Supporting Information

Bullard et al. 10.1073/pnas.0912959107

SI Text

Identifying Ortholog Pairs with Differential Allele-Specific Expression. A natural measure of differential expression between orthologs is the within-lane log-ratio of their per-base sequencing read counts. In the present experiment, by virtue of the paired design (interspecies hybrid assayed within each lane), technical effects, such as differences across lanes, flow-cells, or sequencing centers, were mitigated by forming such relative measures of expression between orthologs within a given lane. Therefore, we employed as statistics for differential expression (DE) the averages of the within-lane started log-ratios of per-base read counts for biological replicates 1 and 2, respectively. The started (natural) logarithm, defined as $slog(x) = ln(x+1)$, was used to handle zero counts.

We sought to generate sequence-specific null distributions that accounted for the influence of differing starting bases of a read (e.g., G and C bases) on its sequenceability. For this purpose, we resampled, for each ortholog pair, the base-level read counts of the two alleles to form "null" ortholog pairs with no differential expression and the same marginal nucleotide distributions as the original orthologs (Fig. 1). Specifically, for a given gene, let L_b and L_c denote, respectively, the lengths of the *S. bayanus* and *S. cerevisiae* orthologs [only uniquely-mappable (2) bases are counted], and let $\pi_b = [\pi_b(A),\pi_b(C),\pi_b(G),\pi_b(T)]$ and $\pi_c = [\pi_c(A),\pi_c(C),\pi_c(G),\pi_c(T)]$ denote, respectively, the marginal nucleotide frequencies of the *S. bayanus* and *S. cerevisiae* orthologs. We began by resampling the original base-level read counts of the *S. bayanus* ortholog. To create null *S. bayanus* read counts, we sampled L_b base-level read counts at random, with replacement, with uniform probability $1/L_b$. Next, to create null *S. cerevisiae* read counts, we sampled, from the original *S. bayanus* ortholog, L_c base-level read counts at random, with replacement, with nucleotide-specific probabilities $\pi_c/(L_b\,\pi_b)$. For each such ortholog pair, we computed a null DE statistic, defined as the average (over all lanes for biological replicates 1 and 2, respectively) log-ratio of *S. bayanus* to *S. cerevisiae* null per-base read counts. Repeating the above procedure 10,000 times yielded a null distribution of DE statistics, based on the original *S. bayanus* read counts, which preserved the marginal nucleotide distributions of the two orthologs, and to which we compared the observed DE statistic to obtain a two-sided *P* value. We then repeated the entire resampling procedure using the base-level read counts of the *S. cerevisiae* ortholog. Finally, we conservatively retained the maximum of the *P* values based on the *S. bayanus* and *S. cerevisiae* resampled counts. The complete procedure yielded, for each of the $J = 4{,}238$ uniquely-mappable (2) ortholog pairs, an average within-lane paired log-ratio f_j of *S. bayanus* to *S. cerevisiae* per-base read counts and its associated *P* value P_j, assessing the significance of the differential allele-specific expression, $j = 1,\ldots,J$. We applied the procedure separately to each of the two biological replicates.

We note that this method is general and modular, in that it can be applied to other DE statistics (e.g., paired *t*-statistics, GLM-based *t*-statistics) and sequence features (e.g., dinucleotide frequencies).

Quantitative RT-PCR. For Fig. S2, pairs of orthologs were amplified separately, each using allele-specific primers designed with Primer3 (http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi). The web edition of Beacon Designer (http://www.premierbiosoft.com/qOligo/Oligo.jsp?PID=1) was used to screen for hairpins and primer dimerization, and BLAST (http://www.ncbi.nlm.nih.gov/BLAST) was used to check that candidate primers were specific to a single locus within the combined *S. bayanus* and *S. cerevisiae* genomes. Primers were synthesized by Elim Biopharmaceuticals.

