Molecular Biology and Evolution

# Polymorphism, Divergence, and the Role of Recombination in *Saccharomyces cerevisiae* Genome Evolution

Asher D. Cutter*[,1,2] and Alan M. Moses[2,3]

[1]Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, ON, Canada

[2]Centre for the Analysis of Genome Evolution and Function, University of Toronto, Toronto, ON, Canada

[3]Department of Cell and Systems Biology, University of Toronto, Toronto, ON, Canada

***Corresponding author**: E-mail: asher.cutter@utoronto.ca.

**Associate editor:** Matthew Hahn

## Abstract

A contentious issue in molecular evolution and population genetics concerns the roles of recombination as a facilitator of natural selection and as a potential source of mutational input into genomes. The budding yeast *Saccharomyces cerevisiae*, in particular, has injected both insights and confusion into this topic, as an early system subject to genomic analysis with subsequent conflicting reports. Here, we revisit the role of recombination in mutation and selection with recent genome-wide maps of population polymorphism and recombination for *S. cerevisiae*. We confirm that recombination-associated mutation does not leave a genomic signature in yeast and conclude that a previously observed, enigmatic, negative recombination–divergence correlation is largely a consequence of weak selection and other genomic covariates. We also corroborate the presence of biased gene conversion from patterns of polymorphism. Moreover, we identify significant positive relations between recombination and population polymorphism at putatively neutrally evolving sites, independent of other factors and the genomic scale of interrogation. We conclude that widespread natural selection across the yeast genome has left its imprint on segregating genetic variation, but that this signature is much weaker than in *Drosophila* and *Caenorhabditis*.

**Key words:** genetic hitchhiking, background selection, recombination rate, mutagenic recombination, budding yeast, nucleotide polymorphism.

Research article

## Introduction

Determining the relative contributions of recombination as a force facilitating efficient natural selection and as a source of mutational heterogeneity within genomes forms a central problem in genome biology. Theory predicts that linkage among simultaneously selected sites will interfere with the ability of selection to fix beneficial mutations or to drive extinct detrimental mutations (Hill and Robertson 1966; Felsenstein 1974; Birky and Walsh 1988; Barton 1995). Several species provide evidence supporting enhanced selection in genomic regions subject to more recombination (Kliman et al. 2003; Presgraves 2005; Connallon and Knowles 2007; Shapiro et al. 2007; Larracuente et al. 2008; Cutter and Choi 2010). The interaction between recombination and positive selection also is predicted to reduce neutral polymorphism in genomic regions that experience low rates of recombination due to genetic hitchhiking of neutral polymorphisms linked to selected sites (Maynard Smith and Haigh 1974; Andolfatto 2001). Indeed, crossover rates commonly do correlate positively with genetic diversity within genomes, which often is interpreted as evidence of recurrent selective sweeps (Hahn 2008; Sella et al. 2009). Background selection against deleterious mutations provides another selective explanation for reduced diversity in low-recombination regions (Charlesworth et al. 1993; Charlesworth 1994), which also has been invoked to explain recombination-associated heterogeneity in patterns of population polymorphism (Hudson and Kaplan 1995; Charlesworth 1996; Jensen et al. 2002; Innan and Stephan 2003; Kaiser and Charlesworth 2009; Cutter and Choi 2010). Background selection is predicted to be particularly potent in asexually-reproducing species (Charlesworth et al. 1993; Charlesworth 1994).

Importantly, higher rates of mutation in high-recombination regions of the genome also could yield a positive correlation between polymorphism and recombination across the genome (Lercher and Hurst 2002). Although recombination itself need not be the ultimate cause of the mutagenicity, experimental studies with mitotically dividing cells in *Saccharomyces cerevisiae* implicate the recombination process itself as being mutagenic (Magni 1964; Strathern et al. 1995; Holbeck and Strathern 1997). More typically, between-species divergence at putatively neutrally-evolving sites is used to test for recombination-associated mutation (RAM) because differences in divergence in this class of sites should only reflect long-term differences in the mutation rate (Kimura 1968; Birky and Walsh 1988). In this way, RAM has been ruled out as a major force in the cases of *Drosophila melanogaster*, *D. simulans*, and *Caenorhabditis briggsae* (Begun and Aquadro 1992; Begun et al. 2007; Cutter and Choi 2010), but its potential to contribute to mutational heterogeneity in the genomes of several other species, including humans, remains controversial (Lercher and Hurst 2002; Hellmann et al. 2003; Filatov 2004;

© The Author 2011. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com



http://mbe.oxfordjournals.org/

Spencer et al. 2006; Hellmann et al. 2008; Kulathinal et al. 2008; Noor 2008a; Cai et al. 2009; McVicker et al. 2009).

Recently, Noor (2008b) concluded that interspecific divergence data in *Saccharomyces* also are not consistent with RAM. And yet, he observed that divergence between homologous *S. cerevisiae* and *S. paradoxus* introns and intergenic sequences was lower for sequences in regions of the genome that experience higher rates of recombination—opposite to the expectation for RAM—an enigmatic result with no predicted cause at neutral sites. This pattern also was reported previously for synonymous sites in coding sequence (Connallon and Knowles 2007) but opposite to the findings of Doniger et al. (2008). A possible complication with these analyses is that the compact genomes and large population sizes of these species may render few sites selectively neutral, in which case more efficient translational selection and purifying selection on noncoding DNA in high-recombination regions might generate a negative relation between divergence and recombination. Analysis of the yeast genome suggests that purifying selection may indeed be more effective in regions of higher recombination (Kliman et al. 2003; Connallon and Knowles 2007). The variant frequency spectrum in yeast also indicates that the strength of purifying selection on noncoding sequence lies intermediate between the stronger selection on replacement coding sites and weaker selection on synonymous coding sites (Liti et al. 2009).

Here, we reconsider the potential for recombination to mediate heterogeneity in polymorphism and divergence in the yeast genome. Taking advantage of whole-genome population polymorphism data for *S. cerevisiae* (Liti et al. 2009) and recent dense maps of *S. cerevisiae* crossover and noncrossover (gene conversion) recombination (Mancera et al. 2008), we test for associations between recombination, polymorphism, and divergence while accounting for ancillary factors and considering different scales of heterogeneity. We find that divergence at the most-neutral sites in the yeast genome (synonymous sites in genes with low codon bias) does not markedly correlate with recombination rates independently of other covariates, across a broad range of scales. In contrast, divergence at synonymous sites considered for all genes regardless of codon bias correlates more strongly and negatively with recombination rate, implying that high-recombination regions enable more effective weak purifying selection on synonymous sites (Kliman et al. 2003). We also demonstrate consistent, but quantitatively weak, positive associations between diversity at putatively neutral sites and recombination, independently of other covarying genomic features. We infer that selection drives recombination rate-dependent patterns of polymorphism across the *S. cerevisiae* genome.

