BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

Local synteny and codon usage contribute to asymmetric sequence divergence of *Saccharomyces cerevisiae* gene duplicates

Lijing Bu, Ulfar Bergthorsson and Vaishali Katju*

Abstract

Background: Duplicated genes frequently experience asymmetric rates of sequence evolution. Relaxed selective constraints and positive selection have both been invoked to explain the observation that one paralog within a gene-duplicate pair exhibits an accelerated rate of sequence evolution. In the majority of studies where asymmetric divergence has been established, there is no indication as to which gene copy, ancestral or derived, is evolving more rapidly. In this study we investigated the effect of local synteny (gene-neighborhood conservation) and codon usage on the sequence evolution of gene duplicates in the *S. cerevisiae* genome. We further distinguish the gene duplicates into those that originated from a whole-genome duplication (WGD) event (ohnologs) versus small-scale duplications (SSD) to determine if there exist any differences in their patterns of sequence evolution.

Results: For SSD pairs, the derived copy evolves faster than the ancestral copy. However, there is no relationship between rate asymmetry and synteny conservation (ancestral-like *versus* derived-like) in ohnologs. mRNA abundance and optimal codon usage as measured by the CAI is lower in the derived SSD copies relative to ancestral paralogs. Moreover, in the case of ohnologs, the faster-evolving copy has lower CAI and lowered expression.

Conclusions: Together, these results suggest that relaxation of selection for codon usage and gene expression contribute to rate asymmetry in the evolution of duplicated genes and that in SSD pairs, the relaxation of selection stems from the loss of ancestral regulatory information in the derived copy.

Background

The appearance of novel biochemical traits contributing to phenotypic diversity is inextricably linked with the constant input of new genetic fodder via gene and genome duplication. However, a mere duplication of an ancestral locus far from guarantees the origin of a novel gene product and the majority of gene duplicates end up being silenced following a brief evolutionary existence [1,2]. For those paralogs that emerge unscathed by deleterious mutations, the first clues as to how paralogs are able to forge an independent evolutionary trajectory may be provided by studying their patterns of expression divergence and relative rates of molecular evolution.

Early studies of DNA sequence divergence between paralogs suggested there was little or no difference between duplicate gene-copies in their rates of evolution [3-7]. These results were used to argue against the hypothesis proposed by Ohno that following gene duplication, one copy is under relaxed selection and begins to accumulate previously 'forbidden' mutations [2]. However, these analyses may have had limited power to detect differences in evolutionary rates, or rate asymmetry, because they analyzed old duplicates, while an increase in the evolutionary rate is easiest to detect in young gene duplicates [8]. Subsequent studies have demonstrated relatively large rate asymmetry between duplicate genes [9-13]. For instance, 20%-30% of paralogous gene in *Saccharomyces cerevisiae* displayed significant differences in evolutionary rate [11] and one or both paralog(s) exhibited accelerated evolution in 17% of the cases [12].

* Correspondence: vkatju@unm.edu
Department of Biology, University of New Mexico, Albuquerque, NM 87131, USA

The phrase "gene duplication" appears to imply that all functionally relevant features of an ancestral gene are duplicated and therefore the two resulting gene copies ought to be functionally equivalent. In fact, there may be numerous differences between the two "copies". The derived copy often does not retain the full regulatory element repertoire of the ancestral copy or has some structural or genomic location differences relative to the ancestral gene [8,14-16]. These differences suggest that the derived copy might be expected to evolve under divergent constraints relative to the progenitor gene, either due to relaxation of natural selection or due to selection for novel attributes. In the majority of studies where asymmetric divergence has been established, there is no indication as to which gene copy, ancestral or derived, is evolving more rapidly. 'Derived' and 'ancestral' in the context of this study refer to the location of the paralogs in the genome rather than function. Recently, a study of gene duplicates in the mouse genome found that relocated gene copies following duplication, and in particular retrotransposed copies, evolved faster than paralogs in their ancestral location [16]. Similarly, a study in four mammalian genomes found that genes that came to reside in a different location following gene duplication were more likely to display evidence of adaptive evolution relative to gene copies that did not relocate [17].

In the case of a new gene-copy originating from a small-scale duplication (SSD) event and relocating some genomic distance from the ancestral copy, the identity of the ancestral and derived copies can be established by conservation of synteny flanking the paralogs or chromosomal location in comparison to a single-copy ortholog in an outgroup genome [15,16]. Distinguishing the ancestral from the derived copy becomes problematic in the case of whole-genome duplication (WGD henceforth). For example, in the instance of a genome resulting from allopolyploidy where duplicate gene-copies result from hybridization rather than gene duplication, naming ancestral and derived genes has no biological relevance.

Here we examine paralogs with low synonymous divergence in the *S. cerevisiae* genome to determine if it is the derived copy that evolves faster than the ancestral copy following gene duplication. Most duplicates in yeast originated from a WGD event [12,18] and for reasons mentioned in the preceding paragraph, it is inappropriate to assign ancestral and derived status to gene copies in the same manner as duplicates arising from SSD events. Gene duplicates that were previously identified as resulting from the WGD event are henceforth referred to as 'ohnologs' and were analysed separately from those resulting from SSD events to test if these two pools of duplicated genes behaved differently with respect to their rates of molecular evolution.

Results

Greater conservation of synteny in ohnologs

We initially commenced the analysis with 43 pairs of ohnologs and 15 SSD-derived gene duplicate pairs. These only included gene pairs that could be unambiguously assigned a single ortholog in an outgroup genome and the identification of local synteny conservation. Despite massive gene loss and genomic rearrangements in the evolutionary period subsequent to the WGD event, ohnologs have more extensive tracts of synteny relative to SSD-originated gene duplicates (Table 1). For instance, the average total upstream and downstream number of syntenic genes in the flanking regions for ohnologs versus SSD pairs is 19.87 and 4.67, respectively. Additionally, Wilcoxon signed-ranks tests revealed no significant difference in the extent of syntenic tracts in the upstream and downstream flanking regions within each population of yeast paralogs (ohnologs and SSD pairs).

Rate of molecular evolution of ohnologs is decoupled from synteny conservation

Nine and zero of 43 ohnolog pairs displayed significant asymmetry based on Tajima's Relative Rate test (uncorrected for multiple comparisons) using DNA (Additional File 1, Table S1) and amino acid sequences (Additional File 2, Table S2), respectively. Of these nine pairs of ohnologs, the faster evolving copy was associated with less synteny conservation in seven instances. This would indicate that the rate of evolution for paralogs formed via polyploidization might be influenced by the degree of preserved synteny. However, a nonparametric rank correlation test testing for association between synteny (sum of upstream and downstream continuous synteny) and the number of unique nucleotide sites was non-significant (*Kendall's tau* = 0.0132; $p = 0.91$). Likewise, we found no significant association between synteny preservation and the number of unique sites at the amino acid level (*Kendall's tau* = 0.0086; $p = 0.94$).

