Assumption of Existing Approximation Methods for Multiple Testing Correction

Figure S1 illustrates the correlation structures of genotype data in yeast segregants (BREM *et al.* 2005) (Figure S1(a-b)), mouse inbred strains (MCCLURG *et al.* 2007) (Figure S1(c-d)), and rare variants in human population (FRAZER *et al.* 2007) (Figure S 1(e-f)). More details of these three datasets have been discussed in the Results Section. The yeast segregants data is a typical genetic dataset from a cross of inbred strains where markers within a chromosome are highly correlated (Figure S1 (a)) and form an approximate banding structure (Figure S1 (b)). Thus this dataset satisfies the assumptions needed for the approximation methods. In contrast, we do not observe such correlation structure in the data of mouse inbred strains and human rare variants. Both mouse inbred strains and human rare variants data are commonly encountered in genetic studies. The insufficiency of the approximation methods for these datasets motivates the recent development of the *exhaustive* methods that calculate the exact resampling-based p-values.

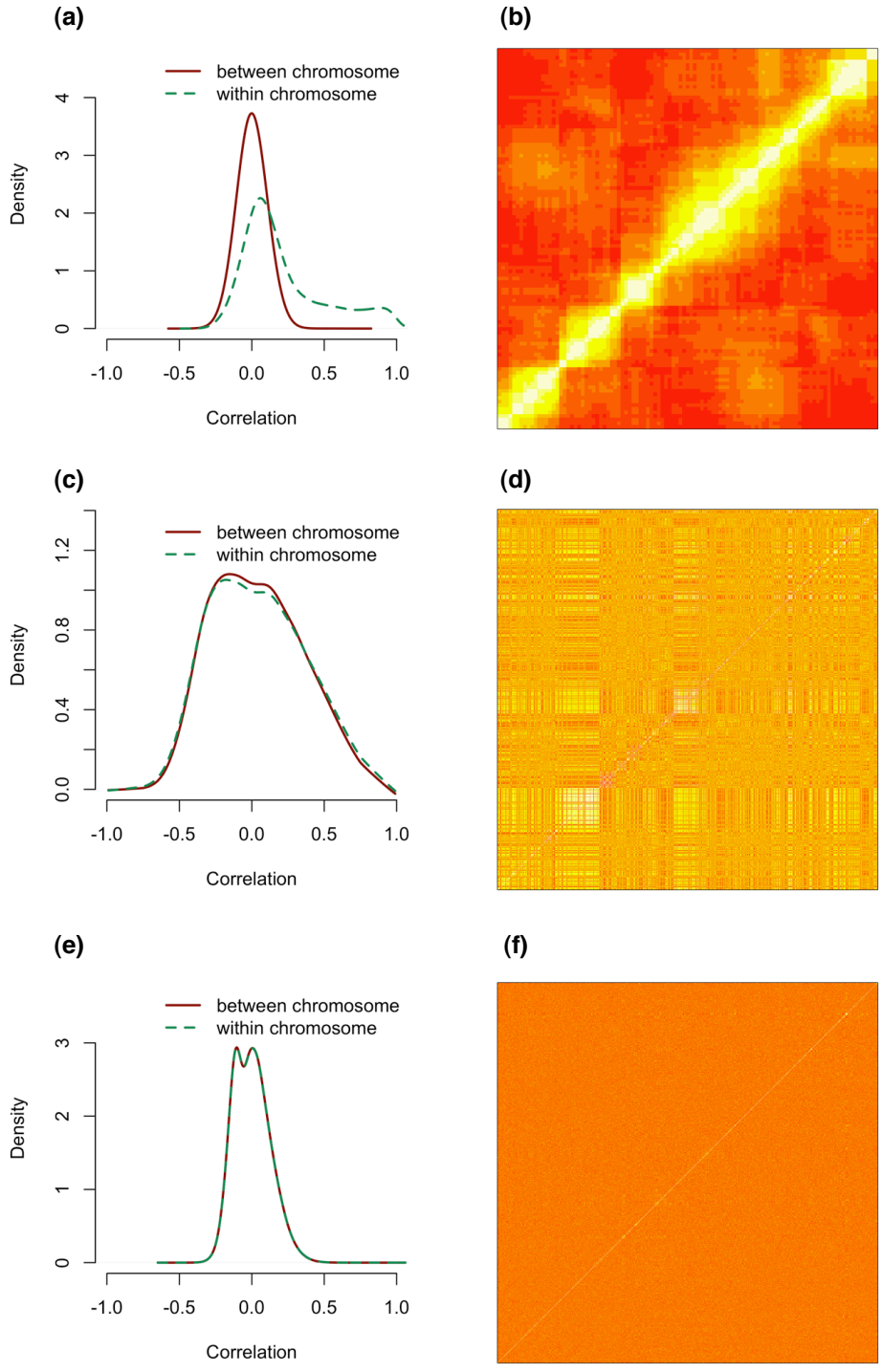


Figure S1: Difference between correlation structures in the yeast [(a) and (b)], inbred mouse [(c) and (d)], and human rare variant [(e), (f)] data sets. (a), (c), (e) compare the correlation density for marker pairs within and between chromosomes. (b), (d), and (f) are the heat maps of correlation matrices in chromosome 12 of the three data sets.

Convexity of Commonly Used Statistical Tests

