Minireview

Evaluating dosage compensation as a cause of duplicate gene retention in *Paramecium tetraurelia*

Timothy Hughes*, Diana Ekman[†], Himanshu Ardawatia*[‡], Arne Elofsson[†] and David A Liberles[‡]

Addresses: *Computational Biology Unit, Bergen Center for Computational Science, University of Bergen, 5020 Bergen, Norway. [†]Department of Biochemistry and Biophysics, Stockholm University, 10691 Stockholm, Sweden. [‡]Department of Molecular Biology, University of Wyoming, Laramie, WY 82071, USA.

Correspondence: David A Liberles. Email: liberles@uwyo.edu

Abstract

The high retention of duplicate genes in the genome of *Paramecium tetraurelia* has led to the hypothesis that most of the retained genes have persisted because of constraints due to gene dosage. This and other possible mechanisms are discussed in the light of expectations from population genetics and systems biology.

Many genomes display extensive gene duplication, which may result from either small-scale duplications or from duplication of the whole genome. What determines whether both copies of a duplicate gene are retained in the genome, and their subsequent evolutionary fate, is still a matter of debate. Aury *et al.* [1] have recently characterized gene duplication in the ciliate *Paramecium tetraurelia*, a unicellular eukaryote, which appears to have undergone multiple rounds of whole-genome duplication with a high level of retention of the duplicate copies. They suggest that this high level of retention is due to constraints arising from gene dosage, rather than other proposed mechanisms. Here we discuss these results in relation to the various models proposed for gene duplication and retention.

When duplication of a gene, or genome, occurs in an individual organism, it will only become part of the species genome if it becomes 'fixed' in the population (that is, becomes part of the genome of all members of the population). If the initial duplication event is evolutionarily neutral, the duplicated genes will become fixed in the population with a probability dependent on the inverse of the effective population size. It has been suggested, however, that the initial duplication event is likely to be deleterious for gene duplicates with functional regulatory regions, because of the metabolic cost of producing extra protein [2]. This would reduce the probability of fixation.

Given that fixation probably occurs much more quickly than the resolution of the fates of the duplicate copies, most work has considered fate determination as an independent step that occurs after the random process of fixation. Once fixation occurs, if there is purely neutral evolution at the protein level, one copy of a duplicated gene will quickly become a pseudogene, leaving a single ancestral copy with an ancestral function. While relaxation of selective constraint is generally thought to occur after gene duplication, negative selection, which discards changes, apparently returns quickly. Negative selection on parts of the gene may also be coupled to positive selection for the evolution of new functions or levels of expression. Relaxation of selective constraint (or a combination of negative and positive selection) that quickly gives way to stronger negative selection has been observed both in *Paramecium* [1] and in computer simulations of the evolution of gene duplicates [3].

Models that aim to explain the retention of duplicated genes include the subdivision of expression profiles or functions of the ancestral gene between the duplicates (subfunctionalization) [4]; the acquisition of new functions by one or both

duplicated copies (neofunctionalization) [5]; selection to increase robustness by maintaining a highly conserved back-up copy [6]; and selection for increased gene dosage or for dosage-compensation effects, as suggested for *Paramecium* (see also [7]).

Selection that depends on gene dosage can involve two different mechanisms. Selection for increased gene dosage involves a positive selection pressure to increase expression from a locus that is already highly expressed and has little mutational capacity to increase its expression or concentration-dependent activity. The dosage-compensation model, on the other hand, invokes a negative selection pressure to retain the function and expression levels of both copies in order to preserve the correct stoichiometry - the appropriate amounts or activity of the proteins in relation to each other or other proteins. Subfunctionalization is a nearly neutral model, with neither positive nor negative selection on gene function during the initial period of preservation, whereas neofunctionalization involves positive selection for the generation of new functions in the retained genes. Selection for redundancy, like that for dosage compensation, is characterized by negative selection. Several of these processes can act at different levels of biological regulation: for example, neofunctionalization and subfunctionalization can occur through changes in protein expression, changes in protein function, or changes in alternative or constitutive splicing. Dosage compensation, on the other hand, is a model in which conservation acts simultaneously on all of these processes.

