adrenal cortex of both E15.5 and E16.5 mice, including miR-34c, miR-21, miR-10a, and let-7d, which play a role in tumorigenesis among other functions (98). They also presented lists of miRNAs specifically down-regulated at each stage.

We analyzed the mRNA microarray expression data (*3.2.2: Mouse adrenal cortex Dicer KO mRNA and miRNA expression profiles*) from both E15.5 and E16.5 embryos using the linear model, miReduce, Sylamer, cWords, and MixMir. When compared to the miRNAs that are down-regulated at both E15.5 and E16.5, we found that most methods were able to find either an exact or offset seed match to let-7d either as the first or second motif returned, with the exception of the linear model, which performed worse. Overall, MixMir ranked true miRNA seeds higher than the other methods in both E15.5 and E16.5 data sets (Table 16). Most notably, MixMir found both miR-34b and miR-34c in the top ranked motifs at E15.5, which no other method was able to do. We also performed a separate analysis of motif ranks and miRNA matches for E15.5 and E16.5 separately, as some miRNAs were found to be significantly down-regulated at one stage and not at another—namely, there were more such miRNAs at E16.5, as expected. We found similar results in this analysis, in particular that MixMir consistently found biologically significant miRNAs, with performance comparable to miReduce for both time points. cWords and the linear model were comparable for E16.5 only (Supplementary Table 3).

	E15.5		E16.5	
	Rank	**miRNAs**	**Rank**	**miRNAs**
MixMir	**2** **5** **8**	[1]miR-34b-3p, [1]miR-34c-3p [2]let-7d-5p, [2]miR-202-3p [3]let-7d	**1** **3** **16**	[1]miR-34b-3p, [1]miR-34c-3p [2]let-7d-5p, [2]miR-202-3p [3]let-7d-5p
miReduce	**1**	[3]let-7d-5p	**1**	[2]let-7d-5p, [2]miR-202-3p
Sylamer	**2** **10**	[3]let-7d-5p [2]let-7d-5p, [2]miR-202-3p	**NA**	
cWords	**1** **2**	[2]let-7d-5p, [2]miR-202-3p [3]let-7d-5p	**3** **9**	[2]let-7d-5p, [2]miR-202-3p [2]miR-107-3p
LM Bin	**9** **13**	[2]let-7d-5p, [2]miR-202-3p [3]let-7d-5p	**3** **9**	[1]miR-34b-3p, [1]miR-34c-3p [3]miR-193a-3p

Table 16

Comparison of all methods in analyses of adrenal cortex Dicer knockout data for mouse embryos at stages E15.5 and E16.5. We present matches to miRNAs found to be experimentally down-regulated in the Dicer KO samples compared to WT in both E15.5 and E16.5 adrenal cortex samples, as reported by the authors. The top 20 motifs returned by each method were analyzed. Column labeled Rank gives the rank of the motif matched; miRNAs are preceded by the match position of the motif, with [2] indicating an exact seed match.

Thus, testing the different methods on additional biological data sets confirms the improvement of MixMir over previous methods. We believe that the most important use of miRNA motif finding methods is to find a small number of miRNAs that are most important in a particular cell type for further experimental validation, since usually there are very few miRNAs that are active in a cell type (110). Therefore, we view the ability of MixMir to improve the predictions by a small number of motifs to be a significant result that is a feature of the biological properties of the miRNA system.

3.5: DISCUSSION

In conclusion, we have presented MixMir, a novel method for microRNA (miRNA) motif discovery from sequence and gene expression data. Our method corrects for pairwise

sequence similarities between 3' UTRs that could confound a motif finding algorithm in a way that is fundamentally different from previous approaches to this problem (e.g. cWords, Sylamer). We applied MixMir to a microarray dataset from wild-type and Dicer knock-out (KO) mouse CD4+CD25- T cells (T conv cells) collected by one of the authors. Since Dicer is required for miRNA biogenesis, we expect that Dicer KO cells do not contain any miRNAs and indeed this point was validated by quantitative PCR for selected miRNAs, showing a greater than 90% decrease in the knock-out (unpublished results). We found that MixMir was more accurate in finding active miRNAs in these cells than three other similar published methods, miReduce, cWords and Sylamer, as well as a simple linear regression model we used as a baseline for comparison. We validated our computational predictions using two independent biological data sets consisting of miRNA expression measurements in this cell type quantified by either miRNA microarrays or single molecule imaging using the nCounter system (Nanostring Technologies).

Importantly we found that miRNA activity was highly but not perfectly correlated with miRNA abundance in the cells, so it is not sufficient to simply measure miRNA expression levels in a cell type to determine the miRNAs that play the largest role in shaping global gene expression in those cells. For example, as in similar analyses for transcription factors, miRNAs could be highly abundant but not highly active in repressing mRNA expression due to their sub-cellular localization or the presence of competing RNA species that could sequester the miRNAs from their mRNA targets (120). Another possibility is that miRNAs may have differential efficiency of loading into the RISC complex or of targeting mRNAs, and certain mRNAs may not be efficiently repressed by miRNAs due to the presence of either stable RNA secondary structures occluding the miRNA binding site or the binding of additional trans-acting factors. An interesting biological finding from our analysis is that the miRNAs that we found to be the most active in T conv cells were in fact exactly

the miRNAs that were more differentially expressed between these cells and CD4+ CD25+ T cells (T reg), based on previously published data from the same cell type (112).

To confirm the performance of MixMir on additional data sets, we tested MixMir against the other methods on five miRNA transfection experiments in HeLa cells, using both microarray and pSILAC quantitative proteomics data previously published by Selbach et al. (100). In all transfection experiments, for the pSILAC data, nearly all methods were able to find the exact seed sequence first, with the exception of the linear model, which failed to do so in three cases, and Sylamer, which failed to do so for miR-30a (Table 13). In the microarray data, MixMir ranked the exact seed sequence of the transfected miRNA first, with the exception of let-7b where it ranked it second. For the let-7b experiment, all of the other methods performed much more poorly than MixMir, demonstrating that MixMir gives a significant improvement on at least one transfection experiment.

We performed a similar analysis with mouse embryonic stem (ES) cell Dicer knockout experiments, for which we also included previously unpublished microarray expression data, and mouse adrenal cortex Dicer knockout experiments, using data obtained from Krill et al. (98). In the former, we found again that most methods we tested were able to identify the seed sequence of the miR-290 cluster known to be highly active in ES cells but that MixMir additionally found more offset seed sequences for this cluster; in the latter, we found that MixMir either identified more true miRNAs or performed comparably to the other methods depending on the time point examined. Note that the adrenal cortex data might be noisier than the other experiments because it was derived from more heterogeneous primary tissue rather than cell cultures.

These experiments demonstrate the general applicability of MixMir on different technologies (microarray and proteomics), species (human and mouse), cell types (cell lines, primary T cells and adrenal cortex tissue) and experiments of varying complexity,

from relatively simple (microRNA transfection or ablation of a single dominant microRNA cluster) to relatively complex (perturbation of many microRNAs in a tissue). Our analysis of HeLa cells also demonstrate the utility of MixMir in a context where miReduce is often used in practice—to verify that a miRNA transfection experiment was carried out successfully.

In our miRNA targeting model, we made several assumptions similar to previous methods, like miReduce. First, we searched over non-degenerate kmer motifs only. Although this does not rule out the possibility of detecting degenerate motifs, it probably biases our search towards non-degenerate seed matches. Although we searched for several published types of degenerate motifs such as G-bulge sites and imperfect sites in our data, we found only a few cases of such sites. We note that many of the analyses of non-canonical miRNA motifs have been performed on Ago HITS-CLIP or PAR-CLIP data and therefore represent biochemical binding events of the miRNAs, which are not necessarily perfectly correlated with repression that is detectable at the mRNA level. Similar observations hold for ChIP-seq data on transcription factors where biochemical binding does not necessarily produce transcription of the target gene. Second, we searched over motifs in 3' UTRs only. This choice was based on previous results in the literature but can be easily changed to examine other sequences, such as coding sequences or 5' UTRs, by users of MixMir. Third, our model assumes that the miRNA regulatory effect is additive, which is supported by previous evidence (48) but is still an approximation to biological reality.

Our approach to the motif discovery problem borrows an idea from genome-wide association studies (GWAS), namely that cryptic relatedness between individuals acts as a confounding factor that causes simple linear models to detect many false positive associations. In GWAS, cryptic relatedness is captured by a kinship matrix representing pairwise similarities between individuals. In the miRNA motif discovery problem, we considered background nucleotide composition similarity, which may affect miRNA binding

in a variety of ways. It may affect binding site accessibility (121), represent other cis-regulatory sites for RNA-binding proteins, or simply be a correlate of paralogy—consider for instance ribosomal genes that are very similar and have similar expression patterns (e.g. due to similar transcriptional regulation) but are not affected by miRNA targeting (122). Such signals can confound a motif finder based on a simple linear model if sequence similarity is not corrected. In particular, we found that the relatedness matrix corrected for high AU content of the 3' UTRs. This observation could be due to the presence of AU-rich elements, which are known to be involved in mRNA regulation, other AU-rich motifs for trans-acting factors or more open secondary structures in the 3' UTR that might increase the efficiency of miRNA binding. Furthermore, it has been shown in microarray analyses that demonstrated AU bias may be caused by underlying array probe bias (115).