## Methods

We computed per-gene measures of divergence at replacement sites ($d_N$) and synonymous sites ($d_S$) of genes, coding length, codon bias (codon adaptation index [CAI]), G + C content at 4-fold degenerate sites, and single nucleotide polymorphism (SNP) diversity at replacement sites ($S_N$) and synonymous sites ($S_S$). Divergence measures are the *S. cerevisiae* lineage-specific values for orthologous groups of 3–4 yeast species (*S. paradoxus* as well as *S. mikatae* and/or *S. bayanus*), using protein-based nucleotide alignments as calculated in the codeml program of PAML (Yang 1997) permitting one $d_N/d_S$ ratio for the *S. cerevisiae* branch and a second ratio for the rest of the tree. This yielded estimates for 5,204 genes. Loci with $d_S > 1$ on the *S. cerevisiae* lineage were excluded from further analysis, as they likely result from errors in ortholog assignment. We then tallied or averaged these per-gene statistics within nonoverlapping windows of the genome of different interval size (5, 10, 20, 40, and 60 kb), computing them separately for all genes in a given region and for only those genes with CAI $\leq$ 0.24. The number of windows (sample size) ranged from 200 for the 60 kb scale to 2,354 for the 5 kb scale, after excluding 21 of the 4,750 windows across the five scales that had outlier measures of polymorphism. As appropriate, these per-window statistics were scaled by the number of replacement or synonymous sites in the window being considered. We refer to the average replacement-site divergence in these regions as $D_N$ and the putatively neutral divergence at synonymous sites for genes with CAI $\leq$ 0.24 as $D_{neu}$; divergence at synonymous sites that included all genes is $D_S$. Coding gene density was computed as the fraction of the interval comprised of coding sequence. SNPs were identified from the aligned sequence data and annotations from SGRP (Liti et al. 2009). Base calls were made at Q40, and a coverage of at least five individuals was required at any position for it to be considered. This yielded 144,164 SNPs in coding regions for use in calculations of polymorphism. SNP diversity was calculated as the number of segregating sites divided by the total number of replacement sites or synonymous sites in the window, using either the entire sample ($n = 40$) of Liti et al. (2009) or just the "wine-European" population subset of 11 samples that they identified. Because sample size varies across sites due to heterogeneity in sequence coverage, we used a diversity statistic ($Q_N$ and $Q_S$) that is analogous to Watterson's $\theta$ (Watterson 1975) as a metric of sequence variation, such that the number of segregating sites in a given nonoverlapping window was scaled by $\log[n-1]$, where $n$ is the average sample size across polymorphic sites. We refer to putatively neutral polymorphism, computed only for synonymous sites from genes with CAI $\leq$ 0.24, as $Q_{neu}$. Similarly, $GC4_{neu}$ represents the G + C content at 4-fold degenerate sites for genes with CAI $\leq$ 0.24. We arrived at this CAI = 0.24 threshold as the highest value of CAI with a nonsignificant correlation between CAI and $d_S$ (see Results; for genes with CAI $\leq$ 0.25, Pearson's $r = -0.03$, $P = 0.038$).

We also calculated the average frequency of derived variants for replacement sites, synonymous sites, and putatively neutral sites. Ancestral state polarization was inferred via majority rule of nucleotide state relative to the 2–3 outgroup species mentioned above for divergence calculations. We then calculated for each gene the average frequency of alleles with a derived state in *S. cerevisiae* for replacement and synonymous sites as well as for the 11 "wine" strains or "all" 40 strains; note that the realized sample size of strains can differ

for different sites within a gene. These per-gene estimates were then averaged across genes in a given genomic interval (5–60 kb) as an index of the derived variant frequency in that interval ($F_{D-N}$, $F_{D-S}$, $F_{D-neu}$ for replacement sites, synonymous sites, and putatively neutral sites, respectively). In some cases, genes overlapped an interval boundary and were included in calculations for both intervals. The medians of our derived variant frequency metric for replacement sites ($F_{D-N} \cong 0.17$) is about half that for putatively neutral sites ($F_{D-neu} \cong 0.30$), across all scales, consistent with previous assessments of the derived allele frequency distribution in *S. cerevisiae* (Liti et al. 2009; Schacherer et al. 2009). For the putatively neutral sites, we also computed separate measures for the frequency of derived variants that have G or C nucleotides and for the derived A/T variants ($F_{Dg/c-neu}$, $F_{Da/t-neu}$).

Using the map of crossover and noncrossover recombination detailed by Mancera et al. (2008) for the yeast genome, we computed recombination frequency in the same nonoverlapping windows used for measuring gene density, codon bias, and sequence polymorphism (scales of 5, 10, 20, 40, and 60 kb). In a given window, the number of crossover and noncrossover recombination events was tallied from Mancera et al.'s (2008) supplemental file "event_intervals.txt" to compute the recombination rate as the number of crossovers per kb per meiosis from fine to coarse scales of recombination (similarly for noncrossovers or total recombination, the sum of crossover and noncrossover events). Parametric analyses were conducted on $\log_{10}$-transformed variables for cases where this improved conformity with the assumption of normality (CAI, $D_N$, $Q_N$, $Q_{neu}$, recombination rates).

## Results

### Identifying a Set of "Neutral" Sites in the Yeast Genome

To investigate the influence of recombination on divergence and polymorphism, it is critical to evaluate nucleotide sites with evolutionary dynamics as close to neutrality as possible. A challenge for the compact genome of *S. cerevisiae* is identifying sites free of selective constraint or functional importance—even in synonymous sites and noncoding sequence. Genes with strong codon bias have reduced synonymous site divergence ($d_S$) in yeast and other taxa as a consequence of translational selection (Bennetzen and Hall 1982; Ikemura 1985; Coghlan and Wolfe 2000; Kliman et al. 2003; Hirsh et al. 2005), and a substantial portion of intergenic sequence is composed of functionally significant regulatory elements (Kellis et al. 2003; Chin et al. 2005; Doniger et al. 2005; Liti et al. 2009). Here, we consider synonymous sites within genes that exhibit little codon usage bias (CAI $\leq$ 0.24) to represent the best candidates for sites evolving in a neutral fashion. Indeed, the correlation between $d_S$ and CAI is undetectable in this set of 4,517 genes (Pearson's $r = -0.02$, $P = 0.14$) compared with the remaining strongly codon-biased set of 554 genes ($r = -0.63$, $P < 0.0001$), indicating that synonymous sites in genes with low codon bias experience little if any direct selection for translational efficiency and/or accuracy (fig. 1). Hereafter, we refer to such synonymous sites in low codon bias genes as neutral sites, for simplicity, recognizing that there may be additional sources of direct selection that remain unaccounted (e.g., regulatory and/or splicing enhancer regions or nucleosome positioning; Parmley et al. 2006; Warnecke et al. 2008).

