Supplemental Material for
Natural selection on *cis* and *trans* regulation in yeasts

J.J. Emerson[1,10], Li-Ching Hsieh[1,2,10], Huang-Mo Sung[3,10], Tzi-Yuan Wang[1,4,10], Chih-Jen Huang[4,5,6], Henry Horng-Shing Lu[7], Mei-Yeh Jade Lu[1,4], Shu-Hsing Wu[8], and Wen-Hsiung Li[1,4,9,11]

1 Genomics Research Center, Academia Sinica, Taipei 115, Taiwan
2 Institute of Information Science, Academia Sinica, Taipei 115, Taiwan
3 Department of Life Sciences, National Cheng Kung University, Tainan 701, Taiwan
4 Biodiversity Research Center, Academia Sinica, Taipei 115, Taiwan
5 Molecular and Biological Agricultural Sciences Program, Taiwan International Graduate Program, Academia Sinica, Taipei 115, Taiwan
6 Graduate Institute of Biotechnology, National Chung Hsing University, Taichung 402, Taiwan
7 Institute of Statistics, National Chiao Tung University, Hsinchu 30010, Taiwan
8 Institute of Plant and Microbial Biology, Academia Sinica, Taipei 115, Taiwan
9 Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637
10 These authors contributed equally to this work and are joint first authors.

11 Corresponding author.
E-mail: whli@uchicago.edu; fax (773) 702-9740

Running title: **Selection on *cis* and *trans* regulation in yeasts**

Key words: expression evolution, gene regulation, *cis*, *trans*, population genomics, evolutionary constraint, natural selection

Supplemental Methods

Mapping IGA reads to the reference genomes

In order to map the cDNA sequence reads to the genomes in each sample, every sequence read was used as a query against each reference yeast genome by using Megablast with the wordsize parameter set to 8. We then recorded all hits with up to two nucleotide mismatches. A mismatch may be due to a sequencing error in the sequence read we obtained, a sequence error in the reference genome(s), or a SNP site between the two reference genomes. In order to distinguish among these possibilities, we developed a set of procedures: (i) For each genome, we classified the Megablast results into three datasets - the perfectly matched sequence reads (BY_{mis0} for the BY genome and RM_{mis0} for the RM genome), the sequence reads with one mismatch (BY_{mis1} and RM_{mis1}) and the sequence reads with two mismatches (BY_{mis2} and RM_{mis2}). (ii) The perfectly matched sequences from BY and from RM were combined together to form a new dataset (All_{mis0}) that included all the perfectly matched sequences. (iii) Each sequence from BY_{mis1} or from RM_{mis1} was searched against All_{mis0}. If the sequence was found in All_{mis0}, it was a sequence which had been perfectly mapped on one of the genomes and had one SNP site against the other genome. On the other hand, if the sequence was not in All_{mis0}, it was a sequence read with one sequence error site. Then all the sequences with one sequence error site were obtained. (iv) The sequences with one sequence error site and the sequences from All_{mis0} were combined together to form a new dataset (All_{mis0+1}) that included all the perfectly matched sequences and all the sequences with one sequence error site. (v) The sequences with two sequence error sites were obtained by using a similar procedure as in (iii). After completing the above procedure, we obtained all mapped reads with up to 2 RNA sequencing errors ($All_{mis0+1+2}$).

The proportions of the sequence reads that were perfectly mapped onto the BY genome and onto the RM genome are shown in Table S1; about 62% of sequence reads were mapped perfectly in every condition. A large proportion of these mapped sequence reads were mapped to single genomic locations and referred to as single-hit reads; the rest were multiple-hit reads. Table S2 shows that about 12% of perfectly matched sequences were multiple-hit reads. Moreover, when up to two

mismatches were allowed during the mapping process, we obtained extra ~5% useful sequence reads in every condition (Table S1). Putting these results together, we found that about two thirds of total reads in each sample were mapped onto either the BY genome or the RM genome (Table S1).

Detecting errors in the genomic sequences of the two strains

A number of genes with SNPs showed an extreme pattern in which the reads mapped on the SNP sites were detected only in one strain but not in the other, no matter what sample the data were from. These exceptional cases could be because these genes were expressed in only one strain or because the SNPs we identified actually reflected genomic sequence errors. There is also the formal possibility that a discrepancy arose from a mutation after the divergence of the strains used in this study and the strains used in the public genomic sequence databases; for simplicity, we called all such events "reference sequencing errors" instead of "recent mutations". We examined the genomic DNA of several of these SNPs by pyrosequencing analysis and confirmed that all the SNPs we tested were caused by reference sequencing errors. Following the pyrosequencing confirmation of these reference sequencing errors, we extended our analysis pipeline to examine all the SNPs sites that might fit the pattern described above. To detect genomic sequence errors in the reference genome sequences at the SNP sites we identified, we used the following bioinformatics approach.

Our strategy is to identify real error sites in the reference genomic sequences. First, if a site showed more than 10-fold change in the expression read count between the two strains, or had zero count in one strain and more than 10 counts in the other, it was regarded as a potential genomic sequence error site. According to these criteria derived from examining the expression data, 1,490 sites qualified as putative reference genome error sites. We then verified these sites against our gDNA IGA-II sequencing data; the data consisted of two channels of hybrid and two channels of co-culture gDNA IGA-II sequencing. The processing methods for the gDNA data were similar to the methods described in the section "Mapping IGA reads to the reference genomes" but only exact matches were allowed and the "wordsize" parameter of Megablast was set to 32 because the length of the gDNA sequence reads was 40 nt. A potential error site is categorized as either: (1) a real error site in a reference genome; (2) a non-error site; or, (3) an uncertain site. If the mapped gDNA showed the same absence of one allele as did the cDNA data, then it was considered a real error site. In this case

the expression polymorphism is not real but due to an error in the reference genomic sequence. On the other hand, if the mapped gDNA reads were detected in both strains, the expression signal observed in the cDNA was considered real. Finally, if there were insufficient gDNA reads mapped on the site in both strains, we regarded the site as ambiguous (uncertain). Of the 1,490 putative genome reference sequence error sites, 893, 540 and 57 sites were identified as real error sites, non-error sites and uncertain sites, respectively. For the 893 error sites, 309 were from the BY strain and 584 were from the RM strain. We then updated our genomic sequence databases for the 2 strains and removed these 893 error sites from the SNP list. We also removed the 57 uncertain sites from our SNP list. We then redid all the analysis using the updated genomic sequence databases and the updated SNP list.