Table 1 Averaged measures of synteny preservation for 43 pairs of ohnologs *versus* 15 SSD pairs in the *S. cerevisiae* genome

Synteny Measure	*Ohnologs*	*SSD pairs*	*p-value*
Upstream continuous	1.41	0.47	0.0002
Downstream continuous	1.50	0.20	< 0.0001
Upstream continuous + Downstream continuous	2.91	0.67	
Upstream total	10.08	3.00	< 0.0001
Downstream total	9.79	1.67	< 0.0001
Upstream total + Downstream total	19.87	4.67	

For all measures of synteny (upstream continuous, downstream continuous, upstream total, and downstream total), the extent of synteny preservation is significantly greater in ohnologs relative to SSD pairs based on Wilcoxon tests.

Derived gene copies originating from SSD events exhibit accelerated rates of molecular evolution

Seven of 15 SSD pairs showed significant asymmetry using a Tajima's Relative Rate test at the nucleotide and amino acid level, respectively (Additional File 3, Table S3 and Additional File 4, Table S4). Six of these seven SSD pairs exhibited rate asymmetry both at the nucleotide and amino acid level. In all seven instances of significant rate asymmetry between paralogs at the nucleotide level, the derived copy exhibited accelerated rates of molecular evolution. In six of the seven instances of significant rate asymmetry at the amino acid level, the derived copy was the faster-evolving paralog. A Wilcoxon signed-ranks test of all 15 SSD pairs showed that collectively, the derived copies tend to possess a greater number of unique sites, suggesting accelerated molecular evolution at the nucleotide level (T = -25.0; p = 0.024) as well as the amino acid level (T = -21.0; p = 0.029).

CAI Results

Codon adaptation index (CAI) is a measure of optimal codon usage and it is positively correlated with levels of gene expression [19]. Following gene or genome duplication, there may be a period of relaxed selection resulting in lower CAI. If relaxation of selection does not apply equally to both paralogs, we may observe greater reduction in the use of optimal codons and CAI in one of the paralogs. We tested for the degree of association between the difference in CAI values between the two paralogs and the degree of rate asymmetry at the nucleotide level (difference in unique sites between the two paralogs generated from the Tajima's Relative Rate test) for both pools of gene duplicates in the *S. cerevisiae* genome. For SSD pairs, the derived paralogs have a significantly lower CAI than the ancestral paralogs (*Wilcoxon signed-ranks test*: T = 39.5; p = 0.011). However, we did not find a significant association between nucleotide rate asymmetry and change in CAI (*Kendall's tau* = 0.226; p = 0.25) (Figure 1). That is, faster-evolving paralogs did not have lower CAI values than slowly-evolving paralogs for SSD pairs. In contrast, we find a strong negative correlation between rate asymmetry and a difference in CAI values among ohnologs (*Kendall's tau* = -0.453; p < 0.0001) (Figure 2). Here, the faster-evolving paralogs resulting from the whole genome duplication event also have lower optimal codon preference.

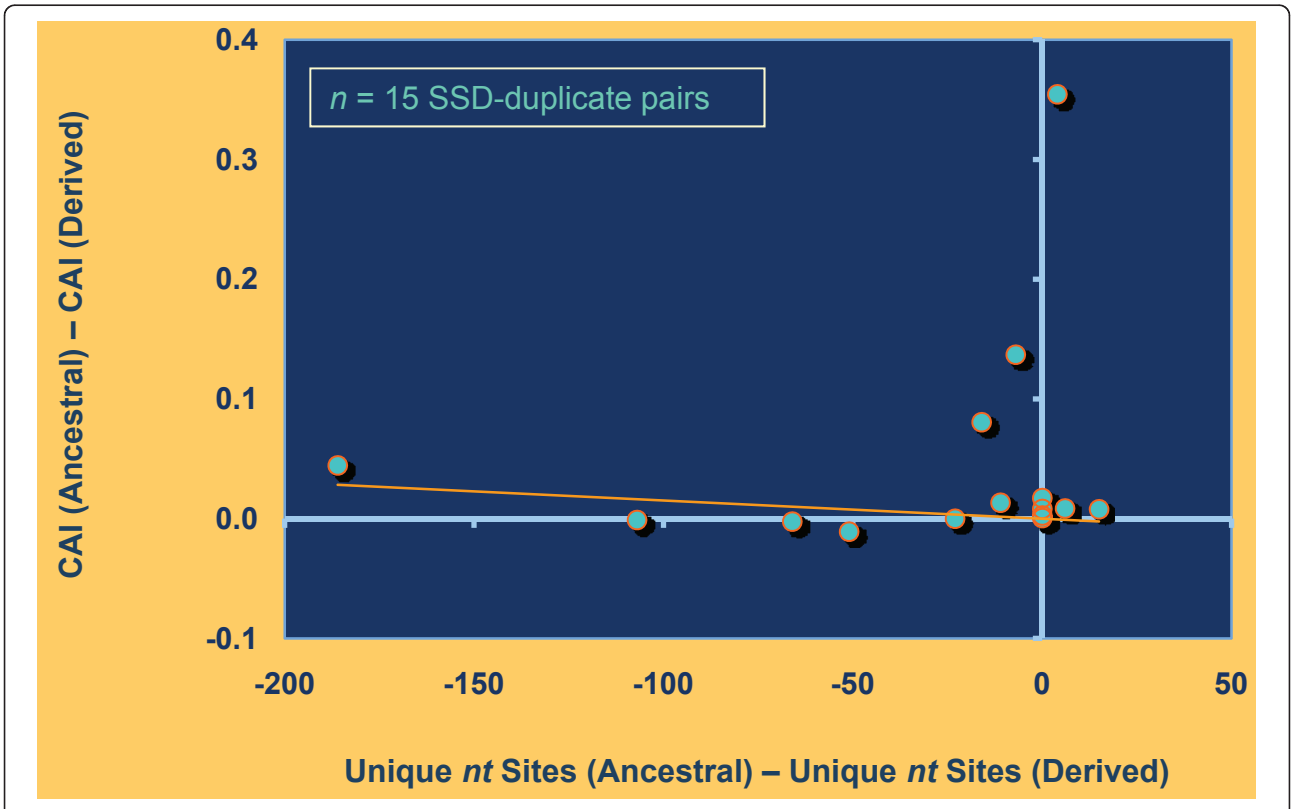

Figure 1 Nucleotide sequence asymmetry and codon adaptation index (CAI) for 15 SSD pairs in the *S. cerevisiae* genome. The sequence asymmetry measure on the *x* axis was calculated as the difference between unique nucleotide sites at the ancestral copy and the derived copy. The *y* axis represents the difference in CAI values between the ancestral copy and the derived copy for the same SSD pair. There was no significant association between differences in rate asymmetry and CAI values for SSD pairs (*Kendall's tau* = 0.226; p = 0.25).