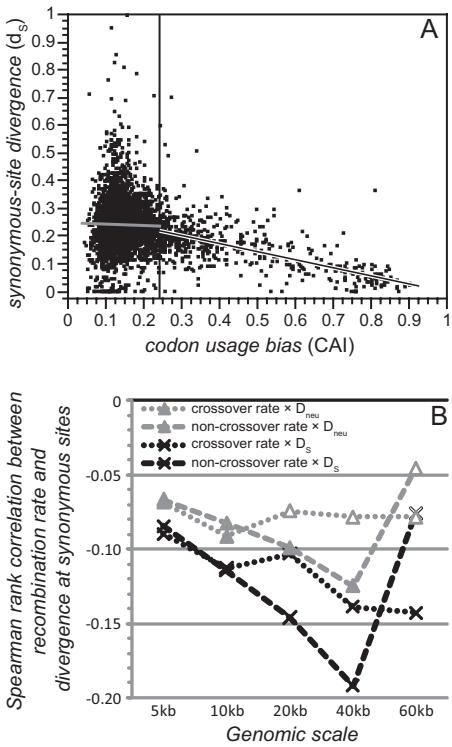


FIG. 1. (*A*) Divergence at synonymous sites ($d_S$) correlates strongly with codon usage bias only for genes with substantial codon usage bias (CAI > 0.24). Data shown for 5,071 genes with *Saccharomyces cerevisiae* lineage-specific divergence estimates $\leq$ 1. Least squares regression $r^2_{adj} = 0.00026$, $P = 0.14$ for $0 \leq$ CAI $\leq 0.24$; $r^2_{adj} = 0.40$, $P < 0.0001$ for CAI > 0.24. In (*B*), pairwise correlations between recombination rate and divergence are weaker when genes with strong codon bias ($D_{neu}$) are compared with synonymous site divergence for all genes ($D_S$). Solid symbols indicate a significant independent effect ($P < 0.05$); empty symbols nonsignificant.

### Testing for RAM

Given this improved designation of putatively neutral sites in the yeast genome, we attempted to resolve whether RAM leaves a general signature across the genome. First, we tested for a positive relationship between rates of recombination and lineage-specific rates of divergence at these most-neutral of sites ($D_{neu}$) across scales from 5 to 60 kb intervals. Importantly, we calculated all genomic variables at the same scale in any given analysis. Note that smaller scales have more nonoverlapping intervals although there is more heterogeneity among intervals. Using nonparametric bivariate correlation, we identified significant—but quantitatively weak—negative correlations between $D_{neu}$ and rates of crossover recombination at small to moderate scales (e.g., Spearman's $\rho = -0.07$ at 5 kb scale, $P = 0.0061$; $\rho = -0.09$ at 10 kb scale, $P = 0.0025$; figs. 1*B* and 2*C*). These findings are consistent with

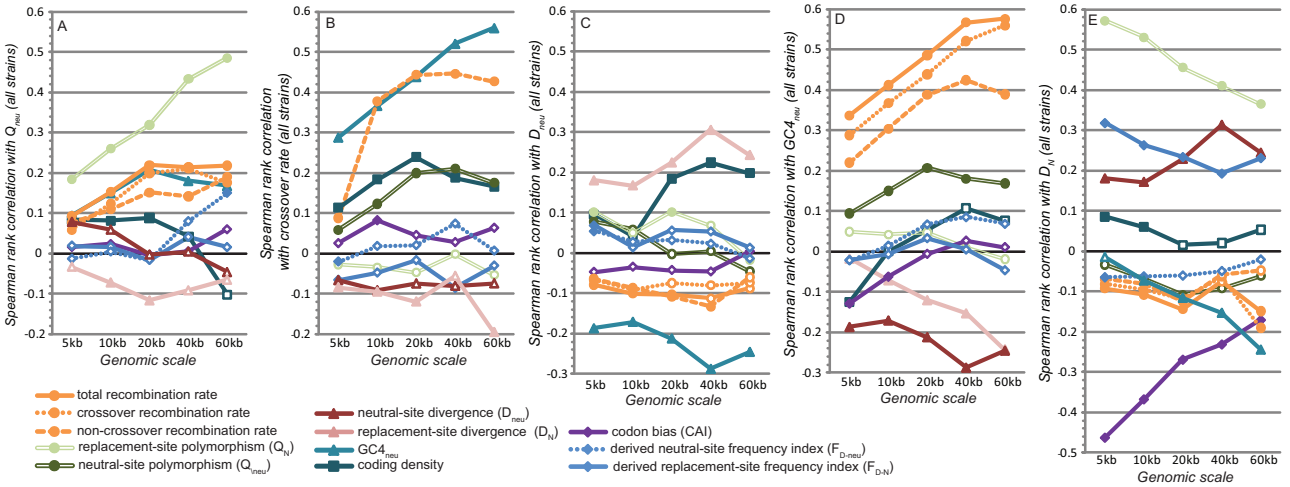


**Fig. 2.** Summary of the pairwise nonparametric correlations of genomic features with (*A*) neutral polymorphism ($Q_{neu}$), (*B*) crossover rate, (*C*) neutral divergence ($D_{neu}$), (*D*) neutral site G + C content ($GC4_{neu}$), and (*E*) replacement site divergence ($D_N$). Population genetic metrics correspond to data from all 40 strains of *Saccharomyces cerevisiae*. Solid symbols indicate a significant effect ($P < 0.05$); empty symbols nonsignificant.

observations from two previous studies (Connallon and Knowles 2007; Noor 2008b) and opposite to the predictions of RAM. The correlation is notably weaker with $D_{neu}$ than with $D_S$ (which includes all synonymous sites, including genes with strong codon bias) (fig. 1*B*), indicative of the effects of direct selection on synonymous sites. Spencer et al. (2006) predicted that RAM should be strongest at small scales, yet the magnitude of the divergence–recombination correlation is especially weak at small scales (fig. 1*B*). Noncrossover recombination (i.e., gene conversion) yielded negative correlations with $D_{neu}$ of similar magnitude as for crossover recombination but significantly so up to the 40 kb scale (figs. 1B and 2*C*). We then performed a multivariate analysis to account for other covarying factors (G + C at putatively neutral sites, codon bias, and the fraction coding sequence in the interval). In these multiple regression models, we detected no significant independent association between recombination rate and $D_{neu}$ at any scale (supplementary table S1, Supplementary Material online). Therefore, we conclude that there is neither a widespread strong signature of RAM nor an enigmatic pattern of divergence in the *S. cerevisiae* genome, after accounting for other covarying factors.

