Genetic Interaction Network as an Important Determinant of Gene Order in Genome Evolution

Yu-Fei Yang,[†,‡,1,2,3] Wenqing Cao,[‡,1,2,3] Shaohuan Wu,[1,2,3] and Wenfeng Qian*[,1,2,3]

[1]State Key Laboratory of Plant Genomics, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China

[2]Key Laboratory of Genetic Network Biology, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China

[3]University of Chinese Academy of Sciences, Beijing, China

[†]Present address: Genetron Health Co., Ltd., Beijing, China

[‡]These authors contributed equally to this work.

***Corresponding author:** E-mail: wfqian@genetics.ac.cn.

Associate editor: Jianzhi Zhang

Abstract

Although it is generally accepted that eukaryotic gene order is not random, the basic principles of gene arrangement on a chromosome remain poorly understood. Here, we extended existing population genetics theories that were based on two-locus models and proposed a hypothesis that genetic interaction networks drive the evolution of eukaryotic gene order. We predicted that genes with positive epistasis would move toward each other in evolution, during which a negative correlation between epistasis and gene distance formed. We tested and confirmed our prediction with computational simulations and empirical data analyses. Importantly, we demonstrated that gene order in the budding yeast could be successfully predicted from the genetic interaction network. Taken together, our study reveals the role of the genetic interaction network in the evolution of gene order, extends our understanding of the encoding principles in genomes, and potentially offers new strategies to improve synthetic biology.

Key words: **gene order, genetic interaction network, genetic recombination, fitness, yeast.**

Article

Introduction

With thousands of genomes being sequenced, it is increasingly being observed that gene order in a genome is not random (Hurst et al. 2004). For example, six genes in the allantoin degradation (*DAL*) pathway formed a cluster on chromosome IX during the evolution of *Saccharomyces cerevisiae* (Wong and Wolfe 2005). More dramatically, three genes in the galactose utilization (*GAL*) pathway formed a cluster in multiple lineages independently during fungal evolution (Slot and Rokas 2010). However, the evolutionary principles underlying such nonrandom gene order are still elusive, except when neighboring genes form an operon (Lawrence 1999; Lawrence 2002; Qian and Zhang 2008; Zaslaver et al. 2011).

A number of hypotheses have been proposed to explain the evolution of gene order. First, the clustering of genes with similar functions in the genome may facilitate their coordinated expression. Although neighboring genes indeed tend to have similar expression profiles (Cho et al. 1998; Cohen et al. 2000; Boutanaev et al. 2002; Spellman and Rubin 2002; Williams and Bowles 2004), such phenomena could also be explained by the "leaky" expression of neighboring genes (Spellman and Rubin 2002; Hurst et al. 2004; Liao and Zhang 2008; Ghanbarian and Hurst 2015). Second, housekeeping or essential genes tend to cluster in a genome (Lercher et al. 2002; Pal and Hurst 2003), a phenomenon that might be explained by natural selection to reduce gene expression noise (Batada and Hurst 2007). However, this theory cannot explain the nonrandom gene order within and between such clusters. Third, mutational bias, such as tandem gene duplication, could also lead to nonrandom gene order; however, after removing tandem duplicate genes, gene order is still nonrandom in the aspects described earlier (Hurst et al. 2004). Together, these observations suggest that additional evolutionary mechanisms exist to explain nonrandom gene order in the genome.

The evolution of gene order may be driven by natural selection to optimize recombination frequencies among genes because gene order determines gene distance (*D*, defined as the number of genes between two genes on a chromosome) and gene distance is highly correlated with recombination frequency (supplementary fig. S1, Supplementary Material online, the budding yeast as an example). Several theoretical analyses suggested that the evolution of recombination frequency between a pair of genes can be influenced by their epistatic interaction (Nei 1967, 1969; Eshel and Feldman 1970; Feldman et al. 1980; Kondrashov 1982, 1988; Charlesworth 1990; Kouyos et al. 2007; Charlesworth and Charlesworth 2011). Here, epistasis, or genetic interaction, refers to the phenomenon that the fitness effects of two mutations on two different genes are not independent (Phillips 2008), and can be quantified as the difference between the relative fitness of the double mutant

(ω_{ab}) and the multiplicative expectation from those of two single mutants ($\omega_{aB}\omega_{Ab}$). Previous theories could be summarized as two effects, the short-term effect and the long-term effect (fig. 1A). On the one hand, the linkage of genetically interacting genes (regardless of sign) is advantageous in the short term because it helps to maintain the epistasis-induced linkage disequilibrium (LD), which is favored by natural selection (Nei 1967, 1969; Eshel and Feldman 1970; Feldman et al. 1980; Kouyos et al. 2007; Charlesworth and Charlesworth 2011) (fig. 1A, the solid line; supplementary note S1, Supplementary Material online). Here, the coefficient of LD is defined as the difference between the genotype frequency of the wild-type individuals (X_{AB}) and the multiplicative expectation from the frequencies of the wild-type alleles (X_AX_B). On the other hand, for two deleterious mutations, genetic recombination breaks the negative LD induced by negative epistasis and thus increases the proportion of double mutants. This increase benefits the population in the long term by facilitating the purge of deleterious mutations, which has been suggested to be related to the origin of sexual reproduction (Feldman et al. 1980; Kondrashov 1982; Charlesworth 1990) (fig. 1A, the dash line).

To summarize, genetic recombination is never favored by natural selection for positively epistatic gene pairs, whereas for negatively epistatic gene pairs, the long- and short-term effects of genetic recombination counteract each other (fig. 1A). Therefore, epistasis could drive the evolution of recombination frequencies among genes on the same chromosome, potentially by altering gene order. In this process, a negative correlation between epistasis and gene distance, hereafter referred to as E–D correlation, evolves. Furthermore, this correlation would be especially strong among positively epistatic gene pairs due to the synergistic combination of the long- and short-term effects (fig. 1A). However, these theoretical predictions have never been rigorously tested with empirical data. In this study, we tested these predictions with both simulation and yeast empirical epistasis data. Our study thus reveals a basic principle in the evolution of gene order and enhances our power to decode information from well-constructed genetic interaction networks and thousands of sequenced genomes.

Results

Negative E–D Correlation Is Observed during in silico Evolution

We first performed in silico evolution in which two genes (*A* and *B*) were considered, each having a wild-type allele (*A* or *B*) and a deleterious allele (*a* or *b*). The relative fitness of the haploid wild-type genotype (ω_{AB}) was defined as 1, and the epistasis (*E*) was defined as $\omega_{ab}-\omega_{aB}\omega_{Ab}$. For two genotypes with different gene distances (*D*) between *A* and *B*, gene flow within this region is strictly prohibited because recombination within the region between *A* and *B* leads to the gain or loss of genes after segregation (Wu and Ting 2004). Furthermore, the "modifier" locus of recombination frequency is completely linked with *A* and *B* (Nei 1967, 1969), making it possible to directly compare the fitness of genotypes with different gene distances. Therefore, to investigate the impact of gene distance on fitness, we compared the average fitness of two populations over generations, one with *D* between *A* and *B* equal to 50 and the other with *D* equal to 0. Based on the empirical data from the budding yeast *S. cerevisiae* (Mancera et al. 2008), these *D* values correspond to recombination frequencies $R = 0.264$ and 0.064, respectively (supplementary fig. S1, Supplementary Material online). In each generation, we calculated the frequency changes of genotypes by considering both natural selection and genetic recombination. Figure 1B shows the results of the first 100 generations of in silico evolution, when long-term effects begin to dominate the evolutionary process. If epistasis between *A* and *B* was positive, the population with $D = 50$ was always outcompeted by the population with $D = 0$ (fig. 1B, $\omega_{D=50} - \omega_{D=0} < 0$). Furthermore, the fitness difference increased with the magnitude of the epistasis value (fig. 1B). In other words, reduced *D* between two genes with positive epistasis is favored by natural selection. By contrast, if the epistasis between *A* and *B* was negative, the population with $D = 50$ exhibited a short-term disadvantage followed by a long-term advantage compared with the population with $D = 0$ (fig. 1B). As expected, the long-term advantage was due to an elevated purging rate of deleterious alleles [fig. 1C, $X_{a\ (D=50)} - X_{a\ (D=0)} < 0$]. A similar trend was observed when we compared a population with $D = 50$ and one with $D = 100$ ($R = 0.464$, supplementary fig. S2, Supplementary Material online).

We also performed in silico evolution in a series of strains in which *D* varied between 0 and 100 ($R = 0.064$ and 0.464, respectively). We recorded the average fitness of each population at the 100th generation (ω_{100}) and identified the optimal *D*, D_{opt}, for each epistasis value (fig. 1D). Given that the effective population size (N_e) in the budding yeast is $\sim 10^7$ (Wagner 2005), the minimal selection coefficient that can be detected for yeast is $\sim 10^{-7}$. In other words, all *D* values that reduce the relative fitness by $< 10^{-7}$ are permitted during evolution. We calculated the mean of all permitted *D* values (dashed line in fig. 1D). As predicted in our model, a strong negative E–D correlation was observed from the results of in silico evolution. Importantly, such negative E–D correlation was also observed at the 50th and 200th generation (supplementary fig. S3, Supplementary Material online). To further test whether the outcome of in silico evolution was sensitive to population genetics parameters, we examined various values for initial allele frequencies and fitness defects. We observed a negative E–D correlation with all parameter sets (supplementary fig. S4, Supplementary Material online).

