Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots

Hyun Min Kang*,**, Chun Ye†**, and Eleazar Eskin‡,§

* Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA 92093

†Bioinformatics Program, University of California San Diego, La Jolla, CA 92093

‡Department of Computer Science, University of California, Los Angeles, CA 90095

§Department of Human Genetics, University of California, Los Angeles, CA 90095

** These authors equally contributed to this work.
Accurate discovery of eQTLs

Corresponding author:
Eleazar Eskin
Department of Computer Science and Department of Human Genetics
University of California, Los Angeles
4732 Boelter Hall
Los Angeles, CA 90095
Telephone: (310) 206-4490
Fax: (310) 825-2273
E-mail: eeskin@cs.ucla.edu
Abstract

In genome wide mapping of expression quantitative trait loci (eQTL), it is widely believed that thousands of genes are \textit{trans}-regulated by a small number of genomic regions called “regulatory hotspots”, resulting in “\textit{trans}-regulatory bands” in an eQTL map. As several recent studies have demonstrated, technical confounding factors such as batch effects can complicate eQTL analysis by causing many spurious associations including spurious regulatory hotspots. Yet little is understood about how these technical confounding factors affect eQTL analyses and how to correct for these factors. Our analysis of datasets with biological replicates suggests that it is this inter-sample correlation structure inherent in expression data that leads to spurious associations between genetic loci and a large number of transcripts inducing spurious regulatory hotspots. We propose a statistical method that corrects for the spurious associations caused by complex inter-sample correlation of expression measurements in eQTL mapping. Applying our Inter-sample Correlation Emended (ICE) eQTL mapping method to mouse, yeast, and human identifies many more \textit{cis} associations while eliminating most of the spurious \textit{trans} associations. The concordances of \textit{cis} and \textit{trans} associations have consistently increased between different replicates, tissues, and populations; demonstrating the higher accuracy of our method to identify real genetic effects.
INTRODUCTION

Genome wide analysis of gene expression data in segregating populations has been widely conducted to understand the genetic basis of regulation in many organisms including yeast (Brem and Kruglyak, 2005), Arabidopsis (Keurentjes et al., 2007), mouse (Chesler et al., 2005; Bystrykh et al., 2005) and human (Cheung et al., 2005; Stranger et al., 2007). In order to understand the complex regulatory network, numerous statistical analysis methods have been proposed including clustering of co-regulated genes (Yvert et al., 2003), multipoint linkage analysis (Storey et al., 2005; Brem et al., 2005), prediction of regulatory modules (Lee et al., 2006; Ghanzalpour et al., 2006), and pathway enrichment analysis (Subramanian et al., 2005; Ye and Eskin, 2007).

Among these “genetical genomics” approaches, the most widely used statistical analysis is expression quantitative trait loci (eQTL) mapping between genetic variation and gene expression levels (Brem et al., 2002). The goal of these studies is to identify associations between an individual genetic variation and the differential expression of a gene that might help explain the transcriptional regulation of the gene. Many recent studies have identified a large number of cis associations between eQTLs and the expression of genes in close proximity. They have also identified many more trans associations between eQTLs and the expression of genes in other regions of the genome (Yvert et al., 2003; Chesler et al., 2005; Hubner et al., 2005). An interesting observation consistent across multiple datasets is that hundreds or even thousands of genes are trans-regulated by a small number of genomic regions called “regulatory hotspots” (Keurentjes et al., 2007; Chesler et al., 2005) and these associations appear as “trans-regulatory bands” in eQTL plots regardless of the normalization method used (Hubner et al., 2005; Chesler et al., 2005; Bystrykh et al., 2005; Peirce et al., 2006; Williams et al., 2006; Chesler et al., 2006).

Recent genetical genomics studies of yeast have provided much evidence supporting the existence of global regulators that induce trans-regulatory bands (Perlstein et al., 2007; Foss et al., 2007). For mammalian expression datasets, although a large numbers of regulatory hotspots have consistently been observed, the locations of these regulatory hotspots are inconsistent between different
datasets (Hubner et al., 2005; Chesler et al., 2005; Peirce et al., 2006). Simulation studies suggest that spurious regulatory hotspots may be frequently observed in outbred populations (Perez-Enciso, 2004; de Koning and Haley, 2005; Wang et al., 2007).

Building on previous studies we examine two first-generation expression datasets of recombinant inbred (RI) mice where their regulatory hotspots have been shown to poorly replicate in previous studies (Peirce et al., 2006). Due to the high degree of systematic confounding inherent in these datasets, it is particularly challenging to distinguish true genetic effects from the spurious associations. The availability of biological replicates in these datasets allows us to compare the level of true positives between different methods. Two observations suggest that many trans-regulatory bands previously identified in these datasets correspond to “spurious” regulatory hotspots not real genetic effects. First, the locations of regulatory hotspots are inconsistent across disjoint sets of biologically replicated samples. Second, stronger trans-regulatory bands frequently appear with randomly permuted SNPs. To understand the cause of this phenomenon, we carefully examined these datasets and identified a surprising pattern of inter-sample correlation where the pairwise correlations of expression arrays between different strains are often stronger than between replicates of the same strain.

Previous studies have shown that many factors contribute to the spurious correlation between microarray samples including systematic bias from sources such as technical variation in microarray manufacturing (Churchill, 2002; Akey et al., 2007), variations introduced during sample preparation such as the time postmortem a sample is collected, and variations introduced during expression measurements such as the batch of reagents used or laboratory ozone levels (Fare et al., 2003; Branham et al., 2007). Such spurious inter-sample correlation are usually not completely resolved by randomized design of the experiment (Churchill, 2002) or through low-level normalization techniques (Irizarry et al., 2003; Yang et al., 2002).

We suspect that when a SNP, by chance, segregates the strains in a manner consistent with the inter-sample correlation, the p-values of associations between that SNP and the transcripts are inflated leading to spurious associations and in extreme cases, a “trans-regulatory band.” To verify this phenomenon, we constructed a set of simulated data to intuitively show how complex inter-
sample correlation structure inherent in expression data leads to associations between genetic loci and a large number of gene transcripts inducing spurious regulatory hotspots. When we generate random expression data using the same inter-sample correlation structure found in the recombinant inbred expression data, we observed exactly the same regulatory hotspots.

Two types of computational approaches have previously been proposed to reduce the effects of confounding factors in gene expression experiments. The first type are methods that correct for known confounding factors, including ComBat (Johnson et al., 2007), which directly estimates the location and scale model parameters that represent the batch effect using an empirical Bayes (EB) approach. The second type are methods that correct for unknown confounding factors, including Surrogate Variable Analysis (SVA) (Leek and Storey, 2007), which identifies, estimates and corrects for principal components of expression heterogeneity.

We propose a statistical method that corrects for confounding effects induced by complex inter-sample correlation of expression measurements in eQTL mapping using a linear mixed model. Our Inter-sample Correlation Emended (ICE) eQTL mapping directly incorporates the complex correlation structure into the statistical model as a variance component accounting for random effects. Compared to ComBat, our approach is not limited by prior knowledge of confounding factors and is capable of capturing the complex correlation structure introduced by multiple known and unknown effects. Compared to SVA which projects confounding effects onto several distinct single dimensional vectors each treated as a fixed effect in the statistical model, our random effects model is not limited by the number of confounding variables because it does not explicitly infer and correct for each confounding variable. Instead, our method only needs to estimate the total correlation between samples and corrects for the cumulative effects over all confounding factors using the inter-sample correlation structure. Furthermore, the statistical power of SVA decreases as the number of confounding variables increases due to the loss of degrees of freedom while our method always uses only one additional degree of freedom for the inter-sample variance component. As a result, our method has the advantage that it is able to correct for a mixture of strong and moderate confounding effects as shown in our simulation studies while SVA is only able to correct for a number of strong confounding factors.
To gain some intuition as to why our random effects model corrects for confounding factors, consider a pair of samples with a differing expression level for a given gene and a marker SNP which segregates the pair of samples. If the remaining gene expression values have similar expression values between the samples, intuitively this pair of samples provides more evidence that the SNP is associated with the gene’s expression level than if the remaining gene expression values differ greatly between the samples. In the later case, the expression difference of the gene between the pair of samples is less informative given the large amount of global differences in expression values between the samples, which may be due to a confounding factor such as a batch effect.

