

Vertebrate evolution: doubling and shuffling with a full deck

Dannie Durand

Department of Biological Sciences, Carnegie Mellon University, 4400 Fifth Avenue, Pittsburgh, PA 15213, USA

The number and role of whole-genome duplications in vertebrate evolution has intrigued evolutionary biologists since Ohno first proposed genome duplication as the force driving the 'big leap' in vertebrate morphological innovation. Attempts to resolve these issues have been thwarted by small and noisy datasets, and by lack of computational accuracy and statistical rigor. Recently, Ken Wolfe and colleagues presented a genome-scale, statistically rigorous analysis of evidence based on the spatial organization of duplicated genes, as well as estimates of duplication times. Their results provide the strongest evidence to date of large-scale duplication throughout the vertebrate genome, consistent with at least one whole-genome duplication.

The vertebrate lineage is characterized by a doubling in gene number and dramatic developmental innovation in a relatively short period of time (Fig. 1). What was responsible for this rapid and substantial increase in gene number? And, how is this increase related to the functional innovation and developmental diversity seen in vertebrates? In 1970, Ohno [1] proposed that whole-genome duplication (polyploidization) provides the raw material from which new genes can evolve *en masse*, fueling major transitions in organismal evolution. Contemporary statements of Ohno's hypothesis focus on the '2R' hypothesis: that is, that two rounds of polyploidization, followed by loss of some genes and functional differentiation of others, occurred early in the vertebrate lineage, driving the evolution of developmental complexity in vertebrates.

Since the early 1990s, the 2R hypothesis has been much debated [2–5], and despite the increasing availability of sequence data, it remains controversial. The results of numerous studies ([6,7] and work surveyed in [2,4,5,8]), some in support of, and others at variance with, 2R have been inconclusive owing to the limited amount of data available, sensitivity of computational methods to noisy data, and lack of methodological rigor. Worst of all, there is lack of agreement about the predictions the 2R hypothesis generates and what constitutes a rigorous test of those predictions.

Testable predictions of 2R

Proposed testable predictions of the 2R hypothesis fall into two categories, spatial and temporal. Spatial analyses are based on the contention that local similarities in gene content should still be discernible in the modern genome, despite extensive rearrangement and gene loss following polyploidization. However, CONSERVED SYNTENY (see Glossary) between PARALOGOUS GENES, presented in early studies, is not constitute rigorous proof of polyploidization [4]. More recent studies [4,7,8] report pairs of paralogous gene pairs. These paralogs are typically not in the same order in both regions, and are interspersed with unrelated genes. If such regions are to be taken as evidence of polyploidization, it is necessary to show that they could not have arisen through a series of independent duplications of individual genes, yet many reports include no statistical analysis. Randomization testing has been used in several recent large-scale studies, and formal statistical tests for putative duplicated regions are beginning to emerge [8].

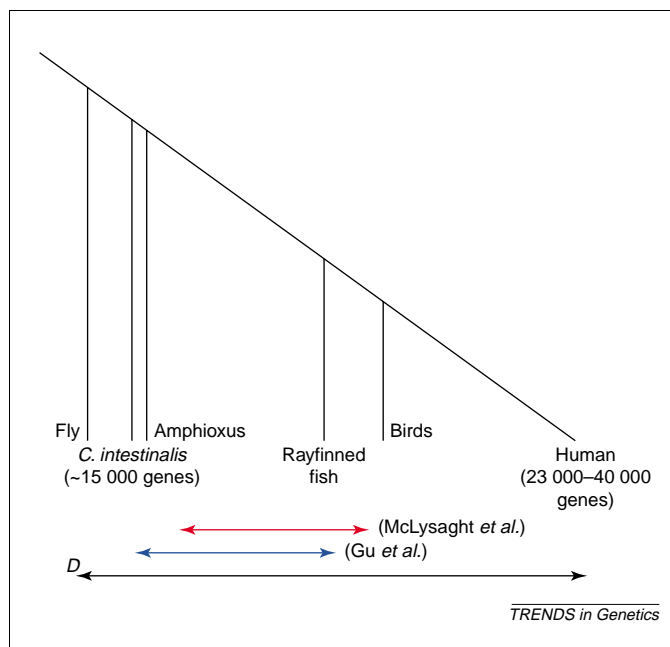


Fig. 1. Evolution of the chordate lineage with fly as an outgroup. The chordate invertebrate *Ciona intestinalis* has roughly 15 000 genes [21], whereas current estimates of human gene number range from 23 000 to 40 000 [6,7]. Conserved synteny between human and fish suggests a roughly twofold increase in gene number occurred before the separation of fish and tetrapods [22,23]. Estimates of D , the time since the fly–vertebrate split, range from 833 Myr [24] to 993 Myr [25]. The excess of gene duplications in the interval $0.4D$ – $0.7D$ (red arrow) observed by McLysaght *et al.* [14] is consistent with the report by Gu *et al.* [9] of a wave of duplications between the *Amphioxus*–vertebrate split and the emergence of bony fish (blue arrow).