We constructed a relatedness matrix analogous to the kinship matrix by representing kmer content similarity between 3' UTRs, which implicitly accounts for 3' UTR length. Our finding k = 6 provided the most correction of the results is intuitive, as this choice of k corrects for motifs of the same length as the seed sequences for which we are searching and synergistic interactions between nearby miRNA binding sites and RNA binding protein binding sites have been previously documented (123,124). It is possible that we are also computing an approximation to the alignment score of the 3' UTRs and that global similarity of 3' UTRs is more important than the presence of short, 6 nt motifs, but we consider this possibility unlikely because very few pairs of 3' UTRs should have any meaningful sequence alignment at all. Most significantly, MixMir6 was able to correctly implicate significant hexamer motifs associated with both known miRNAs as well as with highly expressed miRNAs in our dataset, as indicated using the area under the truncated receiver operating characteristic (AUROC). In particular, on the data sets we tested, MixMir performed better than current state-of-the-art methods of motif discovery, miReduce,

cWords, and Sylamer (104-106) over the entire range of sensitivity settings considered and on several different types of data.

Notably, both cWords and Sylamer correct for 3' UTR length and compositional biases. Overall, cWords performed better than Sylamer, but exhibited strong AU bias in the datasets we examined. These results suggest that background nucleotide composition similarity can strongly affect the ability of a linear model to uncover true motifs, but also that the way in which we correct for background composition can dramatically alter the results. Unlike Sylamer and cWords, MixMir utilizes the expression fold change values instead of just the ranks. Additionally, MixMir makes pairwise comparisons of the entire 3' UTR sequences, thus performing a more direct comparison of sequence context, rather than comparing motif vs. background composition within each 3' UTR like the other methods.

The MixMir software is freely available online at https://github.com/ldiao/MixMir, and utilizes FaST-LMM, a fast mixed linear model solver that can be obtained freely online (see the MixMir README file with the software for details). One limiting factor in our approach is the total amount of 3’ UTR sequence available to construct large correlation matrices for long kmers. Increasing the kmer length to 7mers or higher would make the correlation matrix very sparse and difficult to estimate accurately. Another drawback of MixMir is its relative computational inefficiency: we exhaustively analyzed all 6mers but if we wanted to exhaustively analyze all 7- or 8mers, the runtime and memory requirements for FaST-LMM would make the computation too inefficient for practical use in our experience. However, miReduce suffers from a similar problem of computational inefficiency for values of k greater than about 6, so this is not an issue unique to MixMir.

Finally, we note that our mixed linear model approach is not limited to solving the miRNA motif discovery problem. Like the REDUCE software, MixMir can potentially also be applied to other regulatory element motif detection problems, such as transcription factor

and RNA binding protein motif prediction, by varying the type of sequence input and gene expression fold change input. For example, REDUCE was originally applied to transcription factors but was later applied to miRNAs in miReduce (104), RNA binding proteins in matrixREDUCE (125), degenerate transcription factor motifs in fREDUCE (126) and ChIP-chip data. We believe that MixMir can be similarly applied to many of these types of data and possibly also other data types such PAR-clip (127) as well.

CHAPTER 4: CONCLUSIONS AND FUTURE APPLICATIONS

4.1: CONCLUSIONS

Though the idea of the mixed linear model was first developed by Fisher to account for the effect on phenotype of genetic similarities between individuals (60), the utility of mixed linear models has in a wide variety of biological problems is clear. Their usefulness is due to their versatility and ease of definition: So long as we can define some measure of expected similarity between samples, we can generate the relationship matrix needed to formulate the model. Such relationships are plentiful in biology—while we have focused here on genetic similarities between strains of yeast and *k*mer similarity between 3' UTR sequences as they have applied to the problems at hand, other examples traditionally include longitudinal data models and random batch effects caused by, for example, the random selection of clinics at which to administer a drug. Recently MLMs have also been applied to correct for cell type composition in epigenome-wide association studies (128), and undoubtedly more applications are yet to come.

We have examined the efficacy of MLMs to correct for relatedness in the yeast *Saccharomyces cerevisiae*, a species exhibiting very strong population structure. There is some doubt as to whether genome-wide association studies are feasible in such a sample, as the confounding effects of the complex population structure may be difficult to overcome (86). In *Chapter 2: Genome-wide association studies in highly structured populations* we investigated how to address this issue and posited that the inclusion of fixed local ancestry covariates in a typical mixed linear model may be helpful in removing the confounding effects. Here we defined the relationship matrix by the identity-by-state matrix estimated from the SNPs being interrogated. While we found issues with power in the event that a causal SNP's local ancestry covariates are themselves associated with the phenotype, our

simulation studies showed that when this is not the case, we have greater power to detect the causal SNP than a mixed linear model without the local ancestry covariates. Thus, what we observed was that the local ancestry MLM we tested had a higher variance when it came to ranking planted causal SNPs: They could be ranked very highly, performing better than any other method; or they could be ranked more lowly. On the other hand, MLM methods without fixed ancestry covariates show more consistency in their rankings of causal SNPs than their counterparts. Nonetheless, we showed that in real data with known gene-phenotype associations, the mixed linear model including local ancestry covariates is able to uncover the SNP nearest the causal gene.

After becoming familiar with MLMs and their application to GWAS, we considered another biological problem, which on the surface appears to be quite different but which also benefited from a similar modeling strategy. In this case we were interested in predicting the active microRNAs in a cell type, using only the change in expression between the wild type cell and the cell in which miRNAs were no longer functional (i.e., with Dicer knocked out). While a considerable amount of work has been done on this problem, we took a different approach and modeled the background sequence composition of the 3' UTR, where miRNAs typically bind. Sequence composition is known to have an effect on the efficacy of miRNA binding (96); however, the exact details as to what features might play a part in this role are not known, making the flexibility of the mixed linear model quite attractive. Specifically, we do not have to incorporate the exact sequence features, which are unknown, and instead we defined a measure of *k*mer similarity between 3' UTRs and used this similarity measure in our MLM relationship matrix.

While some other papers have addressed the issue of sequence composition (105,106), our method firstly allows for regression on quantitative change in expression, rather than ranking alone; we also take a relative approach, by comparing log fold change in

expression and sequence similarity pairwise between all samples, instead of only considering the effect of the background composition of a single 3' UTR on the likelihood of observing a given motif. The method we proposed is more computationally demanding due to the processing of an $N \times N$ large matrix, where N is the number of samples, but we found that overall, our method is able to better detect true miRNAs, and rank them more highly, than others that we tested. We believe that making a small sacrifice in computational complexity is worthwhile to uncover more high-confidence miRNAs, and from a biological perspective since our method ranks true miRNAs more highly it can be better used for further experimental verification, similarly to miReduce (103).

4.2: FUTURE DIRECTIONS

In the work presented here we have demonstrated the importance of correcting for population structure in genome-wide association studies, and how feasible it might be to perform GWAS in highly structured populations using a combination of the mixed linear model with local ancestry covariates. We also demonstrated that the mixed linear model can be used to correct for sequence composition factors that affect miRNA binding.

The mixed linear model appears to be particularly well suited to biological problems. Many experimental setups result in groups of samples being more similar in phenotype than might be expected by chance, which is not indicative of any interesting relatedness, but rather of confounding factors we wish to exclude from our analyses. The power of MLMs to correct for batch effects in a flexible and simple way have been exploited recently to correct for issues such as sample heterogeneity in epigenome-wide association studies in blood (128). In sections *4.2.1: miRNA:mRNA target pairs in breast cancer* and *4.2.2: Application to discovery of RNA binding proteins: Beyond miRNAs*, we describe other

potential applications of MLMs, including another type of motif discovery problem and motif target prediction.

4.2.1: miRNA:mRNA TARGET PAIRS IN BREAST CANCER

As mentioned above, MLMs have great potential to correct for batch effects, particularly when these effects are not well-defined. For example, if we had two different labs produce microarray measurements of the exact same biological sample, it's not unusual that replicate measurements within a lab will be more similar to each other than to measurements from the other lab. One simple way to correct for this kind of batch effect is to include a fixed categorical factor to account for the "lab" effect. However, in some cases the effect is less obvious and well-defined. A good example of this would be studies in which tissue samples are used, since tissue samples can be quite heterogeneous. Studies of breast cancer and chronic obstructive pulmonary disease often collect breast and lung samples, where the tissues studied are in fact composed of many different elements—breast tissue as a whole is a composite of fat and muscle cells, as well as gland and duct tissue, and the lungs are also an organ with complex structure composed of a network of alveoli and bronchioles. This makes expression profiles derived from such tissue samples difficult to compare, since different samples likely contain a different composition of tissue subtypes.

Quite some exploration has been done with regard to the role that miRNAs play in cancer (54-57). We hypothesized that performing mRNA:miRNA target pair prediction using a mixed linear model would be more effective than using a linear model with fixed covariates. Such a model has been proposed previously, taking into consideration miRNA, mRNA, AGO, and other types of expression information (58). However, since breast tissue samples (both tumor and normal) are likely to be heterogeneous, we believed that

correcting for this using a relationship matrix defined by overall tissue sample expression could reduce the number of false positives discovered by a linear model, without the need to guess at possible confounding factors.

We downloaded gene expression data for both normal breast tissue and breast cancer samples from The Cancer Genome Atlas (TCGA) (129). We also obtained miRNA expression data derived from genome analyzer and HiSeq platforms. Many mRNA and miRNA samples were paired. Since we had more HiSeq samples, and data derived from HiSeq should be more accurate, we chose to focus first on paired samples with HiSeq miRNA expression data. This consisted of 490 tumor samples and 83 normal samples. Tumor samples contained expression values for 20365 genes and 1046 miRNAs; normal samples contained expression values for 20365 genes and 748 miRNAs, where the set of normal miRNAs is contained in the set of tumor miRNAs. We removed miRNAs for which there were no matched expression values in normal. Of the samples obtained, we removed all miRNAs with no variation in expression.

