Uncovering Genetic Regulatory Network Divergence Between Duplicate Genes Using Yeast eQTL Landscape

YANGYUN ZOU[1], ZHIXI SU[1], JIAN YANG[2], YANWU ZENG[1], AND XUN GU[1,3,4]*
[1]*Institutes of Biomedical Sciences, Fudan University, Shanghai, China*
[2]*Queensland Statistical Genetics, Queensland Institute of Medical Research, Herston, Queensland, Australia*
[3]*Department of Genetics, Development, and Cell Biology, Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, Iowa*
[4]*School of Life Sciences, Center for Evolutionary Biology, Fudan University, Shanghai, China*

ABSTRACT Genetical genomics, a novel approach combining microarray technology and quantitative genetic analysis, aims to identify the expression quantitative trait loci (eQTLs), which may regulate the genome-wide expression pattern. In this article, we have studied yeast genomic eQTL data to investigate how the genetic eQTL regulation of ancestral gene has diverged since gene duplication. Our findings are as follows: (i) Duplicate genes have higher heritability for gene expression than single-copy genes, but little difference in their epistasis and directional effect. (ii) The divergence of *trans*-acting eQTLs between duplicate pairs increases with the evolutionary time since gene duplication. (iii) *Trans*-acting eQTL divergence can explain about 21% of the variation in expression divergence between duplicate pairs with $K_S<2.0$, which increases to 27% when the transcription factor (TF)–target interaction divergence is combined. Moreover, under the partial correlation analysis, *trans*-acting eQTL divergence seems make a bigger contribution to expression divergence than does TF divergence. (iv) *Trans*-acting eQTL divergence between duplicate pairs is correlated with gene ontology categories ''Biological processes'' and ''Cellular components,'' but not with ''Molecular functions,'' and is related to fitness defect under treatment conditions, but not with fitness under normal condition. We conclude that eQTL analysis provides a novel approach to explore the effect of gene duplications on the genetic regulatory network. *J. Exp. Zool. (Mol. Dev. Evol.) 312B:722–733, 2009.* © 2009 Wiley-Liss, Inc.

How to cite this article: Zou Y, Su Z, Yang J, Zeng Y, Gu X. 2009. Uncovering genetic regulatory network divergence between duplicate genes using yeast eQTL landscape. J. Exp. Zool. (Mol. Dev. Evol.) 312B:722–733.

Jansen and Nap (2001) proposed a novel strategy termed as ''genetical genomics'' to address quantitative variation in gene expression. This approach can be used to identify the potential loci (expression quantitative trait loci, eQTLs), which are responsible for regulating the expression of target genes (Gibson and Weir, 2005). Recent technological advances in microarrays have made its large-scale application feasible at the level of genomes. Numerous applications on yeast (Brem et al., 2002; Bing and Hoeschele, 2005; Brem and Kruglyak, 2005), fish (Oleksiak et al., 2002), tree (Kirst et al., 2005), maize (Schadt et al., 2003), mouse (Schadt et al., 2003; Bystrykh et al., 2005) and human (Yan et al., 2002; Cheung et al., 2003; Schadt et al., 2003; Morley et al., 2004) have showed that both *cis*-acting and

Grant sponsor: National High-Tech Research and Development Program of China; Grant number: 2006AA02Z326; Grant sponsor: National Science Foundation of China (NSFC) Oversea Outstanding Young Investigator Award; Grant number: 30428003; Grant sponsor: National Science Foundation of China (NSFC); Grant number: 30700140.

The first two authors contributed equally to this work.

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Xun Gu, Department of Genetics, Development, and Cell Biology, Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, IA 50011. E-mail: xgu@iastate.edu

Received 4 November 2008; Revised 3 February 2009; Accepted 21 March 2009

Published online 16 April 2009 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/jez.b.21290

trans-acting eQTLs play considerable roles in determining detectable genetic variations in gene expression. In particular, the genome-wide information of linking eQTLs to their target genes can provide a global view of the genetic mechanism of regulatory networks, e.g. in yeast (Brem and Kruglyak, 2005).

We have recognized that the eQTL landscape provides a novel genetic framework for studying regulatory evolution at the genome level. Our goal is to understand the underlying genetic pattern of the expression divergence between duplicate pairs. Gene duplication followed by substantial expression divergence has been well demonstrated, thanks to the availability of massive microarray data (Makova and Li, 2003; Papp et al., 2003; Gu, 2004; Huminiecki and Wolfe, 2004; Bing and Hoeschele, 2005). However, much remains unknown about the underlying regulatory evolution that may guide expression divergence between duplicate pairs, largely because of the lack of genomic data. For instance, our understanding of regulatory divergence between yeast duplicate pairs has been compromised by unreliable predicted transcription factor (TF)-binding motifs (Papp et al., 2003; Castillo-Davis et al., 2004; Zhang and Li, 2004) or incomplete TF–target gene interaction data (Lee et al., 2002).

Genome-wide eQTL data may shed some light on solving this problem (Leach et al., 2007). In this study, we investigate how the polygenic eQTL regulatory system in the ancestral gene diverged since gene duplication (Fig. 1). To this end, we combine yeast eQTL data (Brem and Kruglyak, 2005) and other functional genomic data to address the following specific issues: (i) Whether duplicate genes and single-copy genes differ in their eQTL regulation, such as in their heritability, epistasis and the directional effect? (ii) Whether the shared number of eQTLs between duplicate pairs decreases with the increase of evolutionary time since gene duplication? (iii) How much of the variation of expression divergence between duplicate pairs can be explained by the divergence of eQTLs? (iv) Whether the divergence of eQTLs is correlated with other