We applied our statistical model to expression data from two mouse tissues (hematopoetic stem cell and whole brain). In both cases, ICE eQTL mapping outperformed ComBat and SVA in eliminating the spurious trans-regulatory bands while increasing the number of identified cis associations. The remaining trans associations are more likely to be real genetic effects because they are concordant between tissues and between replicates. In yeast, where global regulators have been previous identified, a separate permutation analysis showed that most of the regulatory hotspots are likely to correspond to real genetic effects. Even though yeast regulatory hotspots are likely to be genuine, they globally influence the expression levels and may seriously confound the identification of gene-specific cis or trans associations (Storey et al., 2005; Leek and Storey, 2007). After applying ICE eQTL mapping to correct for the confounding effects from regulatory hotspots, the number of cis associations almost doubled and the concordances of cis and trans associations between disjoint subsets significantly improved. Finally in human lymphoblastoid cell lines, where other known batch effects have been suggested (Akey et al., 2007), our analysis identified more real cis associations than methods that explicitly correct for the batch effects. Our method is publicly available as an R package at http://mouse.cs.ucla.edu/ice
RESULTS

Spurious regulatory hotspots in recombinant inbred mice

We analyzed the expression data from hematopoetic stem cells (HSC) and whole brain tissue collected from BXD mice where prominent trans-regulatory bands have previously been observed (Chesler et al., 2005; Bystrykh et al., 2005). However, most trans-regulatory bands found in these first-generation mouse expression datasets do not overlap with trans-regulatory bands from independent studies (Manly et al., 2005; Peirce et al., 2006). We selected these datasets to evaluate and correct for the systematic confounding effects for two reasons. First, the presence of biological replicates allows us to quantify the level of systematic confounding effects that are heavily imprinted in the datasets. Second, we demonstrate that even in the presence of many complex systematic confounding effects, our method is able to recover true genetic signals better than competing approaches.

We first examined the reproducibility of trans-regulatory bands between different sets of biological replicates. We defined a metric to quantify the strength of a regulatory band allowing us to compare regulatory patterns between datasets. We performed standard eQTL mapping using the t-test and defined the average log p-values across all genes as the regulatory enrichment score. The resulting eQTL map shows that this score correlates well with the prominence of a trans-regulatory band (Figure 1a, 1c). We created two disjoint subsets of expression experiments by picking one of the replicates per strain and compared the enrichment scores between them. Interestingly, the observed patterns of trans-regulatory bands are inconsistent between the subsets (Figure 2a, 2b). The enrichment scores between the replicates are uncorrelated (Spearman \( r = -0.0067 \)) in the BXD brain dataset. On the other hand, the HSC dataset shows relatively high correlation of the enrichment scores (Spearman \( r = 0.30 \)) due to the batch effect shared between two groups of strains.

Next, we examined whether regulatory hotspots are likely to be observed at random SNPs by chance. If we observe stronger trans-regulatory bands with randomly permuted SNP sets than the original trans-regulatory bands, then it suggests that the original hotspots may not correspond to real genetic effects, but are rather caused by co-regulation of a large number of transcripts (Perez-Enciso, 2004; Wang et al., 2007) or other systematic confounding factors. Out of 1,000 random
permutations, we observed hotspots with higher enrichment scores than the strongest hotspot in the original dataset 890 and 643 times, corresponding to genome-wide adjusted p-values of 0.89 and 0.643, for whole brain and HSC dataset, respectively. (Figure 2c, 2d). The inconsistencies between biologically replicated samples and the occurrence of strong trans-regulatory bands with permuted SNPs suggest that the observed trans-regulatory bands correspond to spurious regulatory hotspots which do not correspond to real genetic effects.

**Inter-sample correlation as signatures of systematic confounding effects**

The question remains as to how systematic confounding effects cause spurious regulatory hotspots. To gain intuition of this phenomenon, we examined the pairwise correlations between expression arrays or the inter-sample correlation structure. After normalizing each gene’s expression levels across strains, we computed the correlation between each strain pair and each replicate pair. The normalization ensures that the correlation between truly unrelated strain pairs is expected to approach zero, while the replicated pairs are likely to have higher correlation between them. We observed that most of the inter-sample correlations in recombinant inbred mouse strains do not correspond to real genetic effects. Correlation maps between intra and inter strain replicates show that the diagonals are not pronounced, providing striking evidence that replicated strain pairs are not correlated (Figure 3a,3b). Furthermore, the results show many unrelated strain pairs with much stronger correlation than expected by chance. In the HSC correlation map, there is also a clear division of two groups where members within a group are highly correlated. Upon further analysis, we discovered that the expression measurements for the two groups of individuals were collected in two batches three months apart.

In order to verify that the inter-sample correlation structure effectively captures the systematic confounding effects inducing spurious regulatory hotspots, we created a simulated expression dataset preserving the inter-sample correlation structure. In this dataset, each SNP corresponds to one simulated transcript with cis-regulatory effects accounting for 4% of the variance explained by the SNP. Standard eQTL mapping with simulated data shows almost identical trans-regulatory bands as the original data (Supplementary Figure 9). The reason for this is that the SNPs which segregate
the strains in a manner consistent with the inter-sample correlation structure are more likely to be associated with many expression transcripts. This result strongly supports that most of the trans-regulatory bands are explained by the complex inter-sample correlation structure inherent in expression data.

Furthermore, we evaluated how many transcripts are explained by the inter-sample correlation structure using a variance component model (see Methods). At a false discovery rate (FDR) of 0.05, we observed that 94.1% and 47.9% of the transcripts are significantly associated with the correlation structure in the whole brain and HSC datasets respectively. Since the HSC dataset has an obvious batch effect, we also tested how many transcripts are differentially expressed between the two batches using a t-test. At the FDR threshold of 0.05, only 20.0% of the transcripts are differentially expressed. These results suggest that a significant portion of confounding effects in the HSC dataset are not captured by the known batch effect. When applying SVA to test the significance of surrogates variables explaining the expression levels, 88.7% and 40.9% of the transcripts were significantly associated with the five and six identified surrogate variables of whole brain and HSC dataset, respectively, demonstrating that inter-sample correlation captures more of the systematic confounding than what is captured by surrogate variables with fewer degrees of freedom.

**Inter-sample Correlation Emended (ICE) eQTL mapping**

Motivated by our observation of inter-sample correlation, we propose a new statistical method for identifying eQTLs based on a linear mixed model. Our method first estimates the pairwise correlation between samples which can be accurately estimated since there are thousands of probes in each sample. Instead of assuming independent random variations of expression levels between samples, our method assumes that a gene in pair of samples with globally correlated expression pattern is more likely to have similar expression values than a gene in a pair of globally uncorrelated samples. As a result, the variance component of expression levels at each gene is estimated as a mixture of inter-sample correlation and independent errors. A marker SNP is considered to be significantly associated with a transcript only if it predicts the expression beyond the level suggested by the inter-sample correlation (see Methods). Since spurious regulatory hotspots appear at marker
SNPs consistent with the inter-sample correlation structure, accounting for this correlation in the null model significantly reduces these hotspots.