Similarity matrices were created based on a pairwise Pearson correlation of gene expression vectors across all genes, for each pair of samples, including miRNA expression. This resulted in a matrix with starkly different values when comparing tumor to normal samples, as expected. In fact, Pearson's *r* is primarily negative for tumor-normal pairs, which means the relationship matrix constructed is not positive semidefinite and therefore cannot be used as is in the mixed linear model. Therefore, we chose to analyze tumor and normal tissues separately first.

We tested conserved miRNA:mRNA target pairs as predicted by TargetScan, with a cutoff of $P_{CT} > 0.5$ (32,49). This was primarily a time saving step, as the set of all possible miRNA:mRNA target pairs in our data sets is considerable ($20365 \times 748 = 1.52 \times 10^7$), so that even using the simple linear model, the analysis of all target pairs is expensive. The

number of TargetScan predicted conserved target pairs fitting our criteria is 95,422 in normal and 97,834 in tumor, which made further analysis more feasible.

We tested a simple model described by:

$$miRNA_i = mRNA_j + \varepsilon,$$

where $miRNA_i$ describes the expression value of miRNA *i* and $mRNA_j$ describes the expression value of mRNA *j*. The MLM further contained a relatedness matrix describing similarity of samples.

Initial MLM analyses were performed using a student license version of ASReml, which allowed for quantitative phenotypes as well as the definition of a custom relationship matrix. After the ASReml license expired and is now only available commercially, we used R's lmekin package, which produces identical results. We tested the basic MLM against a simple linear model (LM) which was performed in R. For additional comparison, we also created a randomized relationship matrix, by randomizing entries in the original matrix.

One of the preliminary results we obtained concerned the number of significant target pairs observed as well as the sign of the coefficient of the fixed effect $mRNA_j$. While it is unsurprising that the number of significant target pairs found at $p < 0.05$ after false discovery rate (FDR) correction is much smaller in the mixed linear model compared with the linear model (~120-200 vs. 7,000-10,000), the more interesting note is that in the top 100 most significant target pairs, we find the fixed coefficients in the MLM to be primarily positive, whereas in the linear model and the mixed linear model with random kinship matrix the coefficients are approximately 45% positive (Table 17). The randomized relationship matrix MLM models are more similar to the linear models overall, but more so in normal samples than in tumor (Table 18).

	% Positive over all	**% Positive in top 100**

	target pairs	miRNAs
LM Tumor	44.74	45.75
LM Normal	45.01	47.08
MLM Tumor	43.35	88.73
MLM Normal	50.84	91.16
MLM Tumor rand	46.05	46.79
MLM Normal rand	45.08	46.20

Table 17

Percentage of target pairs with positive fixed coefficients when analyzed with the linear model (LM), the mixed linear model (MLM), for tumor and normal samples separately. The last two rows represent analysis with the MLM with a randomized relationship matrix, which are more similar to the LM results than to the non-randomized MLM.

	MLM Tumor Rand	LM Tumor
MLM Tumor	0.2714	0.2277
MLM Tumor Rand		0.5508
	MLM Normal Rand	**LM Normal**
MLM Normal	0.1497	0.1515
MLM Normal Rand		0.8140

Table 18

Pearson correlation of *p*-values for each target pair, comparing mixed linear models with randomized relationship matrices and those with non-randomized matrices.

We performed preliminary functional analysis by taking the genes implicated in significant miRNA:mRNA target pairs and analyzing them using FuncAssociate 2.0 with $p < 0.30$ (130), which searches for enriched GO terms in gene lists. Analyses using genes

implicated by the linear model produced no results for either tumor or normal, possibly because of the large number of genes found to be significant. Overrepresented attributes found using significant genes implicated in MLM include cell differentiation, chemotaxis, and various developmental processes in tumor samples, and muscle development in normal samples. The most significant target pairs as identified by the MLM are given in Table 19 below. These include well-known targets such as HOXB3 as well as known oncomirs such as miR-22.

miRNA	Gene
hsa-mir-10a	HOXB3
hsa-mir-182	KIAA1324
hsa-mir-22	FBXO46
hsa-mir-30a	LMBR1L
hsa-let-7b	NPEPL1
hsa-mir-30a	ZNRF1
hsa-mir-21	KLHL15
hsa-mir-21	CASKIN1
hsa-mir-30a	SMAP1
hsa-mir-30a	ELFN2
hsa-mir-30a	CARS
hsa-mir-148a	STT3A
hsa-mir-182	CASP2
hsa-mir-30a	STX2
hsa-mir-30a	FOXA1
hsa-mir-148a	EXTL3
hsa-let-7b	ANKRD49
hsa-let-7b	E2F2
hsa-mir-148a	LRCH1
hsa-let-7b	USP44
hsa-mir-22	HSPG2
hsa-mir-30a	PTGFRN
hsa-mir-30a	BSN
hsa-let-7b	PBX2
hsa-mir-182	C9orf80
hsa-mir-22	ERBB4

Table 19

List of most significant miRNA:mRNA target pairs as identified by the mixed linear model applied to tumor samples.

While these results are only exploratory, they suggest that correcting for possible confounders such as tissue type composition is important, and possibly even more so in heterogeneous samples such as those that would be expected in cancer phenotypes. It also begins to address the issue of moving from motif discovery to target prediction, which is a natural next step. Here we used a precompiled database of conserved and predicted targets, but many other target prediction algorithms exist. We believe that the mixed linear model can be applied to this type of problem as well, perhaps to account for the effect of background sequence composition on binding efficacy as we assumed in the development of MixMir. Issues to overcome would include the computational complexity of testing all potential miRNA:mRNA target pairs, especially when data sets are very large and the relationship matrix becomes unwieldy.

4.2.2: Application to discovery of RNA binding proteins: Beyond miRNAs

While we have focused primarily on miRNA motif discovery, there is no reason that MixMir cannot be applied to other types of motif discovery problems. In particular, we also applied MixMir to the discovery of an RNA binding protein (RNAbp) called HuR. While RNAbps are thought to widely perform posttranscriptional gene regulation, the roles that each RNAbp play are still mostly unknown (131). The RNAbp HuR appears to be a player in mRNA stabilization in the cytoplasm (132). The HuR binding motif has been previously described by computational and experimental methods, with slightly differing results (131,132). We were able to recover these motifs using MixMir in our data.

One primary difference between miRNA motif discovery and RNAbp motif discovery is that with miRNAs, the seed region we take for our motif is highly conserved and also does appear to allow variations, with a few exceptions (59,116). However, RNAbp motifs, like

transcription factor (TF) motifs, often allow variations in binding motifs, which result in motif predictions in the form of position weight matrices (PWMs). These are typically depicted in graphical form as a motif logo, which describe the probability of seeing a particular nucleotide at a position in the motif based on the size of the letter (see Figure 14). For more common application of MixMir to TF and RNAbp motifs, we would need to be able to reassemble a PWM using the list of significant motifs returned by the regression model.

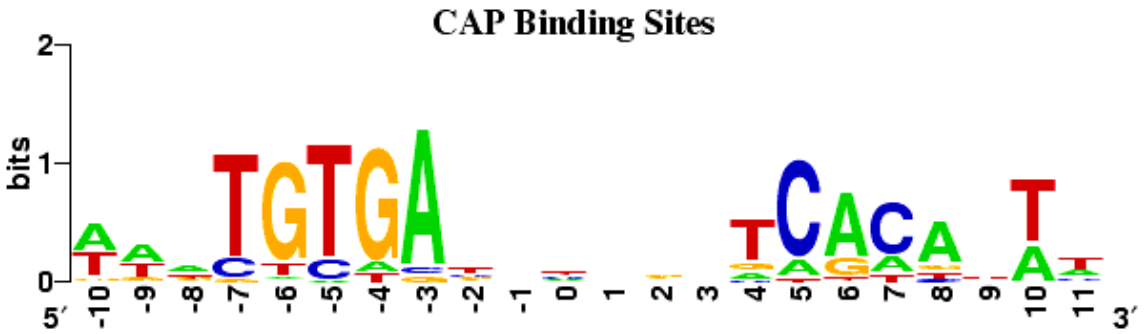


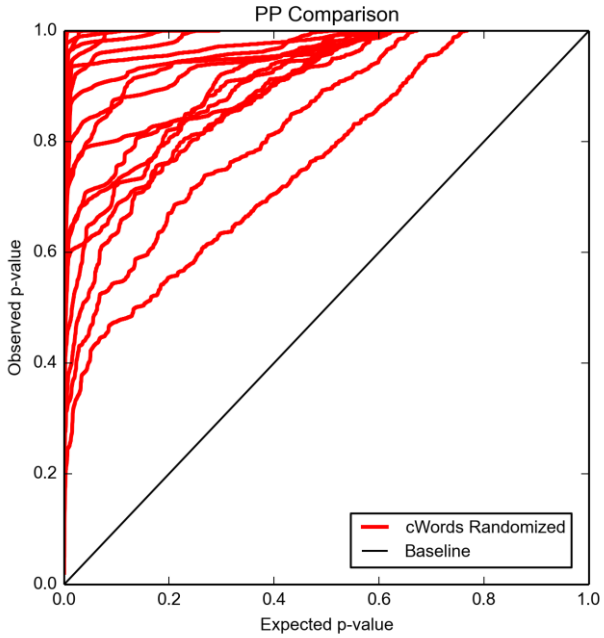
Figure 14

Example motif logo generated by http://weblogo.berkeley.edu/examples.html.

Thus, while we have presented here a framework for discovering significant motifs with consideration for background sequence composition, there are issues that remain to be explored, such as application to miRNA motif prediction and discovery of more complex sets of motifs, such as with transcription factors and RNA binding proteins.