Genome duplication favors the retention of duplicate genes

From examination of a variety of genomes, tandem and segmental gene duplications are known to occur at very high rates (on average 0.01 per gene per million years), similar in magnitude to the rate of mutation per nucleotide site [8,9]. Following such duplications, the average half-life of a gene copy is of the order of a few million years, with only a small fraction of duplicates surviving beyond a few tens of millions of years (TH and DAL, unpublished observations). Following whole-genome duplication, on the other hand, a large proportion of duplicate genes is retained after tens of millions of years (as in *Xenopus laevis* [10]) or even hundreds of millions of years (in teleost fish [11]). For teleost fish, the rate of retention has been reported to be much higher for the products of whole-genome duplication than for those of small-scale duplication [11].

One possible explanation for these differences is that gene fate is shaped by different evolutionary forces, depending on whether a gene is duplicated in a whole-genome event or not. In a whole-genome duplication, unlike a smaller-scale duplication, the entire network of interacting partners is duplicated together (Figure 1). It is unclear to what degree
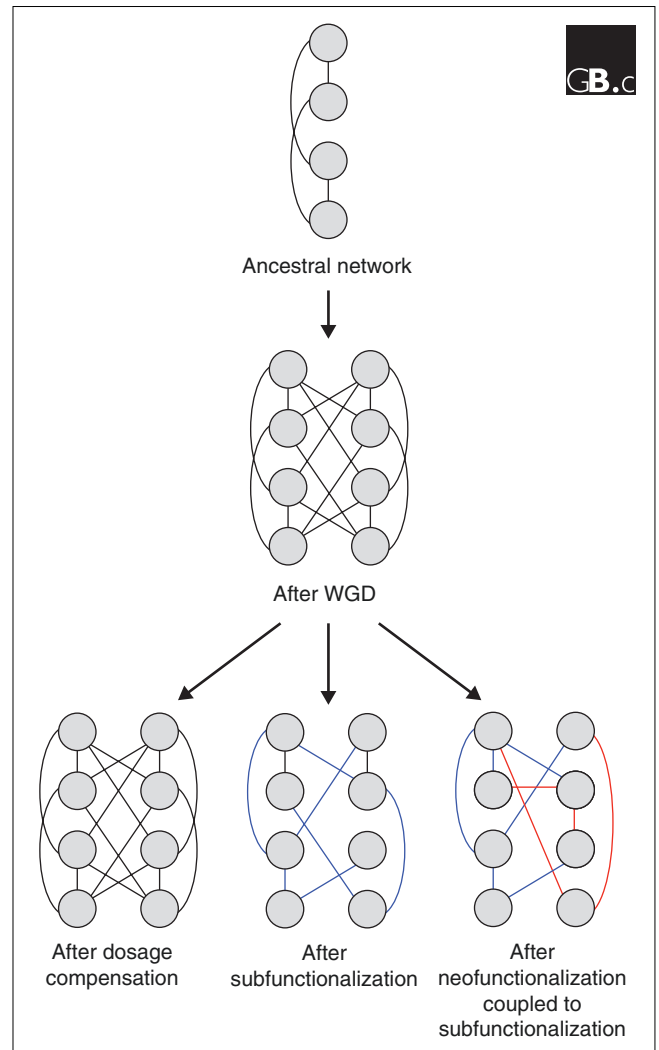