It has been shown that most of the commonly used statistical tests in eQTL studies, such as Pearson's correlation, Student's t-test, analysis of variance (ANOVA F-test), and likelihood ratio test are equivalent for binary genotype data (GATTI *et al.* 2009). Without loss of generality, we show that the ANOVA F-test is a convex function of $\bar{Y}_1$. Recall that $\bar{Y}_1$ is defined as follows: for SNP X_n and a resampled phenotype Y_m^k, $\bar{Y}_1$ represents the sum of the phenotype values of the individuals with rarer alleles (i.e., when X_n equals to 1).

The ANOVA F-test partitions the total sum of squares SS_T into a between-group sum of squares SS_B and a within-group sum of squares SS_W. The F-statistic is $F = cSS_B/SS_W$, where c is a fixed constant for a particular study. Let SS_T be the total sum of squares. We have that $F = cSS_B/SS_W = cSS_B/(SS_T - SS_B)$. For a given resampled phenotype vector Y_m^k, the F-statistic is a monotone function of SS_B. From now on, we will use SS_B as our test statistic. For SNP X_n and resampled phenotype vector Y_m^k,

$$SS_B(X_n, Y_m^k) = \frac{\bar{Y}_0^2}{S_0} + \frac{\bar{Y}_1^2}{S_1} - \frac{\bar{Y}^2}{S},$$

where $\bar{Y}_0$ and $\bar{Y}_1$ are the sums of the phenotype values in Y_m^k when X_n equals to 0 and 1, respectively, S_0 and S_1 are the numbers of 0's and 1's in X_n, respectively, $\bar{Y}$ is the sum of all phenotype values in Y_m^k, and S is the total number of individuals. Clearly, $\bar{Y}_0 + \bar{Y}_1 = \bar{Y}$, $S_0 + S_1 = S$, and thus we can rewrite SS_B as

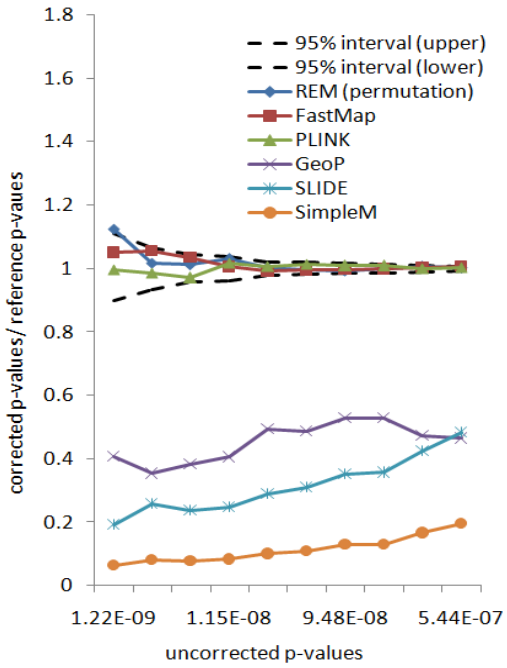
$$SS_B(X_n, Y_m^k) = \frac{(\bar{Y} - \bar{Y}_1)^2}{S - S_1} + \frac{\bar{Y}_1^2}{S_1} - \frac{\bar{Y}^2}{S}. \tag{1}$$

Clearly, $SS_B(X_n, Y_m^k)$ is a convex function (more specifically a quadratic function) of $\bar{Y}_1$.

