Local ancestry corrects for population structure in *S. cerevisiae* genome-wide association studies

Liyang Diao and Kevin C. Chen

BioMaPS Institute for Quantitative Biology and Department of Genetics

Rutgers, The State University of New Jersey

Piscataway, NJ 08854, USA

Copyright 2012.
Running title:
Comparing GWAS Methods in Yeast

Key words:
Genome-wide association studies, Saccharomyces cerevisiae, mixed linear model, population structure, local ancestry

Corresponding author:
Kevin C. Chen
Rutgers, The State University of New Jersey
Dept. of Genetics and BioMaPS Inst. for Quantitative Biology
145 Bevier Road
Piscataway, NJ 08854, USA
Phone: 732 445 1027 ext 40055
Fax: 732 445 1147
kcchen@biology.rutgers.edu
ABSTRACT

Genome-wide association studies (GWAS) have become an important method for mapping the genetic loci underlying complex phenotypic traits in many species. A crucial issue when performing GWAS is to control for the underlying population structure, since this can lead to spurious associations. Population structure is a particularly important issue in non-human species since it is often difficult to control for population structure during the study design phase, so population structure must be corrected statistically after the data has been collected. It has not yet been established if GWAS is a feasible approach in *Saccharomyces cerevisiae*, an important model organism and agricultural species. We thus performed an empirical study of statistical methods for controlling for population structure in GWAS using a set of 201 phenotypic traits measured in multiple resequenced strains of *S. cerevisiae*. We complemented our analysis of real data with an extensive set of simulations. Our main result is that a mixed linear model using the local ancestry of the strain as a covariate is effective at controlling for population structure, consistent with the mosaic structure of many *S. cerevisiae* strains. We further studied the evolutionary forces acting on the GWAS SNPs and found that SNPs associated with variation in phenotypic traits are enriched for low minor allele frequencies, consistent with the action of negative selection on these SNPs. Despite the effectiveness of local ancestry correction, GWAS remains challenging in highly structured populations, such as *S. cerevisiae*. Nonetheless, we found that even after correcting for population structure, there is still sufficient statistical power to recover biologically meaningful associations.
INTRODUCTION

A key question in biology is to understand the nature of the nucleotides underlying the variation in complex phenotypic traits among individuals. Recent breakthroughs in genotyping and sequencing technologies have facilitated an explosion of genome-wide association studies (GWAS) resulting in the discovery of many new associations of SNPs with complex traits, especially in humans (HINDORFF et al. 2009). GWAS is also an increasingly important mapping approach in other non-human species, including plant species such as maize, rice and Arabidopsis [reviewed in (BRACHI et al. 2011)], mouse (PAYSEUR and PLACE 2007), Drosophila (MACKAY et al. 2012) and dog (TSAI et al. 2012). However, the feasibility of GWAS in S. cerevisiae, an important model organism and agricultural species, has not yet been fully studied.

A critical issue that needs to be addressed in any GWAS is the presence of underlying population structure in the individuals studied, since population structure generally leads to spurious associations (PRICE et al. 2010). Accordingly, a number of statistical techniques have been proposed to correct for population structure, including genomic control, which corrects the GWAS test statistics by a variance inflation factor that is constant over all SNPs (DEVLIN and ROEDER 1999; REICH and GOLDSTEIN 2001) and structured association (PRITCHARD et al. 2000b). An elegant approach is to use a linear model with ancestry components estimated from principal components analysis (PRICE et al. 2006) or mixture models (PRITCHARD et al. 2000a) as covariates. Other statistical approaches are also possible when there is pedigree structure underlying the data (THORNTON and McPEEK 2010).

Here we explored the feasibility of using GWAS to map complex traits in S. cerevisiae. We used whole genome resequencing data from multiple S. cerevisiae strains (LITI et al. 2009) and data for 201 phenotypic traits for a significant fraction of these strains (WARRINGER et al. 2011). Previous
population genetic studies of *S. cerevisiae* have made it clear that there is considerable population structure in this species (ELYASHIV et al. 2010; LITI et al. 2009; SCHACHERER et al. 2009). We took an empirical approach to comparing different statistical methods for correcting for population structure in GWAS. Specifically, we assessed the performance of each statistical method by the shape of the quantile-quantile (Q-Q) plot of observed vs. expected P-values. Throughout our analysis we considered only statistical methods based on linear models since these allow the use of population structure data as covariates in a natural and flexible way. We also complemented our empirical study with an extensive set of simulations to assess the performance of the different GWAS methods.
MATERIALS AND METHODS

