# Do disparate mechanisms of duplication add similar genes to the genome?

**Jerel C. Davis and Dmitri A. Petrov**

Department of Biological Sciences, Stanford University, 371 Serra Mall, Stanford, CA 90305-5020, USA

**Gene duplication is the fundamental source of new genes. Biases in duplication have profound implications for the dynamics of gene content during evolution. In this article, we compare genes arising from whole gene duplication (WGD), smaller scale duplication (SSD) and singletons in *Saccharomyces cerevisiae*. Our results demonstrate that genes duplicated by WGD and SSD are similarly biased with respect to codon bias and evolutionary rate, although differing significantly in their functional constituency.**

## Introduction

Gene duplication is the major source of new genes [1] and consequently is a central force affecting genome evolution [2]. Duplications are known to occur on two fundamental scales: whole genome duplication (WGD) and smaller scale duplication (SSD). WGD has been important in the evolutionary history of several animal and plant lineages [1,3–9] and SSDs, involving one or several genes, occur continuously by several mechanisms [2,10,11].

It is known that certain types of genes are more likely to lead to persistent duplicates than others [12–16]. It is unknown, however, whether both WGD and SSD lead to similar compliments of persistent duplicate genes. New duplicates must pass through several sieving stages to become a persistent part of the genome and these sieving stages are somewhat different for WGD and SSD (Box 1). Therefore, these two modes of duplication might influence gene content in distinct ways.

In this article, we investigated whether both SSD and WGD in *Saccharomyces cerevisiae* (WGD occurred ~100 million years ago (Mya) [4,17]) led to similar compliments of persistent duplicate genes by investigating their functional classification, codon bias and rate of evolution before duplication. We found that although both WGD and SSD sets have a greater codon bias and arise from more slowly evolving genes than those that remained as single copies, the two duplicate sets are enriched for different functional classes of genes.

## Identification of genes in the three classes

We identified 2126 singleton genes, 356 WGD duplicates and 626 SSD duplicates in *S. cerevisiae* as described in supplementary material online. For simplicity, we limited both WGD and SSD sets to consist only of duplicate genes with no other paralogs in the genome. The average

### Box 1. The process leading to long-term duplicate survival

The WGD and SSD duplication processes differ in several respects (Figure I). First, in SSD a duplicate gene must undergo an independent mutational event. Second, a duplicated gene must start from being present in one individual in the population to becoming present in the entire population (fixation) – passing through this stage depends largely on whether the duplication is selectively advantageous, deleterious or neutral. Most genes never survive this step [26]. Finally, duplicate genes which are functionally redundant with the ancestral copy must diverge, so they do not completely overlap in function (e.g. Refs [27–29]), for both duplicate copies to persist. Many duplicate genes are never preserved and become quickly silenced over the course of a few million years [19]. Although WGD shares the preservation step, the first two steps of the process differ. First, genes arising by WGD do not duplicate independently but duplicate with the rest of the genome. Second, the genes do not need to fix in the population but must instead survive a period of rapid genome rearrangement and gene loss [4,9,30]. Because these two processes differ, the complements of genes arising from each mechanism can also vary.

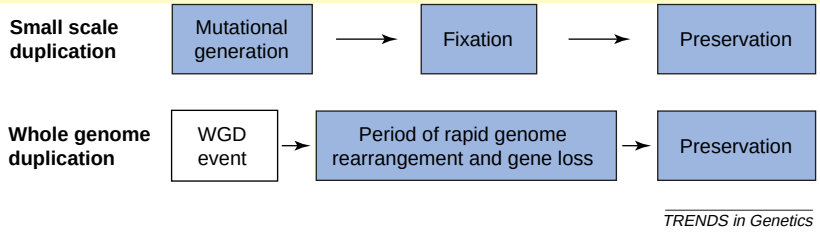


**Figure I.** An illustration of the steps leading to the generation and long-term persistence of duplicate genes. The blue colour identifies those steps that can potentially bias the set of eventual duplicate genes. The only step shared between the two processes is duplicate gene preservation.