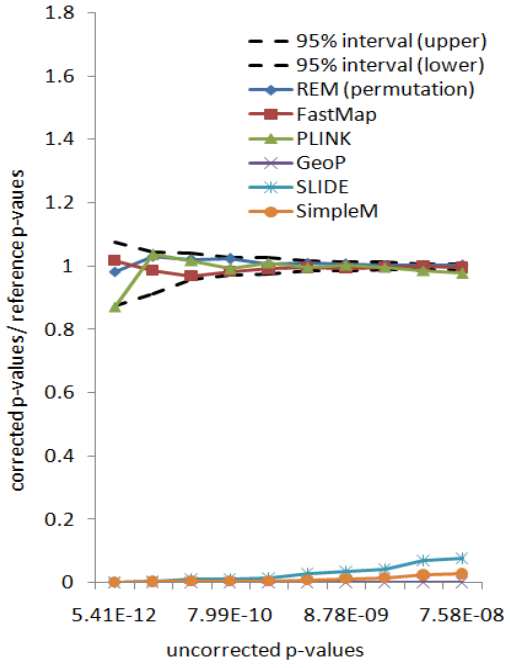
Accuracy Evaluation using Synthetic Phenotypes

To further study the accuracy of the selected methods on the inbred mouse data set, we generate three synthetic phenotypes whose values follow standard normal, exponential, and uniform distributions. For each synthetic phenotype, we use the selected methods to correct the p-values. The uncorrected p-values ranges from 1.2×10^{-9} to 5.4×10^{-7} (normal), 5.4×10^{-12} to 7.6×10^{-8} (exponential), and 2.7×10^{-11} to 2.1×10^{-7} (uniform). After correction (by 100M permutations), the p-values range from 0.00039 to 0.052 (normal), 0.00044 to 0.053 (exponential), and 0.00026 to 0.05 (uniform). Then we apply different methods to estimate the corrected p-values.

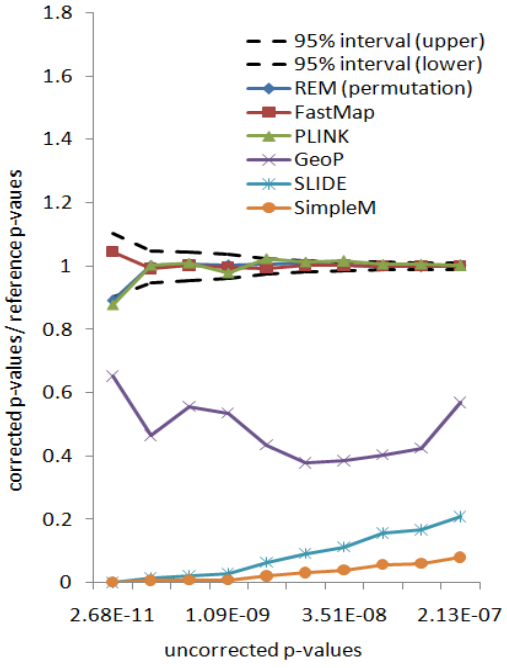
Figures S2(a), S2(b), and S2(c) show the results when using the permutation p-values as reference. From these figures, we can observe a similar trend using the three synthetic datasets to that using the real expression traits data. The approximation methods are anti-conservative. Moreover, they do not provide accurate estimation for the permutation p-values. Their performances vary for phenotypes with different distributions. GeoP does not work for exponentially distributed phenotypes (with all corrected p-values being 0), though it performs better than SLIDE and SimpleM on the other two distributions. This demonstrates that the distribution of the phenotypes plays an important role in the performances of the approximation methods.



