

LBNL-57436: Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate

Paramvir Dehal¹ & Jeffrey L. Boore^{1,2}

¹ Evolutionary Genomics Department

DOE Joint Genome Institute and Lawrence Berkeley National Laboratory

2800 Mitchell Drive

Walnut Creek, CA 94598 USA

² Department of Integrative Biology

3060 Valley Life Sciences Building

University of California

Berkeley, CA 94720 USA

Keywords: Evolution, Genome duplication, Comparative genomics, Vertebrates, Gene clustering

Corresponding authors: Jeffrey Boore, DoE Joint Genome Institute, 2800 Mitchell Drive,

Walnut Creek, CA 94598, phone: 925-296-5691, fax: 925-296-5620, JLBoore@LBL.gov

Paramvir Dehal, DoE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA

94598, phone: 925-296-2855, PSDehal@LBL.gov

Summary

The hypothesis that the relatively large and complex vertebrate genome was created by two ancient, whole genome duplications has been hotly debated, but remains unresolved. We reconstructed the evolutionary relationships of all gene families from the complete gene sets of a tunicate, fish, mouse, and human, then determined when each gene duplicated relative to the evolutionary tree of the organisms. We confirmed the results of earlier studies that there remains little signal of these events in numbers of duplicated genes, gene tree topology, or the number of genes per multigene family. However, when we plotted the genomic map positions of only the subset of paralogous genes that were duplicated prior to the fish-tetrapod split, their global physical organization provides unmistakable evidence of two distinct genome duplication events early in vertebrate evolution indicated by clear patterns of 4-way paralogous regions covering a large part of the human genome. Our results highlight the potential for these large-scale genomic events to have driven the evolutionary success of the vertebrate lineage.

Introduction

It has long been hypothesized that the increased complexity and genome size of vertebrates has resulted from two rounds (2R) of whole genome duplication (WGD) occurring in early vertebrate evolution, thus providing the requisite raw materials [1]. This seemed to be supported by the long-standing speculation that humans have about 100,000 genes, roughly four times the number expected for invertebrates' genomes, but this is now known to be incorrect, with the actual human gene count being closer to 30,000 [2,3]. Conflicting analyses have now made this very controversial, with some

studies supporting 2R (e.g. [4-8]), others seeing only a single round of WGD (e.g. [9-11]), and still others refuting WGD altogether by concluding that nothing greater than limited segmental duplications have occurred (e.g. [12,13]).

The 2R hypothesis had been bolstered by observations that a few gene families, e.g. Hox clusters [14], follow a “4:1 rule” in the numbers of vertebrate to invertebrate genes. However, comparison of the complete genome sequences of human [2,3] and *Drosophila* [15] revealed that less than 5% of homologous gene families follow the 4:1 rule [12]. Further, although two sequential duplications are expected to generate the evolutionary topology (AB)(CD) for the descendent genes, rather than (A)(BCD), in fact, the relationships of vertebrate multigene families do not generally show this pattern, as indicated by early studies using only a few genes [16] and confirmed as complete genome sequences became available [2,13]. (However, for a different view, see [17].) Several studies have incorporated data from sparse sampling of genes from taxa thought to have branched near to these purported duplications, including lamprey [18], hagfish, amphioxus [17,19-22] and *Ciona* [23]; while these results are useful for timing duplications, the conclusions could never be viewed as definitively resolving this issue, since these products could have alternatively been generated by duplications of individual genes or short gene segments rather than by whole genome duplications. Even duplicating all of the genes in a genome individually is quite different from a whole genome duplicating simultaneously.

There are several reasons why this has been a difficult issue to resolve. After duplication, only the minority of gene pairs will adopt a new function (“neofunctionalization”) or partition old functions (“subfunctionalization”) quickly

enough to escape disabling mutations that would lead to their eradication [24]; therefore, rampant gene loss rapidly erases this signal of genome duplication. Further, four-member gene families, even those with the (AB)(CD) topology, can be generated by two rounds of duplications of individual genes or of segments much smaller than the entire genome, generating a condition that cannot be differentiated on this basis from two rounds of WGD followed by many gene losses. This alternative scenario seems especially plausible since recent analyses have shown that gene duplications occur much more frequently than had been thought, with the typical rate being sufficient to duplicate an entire genome equivalent every 100 million years [25,26]. Until recently, no complete genome sequence has been available from an outgroup that is closely related to vertebrates and all methods of phylogenetic reconstructions are less accurate with more distant relatives such as *Drosophila* and yeast [20]. Lastly, there has not been to date a method to accurately and comprehensively cluster genes into homologous families, since methods that rely on sequence similarity alone are highly subject to artifactual association of slowly evolving paralogs and to erroneous exclusion of the more rapidly evolving genes.