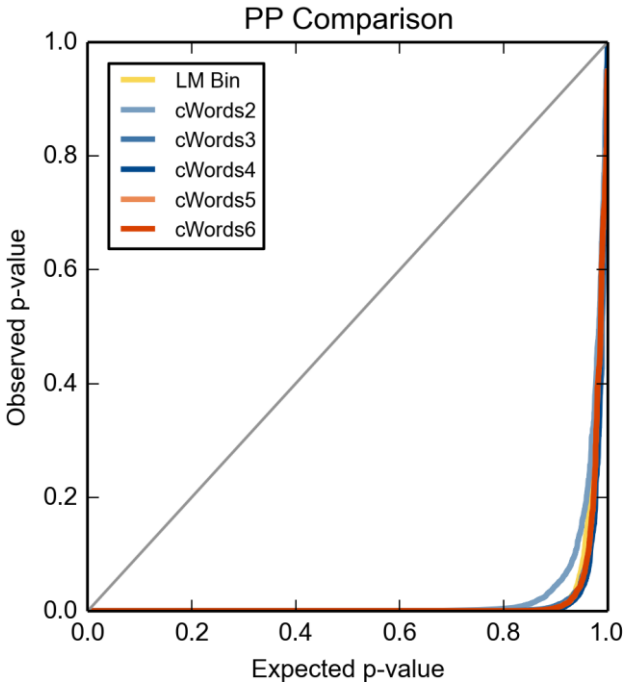
APPENDIX A:

SUPPLEMENTARY FIGURES:



Supplementary Figure 1

PP plots for randomized data, with expected *p*-values obtained from cWords with background Markov Model of order 2. While we would expect to see *p*-values randomly distributed, i.e. falling along the *x*-*y* axis in the plot with randomized data, we in fact see highly inflated *p*-values.



Supplementary Figure 2

Percentile-percentile plot comparing cWords with Markov model of order $k = 2 .. 6$. Expected p-values are found on the x-axis; observed p-values are found on the y-axis. The yellow line is the PP plot for the linear model, which displays an extremely skewed distribution towards many very low p-values. We do not observe better correction in terms of the number of false positives with different choices of k.

SUPPLEMENTARY TABLES:

Method	Rank	Motif	miRNAs matched
LM Bin	13	TGTAAA	[1]mmu-miR-30b-5p, [1]mmu-miR-30c-5p, [1]mmu-miR-30e-5p
	24	TAAACA	[3]mmu-miR-30b-5p, [3]mmu-miR-30c-5p, [3]mmu-miR-30e-5p
cWords2	7	TCAAGT	[2]mmu-miR-26a-5p, [2]mmu-miR-26b-5p
	26	TGTAAA	[1]mmu-miR-30b-5p, [1]mmu-miR-30c-5p, [1]mmu-miR-30e-5p
	30	TTCAAG	[1]mmu-miR-26a-5p, [1]mmu-miR-26b-5p
	35	TAGTTT	[1]mmu-miR-19a-3p
	44	GTGCAA	[2]mmu-miR-19a-3p
	47	AGCAGC	[2]mmu-miR-15b, [2]mmu-miR-195a-5p
Sylamer	8	TCAAGT	[2]mmu-miR-26a-5p, [2]mmu-miR-26b-5p
	23	TAGTGT	[3]mmu-miR-142-3p
miREDUCE	2	GTGCAA	[2]mmu-miR-19a-3p
	6	GTAAAC	[2]mmu-miR-30b-5p, [2]mmu-miR-30c-5p, [2]mmu-miR-30e-5p
	8	TCAAGT	[2]mmu-miR-26a-5p, [2]mmu-miR-26b-5p
	15	GTAGTG	[2]mmu-miR-142-3p
MixMir6	1	GTGCAA	[2]mmu-miR-19a-3p
	2	TCAAGT	[2]mmu-miR-26a-5p, [2]mmu-miR-26b-5p
	4	GTAGTG	[2]mmu-miR-142-3p
	8	TAGTGT	[3]mmu-miR-142-3p
	12	TGTAAA	[1]mmu-miR-30b-5p, [1]mmu-miR-30c-5p, [1]mmu-miR-30e-5p
	17	CTGCAT	[2]mmu-miR-20a-3p
	37	TTCAAG	[1]mmu-miR-26a-5p, [1]mmu-miR-26b-5p

Supplementary Table 1

Performance of each method on miRNA expression data from mouse CD4+ T-cells. The number in square brackets refers to the position of the 6-mer match in the mature miRNA (position 2 is the exact seed match). Selected miRNAs shown are those which are also highly expressed in one of two experimental data sets.

Exact seed match

Rank	LM Bin	cWords2	Sylamer	miReduce	MixMir6
1	TTAAA	TTAAAA	TTTATT	ACAAAA	**GTGCAA**

Offset seed match

LM Bin	cWords2	Sylamer	miReduce	MixMir6
TTAAA				
A	**TTAAAA**	**TTTATT**	ACAAAA	**GTGCAA**

2	TAAAA	TTTAAA	TTAATT	GTGCAA	TCAAGT
3	TATAA	TATAAA	TAAATA	GGGACC	AAAGCA
4	AAAGC	TATATA	AGGGGG	GTACAA	GTAGTG
5	AAGAA	ATATAA	CCCCCC	CCTGGA	GAACAG
6	AAAAA	AATATA	ATAAAT	GTAAAC	GTATCT
7	AAATT	TCAAGT	TTATTA	GATGCT	AAGAAA
8	TTTAC	TAAAAT	TCAAGT	TCAAGT	TAGTGT
9	ATACA	TATACA	TTTAAT	CACGGA	TATAAA
10	AAAAT	ATATAT	CGGCAG	TAGGGT	CAAAGC
11	ATAAA	AAAATA	ATATAG	CCGCGC	GTGGGA
12	TTTAA	TAAAAA	CTTACT	CGGCTT	TGTAAA
13	TGTAA	TGTACA	GGGGGA	GCACTA	TCAATG
14	TTCAA	CTTAAA	CGCGAG	GGATCC	TGTGTG
15	TAAAA	TTTACA	AATAAA	GTAGTG	ATCAAT
16	TTAAA	TTTTAA	TATTGC	TTTGTG	CCAGCG
17	CTTAA	ATATAC	CCTGGG	GGATCG	CTGCAT
18	TACAA	AAAACC	ACGGGT	ACAGTA	CTGCGT
19	TATAC	TACATT	GCACTT	CGCGCC	CTCTGA
20	TTTCA	AACCAA	GCTGCT	GTTCCG	GTCGGC
21	TCAAA	CAAGTT	TCATGT	ACGCTG	TGCAAC
22	AAATA	AATGTT	TCCCCC	TCGATC	GCACTA
23	AAAAA	TTCTAA	TAGTGT	CCGGCT	GCACGC
24	TAAAC	TTTAAG	TAATTA	AACGGG	TTCCAT
25	ATCAA	ACTTAA	CCCTCA	TTCTAT	AAGCAT
26	ATTAA	TGTAAA	CGAAGC	CCGTAA	TCTGCG
27	AAACA	AAATAT	CCGTTT	GCTCCG	CAACGG
28	ATACA	AACTGC	GCACGC	TCGCTC	GCTGGC
29	AAAGA	ATGTAC	TCCCAT	GAACGC	TTAGTA
30	AAAAG	TTCAAG	ACGAAT	CGATGC	AACGGG
31	AAAAG	TTTCAA	TATATG	CGATGG	TCAAAC
32	AATAA	AAACAA	CTACCC	CGTTGG	ATCAAA
33	AAAAC	ATAAAA	TTTAAG	TGGTCC	GGAACA
34	AAAAT	TTCCTA	GGTGAG	ATACAC	AATGCA
35	AAATA	TAGTTT	GGAGGG	AATCTC	CAAACG
36	AAAAT	AAATGT	GGTAAT	ACGAGA	GGCAGC
37	GAAAA	TACATA	AGTATT	AAAGCG	TTCAAG
38	TTTTA	ATACAT	TTATTT	CTACGT	TTCAGC
39	CAAAA	ATACAA	ACGCGT	CACTTA	AGCGCA
40	AAAAA	ACCAAG	TGAAAC	TTCTTC	AAAGTT
41	TAAAT	AATCAA	TTGCTC	AACCGA	AACCGA