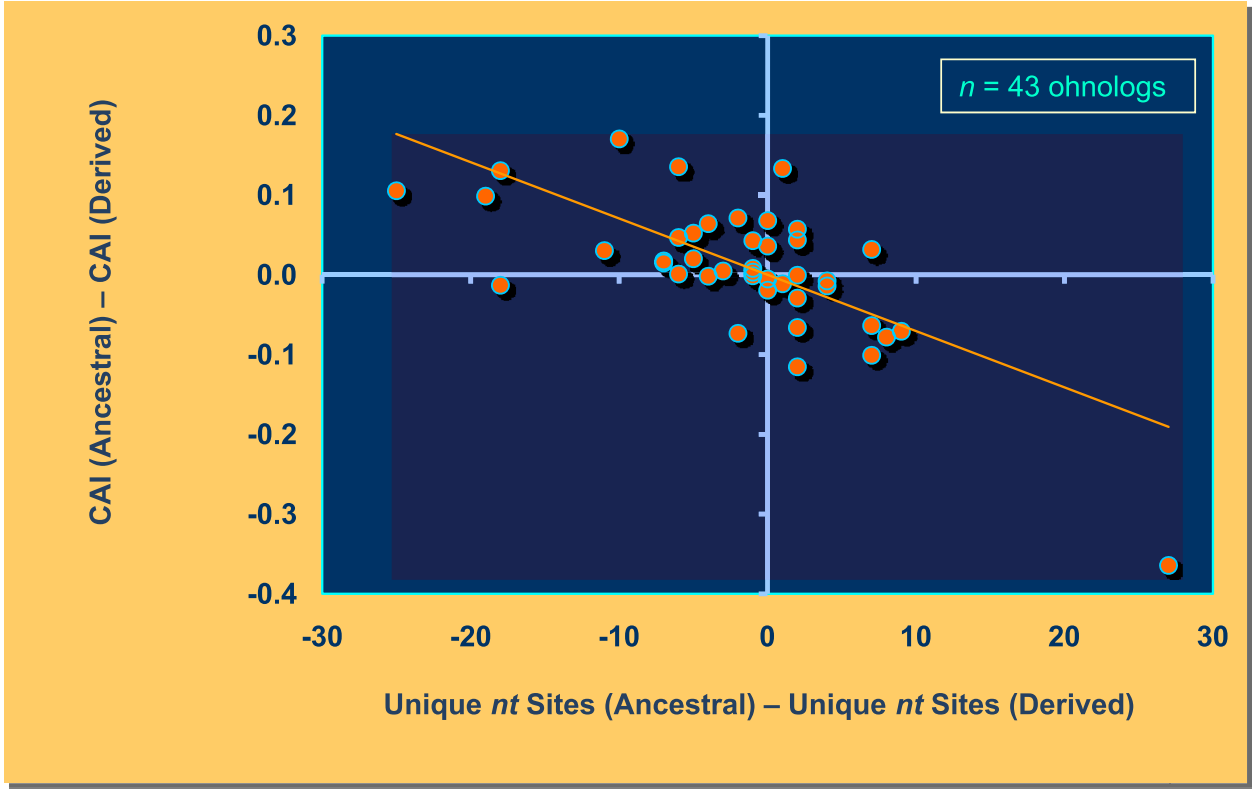

Figure 2 Negative relationship between nucleotide sequence asymmetry and codon adaptation index (CAI) for 43 pairs of ohnologs in the *S. cerevisiae* genome. The sequence asymmetry measure on the *x* axis was calculated as the difference between unique nucleotide sites at the ancestral-like copy and the derived-like copy within an ohnolog pair. The *y* axis represents the difference in CAI values between the ancestral-like copy and the derived-like copy for the same ohnolog pair. There was a significant negative correlation between differences in rate asymmetry and CIA values for ohnologs (*Kendall's tau* = -0.453; $p < 0.0001$).

Ohnologs and SSD duplicate pairs also differ with respect to their CAI values. The median CAI value for ohnologs and SSD pairs are 0.70 and 0.11, respectively. Indeed, CAI values averaged across both paralogs were determined to be significantly greater for ohnologs relative to SSD pairs (*Wilcoxon two-sample test*: Z = -4.723; $p < 0.0001$).

Faster-evolving paralogs have lower mRNA abundance

The preceding CAI results suggest that relaxed selective constraints due to reduced expression of the derived paralog may contribute significantly to rate asymmetry between ancestral and derived paralogs. We find that ancestral paralogs are expressed at significantly higher levels (greater mRNA abundance) than derived paralogs for SSD pairs (*Wilcoxon signed-ranks test*: $T = 37.5$; $p < 0.017$). In contrast, ancestral-like ohnologs with greater syntenic preservation do not differ significantly in their expression levels compared to derived-like ohnologs with lower syntenic preservation (*Wilcoxon signed-ranks test*: $T = 52$; $p = 0.54$).

We additionally tested if there is a relationship between transcription levels of paralogs and their degree of rate asymmetry at the nucleotide level. Figure 3 shows a significant correlation between the ratio of paralog-specific RNA and the ratio of unique sites in derived and ancestral copies of SSD pairs ($r = 0.87$, *Kendall's tau* = 0.74, $p < 0.0002$). Likewise, we find a significant association between the ratio of paralog-specific RNA and the ratio of unique sites in derived and ancestral copies for ohnologs ($r = 0.38$, *Kendall's tau* = 0.225, $p = 0.0343$).

Discussion

Duplicated genes frequently experience an initial increase in their rate of evolution and nonsynonymous substitutions relative to synonymous substitutions. Moreover, recent analyses of young gene duplicates in several eukaryotic genomes indicate that paralogs exhibit asymmetric rates of sequence divergence in the evolutionary period soon after duplication [11,16,20-24]. Together, these observations indicate that initial relaxation of selection, or adaptive evolution, after duplication is limited to one of the paralogs, and that the slower-evolving paralog is more constrained by its ancestral