Modeling gene expression as a discrete sampling process

In order to make inferences on read data provided by the Illumina Genome Analyzer, we formulated our statistical questions in terms of a discrete statistical framework. We sought to frame our questions, so that we could solve both the normalization problem and the expression parameter estimation problem by using the binomial distribution. Normalization is required because differences in total reads between samples occur whenever sampling effort is not evenly distributed between the samples, either due to design or experimental error. Such differences may be considered analogous to the systematic intensity differences encountered in comparing two microarrays. Indeed, some authors even recommend employing methods related to standard quantile normalization (Bolstad et al. 2003) for count data (Balwierz et al. 2009). Since we decided to model our data as a binomial sampling problem, using a normalization procedure that rescales the read counts would discard important information about the variance, which is already encoded in the read counts of a discrete sampling experiment. Additionally, explicitly modeling noise (Balwierz et al. 2009) for unambiguous counts derived from a discrete sampling process introduces a concept designed to account for a physical assay where real sources of noise are well attested, such as detection of light by analog sensors in scanners that cannot completely avoid light contamination. Given that for our data, each observed read that can be mapped is clearly derived from an identifiable region, we have

chosen to abandon legacy frameworks modeled on signal processing insights gained from microarray experiments. There are reasonable physical rationales for why noise exists in array scanning like: photons derived from cross hybridizing probes; strayed photons derived from leaked light hitting a CCD sensor; or photons derived from one probe hitting the part of the sensor devoted to another probe (Tu et al. 2002). In the case of deep sequencing, it is not clear which sequence reads comprise the "signal" and which comprise the "noise".

cis and *trans* parameter estimation

Consider a measurement of expression read counts for two alleles at the same locus in a single experiment. We assume that such counts can be viewed as a binomial sample such that the observed counts are the X, N variables of a binomial experiment having an underlying binomial parameter p. For a given locus, the parameter p characterizes the proportion of the read counts (X) from one allele in the total of read counts N from both alleles. This parameter is influenced both by the number of cells containing each allele and by the relative expression level per cell of each transcript. Let d represent the ratio of the number of cells containing one allele to the number containing the other. Let e represent the ratio of the expression level per cell in cells containing one allele to the expression level per cell in cells containing the other. In this case, the binomial parameter p is:

$$p_j = \frac{d_j e_j}{d_j e_j + 1},$$

where j represents the experiment of interest, and, in this study, can be either a comparison of two alleles measured in the two strains grown in the same culture (the "co-culture" experiment) or a comparison of those same two alleles measured in an F_1 hybrid (the "hybrid" experiment).

Estimation of the relative cell density parameter

To eliminate the identifiability problem between the d and e parameters, we first estimate d independently from genomic DNA with the following maximum likelihood estimator:

$$\hat{d}_j = \frac{\sum_{i=1}^{G} X_{i,j}}{\sum_{i=1}^{G} \left(N_{i,j} - X_{i,j} \right)}, \qquad (1)$$

where *i* indexes each gene in the genome and *G* is the total number of genes studied. For estimating the underlying d_j, the data considered were from the gDNA extracts from the same co-culture and hybrid experiments that produced the cDNA for the expression measurements.

There are two assumptions behind using the contrast between the co-culture and hybrid experiments to investigate *cis* and *trans* variation. The first is the observation that all freely diffusing factors in the hybrid experiment are presented equally to each allele at a heterozygous locus, and that allele-specific expression differences in a hybrid are due to *cis* variation only. This assumption is expressed as follows:

$$e_{Hy} = e_{cis}.$$

The other underlying assumption is that the allele-specific expression differences in the co-culture experiment are some combination of *cis* and *trans* effects such that, when no difference is observed in the hybrid experiment and a difference is observed in the co-culture experiment, all of the differences in the co-culture experiment are attributed to *trans*. Though such an assumption can be satisfied through many functional forms of varying complexity, one of the simplest functional forms is commonly assumed (at least implicitly):

$$e_{Co} = e_{cis}e_{trans}.$$

In fact, when genes are influenced by only a single mutation, the form above is accurate, as one parameter is fixed at unity (0 on a $\log_2$ scale) and the other parameter varies. In this special case, single mutations of the *cis* variety would fall along the diagonal of a $\log_2$ plot of hybrid versus co-culture, whereas mutations of the *trans* variety would fall along the axis defined by $\log_2(e_{Hy})=0$ (Wittkopp et al. 2004; Wittkopp et al. 2008).

Estimating expression parameters

Under the above assumptions, we can examine the hybrid and co-culture experiments in terms of *cis* and *trans* expression parameters e_{cis}, e_{trans} by rewriting the binomial p_{Hy}, p_{Co} parameters as follows:

$$p_{Hy} = \frac{d_{Hy}e_{Hy}}{d_{Hy}e_{Hy}+1} = \frac{d_{Hy}e_{cis}}{d_{Hy}e_{cis}+1},$$

$$p_{Co} = \frac{d_{Co}e_{Co}}{d_{Co}e_{Co}+1} = \frac{d_{Co}e_{cis}e_{trans}}{d_{Co}e_{cis}e_{trans}+1}.$$

Estimating cis

The hybrid experiment can be used to obtain the expectation and confidence interval for the parameter e_{Cis} from the following equation, using standard maximum likelihood methods:

$$L\left(e_{cis}|d_{Hy},X_{Hy},N_{Hy}\right)=\binom{N_{Hy}}{X_{Hy}}p_{Hy}{}^{X_{Hy}}\left(1-p_{Hy}\right)^{N_{Hy}-X_{Hy}}. \qquad (2)$$

We obtain p-values for hypothesis tests using the likelihood ratio test with one degree of freedom with the null hypothesis $e_{Cis} = 1$.