Chromosomal Arrangement of Genes in Star-Like Motifs of Genetic Interaction Networks

The analyses we have described so far were based on two-locus processes. In reality, however, a gene may have genetic interactions with multiple genes that together form a complex genetic interaction network (Boone et al. 2007). In sharp contrast to the network topology, genes are linearly aligned on a limited number of chromosomes, and therefore, the optimization of pairwise gene distances may be restricted

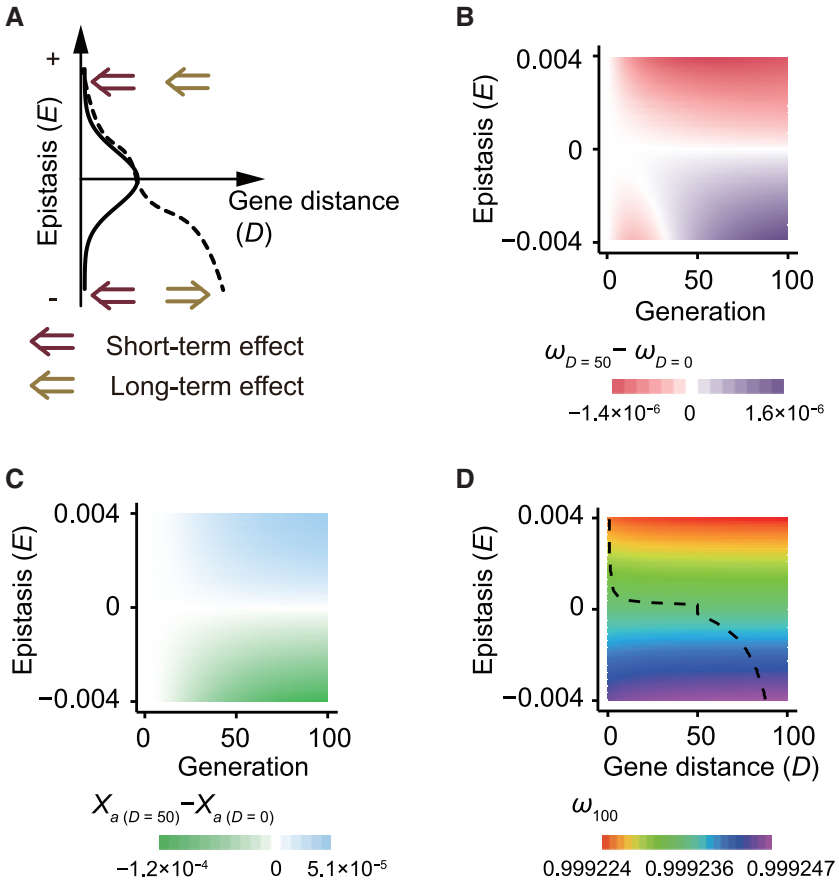

Fig. 1. Theoretical prediction and computational simulations of the negative E–D correlation. (*A*) The short- and long-term effects are in the same direction when epistasis is positive but are in the opposite directions when epistasis is negative. Thus, a negative E–D correlation can be predicted from the view of population genetics. (*B*) Fitness differences between a strain with D=50 and a strain with D=0 are plotted over 100 generations during simulations of in silico evolution. (*C*) The difference in allele *a*'s frequency (X_a) between strains with D=50 and D=0 are plotted over 100 generations during simulations of in silico evolution. (*D*) The average fitness of a population at the 100th generation (ω_{100}) is plotted against epistasis and D. For each epistasis, we defined all permitted D values with their resulting ω_{100}. If ω_{100} is smaller than that of the optimal distance (D_{opt}) by $< 10^{-7}$, the minimal selective coefficient that can be detected by nature given the effective population size ($N_e \approx 10^7$) of yeast, D is permitted. The mean of all permitted D values is plotted against epistasis (dashed line).

by the chromosomal localization of other genes. Thus, it remains unknown whether the negative E–D correlation would evolve in the context of a highly connected network of genetic interactions.

We first examined the impact of epistasis on the chromosomal order of genes in a star-like motif, which is typical in empirical genetic interaction networks (Costanzo et al. 2010). To this end, we built a toy motif in which a hub gene interacts with nine partner genes with different epistasis values ranging from −0.004 to 0.004 (fig. 2*A*). We fixed the chromosomal location of the hub gene and attempted to place partner genes on the same chromosome. We focused on the epistasis and gene distance of hub-containing gene pairs, and therefore, we placed all partner genes on the same side of the hub gene on the chromosome for convenience. We calculated ω_{100} for each of the total (9! =) 362,880 possible gene orders and found that ω_{100} varied among them (fig. 2*B*). Importantly, the gene order with the highest ω_{100} showed a perfect negative E–D correlation ($\rho_{ED} = -1$, fig. 2*C*), whereas the gene order with the lowest ω_{100} showed a perfect positive E–D correlation ($\rho_{ED} = 1$, fig. 2*D*). In fact, we found that ω_{100} was negatively correlated with ρ_{ED} (fig. 2*B*, $\rho = -0.94$, $P < 10^{-100}$, Spearman's correlation), implying that the negative E–D correlation itself is under natural selection. To understand whether the negative correlation between ω_{100} and ρ_{ED} is still present under the parameters derived from empirical data, we randomly chose epistasis values and fitness defects from two genome-wide studies in the budding yeast (Costanzo et al. 2010, 2016) and still observed strong negative correlations between ω_{100} and ρ_{ED} (supplementary fig. S5A and B and table S1, Supplementary Material online). And we also confirmed that the negative correlation between ω_{100} and ρ_{ED} was insensitive to D and initial allele frequencies (supplementary fig. S5C and D and table S1, Supplementary Material online).

Next, we calculated the distance (d) to the fittest gene order shown in figure 2*C*, which was defined as the number of differently placed genes (fig. 2*E*), for each possible gene order. We found that an increase in d reduced ω_{100} (fig. 2*E*), again emphasizing the impact of gene order on fitness. To further investigate the impact of the range of epistasis on ω_{100}, we generated a series of epistasis

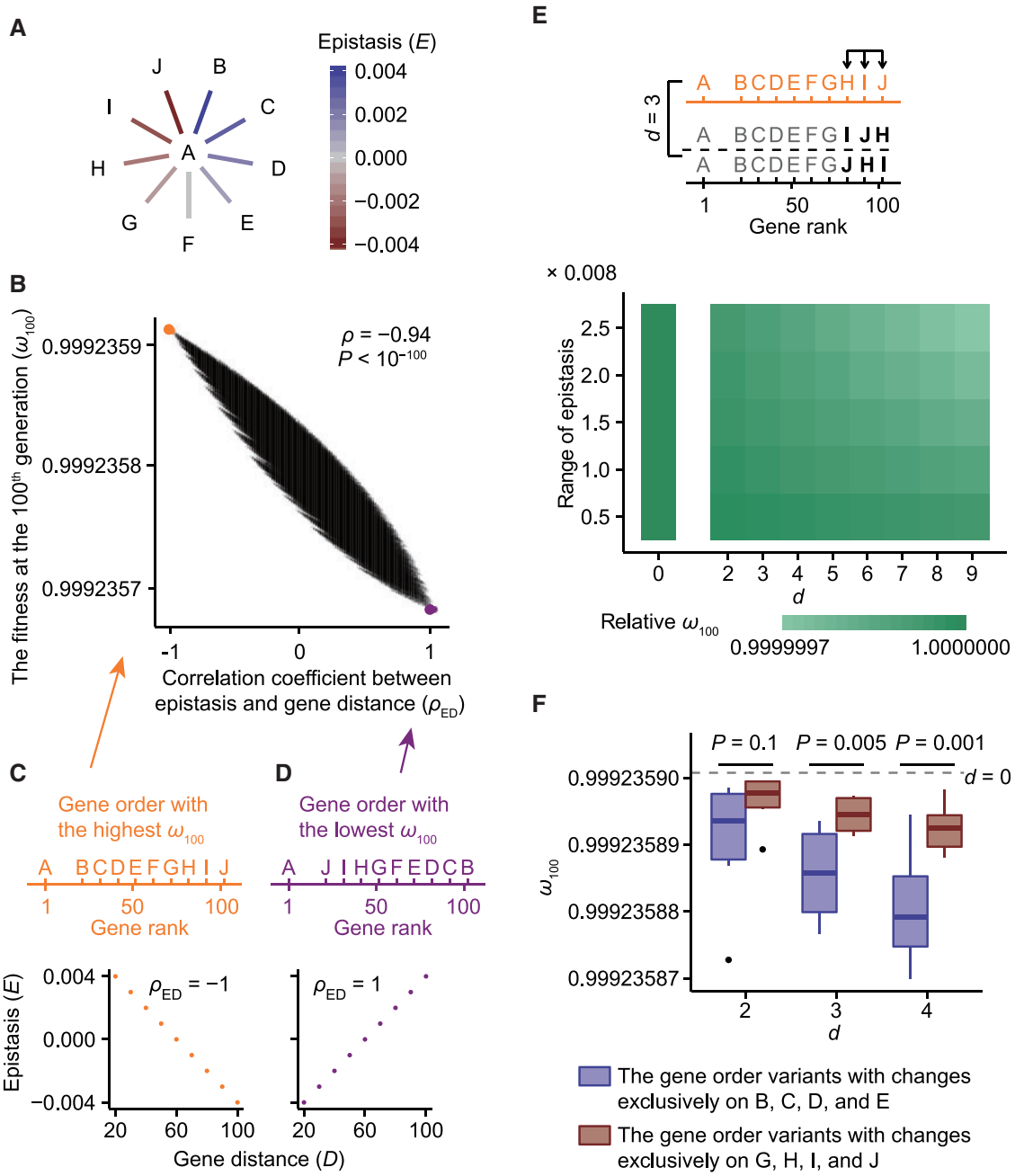

Fig. 2. The negative E–D correlation in star-like motifs of genetic interaction networks. (*A*) A toy model of a star-like motif in which gene A is the hub. Gene A has positive epistasis with genes B, C, D, and E, and negative epistasis with genes G, H, I, and J. The range of epistasis is (0.004 − [−0.004]=) 0.008 in this motif. (*B*) Spearman's correlation coefficient between epistasis and *D* (ρ_{ED}) varies among gene orders. The average fitness at the 100th generation (ω_{100}) is negatively correlated with ρ_{ED}. (*C*) The gene order with the highest ω_{100}. $\rho_{ED} = -1$. (*D*) The gene order with the lowest ω_{100}. ρ_{ED}=1. (*E*) The difference between a gene order and the gene order with the highest ω_{100} (*d*) is defined as the number of differently placed genes. Two examples with $d = 3$ are shown. The heat map of the relative ω_{100} (normalized to the highest ω_{100}) is shown. The average relative ω_{100} decreases with the increase of *d*. The reduction is more dramatic when the range of epistasis values is larger. (*F*) Shuffling among genes B, C, D, and E (positive epistasis with the hub gene A) have larger impact on ω_{100} than shuffling among genes G, H, I, and J (negative epistasis with the hub gene A). *P* values of one-tailed Mann–Whitney *U* test are shown. The gray dashed line indicates the ω_{100} of the optimal gene order in panel (C).