2	TAAAA	TTTAAA	TTAATT	GTGCAA	TCAAGT
3	TATAA	TATAAA	TAAATA	GGGACC	AAAGCA
4	AAAGC	TATATA	AGGGGG	GTACAA	GTACAA
5	AAGAA	ATATAA	CCCCCC	CCTGGA	GAACAG
6	AAAAA	AATATA	ATAAAT	GTAAAC	GTATCT
7	AAATT	TCAAGT	TTATTA	GATGCT	AAGAAA
8	TTTAC	TAAAAT	TCAAGT	TCAAGT	TAGTGT
9	ATACA	TATACA	TTTAAT	CACGGA	TATAAA
10	AAAAT	ATATAT	CGGCAG	TAGGGT	CAAAGC
11	ATAAA	AAAATA	ATATAG	CCGCGC	GTGGGA
12	TTTAA	TAAAAA	CTTACT	CGGCTT	TGTAAA
13	TGTAA	TGTACA	GGGGGA	GCACTA	TCAATG
14	TTCAA	CTTAAA	CGCGAG	GGATCC	TGTGTG
15	TAAAA	TTTACA	AATAAA	GTAGTG	ATCAAT
16	TTAAA	TTTTAA	TATTGC	TTTGTG	CCAGCG
17	CTTAA	ATATAC	CCTGGG	GGATCG	CTGCAT
18	TACAA	AAAACC	ACGGGT	ACAGTA	CTGCGT
19	TATAC	TACATT	GCACTT	CGCGCC	CTCTGA
20	TTTCA	AACCAA	GCTGCT	GTTCCG	GTCGGC
21	TCAAA	CAAGTT	TCATGT	ACGCTG	TGCAAC
22	AAATA	AATGTT	TCCCCC	TCGATC	GCACTA
23	AAAAA	TTCTAA	TAGTGT	CCGGCT	GCACGC
24	TAAAC	TTTAAG	TAATTA	AACGGG	TTCCAT
25	ATCAA	ACTTAA	CCCTCA	TTCTAT	AAGCAT
26	ATTAA	TGTAAA	CGAAGC	CCGTAA	TCTGCG
27	AAACA	AAATAT	CCGTTT	GCTCCG	CAACGG
28	ATACA	AACTGC	GCACGC	TCGCTC	GCTGGC
29	AAAGA	ATGTAC	TCCCAT	GAACGC	TTAGTA
30	AAAAG	TTCAAG	ACGAAT	CGATGC	AACGGG
31	AAAAG	TTTCAA	TATATG	CGATGG	TCAAAC
32	AATAA	AAACAA	CTACCC	CGTTGG	ATCAAA
33	AAAAC	ATAAAA	TTTAAG	TGGTCC	GGAACA
34	AAAAT	TTCCTA	GGTGAG	ATACAC	AATGCA
35	AAATA	TAGTTT	GGAGGG	AATCTC	CAAACG
36	AAAAT	AAATGT	GGTAAT	ACGAGA	GGCAGC
37	GAAAA	TACATA	AGTATT	AAAGCG	TTCAAG
38	TTTTA	ATACAT	TTATTT	CTACGT	TTCAGC
39	CAAAA	ATACAA	ACGCGT	CACTTA	AGCGCA
40	AAAAA	ACCAAG	TGAAAC	TTCTTC	AAAGTT
41	TAAAT	AATCAA	TTGCTC	AACCGA	AACCGA

42	AGAAAA	TTAAGT	GCGTTC	CGGTAT	GAACAC	AGAAAA	TTAAGT	**GCGTTC**	CGGTAT	GAACAC
43	ATTTAA	TTAAGA	**AGGGAG**	TTATCG	**GGGGCA**	ATTTAA	**TTAAGA**	**AGGGAG**	TTATCG	**GGGGCA**
44	**AATTTA**	**GTGCAA**	TTTATA	CGCATA	TTCGGC	**AATTTA**	**GTGCAA**	TTTATA	CGCATA	TTCGGC
45	AAATGT	AAAGCA	TAAATT	CGGACG	AATAAG	AAATGT	**AAAGCA**	TAAATT	CGGACG	AATAAG
46	**TGAAAA**	TGATTT	CGTTCA	**GCTGCG**	CGCTCA	**TGAAAA**	**TGATTT**	CGTTCA	**GCTGCG**	**CGCTCA**
47	AACATA	**AGCAGC**	**GCTGGG**	CGGCGG	CCCATA	**AACATA**	**AGCAGC**	**GCTGGG**	**CGGCGG**	CCCATA
48	AATGCA	ATTCAA	**ATATCA**	TTCGCA	TACCAT	**AATGCA**	ATTCAA	**ATATCA**	TTCGCA	**TACCAT**
49	TTACAA	TACAAT	AAATAA	ACCTGT	**TTATTG**	**TTACAA**	TACAAT	**AAATAA**	**ACCTGT**	**TTATTG**
50	TAAATT	**AACAAA**	ACATGT	TTCGGC	AGTAGT	TAAATT	**AACAAA**	**ACATGT**	TTCGGC	**AGTAGT**

Supplementary Table 2

The top 50 motifs from each of the following methods, along with their miRNA matches in miRBase: LM Bin, cWords2, miReduce, MixMir6. Left: Matches to exact seed sequence. Right: Matches allowing offset seed sequences. Light grey backgrounds indicate a match to miRBase, Orange indicates a match to a highly expressed miRNA found by both experimental data sets (Sommers et al. and Cobb et al). Green indicates a miRNA found by only the Cobb et al. data set. We take as highly expressed the miRNAs corresponding to the top ten unique motifs in each dataset.

	E15.5		**E16.5**	
	Rank	**miRNAs**	**Rank**	**miRNAs**
MixMir	**2** **5** **8**	[1]miR-34b-3p, [1]miR-34c-3p [2]let-7d-5p, [2]let-7g-5p, [2]miR-202-3p [3]let-7d-5p, [3]let-7g-5p	**1** **3** **16** **31** **39**	[1]miR-34b-3p, [1]miR-34c-3p [2]let-7d-5p, [2]miR-202-3p [3]let-7d-5p, [2]miR-196a-5p [2]miR-30e-5p [3]miR-193a-3p, [3]miR-193b-3p
miREDUCE	**1** **4**	[3]let-7d-5p, [3]let-7g-5p [1]miR-672-3p	**1** **4**	[2]let-7d-5p, [2]miR-202-3p [2]miR-362-3p, [1]miR-672-3p
Sylamer	**2** **10** **23**	[3]let-7d-5p, [3]let-7g-5p [2]let-7d-5p, [2]let-7g-5p, [2]miR-202-3p [2]miR-22-5p	**34** **43**	[3]miR-10a-3p [3]miR-18a-5p
cWords	**1** **2**	[2]let-7d-5p, [2]let-7g-5p, [2]miR-202-3p [3]let-7d-5p, [3]let-7g-5p	**3** **9** **35** **46**	[2]let-7d-5p, [2]miR-202-3p [2]miR-107-3p [3]miR-193a-3p [3]miR-34b-3p, [3]miR-34c-3p

Supplementary Table 3

Comparison of all methods in analyses of adrenal cortex Dicer knockout data for mouse embryos at stages E15.5 and E16.5. We present matches to miRNAs found experimentally down-regulated in the Dicer KO compared to WT adrenal cortex samples, broken down for E15.5 and E16.5 separately. As in Table 15, only the top 50 motifs returned by each method were analyzed.

APPENDIX B:

B.1: PLOTTING THE TRUNCATED RECEIVER OPERATOR CURVES

To draw a ROC curve, we must be able to define the true positives. In our case, we chose not to draw ROC curves across all possible motifs while using all miRNAs in miRBase as the true positives, since relatively few of them are expected to be expressed in a particular cell type. Furthermore, the methods did not output the same number of motifs—in particular, miReduce outputs many fewer motifs than the other methods. It is not clear how to best draw ROC curves when the methods do not output the same number of predictions.

We thus chose to truncate the ROC curves to the number of motifs to N = 20 and N = 50, to demonstrate how well the methods perform in the top predictions. The way we truncated the curves produces exactly the same curves as would be obtained by magnifying the top results in a full ROC curve. The remainder of the full ROC curve is expected to approach the random predictor line for all methods and would not give additional information about the performance of the methods. After truncation, we simply scale the X and Y axes to both range from 0 to 1. We caution that this truncated AUC statistic should only be used to compare the different methods to each other and not to the typical baseline value of 0.5 for a random method. To this end, we include in each plot a baseline for a random predictor calculated from the expected true and false positive rate given the total number of hexamers matching miRBase miRNAs and the total number of motifs being tested.

In the truncated ROC curves, we see that the results of the comparisons are robust in that the accuracy of MixMir dominates that of the other methods over almost the entire range of sensitivity settings. We consider this to be the best indicator of MixMir's

performance as the AUC values are really only needed for comparison when ROC curves intersect each other.

B.2: Analysis of the effects of adding a 3' UTR length covariate

Figure 15 plots the percentage of positive associations against motif rank as a PP-plot. The LM Bin model without 3' UTR length as a covariate had nearly all positive associations across all motifs, but when we added the 3' UTR length covariate, this was altered dramatically. There was a less pronounced effect for MixMir, presumably because the relationship matrix implicitly corrects for this UTR length effect, as genes with longer UTRs (with more motifs present) will have lower relatedness to genes with shorter UTRs.

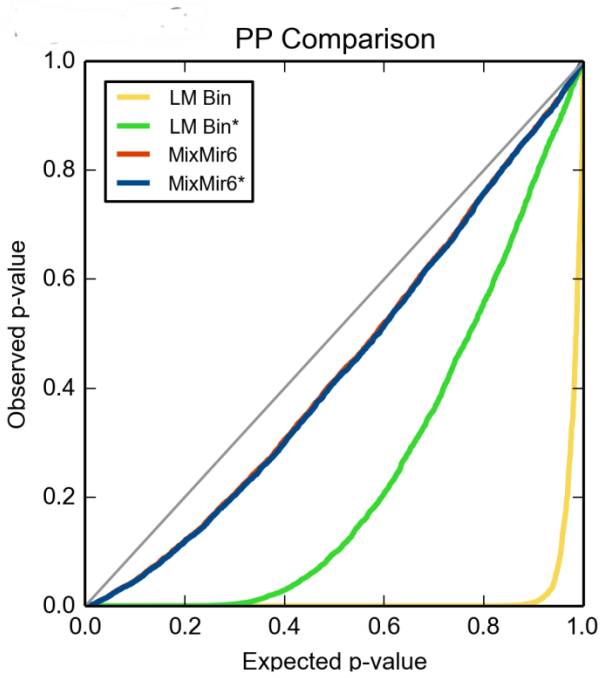


Figure 15

Comparison of PP plots for the linear model and MixMir6, with and without the inclusion of a fixed effect variable to account for 3' UTR length. The model including the 3' UTR length variable is denoted by *. We notice that while the *p*-values of the linear model change dramatically with the inclusion of the additional variable, the results of MixMir do not visibly change.