Parameters for STRUCTURE:
To study the population structure in the *S. cerevisiae* strains, we used the program STRUCTURE which uses a Markov Chain-Monte Carlo approach to assign individuals to ancestral populations and simultaneously infer the allele frequencies of the ancestral populations (FALUSH et al. 2003; PRITCHARD et al. 2000a). The parameters for the preliminary short run of STRUCTURE used for SNP selection (see Methods for SNP Selection) were default parameters except for PLOIDY=1; BURNIN=5000; NUMREPS=5000; LINKAGE=1; ADMBURNIN=2500; SITEBYSITE=1. The long STRUCTURE runs used for the rest of the analysis used the same parameters except for BURNIN=50000; NUMREPS=50000; ADMBURNIN=25000.

Methods for SNP Selection
As described in the main text, 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.

**Sliding window:** We implemented a sliding window procedure in which we selected a small number of SNPs for each window of N consecutive SNPs. To select the SNPs, we calculated linkage disequilibrium (LD), measured by $D^2$, for each pair of SNPs in the window. 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 (LITU et al. 2009). However, the set of SNPs selected was essentially unchanged (data not shown).

**Sequential**: We implemented a sequential SNP selection procedure in which we identified LD blocks along a chromosome and kept one SNP per block. Starting with the first SNP in a chromosome (SNP A), we calculated LD between this SNP and the next SNP (SNP B). If SNP B was in high LD with SNP A (upper bound), 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).

**Methods for correcting for population structure in GWAS**
A general linear model (LM) without covariates does not correct for population structure and therefore serves as a baseline statistical method for comparison with more sophisticated GWAS methods. Several other linear models are commonly used for population correction, including a general linear model with whole-genome ancestry estimates as covariates and a mixed linear model (MLM) with the kinship matrix as a random effect, both with and without whole-genome ancestry covariates. There are several existing programs which implement these methods in an efficient way for genome-scale studies, including TASSEL (Yu et al. 2006) and EMMAX (Kang et al. 2010). TASSEL performs both LM and MLM, while EMMAX performs only 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 used is as follows:

**LM with covariates:**

1. Whole-genome ancestry covariates (obtained from STRUCTURE); implemented in R
2. Local-ancestry covariates (obtained from STRUCTURE); implemented in R
3. Both whole-genome ancestry and local ancestry covariates; implemented in R

**MLM with covariates, implemented in both EMMAX and TASSEL:**

4. Kinship matrix only (kinship matrix computed by EMMAX-KIN from the EMMAX package)
5. Whole-genome ancestry covariates and kinship matrix, implemented in EMMAX and TASSEL
6. Local ancestry covariates and kinship matrix, implemented in EMMAX
7. Local ancestry covariates, whole-genome ancestry covariates and kinship matrix, implemented in EMMAX
We denote whole-genome ancestry covariates by $Q$, following the notation used in STRUCTURE (PRITCHARD et al. 2000a), local ancestry covariates as $LA$ and the kinship matrix as $K$. The statistical methods are summarized below in Table 1.

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Statistical Method</th>
<th>Covariates</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-LM</td>
<td>LM</td>
<td>none</td>
</tr>
<tr>
<td>R-Q</td>
<td>LM</td>
<td>$Q$</td>
</tr>
<tr>
<td>R-LA</td>
<td>LM</td>
<td>$LA$</td>
</tr>
<tr>
<td>R-LAQ</td>
<td>LM</td>
<td>$LA + Q$</td>
</tr>
<tr>
<td>EMMAX-K</td>
<td>MLM</td>
<td>$K$</td>
</tr>
<tr>
<td>EMMAX-QK</td>
<td>MLM</td>
<td>$Q + K$</td>
</tr>
<tr>
<td>TASSEL-K</td>
<td>MLM</td>
<td>$K$</td>
</tr>
<tr>
<td>TASSEL-QK</td>
<td>MLM</td>
<td>$Q + K$</td>
</tr>
<tr>
<td>EMMAX-KLA</td>
<td>MLM</td>
<td>$LA + K$</td>
</tr>
<tr>
<td>EMMAX-KLAQ</td>
<td>MLM</td>
<td>$LA + K + Q$</td>
</tr>
</tbody>
</table>

Table 1. Summary of the GWAS methods used in this study.