DNase-treated total RNA from biological replicate 1 was converted to cDNA using the SuperScript III First Strand Synthesis System with oligo(dT) primers (Invitrogen), following the manufacturer's instructions. RT-PCR reactions were performed using the DyNAmo HS SYBR Green qPCR mix (Finnzymes) in 96-well plates on a Stratagene Mx3000p thermocycler. To control for differences in amplification efficiency between primers, expression levels for each ortholog were normalized based on a primer-specific standard curve of genomic DNA. Genomic DNA was isolated from the OZY27 hybrid as described (3).

Genes to be tested by RT-PCR were selected to span a wide range of expression fold-changes and *P* values while still showing significant differences in allele-specific expression levels according to the procedure in Fig. 1. For each of 22 genes (Table S3), we performed three cDNA amplifications and two standard curve replicates per ortholog. For Fig. S2, the resulting fold-changes from RT-PCR were compared with fold-changes measured by RNA-seq for biological replicate 1.

Testing for Directional Imbalance in *cis*-Regulatory Effects Across Pathways. For a given biological replicate and gene $j = 1,\ldots,J$, a simple and robust measure of directional differential expression is given by the product $sign(f_j)\ I(p_j \leq \alpha)$. This statistic takes on three values: 1 for genes significantly up-regulated in *S. bayanus*, −1 for genes significantly up-regulated in *S. cerevisiae*, and 0 otherwise. We assessed significance at a common single test level $\alpha = 0.05$. We defined a conservative, combined directional DE statistic S_j for the two biological replicates as the common value of the statistic if both replicates agreed and zero otherwise. For analyses of coregulated gene groups, we used the regulons defined in (1), eliminating groups with fewer than five genes and those with at least 50% overlap with another group (for the latter, we retained the larger of the two groups). For the Gene Ontology groups, we used the Biological Process categories from the GO slim annotations maintained by the Saccharomyces Genome Database (http://downloads.yeastgenome.org/literature_curation/go_slim_mapping.tab). For a given pathway, we measured the extent of directional differential expression by adding the gene-level statistics S_j of the pathway members. Let n_k and T_k denote, respectively, the cardinality and directional DE statistic of the *k*th pathway, $k = 1,\ldots,K$. The K=167 gene groups defined by coregulation and the K=38 groups defined by Gene Ontology were analyzed separately. For each set, we assessed the significance of the directional DE for pathway *k* by comparing the observed statistic T_k with a null distribution of 100,000 such statistics obtained by resampling groups of n_k genes at random, without replacement from the full set of *J* genes. Let P_k denote the resulting (unadjusted) two-sided *P* value. At a given unadjusted *P* value threshold P_0, we estimated the expected number of false-positive pathways as the product of P_0 and the number of tests; we set P_0 to attain as close to one false-positive pathway per set as possible. The directional DE statistics T_k and *P* values P_k are reported in Table 1. For comparison, results from the Benjamini–Hochberg multiple testing procedure (2), which controls the false-discovery rate, are given in Table S8.

We note that this directional DE method is general and modular, in that it can be applied to other gene-level statistics [e.g., $Sj = sign(f_j)\ (1 - p_j)$] and gene-group-level statistics (e.g., quantiles of S_j).

1. Gasch AP, et al. (2004) Conservation and evolution of *cis*-regulatory systems in ascomycete fungi. *PLoS Biol* 2:e398.
2. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol* 57:289–300.
3. Hoffman CS, Winston F (1987) A ten-minute DNA preparation from yeast efficiently releases autonomous plasmids for transformation of Escherichia coli. *Gene* 57:267–272.