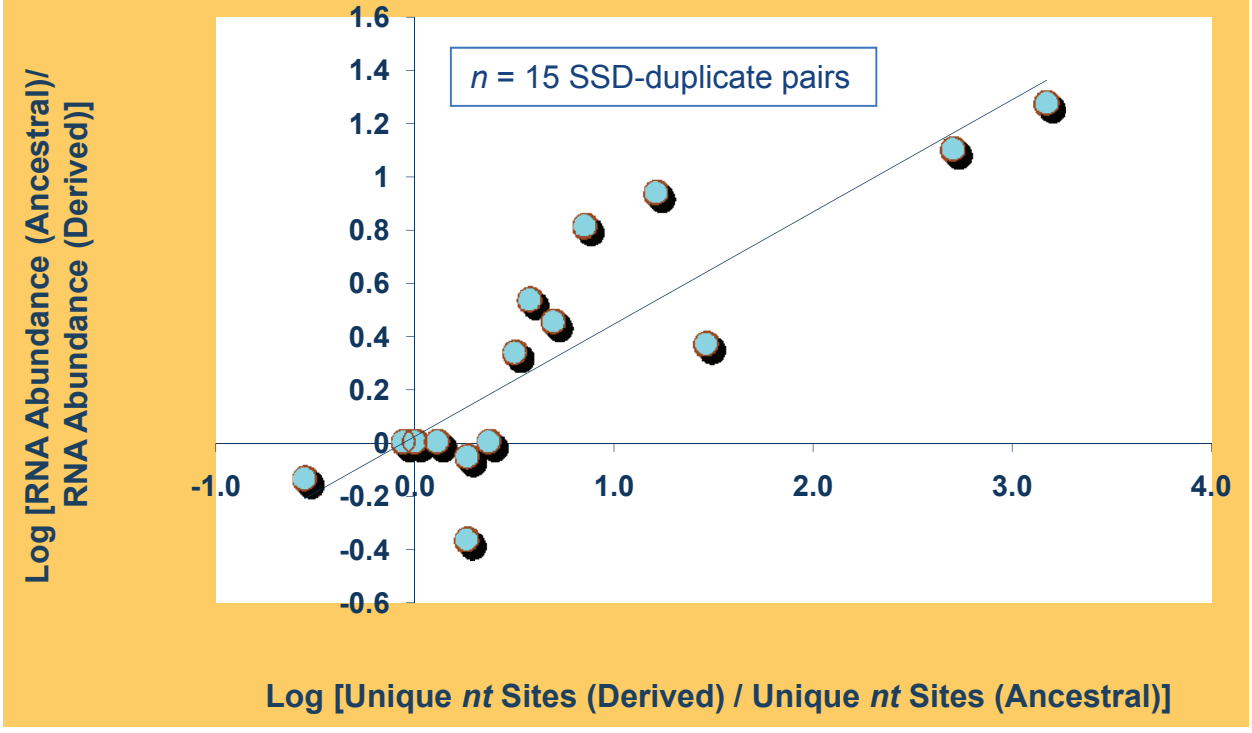

Figure 3 Nucleotide sequence asymmetry and mRNA abundance for 15 SSD pairs in the *S. cerevisiae* genome. The sequence asymmetry at the nucleotide level is expressed as the $\log_{10}$(unique sites in the derived paralog/unique sites in the ancestral paralog) and relative RNA abundance is expressed as the $\log_{10}$(RNA count for ancestral paralog/RNA count for derived paralog). There is a significant correlation between divergence between paralogs at the sequence level and divergence in their expression profiles (as represented by mRNA abundance) (*Kendall's tau* = 0.74; $p < 0.0002$).

function [11,22]. The majority of past studies did not distinguish between the ancestral and derived copies within a gene-duplicate pair, which in turn has precluded an unambiguous assessment of which copy is under stringent versus relaxed selective constraints.

There is some evidence that derived paralogs evolve faster than their counterparts residing at ancestral locations. In their study of evolutionarily young rodent gene duplicates, Cusack and Wolfe [16] assigned ancestral versus derived states to paralogs and demonstrated that genomic relocation of one paralog by retrotransposition engenders rate asymmetry in the sequence evolution of paralogs, commonly manifested as an accelerated rate of sequence evolution in the relocated paralog. Likewise, in bacterial genomes, the majority of paralogs that appear to have moved away from their ancestral gene neighborhood evolved faster than static paralogs [25]. Furthermore, a study of gene duplicates in four mammalian genomes determined that signatures of positive selection were more frequent in the derived copies than genes at their ancestral locations [17].

In this study, we analysed the rate of evolution in yeast paralogs for which an ancestral versus derived status could be assigned by analyzing synteny as manifested in gene-neighborhood conservation. There was significantly greater gene-neighborhood conservation in ohnologs relative to SSD pairs. Although ohnologs originated from an ancient polyploidization event and rampant genome-wide deletions have since restored functional normal ploidy in these *Saccharomyces* species [26,27], it is noteworthy that this extensive gene-neighborhood conservation has persisted. There is no difference in the extent of gene-neighborhood conservation in the upstream and downstream regions of the paralogs for both populations of duplicates (ohnologs and SSD), suggesting, on average, equal rates of preservation/loss of upstream and downstream neighboring genes.

The majority of gene duplicates with low sequence divergence in *S. cerevisiae* stem from an ancient WGD event rather than segmental duplications. Subsequent to the WGD event, there has been extensive loss of genetic material with an estimated 10% of the original ohnologs remaining [12]. Deletions of genetic material within a WGD-derived homology block have the potential to remove or rearrange regulatory sequences for the remaining genes in the block. Therefore, the DNA sequence of a paralog associated with more extensive gene-neighborhood conservation (i.e. local synteny)

might be under stronger purifying selection than a paralog residing in regions that have endured more gene loss and rearrangements. While it is problematic to assign ancestral versus derived states to gene duplicates originating from WGD events, we reasoned that a paralog within an ohnolog pair could be characterized as being ancestral-like or derived-like based on the extent of gene-neighborhood conservation it shared with a single-copy ortholog in an outgroup genome. We then sought to test the hypothesis that ancestral-like gene-copies within ohnolog pairs are more likely to maintain ancestral gene function and therefore exhibit lower rates of sequence evolution. In contrast, gene-copies displaying a reduction in the extent of local synteny relative to the ortholog may be predisposed to accelerated rates of sequence evolution and the resultant fates of neofunctionalization or nonfunctionalization. However, we find no evidence of an association between rate asymmetry in ohnologs and local gene-neighborhood conservation. In other words, for ohnologs, a decline in local gene-neighborhood conservation (derived-like) does not engender accelerated rates of sequence evolution either at the nucleotide or amino acid level. This is in contrast to a study of vertebrate genomes that found a significant correlation between synteny preservation and sequence conservation [28]. We speculate that the greater number of regulatory sites in vertebrate genomes might engender greater sensitivity to syntenic changes relative to yeast. However, ohnologs in yeast do exhibit a strong significant relationship between rate asymmetry and CAI such that the faster-evolving paralogs have lower CAI. The rate asymmetry in ohnologs also seems to be to some degree caused by relaxation of selection for codon usage in one copy.

Among the SSD pairs in our sample, it is the derived copy that evolves faster on average, both at the nucleotide and the amino acid level. This lends credence to Ohno's original hypothesis that duplication enables redundancy, enabling one copy to explore new evolutionary space by accumulating mutations [2]. It is likely that segmental duplications frequently do not capture the full repertoire of regulatory sequences [8] associated with the ancestral genes and/or result in the insertion of the derived copy into a region of the genome with different chromatin structure and potentially under the influence of different regulatory elements. Under these conditions, mutations that interfere with the ancestral gene's original function would still be selected against, whereas the derived copy could be under relaxed or positive selection. For SSD pairs, the rate asymmetry at the nucleotide level is likely due to a regime of relaxed selective constraints as there is a significant reduction in the CAI of the derived paralogs within SSD pairs. The CAI compares the codon usage of a gene to codon usage in highly expressed genes; hence, the reduction in the CAI values of derived paralogs suggests that selection for optimal codon usage has been relaxed in the derived copy. Puzzlingly, we failed to detect any correlation between nucleotide sequence asymmetry of SSD paralogs and changes in their CAI values. This may stem from limited power given the small sample size of available SSD duplicates in the yeast genome.

If the rate asymmetry in paralogs is largely a consequence of relaxation of selection in the derived paralog, it should also be manifested as different levels of expression among the two copies. Previous work has shown that the evolutionary rate in yeast is strongly influenced by gene expression [29,30]. In both the yeast ohnologs and SSD pairs studied here, mRNA abundance is correlated with the rate of evolution. Moreover, within SSD pairs, it is the derived paralogs that have lowered mRNA abundance relative to the ancestral loci. Both the CAI and mRNA abundance suggest that selective constraints on gene expression is a significant driver of evolutionary rate asymmetry in paralogs.