**Computing global ancestry:**

We estimated the global ancestry for each strain by running STRUCTURE with the 3723 SNPs selected by the sequential SNP selection method (see Methods for SNP selection). To determine the most likely number of ancestral populations, we first ran STRUCTURE (see short run parameters in Parameters for STRUCTURE) by varying the value of $K$ - the number of populations - between 3 and 8. Our results indicated that the most likely number of populations is six, consistent with previous studies (LITI et al. 2009; SCHACHERER et al. 2009). We performed a longer run of STRUCTURE using the
linkage model with \( K = 6 \) to get a more refined analysis of the population structure (see long run parameters in Parameters for STRUCTURE).

**Computing local ancestry:**

STRUCTURE produces local ancestry estimates for each nucleotide based on a Hidden Markov Model (Falush *et al.* 2003). We removed SNPs in background LD using the sequential SNP selection scheme described above (see Methods for SNP selection). To verify the local ancestry estimates of STRUCTURE, we used the program WINPOP (Pasaniuc *et al.* 2009) to more carefully analyze the mosaic strains by using as ancestral populations the non-mosaic strains as determined by STRUCTURE, since we expect those lineages to be primarily of a single ancestry across the genome. For this purpose, the North American strains were grouped with the mosaic strains as previously determined by (Liti *et al.* 2009), as they exhibit widespread mosaicism. We used a recombination rate of \( r = 3.5 \times 10^{-6} \) following (Ruderfer *et al.* 2006). While the number of generations since admixture is unknown, one estimate of the number of outcrossing events since the MCRA of two particular *S. cerevisiae* strains is 314 (Ruderfer *et al.* 2006) and the estimate from STRUCTURE was 28 generations. Therefore 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.

**Computing the kinship matrix:**

We computed a kinship matrix for use in the EMMAX and TASSEL MLMs using the EMMAX-KIN program. We used PLINK (Purcell *et al.* 2007) to convert .ped and .map files to EMMAX-readable
.tped and .tmap files. These converted files were then input to EMMAX-KIN. We used the “Identity-
by-State” matrix option, following (Zhao et al. 2007).

**Identifying significant associations:**

We used two methods for multiple hypothesis testing correction. First, we used the Holm correction
which is a Family-Wise Error Rate (FWER) method, similar to the Bonferroni correction. Second, we
used the Bejamini-Hochberg correction, which is a False Discovery Rate (FDR) correction. We used a
threshold of 0.05 in both cases.

To identify genes corresponding to statistically significant SNPs, we used two methods. First, given a
significant SNP, if the SNP fell within a gene, we counted the gene as significant. Otherwise, we
treated both flanking genes as significant. Second, given a significant SNP, we considered the entire
linkage block containing the SNP and all genes which overlapped the linkage block were counted as
significant. Linkage blocks were determined with a similar procedure as the sequential SNP selection
method: starting with the SNP (SNP A), we computed pair-wise LD between the SNP and the next
SNP (SNP B). If LD between SNP A and SNP B were above the threshold $D'^2 > 0.95$ (this threshold
was chosen to be consistent with the section, *Methods for SNP selection*) then we considered SNP B to
be part of the LD block. We did the same for the SNPs before SNP A as well. We preferred the second
method since the manner in which the SNPs were chosen to represent each LD block involved some
randomness.

**Simulations:**

In the first set of simulations, we chose one known phenotype (specifically, Maltose 2% Growth Rate)
from (Warringer et al. 2011). We randomly chose a certain number of causal SNPs , N, from the
3723 SNPs used in our analysis. For each of the strains containing the major allele of each of the causal
SNPs, we added a certain fixed effect, V, to the phenotypic value for the trait. We set N to [3, 20, 100] and we set V to [1, 3], to explore a wide range of genetic architectures. In addition, we performed one simulation with N=3 and V=10 to simulate the case of a few causal SNPs of large effect. After all the effects attributed to these causal SNPs were assigned, we normalized the phenotypic data by subtracting the mean and dividing by the standard deviation across all strains. The mean and variance for the original Maltose 2% Growth Rate trait were 0.9347 and 0.5781 respectively.