(a) Synthetic normally distributed trait



(b) Synthetic exponentially distributed trait

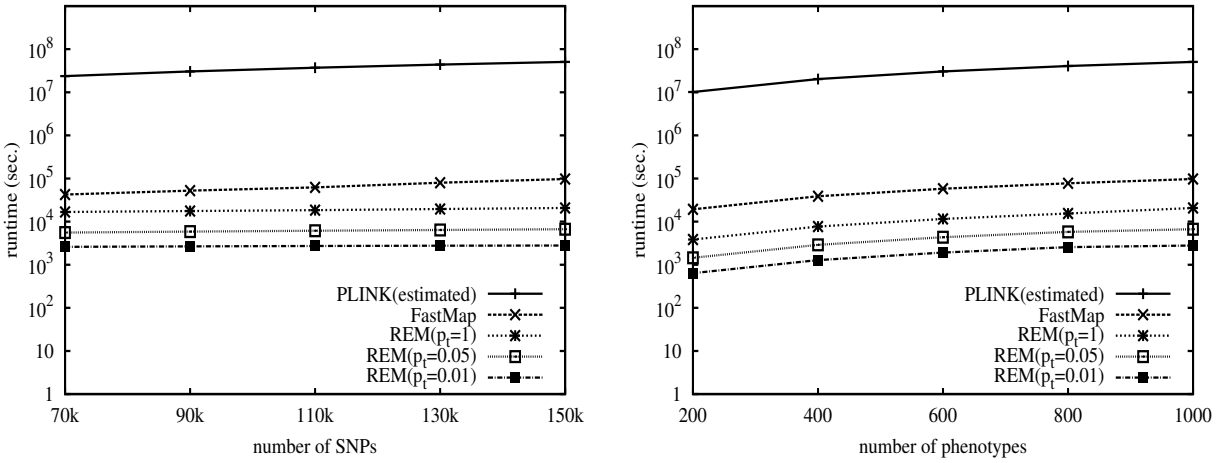


(c) Synthetic uniformly distributed trait

Figure S2: Accuracy evaluation of selected methods on synthetic gene expression traits using inbred mouse data set. (Each line represents the ratio between the corrected p-values and the reference p-values for a method. The reference p-values are obtained using 100M permutations. An accurate method should yield a ratio of 1. In this figure, the reference p-values are estimated by permutation test.)

Computational Efficiency Evaluation when Varying the Size of the Data Set

We randomly sample 1K real gene expression traits for the evaluation. Unless otherwise specified, the default experimental setting is as follows: number of SNPs = 150K, number of traits = 1K, and number of resamplings = 100K. PLINK is not computationally efficient enough for this setting. However, since its runtime is linear to the number of resamplings, we estimate its runtime for 100K resamplings by first running it with 100 resamplings, and then multiplying the runtime by 1000. We examine the runtimes of REM for three different thresholds of corrected p-values, 1, 0.05, and 0.01. When the threshold is set to be 1, REM will find the corrected p-values for all traits. Otherwise REM automatically finds the traits whose corrected p-values are less than the threshold. As shown in Figure S3, FastMap is about two orders of magnitude faster than PLINK. REM further improves the computational efficiency by about two orders of magnitude. The computational efficiency of REM is dramatically improved when the corrected p-value threshold decreases, because REM can filter out insignificant traits in a very early stage of the process.



(a) Varying the number of SNPs

(b) Varying the number of traits

Figure S3: Efficiency evaluation of three exact methods, PLINK, FastMap, and REM, when varying the number of SNPs and the number of traits in the mouse data set. The y-axis (runtime) is in logarithmic scale. The runtime of PLINK is estimated based on small scale experiments. See text for more details.

Pseudo Code of the REM Algorithm

Algorithm S1: REM - Rapid and Exact Multiple testing correction by resampling

Input: SNPs $\{X_1, X_2, \cdots, X_N\}$, gene expression traits $\{Y_1, Y_2, \cdots, Y_M\}$, number of resamples K, and desired resampling-based p-value threshold p_t.

Output: Significant gene expression traits, i.e., the ones whose resampling-based p-values are no greater than p_t.

```
1  index SNPs {X_1, X_2, ..., X_N} by the two-layer indexing structure;
2  for every Y_m (1 ≤ m ≤ M) do
3      scan all SNPs to find maximum statistic T_{Y_m};
4      generate resampled phenotype vectors {Y_m^1, Y_m^2, ..., Y_m^K};
5      count = 0;
6      p_res(Y_m) = (count+1)/(K+1);
7      for every Y_m^k (1 ≤ k ≤ K) do
8          for every e_1i (e_1i is a first layer entry) do
9              if ub(e_1i) > T_{Y_m} then
10                 for every e_2j (e_2j is a second layer entry of e_1i) do
11                     if ub(e_2j) > T_{Y_m} then
12                         for every X_n in entry e_2j do
13                             if T(X_n, Y_m^k) > T_{Y_m} then
14                                 count = count + 1;
15                                 p_res(Y_m) = (count+1)/(K+1);
16                                 goto line 23;
17                             end
18                         end
19                     end
20                 end
21             end
22         end
23         if p_res(Y_m) > p_t then
24             goto line 2;
25         end
26     end
27     return Y_m as significant;
28 end
```