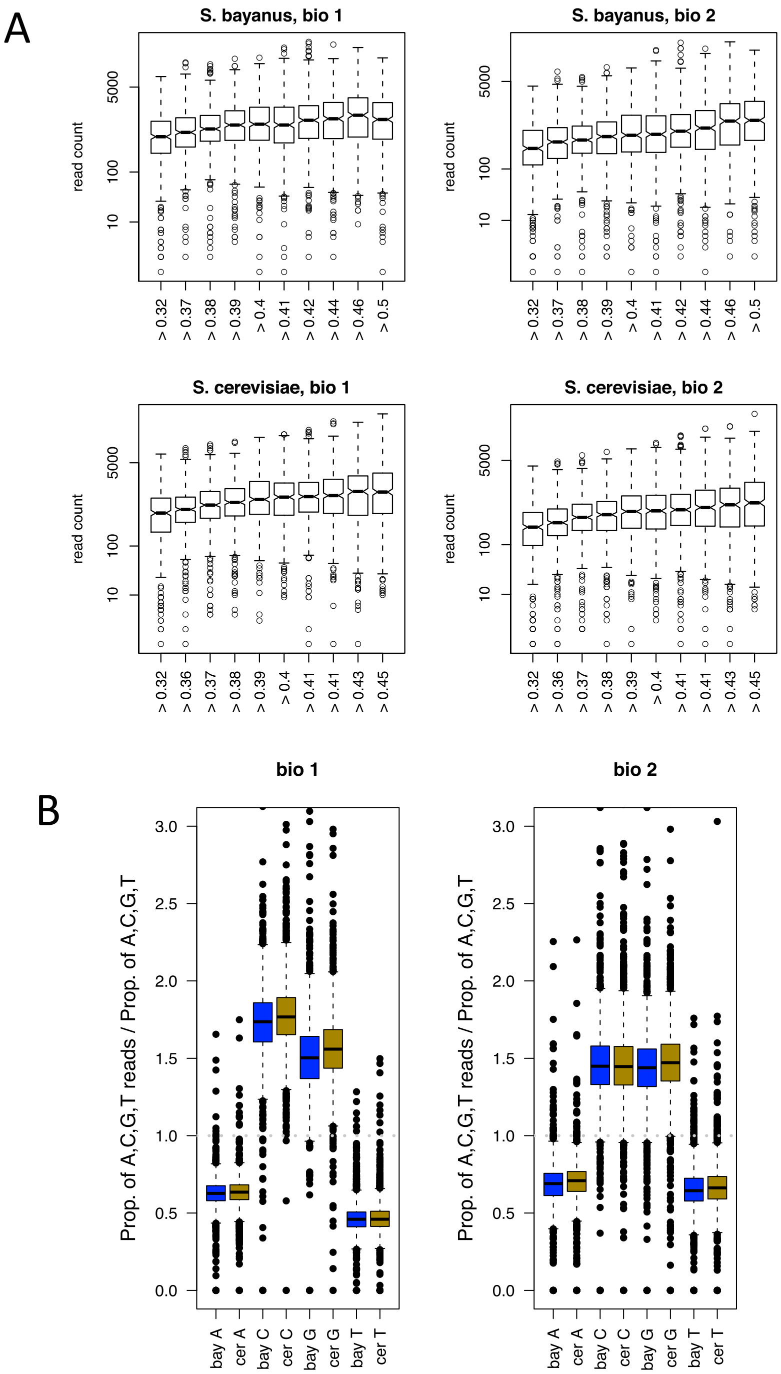

Fig. S1. Effect of base composition on RNA-seq read counts. (*A*) Effect of GC content on RNA-seq read counts. The set of *S. bayanus* and *S. cerevisiae* orthologs for 4,238 genes were each partitioned into 10 equally-sized groups according to GC content. In each boxplot, the *y* axis reports the allele-specific RNA-seq read counts across the orthologs in one such group, and the *x* axis gives the GC content of the group. (*B*) Sequenceability by initial nucleotide. The boxplots display the distribution across orthologs of a ratio of two proportions: For each ortholog, the numerator is the proportion of allele-specific reads starting with each of the four nucleotides, and the denominator is the nucleotide frequency. This ratio was calculated for each of 1,000 randomly chosen ortholog pairs with more than 1,000 reads across all eight lanes. In each panel, "bio1" and "bio2" refer to the two biological replicates.