Glossary

Conserved synten: In a comparative map, the co-occurrence of homologous genes on the same chromosome in both genomes.

Gene conversion: Nonreciprocal recombination resulting in identical sub-sequences on both chromosomes.

Molecular clock: The hypothesis that the accepted mutation rate is roughly constant in all lineages.

Monte Carlo significance testing: An approach to hypothesis testing in which the distribution of the test statistic under the null hypothesis is simulated using randomization.

Paralogous genes: Genes within the same species that originated through duplication of an ancestral gene.

Substitution rate correction: Estimation of the expected number of mutations at each site from observed mutations, correcting for multiple substitutions at the same site.

Tandem duplication: Duplication due to unequal crossing over resulting in paralogs that are in close proximity to each other on the chromosome.

Temporal predictions of 2R are based on estimates of gene duplication times and gene family tree topologies. A first prediction of 2R is that estimated gene duplication times should show two distinct waves early in the vertebrate lineage [9]. Second, it has been proposed [10] that gene families in a 2R genome should exhibit an excess of (AB)(CD) topologies (Fig. 2(a)). Many phylogenetic analyses have not yielded a preponderance of (AB)(CD) topologies, contradicting 2R [2,10]. However, various authors argue that other topologies can also be consistent with 2R, depending on the mechanics and timing of rediploidization [5,11]. Furthermore, if a cluster of paralogs that arose through earlier TANDEM DUPLICATION were duplicated in a round of polyploidization and subsequently

different members of the tandem cluster were lost in each of the resultant blocks, topologies other than (AB)(CD) could result [12].

Temporal analyses of 2R, whether based on time estimates or tree reconstruction, face significant methodological challenges. Both approaches rely on the 'molecular clock' assumption that mutations accumulate at an approximately constant rate in all lineages. At the best of times this assumption must be approached with skepticism, because of mutational saturation and differences in mutation rates under varying selective conditions. It is particularly problematic with gene duplication data [3,5,12]. GENE CONVERSION will lead to homogenization, especially in tandem duplicates, and selective constraints on recently duplicated genes are reduced, leading to an increase in accepted mutations [13]. In tree reconstruction, such rapid mutation can cause 'long branch attraction', erroneously grouping rapidly evolving, but otherwise unrelated, taxa together.

Computational genome scale analysis

In the absence of a complete vertebrate genome sequence, previous studies relied on partial datasets. Now that the whole-genome sequence for human is available, can 2R be resolved? The careful study presented by McLysaght *et al.* [14], based on a comprehensive spatial and temporal approach, addresses many of the problems that plagued earlier analyses. McLysaght *et al.* use MONTE CARLO SIGNIFICANCE TESTING to validate the spatial results, and they validate the temporal evidence by estimating gene duplication times using two different methods.

There are three major steps involved in a genome-scale computational analysis of gene duplication: (1) detecting putative duplicated genes; (2) identifying candidate duplicated regions; and (3) constructing gene families trees and estimating gene duplication times. McLysaght *et al.* constructed an initial set of candidate paralogs with significant similarity using sequence comparison. For the purposes of investigating the 2R hypothesis, the paralogs of interest are genes that arose through duplication of an entire gene early in the chordate lineage. To eliminate ancient duplications as well as genes that share a domain, but not a common ancestor, McLysaght *et al.* required that candidate pairs have a minimum alignment length and be more similar to each other than to the closest invertebrate ancestor. Candidate pairs were also eliminated if they were drawn from families that were too diverse in sequence similarity or too large (≥ 20 members). Putative tandem clusters, defined in [14] as paralogs on the same chromosome separated by at most 30 intervening genes, were collapsed by replacing the cluster by its longest representative.