Estimating trans

All of the information about *trans* differences is found in the co-culture comparison. However, by using only co-culture data, another identifiability problem arises where the two expression parameters e_{cis} and e_{trans} cannot be distinguished or separated:

$$L\left(e_{Co}|d_{Co},X_{Co},N_{Co}\right)=L\left(e_{cis},e_{trans}|d_{Co},X_{Co},N_{Co}\right)=\binom{N_{Co}}{X_{Co}}p_{Co}^{X_{Co}}\left(1-p_{Co}\right)^{N_{Co}-X_{Co}}. \qquad (3)$$

This problem can be solved by simultaneously considering both the hybrid and co-culture data:

$$L\left(e_{cis},e_{trans}|d_{Hy},X_{Hy},N_{Hy},d_{Co},X_{Co},N_{Co}\right)=\binom{N_{Hy}}{X_{Hy}}p_{Hy}^{X_{Hy}}\left(1-p_{Hy}\right)^{N_{Hy}-X_{Hy}}\binom{N_{Co}}{X_{Co}}\left(1-p_{Co}\right)^{N_{Co}-X_{Co}}. \quad (4)$$

Using this formulation, the expectations, the confidence intervals, and hypothesis tests for the expression parameters can be obtained. This equation provides not only results for the *trans* parameters, but also results for the *cis* parameters that are identical to those from Equation 2.

Correlated errors in estimated expression parameters

Both e_{cis} and e_{trans} expression parameters depend on the hybrid data, though the *cis* expression parameter is independent of the co-culture data. As a result, the expression parameter estimates obtained from Equation 4 are negatively covarying, which is an important factor when considering the relationship between *cis* and *trans* estimates. When either *cis* or *trans* estimates are considered alone, the estimates from Equation 4 are appropriate (e.g. Figure 1; the *cis* and *trans* categories in Figure 3; Figure 4; Figure 5; and Table 1). However, when interpretations depend on inferences made simultaneously on *cis* and *trans* estimates, this co-variance creates a problem. Comparing Figure 2B and Supplemental Figure S2B illustrates the negative correlation introduce when the

errors are correlated. It is thus important to also obtain independent estimates. These independent estimates are intended for use whenever there is a direct relationship between *cis* and *trans* in the inferences we are making. For example, the inferences presented in Figure 2, and the *trans* – *cis* categories in Figure 3 involve inferences that could be misled by a direct relationship between *cis* and *trans*.

In order to accomplish these goals, we must make sure that data from which *cis* estimates are made are independent of data from which *trans* estimates are made. Thus, we assign half of the hybrid sequencing data (6 out of 12 total channels) to estimating the *cis* parameters and the other half to estimating the *trans* parameters. In turn, we also obtain two estimates of the *cis* expression parameter, one of which is independent of the estimate of the *trans* parameter and one of which is not. The following two equations illustrate how we formulate the independent estimates of the *cis* and *trans* parameters (the '.1' indicates the results from one partition of the hybrid data and the '.2' indicates the results from the complementary and independent partition):

Independent estimate of cis

$$p_{Hy.1} = \frac{d_{Hy}e_{Hy.1}}{d_{Hy}e_{Hy.1}+1} = \frac{d_{Hy}e_{cis.1}}{d_{Hy}e_{cis.1}+1},$$

$$L\left(e_{cis.1}\middle|d_{Hy.1},X_{Hy.1},N_{Hy.1}\right) = \binom{N_{Hy.1}}{X_{Hy.1}} p_{Hy.1}^{X_{Hy.1}}\left(1-p_{Hy.1}\right)^{N_{Hy.1}-X_{Hy.1}}. \qquad (5)$$

Independent estimation of trans

$$p_{Hy.2} = \frac{d_{Hy}e_{Hy.2}}{d_{Hy}e_{Hy.2}+1} = \frac{d_{Hy}e_{cis.2}}{d_{Hy}e_{cis.2}+1},$$

$$p_{Co} = \frac{d_{Co}e_{Co}}{d_{Co}e_{Co}+1} = \frac{d_{Co}e_{cis.2}e_{trans}}{d_{Co}e_{cis.2}e_{trans}+1},$$

$$L\left(e_{cis.2},e_{trans}\middle|d_{Hy.2},X_{Hy.2},N_{Hy.2},d_{Co},X_{Co},N_{Co}\right) = \binom{N_{Hy.2}}{X_{Hy.2}} p_{Hy.2}^{X_{Hy.2}}\left(1-p_{Hy.2}\right)^{N_{Hy.2}-X_{Hy.2}} \binom{N_{Co}}{X_{Co}} p_{Co}{}^{X_{Co}}\left(1-p_{Co}\right)^{N_{Co}-X_{Co}}. \quad (6)$$