Additionally, the inclusion of the 3' UTR length covariate partially corrected the skewness of the PP plots observed in Figure 12 (see Figure 15). Notably, the simple linear models became much less skewed. Interestingly, MixMir6* showed no improvement over MixMir6. Further, motif rankings produced by the linear model was substantially different when comparing models with and without the added covariate. This shift was much smaller or non-existent in MixMir (Table 20).

The addition of the 3' UTR length covariate provides a strong correction for the overall percentage of positive coefficients in the simple linear models (Table 21). This brings out the enrichment of positive coefficients in the significant motifs for the linear models, to be more in line with what we observe in the mixed linear models.

However, note that these changes in motif rank did not strongly affect our previous ROC results, as the most highly ranked motifs did not change significantly (data not shown). We thus present our results in the main text without the correction for 3' UTR length.

Method	LM Bin*	MixMir6	MixMir6*
LM Bin	0.5666	0.2293	0.2227
LM Bin*		0.3097	0.3088
MixMir6			0.9993

Table 20

Pairwise Pearson correlations of motif rank, comparing LM Bin and MixMir. While motif rank was considerably changed by adding the UTR length covariate to the linear model, MixMir changed much less.

Method	Number of significant motifs ($p < 0.05$)	Percent of positive coefficients	Percent positive coefficients (overall)
LM Bin*	1792	75.56%	61.04%
MLM6*	439	86.33%	66.75%

Table 21

After incorporating a covariate for 3' UTR length (methods including the covariate marked with an asterisk), we found that the number of positive coefficients overall dropped significantly, particularly for the simple linear model (LM Bin). Similar to Table 12, the number of significant motifs in the first column is determined by a cutoff of $p < 0.05$. The second column shows the percentage of motifs from the first column which have positive coefficients, and the third column shows the percentage of all motifs which have positive coefficients. Notably, the overall percentage of positive coefficients has dropped considerably for the linear model. However, MixMir6 has changed very little.

REFERENCES:

1. Diao, L. and Chen, K.C. (2012) Local ancestry corrects for population structure in Saccharomyces cerevisiae genome-wide association studies. *Genetics*, **192**, 1503-1511.
2. Diao, L., Marcais, A., Norton, S. and Chen, K.C. (2014) MixMir: microRNA motif discovery from gene expression data using mixed linear models. *Nucleic Acids Res*.
3. Visscher, P.M., Brown, M.A., McCarthy, M.I. and Yang, J. (2012) Five years of GWAS discovery. *Am J Hum Genet*, **90**, 7-24.
4. Kruglyak, L. (2008) The road to genome-wide association studies. *Nat Rev Genet*, **9**, 314-318.
5. Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science*, **273**, 1516-1517.
6. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860-921.
7. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304-1351.
8. Consortium, I.H. (2003) The International HapMap Project. *Nature*, **426**, 789-796.
9. Hinds, D.A., Stuve, L.L., Nilsen, G.B., Halperin, E., Eskin, E., Ballinger, D.G., Frazer, K.A. and Cox, D.R. (2005) Whole-genome patterns of common DNA variation in three human populations. *Science*, **307**, 1072-1079.
10. Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., Mayne, S.T. *et al.* (2005) Complement factor H polymorphism in age-related macular degeneration. *Science*, **308**, 385-389.
11. Dewan, A., Liu, M., Hartman, S., Zhang, S.S., Liu, D.T., Zhao, C., Tam, P.O., Chan, W.M., Lam, D.S., Snyder, M. *et al.* (2006) HTRA1 promoter polymorphism in wet age-related macular degeneration. *Science*, **314**, 989-992.
12. Menzel, S., Garner, C., Gut, I., Matsuda, F., Yamaguchi, M., Heath, S., Foglio, M., Zelenika, D., Boland, A., Rooks, H. *et al.* (2007) A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. *Nat Genet*, **39**, 1197-1199.
13. Lettre, G., Sankaran, V.G., Bezerra, M.A., Araújo, A.S., Uda, M., Sanna, S., Cao, A., Schlessinger, D., Costa, F.F., Hirschhorn, J.N. *et al.* (2008) DNA polymorphisms at the BCL11A, HBS1L-MYB, and beta-globin loci associate with fetal hemoglobin levels and pain crises in sickle cell disease. *Proc Natl Acad Sci U S A*, **105**, 11869-11874.
14. Uda, M., Galanello, R., Sanna, S., Lettre, G., Sankaran, V.G., Chen, W., Usala, G., Busonero, F., Maschio, A., Albai, G. *et al.* (2008) Genome-wide association study shows BCL11A associated with persistent fetal hemoglobin and amelioration of the phenotype of beta-thalassemia. *Proc Natl Acad Sci U S A*, **105**, 1620-1625.
15. Sankaran, V.G., Menne, T.F., Xu, J., Akie, T.E., Lettre, G., Van Handel, B., Mikkola, H.K., Hirschhorn, J.N., Cantor, A.B. and Orkin, S.H. (2008) Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor BCL11A. *Science*, **322**, 1839-1842.
16. Barrett, J.C., Hansoul, S., Nicolae, D.L., Cho, J.H., Duerr, R.H., Rioux, J.D., Brant, S.R., Silverberg, M.S., Taylor, K.D., Barmada, M.M. *et al.* (2008) Genome-wide association

defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet*, **40**, 955-962.
17. Zeggini, E., Scott, L.J., Saxena, R., Voight, B.F., Marchini, J.L., Hu, T., de Bakker, P.I., Abecasis, G.R., Almgren, P., Andersen, G. *et al.* (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet*, **40**, 638-645.
18. Willer, C.J., Speliotes, E.K., Loos, R.J., Li, S., Lindgren, C.M., Heid, I.M., Berndt, S.I., Elliott, A.L., Jackson, A.U., Lamina, C. *et al.* (2009) Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat Genet*, **41**, 25-34.
19. Wray, N.R., Yang, J., Hayes, B.J., Price, A.L., Goddard, M.E. and Visscher, P.M. (2013) Pitfalls of predicting complex traits from SNPs. *Nat Rev Genet*, **14**, 507-515.
20. Devlin, B. and Roeder, K. (1999) Genomic control for association studies. *Biometrics*, **55**, 997-1004.
21. Pritchard, J.K. and Donnelly, P. (2001) Case-control studies of association in structured or admixed populations. *Theor Popul Biol*, **60**, 227-237.
22. Bacanu, S.A., Devlin, B. and Roeder, K. (2002) Association studies for quantitative traits in structured populations. *Genet Epidemiol*, **22**, 78-93.
23. Patterson, N., Price, A.L. and Reich, D. (2006) Population structure and eigenanalysis. *PLoS Genet*, **2**, e190.
24. Marchini, J., Cardon, L.R., Phillips, M.S. and Donnelly, P. (2004) The effects of human population structure on large genetic association studies. *Nat Genet*, **36**, 512-517.
25. Kang, H.M., Zaitlen, N.A., Wade, C.M., Kirby, A., Heckerman, D., Daly, M.J. and Eskin, E. (2008) Efficient control of population structure in model organism association mapping. *Genetics*, **178**, 1709-1723.
26. Pritchard, J.K., Stephens, M., Rosenberg, N.A. and Donnelly, P. (2000) Association mapping in structured populations. *Am J Hum Genet*, **67**, 170-181.
27. Falush, D., Stephens, M. and Pritchard, J.K. (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567-1587.
28. Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.Y., Freimer, N.B., Sabatti, C. and Eskin, E. (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat Genet*, **42**, 348-354.
29. Ghildiyal, M. and Zamore, P.D. (2009) Small silencing RNAs: an expanding universe. *Nat Rev Genet*, **10**, 94-108.
30. Bartel, D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281-297.
31. Farh, K.K., Grimson, A., Jan, C., Lewis, B.P., Johnston, W.K., Lim, L.P., Burge, C.B. and Bartel, D.P. (2005) The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science*, **310**, 1817-1821.
32. Friedman, R.C., Farh, K.K., Burge, C.B. and Bartel, D.P. (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res*, **19**, 92-105.
33. Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E. and Mello, C.C. (1998) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, **391**, 806-811.
34. Ecker, J.R. and W., D.R. (1986) Inhibition of gene expression in plant cells by expression of antisense RNA. *Proc Natl Acad Sci U S A*, **83**, 5372-5376.
35. Lee, R.C., L., F.R. and V., A. (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, **75**, 843-854.