To demonstrate the effectiveness of our method, we first applied it to the simulated expression datasets presented in the previous section. Although the simulated datasets contain only cis-acting eQTLs, traditional eQTL mapping identified both the cis-acting band and spurious trans-regulatory bands (Supplementary Figure 10a, 10c). The ICE eQTL map shows no trans-regulatory bands and a much stronger cis-regulatory band (Supplementary Figure 10b, 10d). At a FDR level of 0.05, ICE eQTL mapping recovered 8.4% of the simulated whole brain cis-acting eQTLs, which was more than an one hundred fifty fold increase over the standard t-test and more than a three fold increase over SVA. These results illustrate that our method not only eliminates suspicious trans-regulatory bands but also has higher statistical power to recover real eQTLs that might be masked by the correlation structure.

To better understand the relative performance of random effects models versus fixed effects models on this problem, we analyzed our simulated data using the simple t-test, SVA and ICE eQTL mapping. At a SNP effect explaining 5% of phenotypic variations and a systematic confounding effect of 75%, we see that both the fixed effect model (SVA) and our random effects model (ICE eQTL) outperforms the simple t-test in discovering true positives when the samples are correlated in two batches (Figure 4b). When the samples are correlated within smaller groups of size two, we see that ICE eQTL outperforms SVA and the simple t-test (Figure 4d). In the mixture of large group and small group effects, which we expect to see in real datasets, we again see that ICE eQTL outperforms SVA and the simple t-test (Figure 4f). Under complex systematic confounding effects, because the fixed effects model requires a large number of confounding variables to completely correct for the confounding, it loses many degrees of freedom and the estimation of confounding variables becomes less accurate, resulting in the loss of statistical power.

It should be noted that ICE fundamentally differs from traditional mixed model methods such as MANOVA in that it estimates the variance component directly from the expression data. By leveraging the massive number of probes, ICE can accurately estimate the inter-sample correlation. Although we have used block-structure variance components as examples of systematic confounding
in the above simulations, the estimated variance components typically have a much more complex structure. On the other hand, MANOVA uses predefined variance components which are usually block-structured to model random effects specific to groups of samples such as batches, cages, cohorts, or strains, depending on the context of the statistical analysis. Since these variance components are predefined, MANOVA can not correct for unknown confounding factors.

We next applied ICE eQTL mapping to real whole brain and HSC expression datasets from BXD RI mice. In both cases, ICE eQTL mapping eliminated the trans-regulatory bands while enhancing the cis-regulatory bands (Figure 1b,1d). The number of significant cis-acting eQTLs discovered increased dramatically. The enrichment in cis-acting eQTLs serves as a good indicator of the statistical power to identify differential expressions due to true genetic effects, even though some of the cis-associations might be due to polymorphic SNPs residing in the probe sequences(WALTER et al., 2007). For the whole brain dataset, ICE eQTL mapping identified nearly three times as many genes with cis-acting eQTLs (120) as the t-test (43) and 52% more than SVA (79) at a significance level of ten false positives per genome (Table 1, Figure 5). ICE eQTL mapping of the HSC dataset showed fewer significant eQTLs due to the reduced power of having a limited number of strains in the dataset. Nevertheless, similar to the whole brain results, our method consistently identified more genes with cis-acting eQTLs (23) than the t-test (19) and SVA (14). In this case SVA was outperformed by the t-test because having a large number of surrogate variables significantly reduced the degrees of freedom.

Similar to how the number of cis associations detected is a good measure of increased power to identify true genetic effects, another measure is the concordance of association between biologically replicated samples. We leveraged the replicated samples of the BXD datasets to measure the concordances of cis and trans eQTLs between replicates. After ordering the transcripts according to the strength of association for each replicated set, we plotted the concordances of cis and trans associations between the sets using CAT concordance plots(IRIZZARY et al., 2005). In HSC and brain, both cis and trans eQTLs between replicates are significantly more concordant with ICE eQTL mapping (Figure 7a, 7b) than the t-test and SVA. Finally, we compared the results between whole brain and HSC datasets to see if the trans-acting eQTLs are replicable across different tissues.
Previous studies have suggested that most trans-regulatory elements are tissue-specific because they have not been replicated in different tissues (Chesler et al., 2005). We postulate that most trans-regulatory elements were not replicated across tissues in previous studies because they are spurious associations caused by confounding factors. We ordered the transcripts based on the strength of trans-acting eQTLs for each dataset, and computed the Spearman’s rank correlation between the two datasets. The p-values of correlation obtained from the standard \( t \)-test show a slightly negative correlation \((r = -0.012)\), with a one-sided p-value of 0.857. However, the ICE eQTLs show much higher rank-concordance between the tissues with a p-value of \(1 \times 10^{-7}(r = 0.056)\). The CAT concordance plot (Irizarry et al., 2005) also shows that ICE eQTL mapping results are significantly more concordant between tissues (Figure 7c). This suggests that a significant fraction of real trans-acting eQTLs are not tissue-specific, and many of the previously identified trans-acting eQTLs did not replicate since they are largely confounded by spurious associations.

**Some trans-regulatory bands in high quality datasets are likely to correspond to real genetic effects**

In the previous sections, we demonstrated the ability of ICE eQTL mapping to obtain reliable and consistent associations in first generation mouse datasets that have been previously shown to have little reproducibility between independent data sets (Peirce et al., 2006). Second generation datasets collected using better protocols and newer expression chips such as Affymetrix M430v2 are of higher quality, resulting in much higher correlation between replicated samples than between unrelated pairs. Nevertheless, not only do these studies still suffer from moderate levels of inter-sample correlation between unrelated pairs (Supplementary Figure 12), potentially genuine regulatory hotspots globally affect the expression levels and may confound the identification of gene-specific cis or trans associations (Storey et al., 2005; Leek and Storey, 2007). In this section, we analyze one of the classic genetical genomics dataset in yeast where global regulators have been previously reported by several studies (Perlstein et al., 2007; Foss et al., 2007; Brem et al., 2002; Zhu et al., 2008). Under the confounding effects from such genuine regulatory hotspots, we demonstrate that ICE eQTL mapping identifies more cis and trans associations that are consistent.
Yeast expression profiles and genotypes were collected from 112 segregants derived by crossing the lab isogenic BY4716 strain with the wild RM11-1A isolate (Brem et al., 2002; Brem and Kruglyak, 2005). There are several differences between the yeast and the BXD RI datasets. First, the technological differences between the cDNA arrays used for yeast and the Affymetrix GeneChips used for mice may lead to very different patterns of systematic bias. Second, having a larger number of strains increases the number of eQTLs expected at the same significance level due to increased power. Third, since biological replicates are not available in the yeast dataset, it is difficult to determine whether the appearance of regulatory hotspots is caused by a systematic bias or a real genetic effect. Although the dye-swap results provide us with technical replicates, they are not suitable for verifying real hotspots because a dye-swap pair tend to have much stronger correlation than a pair of biological replicates due to smaller environmental or sampling biases between replicates than between unrelated samples (Figure 3c). This correlation may lead to the biased conclusion that most regulatory hotspots are highly reproducible.