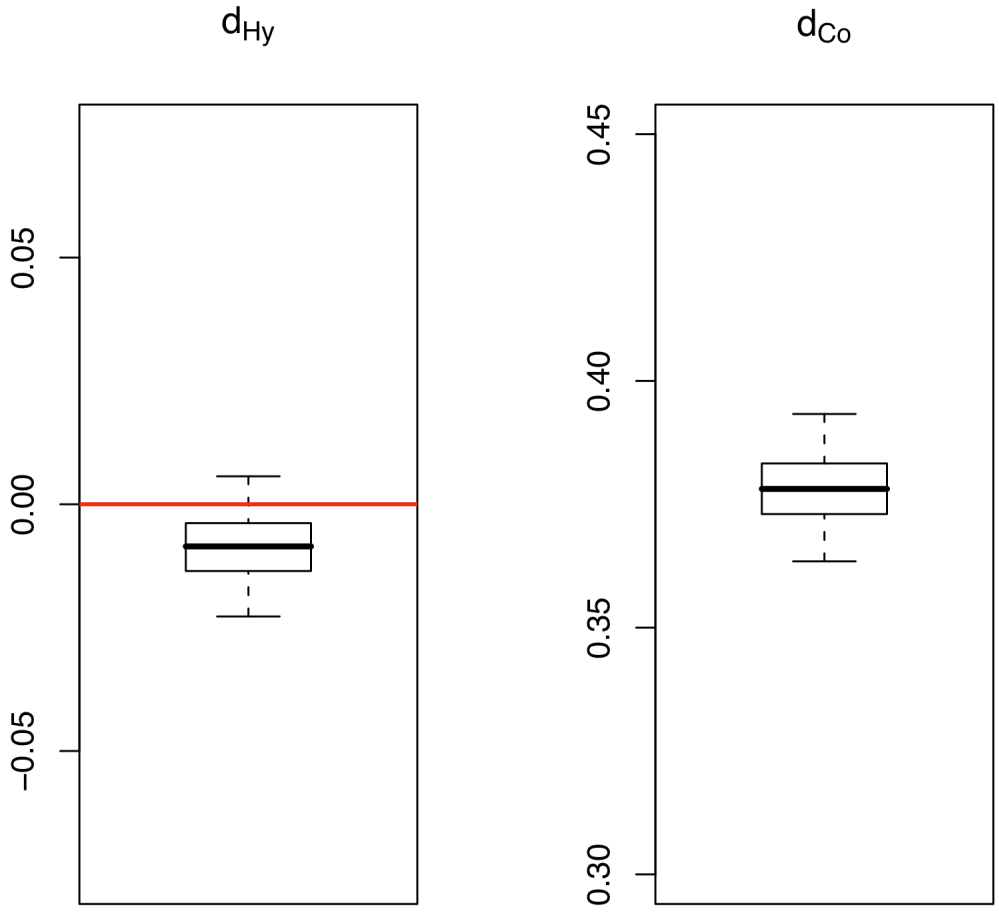
Estimates of *cis* ($e_{cis.1}$) and *trans* (e_{trans}) from Equations 5 & 6 above are used in place of estimates from Equations 3 & 4 above in Figure 2 and conclusions related to Figure 2. However, when

independent inferences are made with regard to the expression parameters, Equations 3 & 4 are preferred due to their greater power.

Supplemental Figures

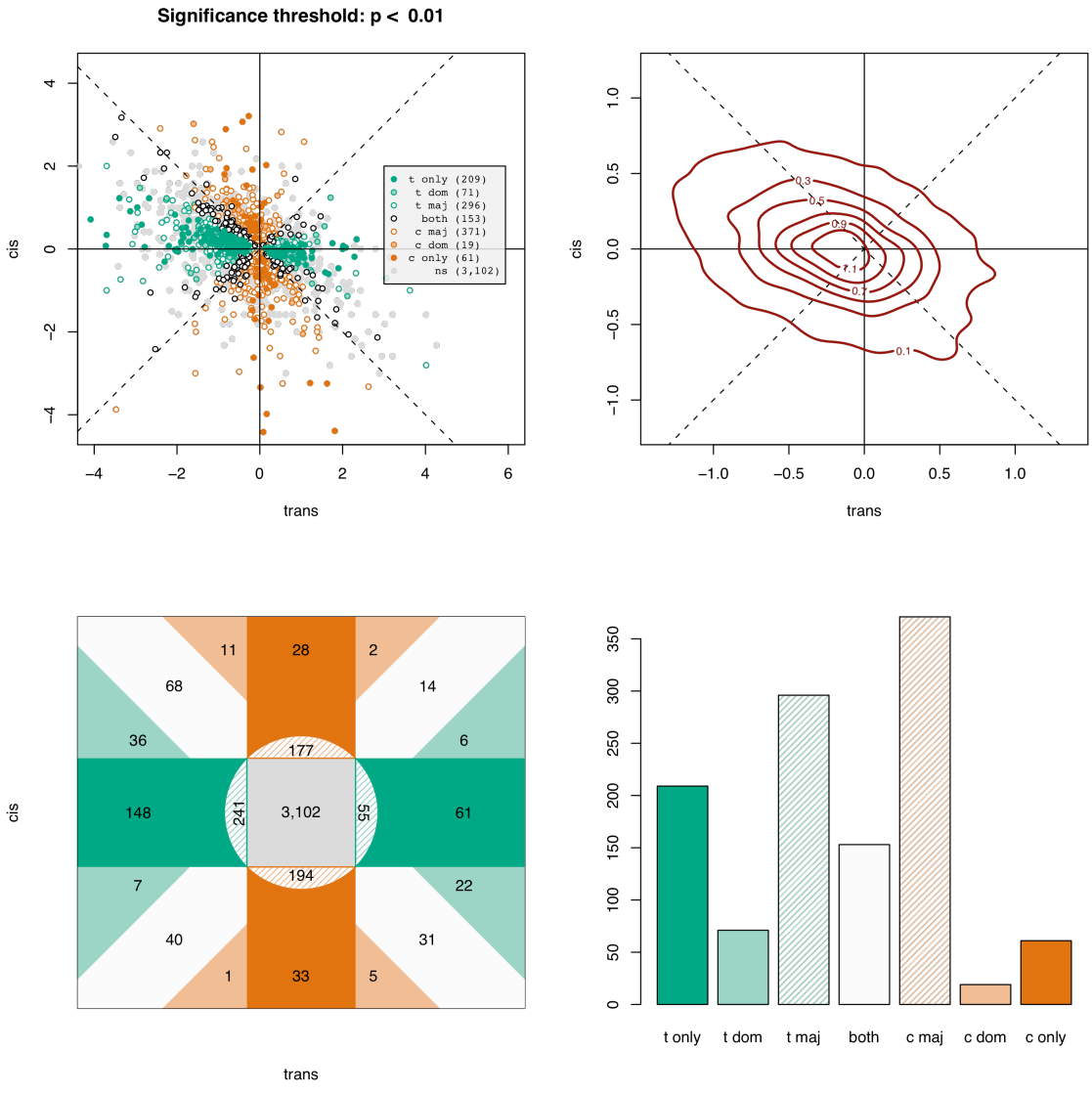
Supplemental Figure S1:

gDNA estimates of *d* for hybrid and co-culture experiments. The maximum likelihood estimate of *d* and its associated confidence intervals are displayed graphically. The boxes indicate the 50% confidence interval, while the whiskers indicate the 95% confidence intervals. The dark lines in the middle of the boxes indicate the MLE. The red horizontal line indicates $\log_2(d) = 0$, i.e., $d = 1$. Confidence intervals were established through bootstrapping. (A) The hybrid experiment. (B) The co-culture experiment.



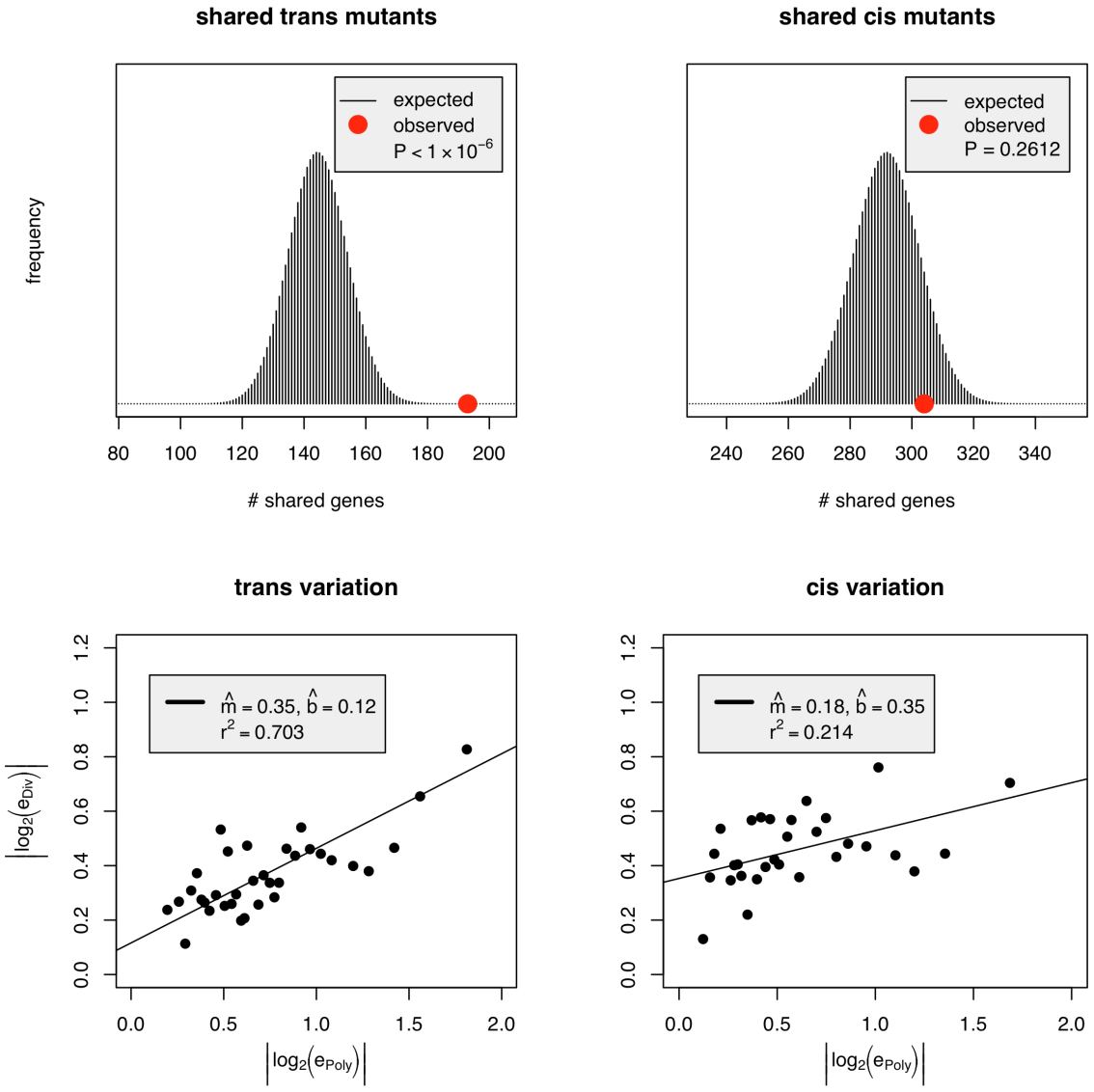
Supplemental Figure S2:

Genome-wide allele-specific expression polymorphism in *S. cerevisiae*. This figure follows Fig. 2 in the main text, except that the estimates follow Equations 3 & 4. The co-variance between *cis* and *trans* is clearly visible in panel B. The data as summarized in panel D shows the same general patterns demonstrated in Fig. 2, except that the two "major" categories are moderately affected by the co-variance between *cis* and *trans* estimates.