Figure 1
Possible outcomes for gene retention after whole-genome duplication. An ancestral network of interacting proteins is shown. Following a whole-genome duplication event, all of the proteins together with their interactions are duplicated. Over time, depending upon the evolutionary forces that are operating on the genome, different interactions are retained, gained or lost. Under the dosage-compensation model (bottom left), all interactions are retained. Under the subfunctionalization model (bottom center), redundant interactions become nonredundant (blue). When this is combined with the neofunctionalization model (bottom right), new interactions are also gained (red). In this figure, all of the duplicated copies have been retained as functional genes, but that is not the most likely outcome with increasing evolutionary time.

this build-up of pleiotropic constraints is a limitation as duplicates diverge, and this question needs to be addressed, potentially using protein structural models. The dosage-compensation model would predict that the build-up of pleiotropic constraint is difficult to resolve without deleterious effects, thus introducing a strong negative selection initially against the loss of genes or interactions. This would lead to gene retention and initial conservation of sequence and expression after whole-genome duplication.

Gene duplication in the *Paramecium* genome

With the sequencing of the genome of *P. tetraurelia* by Aury *et al.* [1], it was found to contain 39,642 genes, more genes than many other completely sequenced genomes. Furthermore, these genes can be grouped into families whose members are very closely related in sequence. Phylogenetic analysis of these gene families points to a recent whole-genome duplication in *P. tetraurelia*, in addition to several older genome duplications. The most recent duplication occurred long enough ago for negative selection to have set in, however.

Aury *et al.* [1] find that duplicate genes for signaling proteins and transcription factors are preferentially retained in the genome, as are duplicated genes for proteins known to form multicomponent complexes, with a positive correlation between retention and the number of components in the complex. A similar correlation between retention and complexity was observed for genes involved in metabolic pathways. More highly expressed genes were also more likely to have been retained.

Interestingly, the co-retained duplicates did not always originate from the same whole-genome duplication. In regard to complex-forming proteins, genes that were co-retained after the most recent whole-genome duplication were not found to be those preferentially retained in the older duplications. In all, Aury *et al.* [1] found that patterns of retention across whole-genome duplications were affected by gene function, and showed a preference for retention of duplicated genes that had not retained a duplicate in an older whole-genome duplication.

The authors conclude that dosage compensation to maintain the stoichiometry of protein complexes and metabolic pathways and keep them functioning correctly plays an important part in the retention of duplicate genes after a whole-genome duplication. From consideration of the traces of the preceding whole-genome duplications they also propose that over time there is a slow progressive loss of duplicates, as gene-expression levels become adapted for stoichiometric reasons, for example.

The dosage-compensation model predicts that duplicates of genes for proteins that do not form complexes or do not have concentration-dependent roles in metabolism will be rapidly lost. In the case of duplicated genes encoding interacting proteins, it predicts strong selection for retention, but if one of the interacting duplicates is lost from the genome, the model predicts that the loss of the remaining duplicate will now be positively selected for. The first part of this prediction is qualitatively satisfied by the observations from the *P. tetraurelia* genome of the retention of genes for complex-forming proteins. On the other hand, the retention patterns and differing profiles of nonsynonymous (K_a) and synonymous (K_s) substitutions (K_a/K_s profiles) for duplicates of different ages do not seem to support dosage compensation

as the driving force for keeping them in the genome. Selection as a result of dosage compensation thus appears to be complex and may have a role in modulating other evolutionary mechanisms. The apparent burst of either positive selection or relaxation of selective constraint in the period shortly after genome duplication implies that selective mechanisms other than dosage compensation are also acting.

Following the most recent whole-genome duplication in *P. tetraurelia*, species radiation occurred, resulting in the *P. tetraurelia* complex of 15 sibling species. Aury *et al.* [1] propose that this burst of speciation is a side-effect of the whole-genome duplication, occurring as a result of differential gene loss in different populations, leading to inviable hybrids and reproductive isolation by Dobzhansky-Muller incompatibility [12]. Such a proposition is consistent with the loss of proteins not under dosage-balance constraint under the dosage-compensation model and in our opinion is most consistent with speciation accompanied by neo-functionalization or subfunctionalization.