After applying traditional eQTL mapping, we observed strong trans-regulatory bands, many of which are consistent with the inter-sample correlation structure (Figure 1e). However, unlike in the BXD recombinant inbred strains, several of the bands remained significant after performing permutation analysis. Three genomic regions in chromosome 2 (521 584kb), 14 (418 502kb), and 15 (171 193kb) had genome wide significant p-values of less than 0.05 with the most significant \( p = 2 \times 10^{-4} \). This suggests that these trans-regulatory bands may be the result of real genetic effects rather than confounding effects. Recently, linkage studies of small-molecule drug response traits with the same set of yeast strains have shown that most of the QTL hotspots of these traits fall into the same genomic region where the bands occur (Perlstein et al., 2007). Since the yeast dataset does not have biological replicates, we instead randomly divided the 112 segregants into two disjoint sets to perform eQTL mappings separately. If the regulatory hotspots are not real genetic effects, it would be unlikely that the same regulatory hotspot consistently appear between the disjoint sets. However, most of hotspots between the sets coincide, suggesting that they correspond to real genetic effects (Supplementary Figure 9d).
We further tried to understand the biological importance of those significant trans-regulatory bands in the yeast data. We listed all 61 genes within 10kb of the significant regulatory hotspots and queried the set of genes in the Comprehensive Yeast Genome Database (CYGD) (Guldener et al., 2005; Robinson et al., 2002). Interestingly, the three regions with significant hotspots on chromosomes 2, 14 and 15, contain IRA1, RAS2, and IRA2, respectively. It has been known that IRA1 and IRA2 genes negatively regulate the RAS2 protein activation state from multiple studies (Tanaka et al., 1990; Colombo et al., 2004). The probability of three genes appearing in a random set of 61 genes is $1.3 \times 10^{-6}$. There are also other genes that encode small GTP-binding proteins of the RAS superfamily such as ARL1, RHO2 and YPT53 near the significant regulatory hotspots. It is possible that the variations in those regions may change the mRNA levels of a large number of genes by perturbing the RAS GTP-binding signal transduction pathway. However, under this interpretation, it is not certain why a portion of mRNA levels are up-regulated while others are down-regulated by the same variant in those regulatory hotspots. In addition, a recent study suggest that MKT1 is the causal regulator that may be responsible for the regulatory hotspot in chromosome 14 (Zhu et al., 2008).

Even though many trans-regulatory bands in yeast are likely to be real genetic effect, they globally influence the expression levels and may seriously confound the identification of gene-specific cis or trans associations (Storey et al., 2005; Leek and Storey, 2007), resulting in the loss of power to identify real cis and trans associations. Correcting for the inter-sample correlation induced by genuine regulatory hotspots may eliminate true trans-regulatory bands, but also can reveal many true regulatory signals obscured by the hotspots. We compared the power of different eQTL mapping methods at identifying true genetic effects by randomly partitioning the dataset as described above. In each partition, the transcripts are ordered by the strength of cis or trans associations, and the concordance between the disjointly partitioned datasets are illustrated using the CAT plot (Figure 7d). The results show that ICE eQTL mapping have higher concordance than the t-test and SVA both for cis-acting and trans-acting eQTLs, despite the loss of true regulatory hotspots.

We applied ICE eQTL mapping to the entire yeast dataset and observed that the trans-regulatory bands are eliminated while the genes with significant cis associations is nearly doubled.
(Figure 1f, Table 1, Figure 5). The number of genes with trans-acting eQTLs are significantly reduced using ICE eQTL mapping due to eliminated regulatory bands, but many new trans-acting genes that have not been identified by the t-test are discovered. For example, among the 363 significant trans-acting genes identified by ICE eQTL mapping at the significance of ten false positives per genome, 25% (89) of them are not identified by the t-test at the same threshold. On the contrary, only 7% (35) of the 506 significant cis-associated genes identified by the t-test are not identified by ICE eQTL mapping at the same significance level. ICE eQTL mapping outperforms SVA in discovering cis-acting eQTLs across different significance thresholds. For trans-acting eQTLs, ICE identifies larger number of eQTLs for conservative significance threshold of FDR less than 0.1, while SVA identifies more eQTLs for higher thresholds. This may be due to the effects from the moderate regulatory hotspots that have not been captured by surrogate variables as appeared in Supplementary Figure 13c. In this dataset, SVA appears to over correct for the trans-regulatory bands and eliminated even the the cis-acting eQTLs in the middle of chromosome 12.

**Correcting for confounding effects in human lymphoblastoid cell line expression**

Finally, we applied our method to a human genetical genomics study of the HapMap individuals where the goal was to determine whether differentially expressed genes between CEU and JPT+CHB populations are caused by allelic or population differences. It is known that the HapMap expression experiments were conducted on different dates for the CEU and JPT+CHB populations and the problems introduced by this batch effect have recently been addressed (Akey et al., 2007; Spielman and Cheung, 2007). While the original paper claimed that 26% of genes are differentially expressed between European and Asian samples at a genome wide Sidak-corrected \( p < 0.05 \), none of them were identified to be significant after controlling for the year in which the sample was processed. In fact, with respect to this batch effect, 28% of the genes were differentially expressed.

We applied ICE eQTL mapping to identify differentially expressed genes. Our method is able to control for the inflated false positives of differentially expressed genes without the prior knowledge of batch information. The p-value distribution appears to be almost uniform (Figure 8). Spielman et al. provided POMZP3 as an example of a differentially expressed gene between the two populations.
to demonstrate that not all of their findings were false positives (SPIELMAN et al., 2007). The gene was associated with a cis-regulatory SNP, whose allele frequency was significantly different between the two populations. We examined how strongly the POMZP3 gene is differentially expressed using three different methods. Without correcting for confounding effects, the gene is significant at a p-value of $1.91 \times 10^{-6}$. However, since numerous other genes are identified to be significant, the strength of the signal is ranked only 943th (23.4%) out of 4,030 genes. After explicitly correcting for the year of the experiment using ComBat (JOHNSON et al., 2007), the gene is no longer significant at a p-value of 0.309. However, the signal is ranked relatively high, 434th (10.8%) out of 4,030 genes. After correcting using SVA, it is ranked only 1992th (49.4%) with a p-value of 0.352. After correcting for the inter-sample correlation pattern using our method, the gene is ranked 6th (0.15%) at a p-value of $3.1 \times 10^{-4}$. Using the same approach, we examined the top 5 genes among the 11 genes reported as differentially expressed genes with concordant cis-eQTLs between populations. Correcting for inter-sample correlation consistently outperformed the other methods at identifying those genes as differentially expressed with higher ranks (Supplementary Table 3).

We next performed ICE eQTL mapping and compared the cis associations with those obtained from t-test based mapping and batch-corrected mapping. We analyzed a total of 3942 genes within 500kb of at least one of the 2 million HapMap SNPs. In both CEU and JPT+CHB populations, the number of genes with cis associations increased significantly with our method (Figure 6a, 6b). eQTL mapping performed after correcting for the known batch effect using ComBat did not significantly outperform the t-test. Furthermore, the concordance of cis-acting genes between populations significantly increased as well, suggesting that ICE association mapping has higher power to identify real genetic effects (Figure 7e).

Finally, we applied our method to identify differentially expressed genes with evidence of concordant cis-acting SNP between populations. We applied a more stringent threshold than previous studies (SPIELMAN et al., 2007) by requiring the cis-acting SNP to have a genome wide p-value of less than $2.5 \times 10^{-8}$ in at least one population and a strong p-value of less than $10^{-5}$ in the other after Bonferroni correction. In addition, we required the minor allele frequency of the SNP to differ by at least 0.1, and the strength of differential expression to be ranked in the top 10% of all genes.
Using these stringent criteria, only two genes are identified using the \( t \)-test, and three genes are identified after explicitly correcting for the batch effect. On the other hand, ICE association mapping successfully identified 10 differentially expressed genes including four previously unreported (Table 2).

**Comparison with previous methods**

A key difference of ICE association mapping from the previous methods using singular value decomposition(Alter *et al.*, 2000; Leek and Storey, 2007) is that previous methods project the systematic confounding onto several distinct single dimensional vectors as fixed effects while ICE association mapping directly incorporates the pairwise correlation as random effects into the statistical model. For previously known confounding variables such as batch effects, both methods can incorporate them as fixed effects in the statistical model. While the singular value decomposition methods infers a number of confounding factors strongly affecting the variations in expression, those with relatively moderate effects may remain uncorrected. Our mixed model approach ICE does not suffer from this shortcoming since it does not explicitly infer a limited number of confounding variables. Instead, the confounding effects from various unknown sources are assumed to be intrinsically imprinted in the expression profiles, specifically as inter-sample correlation.