*Corresponding author:* Davis, J.C. (jerel@stanford.edu).
Available online 11 August 2005

**Table 1. Evolutionary rate estimates for the three sets of study genes obtained by measuring the divergence between outgroup orthologs**

| Gene set | Nematode outgroup lineages[b] | | Fly outgroup lineages[c] | |
|---|---|---|---|---|
| | $K_A$ (all genes) | $K_A$ (without ribsomal genes) | $K_A$ (all genes) | $K_A$ (without ribsomal genes) |
| Singletons | 0.095 | 0.097 | 0.075 | 0.077 |
| WGD duplicates | 0.061 **[d] | 0.089 * | 0.045 ** | 0.061 * |
| SSD duplicates | 0.079 ** | 0.087 * | 0.063 ** | 0.070 * |

[a] The gene studied included singltons, WGD duplicates and SSD duplicates. Both types of duplicates appear to arise from conserved genes.
[b]Pairs of outgroup orthologs were identified in *C. elegans* and *C. briggsae* where possible (see methods in supplementary material online) and mean divergence between these pairs for each group, both with and without ribosomal proteins is shown.
[c]Similar mean divergence estimates for representative pairs from *D. melanogaster* and *D. pseudoobscura* are shown.
[d]Significance levels are * $P<0.05$, ** $P<0.01$ for Mann-Whitney U-test, compared with the set of singleton genes.

divergence between SSD pairs is greater than the divergence between WGD duplicates, probably reflecting their greater age (Figure 1 in the supplementary material online). The variance in non-synonymous divergence is greater for duplicate pairs that arose independently of polyploidization as expected.

## Elevated codon bias of WGD and SSD duplicates

Duplicate genes have previously been shown to have an elevated codon bias in *S. cerevisiae* and *Caenorhabditis elegans* [16], suggesting that highly expressed genes duplicate preferentially. To estimate the level of expression of genes belonging to our three study groups, we measured their codon bias using the frequency of optimal codons (FOP) metric. On average, duplicate genes arising from both WGD (mean FOP=0.548) and SSD (mean FOP=0.495) have a greater codon bias than singletons (mean FOP=0.458) (Mann-Whitney U-test, $P<0.0001$) (Figure 2 in the supplementary material online). The greater FOP values for WGD compared with SSD can be largely attributed to the overabundance of highly codon biased ribosomal proteins in the WGD set. Once ribosomal proteins are removed, the two sets do not have significantly different FOP values (0.485 and 0.471, respectively; $P=0.84$), although the codon bias of both sets remain significantly elevated above singleton genes ($P=0.0002$, $P<0.0001$, respectively).

## The slow evolutionary rate of duplicates arising from WGD and SSD

Extant duplicates in many organisms have arisen primarily from genes that are highly conserved before duplication [15,16]. Here we asked whether duplicates from both WGD and SSD arise from slowly evolving genes. We estimated evolutionary rates for the genes in the three study sets using two pairs of outgroup orthologs (from *C. elegans* and *C. briggsae* and *Drosophila melanogaster* and *D. pseudoobscura*). This method estimates the rate of protein evolution from outgroup lineages in a way that is not biased by the duplication event itself [15,16] (see addendum in the supplementary material online).

Our analysis reveals that orthologs of both sets of duplicate genes evolve significantly more slowly than orthologs of singletons (Table 1). The evolutionary rate of the orthologs of duplicate genes (excluding ribosomal genes) is ~10–15% slower than those for the orthologs of the singleton genes. However, this difference is less pronounced than that estimated for the complete set of all duplicate genes in the *S. cerevisiae* genome (which together possess a ~25% depression in evolutionary rate) [16], possibly indicating that genes that give rise to large gene families evolve even slower than those from families with two members (analyzed here).

## Duplicate gene sets are enriched for different functional classes

We used molecular function gene ontology (GO) annotations [18] to study functional properties of the three sets of study genes. At the second level under 'molecular function' in the GO directed acyclic graph (DAG) the three sets differ significantly from one another (Table 2) ($P \ll 0.0001$, G-test, all pairwise comparisons). Interestingly, the two duplicate sets differ from singletons in distinct ways. For example, only the WGD duplicates are enriched for ribosomal proteins, leading to their over-representation in the 'structural' class. After removing ribosomal genes from the analysis, WGD duplicates possess an overabundance of 'catalytic' proteins, a paucity of 'binding' proteins and 'enzyme regulator' proteins (Table 2). By contrast, the SSD duplicates display a lack of 'transcription regulator' proteins and an overabundance of 'enzyme regulator' proteins.