Candidate duplicated regions were inferred from this set of paralogs using gene positions from the Golden Path assembly of 12/2000. Although there is general agreement that local similarities in gene content and order suggest a history of large-scale duplication [8], there is no consensus as to which definitions for candidate paralogous regions are most appropriate. In their study, McLysaght *et al.* searched for 'paralogons', pairs of regions with a set of paralogs in common. In each region, adjacent paralogs

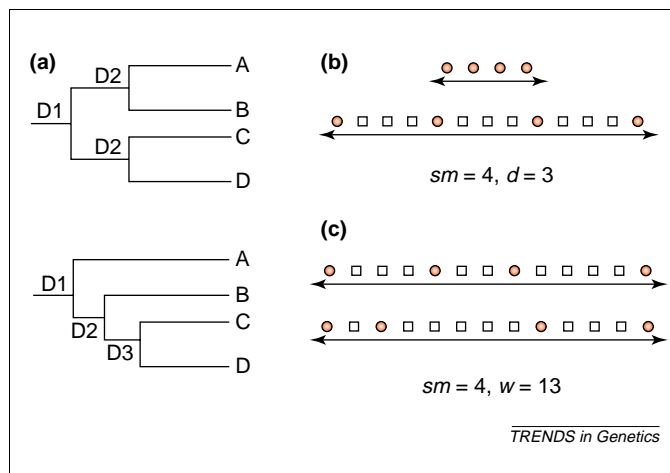


Fig. 2. Testable predictions of 2R. (a) Two possible topologies for a gene family with four vertebrate members, A, B, C and D. Hughes [10] has argued that the (AB)(CD) topology (upper tree) is consistent with two rounds of polyploidization, whereas the lower tree (A(B(CD))) can only be explained by three distinct duplications. Other authors [3,5,11,12] have countered that the lower topology could be recovered from a 2R genome under several scenarios including earlier tandem duplication followed by loss, major disparities in substitution rates, or certain modes of polyploidization and rediploidization. (b) Conserved duplicated regions containing paralogs (red circles) with intervening unduplicated genes (white squares). In [14], paralogons are defined in terms of two parameters: the number of shared paralogs (sm) and the maximum number of interlopers between any pair of adjacent paralogs (d). Under this definition, paralogons with the same parameter values can vary greatly in compactness. Thus, a paralogon with $sm = 4$ and $d = 3$ can be found in a window as small as four or as large as 13. In the statistical analysis, both paralogons are treated identically, yet the former is much less likely to be a chance occurrence than the latter. (c) This problem can be avoided by parameterizing conserved regions using paralog count (sm) and window size (w) instead of gap length (d). These regions, both containing $sm = 4$ genes in a window of size $w = 13$, are comparable, whereas the regions in (b) are not.

were separated by no more than d intervening genes. The parameter value $d = 30$ was chosen empirically. McLysaght *et al.* reported the number of paralogons observed as a function of the number of paralogs they contained (sm). Significance was determined using randomization tests by comparing, for a given value of sm , the number of observed paralogons with the average number of paralogons obtained by applying the same algorithm to shuffled data. From these results, McLysaght *et al.* concluded that the human genome was generated by a pattern of large-scale duplications and that 'any paralogon with $sm \geq 6$ was very likely to have been formed by a single duplication of a chromosomal region'.

Although the former conclusion is well supported by the data, the statement that a particular cluster is significant requires further analysis, because paralogons with the same number of paralogs can vary greatly in sparsity (Fig. 2(b)). Because McLysaght *et al.* do not report the range of actual window sizes in which sm paralogs were observed, it is possible that the distribution of window sizes differed in the observed and shuffled data. Statistics based on the total number of paralogons observed can be used to reject global null hypotheses, such as that the entire genome evolved through independent duplication of individual genes, but this definition is not appropriate for testing the significance of individual clusters. An alternative definition in which both the window size and the number of paralogs found in the window are fixed [7,8] is better suited for systematic studies.

To estimate gene duplication times, McLysaght *et al.* constructed a set of non-overlapping gene family trees with trusted invertebrate outgroups, each containing one fly, one worm and two to ten human sequences. Families for which an accurate SUBSTITUTION RATE CORRECTION [15] could not be obtained or that did not satisfy a MOLECULAR CLOCK test [16] were eliminated from the dataset, as were trees in which the two invertebrate outgroups were not grouped together. Time estimates were obtained from the remaining trees. For two-gene families, these estimates were validated by comparing them with upper and lower bounds on the time of duplication obtained by constructing trees with additional sequences from other vertebrate species.