ranging from 0.004 to 0.020, shuffled the gene order, and recalculated the average ω_{100} of all gene orders with the same *d* (fig. 2*E*). We found that although it was always true that higher *d* led to a reduction of ω_{100}, the increase in the range of epistasis values enlarged the fitness differences among gene orders (fig. 2*E*).

Our model further predicted that positive epistasis should play a more important role in the evolution of gene order (fig. 1), such that altering the order of partner genes that have positive epistasis with the hub gene would lead to a larger fitness reduction. As expected, we observed that the gene-order variants with changes exclusively to the positively

epistatic genes generally had a larger reduction in fitness compared with those with changes exclusively to the negatively epistatic genes (fig. 2F).

Chromosomal Arrangement of Genes in All-Connected Motifs of Genetic Interaction Networks

We further examined the negative correlation between ω_{100} and ρ_{ED} in all-connected motifs. To this end, we built a toy all-connected motif with five nodes and assigned epistasis values in the range of −0.0036 to 0.0036 to edges (fig. 3A). Again, we observed a strong negative correlation between ω_{100} and ρ_{ED} (fig. 3B, $\rho = -0.87$, $P = 3.2 \times 10^{-38}$). The gene order with the highest fitness had a strong negative E–D correlation ($\rho_{ED} = -1.00$, fig. 3C), whereas the gene order with the lowest fitness had a positive E–D correlation ($\rho_{ED} = 0.39$, fig. 3D). Similarly, we confirmed that the negative correlation between ω_{100} and ρ_{ED} was insensitive to epistasis values, fitness defects, gene distances, and initial allele frequencies (supplementary fig. S6 and table S2, Supplementary Material online). Furthermore, we observed that the fitness of a gene order decreased with the increase in its d (fig. 3E), and this trend was stronger when the range of epistasis was larger (fig. 3E).

Negative E–D Correlation in *S. cerevisiae*

To investigate whether the negative E–D correlation is supported by empirical evidence, we retrieved the pairwise epistasis data generated by Costanzo *et al.*, who systematically measured the vegetative growth rates of both single and double mutants in the budding yeast *S. cerevisiae* and estimated epistasis values for ~26 million gene pairs (Costanzo et al. 2010, 2016). As expected, we observed a significant negative E–D correlation among linked genes (fig. 4A, $\rho = -0.15$, $P = 4.0 \times 10^{-8}$, $N = 1{,}254$). Consistent with this trend, unlinked genes on the same chromosome exhibited lower epistasis values (fig. 4A, gray dashed line). As a control, we permutated the gene orders and recalculated the correlation coefficients 1,000 times. We found that the negative E–D correlation disappeared after permutation (fig. 4B, $P < 0.001$, permutation test).

These observations are potentially attributable to a number of confounding factors. The first is mutational bias, such as tandem duplication. However, duplicate genes tend to have negative epistasis (Tischler et al. 2006; Dean et al. 2008; DeLuna et al. 2008; Musso et al. 2008; Vavouri et al. 2008; Qian et al. 2010), which should result in a positive E–D correlation. Nevertheless, we controlled for this mutational bias by randomly keeping only one gene in a gene family and still observed the negative E–D correlation (supplementary fig. S7A, Supplementary Material online, $\rho = -0.17$, $P = 1.8 \times 10^{-5}$, $N = 641$).

Second, genes with coordinated expression are clustered (Cho et al. 1998; Cohen et al. 2000; Boutanaev et al. 2002; Spellman and Rubin 2002; Williams and Bowles 2004). If coordinately expressed genes tend to have positive epistasis, the negative E–D correlation could result from these genes. To control for this effect, we first inferred expression pattern similarity for each pair of genes by calculating the correlation of gene expression levels in multiple conditions

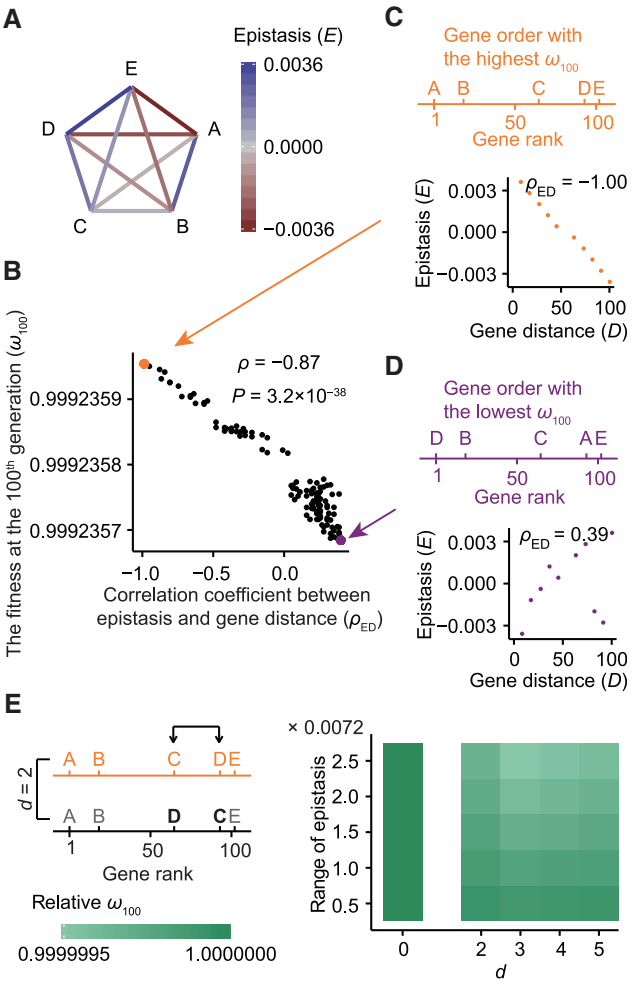

FIG. 3. The negative E–D correlation in all-connected motifs of genetic interaction networks. (*A*) A toy model of an all-connected motif. (*B*) The fitness at the 100th generation (ω_{100}) is negatively correlated with ρ_{ED}. (*C*) The gene order with the highest ω_{100}. $\rho_{ED} = -1.00$. (*D*) The gene order with the lowest ω_{100}. ρ_{ED}=0.39. (*E*) The heat map of the relative ω_{100} (normalized to the highest ω_{100}) is shown. The relative ω_{100} decreases with the increase of d. The reduction is more dramatic when the range of epistasis is larger.

(Qian and Zhang 2014). We did not observe a significant correlation between epistasis and expression similarity ($\rho = -0.015$, $P = 0.6$, $N = 1{,}202$). Nevertheless, we divided these gene pairs into two groups according to expression similarity, recalculated the E–D correlation within each group, and still observed significant negative E–D correlations (supplementary fig. S7B and C, Supplementary Material online). Similar results were obtained when we calculated the partial E–D correlation after controlling for expression similarity (partial $\rho = -0.14$, $P = 6.6 \times 10^{-7}$, $N = 1{,}202$). In addition, we also found that coordinated gene expression occurring through 3D chromatin interactions did not confound our results (supplementary fig. S7D, Supplementary Material online, $\rho = -0.17$, $P = 8.7 \times 10^{-6}$, $N = 704$), which was not unexpected, as 3D chromatin interactions do not influence recombination frequency.