For the second set of simulations, we used genotypic data from (SCHACHERER et al. 2009) which contains a larger set of 63 strains. To make the simulations comparable to our previous analyses using the strains from (LITI et al. 2009), we selected SNPs from the set of all 101343 SNPs reported by (SCHACHERER et al. 2009) using the same SNP selection procedure we used to select SNPs from the strains from (LITI et al. 2009). This procedure selected a total of 12,916 SNPs. To compute the local and global ancestry covariates, we ran STRUCTURE on this set of SNPs for K=6 ancestral populations, which is consistent with the number of populations described by (SCHACHERER et al. 2009) and also our previous analysis of the strains from (LITI et al. 2009). For the mixed linear models, we generated an identity-by-state kinship matrix based on the 12,916 SNPs selected using the EMMAX-KIN program. To simulate the phenotypic data, we first let phenotypes for all 63 strains be 0. Then we proceeded as with the first set of simulations by randomly selecting N causal SNPs and adding to the phenotypic values the phenotypic effect V. As before, we let N = [3,20,100] and V = [1,3], with an additional simulation of N=3 and V=10 to simulate the case of few causal SNPs with a very large effect.

For both types of simulations, 200 iterations were run for each pair of parameters, N and V. For each of these simulations, we ran all GWAS methods analyzed in our previous analyses. To compute the ROC (receiver operating characteristic) curves, for each pair of parameters and each simulation, we
determined the ranks of the N planted SNPs and then took the average ranks over all simulated data sets. That is, we took the average rank of the highest ranked planted SNP, then we took the average rank of the next highest ranked planted SNP etc. Finally we plotted these average ranks in the ROC curves.
RESULTS

We obtained whole genome resequencing data for 38 strains of *S. cerevisiae* (LITI *et al.* 2009) and processed the raw data as previously described (CHEN *et al.* 2010). We obtained phenotypic data for 35 of the strains in 67 different environmental conditions (WARRINGER *et al.* 2011). There were three measurements per environment – growth rate (population doubling time), adaptation (proliferation lag), and efficiency (population density change) – resulting in 201 total phenotypes for each strain.

**Selecting SNPs for population structure analysis:**

We investigated the population structure of the 35 *S. cerevisiae* strains using the program STRUCTURE (FALUSH *et al.* 2003; PRITCHARD *et al.* 2000a). An important issue that has not received much attention in the literature is how to select an appropriate set of SNPs when running STRUCTURE (MICLAUS *et al.* 2009). STRUCTURE requires SNPs that are somewhat linked but not so tightly linked that there is significant background linkage disequilibrium (LD) among the SNPs (i.e. LD present in the ancestral populations prior to admixture) (FALUSH *et al.* 2003). Indeed, we found that the results of STRUCTURE were sensitive to the choice of SNPs (Materials and Methods, Figure S1).

We thus performed a comparison of SNP spacing produced by the two SNP selection procedures – sliding window selection and sequential selection (Materials and Methods). We found that SNPs selected by the sliding window procedure were more uniformly distributed across the genome while sequentially selected SNPs tended to form clusters (Figure S2). Sequential SNPs were chosen in a manner consistent with the recombination landscape of the genome (MANCERA *et al.* 2008), with fewer SNPs chosen around the centromeres where recombination is low and more SNPs chosen around recombination hotspots, while windowed SNPs were more uniform across the genome (Figure S2). 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. To confirm the clustering result of STRUCTURE, we performed principal components analysis using the 3723 SNPs and found that the non-admixed strains found by STRUCTURE cluster together (Figure S3).

**Preliminary analysis of the phenotypic data:**

We started by computing the variance of each phenotype across all strains. We also measured the degree of population structure underlying each phenotype by performing GWAS using a general linear model with no correction for population structure (Methods and Materials). One empirical way to ascertain the existence of population structure in GWAS is to plot the P-values produced by a statistical method as a quantile-quantile (Q-Q) plot. A standard result states that under the null hypothesis (in this case, the hypothesis of no population structure), the P-values are uniformly distributed. Since population structure is expected to produce spurious associations when in fact there are none, in the presence of population structure the observed P-values will tend to be smaller than expected. We defined the amount of population structure remaining after statistical correction by taking the mean squared distance (MSD) between the points generated by the Q-Q plot and the corresponding point on the line $Y = X$ (i.e. the expected P-values).

The top ten conditions that are most genetically structured as measured by the MSD statistic are: Rapamycin 0.5µg/ml Adaptation, Rapamycin 1µg/ml Adaptation, pH 3.5 Adaptation, LiCl 150mM Efficiency, CuCl2 0.75mM Rate, CuCl2 0.375mM Rate, KCl 1.45M Rate, CoCl2 0.015mM Adaptation, Maltose 2% Rate and LiCl 225mM Efficiency. The average phenotypic variance for these ten conditions (0.77) was significantly higher than the average phenotypic variance for all 201 conditions (0.30; P-value 0.0036, 100,000 bootstrap replicates). This suggests that conditions with high levels of population structure (at the genotype level) also have more varied phenotypes. We repeated
the analysis with a different measure of population stratification, the variance inflation factor (Develin and Roeder 1999), instead of the MSD statistic and obtained a nearly identical result (data not shown).