The simulation results under various types of confounding effects we presented above (Figure 4) are largely consistent with those seen with applying random models versus fixed models on the related problem of correcting for population structure in association studies. Previous studies showed that the random effects model corrects for the heterogeneous population structure better than a fixed effects model based on principal component analysis such as EIGENSTRAT for model organisms association mapping(Yu *et al.*, 2006; Zhao *et al.*, 2007; Kang *et al.*, 2008). Although EIGENSTRAT can robustly correct for population structure in human association mapping where an admixture model assuming a small number of distinct ancestral populations accurately describes the structure of the data(Pritchard *et al.*, 2000; Price *et al.*, 2006), in the model organism association mapping involving multi-level population structure, such methods only partially capture the population structure resulting in an inflated number of false positives(Aranzana *et al.*, 2005).
Similarly, we see that fixed effects models can effectively correct for inter-sample correlation where there’s relatively simple confounding structure such as batch effects while the random effects model performs much better when we have more complex and multi-leveled confounding structures we see in simulated and real data sets.

A second intuition why mixed models outperform SVD methods in this case is that a large number of surrogate variables or eigengenes are required to capture the complex expression heterogeneity, resulting in a significant increase in the degrees of freedom which affects the statistical power. These effects can be substantial especially for those datasets with a limited number of samples. For example, in the HSC dataset containing only 22 strains, SVA was shown to be even less powerful than the \( t \)-test in identifying \textit{cis}-acting eQTLs.

Both approaches have potential risk of over correcting true genetic effects, especially for those \textit{trans}-acting eQTLs corresponding to true regulatory hotspots. The concordance plots between replicated and disjointly partitioned datasets consistently show that our ICE association mapping provides higher concordance than standard \( t \)-test at identifying both \textit{cis}-acting and \textit{trans}-acting eQTLs, while SVA method consistently shows lower concordance of \textit{trans}-acting eQTLs than standard \( t \)-test (Figure 7). Although some genuine regulatory hotspots may have been eliminated using ICE eQTL mapping particularly in the yeast dataset, we were able to identify some of the regulatory hotspots as significant through the analysis of replicates and our SNP permutation approach.

In terms of the computational cost, the running time of ICE association mapping is twice as fast as SVA using Efficient Mixed Model Association (EMMA)(Kang \textit{et al.}, 2008).

**DISCUSSION**

We have proposed a novel statistical method, Inter-sample Correlation Emended (ICE) eQTL mapping which corrects for the systematic confounding effects inherent in expression datasets. Using the first-generation RI mouse expression dataset where the problem of systematic confounding effects has already been documented, we have demonstrated that most \textit{trans}-regulatory bands in the dataset correspond to spurious regulatory hotspots through the analysis of biological replicates
and the permutation analysis. Using simulated data that preserves the inter-sample expression correlation structure, we have shown that the inter-sample correlation effectively characterize the systematic biases that are responsible for the spurious associations. Using the same methods, we demonstrated that a number of trans-regulatory bands in yeast correspond to genetic variation in global regulators.

From both differential expression analysis in human and association analysis in recombinant inbred mice and yeast, we conclude that our method is more robust at correcting for systematic confounding factors than previous methods including an explicit batch correction method, ComBat(JOHNSON et al., 2007) and an automated method that corrects for unknown confounding factors, Surrogate Variable Analysis(LEEK and STOREY, 2007). Not only did ICE eQTL mapping identify more cis-acting eQTLs than both methods, those identified cis-acting and trans-acting eQTLs also showed higher concordance between replicated datasets (BXD RI strains), different tissues (BXD RI strains), and disjoint subsets (yeast). These results suggest that our method has higher power to identify associations corresponding to real genetic effects.

Our results also highlight the importance of obtaining independent replicates of expression measurements and the utility of these replicates for analyzing and validating eQTLs. We have shown that different strategies for obtaining replicates have profound effects on the correlation structure between replicates. Technical replicates obtained by either performing a dye-swap (Figure 3c) or running multiple expression arrays in the same sample (Figure 3d) exhibit much higher correlation between replicate pairs than full biological replicates (Figures 3a and 3b). We suspect that confounding factors in the sample preparation are largely responsible for the higher pairwise correlation observed among technical replicates reducing their utility in analysis and validation. Preparing multiple samples from the same individual can help reduce the effect of these confounding factors. In many eQTL studies, it is possible to independently measure expression from genetically identical individuals which can further reduce the effects of these confounding factors.
Gene expression data and genetic maps

We obtained the yeast expression dataset over 112 segregants across 6,216 probes from the GEO database with accession number GSE1990 (BREM and KRUGLYAK, 2005). Each of them has two replicates, and the values are represented as the log ratio between the expression and the average expression of the reference (BY) strains. 5,534 genes with validated genomic annotations were mapped onto the genome to draw the genome wide eQTL maps. For BXD RI datasets, we obtained the hematopoetic stem cell (HSC) data from the GEO database with accession number GSE2031, and the whole brain dataset by request from the authors. Both datasets use the Affymetrix U74Av2 GeneChip platform and contain 12,422 probes. 8,596 probes were mapped onto NCBI build 34 version of the mouse genome using refSeq to draw the eQTL maps. The HSC dataset contains the expression data over 22 strains with duplicates for each strain, and the whole brain dataset contains 64 samples over 28 strains, varying one to four measurements per strain. The second generation whole brain dataset using M430v2 arrays downloaded from GeneNetwork (http://www.genenetwork.org) contains expression profiles over 45102 probes across 30 BXD RI strains with up to 6 replicates per strain. Their expression values were normalized using RMA (IRIZARRY et al., 2003).

We obtained the human lymphoblastoid cell line expression data over the HapMap individuals from the GEO database with accession number GSE5859 (SPIELMAN and CHEUNG, 2007). There are a total of 141 samples, 60 from CEU and 81 from JPT+CHB. Although the Affymetrix Genome Focus Array contains 8,500 annotated genes, we focused on the 4,030 that are expressed in lymphoblastoid cell lines defined the same way as Spielman, et. al (SPIELMAN and CHEUNG, 2007).

The genetic map of 2,956 SNPs of yeast segregants were obtained by request from the authors. The BXD RI SNPs were obtained from the Wellcome Trust Center, containing 13,270 SNPs over the genome, of which 7,413 SNPs are polymorphic between the two parental strains. Sixty-one and 25 SNPs with minor allele frequency less than 5% were discarded in the HSC and the whole
brain datasets respectively. A very small portion of heterogeneous alleles in the RI strains were assumed to have additive effects, and the missing SNPs were not resolved. The genetic map for the HapMap samples were obtained using release 22 of the human HapMap (INTERNATIONAL HAPMAP CONSORTIUM, 2003). We examined a total of 3942 genes that are within 500kb of at least one of the 2 million HapMap SNPs.

**Traditional eQTL mapping and genome wide eQTL maps**

Traditional eQTL mapping was performed by taking the average of expression values of each strain and performing $t$-test between each marker SNP and each transcript. The eQTL mapping using either of the replicates was performed in the same way except that the samples were divided into two disjoint sets of expression experiments by randomly picking one of the replicates. For the seven strains that have only one measurement in the BXD whole brain dataset, they were included in both sets. Missing SNPs or missing expression values were excluded in the test only for the corresponding marker-transcript pair, and the p-value was obtained from the asymptotic $t$-distribution.

Genome-wide eQTL maps were plotted based on the relative genomic positions of each transcripts and marker SNPs. Since the previously suggested transcriptome map (CHESLER et al., 2005) may create artificial horizontal bands due to non-uniform genomic densities of the probes, we used an eQTL map based on the relative positions of markers and probes, simply corresponding each marker-transcript association to one pixel. The degree of redness of each pixel is proportional to the log p-values.

**Explicit batch effect correction and Surrogate Variable Analysis**

We explicitly corrected for known batch effects using the ComBat R package (JOHNSON et al., 2007). We used the default settings and the batch corrected expression levels were used to perform traditional eQTL mapping using the $t$-test.