New results and remaining challenges

The picture that emerges is of widespread, large-scale duplication, be it polyploidy, aneuploidy (chromosomal duplication) or many sub-chromosomal duplications, followed by substantial gene loss. These conclusions are supported by both the spatial and the temporal evidence. Forty-four percent of the genome was covered by 96 paralogons with $sm \geq 6$, and a majority of the genes analyzed were duplicated during the period $0.4D-0.7D$, where D is the time since the fly-vertebrate split. This is consistent with a wave of gene duplication dating from roughly the same period (Fig. 1) observed by Gu *et al.* [9] in their study of 749 vertebrate gene families. Neither the spatial nor the temporal evidence in [14] provides strong support for two rounds of polyploidization.

McLysaght *et al.* present a compelling, well-constructed case, particularly impressive given the age of the events

under study and the degree of genomic rearrangement since the advent of bony fish. Nevertheless, problems still abound. The human genome sequence is noisy, incomplete and suffers from difficulties in gene prediction and whole-genome sequence assembly [17,18]. To obtain the quality exhibited in their analysis, McLysaght *et al.* were obliged to impose extremely stringent requirements on the data included in their study, reducing the number of genes used in their final analysis to a fraction of the genome.

This situation imposes a difficult tradeoff between methodological rigor and potential insight. Only 191 of 758 initial gene families were considered to be robust enough to include in their final analysis of duplication times. Unfortunately, the remaining data were insufficient to estimate the age of individual paralogons by combining estimates of duplication times with gene clusters. And McLysaght *et al.* did not rule out the possibility that the exclusion of these families introduced a systematic bias. However, Gu *et al.* examined the gene families excluded from their study [9] and found no evidence of bias.

The presence of large families of multidomain proteins in higher eukaryotes poses a major difficulty for identifying paralogs. McLysaght *et al.* avoided this problem by eliminating large, diverse families from the analysis. However, recent comparisons of eukaryote proteomes [6] suggest that the vertebrate genomes are characterized by rapid expansion of multidomain families with roles in immune response, regulation and the cytoskeleton, indicating the potential importance of these families in vertebrate developmental and morphological innovations. Thus, the price of stringency is an analysis that could fail to capture a significant aspect of vertebrate evolution.

Can 2R be resolved?

The results presented by McLysaght *et al.*, as well as those of other genome-scale studies [6,7,9], are consistent with at least one round of polyploidization, but could also be explained by many smaller block duplications and do not strongly support 2R. Resolution of the alternative hypotheses will be helped by more sequence data of greater accuracy, including better assembly and gene prediction for the human genome, additional vertebrate genomes, additional invertebrate outgroup genomes and genomes for intermediate species such as *Amphioxus* or tunicate fish. Currently, we cannot guess to what extent the human, and fly and worm genomes are typical of vertebrate and invertebrate genomes, respectively. Algorithms and statistical methods that support automated large-scale analysis and integrate different types of spatial and temporal evidence are also needed.

Finally, a better understanding of the fate of duplicated genes will contribute to the resolution of 2R. Much of the disagreement in this area stems from a lack of consensus over the null and alternative hypotheses [4,5,10-12]: what are the unique properties of an ancient 2R genome? The type and frequency of gene loss and functional differentiation, as well as the interplay of large-scale duplication with tandem duplication and retrotransposition, greatly influences the properties of the null hypothesis we must reject to confirm 2R.

Beyond 2R

Although recent studies have focused on 2R, Ohno's seminal work [1] also presented a broader vision of whole-genome duplication as a force that could enable the leap in morphological and developmental complexity seen in modern vertebrates. Since 1970, genomics has revealed several other mechanisms that promote diversity of protein function on a genomic scale, including alternative splicing [19] and domain shuffling [6]. Widespread segmental duplication and domain accretion (recently reported by Eichler *et al.* [20]) could be evidence of ongoing gene formation through domain shuffling in the modern genome. The challenge for the future is to forge a comprehensive view of the interplay of all of the forces that drive vertebrate evolution.

Acknowledgements

I thank C. Beshers, L. Hurst, E.W. Jones, A.J. Lopez and D. Sankoff for helpful comments. I was supported by NIH grant 1 K22 HG 02451-01 and a David and Lucille Packard Foundation fellowship.