## Can Selection at Linked Sites Account for Variation in Neutral Polymorphism?

To test how genomic features contribute to heterogeneity in levels of neutral polymorphism ($Q_{neu}$) across the yeast genome, we first performed pairwise correlations. Specifically, we tested whether local rates of recombination, coding gene density, and divergence at neutral sites (as a metric of mutation rate) associate with observed levels of neutral polymorphism across five scales. Again, these factors were all calculated at the same scale of estimation for any given analysis. We also considered separately polymorphism for either the subset of 11 wine yeast strains that plausibly correspond to a single population or all 40 yeast strains combined (Liti et al. 2009). We found that nonparametric correlations show a consistently positive association between neutral polymorphism and recombination rate—this pattern holds for both the wine and all strains for nearly all scales of measurement from 5 to 60 kb intervals (figs. 2*A* and 3; supplementary table S2, Supplementary Material online). The magnitudes of correlation are greater for larger scales, consistent with *D. persimilis* and humans (Spencer et al. 2006; Stevison and Noor 2010). This diversity–recombination relationship is consistent with a process of selection acting at linked sites. In contrast to the scale-independent significance of recombination rate, measures of neutral site divergence correlated significantly with polymorphism only at the smallest scales and with relatively weak magnitude of effect (fig. 2*A*). In addition, polymorphism at replacement sites correlated positively with neutral polymorphism at all scales, being most pronounced at larger scales of measurement. Small to moderate scales show positive associations between $Q_{neu}$ and coding density, and moderate genomic scales (10 kb or 20 kb intervals) exhibited significant negative correlations between $Q_{neu}$ and replacement site divergence ($D_N$; fig. 2*A*). The frequency of derived variants at replacement sites ($F_{D-N}$) positively correlated with neutral polymorphism at most scales for the wine strains only, whereas $F_{D-neu}$ correlated positively with neutral polymorphism at the largest scale for the set of all 40 strains (fig. 2*A*). G + C content at putatively neutral sites exhibits strong positive correlations with both crossover and noncrossover recombination rates, as well as $Q_{neu}$, but correlates negatively with $D_{neu}$ (fig. 2).

With the provocative bivariate correlations between neutral polymorphism and recombination rate in mind (figs. 2*A* and 3), which are indicative of a genomewide signature of selection at linked sites, we sought to identify significant independent effects of genomic features on neutral polymorphism using multivariate analysis. We constructed multiple regression models separately for wine and all yeast strains to explain heterogeneity in neutral polymorphism that included the factors mentioned above,

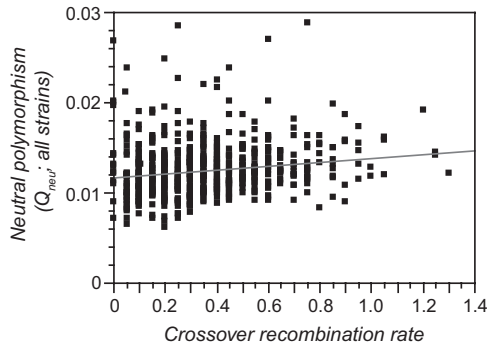


FIG. 3. Crossover recombination rate explains variation in nucleotide polymorphism at putatively neutral sites. Data shown for 20 kb interval genome scale for all 40 strains. Four points with $Q_{neu} > 0.03$ excluded. Gray line indicates least-squares regression line ($F_{1,577} = 16.59$, $P < 0.0001$); multivariate analysis supports this bivariate pattern (supplementary table S3, Supplementary Material online).

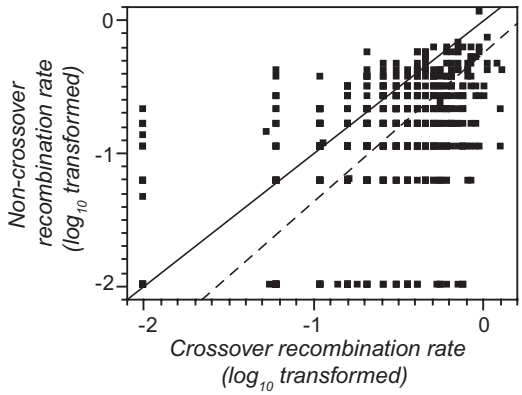


FIG. 4. Crossover and noncrossover recombination rates are highly correlated. Data shown for 20 kb interval genome scale. Recombination rate is the number of crossover or noncrossover events per kb per meiosis. Solid line is the line of unity; dashed line is an orthogonal regression fit on log-transformed values (0.01 added to values before $\log_{10}$-transform; slope not significantly different from 1; correlation = 0.44).

as well as their first-order interactions, and chromosome of origin as a covariate. Despite the covariation between crossover and noncrossover recombination (fig. 4), we tested for independent effects of these alternative modes of recombination. In the full sample of 40 strains, this procedure revealed patterns qualitatively consistent with the bivariate analysis. We found that a significant portion of the variation in neutral polymorphism across the genome is explained independently by crossover recombination rate (5–20 kb scales) and replacement site polymorphism (5–60 kb scales), consistent with pairwise correlations (fig. 5A; supplementary table S3, Supplementary Material online). The magnitude of effect for replacement site polymorphism was consistently strongest (explaining 10–18% of the variation), whereas recombination rate typically explained just $\leq$1% of the heterogeneity in neutral polymorphism independently of other factors. Noncrossover recombination rate was a significant positive factor at the smallest (5 kb) and largest (60 kb) scales, but explaining a similarly small amount of the variation in $Q_{neu}$ as crossover rate. As well, replacement site divergence explained a significant fraction of the variation across all scales (3–7% variance explained), with higher rates of protein-sequence evolution being associated with lower neutral polymorphism. We also detected a significantly positive effect of neutral divergence at most scales ($\leq$1.7%), implying that regions with higher mutations rates contained more polymorphism. $F_{D-neu}$, G + C content at neutral sites, and coding density positively correlated with $Q_{neu}$ at some scales (fig. 5A; supplementary table S3, Supplementary Material online). The 16 chromosomes also differed significantly in levels of neutral polymorphism at nearly all scales, explaining 2–6% of the variation (fig. 5A; supplementary table S3, Supplementary Material online); a variety of interaction terms also exerted significant effects (see supplementary tables, Supplementary Material online). Patterns of significant factors and the sign of effects for the 11 wine strains were largely similar to that of the full set of 40 strains. However, the smaller sample of wine strains revealed no significant independent effects of crossover or noncrossover recombination rate on $Q_{neu}$ (supplementary table S3, Supplementary Material online). Taken together, these results relating polymorphism and crossover rates are consistent with selection at linked sites reducing polymorphism in nearby neutral sites.

## Variant Frequencies Are not Associated with Recombination Rate

Widespread selective sweeps are expected to leave an excess of rare variants in genomic regions that experience little recombination (Braverman et al. 1995; Andolfatto 2001). However, we did not observe such a pattern in the yeast data: Neither crossover nor noncrossover recombination were generally predictive of derived variant frequencies in our multivariate analysis (fig. 2; supplementary table S4, Supplementary Material online). In fact, variation across the genome in the average frequency of derived neutral variants ($F_{D-neu}$) was not consistently and independently explained across scales by any factor, although coding density, codon bias, $D_{neu}$, $Q_N$, and $Q_{neu}$ were significant correlates for at least two scales in either the wine or all strains (supplementary table S4, Supplementary Material online). The average frequency of derived replacement site variants ($F_{D-N}$), however, consistently correlated positively with both $D_N$ and $F_{D-neu}$ across scales, independently of other factors (supplementary table S4, Supplementary Material online). This held for both the set of all yeast strains and for the subset of wine strains and suggests that regions with a history of elevated rates of protein sequence divergence might also experience ongoing selection favoring new replacement mutations. In addition, $F_{D-N}$ for wine strains was consistently positively associated with $Q_{neu}$ and negatively associated with $Q_N$ (supplementary table S4, Supplementary Material online).