For surrogate variable analysis, we used SVA R package downloaded from the author’s website, identifying surrogate variables ignoring the genotype data as suggested (LEEK and STOREY, 2007).
The p-values are obtained using a linear model after correcting for the surrogate variables.

**Genome wide inter-sample correlation**

An inter-sample correlation matrix from a expression dataset is generated as follows. Let $Y$ be a $m \times n$ expression matrix with $n$ arrays for $m$ genes, then the inter-sample correlation matrix is generated as follows. Let $\mu_i, \sigma_i$ be the mean and standard deviation of expression values of $i$-th genes, $(Y_{i1}, Y_{i2}, \cdots, Y_{in})$. Let $Z$ be a $m \times n$ matrix with each element $Z_{ij} = (Y_{ij} - \mu_i)/\sigma_i$, then the inter-sample correlation matrix defined as the covariance matrix of $Z$. It should be noted that we used the covariance matrix $K = \text{Cov}(Z)$ instead of the correlation matrix because the variances are not homogeneous across the strains. Such heterogeneous distribution of variances can be an additional source of systematic confounding but is not emphasized in the main text for the sake of simplicity.

In order to visually compare the consistency between replicated pairs and unrelated pairs, we used the correlation matrix of $Z$ for each replicated dataset because the correlation matrix can be more intuitively to understood than the covariance matrix. Each diagonal element represents the pairwise correlation of a replicated pair. In the upper-triangular region, the correlations between unrelated pairs in one subset of replicates was computed and visualized. In the lower-triangular region, the other subset was computed and visualized. The seven strains without replicates in the BXD whole brain dataset were omitted in the heatmap visualization.

**Simulation studies**

The eQTL mapping with permuted SNPs was performed by permuting the SNPs across the individuals, thereby preserving the correlation between each pair of SNPs. For generating the simulated expression data preserving the genome wide correlation pattern, we assumed the following generalized linear model.

$$y = \alpha g + u$$

(1)
where g represents SNPs encoded by 0 and 1, and u is a multivariate normal random variable sampled from \( N(0, K) \). K = Cov(Z) is the inter-sample correlation matrix defined in the previous section. \( \alpha \) is set so that cis-regulatory effects account for 4% of the phenotypic variation explained by each causal SNP. The number of significant cis-eQTLs are counted using a conservative FDR estimate with \( \pi_0 = 1 \), considering only the SNP and simulated gene pair where the cis-regulatory effects are simulated.

For the comparison of various systematic confounding effects, we simulated expression datasets of 500 genes over 50 samples from three different inter-sample correlation structure described in Figure 4, with 75% of phenotypic variations explained by the confounding effects using a multivariate normal distribution(Kang et al., 2008). We generated a random SNP of minor allele frequency of 0.3 for each gene, and added a SNP effect explaining 5% of phenotypic variation. We performed eQTL mapping using t-test, SVA, and ICE eQTL mapping for all 500 \( \times \) 500 SNP-gene pairs, and computed the true positive rates at each p-value cutoff.

**Variance component test**

We applied the following variance component model to assess the statistical significance of each association in the presence of genome wide correlation.

\[
y = X\beta + u + e \tag{2}
\]

where \( X, \beta \) is the fixed effects of known confounding variables and their coefficients, and \( u \sim N(0, \sigma_g^2 K) \) and \( e \sim N(0, \sigma_e^2 I) \) are random variables accounting for unknown confounding and random errors. \( K = \text{Cov}(Z) \) is the inter-sample correlation matrix and \( I \) is an identity matrix. \( \sigma_g^2 \) and \( \sigma_e^2 \) are coefficients of the two variance components. Under the null hypothesis, \( \sigma_g^2 = 0 \) is assumed. Under the alternative hypothesis, \( \sigma_g^2 > 0 \) is tested. Only the mean is used as the fixed effect in the analysis above. The asymptotic null distribution of the likelihood ratio test statistic follows a 1:1 mixture of \( \chi_0^2 \) and \( \chi_1^2 \) distributions(Stram and Lee, 1994). Efficient Mixed Model Association (EMMA) R package was applied for rapid estimation of variance components and maximum
likelihood to perform likelihood tests (Kang et al., 2008).

We used standard $t$-test to test for the known batch effect for the BXD HSC dataset. When testing the significance of surrogate variables, a standard F test is performed to assess the significance of all surrogate variables using a linear model. In all tests above, FDR is conservatively estimated with $\pi_0 = 1$.

**ICE eQTL mapping**

ICE QTL mapping models the expression levels as the following linear mixed model:

$$y = G\alpha + X\beta + u + e$$  \hspace{1cm} (3)

where $X$, $\beta$ are the fixed effects of known confounding variables and their coefficients, and $u$ and $e$ are random variables accounting for unknown confounding and random errors as described above. $G$ represents the SNPs or other predictor variables to be tested with the coefficients of $\alpha$. EMMA is applied to test for the significance of $\alpha$ using F test as previously suggested based on REML estimates of variance component (Yu et al., 2006; Zhao et al., 2007; Kang et al., 2008).

We classified the eQTLs as cis-acting when the SNP and the probe are no farther than 50kb for yeast, 5Mb for BXD mouse RI strains and 500kb for human. Trans-acting eQTLs are stringently classified with the distance larger than 250kb for yeast, and 50Mb for mouse. The number of expected false positives were computed from 10 null randomized runs of each eQTL mapping setting $\pi_0 = 1$, as suggested in previous studies (Storey and Tibshirani, 2003; Brem and Kruglyak, 2005).

**Assessing the statistical significance of trans-regulatory bands**

The statistical significance of a trans-regulatory band was quantified as the average of log p-values across all probes. We performed 10,000 random permutations of the SNP set with family-wise error correction to evaluate the genome wide corrected p-value of the strength of trans-regulatory bands.
REFERENCES


TANAKA, K., M. NAKAFUKU, T. SATOH, M. S. MARSHALL, J. B. GIBBS, et al., 1990 S.cerevisiae genes IRA1 and IRA2 encode proteins that may be functionally equivalent to mammalian ras gtpase activating protein. Cell 60.


WANG, S., T. ZHENG, and Y. WANG, 2007 Transcription activity hot spot, is it real or an artifact? BMC Proc 1 Supple 1: S94.


**Authors’ Contributions**

HMK, CY, and EE conceived the research, developed the methods, and wrote the paper. HMK and CY analyzed the data.

**Acknowledgment**

We thank A. Jake Lusis at University of California, Los Angeles and Andreas Beyer at the BIOTEC, TU-Dresden for useful discussions. H.M.K is supported by the Samsung Scholarship. H.M.K., C.Y., and E.E. are supported by the National Science Foundation Grants No. 0513612 and No. 0731455, and National Institutes of Health Grant No. 1K25HL080079. Part of this investigation was supported using the computing facility made possible by the Research Facilities Improvement Program Grant Number C06 RR017588 awarded to the Whitaker Biomedical Engineering Institute, and the Biomedical Technology Resource Centers Program Grant Number P41 RR08605 awarded to the National Biomedical Computation Resource, UCSD, from the National Center for Research Resources, National Institutes of Health. Additional computational resources were provided by the California Institute of Telecommunications and Information Technology (Calit2).