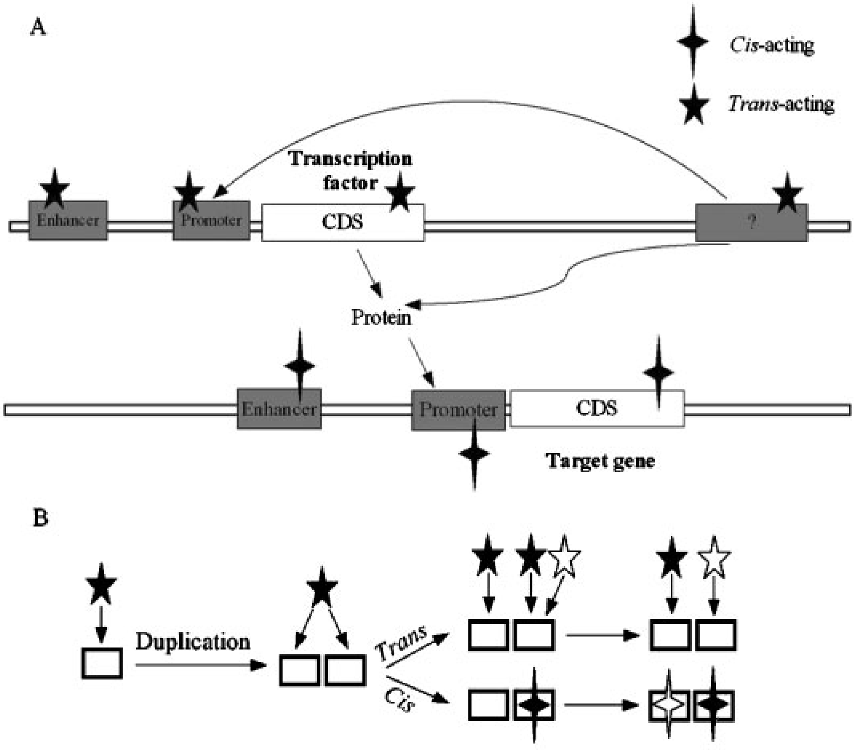

Fig. 1. Schematic illustration for (**A**) genetic variations affecting the expression of target gene, and (**B**) evolution of *trans*- and *cis*-acting eQTLs after gene duplication. Crisscross represents *cis*-acting eQTL and pentacle represents *trans*-acting eQTL. In (A), the region with question mark (?) indicates some hidden regulatory elements for the transcription factor. In (B), blank squares denote the ancestral and duplicate genes. eQTL, expression quantitative trait locus.

measures of functional divergence between duplicate pairs?

DATA AND METHODS

Yeast genomic eQTL data

Data downloaded from Brem and Kruglyak (2005) were based on 2 parent strains of *Saccharomyces cerevisiae*, 6 independent of BY4716 (laboratory) and 12 of RM11-1a (wild), together with 112 segregants and 2,957 markers (oligos with 25 mer from the coding sequence).

eQTL measure: Intuitively, owing to different statistical powers and focusing solutions of various mapping strategies, they may give more or less different eQTL makeup. It may bring bias on what we address. Hence, to take these technical problems into consideration, we conducted two QTL mapping methods to detect eQTLs, one is single marker analysis, Wilcoxon–Mann–Whitney (WMW) test (Brem and Kruglyak, 2005), and the other is multiple interval mapping (MIM) (Kao et al., '99; Zou et al., 2007), both of which apply a false discovery rate (FDR) to determine the joint significance of multiple eQTLs for each expression trait. For MIM analysis, we used a 1.5 LOD support interval around a QTL likelihood ratio test (LRT) peak above the given threshold ($\mathrm{FDR} = 0.05$) to determine the eQTL region. Zou et al. (2007) who used the same data found that the average 1.5 LOD support interval of eQTLs was about 50 kilobases (kb) in the physical map. As the confidence interval is not available in WMW test, we used 50 kb interval to define an eQTL region, which means that any significant linking marker ($\mathrm{FDR} = 0.05$) within 50 kb in the physical map is considered as one eQTL. Moreover, we used several other options to examine whether such fixed-interval procedure may cause some biases in our further analysis: (i) Count genetic markers within a fixed-length region, say, 10, 20, 30 and 40 kb, as a single eQTL. (ii) Invoke the single strongest linking marker per chromosome as one eQTL (Brem and Kruglyak, 2005). As will be shown below, we concluded that these various technical treatments did not change our main results. Hence, in the following, we present analysis based on the fixed 50 kb procedure. Finally, we found at least one eQTL for 2,984 transcripts with the WMW test and 3,326 for the MIM test out of 5,571 transcripts. As expected, the total numbers of eQTLs detected vary with the cutoff for defining the eQTL region (Supplementary Table S1). Once an eQTL affecting the expression level of target gene is detected, a directed link between the target gene and the causal genes within each eQTL was established, which indicates the genetic network regulating the expression of the target gene.

Cis-acting and trans-acting eQTLs: There are two types of eQTLs, *cis*-acting eQTL and *trans*-acting eQTL. A *cis*-acting eQTL is located in the same genomic region of the target gene or nearby, whereas a *trans*-acting eQTL is located in other genomic regions (Fig. 1A). For detailed operation, in WMW test, we defined eQTLs located within 10 kb genomic region around the target gene as *cis*-acting eQTLs, and that outside of this region as *trans*-acting eQTLs. Finally, under the 50 kb procedure, we detected that 13.4% of yeast genes have at least one *cis*-acting eQTL and 50.0% of genes have *trans*-acting eQTLs (for duplicate genes, the number is 14.2 and 54.3%, respectively). Although in the MIM analysis, given the 1.5 support interval around the identified LRT peak, if trait gene locates in its eQTL region, we identified as *cis*-acting eQTL; otherwise, as *trans*-acting eQTL (Zou et al., 2007). We obtained 10.6 and 49.1% of yeast genes, respectively, for possessing *cis*-acting and *trans*-acting eQTLs. We can find that the detected number of *trans*-acting eQTLs is much larger than that of *cis*-acting eQTLs. Moreover, as *cis*-acting eQTLs generally attribute to DNA variations of a gene or its regulatory regions like promoter or anything else that directly influence transcript levels of that gene, there is almost no possibility of sharing *cis*-acting eQTLs between duplicate pairs (see the next paragraph), unless two copies of duplicates share the same regulatory regions or even the structural gene sequences. Therefore, *cis*-acting eQTL is not a good quantitative measure to test the divergence of regulatory network between duplicate pairs. For this reason, we focused on these *trans*-acting eQTLs to understand how two duplicate genes diverge in the genetic regulatory network. The absolute detected numbers of *trans*-acting eQTLs by different procedures are showed in Supplementary Table S1.

Defining shared trans-acting eQTLs between duplicate pairs: Figure 1B illustrates a simple evolutionary scenario to show the role of these two types of eQTLs. For the *trans*-case, shortly after gene duplication, both copies of duplicates have the common *trans*-acting eQTLs. Then, *trans*-acting eQTL of duplicates diverges with the evolutionary time. For the *cis*-case, after gene duplication, there is little chance to share *cis*-acting eQTLs between duplicates. Both copies have their own *cis*-acting eQTLs or do not have any *cis*-acting eQTL, at any stage of gene duplication. Shared *cis*-acting eQTLs between one duplicate pair are obtained if *cis*-acting eQTLs detected in both duplicate copies are mapped to the same relative location of genome (within x kb). The results show that there is almost no shared *cis*-acting eQTLs between duplicate pairs (not shown). It is reasonable and expected, as few duplicates can share the same regulatory regions or structural gene sequences (Fig. 1B). Shared *trans*-acting eQTLs between one duplicate pair were defined if there is overlapping between *trans*-acting eQTL regions, which differ according to the different definition methods of eQTL region. For the WMW analysis, instead of 95 or 99% confidence interval of peaks, we arbitrarily define eQTL region as 10, 20, 30, 40 and 50 kb as mentioned above, whereas for the MIM analysis, we used 1.5 LOD support interval to define the eQTL region, which means that shared *trans*-acting eQTLs were obtained when these eQTL regions around the significant marker location or peak are overlapping between duplicate pairs. To this end, we compared the number of pair that shares the same *trans*-acting eQTL region for duplicate gene pairs and randomized single-copy gene pairs (see the method of detecting yeast duplicate genes). For the fixed 50 kb cutoff, there are 717 duplicate gene pairs that have at least one *trans*-acting eQTL, among which 303 pairs share the same *trans*-acting eQTL regions. This number was reduced to 79 in the case of randomized single-copy gene pairs with the same total pair number ($\chi^2 = 104$, $P < 10^{-15}$, Supplementary Table S2). This implies that duplicate gene pairs share more genetic transcriptional regulatory patterns, in terms of *trans*-acting eQTL, which is consistent with the consensus that duplicate pairs have similar functional roles in biological systems. Consequently, *trans*-acting eQTLs may provide a useful measure for addressing the divergence of genetical control of their expression. Thus, we will use these 717 duplicate gene pairs (the fixed 50 kb cutoff), both of which have at least one *trans*-acting eQTL, for addressing the following issue.