Conclusions

Following gene duplication, there is a general increase in the rate of evolution, and this increase is frequently asymmetric in that one paralog evolves at an accelerated pace. Asymmetry in the rate of molecular evolution after duplication has been variously associated with the evolution of novel functions, change in the number of interactions, and relaxation of selection. Here we address the related question if certain factors predispose one paralog to evolve faster. For instance, segmental duplications may translocate the derived copy to a different regulatory environment where it may evolve under different or reduced constraints [8]. Despite a limited sample of gene-duplicate pairs originating from recent small-scale duplications in *S. cerevisiae*, we find that the derived copy tends to evolve faster and is under reduced selection for codon usage. Accelerated rates in ohnologs are also associated with reduced selection for codon usage. Moreover, the rate of evolution is negatively correlated with mRNA abundance for ohnologs as well as SSD pairs. This adds to the evidence from mammals [17] that genes are not born equal and that the duplication process predisposes the derived copy to an evolutionary trajectory of initially reduced selective constraints and one that is perhaps more conducive to the evolution of new functions.

Methods

Identification of Gene Duplicates in S. cerevisiae with Low Synonymous Divergence

We initially selected gene families in the *S. cerevisiae* genome identified in a preceding study [31] that

comprised only two members and synonymous divergence (K_S) ≤ 0.35. This set had been extracted via the Genome History program [32] using the following parameters: (i) minimum translated ORF length of 100 aa, (ii) minimum number of aligned residues to accept pair being 100 aa, and (iii) using the BLAST matrix BLOSUM62 and acceptance of all BLAST hits with e ≤ 1e-07. The majority of gene duplicates within this initial sample were identified as '*ohnologs*' [33] or duplicates originating from a WGD event [12,34-37]. To further increase representation of gene duplicate pairs originating from small-scale duplication (SSD) events, we raised the K_S cut-off to 1.0 for two-member families and additionally included three-member gene families with K_S cut-off equal to 0.35. Ohnologs and SSD pairs in *S. cerevisiae* were distinguished by consulting Byrne and Wolfe's reconciled ohnolog list from recent comparative genomics studies [36]. The initial dataset after this first set of filtering procedures comprised 47 ohnologs and 31 SSD pairs.

Determination of the extent of synteny preservation with outgroup genomes

Synteny blocks (regions of conserved gene order) were retrieved on the YGOB database http://wolfe.gen.tcd.ie/ygob/). For ohnologs, the single-copy ortholog within the reconstructed ancestor chromosome that is hypothesized to exist immediately before the occurrence of the WGD event 100-200 mya [37] was used as a reference outgroup. For SSD-originating paralogs, the sequence of the most recent ancestor of the paralogs was inferred based on related genes in seven post-WGD yeast species (*Saccharomyces paradoxus, S. mikatae, S. kudriavzevii, S. bayanus, S. castellii, Candida glabrata*, and *Kluyveromyces polyspora*) using the codeml program of PAML by the setting the RateAncestor = 1 [38-40]. Tajima's Relative Rate test was then performed using DNA and protein sequences in triplets containing the two focal *S. cerevisiae* paralogs and their inferred ancestral sequence. In addition, duplications involving more than one gene locus, also referred to as 'linked sets' [31] were treated as a single duplication.

We used two measures to quantify the extent of gene-neighborhood conservation of each *S. cerevisiae* paralog in its upstream and downstream flanking regions. The first measure tallied the number of continuously shared genes with the outgroup genome in both the upstream and downstream directions. The second measure tallied the total number of genes shared with the outgroup genome within a block comprising 20 loci in both the upstream and downstream flanking regions. After excluding duplicate pairs with neither synteny nor outgroup information, the sample size of our study comprised 43 and 15 pairs of ohnologs and SSD-originated duplicates, respectively (Additional Files 1-4, Tables S1-S4).

Determining the degree of asymmetry among paralogs

Tajima's Relative Rate test [41], as implemented in MEGA version 4.0 [42] was used to determine if one of the paralogs was evolving faster. For SSD pairs, the designated outgroup sequence was a single-copy ortholog in an outgroup genome closely-related to *S. cerevisiae*. In the event that multiple outgroup species possessed a single-copy ortholog corresponding to *S. cerevisae*'s paralogs, we selected as outgroup the ortholog in the most closely-related outgroup genome. With respect to three-member gene families, the Tajima's test was only performed for the two most closely-related gene copies. For ohnologs, the outgroup was the phylogenetically closest species that contained a single-copy ortholog to the *S. cerevisiae* duplicate pair and diverged from the *Saccharomyces sensu stricto* group prior to the WGD event.

Genome and protein sequences of 11 fully sequenced yeast species were downloaded from the YGOB http://wolfe.gen.tcd.ie/ygob/ and KEGG http://www.genome.jp/kegg/catalog/org_list.html databases. Outgroup identification was performed using DNA and protein sequences of the paralogs as queries in BLASTN and BLASTP searches against the genomic and protein sequences of the 11 yeast species. The BLAST outputs were filtered and organized using a Perl script. Gene duplicate pairs and their associated outgroup sequences were first aligned with ClustalW 2.0 and then manually checked and improved, when necessary, before the analysis.

The Wilcoxon signed-rank test was used to test if, collectively speaking, the ancestral and derived copies of a gene duplicate pair are evolving at the same rate. Since the ohnolog copies could not be classified as ancestral or derived, this tests if the rate of evolution is associated with the conservation of flanking synteny. Five pairs of ohnologs with equal number of unique sites were excluded from the Wilcoxon signed-rank test to yield a final sample of 38 ohnolog pairs. For SSD pairs, the paralog with the greater upstream synteny compared to the outgroup is taken to be the ancestral copy. In the event that both paralogs have equal continuous synteny, the total synteny gene number within 20 gene loci was further included as a measure of synteny conservation. If the information above was insufficient for distinguishing the ancestral and the derived copies, the total synteny within 20 upstream and downstream gene loci was utilized.

Relationship between codon usage, mRNA abundance and rate asymmetry

The Codon Adaptation Index (CAI) was calculated using the JCat tool http://www.jcat.de[19,43]. The JCat tool uses the method of Carbone and colleagues [44] to select a set of reference genes with optimal codon

usage. In order to determine if differences in the rates of evolution are related to changes in optimal codon usage, we tested for correlation between the difference in number of unique sites (number of unique sites at the ancestral locus - number of unique sites at the derived locus) and the difference in CAI between paralogs (CAI of ancestral locus - CAI of derived locus).

An association between CAI and rate asymmetry between paralogs would suggest that gene expression is imposing differential constraints on the paralogs. As a proxy for gene expression, we obtained mRNA abundance data for all the paralogs in this study from a dataset consisting of transcript counts using single-molecule sequencing [45]. This data was used to test for an association between mRNA abundance and nucleotide rate asymmetry for both SSD pairs (Figure 3) and ohnologs.

Additional material

Additional file 1: *Table S1*. Tajima's Relative Rate Test for Ohnolog DNA sequences.