Fortunately, as has been shown convincingly for the yeast genome and for *Arabidopsis* [27-30], evidence of an ancient genome duplication can be seen in the large-scale pattern of the physical locations of homologous genes, even when the great majority of the duplicated genes have been lost. Studies have shown that the human genome also has multiple regions of colinear paralogous gene copies [4,21,22,31-37], but considered the arrangements of only too small a number of genes and genomic regions to be comprehensive. This approach is now available for a large scale evaluation of the vertebrate 2R hypothesis, since complete (i.e. at least draft quality) genome sequences are

available for the tunicate *Ciona intestinalis* [38] (a basal chordate outgroup), and the vertebrates *Takifugu rubripes* [39] (a pufferfish “fugu”), *Mus musculus* [40] (mouse) and human [2,3]. Figure 1 illustrates how the signal of two rounds of genome duplication could be retained by the large scale pattern in location of duplicated genes, where many tracks of paralogous duplicates (which may not contain identical subsets of genes) each occur at exactly four positions in the genome, i.e. “tetra-paralogons”. No similar signal would be generated by repeated duplications of genes or even large gene segments; only whole genome duplications would result in such global organization of paralogous genes.

Results

Gene Clustering and Duplication Timing

A graph-based method was used with the complete gene sets of the four chordates (98,517 total genes; see Table 1 for details of each step in the analysis) to generate clusters such that each contains all, and only, those genes that descended from a single gene in their common ancestor (Figure 2). A multiple sequence alignment and a maximum likelihood evolutionary tree was constructed for each cluster, then a web browser interface was built so that each can be viewed individually. (For more details and updates that include more taxa, see the “PhIGs” [Phylogenetically Inferred Groups] website at <http://phigs.org/>.) We could then easily determine when each gene duplicated relative to lineage splitting by comparing these gene trees with the known evolutionary relationships of the animals. For example, a gene duplication that is specific to only one animal’s lineage is seen as two genes from the same genome clustered together. A gene that duplicated once in the unique common ancestor of mouse and

human would generate a tree that groups gene copy 1 of human and mouse and, separately, gene copy 2 of human and mouse. Put more generally, gene duplications that are shared by more than one species are seen as a replication of the phylogeny of the descendant organisms for each gene copy. Of course, gene losses and various combinations of these processes are seen as well. Figure 3 shows all possible gene topologies along with how each would be interpreted.

This reveals that 46.6% of the ancestral chordate genes appear in duplicate in one or more of the vertebrate lineages, with 34.5% having at least one duplication before the divergence of fish from tetrapods and 23.5% having at least one duplication afterward. (Some of these are counted twice, having had duplications both before and after the fish-tetrapod split.) This means that there are 3,753 gene duplications placed at the base of Vertebrata, which is remarkable, since the ancestral genome would be reasonably estimated to have had less than 20,000 genes, as is the case for the tunicate as well as other invertebrate outgroups. However, as can be seen in Figure 4, gene duplications are in large numbers on every branch of the tree, making it unclear whether this, in itself, indicates a significant acceleration in duplication rate. Additionally, of the gene clusters with duplications basal to the fish-tetrapod split, 20.6% have had one duplication event, 10.8% have had two, and 5.1% have had more than two, counter to the expectation from 2R, and casting further doubt on the significance of this for the 2R hypothesis.

Gene family membership

An early observation in support of 2R was that several gene families have expanded from a single member in invertebrates to having four members for some vertebrates.

Previous studies have shown that this is not generally true for vertebrate multigene families [12], which is confirmed in this analysis. As can be seen in Figure S1, there is no peak at four for gene family membership for any vertebrate. In fact, even gene duplications do not predominate; for each vertebrate species considered individually, one member per cluster is the largest category, accounting for 55%, 57% and 59% of the fugu, mouse, and human genes, respectively, with 53.4% of the gene clusters having no duplication events whatsoever. Thus, there is no signal of 2R remaining in gene family membership, despite early anecdotal observations to the contrary.