Because functionally related genes are nonrandomly distributed on chromosomes (Wong and Wolfe 2005;

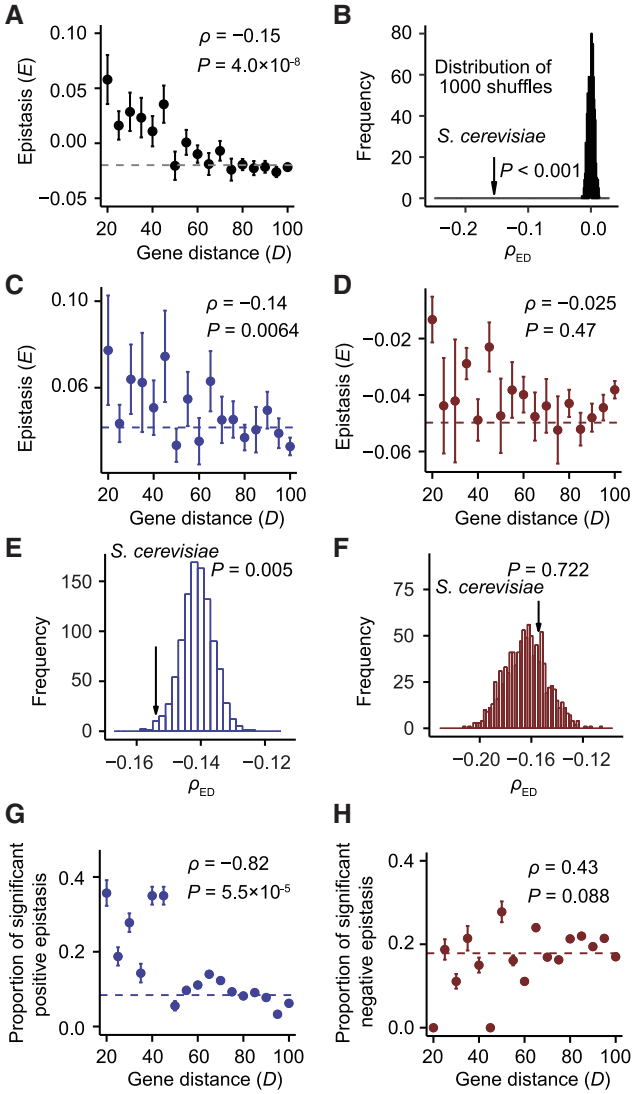

FIG. 4. A negative E–D correlation is observed in the empirical genetic interaction network of the budding yeast *S. cerevisiae*, and positive epistasis plays a more important role in its formation. (*A*) A significant negative E–D correlation is observed in *S. cerevisiae*. Gene pairs are separated into bins based on *D*, with equal width of five genes. The mean value of epistasis and the standard error of the mean (SEM) within each bin are shown. Spearman's correlation coefficient ρ and corresponding *P* values were calculated from the raw data (N=1,254). The gray dashed line shows the average epistasis among unlinked genes (D>100). (*B*) Distribution of E–D correlation coefficients in 1,000 shuffled genomes. The arrow indicates the observed correlation coefficient in *S. cerevisiae*. (*C* and *D*) A significant negative E–D correlation is observed among positively epistatic gene pairs (N=391) but not among negatively epistatic gene pairs (N=863). Spearman's correlation coefficient ρ and the corresponding *P* values are calculated from the raw data. The dashed line shows the average epistasis among unlinked genes. (*E*) The distribution of correlation coefficients in 1,000 artificial genomes in which values of positive epistasis are shuffled. The arrow indicates the E–D correlation coefficient in reality. (*F*) Similar to (E), values of negative epistasis are shuffled. (*G* and *H*) The proportion of gene pairs with significant positive epistasis is significantly correlated with *D*, but that with significant negative epistasis is not. SEMs are estimated based on binomial distribution. The dashed lines show the proportion of gene pairs with significant positive or negative epistasis among unlinked genes.

Slot and Rokas 2010) and functional relationships between genes may lead to epistasis, we next examined whether functional relationships could confound the negative E–D correlation. We observed similar negative E–D correlations for gene pairs with an either high- or low semantic similarity of GO terms in molecular functions (supplementary fig. S7E and F, Supplementary Material online), biological processes (supplementary fig. S7G and H, Supplementary Material online), and cellular components (supplementary fig. S7I and J, Supplementary Material online). Again, we calculated partial correlations controlling for semantic similarity of GO terms in molecular functions (partial $\rho = -0.15$, $P = 7.3 \times 10^{-7}$, $N = 1{,}151$), biological processes (partial $\rho = -0.15$, $P = 3.2 \times 10^{-7}$, $N = 1{,}151$), and cellular components (partial $\rho = -0.15$, $P = 3.2 \times 10^{-7}$, $N = 1{,}151$). All these results indicate that the functional relationship is not a confounding factor in the negative E–D correlation, which is not unexpected given that a large fraction of genetic interactions do not reflect functional relationships (He et al. 2010; Costanzo et al. 2016).

Finally, we investigated the impact of gene expression noise on the negative E–D correlation because it has been proposed that essential genes were colocalized in open chromatin regions to reduce gene expression noise (Batada and Hurst 2007; Chen and Zhang 2016). To control for this effect, we first calculated the average gene expression noise (Newman et al. 2006) for each gene pair. We used the distance of each coefficient of variation (CV) to a running median of CV values (DM) to quantify gene expression noise (Newman et al. 2006) in order to minimize the effect of gene expression magnitude on gene expression noise. We observed a negative E–D correlation for gene pairs with an either high- or low average DM (supplementary fig. S7K and L, Supplementary Material online). Again, we calculated the partial correlation controlling for gene expression noise and still observed a negative E–D correlation (partial $\rho = -0.16$, $P = 9.7 \times 10^{-3}$, $N = 258$).

Positive Epistasis Plays an Important Role in the Origin of the Negative E–D Correlation

Our model further predicted that the reduction of the distance between positively epistatic genes should play a more important role in the formation of the negative E–D correlation (figs. 1 and 2). Indeed, a significant negative E–D correlation was observed among positively epistatic gene pairs in *S. cerevisiae* (fig. 4C, $\rho = -0.14$, $P = 0.0064$, $N = 391$), whereas no significant correlation was observed among negatively epistatic gene pairs (fig. 4D, $\rho = -0.025$, $P = 0.47$, $N = 863$). We further verified the role of positive epistasis by shuffling epistasis values among all 391 positively epistatic gene pairs in *S. cerevisiae*. As expected, the E–D correlation was significantly weakened after the permutation (fig. 4E, $P = 0.005$, one-tailed permutation test). By contrast, no significant difference was observed after shuffling negative epistasis values (fig. 4F, $P = 0.722$, one-tailed permutation test), even though the latter analysis shuffled more gene pairs ($N = 863$).

In studies by Costanzo *et al.*, epistasis was classified into three categories: significantly positive, significantly negative, and nonsignificant (Costanzo et al. 2010, 2016). A strong negative correlation was observed between gene distance and the proportion of gene pairs with significant positive epistasis (fig. 4G, $\rho = -0.82$, $P = 5.5 \times 10^{-5}$), whereas no significant correlation was observed between gene distance and the proportion of gene pairs with significant negative epistasis (fig. 4H, $\rho = 0.43$, $P = 0.088$). All these observations emphasize the important role of positive epistasis in the origin of the negative E–D correlation.

Epistasis-Driven Evolution of Gene Order after the Whole Genome Duplication in Yeast

Thus far, we have observed a negative E–D correlation and demonstrated that the correlation was mainly attributable to positively epistatic gene pairs. The aforementioned theories, simulations, and empirical evidence led us to propose a hypothesis of epistasis-driven evolution of gene order in *S. cerevisiae*. The whole genome duplication (WGD, fig. 5A) and the subsequent extensive gene losses substantially changed the epistatic relations among genes and markedly rewired the genetic interaction network in yeast (Kellis et al. 2004; Dixon et al. 2008; Tischler et al. 2008; VanderSluis et al. 2010). For instance, only 29% of the identified synthetic lethality is conserved between *Schizosaccharomyces pombe* and *S. cerevisiae* (Dixon et al. 2008). At the same time, extensive chromosome rearrangement events occurred. For example, in a reconstructed ancestral species before the WGD (Byrne and Wolfe 2005), we identified the gene pairs that were located on the same chromosome, and surprisingly, 84.6% of them are localized on different chromosomes in *S. cerevisiae*. More strikingly, among gene pairs that are localized on the same chromosome in both the ancestral species and *S. cerevisiae*, 99.6% differ in gene distance. Based on these observations, we proposed that the rewired genetic interaction network drove the evolution of gene order, resulting in numerous chromosome rearrangement events (Kellis et al. 2004). When gene losses ceased, the rewiring of genetic interactions slowed, and the evolutionary force on gene distance also diminished. Consistently, synteny relationships are strongly conserved in the species of the *Saccharomyces sensu stricto* group (Kellis et al. 2003).

Our model predicted that the negative E–D correlation should be weaker if the gene order in *S. cerevisiae* has been unchanged since the WGD. The reason is that the gene order in the ancestor was not subject to the natural selection imposed by the genetic interaction network of the current *S. cerevisiae* genome. Furthermore, given the massive gene losses after the WGD, the genetic interaction network cannot be 100% conserved. To test this prediction, we calculated the gene distances in the reconstructed ancestral species mentioned earlier (Byrne and Wolfe 2005). Indeed, we found that the negative E–D correlation disappeared when the gene distances in *S. cerevisiae* were replaced by those in the ancestral species (fig. 5B, $\rho = 6.9 \times 10^{-3}$, $P = 0.26$, $N = 26{,}630$). This observation indicates that the negative E–D correlation in *S. cerevisiae* formed during the evolution of gene order after

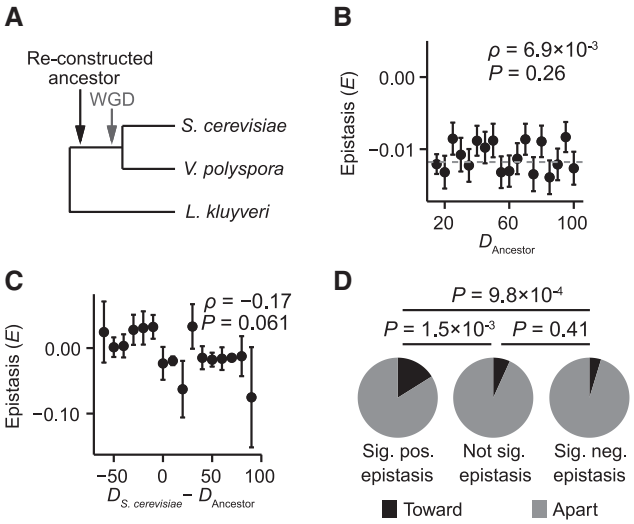

Fig. 5. The origin of the negative E–D correlation after the WGD in yeast. (*A*) Phylogenetic relationship among yeast species. The black arrow indicates the reconstructed ancestor and the gray arrow indicates the WGD event. (*B*) Negative E–D correlation is not observed when the gene order in *S. cerevisiae* is replaced by that in the reconstructed ancestor. Gene pairs are separated into bins based on *D*, with equal width of five genes. The mean and SEM of epistasis within each bin are shown. Spearman's correlation coefficient ρ and the corresponding *P* values were calculated from the raw data ($N = 26{,}630$). The dashed line shows the average epistasis among unlinked genes. (*C*) The change in *D* (*S. cerevisiae*—the reconstructed ancestor) is negatively correlated with the epistasis in *S. cerevisiae* ($N = 127$). (*D*) The *D* between a gene pair in *S. cerevisiae* is compared with that in the reconstructed ancestor. Proportions of gene pairs moving toward and away from each other among gene pairs with significant positive epistasis (left), nonsignificant epistasis (middle), and significant negative epistasis (right) are shown.

the WGD. Consistently, we observed that positively epistatic gene pairs decreased their distances whereas negatively epistatic gene pairs increased their distances during evolution (fig. 5C, $\rho = -0.17$, $P = 0.061$, $N = 127$).