**Competing Interests** The authors declare that they have no competing financial interests.

**Correspondence** Correspondence and requests for materials should be address to E.E. (email: eeskin@cs.ucla.edu)
### Table 1: Number of genes with significant \textit{cis} and \textit{trans} eQTLs in three datasets at different number of expected false positives. Total of 5,534 genes are tested in the yeast dataset, and 8,596 genes are tested in two BXD mouse datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>type</th>
<th>method</th>
<th>E(false positives)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Yeast</td>
<td>\textit{cis}</td>
<td>\textit{t}-test</td>
<td>646</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>506</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>398</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SVA</td>
<td>817</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>617</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>490</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ICE</td>
<td>1118</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>935</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>757</td>
</tr>
<tr>
<td>BXD whole brain</td>
<td>\textit{cis}</td>
<td>\textit{t}-test</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>43</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SVA</td>
<td>148</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>79</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>42</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ICE</td>
<td>193</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>120</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>69</td>
</tr>
<tr>
<td>BXD HSC</td>
<td>\textit{cis}</td>
<td>\textit{t}-test</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>19</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SVA</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>14</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ICE</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>23</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>\textit{trans}</td>
<td>\textit{t}-test</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SVA</td>
<td>102</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>11</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ICE</td>
<td>103</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Gene Name</td>
<td>t-test</td>
<td>Quantiles (Rank)</td>
<td>ELG</td>
</tr>
<tr>
<td>-----------</td>
<td>--------</td>
<td>------------------</td>
<td>-----</td>
</tr>
<tr>
<td>TECA1</td>
<td>21%</td>
<td>(840) 30% (1227)</td>
<td>0.55%</td>
</tr>
<tr>
<td>UGT2B17*</td>
<td>1.4%</td>
<td>(58) 7.2% (292)</td>
<td>0.024%</td>
</tr>
<tr>
<td>CSTB*</td>
<td>14%</td>
<td>(559) 6.9% (277)</td>
<td>1.2%</td>
</tr>
<tr>
<td>DFNA5</td>
<td>7.0%</td>
<td>(284) 2.4% (96)</td>
<td>2.0%</td>
</tr>
<tr>
<td>PEX6*</td>
<td>21%</td>
<td>(866) 19% (773)</td>
<td>0.84%</td>
</tr>
<tr>
<td>POMZP3*</td>
<td>23%</td>
<td>(943) 11% (434)</td>
<td>0.15%</td>
</tr>
<tr>
<td>RABGCTA</td>
<td>50%</td>
<td>(2013) 24% (985)</td>
<td>3.8%</td>
</tr>
<tr>
<td>RRM1</td>
<td>16%</td>
<td>(633) 12% (487)</td>
<td>5.8%</td>
</tr>
<tr>
<td>AP3S2*</td>
<td>6.5%</td>
<td>(262) 5.8% (255)</td>
<td>10%</td>
</tr>
<tr>
<td>NUBP2*</td>
<td>10%</td>
<td>(416) 11% (452)</td>
<td>6.4%</td>
</tr>
<tr>
<td>DNAJC15*</td>
<td>19%</td>
<td>(771) 10% (405)</td>
<td>3.4%</td>
</tr>
</tbody>
</table>

Table 2: List of differentially expressed genes with concordant cis associations across populations. The bold genes have (1) cis associations with $p < 2.5 \times 10^{-8}$ in at least one population, (2) cis associations with $p < 10^{-5}$ at the same SNP in the other population, (3) the difference of minor allele frequency (MAF) between populations is greater than 0.1, and (4) the strength of differential expressions between populations ranked among the top 10% out of 4,030 genes, using three different methods. Only two and three genes pass the criteria using the standard $t$-test and $t$-test after batch correction, respectively. 10 genes are identified to be significant using ICE eQTL approach, including four previously unreported differentially expressed genes. The genes are ordered by the difference of MAF between populations.

* : Previously identified by Spielman et. al.(Spielman et al., 2007)
† : If multiple methods identify the same gene with different strongest SNP, the SNP is selected according to the method providing the strongest rank in differential expression.
** : Although two methods have ranked the gene as among the top 10% in terms of differential expressions, no cis associations pass the p-value threshold described above.
Figures

List of Figures

1. Genome wide eQTL maps using simple $t$-test and ICE eQTL mapping .......................... 35
2. Comparison of regulatory hotspots in BXD datasets .................................................. 37
3. Heatmap visualizations of genome wide inter-sample correlation with replicated samples 38
4. Statistical power of different methods under various systematic confounding effects . 39
5. Number of genes with significant eQTLs ................................................................. 40
6. Cis associations within each HapMap population ....................................................... 41
7. Concordance of eQTLs between replicates, tissues, and populations ....................... 42
8. Distribution of p-values testing for differential expressions between populations using different methods ................................................................. 43
9. (Supplementary) Comparison of regulatory hotspots generated from original dataset and simulated datasets ................................................................. 45
10. (Supplementary) eQTL maps of simulated expression datasets from the inter-sample correlation in mouse datasets ......................................................... 46
11. (Supplementary) Different patterns of $trans$-regulatory bands under various systematic confounding effects ................................................................. 47
12. (Supplementary) Heatmap visualizations of inter-sample correlation with replicated samples of whole brain with M430v2 chip ......................................................... 48
13. (Supplementary) Genome wide eQTL maps using Surrogate Variable Analysis ..... 50
14. (Supplementary) Quantile-Quantile plot of differential expressions between populations 51
Figure 1: Refer to the next page for the details.
Figure 1: Genome wide eQTL maps from (A, B) BXD recombinant inbred whole brain dataset, (C, D) BXD recombinant inbred HSC dataset, and (E, F) yeast dataset, using (A, C, E) standard t-test and (B, D, F) ICE eQTL mapping. The SNPs (horizontal) and probes (vertical) are sorted according to genomic position, mapping each pixel to the linkage between a marker and a gene. The color of each pixel represents the strength of the linkage signal, with red being the strongest signal and white being the weakest. The yellow graph on top represents the strength of trans-regulatory bands, quantified as the average log-p values at each SNP across all genes. There are clear vertical trans-regulatory bands using the standard t-test to perform eQTL mapping. Those bands are eliminated using ICE eQTL mapping. A total of 8,596 probes and 7,413 SNPs are mapped in the two mouse datasets, and 5,534 probes and 2,956 SNPs in the yeast dataset.
a: Between different subsets of replicates of BXD whole brain dataset