36. Siomi, M.C., Sato, K., Pezic, D. and Aravin, A.A. (2011) PIWI-interacting small RNAs: the vanguard of genome defence. *Nat Rev Mol Cell Biol*, **12**, 246-258.
37. Kawaji, H. and Hayashizaki, Y. (2008) Exploration of small RNAs. *PLoS Genet*, **4**, e22.
38. Gebetsberger, J. and Polacek, N. (2013) Slicing tRNAs to boost functional ncRNA diversity. *RNA Biol*, **10**, 1798-1806.
39. Peng, H., Shi, J., Zhang, Y., Zhang, H., Liao, S., Li, W., Lei, L., Han, C., Ning, L., Cao, Y. *et al.* (2012) A novel class of tRNA-derived small RNAs extremely enriched in mature mouse sperm. *Cell Res*, **22**, 1609-1612.
40. Wei, C., Salichos, L., Wittgrove, C.M., Rokas, A. and Patton, J.G. (2012) Transcriptome-wide analysis of small RNA expression in early zebrafish development. *RNA*, **18**, 915-929.
41. Keam, S.P., Young, P.E., McCorkindale, A.L., Dang, T.H., Clancy, J.L., Humphreys, D.T., Preiss, T., Hutvagner, G., Martin, D.I., Cropley, J.E. *et al.* (2014) The human Piwi protein Hiwi2 associates with tRNA-derived piRNAs in somatic cells. *Nucleic Acids Res*.
42. Lee, Y.S., Shibata, Y., Malhotra, A. and Dutta, A. (2009) A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). *Genes Dev*, **23**, 2639-2649.
43. Lee, Y., Kim, M., Han, J., Yeom, K.H., Lee, S., Baek, S.H. and Kim, V.N. (2004) MicroRNA genes are transcribed by RNA polymerase II. *EMBO J*, **23**, 4051-4060.
44. Borchert, G.M., Lanier, W. and Davidson, B.L. (2006) RNA polymerase III transcribes human microRNAs. *Nat Struct Mol Biol*, **13**, 1097-1101.
45. Auyeung, V.C., Ulitsky, I., McGeary, S.E. and Bartel, D.P. (2013) Beyond secondary structure: primary-sequence determinants license pri-miRNA hairpins for processing. *Cell*, **152**, 844-858.
46. Lau, N.C., Lim, L.P., Weinstein, E.G. and Bartel, D.P. (2001) An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science*, **294**, 858-862.
47. Yang, J.S., Phillips, M.D., Betel, D., Mu, P., Ventura, A., Siepel, A.C., Chen, K.C. and Lai, E.C. (2011) Widespread regulatory activity of vertebrate microRNA* species. *RNA*, **17**, 312-326.
48. Bartel, D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, **136**, 215-233.
49. Lewis, B.P., B., B.C. and P., B.D. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15-20.
50. Chen, C.Z., Li, L., Lodish, H.F. and Bartel, D.P. (2004) MicroRNAs modulate hematopoietic lineage differentiation. *Science*, **303**, 83-86.
51. Li, J., Wan, Y., Ji, Q., Fang, Y. and Wu, Y. (2013) The role of microRNAs in B-cell development and function. *Cell Mol Immunol*, **10**, 107-112.
52. Li, X., Zhang, J., Gao, L., McClellan, S., Finan, M.A., Butler, T.W., Owen, L.B., Piazza, G.A. and Xi, Y. (2012) MiR-181 mediates cell differentiation by interrupting the Lin28 and let-7 feedback circuit. *Cell Death Differ*, **19**, 378-386.
53. Pauley, K.M., Cha, S. and Chan, E.K. (2009) MicroRNA in autoimmunity and autoimmune diseases. *J Autoimmun*, **32**, 189-194.
54. Calin, G.A. and Croce, C.M. (2006) MicroRNA signatures in human cancers. *Nat Rev Cancer*, **6**, 857-866.
55. Farazi, T.A., Spitzer, J.I., Morozov, P. and Tuschl, T. (2011) miRNAs in human cancer. *J Pathol*, **223**, 102-115.

56. Lu, J., Getz, G., Miska, E.A., Alvarez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., Ebert, B.L., Mak, R.H., Ferrando, A.A. *et al.* (2005) MicroRNA expression profiles classify human cancers. *Nature*, **435**, 834-838.
57. Dvinge, H., Git, A., Gräf, S., Salmon-Divon, M., Curtis, C., Sottoriva, A., Zhao, Y., Hirst, M., Armisen, J., Miska, E.A. *et al.* (2013) The shaping and functional consequences of the microRNA landscape in breast cancer. *Nature*, **497**, 378-382.
58. Stanhope, S.A., Sengupta, S., den Boon, J., Ahlquist, P. and Newton, M.A. (2009) Statistical use of argonaute expression and RISC assembly in microRNA target identification. *PLoS Comput Biol*, **5**, e1000516.
59. Chi, S.W., Hannon, G.J. and Darnell, R.B. (2012) An alternative mode of microRNA target recognition. *Nat Struct Mol Biol*, **19**, 321-327.
60. Fisher, R.A. (1919) XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, **52**, 399-433.
61. Norton, B. and Pearson, E.S. (1976) A note on the background to, and refereeing of, R. A. Fisher's 1918 paper 'On the correlation between relatives on the supposition of Mendelian inheritance'. *Notes Rec R Soc Lond*, **31**, 151-162.
62. Zhou, X. and Stephens, M. (2012) Genome-wide efficient mixed-model analysis for association studies. *Nat Genet*, **44**, 821-824.
63. Jiang, J. (2007) *Linear and Generalized Linear Mixed Models and Their Applications*. Springer Science+Business Media, LLC, New York, NY.
64. Searle, S.R., Casella, G. and McCulloch, C.E. (1992) *Variance components*. Wiley, New York.
65. Liti, G., Carter, D.M., Moses, A.M., Warringer, J., Parts, L., James, S.A., Davey, R.P., Roberts, I.N., Burt, A., Koufopanou, V. *et al.* (2009) Population genomics of domestic and wild yeasts. *Nature*, **458**, 337-341.
66. Warringer, J., Zörgö, E., Cubillos, F.A., Zia, A., Gjuvsland, A., Simpson, J.T., Forsmark, A., Durbin, R., Omholt, S.W., Louis, E.J. *et al.* (2011) Trait variation in yeast is defined by population history. *PLoS Genet*, **7**, e1002111.
67. Pritchard, J.K., Stephens, M. and Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945-959.
68. Schacherer, J., Shapiro, J.A., Ruderfer, D.M. and Kruglyak, L. (2009) Comprehensive polymorphism survey elucidates population structure of Saccharomyces cerevisiae. *Nature*, **458**, 342-345.
69. Pasaniuc, B., Sankararaman, S., Kimmel, G. and Halperin, E. (2009) Inference of locus-specific ancestry in closely related populations. *Bioinformatics*, **25**, i213-221.
70. Sankararaman, S., Sridhar, S., Kimmel, G. and Halperin, E. (2008) Estimating local ancestry in admixed populations. *Am J Hum Genet*, **82**, 290-303.
71. Ruderfer, D.M., Pratt, S.C., Seidel, H.S. and Kruglyak, L. (2006) Population genomic analysis of outcrossing and recombination in yeast. *Nat Genet*, **38**, 1077-1081.
72. Liti, G., Carter, D., Moses, A., Warringer, J., Parts, L. and James, S. (2009) Population genomics of domestic and wild yeasts. *Nature*, **458**, 337-341.
73. Mancera, E., Bourgon, R., Brozzi, A., Huber, W. and Steinmetz, L.M. (2008) High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature*, **454**, 479-485.
74. Yu, J., Pressoir, G., Briggs, W.H., Vroh Bi, I., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B. *et al.* (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet*, **38**, 203-208.

75. Zhao, K., Aranzana, M.J., Kim, S., Lister, C., Shindo, C., Tang, C., Toomajian, C., Zheng, H., Dean, C., Marjoram, P. *et al.* (2007) An Arabidopsis example of association mapping in structured samples. *PLoS Genet*, **3**, e4.
76. Falush, D., Stephens, M. and Pritchard, J.K. (2007) Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol Ecol Notes*, **7**, 574-578.
77. Devlin, B., Bacanu, S.A. and Roeder, K. (2004) Genomic Control to the extreme. *Nat Genet*, **36**, 1129-1130; author reply 1131.
78. Reich, D.E. and B., G.D. (2001) Detecting association in a case-control study while correcting for population stratification. *Genet Epidemiol*, **20**, 4-16.
79. Berriz, G.F., King, O.D., Bryant, B., Sander, C. and Roth, F.P. (2003) Characterizing gene sets with FuncAssociate. *Bioinformatics*, **19**, 2502-2504.
80. Cubillos, F.A., Billi, E., Zörgö, E., Parts, L., Fargier, P., Omholt, S., Blomberg, A., Warringer, J., Louis, E.J. and Liti, G. (2011) Assessing the complex architecture of polygenic traits in diverged yeast populations. *Mol Ecol*, **20**, 1401-1413.
81. Chen, K., van Nimwegen, E., Rajewsky, N. and Siegal, M.L. (2010) Correlating gene expression variation with cis-regulatory polymorphism in Saccharomyces cerevisiae. *Genome Biol Evol*, **2**, 697-707.
82. Yang, J., Zaitlen, N.A., Goddard, M.E., Visscher, P.M. and Price, A.L. (2014) Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet*, **46**, 100-106.
83. Verdu, P. and Rosenberg, N.A. (2011) A general mechanistic model for admixture histories of hybrid populations. *Genetics*, **189**, 1413-1426.
84. Ohashi, J. and Tokunaga, K. (2001) The power of genome-wide association studies of complex disease genes: statistical limitations of indirect approaches using SNP markers. *J Hum Genet*, **46**, 478-482.
85. Ehrenreich, I.M., Gerke, J.P. and Kruglyak, L. (2009) Genetic dissection of complex traits in yeast: insights from studies of gene expression and other phenotypes in the BYxRM cross. *Cold Spring Harb Symp Quant Biol*, **74**, 145-153.
86. Connelly, C.F. and M., A.J. (2012) On the Prospects of Whole-Genome Association Mapping in Saccharomyces cerevisiae. *Genetics*, **191**, 1345-1353.
87. Ohya, Y., Sese, J., Yukawa, M., Sano, F., Nakatani, Y., Saito, T.L., Saka, A., Fukuda, T., Ishihara, S., Oka, S. *et al.* (2005) High-dimensional and large-scale phenotyping of yeast mutants. *Proc Natl Acad Sci U S A*, **102**, 19015-19020.
88. Payseur, B.A. and Place, M. (2007) Prospects for association mapping in classical inbred mouse strains. *Genetics*, **175**, 1999-2008.
89. Brachi, B., Morris, G.P. and Borevitz, J.O. (2011) Genome-wide association studies in plants: the missing heritability is in the field. *Genome Biol*, **12**, 232.
90. Ranc, N., Muños, S., Xu, J., Le Paslier, M.C., Chauveau, A., Bounon, R., Rolland, S., Bouchet, J.P., Brunel, D. and Causse, M. (2012) Genome-wide association mapping in tomato (Solanum lycopersicum) is possible using genome admixture of Solanum lycopersicum var. cerasiforme. *G3 (Bethesda)*, **2**, 853-864.
91. Tsai, K.L., Noorai, R.E., Starr-Moss, A.N., Quignon, P., Rinz, C.J., Ostrander, E.A., Steiner, J.M., Murphy, K.E. and Clark, L.A. (2012) Genome-wide association studies for multiple diseases of the German Shepherd Dog. *Mamm Genome*, **23**, 203-211.
92. Mackay, T.F., Richards, S., Stone, E.A., Barbadilla, A., Ayroles, J.F., Zhu, D., Casillas, S., Han, Y., Magwire, M.M., Cridland, J.M. *et al.* (2012) The Drosophila melanogaster Genetic Reference Panel. *Nature*, **482**, 173-178.
93. Seldin, M.F., Pasaniuc, B. and Price, A.L. (2011) New approaches to disease mapping in admixed populations. *Nat Rev Genet*, **12**, 523-528.