In evaluating alternative explanations of the retention profiles for duplicates in the paramecium genome, effective population size may be an important consideration. Effective population size (together with mutation rate) as a modulator of the strength of selection has been implicated as an important switch between subfunctionalization as a purely neutral process and neofunctionalization or, potentially, dosage compensation as mechanisms involving selection [4,8,9]. *Paramecium* has been shown to have a relatively large effective population size, making mechanisms that involve selection possible [13]. However, it has been shown that binding interactions as well as regulatory modules can subfunctionalize in the preservation of duplicate genes [3,14], and so the subfunctionalization model for gene duplicate retention may also be consistent with a dependence on the number of interacting protein partners, where the probability of subfunctionalization might be expected to be proportional to the number of ways of subfunctionalizing the interactions with partners. This is a different mechanism of gene retention from dosage compensation, but this characteristic of subfunctionalization has not been evaluated to show that it has the same potential to retain duplicate genes in such high numbers as dosage compensation appears to be able to do. Eventually, quantitative models characterizing these various processes can be tested against the data to extend our understanding of the process of gene retention.

Where does dosage compensation fit in?

Dosage compensation may indeed affect the short-term retention rate of duplicate genes after whole-genome duplication. Over longer time frames, however, proteins involved in complexes and pathways are not preferentially retained in the duplicate pairs originating from whole-genome duplications, neither in *P. tetraurelia*, as indicated

by Aury *et al.* [1], nor in yeast [15] (except for ribosomal proteins [16]). In fact, whereas 17% of highly connected proteins (hubs) in the yeast protein-protein interaction network belong to a pair originating from the relatively ancient whole-genome duplication that has occurred in *Saccharomyces cerevisiae*, only 5% of the party hubs, which are coexpressed with their interaction partners, are part of such a pair [15]. Homologous complexes in yeast appear to have been created through stepwise partial duplications and not through whole-genome duplication [17].

The results of Aury *et al.* [1] do suggest that after more recent whole-genome duplication events, the duplicate proteins belonging to complexes and pathways are initially retained to a greater extent than other proteins. According to this view, although dosage sensitivity is not sufficient for the long-term fixation of duplicates in the genome, it may be important in the first phase following the whole-genome duplication. One might postulate dosage compensation as a mechanism for holding duplicated genes in the genome for some time, to give an opportunity for eventual neofunctionalization (as has been suggested for subfunctionalization [3]). However, even in the period immediately following duplication, stoichiometric issues will be dependent on the interplay between expression and sequence as well as selective pressures for concentration dictated by metabolism and systems-level constraints. Further modeling work is needed to understand the mechanism, as the suggestions by Aury *et al.* [1] and alternative suggestions (such as subfunctionalization of binding interactions) are part of an ongoing synthesis to understand the process of gene duplication and its relationship to the evolution of gene function.

Considering the case of metabolic networks, the patterns of retention or modification have been observed to be influenced by network structure, topology and function, and the positioning of duplicate genes at key points in the network. Genes coding for enzymes involved in directing higher metabolic fluxes are subject to greater evolutionary constraints as a gene duplication event would increase the flux through an enzyme-catalyzed reaction. It has been observed in *S. cerevisiae* that genes encoding highly connected enzymes in metabolic pathways have a higher likelihood of maintaining duplicates [18]. Thus, duplication of genes encoding enzymes carrying high metabolic fluxes are more likely to be retained compared to genes encoding enzymes carrying lower metabolic fluxes.

Enzymes in a pathway can evolve with different functional requirements, which can lead to mismatches in the enzyme activities upon duplication [19]. This means that upregulation of individual enzymes can increase or decrease the flux capacity of the pathway and by different amounts. Hence, if only certain proteins increase the performance of the pathway, the duplicates of the other proteins in the pathway will not provide extra fitness to the organism. This also has

implications for the retention of duplicate copies based upon an entire pathway being duplicated, indicating that the negative selective pressure for retention of each duplicate in a pathway would not be equally strong. Interestingly, it has been argued that the neutral expectation for biological networks involves a more complex network than that minimally required for function, without necessarily invoking robustness as a driving force for this non-minimal network [20].