Time Complexity of REM

Supposed that we have S individuals, N SNPs, M Phenotypes, and K Permutations/bootstrapps. In Line 1 of Algorithm S1, the overall time complexity for indexing the SNPs is $O(NS)$. In Line 9, the total number of upper bounds in the first layer we need to check is $(S/2)^2/2$. The complexity of each check is $O(1)$. So the overall time complexity for searching the first layer is $O(KMS^2)$. In Line 11, in the worse case, each first layer entry has $O(S^2)$ second layer entries. However, the total number of secondary entries cannot be larger than the total number of SNPs N. Thus, the worst case time complexity for searching the second layer is $O(KMN)$. Moreover, in practice, for a first layer entry, a second layer indexing is only needed when its number of SNPs is larger than the possible number of second layer entries. Only a small portion of the first layer entries will actually have the second layer indexing. The overall time complexity of REM is $O(NS + KMS^2 + KMN)$.

Note that the complexity analysis only provides an asymptotic description of the worst case performance of the algorithm. The actual performance of the algorithm heavily depends on the tightness of the upper bound, which has been demonstrated by extensive experimental evaluation.

Applying REM to Large Sample Study through Meta-Analysis

REM can be effectively applied to large sample study through meta-analysis. In a meta-analysis, the samples are partitioned into several groups. The resampling-based p-values within each group are calculated using REM. The p-values are then combined by applying Fisher's method (FISHER 1925).

We simulate data sets of large samples to demonstrate the efficacy of REM. We use SNPs in chromosome 22 of 1000 randomly selected individuals from the genome-wide association study of Schizophrenia (SHI *et al.* 2009). At each locus, the heterozygous genotype is combined with the homozygous genotype of major allele. There are 6,679 SNPs of MAF no less than 0.05. The phenotypes Y are simulated by a linear model: $Y = Xb + \epsilon$, where X is the genotype of a SNP, b is the coefficient, and ϵ is the residual error. In our experiments, b varies from 0.3 to 0.7, and ϵ follows a standard Gaussian distribution.

The square of Pearson's correlation, R^2, is used as the test statistic. We denote the maximum R^2 of the original phenotype to be r_0^2. The permutation p-value across the 1000 individuals is calculated as the proportion of the permutations with maximum R^2 larger than r_0^2.

For meta-analysis, we randomly partitioned the data into two groups, each of which has 500 samples. In each group, we calculated the permutation p-value as the proportion of the permutations whose maximum R^2 are larger than fr_0^2, where f is a constant. We then apply Fisher's method to combine the permutation p-values of every group to obtain the meta permutation p-value.

We repeat the above simulation 1000 times and compare the permutation p-values from the whole group (the 1000 individuals) to the meta permutation p-values. Figure S4 depicts that they are highly correlated. Specifically, when the factor f equals to 2.0 (red points in the figure), the correlation is 0.99.

Meta-analysis using 5 and 10 groups are also performed. The results are similar to that of 2 groups. In particular, the correlation is 0.98 for 5 groups, and 0.96 for 10 groups.

The nearly perfect correlation between the permutation p-value (of the whole group) and the meta permutation p-value enables us to apply REM to studies with large samples effectively. Specifically, we first partition the samples into smaller groups and apply REM to get the permutation p-values for each group. We then combine these p-values by the Fisher's method to get a meta permutation p-value. Finally, we map the meta permutation p-value to the original permutation p-value following the estimated relationship (e.g., the line corresponding to factor 2 in Figure S4) between these two values. Note that the relationship between the original permutation p-value and the meta permutation p-value can be estimated by using a small number (e.g., tens) of phenotypes.

The exact value of the factor f is not essential to the success of our method. This is because, for any given factor f, we can always estimate the relationship between the meta permutation p-value and the original permutation p-value. For example, the three lines in Figure S4 correspond to three different values of f. Any one of them can be used to

map between the two p-values. In theory, however, it is interesting to investigate whether there exists an optimal f value that gives the highest correlation between the two p-values. This is among our future research directions.