94. Shriner, D., Adeyemo, A. and Rotimi, C.N. (2011) Joint ancestry and association testing in admixed individuals. *PLoS Comput Biol*, **7**, e1002325.
95. Kozomara, A. and Griffiths-Jones, S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res*, **39**, D152-157.
96. Grimson, A., Farh, K.K., Johnston, W.K., Garrett-Engele, P., Lim, L.P. and Bartel, D.P. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell*, **27**, 91-105.
97. Cobb, B.S., Nesterova, T.B., Thompson, E., Hertweck, A., O'Connor, E., Godwin, J., Wilson, C.B., Brockdorff, N., Fisher, A.G., Smale, S.T. *et al.* (2005) T cell lineage choice and differentiation in the absence of the RNase III enzyme Dicer. *J Exp Med*, **201**, 1367-1373.
98. Krill, K.T., Gurdziel, K., Heaton, J.H., Simon, D.P. and Hammer, G.D. (2013) Dicer deficiency reveals microRNAs predicted to control gene expression in the developing adrenal cortex. *Mol Endocrinol*, **27**, 754-768.
99. Nesterova, T.B., Popova, B.C., Cobb, B.S., Norton, S., Senner, C.E., Tang, Y.A., Spruce, T., Rodriguez, T.A., Sado, T., Merkenschlager, M. *et al.* (2008) Dicer regulates Xist promoter methylation in ES cells indirectly through transcriptional control of Dnmt3a. *Epigenetics Chromatin*, **1**, 2.
100. Selbach, M., Schwanhäusser, B., Thierfelder, N., Fang, Z., Khanin, R. and Rajewsky, N. (2008) Widespread changes in protein synthesis induced by microRNAs. *Nature*, **455**, 58-63.
101. Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520-562.
102. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, D. (2002) The human genome browser at UCSC. *Genome Res*, **12**, 996-1006.
103. Sood, P., Krek, A., Zavolan, M., Macino, G. and Rajewsky, N. (2006) Cell-type-specific signatures of microRNAs on target mRNA expression. *Proc Natl Acad Sci U S A*, **103**, 2746-2751.
104. Bussemaker, H.J., Li, H. and Siggia, E.D. (2001) Regulatory element detection using correlation with expression. *Nat Genet*, **27**, 167-171.
105. van Dongen, S., Abreu-Goodger, C. and Enright, A.J. (2008) Detecting microRNA binding and siRNA off-target effects from expression data. *Nat Methods*, **5**, 1023-1025.
106. Rasmussen, S.H., Jacobsen, A. and Krogh, A. (2013) cWords - systematic microRNA regulatory motif discovery from mRNA expression data. *Silence*, **4**, 2.
107. Bartonicek, N. and Enright, A.J. (2010) SylArray: a web server for automated detection of miRNA effects from expression data. *Bioinformatics*, **26**, 2900-2901.
108. Thomas-Chollier, M., Sand, O., Turatsinze, J.V., Janky, R., Defrance, M., Vervisch, E., Brohée, S. and van Helden, J. (2008) RSAT: regulatory sequence analysis tools. *Nucleic Acids Res*, **36**, W119-127.
109. Lippert, C., Listgarten, J., Liu, Y., Kadie, C.M., Davidson, R.I. and Heckerman, D. (2011) FaST linear mixed models for genome-wide association studies. *Nat Methods*, **8**, 833-835.
110. Sinkkonen, L., Hugenschmidt, T., Berninger, P., Gaidatzis, D., Mohn, F., Artus-Revel, C.G., Zavolan, M., Svoboda, P. and Filipowicz, W. (2008) MicroRNAs control de novo DNA methylation through regulation of transcriptional repressors in mouse embryonic stem cells. *Nat Struct Mol Biol*, **15**, 259-267.

111. Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., Pfeffer, S., Rice, A., Kamphorst, A.O., Landthaler, M. *et al.* (2007) A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, **129**, 1401-1414.
112. Cobb, B.S., Hertweck, A., Smith, J., O'Connor, E., Graf, D., Cook, T., Smale, S.T., Sakaguchi, S., Livesey, F.J., Fisher, A.G. *et al.* (2006) A role for Dicer in immune regulation. *J Exp Med*, **203**, 2519-2527.
113. Sommers, C.L., Rouquette-Jazdanian, A.K., Robles, A.I., Kortum, R.L., Merrill, R.K., Li, W., Nath, N., Wohlfert, E., Sixt, K.M., Belkaid, Y. *et al.* (2013) miRNA signature of mouse helper T cell hyper-proliferation. *PLoS One*, **8**, e66709.
114. Huang, B., Zhao, J., Lei, Z., Shen, S., Li, D., Shen, G.X., Zhang, G.M. and Feng, Z.H. (2009) miR-142-3p restricts cAMP production in CD4+CD25- T cells and CD4+CD25+ TREG cells by targeting AC9 mRNA. *EMBO Rep*, **10**, 180-185.
115. Elkon, R. and R., A. (2008) Removal of AU bias from microarray mRNA expression data enhances computational identification of active microRNAs. *PLoS Comput Biol*, **4**, e1000189.
116. Shin, C., Nam, J.W., Farh, K.K., Chiang, H.R., Shkumatava, A. and Bartel, D.P. (2010) Expanding the microRNA targeting code: functional sites with centered pairing. *Mol Cell*, **38**, 789-802.
117. Martin, H.C., Wani, S., Steptoe, A.L., Krishnan, K., Nones, K., Nourbakhsh, E., Vlassov, A., Grimmond, S.M. and Cloonan, N. (2014) Imperfect centered miRNA binding sites are common and can mediate repression of target mRNAs. *Genome Biol*, **15**, R51.
118. Jing, Q., Huang, S., Guth, S., Zarubin, T., Motoyama, A., Chen, J., Di Padova, F., Lin, S.C., Gram, H. and Han, J. (2005) Involvement of microRNA in AU-rich element-mediated mRNA instability. *Cell*, **120**, 623-634.
119. Helfer, S., Schott, J., Stoecklin, G. and Förstemann, K. (2012) AU-rich element-mediated mRNA decay can occur independently of the miRNA machinery in mouse embryonic fibroblasts and Drosophila S2-cells. *PLoS One*, **7**, e28907.
120. Salmena, L., Poliseno, L., Tay, Y., Kats, L. and Pandolfi, P.P. (2011) A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell*, **146**, 353-358.
121. Grimson, A., K., F.K., Johnston, W.K., Garrett-Engele, P., Lim, L.P. and Bartel, D.P. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell*, **27**, 91-105.
122. Stark, A., Brennecke, J., Bushati, N., Russell, R.B. and Cohen, S.M. (2005) Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. *Cell*, **123**, 1133-1146.
123. Jacobsen, A., Wen, J., Marks, D.S. and Krogh, A. (2010) Signatures of RNA binding proteins globally coupled to effective microRNA target sites. *Genome Res*, **20**, 1010-1019.
124. Saetrom, P., Heale, B.S., Snøve, O., Aagaard, L., Alluin, J. and Rossi, J.J. (2007) Distance constraints between microRNA target sites dictate efficacy and cooperativity. *Nucleic Acids Res*, **35**, 2333-2342.
125. Foat, B.C., Morozov, A.V. and Bussemaker, H.J. (2006) Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics*, **22**, e141-149.
126. Wu, R.Z., Chaivorapol, C., Zheng, J., Li, H. and Liang, S. (2007) fREDUCE: detection of degenerate regulatory elements using correlation with expression. *BMC Bioinformatics*, **8**, 399.
127. Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jungkamp, A.C., Munschauer, M. *et al.* (2010) PAR-CliP--a

method to identify transcriptome-wide the binding sites of RNA binding proteins. *J Vis Exp*.

128. Zou, J., Lippert, C., Heckerman, D., Aryee, M. and Listgarten, J. (2014) Epigenome-wide association studies without the need for cell-type composition. *Nat Methods*, **11**, 309-311.
129. Network, C.G.A. (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61-70.
130. Berriz, G.F., Beaver, J.E., Cenik, C., Tasan, M. and Roth, F.P. (2009) Next generation software for functional trend analysis. *Bioinformatics*, **25**, 3043-3044.
131. Lebedeva, S., Jens, M., Theil, K., Schwanhäusser, B., Selbach, M., Landthaler, M. and Rajewsky, N. (2011) Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR. *Mol Cell*, **43**, 340-352.
132. López de Silanes, I., Zhan, M., Lal, A., Yang, X. and Gorospe, M. (2004) Identification of a target RNA motif for RNA-binding protein HuR. *Proc Natl Acad Sci U S A*, **101**, 2987-2992.