To further test the role of positive epistasis in the evolution of gene order, we identified genes that were ancestrally linked (i.e., $D \leq 100$ in the reconstructed ancestor) and examined whether they moved toward or away from each other during evolution. Consistent with our model, genes with significant positive epistasis were more likely to move toward each other than genes without significant epistasis (fig. 5D, $P = 1.5 \times 10^{-3}$, two-tailed Fisher's exact test), whereas the difference between gene pairs with significant negative epistasis and those without significant epistasis was not significant (fig. 5D, $P = 0.41$, two-tailed Fisher's exact test). Together, these observations support our hypothesis of epistasis-driven evolution of gene order in yeast.

Genetic Interaction Network Accurately Predicts Gene Order in Yeast

Finally, we determined whether the gene order in *S. cerevisiae* could be successfully predicted by the empirical data of genetic interaction networks (Costanzo et al. 2010, 2016). To this end, we identified 22 all-connected three-node motifs in

which all three genes are localized within a 100-gene range on a chromosome. An example is shown in figure 6A, in which *VMA22* positively interacts with *SRB2* and negatively interacts with *AIM17*, and the epistatic interaction between *SRB2* and *AIM17* is weak. All six possible gene orders are enumerated in figure 6B. Among them, the first gene order exhibits a perfect E–D anticorrelation ($\rho = -1$, fig. 6C), which is exactly the prediction of our model. In fact, it is also the real order of these genes on chromosome VIII. The second gene order successfully places positively epistatic genes close to each other and negatively epistatic genes far from each other. Therefore, it is generally consistent with our prediction (fig. 6C). Either gene order was considered as being successfully predicted by our model if it actually occurred in the yeast genome.

We examined the accuracy of our prediction at the genomic scale. We first divided these 22 motifs into two groups based on the magnitude of $\sigma_{epistasis}$. Group L contains 11 motifs with larger $\sigma_{epistasis}$ and Group S contains 11 motifs with smaller $\sigma_{epistasis}$ (fig. 6D). We found that the gene orders of seven (out of 11) motifs in Group L were precisely predicted by our model, and the accuracy was significantly higher than the random expectation (17.6%, fig. 6E, $P = 0.001$, permutation test). By contrast, the predictive accuracy in Group S (2 out of 11) was not significantly different from the random expectation (fig. 6E, $P = 0.607$, permutation test) because low variation among epistasis values in a motif reduces the selective coefficients (figs. 2E and 3E). More broadly, the proportion of successful predictions (first and second gene orders in fig. 6B) is 100% in Group L, significantly higher than the random expectation (33.3%, fig. 6F, $P < 0.001$, permutation test). This high predictive power suggests that epistasis plays a vital role in driving the evolution of gene order. We further identified 243 and 1,302 all-connected motifs within the range of 150 and 200 genes on the same chromosome, respectively, and again confirmed the predictive power of the genetic interaction network (supplementary fig. S8A–F, Supplementary Material online). Moreover, the predictive power of epistasis on gene order is independent of expression similarity because the latter could not accurately predict gene order (supplementary fig. S9A–I, Supplementary Material online).

We then determined whether gene order could be predicted when the genetic interaction network is incomplete. To this end, we identified 92 star-like motifs in which all genes are on the same chromosome and at least one gene with $D \leq 40$ to the hub gene. For example, *SFH1*, which encodes a component of a chromatin remodeling complex, genetically interacts with 90 genes on the same chromosome (fig. 6G shows 20 genes with $D < 100$ to *SFH1*). Because the negative E–D correlation is mainly contributed by positively epistatic gene pairs in theory (figs. 1 and 2), our model predicted that genes having strong positive epistasis with *SFH1* should be located close to it on the chromosome. Indeed, these genes (*VRP1*, *FKS1*, *RPL26A*, *VPS38*, *DCR2*) are located close to *SFH1*. Specifically, the gene (*VRP1*) having the strongest positive epistatic interaction with *SFH1* was the closest gene to *SFH1* on the chromosome (fig. 6H).

To examine the predictive accuracy at the genomic scale, we divided these 92 star-like motifs into two groups based on the difference between the top two highest epistasis values in the motif ($Diff_{epistasis}$, fig. 6J). We found that the proportion of successful predictions was 32.6% in Group L and 21.7% in Group S, both of which were significantly higher than random expectation (fig. 6K, $P < 0.001$ for Group L, and $P = 0.001$ for Group S, permutation test). Furthermore, despite the lower predictive power, our model could still predict gene locations in 707 star-like motifs in which the closest gene to the hub was within $D = 90$ (supplementary fig. S8G and H, Supplementary Material online, $P = 0.002$ for Group L and $P = 0.036$ for Group S, permutation test).

Discussion

In our study, we provided results from both computational simulation and empirical data analysis to support the role of genetic interaction networks in driving the evolution of gene order. We performed simulations with different sets of parameters, some of which were from empirical data or within the range of empirical data (Schacherer et al. 2009; Costanzo et al. 2010). The negative E–D correlation was observed with all parameter sets (supplementary figs. S4–S6, Supplementary Material online), consistent with both theoretical predictions (fig. 1A) and empirical results (fig. 4).

The epistasis values were estimated mainly from null alleles (Costanzo et al. 2010); thus, it remains unknown whether our conclusion applies to other deleterious alleles. Fortunately, Costanzo *et al.* also estimated 7,786,453 pairwise epistasis values for either decreased abundance by mRNA perturbation (DAmP) alleles or temperature-sensitive (*ts*) alleles with mutations typically changing coding sequences, which offered us the opportunity to address this question (Costanzo et al. 2010, 2016). We found that the epistasis values of null alleles were significantly correlated with those of other deleterious alleles of the same gene (supplementary fig. S10A–L, Supplementary Material online), suggesting that the negative E–D correlation is likely a general phenomenon for various deleterious alleles. More importantly, because DNA deletion events that lead to null alleles are frequently observed in yeast natural populations (Schacherer et al. 2009), epistasis among null alleles by itself may be sufficiently strong to drive the evolution of gene distance.

In principle, if the frequencies of deleterious alleles in a population are significantly reduced by natural selection, the advantage of genetic linkage could be small for this population. However, numerous deleterious mutations (e.g., deletion of a whole gene, nonsense mutations, missense mutations, and mutations altering start codons) have been observed in natural populations (Liti et al. 2009; Schacherer et al. 2009), suggesting that the frequencies of deleterious alleles may not be low. This phenomenon is probably due to the antagonistic pleiotropic effects of a mutation in multiple environments (Qian et al. 2012).

It is worth noting that theoretical analysis predicts that epistasis can drive the evolution of recombination frequency rather than gene distance. In addition to gene distance, 1) the