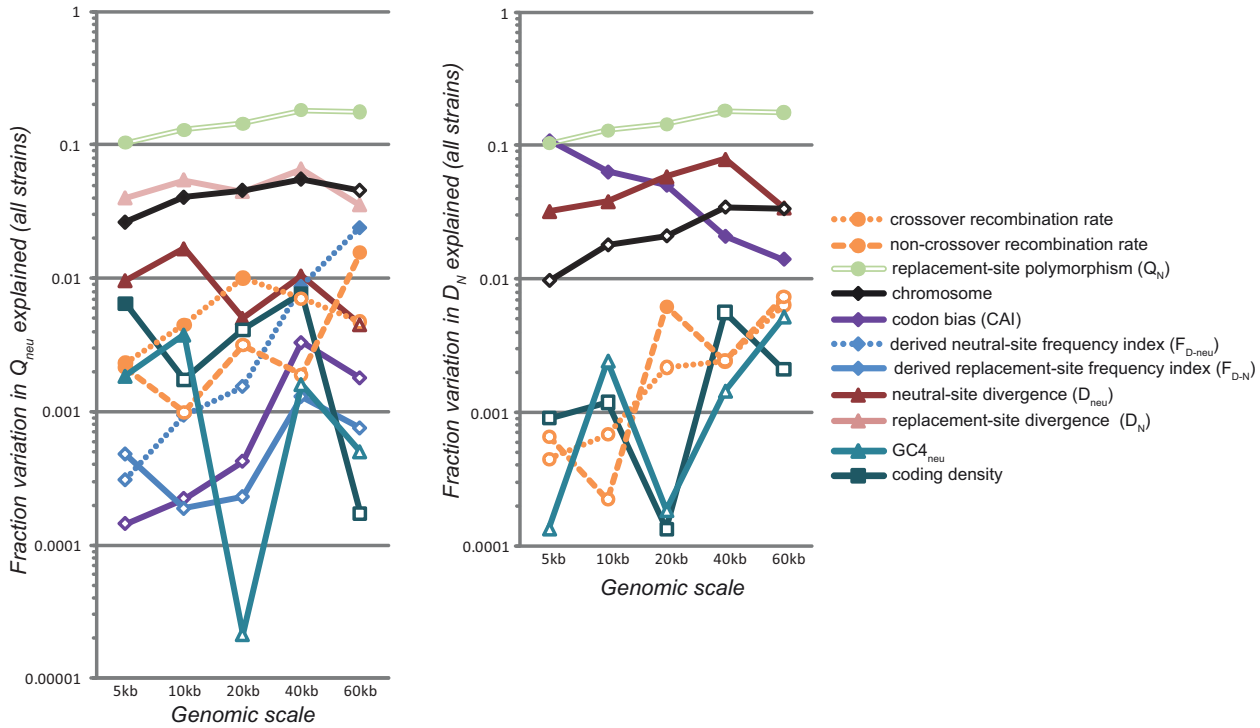


**Fig. 5.** Summary of multivariate analyses to explain variation in (*A*) neutral polymorphism ($Q_{neu}$) and in (*B*) replacement site divergence ($D_N$), across six scales of estimation. Solid symbols indicate a significant independent effect ($P < 0.05$); empty symbols nonsignificant. Population genetic metrics correspond to data from all 40 strains of *Saccharomyces cerevisiae*. Interaction terms that were included in the model are not shown; see also supplementary tables S1–S3, Supplementary Material online. Recombination rates, codon bias (CAI), polymorphism ($Q_{neu}$, $Q_N$), and replacement site divergence ($D_N$) were $\log_{10}$-transformed in analysis.

## Does Recombination Facilitate Protein Evolution?

Crossover recombination can play a role in facilitating selection by reducing interference between loci that are subject to selection simultaneously (Hill and Robertson 1966; Felsenstein 1974; Birky and Walsh 1988; Barton 1995). We tested for such an effect on protein sequence evolution in the yeast genome using both bivariate and multivariate approaches. Pairwise nonparametric correlation identified significant negative associations between $D_N$ and crossover rate at all scales except 40 kb (Spearman $\rho > |-0.08|$, $P \leq 0.007$; fig. 2*D*). The magnitude of these nonparametric bivariate correlations is comparable with previous reports (Pal et al. 2001a; Connallon and Knowles 2007; Weber and Hurst 2009), although our analyses were not conducted on a gene-by-gene basis and we did not partition genes according to "essentiality." However, a multiple regression model did not identify crossover rate as an independent explanatory variable for variation in $D_N$ at any scale. Thus, we do not verify some previous conclusions for yeast that recombination facilitates purifying selection on proteins (Connallon and Knowles 2007; Weber and Hurst 2009), consistent with an earlier finding (Pal et al. 2001a). CAI is a potent explanatory variable for rates of protein evolution (fig. 5*B*; supplementary table S1, Supplementary Material online), consistent with the previous finding of expression level exerting a strong effect on the evolution of amino acid sequences (Drummond et al. 2005). The positive $D_N$–$D_{neu}$ association that also persists in both bivariate and multivariate analysis (fig. 5*B*; supplementary table S1, Supplementary Material online) implies that the genome varies widely in mutation rate, independently of recombination rate, and that this mutational heterogeneity influences protein sequence divergence; this reasoning presumes that selection on replacement sites is uncoupled from any residual selection on the synonymous sites we defined as putatively neutral. Consistent with previous observations (Pal et al. 2001b; Kliman et al. 2003), however, we

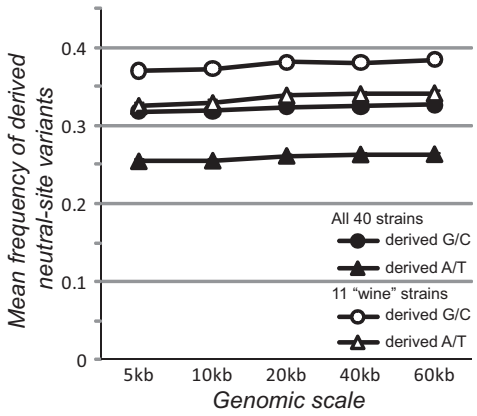


**Fig. 6.** The average frequency of derived G or C variants at neutral sites ($F_{Dg/c-neu}$) is higher than for A or T variants ($F_{Da/t-neu}$). The difference between $F_{Dg/c-neu}$ and $F_{Da/t-neu}$ holds for both the complete set of 40 strains and the subset of 11 wine yeast strains and across all scales of measurement (paired *t*-test, all $P < 0.0001$). Bars indicating standard error of the mean are hidden by the symbols.

saw that genomic heterogeneity in codon bias is partially explained by the effects of crossover recombination rate but only at the 10 kb scale (supplementary table S1, Supplementary Material online). Replacement site divergence provided a much stronger explanatory factor for CAI at most scales, and G + C content also contributed to variation in CAI at small scales (supplementary table S1, Supplementary Material online).