Other eQTL measures: In addition to the number of eQTLs per gene, the complicated genetic architecture of expression traits can be indexed by the *heritability* (relative contributions of genetic and environmental factors to the gene expression), the *epistasis* (interaction between eQTLs) and the *directional effect* (the same direction of all contributing eQTLs that regulate the expression of the target gene). The heritability of each transcript was calculated by the formula $h_2 = (\sigma_s^2 - \sigma_p^2)/\sigma_s^2$, where σ_s^2 is the variance of expression values among the segregants, whereas σ_p^2 is the pooled variance (Zar, '99) of expression values in parental measurements. Permutation was used to determine the significance of heritability of each transcript (Brem and Kruglyak, 2005). Epistasis effect addresses the interaction consequence between one and another eQTL, the tests of which were implemented on a difference between mean expression levels of segregants and parents. The detailed methods were described in Brem and Kruglyak (2005). Similar to Brem and Kruglyak (2005), FDR = 0.05, corresponding to $-\ln(P) = 6.55$, was used to determine the significance. Directional test was conducted as follows (Brem and Kruglyak, 2005): Identify all genes differentially expressed in the parents; for each of these genes, count the number of segregants with expression values between expression means of the parents; test the significance (P-value) by a cumulative binomial test with expected probability of 0.5 (for Mendelian traits). FDR = 0.05 ($P = 0.0084$) was used as the threshold. Finally, we totally obtained 3,455 significant heritable transcripts, which met the criterion of FDR = 0.05 ($h^2 > 0.687$), and found 852 and 707 genes showing significant epistasis and directional effect, respectively, out of 5,571 yeast genes. We have used these measures to study the evolution of genetic regulation after gene duplication.

Yeast genomic data sets

Gene ontology (*GO*) *annotation*: GO annotation data were downloaded from *Saccharomyces* genome database (http://www.yeastgenome.org/). All GO terms of the three categories, biological process, molecular function and cellular component, were accessible for analysis.

Phenotype data: The single-gene-deletion fitness data of *S. cerevisiae* in rich medium were obtained from Steinmetz et al. (2002). Growth rates of each strain with a nearly complete collection of single-gene-deletion mutants cultured in fermentable (YPD, YPDGE) and nonfermentable substrates (YPG, YPE, YPL) were measured, and fitness values were calculated from the growth rate of the deletion strain relative to the pool average for that culture condition, which indicates the capability of the single-gene-deletion strain for survival and reproduction. The larger the fitness value, the smaller the contribution of the very gene to the survival or reproduction of the strain. The lowest value of fitness in these five growth media was obtained. According to the fitness value (f), we divided studied genes into four categories: (1) weak effect (W) if $f \geq 0.95$; (2) moderate effect (M) if $0.80 \leq f < 0.95$; (3) strong effect (S) if $0 < f < 0.8$; (4) lethal (L) if $f = 0$. In addition, another fitness data, which were not measured in normal rich condition, but in drug and environment stress perturbed conditions, were also obtained, that is, the homozygous deletion fitness data of *S. cerevisiae* from Hillenmeyer et al. (2008) (http://chemogenomics.stanford.edu:16080/supplements/global/download.html). There are 4,769 homozygous gene deletion strains for 418 separate conditions perturbed by drug and environment stress. Fitness defect (fd) of the strains in every condition was judged by calculating the P-value from the z-score using the t-distribution (Hillenmeyer et al., 2008). The significance threshold $P < 1 \times 10^{-5}$ was used to decide the sensitivity of growth defect in certain treatments (Hillenmeyer et al., 2008). Fitness defect (fd) is the counterpart to f.

Yeast TF–target gene interaction data: The original data were derived from the large-scale chromatin immunoprecipitation analysis conducted by Lee et al. (2002), including the binding affinity of 141 well-known yeast TFs to the regulatory regions of all *S. cerevisiae* genes (targets). For each gene, a P-value was assigned to measure the probability of true TF–target interaction; a smaller P-value means the interaction is more likely. Here, we used the commonly used significance level of 0.01 as the cutoff to determine the status of TF–target gene interaction.

Microarray data: A total of 504 cDNA microarray data points were used in the study (Supplementary Table S4). Expression changes attributing to environmental stress and internal stress like developmental or physiological conditions are all included. We used background-corrected log ratio (base 2) expression values, and then normalized these expression data to follow the normal distribution with a mean of 0 and a standard deviation of 1.

Detecting yeast duplicate genes

The method of Gu et al. (2002) was applied to identify duplicate genes. As the criterion of 80% alignable regions between protein sequences is too stringent, and may miss some duplicate genes, we reduced this criterion to 50%. The reminders of *S. cerevisiae* genes were considered as singleton genes. The rate of synonymous substitutions (K_S) and nonsynonymous substitutions (K_A) between duplicate pairs was estimated using phylogenetic analysis by maximum likelihood (PAML) (Yang and Bielawski, 2000) with default parameters. The distribution of K_S and K_A is represented in Supplementary Figure S1.