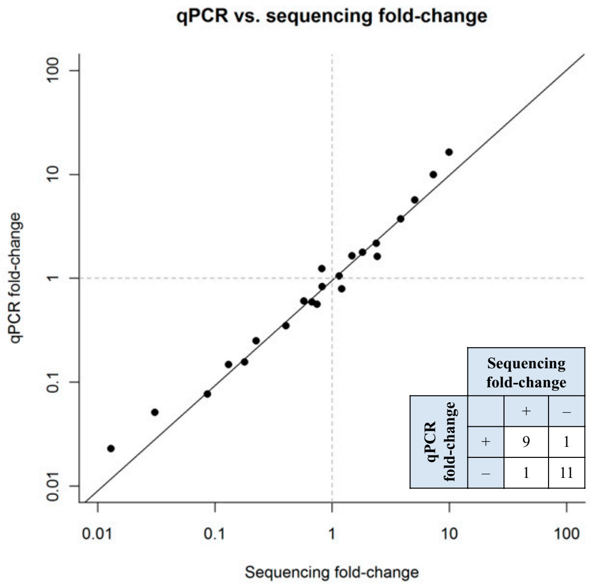

Fig. S2. Comparison of differential allele-specific expression measures from RNA-seq and RT-PCR. Each point represents a gene with differential expression of the *S. bayanus* and *S. cerevisiae* alleles at significance level 0.05 according to the procedure in Fig. 1. The *x* axis corresponds to the fold-change measured by RNA-seq, defined as the average across lanes of the ratio of *S. bayanus* to *S. cerevisiae* per-base read counts. The *y* axis corresponds to the fold-change between the alleles measured by RT-PCR. Raw data are given in Table S3. (*Inset*) Summary of data from the main plot in terms of the concordance of the inferred direction of differential expression between alleles in a given ortholog pair for the two assays. +, up-regulation of the *S. bayanus* allele; –, up-regulation of the *S. cerevisiae* allele. For instance, the top left cell indicates that at nine genes, the *S. bayanus* allele was associated with higher expression by both RT-PCR and RNA-seq.