To identify more specific functional differences between the two duplicate sets and the singletons we compared their third-level GO functional annotations (Table 3). This analysis reveals that the excess of 'transferase' and

**Table 2. The percentage of genes in each paralogy class belonging to each high-level GO functional category[a]**

| GO function | Including ribosomal genes | | | Without ribosomal genes | | |
|---|---|---|---|---|---|---|
| | Singletons | WGD pairs | SSD pairs | Singletons | WGD pairs | SSD pairs |
| Catalytic | 48.54% | 45.21% | 48.74% | 51.64% | 63.06% *[b] | 54.19% ns |
| Binding | 24.73% | 21.00% | 26.38% | 26.19% | 21.66% ns | 28.77% ns |
| Transcription regulator | 10.48% | 9.13% | 4.77% | 11.24% | 12.74% ns | 5.31% * |
| Structural molecule | 10.39% | 30.59% | 11.81% | 9.54% | 4.46% * | 10.89% ns |
| Transporter | 8.90% | 3.20% | 9.81% | 3.92% | 3.18% ns | 1.96% ns |
| Enzyme regulator | 3.56% | 2.74% | 6.78% | 3.82% | 3.82% ns | 7.54% * |
| Signal transducer | 0.69% | 1.83% | 2.01% | 1.80% | 1.91% ns | 1.40% ns |

[a]Only the seven largest functional classes are shown, genes of unknown function have been removed. The total percentages can exceed 100% because a gene can be annotated as belonging to more than one class.
[b]Significance levels are ** $P<0.001$, * $P<0.05$ and not significant (ns) compared with the set of singleton genes.

**Table 3. Most significant differences at the second level of the GO directed acyclic graph for each subset of study genes[a]**

| GO function | Singletons | WGD pairs | SSD pairs |
|---|---|---|---|
| Transferase | 20.47% | 32.00% **[b] | 21.37% ns |
| RNA polymerase II activity | 4.89% | 4.00% ns | 2.28% * |
| Ion transporter | 4.45% | 0.00% ** | 3.13% ns |
| Kinase activity | 2.78% | 6.00% * | 4.84% ns |
| Transcription factor activity | 1.11% | 4.67% ** | 0.57% ns |
| Kinase regulator activity | 0.56% | 1.33% ns | 2.56% * |
| Electron transporter activity | 0.44% | 2.67% * | 0.57% ns |

[a]Ribosomal genes and genes with an unknown function were removed from each of the sets before comparison.
[b]Significance levels are ** $P<0.001$, * $P<0.05$ and not significant (ns) compared with the set of singleton genes.

'kinase' enzymes among WGD genes can account for the enrichment of 'catalytic' proteins. In addition, the scarcity of 'binding' proteins in the WGD class might be explained by a lack of 'protein binding' functional annotations. The paucity of 'transcription regulator' annotations in the SSD set can be attributed to the absence of 'RNA Pol II transcription factor' annotations and the abundance of 'enzyme regulatory' proteins is largely attributable to 'kinase regulators'.

## Stoichiometry and functional differences

What explains the different functional complements of the two duplicate sets? The different stoichiometric environments following WGD and SSD might provide an explanation for some of these differences; the stoichiometric environment is particularly relevant to genes that must be expressed at a precise ratio with respect to other loci, either because they participate in multi-protein complexes or function in related biochemical pathways. Such genes are unlikely to enter the genome by SSD because of deleterious consequences for stoichiometry. By contrast, the loss of such duplicates can be deleterious following WGD, leading to their selective maintenance [19]. A second possibility is that an increase in dosage of particular protein complexes might be selectively favored, but a functional dosage increase might require simultaneous duplication of all the constituents of the complex. In either case, WGD is predicted to lead to the maintenance of dosage sensitive sets of genes and complexes, whereas SSD is not expected to lead to the duplication of such genes.