Functional distances between duplicate pairs

Expression distance (E): We used evolutionary distance (E) defined by Gu et al. (2005) as the measure of expression divergence between duplicate pairs. Specifically, for any duplicate genes 1 and 2, let x_{1k} and x_{2k} be their expression levels, respectively, in the kth microarray experiment, and $\bar{x}_1$ and $\bar{x}_2$ be the mean of the expression level in the kth microarray experiment, respectively, where $k = 1, \ldots, m$. The formula of expression distance (E) between genes 1 and 2 is as follows:

$$\hat{E}_{12} = \sum_{k=1}^{m} [(x_{1k} - \bar{x}_1) - (x_{2k} - \bar{x}_2)]^2/(m-1) \quad (1)$$

Trans-acting eQTL distance ($D_{t-\mathrm{eQTL}}$): We used the *Czekanowski–Dice* formula (Martin et al., 2004) to calculate the distance of *trans*-acting eQTLs between duplicate pairs 1 and 2. Let Δ_{12} be the number of eQTLs that differ between duplicate pairs, $y_1 \cup y_2$ be the number of eQTLs that regulate at least one of the duplicate pairs and $y_1 \cap y_2$ be the number of shared eQTLs between duplicate pairs. Then, the *trans*-acting eQTL distance between duplicate genes 1 and 2 is defined as follows:

$$D_{t-\mathrm{eQTL}}(1,2) = \Delta_{12}/[y_1 \cup y_2 + y_1 \cap y_2] \quad (2)$$

Apparently, $D_{t-\mathrm{eQTL}}(1,2) = 0$ when $\Delta_{12} = 0$. The maximum value of $D_{t-\mathrm{eQTL}}(1,2)$ is 1 when $y_1 \cap y_2 = 0$ (no overlapping eQTLs), resulting in $\Delta_{12} = y_1 \cup y_2$.

Czekanowski–Dice distance for other functional measures: Similar to the *trans*-acting eQTL distance, the *Czekanowski–Dice* formula was also applied to measure the functional divergence based on GO annotations or TF–target interactions between duplicate pairs. In these cases, y_1 and y_2 in Equation (2) are the lists of the given GO terms or TF–target interactions in duplicate genes 1 and 2, respectively, and Δ_{12} is the number of differences between duplicate pairs. Three GO categories, biological process (B), molecular function (F) and cellular component (C), were utilized to estimate the functional distances, denoted by D_B, D_F and D_C, respectively. The functional distance based on TF–target interactions is denoted by D_{TF}. Similarly, the larger the values, the greater the functional divergence between duplicate pairs.

Phenotype distance: Distance of fitness in rich medium (D_f) between duplicate pairs was assigned as binary values (0, 1) according to their classification of four categories (W, M, S, L). If both genes were classified into the same categories, D_f was assigned as zero, denoted as D_f_0, otherwise, 1 (D_f_1), whereas the distance of the fitness defect in drug and environment stress perturbed conditions (D_{fd}) was defined by the proportion of conditions that have different statuses between duplicate pairs, i.e. one is normal, whereas the counterpart is the growth defect.

RESULTS

Duplicate genes have on average higher heritability than single-copy genes, but little difference in epistatis and directional effect

By the method from Brem and Kruglyak (2005), we calculated the heritability of each transcript, and obtained 3,455 significant heritable transcripts, which met the criterion of FDR = 0.05 ($h^2 > 0.687$), out of 5,571 genes. In these highly heritable transcripts, we found that 1,091 were duplicates (1,607 total duplicate genes) and 2,364 single-copy genes (3,964 total single-copy genes). Compared with all yeast genes studied, the proportion of duplicates in the set of highly heritable transcripts is about 1.14-fold increased ($\chi^2 = 32.7$, d.f. = 1, $P < 10^{-7}$). Moreover, we found that, on average, the mean heritability in these highly heritable duplicates (0.84 ± 0.002) is significantly higher than that in single-copy genes (0.82 ± 0.002) (unpaired *t*-test, $P < 10^{-9}$, Fig. 2A). However, some studies say that duplicate genes in yeast tend to have high expression levels (Seoighe and Wolfe, '99). Then, is it possible that higher heritability in duplicates is largely owing to the higher expression of genes. To address this issue, we explored the relationship between the heritability and the expression level of genes, where expression levels were obtained from the mean values of the expression level of segregants, and observed that heritability of genes is negatively correlated with the expression level of genes

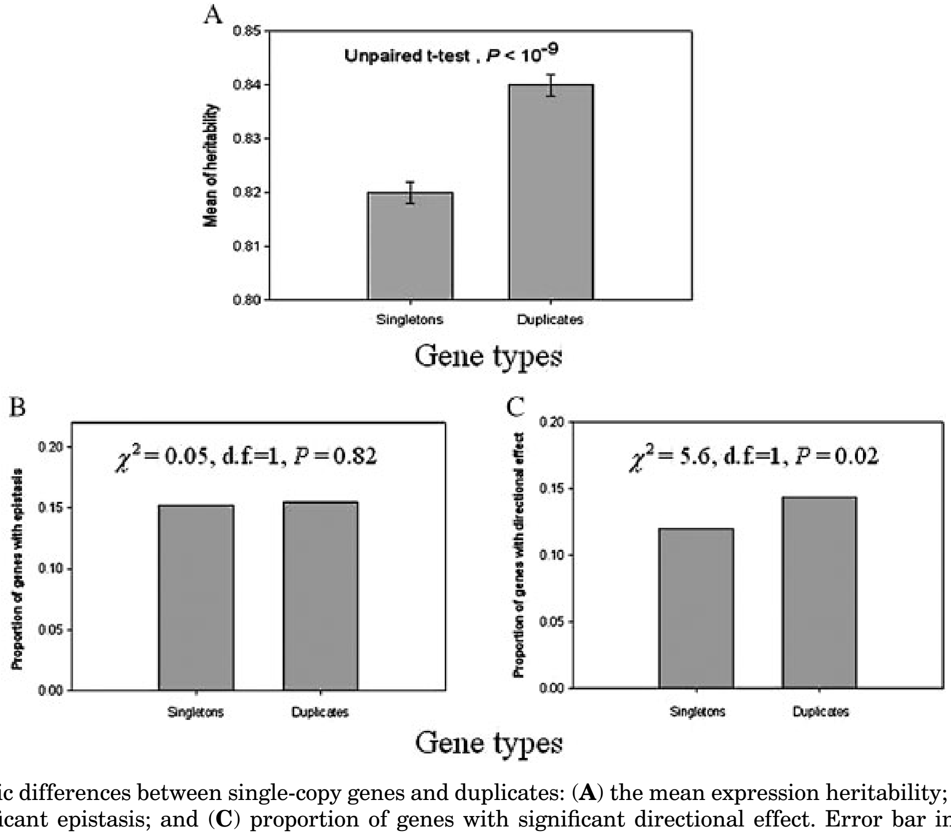

Fig. 2. Genetic differences between single-copy genes and duplicates: (**A**) the mean expression heritability; (**B**) proportion of genes with significant epistasis; and (**C**) proportion of genes with significant directional effect. Error bar indicates standard error.

(Pearson's product–moment correlation, $R = -0.04$, $P = 0.02$). Thus, we may presume that duplicate transcripts tend to be more highly heritable, implying that gene duplication may increase the genetic capacity against the perturbation of external environments on the gene expression. However, the difference of heritability between duplicates and singletons is very small (0.84 vs. 0.82), and the correlation between the heritability and the expression level of genes is not so large ($R = -0.04$). The influence of gene duplication on gene expression by increasing the genetic capacity against environment perturbations is existing, but small.