Additional file 2: *Table S2*. Tajima's Relative Rate Test for Ohnolog amino acid sequences.

Additional file 3: *Table S3*. Tajima's Relative Rate Test for DNA sequences of SSD pairs using a maximum-likelihood generated ancestral sequence as outgroup.

Additional file 4: *Table S4*. Tajima's Relative Rate Test for amino acid sequences of SSD pairs using a maximum-likelihood generated ancestral sequence as outgroup.

Acknowledgements
The authors gratefully acknowledge the comments of two anonymous reviewers whose comments helped improve the final manuscript. VK was supported by the National Science Foundation (NSF) Fellowship in Biological Informatics (DBI 0532735). UB and VK were supported by National Science Foundation (NSF) grant DEB-0952342.

Authors' contributions
LB was responsible for the data collection. VK and UB conceived the study. All authors conducted statistical analyses and participated in the preparation of the manuscript. All authors read and approved the final manuscript.

Received: 6 June 2011 Accepted: 28 September 2011
Published: 28 September 2011

References
1. Haldane JBS: **The part played by recurrent mutation in evolution.** *Am Nat* 1933, **67**:679-682.
2. Ohno S: *Evolution by Gene Duplication* Springer-Verlag; 1970.
3. Hughes MK, Hughes AL: **Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*.** *Mol Biol Evol* 1993, **10**:1360-1369.
4. Cronn RC, Small RL, Wendel JF: **Duplicated genes evolve independently after polyploid formation in cotton.** *Proc Natl Acad Sci USA* 1999, **96**:14406-14411.
5. Robinson-Rechavi M, Laudet V: **Evolutionary rate of duplicate genes in fish and mammals.** *Mol Biol Evol* 2001, **18**:681-683.
6. Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV: **Selection in the evolution of gene duplications.** *Genome Biol* 2002, **3**:RESEARCH0008.
7. Zhang L, Vision TJ, Gaut BS: **Patterns of nucleotide substitution among simultaneously duplicated gene pairs in *Arabidopsis thaliana*.** *Mol Biol Evol* 2002, **19**:1464-1473.
8. Lynch M, Katju V: **The altered evolutionary trajectories of gene duplicates.** *Trends Genet* 2004, **20**:544-549.
9. Van de Peer Y, Taylor J, Braasch I, Meyer A: **The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes.** *J Mol Evol* 2001, **53**:436-446.
10. Nembaware V, Crum K, Kelso J, Seoighe C: **Impact of the presence of paralogs on sequence divergence in a set of mouse-human orthologs.** *Genome Res* 2002, **12**:1370-1376.
11. Conant GC, Wagner A: **Asymmetric sequence divergence of duplicate genes.** *Genome Res* 2003, **13**:2052-2058.
12. Kellis M, Birren BW, Lander ES: **Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisae*.** *Nature* 2004, **428**:617-624.
13. Kim S-O, Yi SV: **Correlated asymmetry of sequence and functional divergence between duplicate proteins of *Saccharomyces cerevisiae*.** *Mol Biol Evol* 2006, **23**:1068-1075.
14. Katju V, Lynch M: **The structure and early evolution of recently arisen gene duplicates in the *Caenorhabditis elegans* genome.** *Genetics* 2003, **165**:1793-1803.
15. Katju V, Lynch M: **On the formation of novel genes by duplication in the *Caenorhabditis elegans* genome.** *Mol Biol Evol* 2006, **23**:1056-1067.
16. Cusack BP, Wolfe KH: **Not born equal: increased asymmetry in relocated and retrotransposed rodent gene duplicates.** *Mol Biol Evol* 2007, **24**:679-686.
17. Han MV, Demuth JP, McGrath CL, Casola C, Hahn MW: **Adaptive evolution of young gene duplicates in mammals.** *Genome Res* 2009, **19**:859-867.
18. Wolfe KH, Shields DC: **Molecular evidence for an ancient duplication of the entire yeast genome.** *Nature* 1997, **387**:708-713.
19. Sharp PM, Li WH: **The Codon Adaptation Index - a measure of directional synonymous codon usage bias and its potential applications.** *Nucl Acids Res* 1987, **11**:1281-1295.
20. Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV: **Selection in the evolution of gene duplications.** *Genome Biol* 2002, **3**:0008.1.
21. Wagner A: **Asymmetric functional divergence of duplicate genes in yeast.** *Mol Biol Evol* 2002, **19**:1760-1768.
22. Zhang P, Gu Z, Li W-H: **Different evolutionary patterns between young gene duplicates in the human genome.** *Genome Biol* 2003, **4**:R56.
23. Scannell DR, Wolfe KH: **A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast.** *Genome Res* 2008, **18**:137-147.
24. Panchin AY, Gelfand MS, Ramensky VE, Artamanova IE: **Asymmetric and non-uniform evolution of recently duplicated human genes.** *Biol Dir* 2010, **5**:54.
25. Notebaart RA, Huynen MA, Teusink B, Siezen RJ, Snel B: **Correlation between sequence conservation and the genomic context after gene duplication.** *Nucl Acids Res* 2005, **33**:6164-6171.
26. Cliften P, Fulton RS, Wilson RK, Johnston M: **After the duplication: gene loss and adaptation in *Saccharomyces* genomes.** *Genetics* 2006, **172**:863-872.
27. Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH: **Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts.** *Nature* 2006, **440**:341-345.
28. Abi-Rached L, Gilles A, Shiina T, Pontarotti P, Inoko H: **Evidence of *en bloc* duplication in vertebrate genomes.** *Nat Genetics* 2002, **31**:100-105.
29. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH: **Why highly expressed proteins evolve slowly.** *Proc Natl Acad Sci USA* 2005, **102**:14338-14343.
30. Drummond DA, Raval A, Wilke CO: **A single determinant dominates the rate of yeast protein evolution.** *Mol Biol Evol* 2006, **23**:327-337.
31. Katju V, Farslow JC, Bergthorsson U: **Variation in gene duplicates with low synonymous divergence in *Saccharomyces cerevisiae* relative to *Caenorhabditis elegans*.** *Genome Biol* 2009, **10**:R75.
32. Conant GC, Wagner A: **GenomeHistory: a software tool and its application to fully sequenced genomes.** *Nucl Acids Res* 2002, **30**:3378-3386.
33. Wolfe K: **Robustness - it's not where you think it is.** *Nat Genet* 2000, **25**:3-4.
34. Wong S, Butler G, Wolfe KH: **Gene order evolution and paleopolyploidy in hemiascomycete yeasts.** *Proc Natl Acad Sci USA* 2002, **99**:9272-9277.
35. Dietrich FS, Voegeli S, Brachat S, Lerch A, Gates K, Steiner S, Mohr C, Pöhlmann R, Luedi P, Choi S, Wing RA, Flavier A, Gaffney TD, Philippsen P:

The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science* 2004, **304**:304-307.
36. Byrne KP, Wolfe KH: **The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species.** *Genome Res* 2005, **15**:1456-1461.
37. Gordon JL, Byrne KP, Wolfe KH: **Additions, losses and rearrangements on the evolutionary route from a reconstructed ancestor to the modern *Saccharomyces cerevisiae* genome.** *PLoS Genet* 2009, **5**:e1000485.
38. Yang Z, Kumar S, Nei M: **A new method of inference of ancestral nucleotide and amino acid sequences.** *Genetics* 1995, **141**:1641-1650.
39. Koshi JM, Goldstein RA: **Probabilistic reconstruction of ancestral protein sequences.** *J Mol Evol* 1996, **42**:313-320.
40. Yang Z: *Computational Molecular Evolution* Oxford University Press, Oxford, England; 2006.
41. Tajima F: **Simple methods for testing the molecular evolutionary clock hypothesis.** *Genetics* 1993, **135**:599-607.
42. Tamura K, Dudley J, Nei M, Kumar S: **MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0.** *Mol Biol Evol* 2007, **24**:1596-1599.
43. Grote A, Hiller K, Scheer M, Münch R, Nörtemann B, Hempel DC, Jahn D: **JCat: a novel tool to adapt codon usage of a target gene to its potential expression host.** *Nucl Acids Res* 2005, **33**:W526-W531.
44. Carbone A, Zinovyev A, Képès F: **Codon adaptation index as a measure of dominating codon bias.** *Bioinformatics* 2003, **19**:2005-2015.
45. Lipson D, Raz T, Kieu A, Jones DR, Giladi E, Thayer E, Thompson JF, Letovsky S, Milos P, Causey M: **Quantification of the yeast transcriptome by single-molecule sequencing.** *Nat Biotech* 2009, **27**:652-658.

doi:10.1186/1471-2148-11-279
Cite this article as: Bu *et al.*: Local synteny and codon usage contribute to asymmetric sequence divergence of *Saccharomyces cerevisiae* gene duplicates. *BMC Evolutionary Biology* 2011 **11**:279.

Table S1: Tajima's Relative Rate Test for Ohnolog DNA sequences.

	Ancestral Paralog (A)	*Derived Paralog (B)*	*Outgroup (C)*	χ^2	*p-value*	*Unique Sites*		
						A	**B**	**C**
1	YBL027W	YBR084C-A	kla:KLLA0E12463g	0.03	0.85746	15	16	54
2	YBL072C	YER102W	kla:KLLA0E20559g	3.60	0.05778	2	8	72
3	YBR031W	YDR012W	kla:KLLA0B07139g	0.11	0.73888	5	4	155
4	YBR048W	YDR025W	kla:KLLA0A10483g	5.76	0.01638	5	16	41
5	YDL131W	YDL182W	kla:KLLA0F05489g	5.59	0.01810	20	38	176
6	YDL191W	YDL136W	kla:KLLA0F05247g	0.33	0.56370	1	2	39
7	YDR342C	YHR092C	kla:KLLA0D13310g	1.00	0.31731	84	90	210
8	YDR447C	YML024W	kla:KLLA0B01474g	0.22	0.63735	8	10	41
9	YEL034W	YJR047C	kla:KLLA0E22286g	0.02	0.87590	21	20	37
10	YER074W	YIL069C	kla:KLLA0C07755g	0.33	0.56370	5	7	33
11	YFR031C-A	YIL018W	kla:KLLA0D16027g	1.81	0.17793	10	17	55
12	YGL031C	YGR148C	kla:KLLA0E10857g	0.50	0.47950	14	18	44
13	YGR034W	YLR344W	kla:KLLA0B05742g	13.5	0.00024	3	21	34
14	YGR118W	YPR132W	kla:KLLA0B11231g	5.40	0.02014	12	3	28
15	YGR138C	YPR156C	kla:KLLA0E03729g	0.40	0.52454	64	57	444
16	YGR192C	YJR009C	ago:AGOS_AER031C	0.93	0.33592	11	16	169
17	YHL033C	YLL045C	kla:KLLA0E00506g	0.50	0.47950	18	14	88
18	YHR066W	YDR312W	kla:KLLA0C14586g	0.64	0.42503	35	42	349
19	YHR141C	YNL162W	kla:KLLA0D07832g	0.00	1.00000	3	3	26
20	YHR203C	YJR145C	kla:KLLA0B03652g	0.07	0.79625	7	8	79
21	YKL006W	YHL001W	kla:KLLA0B13409g	0.29	0.59298	8	6	53
22	YKR059W	YJL138C	kla:KLLA0A05731g	0.20	0.65472	2	3	188
23	YLR333C	YGR027C	kla:KLLA0B06193g	2.13	0.14440	15	8	37
24	YML026C	YDR450W	kla:KLLA0B01562g	0.20	0.65472	11	9	25
25	YML063W	YLR441C	kla:KLLA0B05060g	6.12	0.01338	20	39	59
26	YML073C	YLR448W	kla:KLLA0B04686g	0.38	0.53709	19	23	67
27	YMR121C	YLR029C	kla:KLLA0F17633g	18.69	0.00002	33	6	35
28	YMR142C	YDL082W	kla:KLLA0E22099g	16.03	0.00006	7	32	50
29	YMR143W	YDL083C	kla:KLLA0E22077g	0.14	0.70546	15	13	27
30	YMR186W	YPL240C	kla:KLLA0D12958g	0.03	0.86853	74	72	255
31	YMR230W	YOR293W	kla:KLLA0B08173g	2.58	0.10829	13	6	44
32	YNL209W	YDL229W	kla:KLLA0D19041g	0.10	0.75762	22	20	189
33	YOL120C	YNL301C	kla:KLLA0A07227g	3.85	0.04986	8	18	54
34	YOL121C	YNL302C	kla:KLLA0A07194g	0.00	1.00000	10	10	39
35	YOR133W	YDR385W	kla:KLLA0E02926g	3.00	0.08326	0	3	227
36	YOR182C	YLR287C-A	kla:KLLA0C04809g	4.50	0.03389	1	7	21
37	YOR312C	YMR242C	kla:KLLA0F08657g	0.62	0.43277	15	11	47
38	YPL079W	YBR191W	kla:KLLA0E23727g	0.14	0.70546	15	13	32
39	YPL090C	YBR181C	kla:KLLA0E24090g	0.00	1.00000	3	3	73
40	YPL198W	YGL076C	kla:KLLA0D03410g	3.20	0.07364	14	6	100
41	YPL220W	YGL135W	kla:KLLA0B02002g	0.00	1.00000	0	0	69
42	YPR080W	YBR118W	kla:KLLA0B08998g	0.00	1.00000	1	1	83
43	YPR102C	YGR085C	kla:KLLA0F08261g	1.67	0.19671	5	10	48

Table S2: Tajima's Relative Rate Test for Ohnolog amino acid sequences.