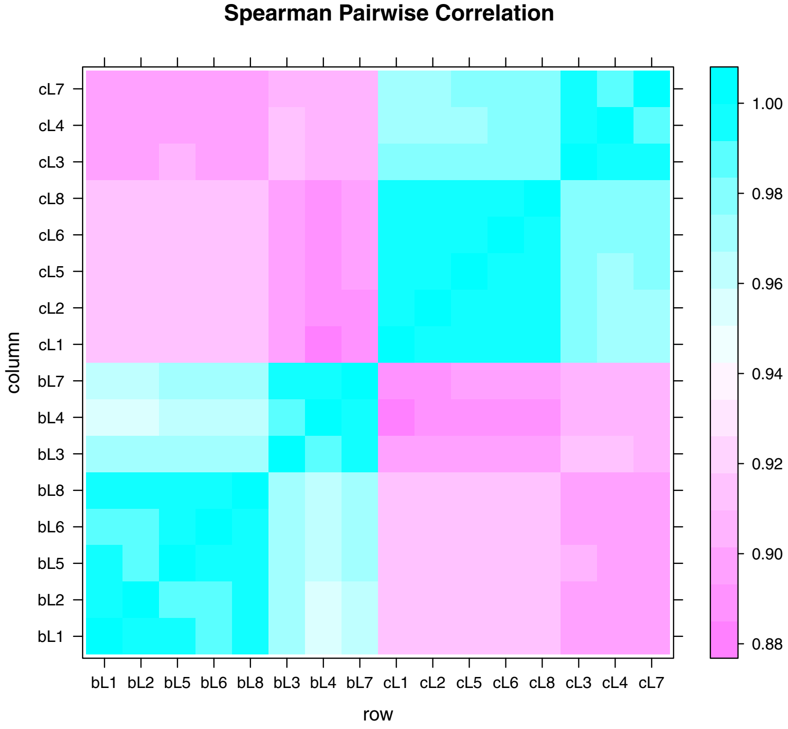

Fig. S3. Technical and biological variation in RNA-seq allele-specific expression data. Colors represent pairwise Spearman correlation coefficients between allele-specific RNA-seq read counts for 4,238 genes, for different combinations of orthologs, lanes, biological replicates, and sequencing centers. Legend is at far right. b, reads mapping to *S. bayanus*; c, reads mapping to *S. cerevisiae*; L, lane. Lanes 1, 2, 5, 6, and 8 were samples from biological replicate 1, and the remainder from biological replicate 2. Lanes 1–5 were run at the Vincent J. Coates Genomic Sequencing Laboratory, University of California-Berkeley, and the remainder were run at the Core Instrumentation Facility, University of California-Riverside.

Other Supporting Information Files

Table S1. Comparing methods to assess significance of differential expression in RNA-seq

Table S1 (DOC)

Each cell represents the number of ortholog pairs with the indicated qualitative measure of differential expression between alleles as called by either of two procedures to assess statistical significance. “Resampling” gives data from the procedure shown in Fig. 1 of the main text. “glm” gives results from an analysis in which *P* values for differential expression between the orthologs were assigned based on a Poisson generalized linear model (R function glm) taking as input the number of reads mapping to each ortholog and the total number of mapped reads in each Solexa sequencing lane. +1 represents ortholog pairs for which the *S. bayanus* allele was expressed significantly higher than the *S. cerevisiae* allele at $P < 0.05$, −1 represents ortholog pairs for which the *S. cerevisiae* allele was expressed significantly higher than the *S. bayanus* allele, and 0 represents all other ortholog pairs.

Table S2. RNA-seq data

Table S2 (DOC)

Bay, total reads mapping uniquely to *S. bayanus* in the indicated Solexa lane (assignment of samples to lanes is listed in the caption to Fig. S3); Bay.bio, sum of all reads mapping uniquely to *S. bayanus* across all lanes of the indicated biological replicate; benj.p, adjusted *P* value for max(bio.1.marginal, bio.2.marginal) from the Benjamini–Hochberg procedure; Cer, Total reads mapping uniquely to *S. cerevisiae* in the indicated Solexa lane; Cer.bio, sum of all reads mapping uniquely to *S. cerevisiae* across all lanes of the indicated biological replicate; gc.diff, proportion of bases in the *S. cerevisiae* ortholog that are Gs or Cs, subtracted from the proportion of bases in the *S. bayanus* ortholog that are Gs or Cs; Logratio, started logarithm of the ratio of average per-base allele-specific read counts across lanes for *S. bayanus* over that for *S. cerevisiae*, for the indicated biological replicate; Mappable lengths, number of bases in the indicated ortholog whose mapped reads were usable according to the criteria described in *Materials and Methods*; Marginal, *P* value for differential allele-specific expression from the indicated biological replicate according to the procedure described in *Materials and Methods*; Ortholog.length, raw number of bases of the indicated ortholog before processing; sc_0.05, indicator variable for significance across replicates at $P = 0.05$; Stats, sign of differential expression from the indicated biological replicate: +1 if level is higher in the *S. bayanus* allele than in *S. cerevisiae*, −1 if level is higher in *S. cerevisiae* than in *S. bayanus*.

Table S3. RT-PCR data

Table S3 (DOC)

qPCR fold-change: ratio of inferred concentration of *S. bayanus* allele relative to inferred concentration of *S. cerevisiae* allele measured from RT-PCR. RNA-seq *P* value: bio.1.marginal in Table S2; RNA-seq fold-change: per-base read counts for the *S. bayanus* allele, divided by the per-base read counts for the *S. cerevisiae* allele, averaged across lanes for biological replicate 1.