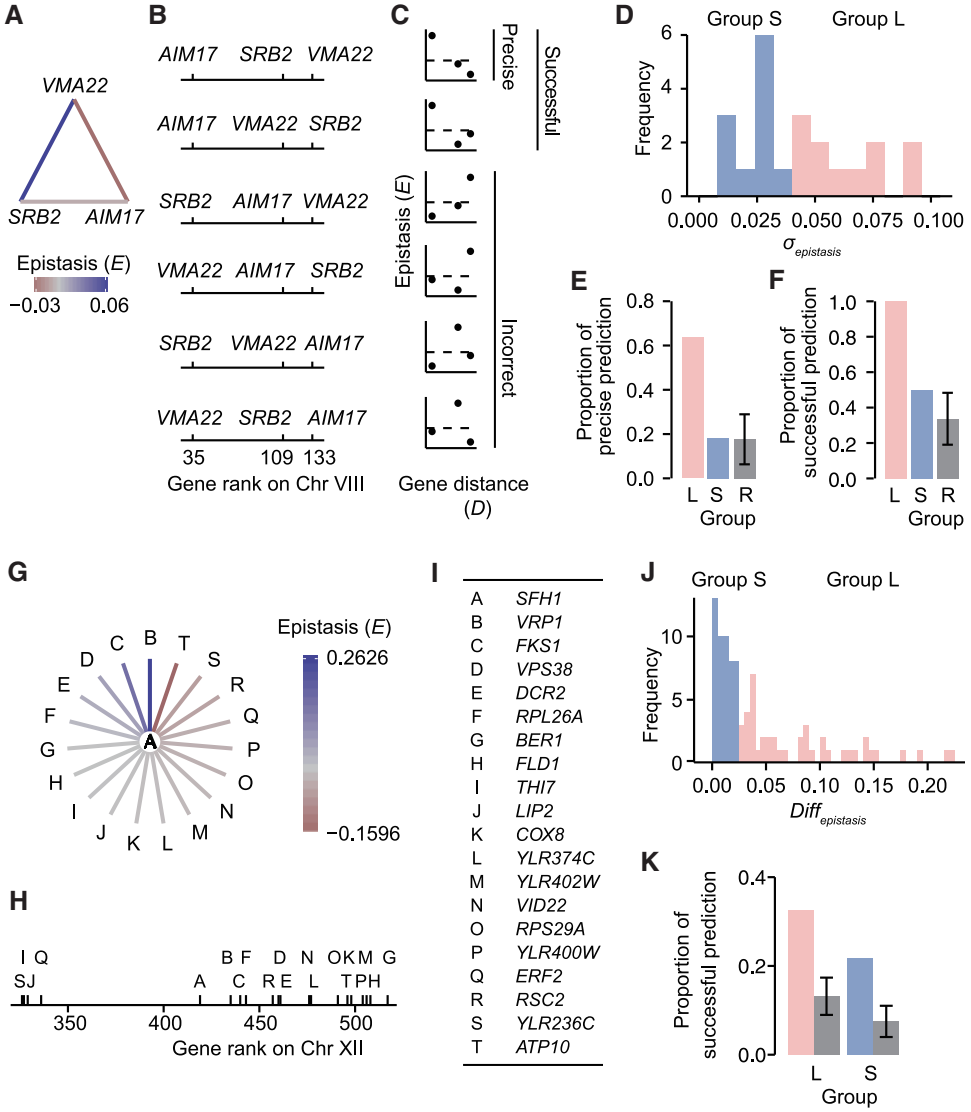

Fig. 6. Gene orders can be predicted from the yeast genetic interaction network. (*A*) A three-node all-connected motif in the yeast genetic interaction network. The standard deviation of epistasis values ($\sigma_{epistasis}$) in this motif is 0.05. (*B*) Six possible gene orders are listed. The first is the observed gene order on chromosome VIII in yeast. (*C*) The E–D relationship for the six possible gene orders in (B). We consider the prediction successful if the first two gene orders are observed because they exhibit negative E–D correlations. In particular, the prediction is precise if the first gene order is observed. The dashed line represents $E = 0$. (*D*) A histogram of the standard deviation of epistasis values ($\sigma_{epistasis}$) based on which genes are divided into two groups with similar sizes: Group S and Group L. (*E*) The proportions of precise prediction of Group S, Group L, and random (R) expectations based on permutation. Error bars represent standard deviations among 1,000 permutations. (*F*) The proportions of successful predictions, similar to (E). (*G–I*) A star-like motif in the yeast genetic interaction network and the gene order on chromosome XII in yeast. The hub gene *SFH1* and 19 genes with the distance <100 to *SFH1* are shown. (*J*) A histogram of differences in the two largest epistasis values in the motif ($Diff_{epistasis}$) based on which genes are divided into two groups with similar sizes: Group S and Group L. (*K*) Proportion of successful predictions of the closest gene in Group S and Group L. Their respective random expectations based on permutation are shown in gray. Error bars represent standard deviations among 1,000 permutations.

lengths of intergenic regions and 2) recombination hot/cold spots could also affect recombination frequency. Nevertheless, a strong correlation between gene distance and physical distance, that is, the number of nucleotides between two genes, was observed ($\rho = 0.998$, $P < 10^{-100}$, $N = 1{,}624{,}935$), suggesting that the effect of (1) is negligible. Furthermore, physical distance is highly correlated with recombination frequency when the physical distance is < 180 kb (supplementary fig. S1C, Supplementary Material online, $\rho = 0.88$, $P < 10^{-100}$, $N = 336{,}227$), suggesting that the effect of (2) is also limited. As a result, recombination frequency and gene distance are highly correlated (supplementary fig. S1B, Supplementary Material online, $\rho = 0.87$, $P < 10^{-100}$, $N = 337{,}892$, for gene pairs with $D \leq 100$). Because the recombination frequency was only measured for ~58% of gene pairs in *S. cerevisiae* (Mancera et al. 2008) and was unknown for the reconstructed ancestral species, we used gene distance to approximate recombination frequency in this study. The optimization of gene distances among

multiple gene pairs may often involve the rearrangement of genes on a chromosome.

It is also worth noting that synthetic genetic array (SGA), the experimental strategy used to generate double mutants, is based on recombination between two null alleles (Tong et al. 2001; Costanzo et al. 2010). Thus, double mutants for linked genes may have a smaller initial frequency, which may lead to inaccuracy in estimating fitness values for these mutants. To determine whether such potential experimental bias has any impact on our results, we calculated the partial E–D correlation after controlling for the double mutant fitness values. Again, we observed a strong negative E–D correlation (partial $\rho = -0.17$, $P = 1.3 \times 10^{-9}$, $N = 1{,}254$). This result was not unexpected because the fitness of double mutants and gene distance was only weakly correlated ($\rho = 0.086$, $P = 0.002$, $N = 1{,}254$). More discussion on the caveats in the analysis of empirical data is included in supplementary note S2, Supplementary Material online.

In a recently study, ~1,800 genetic suppression interactions were identified in the budding yeast (van Leeuwen et al. 2016), which can be regarded as an extreme form of positive genetic interactions. Among them, nine gene pairs are tightly linked ($D \leq 2$, supplementary table S3, Supplementary Material online). This number is significantly greater than the random expectation (permutation test, $P = 0.01$), again supporting our model.

The selection on clusters of locally adapted alleles, which sometimes were captured in chromosomal inversions, has been studied previously (Yeaman 2013; Kirkpatrick 2017). In our study, we proposed a hypothesis (the negative E–D correlation) on deleterious alleles and tested this hypothesis with analyses on empirical data and simulation, both in the context of genetic interaction networks. Therefore, our results provide additional clues for understanding the basic principles of genome organization. In particular, the origin and maintenance of clusters of genes in the same metabolic pathway, which have puzzled scientists for years (Wong and Wolfe 2005; Slot and Rokas 2010; Lang and Botstein 2011), could be well explained by genetic interactions. Genes in a linear metabolic pathway exhibit positive epistasis because 1) double mutants and single mutants have identical impacts on destroying the function of the pathway (He et al. 2010), and 2) the accumulation of deleterious intermediate products resulting from a loss-of-function mutation in a downstream gene of a metabolic pathway may be prevented by a loss-of-function mutation in an upstream gene (Wong and Wolfe 2005; Slot and Rokas 2010; Lang and Botstein 2011). Because genetic linkage is advantageous among genes with positive epistasis, the clustering of genes in the same linear metabolic pathway, such as genes in the galactose utilization pathway or allantoin degradation pathway, is favored by natural selection (Wong and Wolfe 2005; Slot and Rokas 2010).

Many factors have been reported to influence the evolution of gene order, such as tandem gene duplication, position effects on gene expression noise (Batada and Hurst 2007; Chen and Zhang 2016), coordinated gene expression among neighboring genes (Cho et al. 1998; Cohen et al. 2000; Boutanaev et al. 2002; Spellman and Rubin 2002; Williams and Bowles 2004), clustering of functionally related genes (Wong and Wolfe 2005; Slot and Rokas 2010), among many others. In this study, we provided evidence that in addition to these factors, the genetic interaction network also played an important role in driving the evolution of gene order. Because the empirical data of epistasis (fitness values) are available in the budding yeast, an evolutionary simulation integrating many genetic interactions is possible, which makes epistasis unique among all factors driving the evolution of gene order. Based on the empirical data of yeast genetic interaction network, our simulation indicates that the selective coefficient is on the order of 10^{-7}, suggesting that epistasis may play an important role in determining gene order in yeast, a species with a relative large effective population size (10^7). A negative E–D correlation is expected to be observed in species with a smaller effective population size only if the range of epistasis is larger than that in yeast (figs. 2 and 3). Because the genome-wide empirical data of epistasis are unavailable in species with a smaller effective population size (e.g., humans and flies), it requires further investigations whether the genetic interaction network plays a role in the evolution of gene order in these species in the future.

Materials and Methods

Genomes

The genome annotation of *S. cerevisiae* was downloaded from the *Saccharomyces* Genome Database (SGD, http://www.yeastgenome.org, version R64). The gene order of the reconstructed ancestor before the WGD (Gordon et al. 2009) was downloaded from the Yeast Gene Order Browser (YGOB) (Byrne and Wolfe 2005).