There are a total of 67 types of environments, for example, environments with various levels of pH or glucose concentration. Five environment types were significantly structured compared to background (Table 2). We conclude that it is important to account for population structure when performing GWAS in *S. cerevisiae*.

<table>
<thead>
<tr>
<th>Environment</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rapamycin</td>
<td>0.0006</td>
</tr>
<tr>
<td>CuCl2</td>
<td>0.0024</td>
</tr>
<tr>
<td>LiCl</td>
<td>0.026</td>
</tr>
<tr>
<td>Kcl</td>
<td>0.084</td>
</tr>
<tr>
<td>pH</td>
<td>0.095</td>
</tr>
</tbody>
</table>

Table 2. Environment types that are significantly structured compared to the background. The level of population structure was measured by the MSD statistic from the Q-Q plot produced by a general linear model.

**Comparison of statistical methods for GWAS:**

To assess the performance of different statistical methods for correcting for population structure, we carried out GWAS analysis for all 201 phenotypes using the methods listed in Table 1 (Methods and Materials). We found a wide range of significant SNPs called by the different statistical methods, highlighting the importance of considering different statistical methods in GWAS (Table 3).
combined the best scoring method (according to the MSD statistic) for each phenotype into a meta-statistical method we will refer to as the “BEST” method.

Some SNPs were found to be statistically significant in multiple phenotypes and we call these “pleiotropic SNPs”. When removing pleiotropic SNPs, we still found a wide range of numbers of SNPs found by the different statistical methods (Table 3). The number of GWAS loci detected under the Holm correction, a commonly used Bonferroni-type multiple testing correction, was small (<2 per phenotype for the general linear model). This is consistent with the power of the study based on the number of strains and we discuss this point at greater length later (see Discussion).

<table>
<thead>
<tr>
<th>Method</th>
<th>Including duplicates</th>
<th>Unique SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Holm</td>
<td>BH</td>
</tr>
<tr>
<td>R-LM</td>
<td>389</td>
<td>2662</td>
</tr>
<tr>
<td>R-Q</td>
<td>249</td>
<td>1024</td>
</tr>
<tr>
<td>EMMAX-K</td>
<td>235</td>
<td>353</td>
</tr>
<tr>
<td>EMMAX-QK</td>
<td>68</td>
<td>97</td>
</tr>
<tr>
<td>R-LA</td>
<td>431</td>
<td>2154</td>
</tr>
<tr>
<td>R-LAQ</td>
<td>191</td>
<td>1261</td>
</tr>
<tr>
<td>EMMAX-KLA</td>
<td>637</td>
<td>1477</td>
</tr>
<tr>
<td>EMMAX-KLAQ</td>
<td>90</td>
<td>144</td>
</tr>
<tr>
<td>BEST</td>
<td>162</td>
<td>359</td>
</tr>
</tbody>
</table>
Table 3. Number of SNPs found significant across all phenotypes after correction for Family-Wise Error Rate (Holm) and False Discovery Rate (BH, for Benjamini-Hochberg), both at a threshold of 0.05. The multiple testing corrections were applied to each condition separately, not to all 201 phenotypes all together. SNPs that are found to be significant in multiple phenotypes are included multiple times in the “Including duplicates” columns but only once in the “Unique SNPs” columns.

We compared the P-values produced by the different statistical methods (Table 4) and found a strong Pearson correlation between the MLM methods implemented in EMMAX and TASSEL when only the kinship matrix K was used (R = 0.82). However, this correlation was much lower with the addition of the global ancestry covariate Q (R = 0.25). Additionally, we found that the methods R-LA and R-LAQ both exhibited relatively low correlations to the other methods. Surprisingly, R-LM, which used no covariates for correction, had a fairly high correlation with EMMAX-K. We are not sure why the EMMAX and TASSEL implementations of the MLM differ. To the best of our knowledge, the only difference between the two programs is the implementation of the compression method and this might be the reason for the different results.