In addition, we studied whether the degree of directional or epistasis inheritance differs between duplicate genes and single-copy genes. In both cases, we calculated the proportion of genes that have significant directional or epistasis inheritance (see section Data and Methods) among duplicate genes, as well as that among single-copy genes. The difference of these two measures between duplicates and singletons is very small (Fig. 2B, C), despite in the case of directional effect, where the difference is significant. It seems that gene duplications do not dramatically change the interaction nature in the genetic architecture of transcription.

Trans-acting eQTL divergence between duplicate gene pairs increases with the evolutionary time

We used *Czekanowski–Dice* distance (Martin et al., 2004) to estimate the relative divergence of *trans*-acting eQTLs between duplicate gene pairs, denoted by $D_{t\text{-eQTL}}$ (see section Data and Methods). Ranging from 0 to 1, a large $D_{t\text{-eQTL}}$ indicates great divergence of *trans*-acting eQTLs between duplicate gene pairs. Using the synonymous distance (K_S) between duplicate pairs as a proxy to the evolutionary time of duplication event, we found a positive correlation between $D_{t\text{-eQTL}}$ and K_S (Pearson's product–moment correlation for all data, $R = 0.60$, $P < 10^{-15}$). When the nonsynonymous distance (K_A) was used, a similar positive correlation was obtained. As the distance estimation is statistically unreliable when K_S or K_A becomes larger, owing to the repeated substitutions at the same site, we selected the duplicate pairs with $K_S < 2.0$ or $K_A < 0.5$, and then carried out the analysis. In each case we observed a similar result (Pearson's product–moment correlation for all data, $R = 0.66$, $P < 10^{-15}$; $R = 0.61$, $P < 10^{-15}$, respectively, Fig. 3).

Hence, we conclude that the divergence of *trans*-acting eQTLs between duplicate pairs increases with the evolutionary time after gene duplication. We have examined other distance measures for *trans*-acting eQTL divergence, e.g. $1-r$ (r is the coefficient of correlation) or $1-c$ (c is the proportion of overlapping eQTLs between duplicate pairs), and found that this conclusion still holds (not shown). Extensive microarray analyses (Makova and Li, 2003; Castillo-Davis et al., 2004; Zhang and Li, 2004; Bing and Hoeschele, 2005) have shown that expression divergence between duplicate pairs increases with evolutionary time, as does regulatory divergence measured by the predicted TF-binding motifs (Papp et al., 2003; Castillo-Davis et al., 2004; Zhang and Li, 2004) or the TF–target gene interactions (Bing and Hoeschele, 2005). All this evidence together supports the notion of genetic regulatory divergence between duplicate pairs. Different procedures defining eQTL and mapping methods bring little bias (Supplementary Table S3).

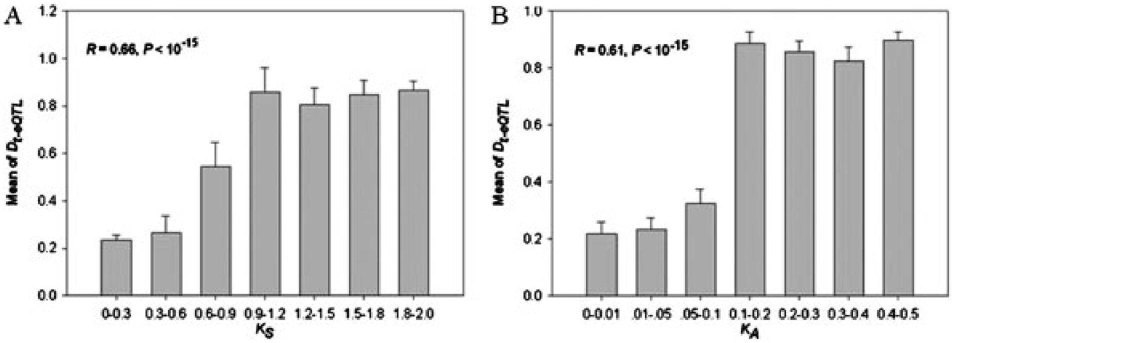

Fig. 3. Divergence of *trans*-acting eQTLs between duplicate pairs ($D_{t\text{-eQTL}}$) increases with synonymous distance K_S (panel **A**) or nonsynonymous distance K_A (panel **B**) between duplicate pairs. Error bar indicates standard error. eQTL, expression quantitative trait locus.

Trans-acting eQTL divergence underlies expression divergence between duplicate gene pairs

In spite of well-demonstrated divergence of gene expression between duplicate pairs, the mechanism remains a subject of debate (Li et al., 2005). Our working hypothesis is that the divergence in transcriptional regulatory network underlies the expression divergence between duplicate pairs. We therefore examined whether the divergence of *trans*-acting eQTLs is correlated with the expression divergence between duplicate pairs. In our analysis, the expression distance (E) was used to measure the expression divergence between duplicate pairs, whereas the *trans*-acting eQTL distance ($D_{t-\mathrm{eQTL}}$) to measure the genetic *trans*-regulatory divergence (see section Data and Methods). We observed a significantly positive correlation between $D_{t-\mathrm{eQTL}}$ and E (Pearson's product–moment correlation for all data, $R=0.37$, $P<10^{-15}$, Fig. 4A) under the 50 kb procedure. Hence, approximately $R^2=13.8\%$ of variation in expression divergence between all duplicate pairs can be explained by their *trans*-acting eQTL divergence. We noticed a higher percentage ($R^2=21.3\%$) when duplicate pairs with $K_S<2.0$ are considered.

Another well-known data set for deciphering transcriptional regulatory networks is the yeast TF–target interactions (Lee et al., 2002). Similar to *trans*-acting eQTLs, we calculated the *Czekanowski–Dice* distance of TF–target interactions between duplicate pairs, denoted by D_{TF} (see section Data and Methods). For all duplicate pairs, we observed that $D_{t-\mathrm{eQTL}}$ and D_{TF} are significantly correlated (Pearson's product–moment correlation for all data, $R=0.35$, $P<10^{-15}$, Fig. 4B). As $D_{t-\mathrm{eQTL}}$ and D_{TF} are associated, we reconducted the analysis of partial correlation between $D_{t-\mathrm{eQTL}}$ and E, controlling for D_{TF}, and obtained the partial correlation coefficient as 0.20, with $P<0.001$. In addition, D_{TF} is also positively correlated with the expression distance E between duplicate pairs (Pearson's product–moment correlation for all data, $R=0.27$, $P<10^{-10}$), whereas controlling for $D_{t-\mathrm{eQTL}}$, the partial correlation coefficient between D_{TF} and E drops to 0.15. This tells us that genetic regulatory network divergence, in terms of *trans*-acting eQTL divergence, may have a greater influence on expression divergence than do the effects of divergence in TFs.