Table S4. McDonald–Kreitman gene tests (*A*) and McDonald–Kreitman-like gene tests on upstream regions (*B*)

Table S4 (DOC)

(*A*) NSfix, number of nonsynonymous fixed sites; NSpoly, number of nonsynonymous polymorphisms; P, *P* value from Fisher’s exact test; Sfix, number of synonymous fixed sites; Spoly, number of synonymous polymorphisms. (*B*) noncoding_fixed, number of fixed differences in the region 200 bp upstream of coding start; noncoding_polymorph, number of polymorphisms in the region 200 bp upstream of coding start; P, *P* value from Fisher’s exact test; silent_fixed, number of fixed synonymous changes in the ORF; silent_polymorph, number of synonymous polymorphisms in the ORF.

Table S5. Enrichment in coregulated gene groups for genes with significant evidence of nonneutral evolution (*A*) and enrichment in yeast GO_slim groups for genes with significant evidence of nonneutral evolution (*B*)

Table S5 (DOC)

(*A*) Group compositions are from (1). benj.p, adjusted *P* value from the Benjamini–Hochberg procedure; enrich, the ratio between n.signif.gasch and the number of genes in all pathways scored as significant (= 61), divided by the ratio between n.gasch.tested and the number of genes in all pathways in the McDonald–Kreitman dataset (= 1,366); n.gasch.tested, number of all genes in the McDonald–Kreitman dataset falling into the indicated pathway; n.signif.gasch, number of genes in the indicated pathway whose McDonald–Kreitman *P* value (Table S4*A*) fell below 0.005; p.hyper, results of a hypergeometric test for overrepresentation in the indicated pathway of genes whose McDonald–Kreitman *P* value (Table S4*A*) fell below 0.005. (*B*). Group compositions are from http://downloads.yeastgenome.org/literature_curation/go_slim_mapping.tab. enrich, the ratio between n.signif.GO and the number of genes in all pathways scored as significant (= 91), divided by the ratio between n.GO.tested and the number of genes in all pathways in the McDonald–Kreitman dataset (= 1,956); benj.p, adjusted *p*-value from the Benjamini–Hochberg procedure; n.GO.tested, number of all genes in the McDonald–Kreitman dataset falling into the indicated pathway; n.signif.GO, number of genes in the indicated pathway whose McDonald–Kreitman *P* value (Table S4*A*) fell below 0.005; p.hyper, results of a hypergeometric test for overrepresentation in the indicated pathway of genes whose McDonald–Kreitman *P* value (Table S4*A*) fell below 0.005.

Table S6. McDonald–Kreitman gene tests and McDonald–Kreitman-like tests on upstream regions, eliminating singleton sites

Table S6 (DOC)

Columns and categories are as for Table S5, except that McDonald–Kreitman tests were performed on the subset of sites at which the minor allele was present in >1 *S. cerevisiae* strain.

Table S7. Enrichment in pathways for genes with significant evidence of nonneutral evolution, eliminating singleton sites

Table S7 (DOC)

Columns and categories are as for Table S5, except that 56 genes were used that had $P < 0.005$ in McDonald–Kreitman tests of coding variation when singleton sites were eliminated (see Table S6*A*).

Table S8. Testing coregulated gene groups for directional evolution of gene expression (*A*) and testing yeast GO_slim groups for directional evolution of gene expression (*B*)

Table S8 (DOC)

(*A*) Group compositions are from ref. 1. benj.p, adjusted *p*-value from the Benjamini–Hochberg procedure; Ngenes, number of genes in the group with observable allele-specific expression; q.005, value of the statistic corresponding to the 5% quantile in resampled data; q.995, value of the statistic corresponding to the 99.5% quantile in resampled data; raw.p, *P* value from resampling for significance of imbalance across the gene group, as described in *Materials and Methods*; Stat, sum of direction scores across genes in the group. (*B*) Group compositions are from http://downloads.yeastgenome.org/literature_curation/go_slim_mapping.tab. Column headings are as in *A*.

1. Gasch AP, et al. (2004) Conservation and evolution of *cis*-regulatory systems in ascomycete fungi. *PLoS Biol* 2:e398.