<table>
<thead>
<tr>
<th></th>
<th>R-LM</th>
<th>R-Q</th>
<th>EMMAX-K</th>
<th>EMMAX-QK</th>
<th>EMMAX-KLA</th>
<th>EMMAX-KLAQ</th>
<th>EMMAX-KL</th>
<th>EMMAX-QKL</th>
<th>R-LA</th>
<th>R-LAQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-LM</td>
<td>1</td>
<td>0.1684</td>
<td>0.7341</td>
<td>0.2024</td>
<td>0.0753</td>
<td>0.0936</td>
<td>0.6397</td>
<td>0.1460</td>
<td>0.0850</td>
<td>0.0566</td>
</tr>
<tr>
<td>R-Q</td>
<td>1</td>
<td>0.2540</td>
<td>0.2847</td>
<td>0.1686</td>
<td>0.0557</td>
<td>0.3073</td>
<td>0.8103</td>
<td>0.2886</td>
<td>0.0989</td>
<td></td>
</tr>
<tr>
<td>EMMAX-K</td>
<td>1</td>
<td>0.3338</td>
<td>0.1374</td>
<td>0.1241</td>
<td>0.8153</td>
<td>0.2603</td>
<td>0.0977</td>
<td>0.0619</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EMMAX-QK</td>
<td>1</td>
<td>0.2493</td>
<td>0.4173</td>
<td>0.2888</td>
<td>0.2542</td>
<td>0.1190</td>
<td>0.0492</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EMMAX-KLA</td>
<td>1</td>
<td>0.2683</td>
<td>0.1084</td>
<td>0.1289</td>
<td>0.2208</td>
<td>0.1027</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EMMAX-KLAQ</td>
<td>1</td>
<td>0.0869</td>
<td>0.0408</td>
<td>0.0849</td>
<td>0.0738</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TASSEL-K</td>
<td>1</td>
<td>0.3425</td>
<td>0.1039</td>
<td>0.0630</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
For each GWAS method and phenotype, we generated a Q-Q plot (Figure 1). Using the MSD statistic, we found that the EMMAX-KLA method was the most effective overall at correcting for population structure, followed by R-LAQ (Table 5; Table S1; Figure 2). Our data highlights the importance of correcting for the local ancestry of a strain when performing GWAS, not just the global ancestry.

<table>
<thead>
<tr>
<th>Statistical method</th>
<th>No. of phenotypes where it performed best</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-LM</td>
<td>11</td>
</tr>
<tr>
<td>R-Q</td>
<td>22</td>
</tr>
<tr>
<td>R-LA</td>
<td>14</td>
</tr>
<tr>
<td>R-LAQ</td>
<td>49</td>
</tr>
<tr>
<td>EMMAX-K</td>
<td>36</td>
</tr>
<tr>
<td>EMMAX-QK</td>
<td>5</td>
</tr>
<tr>
<td><strong>EMMAX-KLA</strong></td>
<td><strong>62</strong></td>
</tr>
<tr>
<td>EMMAX-KLAQ</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 5. For each statistical method, the number of phenotypes where it performed best is shown.

The results from the GWAS analysis reported above used the set of 3723 SNPs produced by our sequential SNP selection procedure (Methods). While removing SNPs in high LD is not required for GWAS, it reduces the multiple hypothesis testing burden while retaining most of the statistical power.
to detect associations - a similar strategy is used when selecting tag SNPs in human GWAS. To test the robustness of our choice of SNPs for the GWAS analysis, we selected an expanded set of 15,812 SNPs by choosing one SNP for each window of 10 SNPs. We then performed GWAS with the linear model with no covariates for the six copper tolerance phenotypes. While using the expanded set of SNPs resulted in more SNPs declared significant, the end result was similar because of the higher multiple testing burden for the expanded set of SNPs (data not shown).

We repeated our analyses using the variance inflation factor (VIF) (Devin and Roeder 1999) as a measure of the degree of population stratification instead of the MSD statistic. We note that the mean of the test statistics has been proposed as a good estimator for the variance inflation factor (Devin et al. 2004; Reich and Goldstein 2001) so our MSD statistic has some basis in formal statistical theory. Overall EMMAX-KLA still performed better than the other GWAS methods under the VIF statistic (data not shown). We also checked the genes found statistically significant by the “BEST” method where we now have two “BEST” methods: “BEST-VIF” and “BEST-MSD”. After FDR multiple testing correction, “BEST-MSD” found 58 conditions with significant genes while “BEST-VIF” found 48. In about half the conditions, there were identical sets of significant genes. We observed very similar results for the Holm correction (data not shown). Thus our use of the MSD statistic did not affect our overall conclusions.

Functional analysis of GWAS SNPs:

For the remaining analyses, we took the best correction method for each phenotype and used the P-values generated by that method. We refer to this meta-statistical method as the “BEST” method. To investigate the functional significance of the statistically significant SNPs, we examined the fraction of significant SNPs contained in genes. We found that 75% of SNPs called statistically significant under
the FDR correction were in genes compared to 63% of all SNPs used in our analysis (P-value < 1e-4).