Several of the functional differences we observe can be explained by the prediction that multidomain complex genes and stoichiometrically sensitive genes will be more likely to arise by WGD. Ribosomal subunit proteins, which are over-abundant only in the WGD set, represent one prime example. Similarly, a close analysis reveals that the dearth of 'transcription regulator' proteins in the SSD set results from the absence of regulatory complex proteins such as transcription factor II (TAFII) polymerase recruitment complexes, SNF–SWI complexes and polymerase II holoenzyme subunits. Such proteins are present only in the singleton and WGD sets (analysis of fourth GO-level, data not shown). Moreover, the enrichment of 'transferase' proteins following WGD can result from the maintenance of stoichiometry along biochemical pathways. For example, TKL1 (and TKL2) and TAL1 (and YGR043C) represent two pairs of transferase proteins that catalyze adjacent steps of the pentose-phosphate pathway. However, another example of duplication of catalytic proteins by WGD is the RNR complex, which catalyzes the rate-limiting step in dNTP synthesis. Both the large and the small subunit of this complex (YIL066C/YER070W and YGR180C/YJL026W) have been maintained since WGD occurred. Taken together, these examples support the idea that stoichiometric differences can affect the propensity of genes to duplicate [20–22] by SSD and WGD mechanisms.

## Concluding remarks and future work

Our results show that the two sets of duplicate genes in *S. cerevisiae* (WGD and SSD duplicates) are similar with respect to their elevated codon biased and their slow rate of evolution before duplication but differ in their functional constituency. Future work will help determine which steps in the process of duplication (Box 1) are responsible for the differences and similarities in the two sets. Differences in the two duplicate sets (e.g. function class differences) can potentially result from distinct steps in the two processes. However, similarities between the two sets might be the result of shared biases in the two duplication processes (see addendum in the supplementary material online). Future research comparing the functional classes and molecular attributes of the genes that have duplicated in other organisms, particularly in well-curated genomes in which WGD and SSD genes can be cleanly separated, will be beneficial in continuing to elucidate the dynamics and determinants of genic diversity.

## Supplementary data

A complete methods section can be found in the supplementary material. Briefly, we identified singleton genes, WGD genes of family size two, and SSD genes of family size in *S. cerevisiae* using data provided by previous work [17] and conventional methods using BLAST [23]. Amino acid substitutions were estimated with codeml [24], codon bias (measured as FOP) was calculated using CodonW, written by J. Peden, and GO analysis was performed using FatiGo [25]. Supplementary data associated with this article can be found at doi:10.1016/j.tig.2005.07.008

## References

1 Ohno, S. (1970) *Evolution by Gene Duplication*, Springer Verlag
2 Wolfe, K.H. and Li, W.H. (2003) Molecular evolution meets the genomics revolution. *Nat. Genet.* 33 (Suppl.), 255–265
3 Friedman, R. and Hughes, A.L. (2001) Gene duplication and the structure of eukaryotic genomes. *Genome Res.* 11, 373–381
4 Wolfe, K.H. and Shields, D.C. (1997) Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387, 708–713
5 Semple, C. and Wolfe, K.H. (1999) Gene duplication and gene conversion in the *Caenorhabditis elegans* genome. *J. Mol. Evol.* 48, 555–564