In spite of many other factors that may be involved, our analyses in *trans*-acting eQTLs and TF–target interactions between duplicate pairs support the following evolutionary scenario. After gene duplication, two duplicated copies may have differential responses to the transcriptional regulatory network, initiating the divergence of expression pattern. Subfunctionization (Force et al., '99) may be the likely mechanism, resulting in differential responses to the regulatory signaling. To support this notion, we conducted a multiple regression analysis to see how much variation of expression divergence between duplicate pairs can be explained by eQTL divergence ($D_{t-\mathrm{eQTL}}$) and TF–target interaction divergence (D_{TF}) taken together. We observed the overall coefficient of correlation $R^2=16.9\%$, after fitting the multiple regression model ($E=D_{t-\mathrm{eQTL}}+D_{\mathrm{TF}}+\varepsilon$), whereas for duplicate pairs with $K_S<2.0$ and $K_S>2.0$, the number changes to 27.4 and 0.9%,

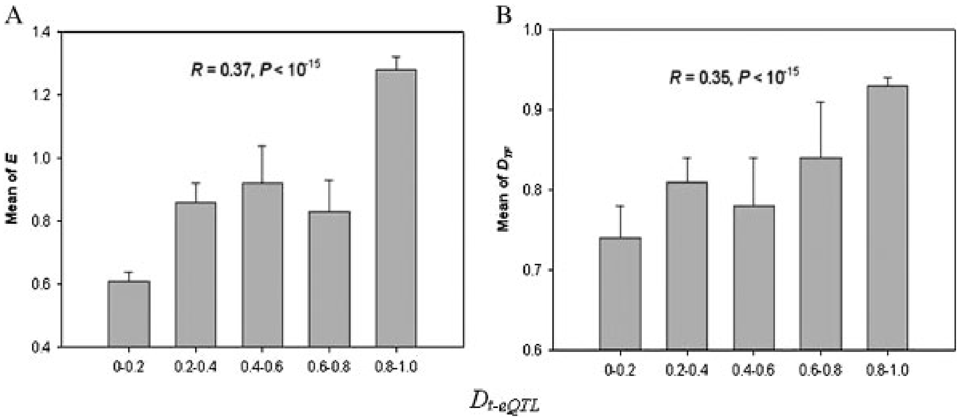

Fig. 4. *Trans*-acting eQTL divergence underlying expression divergence between duplicate pairs. (**A**) Correlation between *trans*-eQTL distance ($D_{t-\mathrm{eQTL}}$) and expression distance (E). (**B**) Correlation between $D_{t-\mathrm{eQTL}}$ and the distance of TF–target interactions (D_{TF}). Error bar indicates standard error. eQTL, expression quantitative trait locus.

respectively. Other procedures and mapping methods show a similar result (Supplementary Table S3).

Trans-acting eQTL divergence coupled with other functional divergences

GO provides valuable resources for understanding the gene function. Gene Ontology Consortium (Ashburner et al., 2000) provides three aspects of gene function, ''Molecular function'' for biochemical activity or function of gene, ''Biological process'' for cellular function, which is accomplished by ordered assemblies of molecular function of a series of genes, and ''Cellular component'' for cell location of protein. In each of three categories, every gene has one or more GO terms with hierarchically different levels, which are structured as directed acyclic graphs (DAGs). In our study, we attempted to explore the interrelation between eQTL function and these three aspects of gene function classified by Gene Ontology Consortium. All GO terms from the root term to leaf term of DAG of three categories were used for the analysis. For a pair of very recent duplicate genes, one may expect very similar GO terms. Functionally diversified duplicate genes would be expected to have a low level of overlapping GO terms. In other words, differences in GO terms can be used to measure the functional divergence between duplicate pairs. As suggested by Martin et al. (2004), we used the *Czekanowski–Dice* distance between duplicate pairs for each of the three GO functional categories: biological process (D_{B}), molecular function (D_{F}) and cellular component (D_{C}), respectively, for further analysis.

We studied the correlation between the *trans*-acting eQTL distance $D_{t-\mathrm{eQTL}}$ and three GO distances (D_{B}, D_{F} and D_{C}). Interestingly, with the 50 kb procedure, we found that $D_{t-\mathrm{eQTL}}$ has statistically highly significant correlations with D_{B} ($R = 0.27$, $P < 10^{-11}$) and D_{C} ($R = 0.23$, $P < 10^{-8}$), respectively, but a very weak correlation with D_{F} ($R = 0.08$, $P = 0.04$). As *trans*-acting eQTL divergence implies a diversified genetic regulatory network between duplicate pairs, our finding indicates that divergence in *trans*-acting eQTLs between duplicate pairs may be coupled with divergence in physiological roles, as implied by GO biological processes or cellular components. Indeed, functional divergence between duplicate pairs usually does not dramatically change the molecular function in the GO. A similar conclusion is drawn when implementing other eQTL defined procedures and mapping methods (Supplementary Table S3).

Another type of functional divergence between a duplicate pair is phenotypic divergence, which is also the most direct presentation of evolutionary scenario of the genes. Curiously, may genetic regulatory network divergence between duplicate pairs contribute to the phenotype change of genes? To answer the question, we explored the relationship between *trans*-acting eQTL distance $D_{t-\mathrm{eQTL}}$ and phenotype distance (D_f and D_{fd}), and found that $D_{t-\mathrm{eQTL}}$ was significantly correlated with D_{fd} (fitness defect in drug and environment stress treatment) (Supplementary Table S3), except for the method of MIM, whereas no significant correlation between $D_{t-\mathrm{eQTL}}$ and D_f (fitness in rich medium) was detected, summing up from the fact that $D_{t-\mathrm{eQTL}}$ presents no difference between the groups of D_f_0 and D_f_1 (Wilcoxon rank sum test, Supplementary Table S3). The result is reasonably expected, as the environment change, like drug or environment stress perturbing, offers more probability to generate genetic mutants compared with normal condition.

DISCUSSION

In this article, we have investigated the yeast eQTL data set (Brem and Kruglyak, 2005) to study the evolutionary pattern of gene regulation after gene duplication. Our analysis revealed that duplicate genes have higher expression heritability than single-copy genes, but little difference for epistasis and directional effect in their expression phenotypes. After calculating the *trans*-acting eQTL distance ($D_{t-\mathrm{eQTL}}$) and the synonymous distance (K_{S}) between duplicate pairs, we found a highly significant correlation between them, supporting the notion of *trans*-acting eQTL divergence after the gene duplication. Interestingly, we estimated that *trans*-acting eQTL divergence can explain about 21% of the variation of expression divergence between duplicate pairs with $K_{\mathrm{S}} < 2.0$. Moreover, when taken together with the TF–target interaction divergence, the explainable variation becomes 27%. We hence conclude that genetic control of gene expression may partially underlie the expression divergence between duplicate pairs. The correlation of *trans*-acting eQTL divergence with other functional divergence measures implies that genetic *trans*-regulatory divergence may be coupled with the divergence of physiological roles between duplicate pairs. Our analysis of *trans*-acting eQTLs is also helpful to generalize the evolutionary model of expression divergence (Evangelisti and Wagner, 2004; Gu, 2004), which integrates the effect of genetic *trans*-regulatory network.