We also examined the biological functions of the GWAS SNPs using the GO term enrichment program FuncAssociate (BERRIZ et al. 2003). We found several interesting enriched functions, including biotin biosynthesis for the phenotype pH 3.5 adaptation (Fisher 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 exact test, uncorrected P-value 0.003) and oligosaccharide metabolic process for the phenotype Glucose 0.5% rate (Fisher 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 previously detected associations (CUBILLOS et al. 2011; WARRINGER et al. 2011). Those papers 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 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 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 (Table S2). 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 the ENA1 gene 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 1743bp 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 (Table S2). We conclude that the GWAS methods are often able to recover previously known associations.

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 of the SNPs which 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 (Chen et al. 2010).

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). These results were robust whether we used an FWER or FDR multiple testing correction method and whether we considered non-pleiotropic 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.
Table 6. Minor allele frequencies for non-pleiotropic and pleiotropic SNPs found by the BEST method (i.e. 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.

Simulation study of *S. cerevisiae* GWAS:

To complement our GWAS analysis on real data, we performed two sets of simulations over a wide range of genetic architectures, one set based on the existing phenotypic data from (WARRINGER et al. 2011) and one set with purely simulated phenotypes (Methods). To determine the performance of each GWAS method, we plotted ROC curves to compare the average ranks of the randomly chosen causal SNPs (Methods). In general, all the GWAS methods performed better in the simulations with simpler genetic architectures (i.e. fewer causal SNPs of larger effect). All methods performed close to random in the simulations involving 100 SNPs (Figure S4) but they performed much better with three SNPs and large phenotypic value (Figure S5). The purely simulated phenotypes generally performed better than the phenotypes based in the existing phenotype (Figure S6), perhaps due to the larger sample size in the analyses involving the dataset of (SCHACHERER et al. 2009). These simulation results are consistent with the recent results of (CONNELLY and AKEY 2012).
In most cases, EMMAX-K performed the best in terms of the ROC curves, having a lower rate of false positives. To address why the methods based on local ancestry appear to be performing worse according to the ROC curves, we computed the variance of the ranks in the simulations. We observed that the methods based on local ancestry have higher variance (data not shown), presumably because when randomly selecting SNPs to be causal SNPs, some of the SNPs chosen were correlated with local ancestry so their effect was inadvertently corrected by the GWAS method. This raises an important broader point: while we have concentrated on trying to reduce false positive associations so far, any such method will necessarily also reduce the power of the statistical approach. Nonetheless, we have shown above that the statistical power of our local ancestry approach is not extremely low because it still is able to recover biologically meaningful associations previously published in (CUBILLOS et al. 2011; WARRINGER et al. 2011).
DISCUSSION

Genome-wide association studies (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, \textit{S. cerevisiae}. Our main results are that GWAS is indeed a feasible approach in \textit{S. cerevisiae} and that it is important to take into account the local ancestry of an \textit{S. cerevisiae} strain when performing GWAS. At a practical level, the EMMAX mixed linear model implementation (KANG et al. 2010) using an Identity-by-State kinship matrix as a random effect and local ancestry inferred by STRUCTURE (FALUSH et al. 2003) as a fixed effect performed best in our experiments. Importantly, our work also shows that existing methods for detecting local ancestry, such as STRUCTURE (FALUSH et al. 2003) and WINPOP (PASANIUC et al. 2009), are effective in \textit{S. cerevisiae}, at least for the purposes of GWAS. Nonetheless, the demographic history of \textit{S. cerevisiae} is complex (LITI et al. 2009; SCHACHERER et al. 2009) 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 (VERDU and ROSENBERG 2011).

There are many differences between performing GWAS in \textit{S. cerevisiae} and humans. First, the burden of multiple hypothesis testing correction is much lower in \textit{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 \textit{S. cerevisiae} GWAS to be far more powerful than a comparable GWAS in humans. Second, the extent of linkage
disequilibrium is much shorter 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, compared to testing a tag SNP (OHASHI and TOKUNAGA 2001). 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, that are impossible or unethical 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 (EHRENREICH *et al.* 2009). 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* (CUBILLOS *et al.* 2011; WARRINGER *et al.* 2011) are promising in this regard. A recent study of GWAS in *S. cerevisiae* also found similar results to our study (CONNELLY and AKEY 2012). In particular, they showed through
simulations on the same set of *S. cerevisiae* strains that GWAS in *S. cerevisiae* 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 2012) 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. 2009) 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 (~12Mb), the decreasing cost of DNA sequencing, and the relative tractability of high-throughput phenotyping in yeast (Ohyu et al. 2005).

