Minireview

Selection and gene duplication: a view from the genome

Andreas Wagner

Address: Department of Biology, University of New Mexico, 167A Castetter Hall, Albuquerque, NM 817131-1091, USA.
E-mail: wagnera@unm.edu

Published: 15 April 2002

Genome ***Biology*** 2002, **3(5)**:reviews1012.1–1012.3

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/5/reviews/1012

Abstract

Immediately after a gene duplication event, the duplicate genes have redundant functions. Is natural selection therefore completely relaxed after duplication? Does one gene evolve more rapidly than the other? Several recent genome-wide studies have suggested that duplicate genes are always under purifying selection and do not always evolve at the same rate.

When a gene duplication event occurs, the duplicate genes have redundant functions. Many deleterious mutations may then be harmless, because even if one gene suffers a mutation, the redundant gene copy can provide a back-up function. Put differently, after gene duplication - which can arise through polyploidization (whole-genome duplication), non-homologous recombination, or through the action of retrotransposons - one or both duplicates should experience relaxed selective constraints that result in elevated rates of evolution. This hypothesis originated as least as early as Ohno's seminal book [1], which emphasized the importance of gene duplications in organismal evolution. But for decades any test of the hypothesis had to rely on small numbers of gene duplicates; doubts thus remained over whether conclusions derived from such case studies were representative of all genes in a genome. This changed with the availability of complete genome sequences from multiple organisms. Such sequence information can address not only this question but also many others related to the influence of selection on gene families. For instance, does one duplicate evolve faster and thus acquire new functions more rapidly than the other? How frequent are beneficial mutations that generate new and advantageous functions? And how frequent is gene conversion of duplicate genes, in which recombination and DNA repair between very similar genes convert the sequence of one to that of the other?

To address such questions, one can use nucleotide alignments of duplicates to calculate two key parameters of molecular evolution [2]: the fractions per nucleotide site, first, of synonymous (silent) nucleotide substitutions, K_s, and second, of non-synonymous nucleotide substitutions (which change the encoded amino acid), K_a (see Box 1). The ratio K_a/K_s provides a measure of the selection pressure to which a gene pair is subject. If a duplicate gene pair shows a K_a/K_s ratio of about 1, that is, if amino-acid replacement substitutions occur at the same rate as synonymous substitutions, then few or no amino-acid replacement substitutions have been eliminated since the gene duplication. In other words, the duplicate genes are under few or no selective constraints. The gene pair is said to be under 'purifying selection' if $K_a/K_s < 1$: some replacement substitutions have been purged by natural selection, presumably because of their deleterious effects. The smaller the K_a/K_s ratio is, the greater the number of eliminated substitutions and the greater the selective constraint under which the two genes have evolved. The converse case, $K_a/K_s > 1$, indicates that replacement substitutions occur at a rate higher than expected by chance alone, so advantageous mutations have occurred in the evolution of the two duplicates.

Purifying or completely relaxed selection?

Two recent studies [3,4] analyzed these ratios in multiple fully sequenced and several partially sequenced genomes. The results are unequivocal: the vast majority of duplicate genes experience purifying selection. Even very closely related gene duplicates, no older than a few million years, experience selective constraints - the ratio K_a/K_s is smaller than one even in these cases. Recent duplicates appear to

Box 1

Key parameters in the evolution of duplicate genes

K_s	The fraction of synonymous (silent) nucleotide substitutions that occurred per synonymous DNA site since duplication
K_a	The fraction of non-synonymous (amino-acid replacement) nucleotide substitutions that occurred per non-synonymous DNA site since duplication
$K_a/K_s \approx 1$	**Neutral evolution**: an equal number of silent and amino-acid replacement substitutions have been preserved since duplication
$K_a/K_s < 1$	**Purifying selection**: more amino-acid replacement substitutions than silent substitutions have been eliminated since duplication, indicating that some amino-acid changes had deleterious effects
$K_a/K_s > 1$	**Positive, directional selection**: more amino-acid replacement substitutions than silent substitutions have been preserved, indicating an abundance of replacement substitutions that confer a selective advantage

tolerate more replacement amino-acid substitutions than older duplicates, however. For duplicates that differ at less than 5% of synonymous sites, between one in two and one in three substitutions are amino-acid replacement substitutions. For old duplicates, this number falls to between one in ten and one in twenty replacement substitutions [3]. But the variation across gene pairs is huge. Even a fine-grained statistical model that allows for differences in K_a/K_s among young and old duplicates may explain only 50% of the variance in evolutionary rates. In addition, there may be species-specific differences in K_a/K_s, but detection of such differences is sensitive to how information on gene duplicates is extracted from genomes and on how K_a and K_s are estimated. For example, one of the above studies [4] suggests that recent mammalian duplicates (K_a/K_s = 0.45 for genes with K_s between 0.05 and 0.5) appear to be under lower selective constraints than recent duplicates of *Drosophila melanogaster*, *Caenorhabditis elegans*, or *Arabidopsis thaliana*, where $K_a/K_s < 0.3$, whereas the other study [3] suggests no such differences.

To determine whether one duplicate evolves faster than the other, one can compare the sequences of both duplicates with that of a related but distant 'outgroup' gene and determine whether one duplicate has diverged to a greater extent than the other. The results may again depend on the organism studied. For example, in bacteria and mammals fewer than 10% of duplicates seem to evolve at different rates [4]. In contrast, a recent study focusing on ancient zebrafish duplicates - most of them developmental genes - found that about 50% of duplicates differ in their rates of evolution [5]. Despite such differences, these results show that it is not generally the case that one duplicate 'holds down the fort', and retains the original function while the other can evolve freely.