Some caveats in eQTL analysis

In spite of these interesting findings, one should note that the eQTLs identified may only provide an incomplete overview of the genetical control of gene expression because of the limitations of the current yeast data set (samples, markers and defined population or strains) and QTL mapping methods. Owing to the limited statistical resolution, eQTLs with the small effect on expression variation cannot be detected. In the case of yeast genomic eQTL analysis (Brem and Kruglyak, 2005), for more than half of mapped eQTLs, only $<30\%$ of genetic variance is explained. Therefore, the detected eQTLs that we used in the study are likely to be underestimated. However, as the measure we used is the relative difference between eQTLs, like $D_{t-\mathrm{eQTL}}$ between duplicate pairs, the effect of statistical bias should be largely canceled out, not only theoretically (Dan Nettleton, Department of Statistics, Iowa State University, personal communication) but also practically. To address this issue, we tried two QTL mapping methods, WMW and MIM. Moreover, in the WMW analysis, several procedures defining eQTLs, say, 10,…,50 kb, were used to test. As expected, we found virtually the same results (Supplementary Tables S1–S3).

In addition, this study was mostly based on the single-locus tests carried out by WMW (Brem and Kruglyak, 2005), which did not consider the genetic interactions known as epistasis in the eQTL inference. Although, in the MIM analysis, genetic interactions were addressed, very few interactions were detected (only 49 interaction pairs in 5,182 detected eQTLs). However, in a separate study, the same authors (Brem et al., 2005) showed that these genetic interactions (e.g. additive–additive, additive–dominant and dominant–dominant eQTL interactions) exist in the inheritance of approximately 57% of yeast expressions. In this sense, our study only took the first-degree approximation of genetic complexity into account. The next study will be further to investigate the effect of genetic interaction between eQTL–eQTL pairs (epistasis) on the regulatory divergence between duplicate pairs. In addition, the gene by environment interaction is another factor that may contribute to the genetic variation of expression, which means that the effect of eQTL may differ in sexes (Jin et al., 2001), organs (Wall et al., 2005) or tissues (Whitehead and Crawford, 2005). Therefore, with the development of QTL mapping methods in the future, the evolution pattern of genetic regulatory network after gene duplication may be largely explored.

The ability to detect eQTLs will also depend on the variability between the two strains that were crossed for the analysis. In other words, some interesting loci may be missed owing to the variation between the strains crossed. Therefore, the eQTLs analyzed just provide the partial insights into genetic regulatory network, which may also more or less answer the doubts that *trans*-acting eQTL divergence is not highly correlated with other functional divergences, such as GO functional divergence and phenotype divergence, though most of them are significantly coupled.

eQTLs and transcription module divergence between duplicate pairs

A typical transcription module for gene regulation includes TFs (*trans*-acting regulators), regulatory and binding regions (*cis*-acting elements) and the target coding sequence. Although some insightful findings have been reported for understanding the roles of these elements in regulatory divergence between duplicate pairs (Papp et al., 2003; Castillo-Davis et al., 2004; Evangelisti and Wagner, 2004; Teichmann and Babu, 2004; Zhang and Li, 2004; Bing and Hoeschele, 2005; Tanay et al., 2005), some difficulties also have been encountered. As *cis*-regulatory motifs (promoters, enhancers, etc.) are sparsely located in the noncoding region with varying lengths of nucleotides, their complete and accurate detection is a difficult task, both experimentally and computationally. For instance, unreliable prediction of TF-binding motifs in the 5′-noncoding region could make the interpretation of expression-motif analysis more difficult (Papp et al., 2003; Castillo-Davis et al., 2004; Zhang and Li, 2004) The TF–target interaction data (Lee et al., 2002) partially solved this problem, as only a limited number of yeast TFs have been experimentally examined. Instead, using yeast eQTLs to study the transcriptional divergence between duplicate pairs provides a genome-wide view of the genetic architecture. In particular, Yvert et al. (2003) observed that most of the *trans*-acting detected eQTLs in yeast do not reside in TFs (Yvert et al., 2003), which suggested that in addition to genetic variations in

TFs, *trans*-acting eQTL strategy may address other signaling-related molecules (Yvert et al., 2003) (Fig. 1A). However, a comprehensive analysis, combining different regulatory genomic data sets, is crucial in order to investigate the pattern of evolutionary genomics in detail.

In addition, of these detected eQTLs, some may be owing to the auxotrophic markers used in the cross, like LEU2, URA3, LYS2, etc., whose alleles do not segregate in natural populations, and thus may bring bias on the result. However, compared with the number of eQTLs we detected (5,632 in the fixed 50 kb procedure), the number of auxotrophic markers in the cross is extremely small, *MATα*, *lys2Δ0* in parent strain BY4716 (S288C derivative), and *MATα*, *leu2Δ0*, *ura3Δ0*, *ho::KAN* in RM11-1a. In practice, after removing these eQTLs, the recalculated results were similar. To avoid repeatedly representing the result, data here were not shown.

Selection of duplicate pairs

Our study used the procedure developed by Gu et al. (2002) to define duplicate pairs in the yeast, with some minor modifications (see section Data and Methods). We have examined several cutoffs that define a duplicate pair, and found that the classification is very similar. Another issue is that some duplicate pairs are interrelated when they belong to a large gene family. We therefore reselected independent pairs of duplicate genes by (i) sorting the duplicate gene set (at least three members) based on the pairwise synonymous distances K_S; (ii) selecting the first duplicate gene pair with the smallest K_S and (iii) then selecting the second independent duplicate gene pair after removing the first selected gene pair, and so forth. We have conducted similar analyses on these independent duplicate gene pairs, and verified our main results.

One of the challenges in measuring the mRNA expression level by microarrays is the cross-hybridization between the genes with similar sequence, because of the ununique probe target. This problem is more prevalent in cDNA arrays than oligonucleotide arrays. As the gene expression values for eQTL detecting in yeast data set and in our further study were measured by cDNA microarray, and the genes we interested in are duplicate genes, which are more likely to have similar sequence, it is necessary to take the cross-hybridization into consideration. We explored the sequence distance distribution between duplicate gene pairs we used by the rate of synonymous substitutions (K_S) and nonsynonymous substitutions (K_A) (Supplementary Figure S1), and removed duplicate pairs with very low sequence divergence ($K_S<0.1$, $K_A<0.01$). Then, after reanalyzing, we observed the similar result. Besides, other factors like codon-usage bias, gene conversion, functional constraints, etc. may also impact the estimation of K_S and K_A (Lin et al., 2006), and thus further exploration may need to be under the way.