Simulation

To examine the fitness effect of gene distance (D), two-locus dynamics under selection were simulated for populations with different D (from 0 to 100 in 1-gene increments), as well as a series of epistasis values (E, from −0.004 to 0.004 in 0.001 increments) between two loci. The recombination frequency (R) between these two loci was estimated from D using the equation below:

$$R = D \times 0.004 + 0.064$$

which is the linear relationship estimated from the distance between two genes (D) and the empirical recombination frequencies (R) between them quantified in a previous study (Mancera et al. 2008). The simulation was performed in haploid organisms, in order to be in alignment with the analyses of empirical data, in which epistasis values were quantified in haploid yeast (Costanzo et al. 2010, 2016). A and B are wild-type alleles of two di-allelic loci on the same chromosome; a and b are their deleterious alleles with fitness values of $\omega_{aB} = \omega_{Ab} = 0.992$. The fitness of the wild-type (ω_{AB}) was defined as 1, and the epistasis E was defined as $\omega_{ab} - \omega_{aB}\omega_{Ab}$. The initial allele frequencies of a and b were both 0.1. The two loci were initially under linkage equilibrium, and therefore the frequencies of $AB(X_{AB})$, $Ab(X_{Ab})$, $aB(X_{aB})$, and $ab(X_{ab})$ were 0.81, 0.09, 0.09, and 0.01, respectively. After random mating,

selection, and recombination, the frequencies of the four genotypes in the next generation were calculated with the following equations (Nei 1967):

$$X'_{AB} = \frac{X_{AB}\omega_{AB}}{\bar{\omega}} - \frac{R(X_{AB}\omega_{AB}X_{ab}\omega_{ab} - X_{aB}\omega_{aB}X_{Ab}\omega_{Ab})}{\bar{\omega}^2}$$

$$X'_{Ab} = \frac{X_{Ab}\omega_{Ab}}{\bar{\omega}} + \frac{R(X_{AB}\omega_{AB}X_{ab}\omega_{ab} - X_{aB}\omega_{aB}X_{Ab}\omega_{Ab})}{\bar{\omega}^2}$$

$$X'_{aB} = \frac{X_{aB}\omega_{aB}}{\bar{\omega}} + \frac{R(X_{AB}\omega_{AB}X_{ab}\omega_{ab} - X_{aB}\omega_{aB}X_{Ab}\omega_{Ab})}{\bar{\omega}^2}$$

$$X'_{ab} = \frac{X_{ab}\omega_{ab}}{\bar{\omega}} - \frac{R(X_{AB}\omega_{AB}X_{ab}\omega_{ab} - X_{aB}\omega_{aB}X_{Ab}\omega_{Ab})}{\bar{\omega}^2}$$

$\bar{\omega}$ is the average fitness of a population and was calculated as follows:

$$\bar{\omega} = X_{AB}\omega_{AB} + X_{ab}\omega_{ab} + X_{aB}\omega_{aB} + X_{Ab}\omega_{Ab}$$

For each population, the average fitness ($\bar{\omega}$) and the frequencies of deleterious alleles (X_a and X_b) were recorded in each generation of the simulation. The difference between the average fitness of two populations with different D ($D = 50$ and $D = 0$) reflects whether linkage is favored by natural selection or not. The frequency difference of the deleterious alleles between these two populations is related to the long-term effects of recombination. We also performed the simulation with a variety of parameter values (supplementary fig. S4, Supplementary Material online).

The average fitness of a population at the 100th generation (ω_{100}) was used to infer the fitness effect of gene distance at a given epistasis value. For each epistasis value, we could identify an optimal gene distance (D_{opt}) as well as a series of permitted gene distances ($D_{permitted}$). The difference in ω_{100} between $D_{permitted}$ and D_{opt} was $<10^{-7}$. The mean of all $D_{permitted}$ values was calculated. In addition, our simulation did not introduce new mutations to the population, and therefore, the frequencies of deleterious alleles reduced over generations. Nevertheless, the negative E–D correlation was also observed at the 50th and 200th generation (supplementary fig. S3, Supplementary Material online).

Toy Motifs

We built a toy motif in a genetic interaction network to simulate the fitness effect of gene distance. In the star-like motif (fig. 2A), a hub gene (A) interacted with nine partners (genes B, C, D, E, F, G, H, I, and J) with nine epistasis values ranging from 0.004 to −0.004 in −0.001 increments. We attempted to place nine partner genes on the same side of the hub gene on a chromosome. Because we focused on the epistasis between the hub gene and the partners, placing the partners on different sides would not affect the results. We fixed the chromosomal location of the hub gene, and the partner genes were placed at $D = 20$, 30, 40, 50, 60, 70, 80, 90, and 100 from the hub gene. The locations of the nine partners were shuffled while keeping the epistasis values unchanged. The fitness of a gene order was defined as the average ω_{100} for all nine hub-containing gene pairs. Thus, we could calculate the fitness for a total of (9! =) 362,880 possible gene orders. Other parameters (epistasis, fitness defect, D, and initial allele frequency) were also used to test whether the simulation result was parameter sensitive (supplementary fig. S5, Supplementary Material online). Here, epistasis values and fitness defects were randomly chosen from the empirical data in previous studies (Costanzo et al. 2010, 2016).

For the all-connected motif (fig. 3A), five genes (A, B, C, D, and E) interacted with each other with ten epistasis values ranging from −0.0036 to 0.0036 in 0.0008 increments. The five genes were placed on the same chromosome, and the gene locations were set as 1, 19, 65, 93, and 102, so that the maximum D in the motif was 100. The locations of these five genes were shuffled, whereas the epistasis values were kept unchanged. Similar to the star-like motifs, the fitness of a gene order was defined as the average ω_{100} for all ten gene pairs. The fitness was calculated for a total of (5! =) 120 possible gene orders. Other parameters (epistasis, fitness defect, D, and initial allele frequency) were also used to determine whether the simulation result was parameter sensitive (supplementary fig. S6, Supplementary Material online).

Epistasis

Epistasis values were retrieved from the studies of Costanzo *et al.* (Costanzo et al. 2010, 2016). Because the epistasis values of the overlapping gene pairs in these studies were highly correlated ($r = 0.71$, $P < 10^{-100}$, Pearson's correlation, $N = 2{,}604{,}539$), we merged their epistasis values. A pair of genes was filtered if (a) the epistasis value between their null mutations was not a number (NaN) or (b) the epistasis values of the same null mutation pair had opposite signs in reciprocal crosses or in different studies. We also removed the gene pairs with the null mutation in at least one gene resulting in a higher fitness value than that of the wild-type. This filtering was performed because only a small number of antagonistic pleiotropic genes were detected in rich media (Qian et al. 2012). Thus, the elevated fitness observed upon gene deletion in Costanzo *et al.* was likely due to inaccurate estimation of fitness. If epistasis between a pair of genes was examined multiple times, we used the epistasis value with the smallest P value. Following Costanzo *et al.* (2010), epistasis values with $P < 0.05$ were classified as significant.

In addition to the two data sets from Costanzo *et al.*, several other studies have also estimated epistasis values using high-throughput strategies (Boone et al. 2007). In principle, we could have included all of them in this study. However, for practical reasons these data sets were not suitable for our study. First, it would not be appropriate to combine other data sets with that of Costanzo *et al.* because epistasis values from multiple studies potentially have different sources of error, especially if they followed different protocols. Second, the sample sizes of epistasis data of linked genes from other studies were not sufficiently large to examine the E–D correlation. Because the data set from Costanzo *et al.* represents the largest available so far, containing ~27 times more

significant epistasis than the total of all other studies, we focused on these epistasis data in this study.

Recombination Frequency

Genotypes of meiosis products from 46 tetrads were downloaded from Mancera et al. (2008) at http://www.ebi.ac.uk/~huber/recombination/. A pair of DNA markers was filtered if the genotypes were available in less than half of the spores (< 92 spores). For each pair of markers, spores with parental genotypes (N_p) and nonparental genotypes (N_{np}) were counted. The recombination frequency of a pair of markers (R_m) was calculated as follows:

$$R_m = N_{np}/(N_p + N_{np})$$

The recombination frequency between a pair of genes (R_g) was defined as the average recombination frequency of all marker pairs within the pair of genes, which was calculated as follows:

$$R_g = \sum_{i=1} \sum_{j=1} R_{m\ ij}/ij$$

Subscripts i and j represent ith and jth markers in each gene, respectively.

Identification of Duplicate Genes

All-against-all BlastP was performed to search for duplicate genes. Gene pairs with E values $< 10^{-10}$ were defined as duplicate genes.

Expression Similarity

Expression profiles of 6,359 genes in 40 studies were compiled by a previous study (Kafri et al. 2005). Pearson's correlation coefficient for two genes was calculated within each study in the compiled data set, and expression similarity between these two genes was defined as the average correlation coefficient among studies.

Other Yeast Functional Genomic Data

Protein–protein interaction data were obtained from the SGD. Three-dimensional chromatin colocalization data were retrieved from Duan et al. (2010). Genome-wide gene expression noise data were downloaded from Newman et al. (2006). A list of essential genes was downloaded from the Database of Essential Genes (DEG, version 10.6) (Luo et al. 2014). Semantic similarity of Gene Ontology (GO) terms was calculated using the R package GOSemSim (version 1.22.0) (Yu et al. 2010).

Shuffling of Gene Positions or Epistasis Values

To obtain the null distribution of E–D correlation coefficients, we shuffled gene positions while keeping the genome structure of *S. cerevisiae* unchanged (i.e., number of chromosomes and number of genes on each chromosome). The epistasis value of each gene pair was also kept unchanged. Gene distances were then calculated based on the new genomic locations.

To estimate the relative importance of different epistasis categories (positive and negative) in shaping the E–D correlation, we shuffled epistasis values of one category while keeping the other category unchanged. Gene positions were not changed during this process.

Estimating the Predictive Accuracy of Gene Order

To examine the accuracy of our prediction of gene order, we estimated the success rate by chance (gray bars in fig. 6 and supplementary figs. S8 and S9, Supplementary Material online) by randomly assigning the positions of genes 1,000 times. For each permutation, the average proportion of successful predictions among motifs was estimated. The mean and standard deviation among the 1,000 permutations were then calculated.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Author Contributions

Y.-F.Y., W.C., and W.Q. conceived the research; Y.-F.Y., W.C., and W.Q. analyzed the data; and Y.-F.Y., S.W., and W.Q. wrote the article.

Acknowledgments

We thank Bin He and Weiwei Zhai for discussion. This work was supported by grants from the National Natural Science Foundation of China to W.Q. (91731302).

References

Batada NN, Hurst LD. 2007. Evolution of chromosome organization driven by selection for reduced gene expression noise. *Nat Genet.* 39(8):945–949.