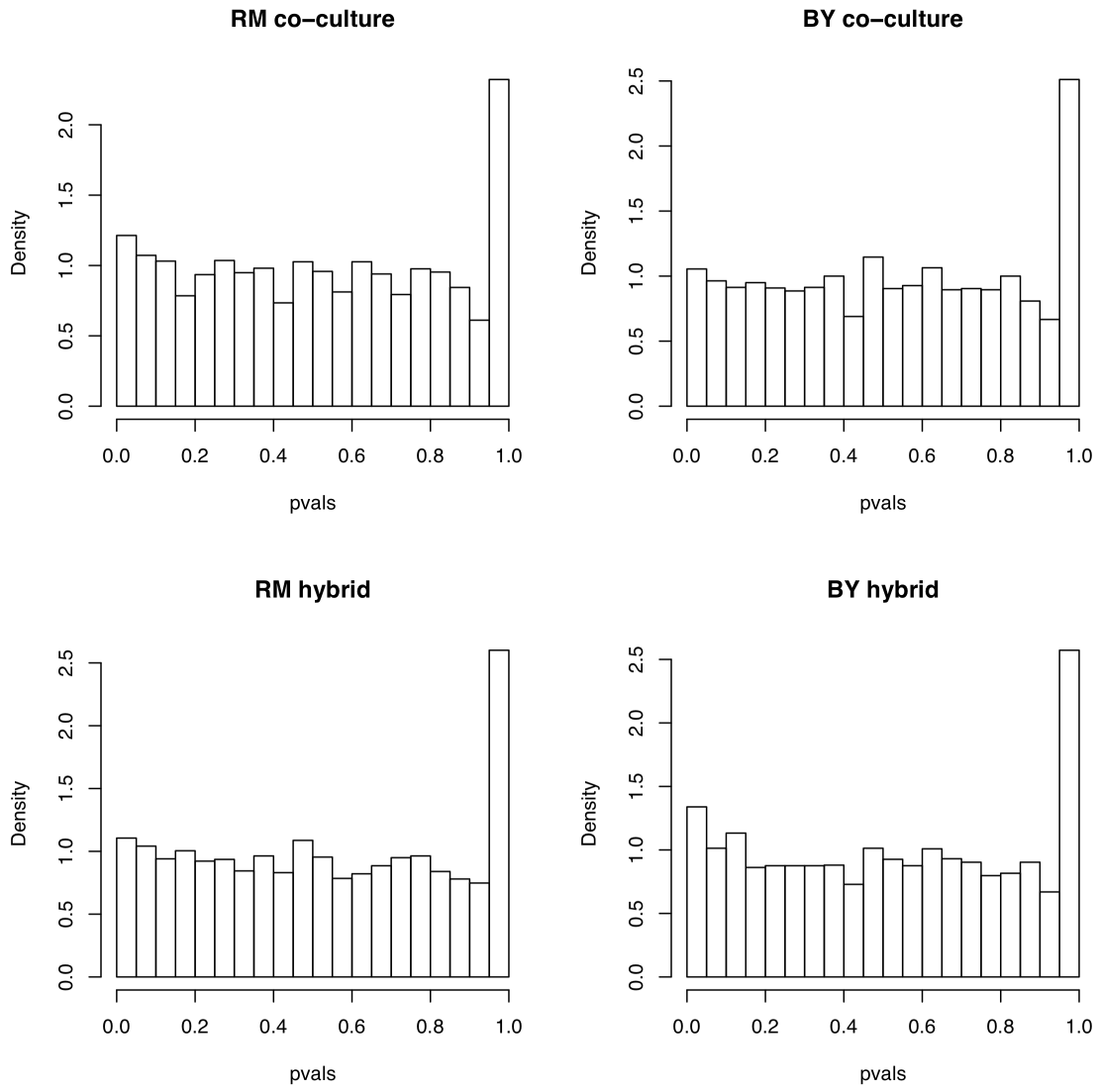
Supplemental Figure S3

Following Figure 6, except that the *trans* hotspot genes were not discarded.



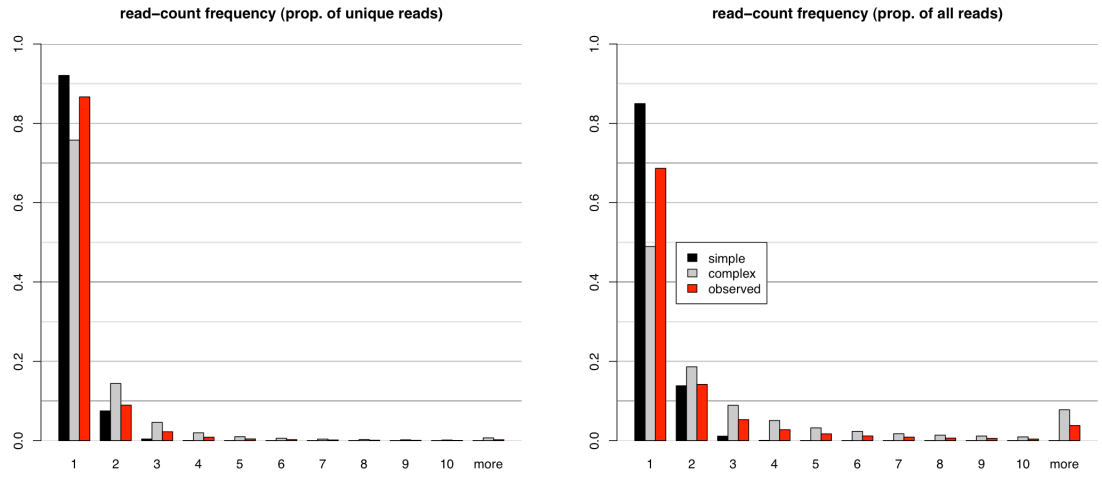
Supplemental Figure S4

Rejection rates for the null hypothesis for IGA transcriptome data. Each figure shows the p-values for conducting binomial tests between 2 lots of 6 channels for each experiment × strain combination. The most notable feature is the excess of p-values near 1, indicating a failure to reject the null hypothesis. This is due to the conservative nature of the binomial test. Otherwise, there is no indication that IGA transcriptome data leads to excessive rejection of the null hypothesis when the data is controlled such that there is no expression variation.