	Ancestral Paralog (A)	*Derived Paralog (B)*	*Outgroup (C)*	χ^2	*p-value*	*Unique Sites*		
						A	**B**	**C**
1	YBL027W	YBR084C-A	kla:KLLA0E12463g	0.00	1.00000	0	0	24
2	YBL072C	YER102W	kla:KLLA0E20559g	0.00	1.00000	0	0	24
3	YBR031W	YDR012W	kla:KLLA0B07139g	0.00	1.00000	0	0	50
4	YBR048W	YDR025W	kla:KLLA0A10483g	0.00	1.00000	0	0	15
5	YDL131W	YDL182W	kla:KLLA0F05489g	2.13	0.14440	8	15	13
6	YDL191W	YDL136W	kla:KLLA0F05247g	0.00	1.00000	0	0	13
7	YDR342C	YHR092C	kla:KLLA0D13310g	1.98	0.15990	16	25	108
8	YDR447C	YML024W	kla:KLLA0B01474g	1.00	0.31730	0	1	15
9	YEL034W	YJR047C	kla:KLLA0E22286g	0.08	0.78150	7	6	11
10	YER074W	YIL069C	kla:KLLA0C07755g	0.00	1.00000	0	0	11
11	YFR031C-A	YIL018W	kla:KLLA0D16027g	0.00	1.00000	0	0	15
12	YGL031C	YGR148C	kla:KLLA0E10857g	0.00	1.00000	2	2	19
13	YGR034W	YLR344W	kla:KLLA0B05742g	1.00	0.31730	0	1	10
14	YGR118W	YPR132W	kla:KLLA0B11231g	0.00	1.00000	0	0	3
15	YGR138C	YPR156C	kla:KLLA0E03729g	0.18	0.66980	12	10	104
16	YGR192C	YJR009C	ago:AGOS_AER031C	1.60	0.20590	3	7	42
17	YHL033C	YLL045C	kla:KLLA0E00506g	1.00	0.31730	3	1	37
18	YHR066W	YDR312W	kla:KLLA0C14586g	0.82	0.36570	4	7	104
19	YHR141C	YNL162W	kla:KLLA0D07832g	0.00	1.00000	0	0	10
20	YHR203C	YJR145C	kla:KLLA0B03652g	0.00	1.00000	0	0	19
21	YKL006W	YHL001W	kla:KLLA0B13409g	1.00	0.31730	1	0	18
22	YKR059W	YJL138C	kla:KLLA0A05731g	0.00	1.00000	0	0	61
23	YLR333C	YGR027C	kla:KLLA0B06193g	1.00	0.31730	1	0	14
24	YML026C	YDR450W	kla:KLLA0B01562g	0.00	1.00000	0	0	10
25	YML063W	YLR441C	kla:KLLA0B05060g	1.29	0.25680	2	5	14
26	YML073C	YLR448W	kla:KLLA0B04686g	0.50	0.47950	5	3	27
27	YMR121C	YLR029C	kla:KLLA0F17633g	0.00	1.00000	1	1	8
28	YMR142C	YDL082W	kla:KLLA0E22099g	0.00	1.00000	0	0	23
29	YMR143W	YDL083C	kla:KLLA0E22077g	0.00	1.00000	0	0	6
30	YMR186W	YPL240C	kla:KLLA0D12958g	2.57	0.10880	4	10	65
31	YMR230W	YOR293W	kla:KLLA0B08173g	2.00	0.15730	2	0	19
32	YNL209W	YDL229W	kla:KLLA0D19041g	0.33	0.56370	2	1	54
33	YOL120C	YNL301C	kla:KLLA0A07227g	0.00	1.00000	0	0	18
34	YOL121C	YNL302C	kla:KLLA0A07194g	1.00	0.31730	0	1	17
35	YOR133W	YDR385W	kla:KLLA0E02926g	0.00	1.00000	0	0	60
36	YOR182C	YLR287C-A	kla:KLLA0C04809g	0.00	1.00000	0	0	7
37	YOR312C	YMR242C	kla:KLLA0F08657g	0.00	1.00000	0	0	15
38	YPL079W	YBR191W	kla:KLLA0E23727g	2.00	0.15730	2	0	9
39	YPL090C	YBR181C	kla:KLLA0E24090g	0.00	1.00000	0	0	29
40	YPL198W	YGL076C	kla:KLLA0D03410g	2.00	0.15730	2	0	27
41	YPL220W	YGL135W	kla:KLLA0B02002g	0.08	0.78150	0	0	19
42	YPR080W	YBR118W	kla:KLLA0B08998g	0.00	1.00000	0	0	17
43	YPR102C	YGR085C	kla:KLLA0F08261g	1.00	0.31730	0	1	16

Table S3: Tajima's Relative Rate Test for DNA sequences of SSD pairs using a maximum-likelihood generated ancestral sequence as outgroup.

	Ancestral paralog (A)	*Derived paralog (B)*	χ^2	*p-value*	*Unique Sites*		
					A	**B**	**C (ancestral sequence)**
1	YDL075W	YLR406C	2.88	0.0896	5	12	2
2	YDR039C	YDR038C	0.00	1.0000	0	0	15
3	YDR533C	YOR391C	6.74	0.0094	11	27	7
4	YFL009W	YER066W	49.50	0.0000	11	77	12
5	YFL058W	YNL332W	0.00	1.0000	2	2	10
6	YGL258W	YOR387C	5.76	0.0164	5	16	17
7	YHR055C	YHR053C	0.00	1.0000	0	0	8
8	YHR056C	YHR054C	0.00	1.0000	0	0	108
9	YLR044C	YLR134W	160.17	0.0000	15	201	5
10	YNL067W	YGL147C	1.00	0.3173	21	15	8
11	YOL055C	YPL258C	0.97	0.3258	124	109	82
12	YOL086C	YMR303C	97.85	0.0000	5	112	0
13	YOR388C	YPL276W_275W	19.59	0.0000	2	25	56
14	YOR389W	YPL277C_278C	28.58	0.0000	20	71	99
15	YPL279C	YOR390W	2.00	0.1573	6	2	0

Note: Cells containing two gene IDs comprise cases where the exon-intron structure of the original locus has been altered to comprise two genes.

Table S4: Tajima's Relative Rate Test for amino acid sequences of SSD pairs using a maximum-likelihood generated ancestral sequence as outgroup.

	Ancestral paralog (A)	*Derived paralog (B)*	χ^2	*p-value*	*Unique Sites*		
					A	**B**	**C (ancestral sequence)**
1	YDL075W	YLR406C	1.00	0.3173	0	1	0
2	YDR039C	YDR038C	0.00	1.0000	0	0	0
3	YDR533C	YOR391C	13.00	0.0003	0	13	0
4	YFL009W	YER066W	37.00	0.0000	0	37	0
5	YFL058W	YNL332W	0.00	1.0000	0	0	0
6	YGL258W	YOR387C	3.57	0.0588	1	6	2
7	YHR055C	YHR053C	0.00	1.0000	0	0	3
8	YHR056C	YHR054C	0.00	1.0000	0	0	42
9	YLR044C	YLR134W	59.24	0.0000	2	65	0
10	YNL067W	YGL147C	0.00	1.0000	2	2	2
11	YOL055C	YPL258C	36.94	0.0000	36	25	27
12	YOL086C	YMR303C	16.67	0.0000	2	22	0
13	YOR388C	YPL276W_275W	11.00	0.0009	0	11	7
14	YOR389W	YPL277C_278C	10.31	0.0013	8	27	26
15	YPL279C	YOR390W	3.00	0.0833	3	0	0

Note: Cells containing two gene IDs comprise cases where the exon-intron structure of the original locus has been altered to comprise two genes.