Boone C, Bussey H, Andrews BJ. 2007. Exploring genetic interactions and networks with yeast. *Nat Rev Genet.* 8(6):437–449.

Boutanaev AM, Kalmykova AI, Shevelyov YY, Nurminsky DI. 2002. Large clusters of co-expressed genes in the Drosophila genome. *Nature* 420(6916):666–669.

Byrne KP, Wolfe KH. 2005. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res.* 15(10):1456–1461.

Charlesworth B. 1990. Mutation-selection balance and the evolutionary advantage of sex and recombination. *Genet Res.* 55(3):199–221.

Charlesworth D, Charlesworth B. 2011. Mimicry: the hunting of the supergene. *Curr Biol.* 21(20):R846–R848.

Chen X, Zhang J. 2016. The genomic landscape of position effects on protein expression level and noise in yeast. *Cell Syst.* 2(5):347–354.

Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, et al. 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* 2(1):65–73.

Cohen BA, Mitra RD, Hughes JD, Church GM. 2000. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat Genet.* 26(2):183–186.

Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, Ding H, Koh JL, Toufighi K, Mostafavi S, et al. 2010. The genetic landscape of a cell. *Science* 327(5964):425–431.

Costanzo M, VanderSluis B, Koch EN, Baryshnikova A, Pons C, Tan G, Wang W, Usaj M, Hanchard J, Lee SD, et al. 2016. A global genetic interaction network maps a wiring diagram of cellular function. *Science* 353(6306):aaf1420.

Dean EJ, Davis JC, Davis RW, Petrov DA. 2008. Pervasive and persistent redundancy among duplicated genes in yeast. *PLoS Genet.* 4(7):e1000113.

DeLuna A, Vetsigian K, Shoresh N, Hegreness M, Colon-Gonzalez M, Chao S, Kishony R. 2008. Exposing the fitness contribution of duplicated genes. *Nat Genet.* 40(5):676–681.
Dixon SJ, Fedyshyn Y, Koh JL, Prasad TS, Chahwan C, Chua G, Toufighi K, Baryshnikova A, Hayles J, Hoe KL, et al. 2008. Significant conservation of synthetic lethal genetic interaction networks between distantly related eukaryotes. *Proc Natl Acad Sci U S A.* 105(43):16653–16658.
Duan Z, Andronescu M, Schutz K, McIlwain S, Kim YJ, Lee C, Shendure J, Fields S, Blau CA, Noble WS. 2010. A three-dimensional model of the yeast genome. *Nature* 465(7296):363–367.
Eshel I, Feldman MW. 1970. On the evolutionary effect of recombination. *Theor Popul Biol.* 1(1):88–100.
Feldman MW, Christiansen FB, Brooks LD. 1980. Evolution of recombination in a constant environment. *Proc Natl Acad Sci U S A.* 77(8):4838–4841.
Ghanbarian AT, Hurst LD. 2015. Neighboring genes show correlated evolution in gene expression. *Mol Biol Evol.* 32(7):1748–1766.
Gordon JL, Byrne KP, Wolfe KH. 2009. Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern *Saccharomyces cerevisiae* genome. *PLoS Genet.* 5(5):e1000485.
He X, Qian W, Wang Z, Li Y, Zhang J. 2010. Prevalent positive epistasis in *Escherichia coli* and *Saccharomyces cerevisiae* metabolic networks. *Nat Genet.* 42(3):272–276.
Hurst LD, Pal C, Lercher MJ. 2004. The evolutionary dynamics of eukaryotic gene order. *Nat Rev Genet.* 5(4):299–310.
Kafri R, Bar-Even A, Pilpel Y. 2005. Transcription control reprogramming in genetic backup circuits. *Nat Genet.* 37(3):295–299.
Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* 428(6983):617–624.
Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423(6937):241–254.
Kirkpatrick M. 2017. The evolution of genome structure by natural and sexual selection. *J Hered.* 108(1):3–11.
Kondrashov AS. 1982. Selection against harmful mutations in large sexual and asexual populations. *Genet Res.* 40(3):325–332.
Kondrashov AS. 1988. Deleterious mutations and the evolution of sexual reproduction. *Nature* 336(6198):435–440.
Kouyos RD, Silander OK, Bonhoeffer S. 2007. Epistasis between deleterious mutations and the evolution of recombination. *Trends Ecol Evol.* 22(6):308–315.
Lang GI, Botstein D. 2011. A test of the coordinated expression hypothesis for the origin and maintenance of the GAL cluster in yeast. *PLoS One* 6(9):e25290.
Lawrence J. 1999. Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes. *Curr Opin Genet Dev.* 9(6):642–648.
Lawrence JG. 2002. Shared strategies in gene organization among prokaryotes and eukaryotes. *Cell* 110(4):407–413.
Lercher MJ, Urrutia AO, Hurst LD. 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat Genet.* 31(2):180–183.
Liao BY, Zhang J. 2008. Coexpression of linked genes in Mammalian genomes is generally disadvantageous. *Mol Biol Evol.* 25(8):1555–1565.
Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP, Roberts IN, Burt A, Koufopanou V, et al. 2009. Population genomics of domestic and wild yeasts. *Nature* 458(7236):337–341.
Luo H, Lin Y, Gao F, Zhang CT, Zhang R. 2014. DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. *Nucleic Acids Res.* 42(Database issue):D574–D580.
Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM. 2008. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature* 454(7203):479–485.
Musso G, Costanzo M, Huangfu M, Smith AM, Paw J, San Luis BJ, Boone C, Giaever G, Nislow C, Emili A, et al. 2008. The extensive and condition-dependent nature of epistasis among whole-genome duplicates in yeast. *Genome Res.* 18(7):1092–1099.
Nei M. 1967. Modification of linkage intensity by natural selection. *Genetics* 57(3):625–641.
Nei M. 1969. Linkage modifications and sex difference in recombination. *Genetics* 63(3):681–699.
Newman JR, Ghaemmaghami S, Ihmels J, Breslow DK, Noble M, DeRisi JL, Weissman JS. 2006. Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature* 441(7095):840–846.
Pal C, Hurst LD. 2003. Evidence for co-evolution of gene order and recombination rate. *Nat Genet.* 33(3):392–395.
Phillips PC. 2008. Epistasis – the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet.* 9(11):855–867.
Qian W, Liao BY, Chang AY, Zhang J. 2010. Maintenance of duplicate genes and their functional redundancy by reduced expression. *Trends Genet.* 26(10):425–430.
Qian W, Ma D, Xiao C, Wang Z, Zhang J. 2012. The genomic landscape and evolutionary resolution of antagonistic pleiotropy in yeast. *Cell Rep.* 2(5):1399–1410.
Qian W, Zhang J. 2008. Evolutionary dynamics of nematode operons: easy come, slow go. *Genome Res.* 18(3):412–421.
Qian W, Zhang J. 2014. Genomic evidence for adaptation by gene duplication. *Genome Res.* 24(8):1356–1362.
Schacherer J, Shapiro JA, Ruderfer DM, Kruglyak L. 2009. Comprehensive polymorphism survey elucidates population structure of Saccharomyces cerevisiae. *Nature* 458(7236):342–345.
Slot JC, Rokas A. 2010. Multiple GAL pathway gene clusters evolved independently and by different mechanisms in fungi. *Proc Natl Acad Sci U S A.* 107(22):10136–10141.
Spellman PT, Rubin GM. 2002. Evidence for large domains of similarly expressed genes in the Drosophila genome. *J Biol.* 1(1):5.
Tischler J, Lehner B, Chen N, Fraser AG. 2006. Combinatorial RNA interference in *Caenorhabditis elegans* reveals that redundancy between gene duplicates can be maintained for more than 80 million years of evolution. *Genome Biol.* 7(8):R69.
Tischler J, Lehner B, Fraser AG. 2008. Evolutionary plasticity of genetic interaction networks. *Nat Genet.* 40(4):390–391.
Tong AH, Evangelista M, Parsons AB, Xu H, Bader GD, Page N, Robinson M, Raghibizadeh S, Hogue CW, Bussey H, et al. 2001. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* 294(5550):2364–2368.
van Leeuwen J, Pons C, Mellor JC, Yamaguchi TN, Friesen H, Koschwanez J, Usaj MM, Pechlaner M, Takar M, Usaj M, et al. 2016. Exploring genetic suppression interactions on a global scale. *Science* 354(6312):aag0839.
VanderSluis B, Bellay J, Musso G, Costanzo M, Papp B, Vizeacoumar FJ, Baryshnikova A, Andrews B, Boone C, Myers CL. 2010. Genetic interactions reveal the evolutionary trajectories of duplicate genes. *Mol Syst Biol.* 6:429.
Vavouri T, Semple JI, Lehner B. 2008. Widespread conservation of genetic redundancy during a billion years of eukaryotic evolution. *Trends Genet.* 24(10):485–488.
Wagner A. 2005. Energy constraints on the evolution of gene expression. *Mol Biol Evol.* 22(6):1365–1374.
Williams EJ, Bowles DJ. 2004. Coexpression of neighboring genes in the genome of *Arabidopsis thaliana*. *Genome Res.* 14(6):1060–1067.
Wong S, Wolfe KH. 2005. Birth of a metabolic gene cluster in yeast by adaptive gene relocation. *Nat Genet.* 37(7):777–782.
Wu CI, Ting CT. 2004. Genes and speciation. *Nat Rev Genet.* 5(2):114–122.
Yeaman S. 2013. Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proc Natl Acad Sci U S A.* 110(19):E1743–E1751. 17
Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S. 2010. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* 26(7):976–978.
Zaslaver A, Baugh LR, Sternberg PW. 2011. Metazoan operons accelerate recovery from growth-arrested states. *Cell* 145(6):981–992.