### Does Noncrossover Recombination Affect Patterns of Molecular Evolution?

Biased gene conversion favoring G or C variants should manifest as a positive correlation between neutral site G + C content ($GC4_{neu}$) and noncrossover recombination rates and also as derived G or C variants occurring at higher frequencies ($F_{Dg/c-neu}$) in populations than derived A or T variants ($F_{Da/t-neu}$) (Lercher et al. 2002; Marais 2003). We find support for both predictions: $F_{Dg/c-neu}$ is greater than $F_{Da/t-neu}$ at all scales (paired *t*-test, all $P < 0.0001$; fig. 6) and $GC4_{neu}$ correlates positively with recombination rates (fig. 2*D*). Biased gene conversion also predicts that G + C content evolution is primarily a by-product of noncrossover rather than crossover recombination and that its effect on variant frequencies should be elevated in regions with more recombination. However, $GC4_{neu}$ correlates more strongly with rates of crossover than noncrossover recombination (fig. 2*D*), and the frequency of derived G or C variants is not consistently higher in genomic regions with more recombination (Spearman correlation $P > 0.1$ for all scales, for both crossover and noncrossover recombination, and for wine or all strains). The total recombination rate (crossovers and noncrossovers) is more strongly correlated with $GC4_{neu}$ than is each component separately (fig. 2*D*). As noted above, we found higher $Q_{neu}$ to associate with higher $GC4_{neu}$ at small genomic scales, independently of other factors (fig. 5). This suggests that biased gene conversion also contributes to observed patterns of polymorphism, although the magnitude of effect is small (fig. 5).

## Discussion

Understanding the roles of recombination in facilitating natural selection and as a potentially mutagenic source is crucial for constructing an integrated view of genome evolution and function. Previous analyses of the budding yeast *S. cerevisiae* genome, however, have produced some conflicting and counterintuitive results (Connallon and Knowles 2007; Doniger et al. 2008; Noor 2008b). Using genome-scale data on recombination, interspecific divergence, and intraspecific polymorphism, we reconsider and clarify this issue in *S. cerevisiae* with a multivariate approach that considers polymorphism estimated from population samples in the most-neutral of sites, using a range of consistent genomic scales of measurement.

Experimental studies in yeast have demonstrated that mitotic recombination can generate mutations (Strathern et al. 1995), which, if general to meiotic recombination as well, should manifest in greater sequence divergence at selectively neutral sites in regions of the genome that experience more recombination, by virtue of an elevated mutation rate in such regions. However, we find that sequence divergence at the most-neutral of sites is not elevated in regions of higher recombination across a range of scales from 5 to 60 kb. This indicates that any RAM does not manifest as consistent heterogeneity in divergence between species, provided that recombination rates themselves have not diverged radically since the common ancestor of *S. cerevisiae* with *S. paradoxus*. We observe the negative bivariate correlation between recombination and divergence reported previously (Connallon and Knowles 2007; Noor 2008b), although this pattern is weaker when genes with high codon bias are excluded. Moreover, the correlation disappears when other genomic features are taken into account. Consequently, we propose that the previously enigmatic finding of a negative recombination-divergence association (Connallon and Knowles 2007; Noor 2008b) largely reflects weak purifying selection on sites that were presumed to evolve neutrally in combination with covariation with other genomic characteristics. We therefore conclude that RAM may be excluded as an underlying cause of broad patterns of difference in nucleotide polymorphism between regions of high and low recombination in the yeast genome, which is consistent with inferences from a variety of other organisms (reviewed in Cutter and Choi 2010; Stevison and Noor 2010).

Recent work documents a weak role for recombination affecting nucleotide G + C composition change in yeast (Marsolier-Kergoat and Yeramian 2009; Harrison and Charlesworth 2011) despite clear evidence of biased gene conversion (Birdsell 2002; Mancera et al. 2008). We corroborate the signatures of biased gene conversion, by way of elevated frequencies of derived G or C variants relative to derived A or T variants, and by positive associations between G + C content at putatively neutral sites with rates of recombination (both crossover and noncrossover recombination). Curiously, however, higher rates of recombination do not seem to lead to higher derived G/C variant frequencies. Finally, multivariate analyses suggest that biased gene conversion might contribute to variation in nucleotide polymorphism at neutral sites, seen by the independent effect of G + C content on neutral polymorphism, but with only a small magnitude of effect.

The *S. cerevisiae* genome demonstrates positive associations between recombination rate and population sequence polymorphism at the most-neutral of sites. This result is robust across a broad range of scales and persists after accounting for covarying genomic factors. Given our evidence from sequence divergence for no consistent effect of recombination-related mutation across the genome, we conclude that the positive relation between recombination and population diversity likely results from selection at linked sites (recurrent genetic hitchhiking and/or background selection). However, the magnitude of effect in yeast appears much weaker than in some other organisms, such as species of *Drosophila* (Begun and Aquadro 1992; Kulathinal et al. 2008; Sella et al. 2009) and *Caenorhabditis*

(Cutter and Payseur 2003, Cutter and Choi 2010). Doniger et al. (2008) reported an even weaker bivariate correlation between polymorphism and recombination rate than we identify, although their study contrasted just three yeast strains and considered only a small scale (4.5 kb interval) that tends to exhibit a particularly weak effect (fig. 2). Another study in yeast actually reported a negative correlation between recombination rate and the incidence of intraspecific sequence differences (Noor 2008b), but this study was limited to a comparison of just two strains and might be subject to confounding with selection on synonymous sites and other genomic covariates (Noor 2008b). One potential reason for only a weak association between recombination rate and levels of polymorphism is selection being inconsistent over time and varying across populations, for which there is now some evidence in yeast (Elyashiv et al. 2010). In addition, despite relatively rare sexual reproduction in nature (Ruderfer et al. 2006), high crossover rates in *S. cerevisiae* (Mancera et al. 2008) may restrict the extent to which selection affects neighboring regions of the genome compared with other taxa. We quantified population genetic and molecular evolutionary parameters across a range of scales because there is little a priori expectation for what genomic scale might best reveal associations among the parameters. Given the high rates of recombination and small chromosomes in yeast compared with many other eukaryotes, it also is important to recognize that the most evolutionarily relevant scale may differ drastically among organisms (c.f. humans; Spencer et al. 2006).

It is difficult to determine whether recurrent selective sweeps or background selection against deleterious mutations might be most responsible for the recombination–polymorphism relationship in yeast. With the emerging acknowledgement of the importance of selective sweeps deriving from standing genetic variation (Pritchard et al. 2010), more sophisticated treatment is needed to infer the relative roles of these two forms of selection. For example, recurrent hitchhiking models generally assume "hard" sweeps of new beneficial mutations (Maynard Smith and Haigh 1974; Wiehe and Stephan 1993), yet recurrent "soft" sweeps on segregating variation lead largely to an expectation of near-independence between effective population size and neutral polymorphism (Gillespie 2001; Innan and Kim 2004; Przeworski et al. 2005; Pennings and Hermisson 2006). Selection at linked sites models operate in a manner analogous to a reduction in effective population size for genomic regions with low rates of recombination. If recurrent soft selective sweeps dominate the process of positive selection, with hard sweeps occurring with negligible frequency, then observations of positive correlations between neutral polymorphism and recombination rate might be best explained by background selection against deleterious mutations. The yeast genome reveals only a very weak recombination-polymorphism association; might adaptive evolution in yeast primarily involve soft sweeps? Subpopulation-specific selective forces might also figure into the answer to this question (Elyashiv et al. 2010).