In addition there are many other studies of GWAS in other model organisms that are similar to our study, including studies in mouse (Payseur and Place 2007), Arabidopsis (Zhao et al. 2007), maize and rice (Brachi et al. 2011), tomato (Ranc et al. 2012), dog (Tsai et al. 2012) and *Drosophila melanogaster* (Mackay et al. 2012). 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 (Pool 2012). 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 (Seldin et al. 2011; Shrinker et al. 2011). Thus we believe that our results will also be useful for GWAS analyses in humans and other model systems as well.
ACKNOWLEDGEMENTS

We thank Steve Buyske for comments on the manuscript, Jonathan Flowers for advice on running the STRUCTURE program, David Gould for assistance with running the WINPOP program and Jonas Warringer for giving us the yeast phenotype data. This work was partially funded by the National Institutes of Health (R00HG004515 to KC).


PRICE, A., N. PATTERSON, R. PLENGE, M. WEINBLATT, N. SHADICK et al., 2006 Principal components
REICH, D., and D. GOLDSTEIN, 2001 Detecting association in a case-control study while correcting for population stratification. Genet Epidemiol 20: 4-16.
### Table S1

This table gives summary statistics of the mean squared distance (MSDs) of all GWAS methods used over all phenotypes. This data is also shown in Figure 2.

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean</th>
<th>Median</th>
<th>Std dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-LM</td>
<td>0.010</td>
<td>0.0060</td>
<td>0.012</td>
</tr>
<tr>
<td>R-Q</td>
<td>0.0040</td>
<td>0.0016</td>
<td>0.0063</td>
</tr>
<tr>
<td>EMMAX-K</td>
<td>0.0028</td>
<td>0.0017</td>
<td>0.0032</td>
</tr>
<tr>
<td>EMMAX-QK</td>
<td>0.0073</td>
<td>0.0040</td>
<td>0.0088</td>
</tr>
<tr>
<td>R-LA</td>
<td>0.0065</td>
<td>0.0024</td>
<td>0.0100</td>
</tr>
<tr>
<td>R-LAQ</td>
<td>0.0036</td>
<td>0.0014</td>
<td>0.0058</td>
</tr>
<tr>
<td>EMMAX-KLA</td>
<td>0.0031</td>
<td>0.00095</td>
<td>0.0067</td>
</tr>
<tr>
<td>EMMAX-KLAQ</td>
<td>0.015</td>
<td>0.010</td>
<td>0.014</td>
</tr>
<tr>
<td>TASSEL-K</td>
<td>0.0038</td>
<td>0.0027</td>
<td>0.0035</td>
</tr>
<tr>
<td>TASSEL-QK</td>
<td>0.0050</td>
<td>0.0024</td>
<td>0.0076</td>
</tr>
</tbody>
</table>

Table S2. This table shows a comparison between the SNPs found by our GWAS analysis and the previously published causal SNPs in (CUBILLOS et al. 2011; WARRINGER et al. 2011). For each condition and each GWAS method we list (1) the number of SNPs found and (2) what the fraction of all significant SNPs at a nominal P-value threshold of 0.05 were associated with the phenotype. The highest ranking method according to criterion (2) is shown in bold.
Figure S1. This figure shows a comparison of representative runs of the program, STRUCTURE, using SNPs selected according to the windowed (top) and sequential (bottom) methods. Although some populations, such as the Malaysian (MA) and West African (WA) strains, separate well in both cases, the European and mosaic strains are effectively indistinguishable when using the windowed SNPs.

Figure S2. This figure shows that 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 centromere position, bright green rectangles represent crossover recombination hotspots, cyan rectangles represent non-crossover recombination hotspots, and orange represents overall recombination hotspots.

Figure S3. Each *S. cerevisiae* strain is plotted on the first two principal components, after performing principal components analysis on the genotype matrix.

Figure S4. ROC curve of simulations with 100 causal SNPs each adding phenotypic value 1.

Figure S5. ROC curve of simulations with 3 causal SNPs each adding phenotypic value 10.

Figure S6. ROC curve of simulations with 3 causal SNPs, each adding phenotypic value 10, with a base phenotypic value given by one trait from (WARRINGER *et al.* 2011).