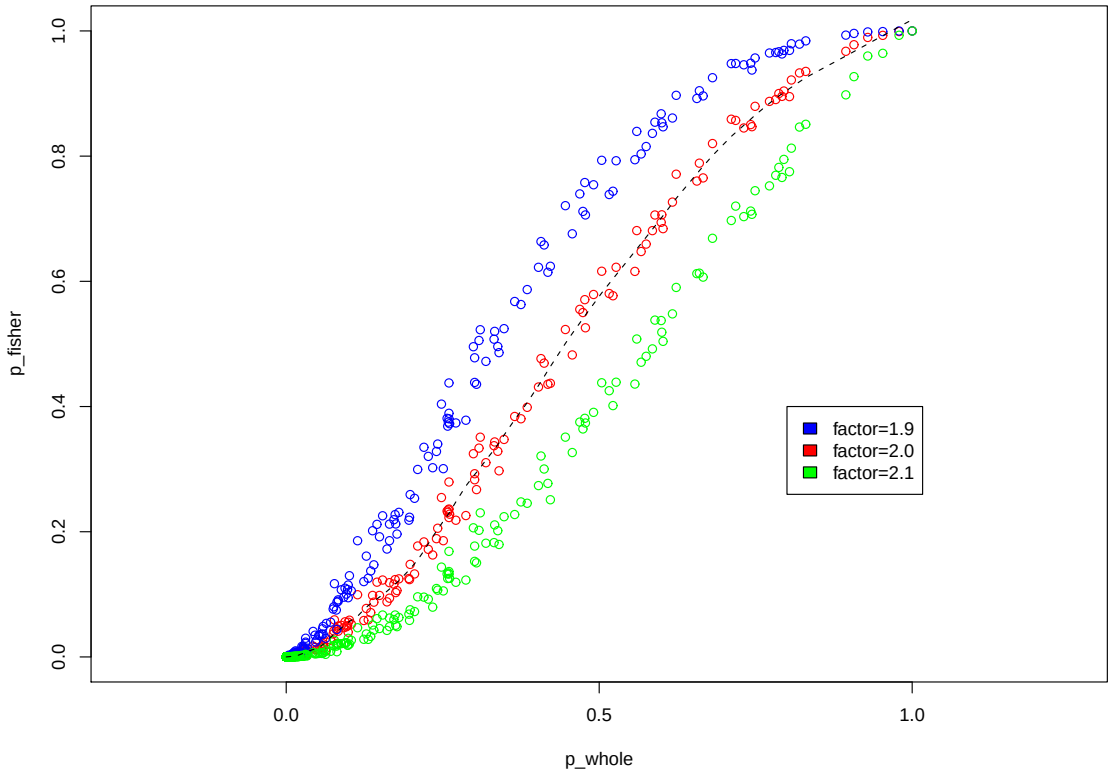


Figure S4: Relationship between the combined p-values and the original p-values

LITERATURE CITED

BREM, R. B., J. D. STOREY, J. WHITTLE, and L. KRUGLYAK, 2005 Genetic interactions between polymorphisms that affect gene expression in yeast. Nature **436(7051):** 701–703.

FISHER, R., 1925 *Statistical Methods for Research Worker*. Oliver and Boyd (Edinburg).

FRAZER, K., D. BALLINGER, D. COX, D. HINDS, L. STUVE, R. GIBBS, J. BELMONT, A. BOUDREAU, P. HARDENBOL, S. LEAL, and OTHERS, 2007 A second generation human haplotype map of over 3.1 million SNPs. Nature *449*(7164)**:** 851–861.

GATTI, D. M., A. A. SHABALIN, T.-C. LAM, F. A. WRIGHT, I. RUSYN, and A. B. NOBEL, 2009 FastMap: Fast eQTL mapping in homozygous populations. Bioinformatics **25(4):** 482–489.

MCCLURG, P., J. JANES, C. WU, D. DELANO, J. WALKER, S. BATALOV, J. TAKAHASHI, K. SHIMOMURA, A. KOHSAKA, J. BASS, T. WILTSHIRE, and A. SU, 2007 Genomewide Association Analysis in Diverse Inbred Mice: Power and Population Structure. Genetics **176(1):** 675–683.

SHI, J., D. LEVINSON, J. DUAN, A. SANDERS, Y. ZHENG, I. PE'ER, F. DUDBRIDGE, P. HOLMANS, A. WHITTEMORE, B. MOWRY, A. OLINCY, F. AMIN, C. CLONINGER, J. SILVERMAN, N. BUCCOLA, W. BYERLEY, D. BLACK, R. CROWE, J. OKSENBERG, D. MIREL, K. KENDLER, R. FREEDMAN, and P. GEJMAN, 2009 Common variants on chromosome 6p22.1 are associated with schizophrenia. Nature **460(7256):** 753–757.