References

- Ohno, S. (1970) *Evolution by Gene Duplication*, Springer
- Makalowski, W. (2001) Are we polyploids? A brief history of one hypothesis. *Genome Res.* 11, 667–670
- Taylor, J.S. and Brinkmann, H. (2001) 2R or not 2R? *Trends Genet.* 17, 488–489
- Skrabaneck, L. and Wolfe, K.H. (1998) Eukaryote genome duplication – where's the evidence? *Curr. Opin. Genet. Dev.* 8, 559–565
- Wolfe, K.H. (2001) Yesterday's polyploids and the mystery of diploidization. *Nat. Rev. Genet.* 2, 33–41
- International Human Genome Sequencing Consortium, (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
- Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351
- Durand, D. and Sankoff, D. (2002) Tests for gene clustering. In *RECOMB2002, Sixth Annual International Conference on Computational Molecular Biology* (Myers, G. *et al.*, eds), pp. 144–154, ACM Press
- Gu, X. *et al.* (2002) Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. *Nat. Genet.* 31, 205–209
- Hughes, A.L. (1999) Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history. *J. Mol. Evol.* 48, 565–576
- Gibson, T.J. and Spring, J. (2000) Evidence in favour of ancient octaploidy in the vertebrate genome. *Biochem. Soc. Trans.* 2, 259–264
- Smith, N.G.C. *et al.* (1999) Vertebrate genome evolution: a slow shuffle or a big bang? *BioEssays* 21, 697–703
- Lynch, M. and Conery, J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290, 1151–1155
- McLysaght, A. *et al.* (2002) Extensive genomic duplication during early chordate evolution. *Nat. Genet.* 31, 200–204
- Gu, X. and Zhang, J. (1997) A simple method for estimating the parameter of substitution rate variation among sites. *Mol. Biol. Evol.* 14, 1106–1113
- Takezaki, N. *et al.* (1995) Phylogenetic test of the molecular clock and linearized trees. *Mol. Biol. Evol.* 12, 823–833
- Eichler, E.E. (2001) Segmental duplications: what's missing, misassembled, and misassembled – and should we care? *Genome Res.* 5, 653–656
- Li, S. *et al.* (2002) Comparative analysis of human genome assemblies reveals genome-level differences. *Genomics* 80, 138–139
- Modrek, B. and Lee, C. (2002) A genomic view of alternative splicing. *Nat. Genet.* 30, 13–19
- Eichler, E.E. (2001) Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet.* 17, 661–669
- Simmen, M.W. *et al.* (1998) Gene number in an invertebrate chordate, *Ciona intestinalis*. *Proc. Natl Acad. Sci. USA* 95, 4437–4440
- McLysaght, A. *et al.* (2000) Estimation of synteny conservation and genome compaction between pufferfish (Fugu) and human. *Yeast* 17, 22–36
- Postlethwaite, J.H. *et al.* (2000) Zebrafish comparative genomics and the origins of vertebrate chromosomes. *Genome Res.* 10, 1890–1902
- Nei, M. *et al.* (2001) Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. *Proc. Natl Acad. Sci. USA* 98, 2497–2502
- Wang, D.Y. *et al.* (1999) Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. *Proc. R. Soc. Lond. Ser. B* 266, 163–171

The genes of the Green Revolution

Peter Hedden

Long Ashton Research Station, Department of Agricultural Sciences, University of Bristol, Long Ashton, Bristol BS41 9AF, UK

The spectacular increases in wheat and rice yields during the 'Green Revolution', were enabled by the introduction of dwarfing traits into the plants. Now, identification of the genes responsible for these traits shows that they interfere with the action or production of the gibberellin (GA) plant hormones. We knew that the wheat *Rht* genes encode growth repressors that are normally suppressed by GA, and recent work shows that the rice *sd1* gene encodes a defective enzyme in the GA-biosynthetic pathway.

In the past 40 years the population of the world has doubled to more than 6.1 billion people. Fortunately, this increase has been more than matched by increases in global cereal production so, despite gloomy predictions of impending famine [1], world-wide food production per capita is higher than it was in 1960 [2]. A major factor that enabled these impressive yield increases was the introduction of high-yielding varieties of wheat and rice, in combination with the application of large amounts of fertilizer and pesticides. The impact of these advances was termed the 'Green Revolution'.

The introduction of dwarfing genes into cereal crops was crucial to this revolution. The stems of tall wheat and rice plants were not strong enough to support the heavy

Corresponding author: Peter Hedden (peter.hedden@bbsrc.ac.uk).