The findings by Aury *et al.* [1] lend further support to the idea that dosage compensation can play a role in the retention of duplicated genes in a genome. Whole-genome duplication events in additional lineages representing different time points will enable a fuller testing of this and other hypotheses, as well as their functional implications for systems biology.

References

1. Aury J-M, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Ségurens B, Daubin V, Anthouard V, Aiach N, *et al.*: **Global trends of whole genome duplications revealed by the ciliate *Paramecium tetraurelia*.** *Nature* 2006, **444:**171-178.
2. Wagner A: **Energy constraints on the evolution of gene expression.** *Mol Biol Evol* 2005, **22:**1365-1374.
3. Rastogi S, Liberles DA: **Subfunctionalization of duplicated genes as a transition state to neofunctionalization.** *BMC Evol Biol* 2005, **5:**28.
4. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J: **Preservation of duplicate genes by complementary, degenerative mutations.** *Genetics* 1999, **151:**1531-1545.
5. Ohno S: *Evolution by Gene Duplication.* New York: Springer-Verlag, 1970.
6. Kuepfer L, Sauer U, Blank LM: **Metabolic functions of duplicate genes in *Saccharomyces cerevisiae*.** *Genome Res* 2005, **15:**1421-1430.
7. Withers M, Wernisch L, dos Reis M: **Archaeology and evolution of transfer RNA genes in the *Escherichia coli* genome.** *RNA* 2006, **12:**933-942.
8. Lynch M, Conery JS: **The evolutionary fate and consequences of duplicate genes.** *Science* 2000, **290:**1151- 1155.
9. Lynch M, Conery JS: **The origins of genome complexity.** *Science* 2003, **302:**1401-1404.
10. Hughes MK, Hughes AL: **Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*.** *Mol Biol Evol* 1993, **10:**1360-1369.
11. Blomme T, Vandepoele K, de Bodt S, Simillion C, Maere S, van de Peer Y: **The gain and loss of genes during 600 million years of vertebrate evolution.** *Genome Biol* 2006, **7:**R43.
12. Orr HA: **Dobzhansky, Bateson, and the genetics of speciation.** *Genetics* 1996, **144:**1331-1335.
13. Snoke MS, Berendonk TU, Barth D, Lynch M: **Large global effective population sizes in *Paramecium*.** *Mol Biol Evol* 2006, **23:** 2474-2479.
14. Braun FN, Liberles DA: **Retention of enzyme gene duplicates by subfunctionalization.** *Int J Biol Macromol* 2003, **33:**19-22.
15. Ekman D, Light S, Bjorkman AK, Elofsson A: **What properties characterize the hub proteins of the protein-protein interaction network of *Saccharomyces cerevisiae*?** *Genome Biol* 2006, **7:**R45.
16. Papp B, Pal C, Hurst LD: **Dosage sensitivity and the evolution of gene families in yeast.** *Nature* 2003, **424:**194-197.
17. Pereira-Leal JB, Teichmann SA: **Novel specificities emerge by stepwise duplication of functional modules.** *Genome Res* 2005, **15:**552-559.
18. Vitkup D, Kharchenko P, Wagner A: **Influence of metabolic network structure and function on enzyme evolution.** *Genome Biol* 2006, **7:**R39.
19. Salvador A, Savageau MA: **Evolution of enzymes in a series is driven by dissimilar functional demands.** *Proc Natl Acad Sci USA* 2006, **103:**2226-2231.
20. Soyer OS, Bonhoeffer S: **Evolution of complexity in signaling pathways.** *Proc Natl Acad Sci USA* 2006, **103:**16337-16342.