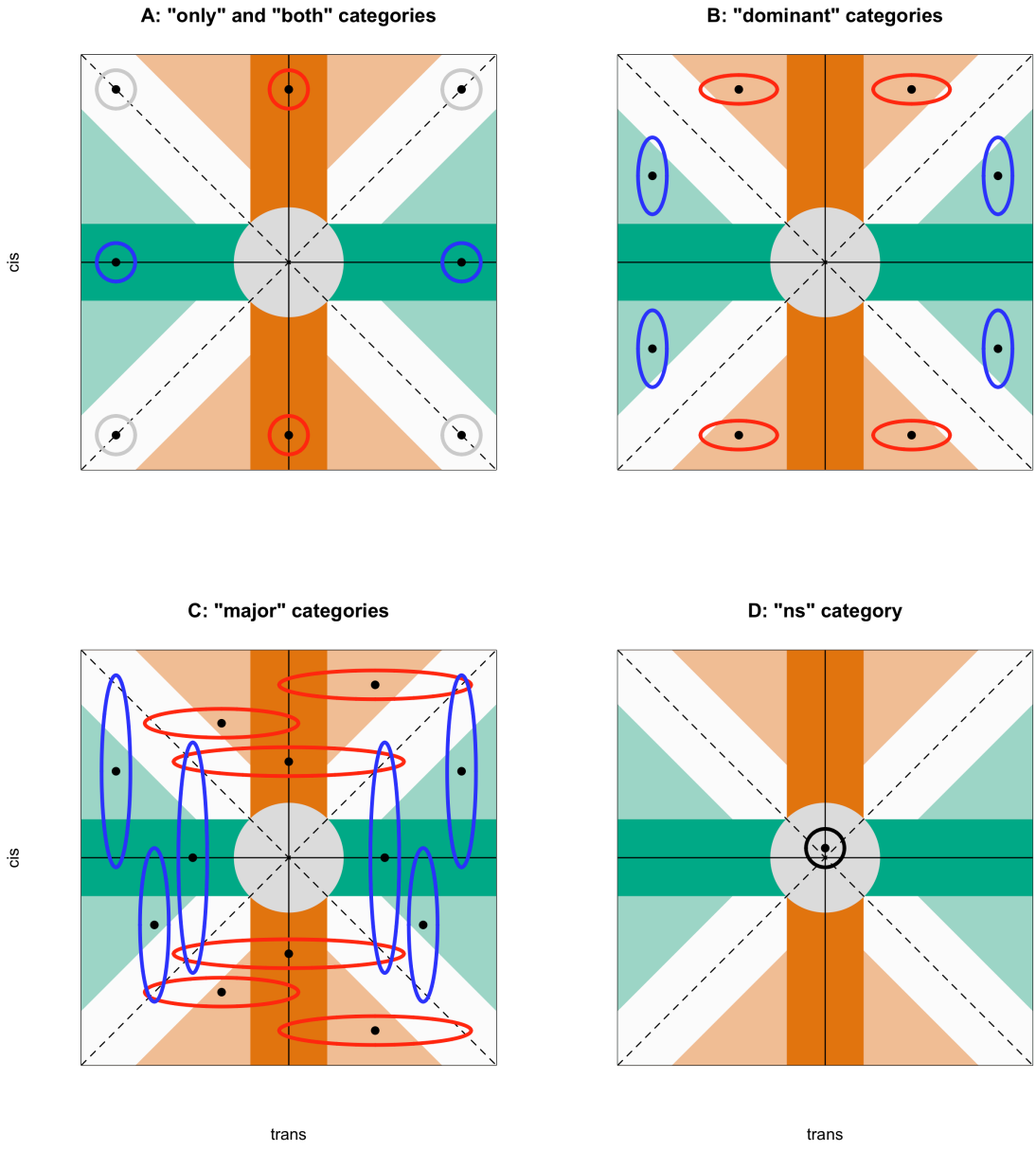
Supplemental Figure S5

Histograms of identical reads. Each histogram is a different representation of the same probabilities. The "simple" category is for an expression model that assumes that all genes have uniform expression and that no sites are more likely to be sequenced than other sites. The magnitudes of these bins are proportional to a Poisson distribution for k > 0. In a uniform sample of N reads from a genome of L nucleotides, the probability that a given nucleotide in the genome serves as the start position for an IGA read k times can be modeled by a Poisson distribution with parameter $\lambda = N/L$. The "complex" category is for an expression model that assumes that reads are sampled from individual genes in proportion to expression levels observed in our actual dataset, but a sampled read's position within each transcript is assumed to be distributed uniformly. Because this distribution cannot be modeled by a simple probability distribution function, as is the case for the "simple" model, the probabilities were calculated through simulation. The "observed" category is what we actually observe in our dataset for read counts for particular sequences. (A) Each bin represents a probability proportional to the number of times a unique read sequence is represented. Thus, if a unique read is present 1 time, a single count is added to the 1 bin and if a unique read is represented 5 times, a single count is added to the 5 bin. Thus, the normalizing factor is the total number of unique reads (as opposed to the total number of reads). (B) Each bin represents a probability proportional to the number of times a read sequence is represented. If a read is present 1 time, a single count is added to the 1 bin and if it is represented 5 times, 5 counts are added to the 5 bin. Thus, the normalizing factor is the total of all reads, not just the total number of unique reads as was done for panel A. Distributions for both panels have been scaled so that the probabilities for k > 0 sum to unity.