6 McLysaght, A. *et al*. (2002) Extensive genomic duplication during early chordate evolution. *Nat. Genet.* 31, 200–204
7 Meyer, A. and Schartl, M. (1999) Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. *Curr. Opin. Cell Biol.* 11, 699–704
8 Vision, T.J. *et al*. (2000) The origins of genomic duplications in *Arabidopsis*. *Science* 290, 2114–2117
9 Wendel, J.F. (2000) Genome evolution in polyploids. *Plant Mol. Biol.* 42, 225–249
10 Sankoff, D. (2001) Gene and genome duplication. *Curr. Opin. Genet. Dev.* 11, 681–684
11 Long, M. *et al*. (2003) The origin of new genes: glimpses from the young and old. *Nat. Rev. Genet.* 4, 865–875
12 Kondrashov, F.A. *et al*. (2002) Selection in the evolution of gene duplications. *Genome Biol*. doi: 10.1186/gb-2002-3-2-research0008 (http://genomebiology.com/2002/3/2/research/0008)
13 Gu, Z. *et al*. (2002) Extent of gene duplication in the genomes of *Drosophila*, nematode, and yeast. *Mol. Biol. Evol.* 19, 256–262
14 Seoighe, C. and Wolfe, K.H. (1999) Yeast genome evolution in the post-genome era. *Curr. Opin. Microbiol.* 2, 548–554
15 Jordan, I.K. *et al*. (2004) Duplicated genes evolve slower than singletons despite the initial rate increase. *BMC Evol. Biol*. doi: 10.1186/1471-2148-4-22 (http://www.biomedcentral.com/1471-2148/4/22)
16 Davis, J.C. and Petrov, D.A. (2004) Preferential duplication of conserved proteins in eukaryotic genomes. *PLoS Biol.* 2. doi: 10.1371/journal.pbio.0020055 (http://biology.plosjournals.org)
17 Kellis, M. *et al*. (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature* 428, 617–624
18 Ashburner, M. and Lewis, S. (2002) On ontologies for biologists: the Gene Ontology–untangling the web. *Novartis Found Symp* 247, 66–80 discussion 80–83, 84–90, 244–252
19 Lynch, M. and Conery, J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290, 1151–1155
20 Veitia, R.A. (2002) Exploring the etiology of haploinsufficiency. *BioEssays* 24, 175–184
21 Papp, B. *et al*. (2003) Dosage sensitivity and the evolution of gene families in yeast. *Nature* 424, 194–197
22 Yang, J. *et al*. (2003) Organismal complexity, protein complexity, and gene duplicability. *Proc. Natl. Acad. Sci. U. S. A.* 100, 15661–15665
23 Altschul, S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402
24 Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* 13, 555–556
25 Al-Shahrour, F. *et al*. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* 20, 578–580
26 Lynch, M. *et al*. (2001) The probability of preservation of a newly arisen gene duplicate. *Genetics* 159, 1789–1804
27 Hughes, A.L. (1994) The evolution of functionally novel proteins after gene duplication. *Proc. Biol. Sci.* 256, 119–124
28 Force, A. *et al*. (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151, 1531–1545
29 Gu, Z. *et al*. (2002) Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet.* 18, 609–613
30 Seoighe, C. and Wolfe, K.H. (1998) Extent of genomic rearrangement after genome duplication in yeast. *Proc. Natl. Acad. Sci. U. S. A.* 95, 4447–4452

0168-9525/$ - see front matter 
doi:10.1016/j.tig.2005.07.008

Letter

# Non-mammalian c-integrases are encoded by giant transposable elements

**Cédric Feschotte and Ellen J. Pritham**

Department of Biology, The University of Texas at Arlington, TX 76019, USA

In a recent report in *Trends in Genetics*, Gao and Voytas [1] described a new family of integrase genes that were identified in diverse eukaryotic species, including slime mold, *Caenorhabditis elegans*, *C. briggsae*, zebrafish, *Takifugu rubripes*, *Xiphophorus maculates*, cow, dog and humans. These genes potentially encode proteins containing ~400 amino acids with homology to retroviral integrases and transposases, including the integrase-like proteins encoded by Tlr, a family of atypical mobile elements with long terminal inverted repeats (TIRs) from the ciliate *Tetrahymena thermophila* [2,3].

Despite the similarity of c-integrases to the Tlr integrases and their presence in multiple copies in some genomes, Gao and Voytas found no evidence linking the c-integrase genes to retroelements [1]. The authors did not exclude the possibility that the c-integrases might be part of an unusual type of mobile element, but instead proposed that the c-integrases are 'host' genes. Two observations supported this hypothesis: (i) an excess of synonymous to nonsynonymous substitutions among c-integrase genes, indicative of purifying selection, and; (ii) the distant relationship of c-integrases to Fob1p, a protein from *Saccharomyces cerevisiae* involved in rDNA metabolism.

Two observations prompted us to investigate the origin of c-integrases. First, we noticed that the ESTs encoding *D. discoideum* c-integrases reported by Gao and Voytas displayed >99% nucleotide identity with the currently active mobile element Tdd-4 (GenBank accession number U57081). Tdd-4 elements essentially consist of the integrase gene flanked by ~125-bp TIRs, a structure reminiscent of DNA transposons [4]. Second, we were intrigued that they found two distinct pairs of c-integrases from *C. elegans* to be part of larger duplicated genomic regions flanked by large inverted-repeats (IRs; see Figure 3 in Ref. [1]). In light of the relationship between *D. discoideum* c-integrases and the TIR-containing Tdd-4 transposons, the presence of IRs flanking the *C. elegans* c-integrase genes might indicate that these genes are part of larger mobile elements.

A hallmark of mobile-element transposition is the duplication of a short genomic sequence at the site of

*Corresponding author:* Feschotte, C. (cedric@uta.edu).
Available online 9 August 2005