How else do positive, directional selection and negative, purifying selection interact with recombination rate heterogeneity in the yeast genome? Unlike species of *Drosophila* (Sella et al. 2009), the yeast genome has not revealed widespread positive selection acting on coding sequences (Connallon and Knowles 2007; Doniger et al. 2008; Liti et al. 2009), although there is evidence of a history of positive selection on gene regulation (Fraser et al. 2010) (see also Mustonen et al. 2008; Mustonen and Lassig 2009). Genetic hitchhiking could lead to an excess of high frequency derived neutral variants in regions of low recombination (Braverman et al. 1995; Andolfatto 2001). However, we do not find statistical support for such a pattern. The presence of some population structure may limit the power to draw strong inferences from our derived variant frequency metrics (Liti et al. 2009; Elyashiv et al. 2010). Interestingly, a substantial portion of the variation in neutral polymorphism is explained by heterogeneity across the genome in regional rates of protein divergence: Regions with low neutral polymorphism have relatively high protein divergence. This robust association could represent a signature of positive selection, despite the lack of a consistent pattern for derived allele frequency metrics (which are difficult to estimate accurately)—although more formal McDonald–Kreitman tests (McDonald and Kreitman 1991) have not previously revealed much in the way of positive selection in coding regions in yeast (Liti et al. 2009). Our multivariate analysis also does not support a strong role for recombination in facilitating purifying selection over the long term (i.e., no independent negative association between recombination and $D_N$), despite the presence of bivariate patterns consistent with this interpretation that has been suggested previously (Pal et al. 2001a; Connallon and Knowles 2007). However, we do find some corroboration for recombination facilitating weak purifying selection on codon usage (c.f., Kliman et al. 2003): Synonymous site divergence is more strongly associated with recombination rate when genes having strong codon bias are included than when they are excluded, and recombination rate significantly explains variation in codon bias at the 10 kb scale, independently of other factors. Overall, we conclude that recombination enables a genome-wide influence of selection on patterns of diversity across the *S. cerevisiae* genome, but that its role is complicated by the correlated effects of other genomic characteristics and is weak compared with Drosophila and Caenorhabditis.

## Supplementary Material

Supplementary tables S1–S4 are available at *Molecular Biology and Evolution*online (http://www.mbe.oxfordjournals.org/).

## Acknowledgments

We thank A. Haudry, S. Wright and J. Fay for insightful discussion and the comments of two anonymous reviewers. A.D.C and A.M.M. are supported by Discovery Grants from the Natural Sciences and Engineering Research Council of

Canada and by infrastructure obtained using funds from the Canadian Foundation for Innovation.

## References


Andolfatto P. 2001. Adaptive hitchhiking effects on genome variability. *Curr Opin Genet Dev*. 11:635–641.

Barton NH. 1995. Linkage and the limits to natural selection. *Genetics* 140:821–841.

Begun DJ, Aquadro CF. 1992. Levels of naturally-occurring DNA polymorphism correlate with recombination rates in *Drosophila melanogaster*. *Nature* 356:519–520.

Begun DJ, Holloway AK, Stevens K, et al. (13 co-authors). 2007. Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol*. 5:e310.

Bennetzen JL, Hall BD. 1982. Codon selection in yeast. *J Biol Chem*. 257:3026–3031.

Birdsell JA. 2002. Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. *Mol Biol Evol*. 19:1181–1197.

Birky CW, Walsh JB. 1988. Effects of linkage on rates of molecular evolution. *Proc Natl Acad Sci U S A*. 85:6414–6418.

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. 1995. The hitchhiking effect on the site frequency-spectrum of DNA polymorphisms. *Genetics* 140:783–796.

Cai JJ, Macpherson JM, Sella G, Petrov DA. 2009. Pervasive hitchhiking at coding and regulatory sites in humans. *PLoS Genet*. 5:e1000336.

Charlesworth B. 1994. The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genet Res*. 63:213–227.

Charlesworth B. 1996. Background selection and patterns of genetic diversity in *Drosophila melanogaster*. *Genet Res*. 68:131–149.

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* 134:1289–1303.

Chin C-S, Chuang JH, Li H. 2005. Genome-wide regulatory complexity in yeast promoters: separation of functionally conserved and neutral sequence. *Genome Res*. 15:205–213.

Coghlan A, Wolfe KH. 2000. Relationship of codon bias to mRNA concentration and protein length in Saccharomyces cerevisiae. *Yeast* 16:1131–1145.

Connallon T, Knowles LL. 2007. Recombination rate and protein evolution in yeast. *BMC Evol Biol*. 7:235.

Cutter AD, Choi JY. 2010. Natural selection shapes nucleotide polymorphism across the genome of the nematode *Caenorhabditis briggsae*. *Genome Res*. 20:1103–1111.

Cutter AD, Payseur BA. 2003. Selection at linked sites in the partial selfer *Caenorhabditis elegans*. *Mol Biol Evol*. 20:665–673.

Doniger SW, Huh J, Fay JC. 2005. Identification of functional transcription factor binding sites using closely related Saccharomyces species. *Genome Res*. 15:701–709.

Doniger SW, Kim HS, Swain D, Corcuera D, Williams M, Yang SP, Fay JC. 2008. A catalog of neutral and deleterious polymorphism in yeast. *PLoS Genet*. 4:e1000183.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci U S A*. 102:14338–14343.

Elyashiv E, Bullaughey K, Sattath S, Rinott Y, Przeworski M, Sella G. 2010. Shifts in the intensity of purifying selection: an analysis of genome-wide polymorphism data from two closely related yeast species. *Genome Res*. 20:1558–1573.

Felsenstein J. 1974. The evolutionary advantage to recombination. *Genetics* 78:737–756.

Filatov DA. 2004. A gradient of silent substitution rate in the human pseudoautosomal region. *Mol Biol Evol*. 21:410–417.

Fraser HB, Moses AM, Schadt EE. 2010. Evidence for widespread adaptive evolution of gene expression in budding yeast. *Proc Natl Acad Sci U S A*. 107:2977–2982.

Gillespie JH. 2001. Is the population size of a species relevant to its evolution? *Evolution* 55:2161–2169.

Hahn MW. 2008. Toward a selection theory of molecular evolution. *Evolution* 62:255–265.

Harrison RJ, Charlesworth B. Forthcoming 2011. Biased gene conversion affects patterns of codon usage and amino acid usage in the *Saccharomyces sensu* stricto group of yeasts. *Mol Biol Evol*. 28:117–129.

Hellmann I, Ebersberger I, Ptak SE, Paabo S, Przeworski M. 2003. A neutral explanation for the correlation of diversity with recombination rates in humans. *Am J Hum Genet*. 72:1527–1535.

Hellmann I, Mang Y, Gu Z, Li P, de la Vega FM, Clark AG, Nielsen R. 2008. Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. *Genome Res*. 18:1020–1029.