Determination of concordantly duplicated regions

To test the extent to which the 3,753 early duplication events that are timed to the base of Vertebrata were generated as part of larger scale, multigene duplications, we examined the relative positions of these resulting paralogs in the human genome (which is currently the best assembled and annotated vertebrate genome). These results are visualized in Figure 5 and more comprehensively in Figure S2, where the linear array of genes for each chromosome is used to query for paralogs generated by any duplication event prior to the fish-tetrapod split. It is apparent from these figures that there is a large-scale pattern of genome segments that are concordant in having similar arrangements of paralogous genes. In order to quantify this, we performed a sliding window analysis, considering whether there are two or more early-duplicating genes found in each subject chromosome that are within 100 genes of each other that are paralogous to genes found both 50 genes upstream and 50 genes downstream of each query gene. There is a distinct pattern of having multiple chromosomes matching with long linear stretches of

paralogous genes. This indicates that these duplications occurred in very large segments, consistent with the hypothesis of whole genome duplication(s). Having matches to three other chromosomal segments is the dominant category, as can be seen by the darker coloring in Figures 5 and S2 and in the histogram of Figure 6. These patterns, with each genomic region corresponding in gene arrangement to sets of paralogs in three other genomic segments, is strong support for the 2R hypothesis.

Although the four-fold (i.e., including the query segment) category is the most prevalent, it accounts for only 25% of the genome. Nonetheless, it is striking that this remains the largest category despite approximately 450 million years of evolution. This constitutes a strong signal of two rounds of whole genome duplication, and could not reasonably have been generated by a series of smaller duplication events. For the latter to have generated this pattern, multiple duplications would have to have occurred of the same region (or its resulting duplicates) three times, and have done so for many regions throughout the genome. We would expect, rather, that independent, random duplications would follow a Poisson distribution; this contrasting situation is seen when the same analysis is done with all human gene paralogs generated by duplication after the split of fish and tetrapods (not shown). Even if we were to consider the alternative of a single WGD followed by subsequent independent, large segmental duplications, it would be difficult to explain why these would have been predominantly two-fold for previously duplicated regions. The most parsimonious explanation for the observed pattern can only be two rounds of WGD.

Tetra-paralogue Detection

To further establish 2R, we evaluated these sets of paralogs for whether this four-fold matching indicates that they fall into tetra-paralogons, as illustrated in Figure 1. We formalized this by first identifying paralogue (paralogous genomic segments) containing the same set of at least two duplicated gene pairs, while allowing for a maximum of 100 unduplicated genes in between (similar to the approach in [10]). (The allowance of 100 genes is arbitrary, but the results are not critically dependent on this number, which is only used to find the blocks of paralogous genes.) We infer that duplicated genes in paralogue are likely to have arisen from a single duplication involving all contained, duplicated genes, and that the unique, intervening genes have resulted from differential gene deletions and subsequent genome rearrangements.

We identified 2,953 paralogous gene pairs in human that are inferred to have resulted from 1,912 genes that duplicated prior to the divergence of the fish and tetrapod lineages (with some gene losses also). Of these paralogous genes, 32.4% are still in 386 detectable paralogue comprising 772 individual genomic segments, containing from two to 42 gene pairs (Table S1). Of these 772 genomic segments, 454 comprise tetra-paralogue (Figures 7A and S2, Table 2) as shown hypothetically in Figure 1, where overlapping sets of paralogs fall into fourfold groups. (Unfortunately, it was not possible for us to evaluate the hypothesis of an additional genome duplication unique to ray-finned fish [41,42] because of the generally poor contiguity of the fugu draft assembly.)

In contrast, when looking at the gene pairs that arose from a duplication event after the divergence of the fish and mammal lineages (Figure 4), we find only 11% are detected in paralogue in the human genome, indicating that these duplications have less

commonly included large segments of the genome (Figure 7B). This is especially interesting in that their relative recency would make it more likely that any large duplications would remain detectable, reinforcing the contrast with the large scale structure of those earlier duplications. By looking specifically for tandemly duplicated genes by defining them as paralogs on the same chromosome that are separated by fewer than 10 intervening genes, we can recognize that 50% of these human gene pairs arose from tandem duplication, compared with 6% for the human gene pairs that arose before the divergence of the fish and tetrapod lineages.