ACKNOWLEDGMENTS

We are grateful to Dr. Rachel Brem for helping us in handling the original data, and being patient to explain the data set and answer our questions when we were treating the data. We also give great thanks to two anonymous reviewers' constructive comments, conducing to refine our study.

LITERATURE CITED

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppiq JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Rinqwald M, Rubin GM, Sherlock G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29.

Bing N, Hoeschele I. 2005. Genetical genomics analysis of a yeast segregant population for transcription network inference. Genetics 170:533–542.

Brem RB, Kruglyak L. 2005. The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proc Natl Acad Sci USA 102:1572–1577.

Brem RB, Yvert G, Clinton R, Kruglyak L. 2002. Genetic dissection of transcriptional regulation in budding yeast. Science 296:752–755.

Brem RB, Storey JD, Whittle J, Kruglyak L. 2005. Genetic interactions between polymorphisms that affect gene expression in yeast. Nature 436:701–703.

Bystrykh L, Weersing E, Dontje B, Sutton S, Pletcher MT, Wiltshire T, Su AI, Vellenga E, Wang J, Manly KF, Lu L, Chesler EJ, Alberts R, Jansen RC, Williams RW, Cooke MP, de Haan G. 2005. Uncovering regulatory pathways that affect hematopoietic stem cell function using "genetical genomics." Nat Genet 37:225–232.

Castillo-Davis CI, Hartl DL, Achaz G. 2004. *cis*-Regulatory and protein evolution in orthologous and duplicate genes. Genome Res 14:1530–1536.

Cheung VG, Conlin LK, Weber TM, Arcaro M, Jen K-Y, Morley M, Spielman RS. 2003. Natural variation in human gene expression assessed in lymphoblastoid cells. Nat Genet 33:422–425.

Evangelisti AM, Wagner A. 2004. Molecular evolution in the yeast transcriptional regulation network. J Exp Zool B Mol Dev Evol 302:392–411.

Force A, Lynch M, Pickett FB, Amores A, Yan Y-L, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151: 1531–1545.

Gibson G, Weir B. 2005. The quantitative genetics of transcription. Trends Genet 21:616–623.

Gu X. 2004. Statistical framework for phylogenomic analysis of gene family expression profiles. Genetics 167:531–542.

Gu Z, Cavalcanti A, Chen F-C, Bouman P, Li W-H. 2002. Extent of gene duplication in the genomes of Drosophila, nematode, and yeast. Mol Biol Evol 19:256–262.

Gu X, Zhang Z, Huang W. 2005. Rapid evolution of expression and regulatory divergences after yeast gene duplication. Proc Natl Acad Sci USA 102:707–712.

Hillenmeyer ME, Fung E, Wildenhain J, Pierce SE, Hoon S, Lee W, Proctor M, St. Onge RP, Tyers M, Koller D, Altman RB, Davis RW, Nislow C, Giaever G. 2008. The chemical genomic portrait of yeast: uncovering a phenotype for all genes. Science 320:362–365.

Huminiecki L, Wolfe KH. 2004. Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. Genome Res 14:1870–1879.

Jansen RC, Nap J-P. 2001. Genetical genomics: the added value from segregation. Trends Genet 17:388–391.

Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G, Gibson G. 2001. The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. Nat Genet 29:389–395.

Kao CH, Zeng ZB, Teasdale RD. 1999. Multiple interval mapping for quantitative trait loci. Genetics 152:1203–1216.

Kirst M, Basten CJ, Myburg AA, Zeng Z-B, Sederoff RR. 2005. Genetic architecture of transcript-level variation in differentiating xylem of a eucalyptus hybrid. Genetics 169:2295–2303.

Leach LJ, Zhang Z, Lu C, Kearsey MJ, Luo Z. 2007. The role of *cis*-regulatory motifs and genetical control of expression in the divergence of yeast duplicate genes. Mol Biol Evol 24:2556–2565.

Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne J-B, Volkert TL, Fraenkel E, Gifford DK, Young RA. 2002. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science 298:799–804.

Li WH, Yang J, Gu X. 2005. Expression divergence between duplicate genes. Trends Genet 21:602–607.

Lin Y-S, Byrnes JK, Hwang J-K, Li W-H. 2006. Codon-usage bias versus gene conversion in the evolution of yeast duplicate genes. Proc Natl Acad Sci USA 103:14412–14416.

Makova KD, Li W-H. 2003. Divergence in the spatial pattern of gene expression between human duplicate genes. Genome Res 13:1638–1645.

Martin D, Brun C, Remy E, Mouren P, Thieffry D, Jacq B. 2004. GOToolBox: functional analysis of gene datasets based on gene ontology. Genome Biol 5:R101.

Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG. 2004. Genetic analysis of genome-wide variation in human gene expression. Nature 430:743–747.

Oleksiak MF, Churchill GA, Crawford DL. 2002. Variation in gene expression within and among natural populations. Nat Genet 32:261–266.

Papp B, Pal C, Hurst LD. 2003. Evolution of *cis*-regulatory elements in duplicated genes of yeast. Trends Genet 19:417–422.

Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, Linsley PS, Mao M, Stoughton RB, Friend SH. 2003. Genetics of gene expression surveyed in maize, mouse and man. Nature 422:297–302.

Seoighe C, Wolfe KH. 1999. Yeast genome evolution in the post-genome era. Curr Opin Microbiol 2:548–554.

Steinmetz LM, Scharfe C, Deutschbauer AM, Mokranjac D, Herman ZS, Jones T, Chu AM, Giaever G, Prokisch H, Oefner PJ, Davis RW. 2002. Systematic screen for human disease genes in yeast. Nat Genet 31:400–404.

Tanay A, Regev A, Shamir R. 2005. Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. Proc Natl Acad Sci USA 102:7203–7208.

Teichmann SA, Babu MM. 2004. Gene regulatory network growth by duplication. Nat Genet 36:492–496.

Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaever G, Eisen MB, Feldman MW. 2005. Functional genomic analysis of the rates of protein evolution. Proc Natl Acad Sci USA 102:5483–5488.

Whitehead A, Crawford D. 2005. Variation in tissue-specific gene expression among natural populations. Genome Biol 6:R13.

Yan H, Yuan W, Velculescu VE, Vogelstein B, Kinzler KW. 2002. Allelic variation in human gene expression. Science 297:1143.

Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15:496–503.

Yvert G, Brem RB, Whittle J, Akey JM, Foss E, Smith EN, Mackelprang R, Kruglyak L. 2003. *Trans*-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. Nat Genet 35:57–64.

Zar JH. 1999. Biostatistical analysis. Upper Saddle River, NJ: Prentice-Hall.

Zhang L, Li W-H. 2004. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. Mol Biol Evol 21:236–239.

Zou W, Aylor D, Zeng Z-B. 2007. eQTL viewer: visualizing how sequence variation affects genome-wide transcription. BMC Bioinformatics 8:7.