Hill WG, Robertson A. 1966. Effect of linkage on limits to artificial selection. *Genet Res*. 8:269–294.

Hirsh AE, Fraser HB, Wall DP. 2005. Adjusting for selection on synonymous sites in estimates of evolutionary distance. *Mol Biol Evol*. 22:174–177.

Holbeck SL, Strathern JN. 1997. A role for REV3 in mutagenesis during double-strand break repair in *Saccharomyces cerevisiae*. *Genetics* 147:1017–1024.

Hudson RR, Kaplan NL. 1995. Deleterious background selection with recombination. *Genetics* 141:1605–1617.

Ikemura T. 1985. Codon usage and transfer-RNA content in unicellular and multicellular organisms. *Mol Biol Evol*. 2:13–34.

Innan H, Kim Y. 2004. Pattern of polymorphism after strong artificial selection in a domestication event. *Proc Natl Acad Sci U S A*. 101:10667–10672.

Innan H, Stephan W. 2003. Distinguishing the hitchhiking and background selection models. *Genetics* 165:2307–2312.

Jensen MA, Charlesworth B, Kreitman M. 2002. Patterns of genetic variation at a chromosome 4 locus of *Drosophila melanogaster* and *D. simulans*. *Genetics* 160:493–507.

Kaiser VB, Charlesworth B. 2009. The effects of deleterious mutations on evolution in non-recombining genomes. *Trends Genet*. 25:9–12.

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423:241–254.

Kimura M. 1968. Evolutionary rate at molecular level. *Nature* 217:624–626.

Kliman RM, Irving N, Santiago M. 2003. Selection conflicts, gene expression, and codon usage trends in yeast. *J Mol Evol*. 57:98–109.

Kulathinal RJ, Bennett SM, Fitzpatrick CL, Noor MA. 2008. Fine-scale mapping of recombination rate in Drosophila refines its correlation to diversity and divergence. *Proc Natl Acad Sci U S A*. 105:10051–10056.

Larracuente AM, Sackton TB, Greenberg AJ, Wong A, Singh ND, Sturgill D, Zhang Y, Oliver B, Clark AG. 2008. Evolution of protein-coding genes in Drosophila. *Trends Genet*. 24:114–123.

Lercher MJ, Hurst LD. 2002. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet*. 18:337–340.

Lercher MJ, Smith NG, Eyre-Walker A, Hurst LD. 2002. The evolution of isochores: evidence from SNP frequency distributions. *Genetics* 162:1805–1810.

Liti G, Carter DM, Moses AM, et al. 2009. Population genomics of domestic and wild yeasts. *Nature* 458:337–341.

Magni GE. 1964. Origin and nature of spontaneous mutations in meiotic organisms. *J Cell Comp Physiol*. 64:165–171.

Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM. 2008. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature* 454:479–485.

Marais G. 2003. Biased gene conversion: implications for genome and sex evolution. *Trends Genet*. 19:330–338.

Marsolier-Kergoat M-C, Yeramian E. 2009. GC content and recombination: reassessing the causal effects for the *Saccharomyces cerevisiae* genome. *Genetics* 183:31–38.

Maynard Smith J, Haigh J. 1974. Hitch-hiking effect of a favorable gene. *Genet Res*. 23:23–35.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the *Adh* locus in Drosophila. *Nature* 351:652–654.

McVicker G, Gordon D, Davis C, Green P. 2009. Widespread genomic signatures of natural selection in Hominid evolution. *PLoS Genet*. 5:e1000471.

Mustonen V, Kinney J, Callan CG, Lassig M. 2008. Energy-dependent fitness: a quantitative model for the evolution of yeast transcription factor binding sites. *Proc Natl Acad Sci U S A*. 105:12376–12381.

Mustonen V, Lassig M. 2009. From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. *Trends Genet*. 25:111–119.

Noor MA. 2008a. Connecting recombination, nucleotide diversity and species divergence in Drosophila. *Fly* 2:1–2.

Noor MAF. 2008b. Mutagenesis from meiotic recombination is not a primary driver of sequence divergence between Saccharomyces species. *Mol Biol Evol*. 25:2439–2444.

Pal C, Papp B, Hurst LD. 2001a. Does the recombination rate affect the efficiency of purifying selection? The yeast genome provides a partial answer. *Mol Biol Evol*. 18:2323–2326.

Pal C, Papp B, Hurst LD. 2001b. Highly expressed genes in yeast evolve slowly. *Genetics* 158:927–931.

Parmley JL, Chamary JV, Hurst LD. 2006. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol Biol Evol*. 23:301–309.

Pennings PS, Hermisson J. 2006. Soft sweeps III: the signature of positive selection from recurrent mutation. *PLoS Genet*. 2:e186.

Presgraves DC. 2005. Recombination enhances protein adaptation in *Drosophila melanogaster*. *Curr Biol*. 15:1651–1656.

Pritchard JK, Pickrell JK, Coop G. 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr Biol*. 20:R208–R215.

Przeworski M, Coop G, Wall JD. 2005. The signature of positive selection on standing genetic variation. *Evolution* 59:2312–2323.

Ruderfer DM, Pratt SC, Seidel HS, Kruglyak L. 2006. Population genomic analysis of outcrossing and recombination in yeast. *Nat Genet*. 38:1077–1081.

Schacherer J, Shapiro JA, Ruderfer DM, Kruglyak L. 2009. Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. *Nature* 458:342–346.

Sella G, Petrov DA, Przeworski M, Andolfatto P. 2009. Pervasive natural delection in the Drosophila genome? *PLoS Genet*. 5:e10000495.

Shapiro JA, Huang W, Zhang C, et al. (12 co-authors). 2007. Adaptive genic evolution in the Drosophila genomes. *Proc Natl Acad Sci U S A*. 104:2271–2276.

Spencer CCA, Deloukas P, Hunt S, Mullikin J, Myers S, Silverman BW, Donnelly P, Bentley D, McVean G. 2006. The influence of recombination on human genetic diversity. *PLoS Genet*. 2:1375–1385.

Stevison L, Noor M. 2010. Genetic and evolutionary correlates of fine-scale recombination rate variation in *Drosophila persimilis*. *J Mol Evol*. 71:332–345.

Strathern JN, Shafer BK, McGill CB. 1995. DNA-synthesis errors associated with double-strand-break repair. *Genetics* 140:965–972.

Warnecke T, Batada NN, Hurst LD. 2008. The impact of the nucleosome code on protein-coding sequence evolution in yeast. *PLoS Genet*. 4:e1000250.

Watterson GA. 1975. Number of segregating sites in genetic models without recombination. *Theor Popul Biol*. 7:256–276.

Weber CC, Hurst LD. 2009. Protein rates of evolution are predicted by double-strand break events, independent of crossing-over rates. *Genome Biol Evol*. 1:340–349.

Wiehe THE, Stephan W. 1993. Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from *Drosophila melanogaster*. *Mol Biol Evol*. 10:842–854.

Yang ZH. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*. 13:555–556.