b: Between different subsets of replicates of BXD HSC dataset

c: Between original SNPs and permuted SNPs for BXD whole brain dataset

d: Between original SNPs and permuted SNPs for BXD HSC dataset

Figure 2: (A) and (B) compare the strength of trans-regulatory bands between replicated subsets, showing that the hotspots are inconsistent between replicates. (C) and (D) compare between using permuted SNPs and using original SNPs, illustrating that even stronger trans-regulatory bands are frequently observed using permuted SNPs. The horizontal axis is the genomic positions of the markers in megabases, and the vertical axis is the strength of regulatory hotspots quantified as the average log-p values at each marker across all genes. Each peak represents the strength of the trans-regulatory band at a particular marker SNP. The taller the peak, the more pronounced a trans-regulatory band is in an eQTL map.
Figure 3: Genome wide correlation coefficients were computed for each pair of samples after standardizing each gene across the samples in BXD RI HSC and whole brain datasets, yeast dataset and HapMap dataset. The x-axis represents one subset of replicates and the y-axis represents the other. Each axis is ordered by strain (mouse), segregant (yeast) or individual name (human). Each diagonal element represents the strength of correlation between the replicated samples. Lower-triangular and upper-triangular regions show the correlation coefficients among two disjoint subsets of replicated samples, covering (A) 22 replicated strains for HSC, (B) 21 replicated strains (excluding 7 non-replicated strains) for whole brain, (C) 112 yeast segregant samples, and (D) 42 HapMap CEU samples. In the plot of BXD datasets where “biological” replicates were collected by measuring expression in two animals from the same strain, neither diagonal is pronounced suggesting that the replicated pairs are not correlated with each other. In yeast and human dataset where “technical” replicates were collected by either performing a dye-swap or running multiple expression arrays on the same sample, the diagonals are moderately pronounced suggesting that the technical replicates are more correlated than unrelated samples. In the HSC data, two distinct groups of correlated samples (strains 1-10 and strains 11-22) suggest a clear batch effect.
Figure 4: Statistical power under systematic confounding from (A) large group (batch) correlation, (C) small group (pairwise) correlation and (E) a combination of batch and pairwise correlation structures in gene expression, using t-test, SVA, and ICE eQTL mapping, presented in (B), (D), and (F), respectively. All the p-values in the eQTL map are ranked and the fraction of true positives is plotted across different quantile of the p-values. For example, in among top 0.1% of p-values in (B), 27% of the signals are true positives with ICE-eQTL mapping, and 12% and 10% are true positives with SVA and t-test.
Figure 5: Number of genes with significant \textit{cis}-acting and \textit{trans}-acting eQTLs at various p-value thresholds in the (A) BXD whole brain dataset, (B) BXD HSC dataset, (C) yeast dataset. The horizontal axis represents the genome wide p-values of the most significant \textit{cis} association or \textit{trans} association for each gene, adjusted by Bonferroni correction using the number of non-redundant SNPs. The vertical axis represents the number of genes with significant associations at a given p-value threshold. In all three datasets, ICE eQTL mapping outperformed traditional eQTL mapping by consistently finding more \textit{cis} associations at all p-value cutoffs. It also consistently found more real \textit{trans} associations at low p-value cutoffs while fewer spurious \textit{trans} associations at moderate p-value cutoffs in mouse.
Figure 6: Number of genes with significant cis associations at various p-value threshold for (A) European-derived and (B) Asian-derived populations among HapMap samples. The standard t-test without correction, t-test after correcting for the previously known batch effect (year of experiment), and ICE association mapping are compared. The horizontal axis represents the Bonferroni adjusted p-values of the most significant cis association for each gene, and the vertical axis represents the number of genes with significant associations at a given p-value threshold. In both populations, ICE eQTL mapping outperformed traditional eQTL mapping by consistently finding more cis associations at all p-value cutoffs. Batch corrected eQTL mapping does not outperform traditional eQTL mapping.
Figure 7: Each gene is ranked according to the strength of *cis* and *trans* associations, and the fraction of genes concordantly ranked within a certain rank is plotted in CAT format (Irizarry et al., 2005). The horizontal axis represents the size of top ranking genes according to the strongest *cis* or *trans* eQTLs, and the vertical axis represents the number of genes concordantly ranked within the size divided by the size of the list. If the ranks are completely independent, the curve will follow the null distribution line. The ICE eQTL mapping shows higher concordance of both *cis* and *trans* eQTLs than the standard t-test across all four comparisons, including (A) between biological replicates of BXD whole brain dataset, (B) between biological replicates of BXD HSC dataset, (C) between whole brain and HSC of BXD datasets, (D) between disjoint subsets of yeast datasets, and (E) between European and Asian HapMap populations (*cis* associations only).
Figure 8: Differential expressions between two HapMap populations are tested using three different methods: (A) t-test using the original uncorrected expressions, (B) Batch correction by the year of experiment, (C) Surrogate Variable Analysis, and (D) ICE association mapping. The horizontal axis represents the p-values of differential expression for each gene using three different methods, and the vertical axis is the frequency of each interval out of 4,030 genes.
## Supplementary Tables

Table 3: The relative strengths of differential expression in two different populations of the top 5 genes reported by Spielman et al. (Spielman et al., 2007), with significant *cis*-eQTLs and significant MAF differences between the populations.

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Quantiles (Rank)</th>
<th>MAF of <em>cis</em>-acting SNP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>t</em>-test</td>
<td>batch-corrected</td>
</tr>
<tr>
<td>UGT2B17</td>
<td>1.4% (58)</td>
<td>7.2% (292)</td>
</tr>
<tr>
<td>POMZP3</td>
<td>23% (943)</td>
<td>11% (434)</td>
</tr>
<tr>
<td>PEX6</td>
<td>21% (866)</td>
<td>19% (773)</td>
</tr>
<tr>
<td>PSPHL</td>
<td>28% (1136)</td>
<td>6.2% (249)</td>
</tr>
<tr>
<td>CSTB</td>
<td>14% (559)</td>
<td>6.9% (277)</td>
</tr>
</tbody>
</table>
Figure 9: Expression datasets were simulated preserving inter-sample correlation structure of expressions, and the regulatory hotspots computed from the simulated dataset were compared to those from the original dataset, in (A) the BXD whole brain dataset (B) the BXD HSC dataset, and (C) the yeast dataset. (D) is the comparison between disjoint subsets of yeast strains, showing that regulatory hotspots are reproduced with independent sets of samples. The simulated datasets almost perfectly reproduced the original regulatory hotspots, suggesting that the inter-sample correlation is the primary source of spurious regulatory hotspots. The horizontal axis is genomic positions of the markers in megabases, and the vertical axis is the strength of regulatory hotspots quantified as the average log-p values at each marker across all the genes. (See Figure 2).
Figure 10: Refer to next page for details.
Figure 11: Different systematic confounding effects lead to different pattern of *trans*-regulatory bands. (A) large group (batch) correlation, (B) small group (pairwise) correlation and (C) a combination of pairwise and batch correlation structures. The combination of pairwise and batch correlation effects result in an eQTL map similar to those observed in real data sets.
Figure 12: Genome wide correlation coefficients were computed for each pair of samples after standardizing each gene across the samples in BXD RI whole brain datasets using M430v2 arrays (Williams RW, unpublished). The x-axis represents one subset of replicates and the y-axis represents the other. Each axis is ordered by strain (mouse), and diagonal elements represent the strength of correlation between the replicated samples. Lower-triangular and upper-triangular regions show the correlation coefficients among two disjoint subsets of replicated samples, covering 30 biologically replicated strains. The diagonal is moderately pronounced suggesting that these dataset is of higher quality than the previous dataset.
Figure 12: Simulated expression datasets are generated preserving inter-sample correlation structure of expressions, and traditional and ICE eQTL mapping is applied to the datasets. Six eQTL maps are plotted from (A,B) BXD recombinant inbred whole brain dataset, (C,D) BXD recombinant inbred HSC dataset, and (E,F) yeast dataset, using (A,C,E) standard t-test and (B,D,F) ICE eQTL mapping. There are almost identical pattern of regulatory hotspots observed in the original dataset with standard t-test, and they are eliminated after applying ICE eQTL mapping. Each SNP (horizontal) and probe (vertical) are sorted according to genomic positions, mapping each pixel to the linkage between a marker and a gene. The color of each pixel represents the strength of linkage signal. The yellow graph on the top represents the strength of trans-regulatory bands, quantified as the average log-p values at each SNP across all the genes. A total of 7,413 probes and 7,413 SNPs are simulated and mapped in two mouse datasets, and 2,956 probes and 2,956 SNPs in the yeast dataset.
Figure 13: Genome wide eQTL maps from (A) BXD recombinant inbred whole brain dataset, (B) BXD recombinant inbred HSC dataset, and (C) yeast dataset, using Surrogate Variable Analysis (Leek and Storey, 2007). The SNPs (horizontal) and probes (vertical) are sorted according to genomic position, mapping each pixel to the linkage between a marker and a gene. The color of each pixel represents the strength of the linkage signal, with red being the strongest signal and white being the weakest. The yellow graph on top represents the strength of trans-regulatory bands, quantified as the average log-p values at each SNP across all genes. Unlike ICE eQTL mapping, some vertical trans-regulatory bands remain using SVA. But even more problematic is the elimination of cis-associations on chromosome 12 of the yeast data due to over correction. A total of 8,596 probes and 7,413 SNPs are mapped in the two mouse datasets, and 5,534 probes and 2,956 SNPs in the yeast dataset.
Figure 14: Q-Q plot of differential expression between two HapMap populations tested using the standard $t$-test, batch-corrected $t$-test, and ICE $t$-test. The horizontal axis is the negative logarithm of the quantiles of the observed p-values, and the vertical axis represents the negative log p-values of differential expression.