Supplemental Figure S6

Diagrams explaining how classifications in the main text are made. The points represent the MLE values for the (e_{trans}, e_{cis}) coordinate pairs. The ellipses encircling the points represent the confidence intervals for the expression difference estimates. (A) The expression difference confidence intervals of genes placed into the "only" or "both" categories must overlap one and only one axis or diagonal. red: "*cis* only" category; blue: "*trans* only" category; gray: "both" category. For the "only" genes, we have good evidence for variation in either *cis* or *trans*, but no evidence for the other. For the "both" genes, we have good evidence for variation in both, but we cannot determine which of *cis* or *trans* is larger. (B) The expression difference confidence intervals of genes placed into the "dominant" category must not overlap any of the axes or the diagonals. red: "*cis* dominant" category; blue: "*trans* dominant" category. For these genes, there is evidence for variation in both *cis* and *trans*. Additionally, we have evidence that one of the two is of greater magnitude than the other. (C) The expression difference confidence intervals of genes placed into the "major" category must overlap one and only one of the axes and at least one of the diagonals. red: "*cis* major" category; blue: "*trans* major" category. For these genes, there is clear evidence for variation in either *cis* or *trans* and equivocal evidence for the other. For example, for "*cis* major" genes, the confidence intervals for *trans* variation is sufficiently large that they neither reject $|\log_2(e_{cis})| = |\log_2(e_{trans})|$ nor $\log_2(e_{trans}) = 0$. (D) The expression difference confidence intervals of genes placed into the "ns" category (not significant) must overlap both of the axes. For these genes, there is no evidence of expression variation.


A: "only" and "both" categories
B: "dominant" categories
C: "major" categories
D: "ns" category
cis
cis
trans
trans

Supplemental Tables

Supplemental Table S1. Statistics of the reads that can be mapped on the genomes

Library	Reference genome	% of total reads	
		Mis_0[a]	**Mis_{0+1+2}[b]**
Hybrid	**BY**	62.84	67.05
	RM	62.53	66.73
Co-culture	**BY**	62.93	66.73
	RM	62.78	66.60

[a] Perfectly mapped reads
[b] Mapped reads allowing up to two mismatches

Supplemental Table S2. Proportions of single-hit and multiple-hit reads for each condition

Library	Reference genome	% of mapped reads	
		Single[a]	**Multiple[b]**
Hybrid	**BY**	87.81	12.19
	RM	88.00	12.00
Co-culture	**BY**	87.06	12.04
	RM	87.91	12.09

[a] Single-hit reads
[b] Multiple-hit reads

Supplemental Table S3
Summary of the results of applying the bioinformatics filters to the number of genes in our sample. The table describes the series of bioinformatics filters applied to the read count data before it was subjected to statistical analysis. The largest number of genes lost prior to analysis is for genes that lack SNPs (1,294). Absence of variation in the transcript of interest is the largest challenge in ASE experiments for very closely related strains.

Gene number	Filter criteria
6,604	Genes with UTR information (from Nagalakshmi 2008)
6,520	After discarding genes lacking unambiguous homology and reliable alignment between BY and RM (84)
6,307	After discarding genes lacking good match coverage, unambiguous one-to-one orthology or good "synteny" relationship (213)
6,075	After discarding lowly expressed genes (232)
5,860	After discarding embedded genes (215)
4,566	After discarding genes lacking SNPs in the alignment (1,294)
4,442	After discarding genes represented by a non-negligible proportion of reads mapped to multiple genomic locations (124)

Supplemental Table S4:

This table follows Table 1 in the main text, except that the uncorrelated estimates are used instead. (A) P-value < 2.2×10^{-16}. (B) P-value = 0.01745. (C) P-value = 1.

A:

	Polymorphism	Divergence
Cis	248	1,270
Trans	314	541

B:

Trans	Poly. Sig	Poly. Nsig
Div. sig	89 (72.2)	452 (468.8)
Div. nsig	225 (242.8)	1,586 (1,569.2)

C:

Cis	Poly. Sig	Poly. Nsig
Div. sig	134 (133.9)	1,136 (1,136.1)
Div. nsig	114 (114.1)	968 (967.9)

Dataset 1 – Dependent Estimates:
The supplemental dataset is a tab-delimited text file. The first line is a header line naming each of the fields. The following is a list of the fields and a description of its contents:

name: the ORF name following the SGD naming convention;
etrans: MLE for $\log_2(e_{trans})$;
ecis: MLE for $\log_2(e_{cis})$;
cat: significance category that the genes fall into, following Figure 2 and Supplemental Figure S2;
X.RM.Co: the read counts for the RM allele in the co-culture experiment summed over all 12 lanes;
X.BY.Co: the read counts for the BY allele in the co-culture experiment summed over all 12 lanes;
X.RM.Hy: the read counts for the RM allele in the hybrid experiment summed over all 12 lanes;
X.BY.Hy: the read counts for the BY allele in the hybrid experiment summed over all 12 lanes.

Dataset 2 – Independent Estimates:
The description of Supplemental Dataset 2 follows that of Supplemental Dataset 1 with the following differences:

X.RM.Hy.1: the read counts for the RM allele in the hybrid experiment summed over the 6 lanes of sequencing used to estimate e_{cis} independently;
X.BY.Hy.1: the read counts for the BY allele in the hybrid experiment summed over the 6 lanes of sequencing used to estimate e_{cis} independently;
X.RM.Hy.2: the read counts for the RM allele in the hybrid experiment summed over the 6 lanes of sequencing used to estimate e_{trans} independently;
X.BY.Hy.2: the read counts for the BY allele in the hybrid experiment summed over the 6 lanes of sequencing used to estimate e_{trans} independently.