Gene conversion

Tandemly duplicated genes are known to be subject to gene conversion events that homogenize their sequences [6]. If rampant, gene conversion could substantially distort inferences of selection pressures after gene duplication. How prevalent is gene conversion for non-tandemly duplicated genes? Increasing amounts of sequence information prove helpful in answering this question as well. One group of genes with extremely slow rates of evolution, the histone H3 genes, has received recent attention in this regard [7]. With only three amino-acid differences between animal and plant histone H3 proteins, for example, histones are among the most highly conserved proteins. Does gene conversion contribute to their homogeneity? If so, one would expect that values of K_s between histone gene duplicates would be small - reflecting recent gene conversion - and not dramatically greater than values of K_a. But in organisms ranging from fungi to mammals, K_a and K_s differ by as much as a factor of 60 between non-tandemly clustered histone H3 genes [7], so evolution by gene conversion is unlikely to be frequent in this family. Another study [8] asked whether yeast (*Saccharomyces cerevisiae*) gene duplicates show evidence of gene conversion. Part of the assay in this study was based on the observation that measures of codon-usage bias are strongly correlated with the rate of synonymous divergence of yeast genes (because mutations in a highly expressed gene to a synonymous codon for which the respective transfer RNA is rare are deleterious). Only 4 out of 160 yeast duplicates had a synonymous divergence (K_s) less than expected on the basis of their codon-usage bias, showing that gene conversion is rare. In summary, although gene conversion is potentially rampant for some genes, it is most likely to be rare for the vast majority of genes.

Perhaps the most difficult questions about the influence of selection after gene duplication is how frequently beneficial mutations occur. Large amounts of genome sequence information lend themselves to the establishment of databases that document the gene families that have elevated K_a/K_s ratios [9]. Mere sequence analysis will probably have a limited impact on answering this question, however, because finding genes with $K_a/K_s > 1$ is usually not quite enough to make a case for positive selection. Although a

particular genome may contain many duplicates with K_a/K_s apparently above one, the observed difference from unity often does not withstand statistical scrutiny. Does this indicate the absence of positive selection after gene duplication? It does not, because positively selected amino-acid substitutions often occur only in a small region of the coding region, too small to be detectable by an elevated K_a/K_s ratio. And several case studies suggest the existence of positive selection for individual gene families, including the opsin visual pigments, primate ribonuclease genes, and triosephosphate isomerases [10-13]. These studies also show that a strong case for positive selection generally requires integration of information on gene divergence, phylogeny, and protein structure and function.

In summary, genome-scale surveys of gene duplication have the great merit of answering questions about molecular evolution without lingering doubts of statistical bias caused by small samples. They can assess to what extent selection is relaxed after gene duplication, to what extent gene duplicates diverge at different rates, and how abundant gene conversion events are. But their biggest strength - providing summary information about thousands of gene pairs - is also their biggest weakness. Some questions, such as the abundance of beneficial mutations, generally require more information than a crude view of the whole genome can provide. Genome-scale surveys thus draw our attention to their own limitations, which call for an integration of a variety of approaches to understand genome evolution.

References

1. Ohno S: *Evolution by gene duplication*. New York: Springer; 1970.
2. Li W-H: *Molecular Evolution*. Sunderland, MA: Sinauer; 1997.
3. Lynch M, Conery JS: **The evolutionary fate and consequences of duplicate genes.** *Science* 2000, **290:**1151-1155.
4. Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV: **Selection in the evolution of gene duplication.** *Genome Biol* 2002, **3:**research0008.1-0008.9.
5. Van de Peer Y, Taylor JS, 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.
6. Hillis DM, Dixon MT: **Ribosomal DNA: molecular evolution and phylogenetic inference.** *Q Rev Biol* 1991, **66:**410-453.
7. Rooney AP, Piontkivska H, Nei M: **Molecular evolution of the nontandemly repeated genes of the histone 3 multigene family.** *Mol Biol Evol* 2002, **19:**68-75.
8. Pal C, Papp B, Hurst LD: **Highly expressed genes in yeast evolve slowly.** *Genetics* 2001, **158:**927-931.
9. Liberles DA, Schreiber DR, Govindarajan S, Chamberlin SG, Benner SA: **The adaptive evolution database (TAED).** *Genome Biol* 2001, **2:**research0028.1-0028.9.
10. Merritt TJS, Quattro JM: **Evidence for a period of directional selection following gene duplication in a neurally expressed locus of triosephosphate isomerase.** *Genetics* 2001, **159:**689-697.
11. Briscoe AD: **Functional diversification of lepidopteran opsins following gene duplication.** *Mol Biol Evol* 2001, **18:**2270-2279.
12. Johnson ME , Viggiano L, Bailey JA, Abdul-Rauf M, Goodwin G, Rocchi M, Eichler EE: **Positive selection of a gene family during the emergence of humans and African apes.** *Nature* 2001, **413:**514-519.
13. Zhang JZ, Rosenberg HF, Nei M: **Positive Darwinian selection after gene duplication in primate ribonuclease genes.** *Proc Natl Acad Sci USA* 1998, **95:**3708-3713.