Discussion

No detectible signal of WGD exists in the analysis of gene family membership. There is no peak at four genes per family for any of the vertebrates (Figure S1) as might result from 2R. Presumably this results from a great number of subsequent gene losses that have erased this signal. Likewise, the phylogenetic timing of the duplication events is also inconclusive, since duplications are common on every branch (Figure 4). Although there is a somewhat greater number assigned to the base of vertebrates, there is no reliable way to evaluate the significance of this. In fact, even if this larger number could be found to be statistically significant, it may simply indicate that this was a period with an accelerated duplication of individual genes or multigene segments or a reduction in the rate of gene loss, rather than indicating WGD.

Conclusive evidence for 2R is seen only when data from gene families, phylogenetic trees, and genomic map position are all taken together, as has been advocated by others [21,32,43]. When examining the genomic map position of only those

genes in the human genome which trace their ancestry back to a duplication event at the base of vertebrates, a clear pattern of tetra-paralogons emerges, indicating that two rounds of WGD occurred at the base of vertebrates. This signal remains most clearly in 25% of the human genome that forms the largest category in the analysis shown in Figures 5 and 6, but we also find that 72% of all human genes are included in the total extent of all of the paralogons that overlap with these regions, providing the least constrained estimate of the portion of the human genome still retaining structure from the two rounds of WGD. This is the outside estimate, because some portion could have as well been the result of segmental duplications of regions earlier established by WGD. This is in contrast to the pattern seen for the many other gene duplications, which generated paralogs that are predominantly arranged in tandem.

This is particularly compelling considering that this signal has survived more than 450 MY of genome rearrangements and the loss of many genes. We can imagine the effect that duplications, translocations, inversions, and deletions (and combinations thereof) would have had on this analysis: (1) Duplications would cause an increase beyond the four-fold category. (2) Translocations would decrease the four-fold category if they are pervasive enough to clear large regions of paralogs. (3) Inversions can either cause a decrease in the number of chromosomes hit by moving paralogous genes beyond the detection of the sliding window analysis or cause an increase by spreading some paralogous genes across the boundaries into adjacent segments; this can be exacerbated by gene translocations that blur the edges of the corresponding regions. (4) Deletions would generally increase the three-fold chromosome category at the expense of the four-fold category and a deletion that occurred between the two whole genome duplications

would increase the two-fold chromosome category. Additionally, in some cases, a few individual gene deletions or translocations may have eliminated the links between pairs of duplicated genes. Through these, and combinations of these events, the original four-fold colinearity established by two rounds of whole genome duplication (or something less than the perfect four-fold pattern, if these duplications were long separated) has been eroded.

These tetra-paralogons are spread across nearly all human chromosomes (Table 2). Notably, chromosome Y does not have any tetra-paralogons, perhaps due to its relatively recent origin and small number of genes, or perhaps this indicates a more rapid rate of gene movement. Chromosome 21 also has no tetra-paralogons and chromosome 18 has only one that is small. These chromosomes, and other regions without tetra-paralogons, could be of recent origin or they could have undergone multiple rearrangement events that would have destroyed the signal.

Although our study does not specifically address the effect that 2R of WGD has had on vertebrate evolution, we note two interesting observations. First, the vast majority of duplicated genes were subsequently deleted, indicating that relatively few genes may have been responsible for the increased complexity seen in vertebrates. Second, it is possible that many genes were loosed from constraint after the genome duplications and experienced an accelerated rate of sequence change before returning to single copy, and it is possible that this has played some role in the evolution of vertebrate complexity [44].

The mechanism of these genome duplication events, whether two separate rounds of either auto- or allo-tetraploidy or a single octoploidy, remains uncertain. We speculate that the most likely scenario is two rounds of closely spaced auto-tetraploidization events,

based on the following observations. For most sets of tetra-paralogs, some pairs within the set extend over a longer region than others, indicating two distinct duplication events. If, alternatively, there had been a single octoploidy, then we would have to hypothesize multiple occasions where two of the four descendant genomic segments lost the same sets of genes independently, which seems unlikely. The phylogenetic trees for the gene families are not consistently nested, as would be expected in the case of allo-tetraploidy or two widely spaced auto-tetraploidy events. Finally, tree topologies of genes within paralogy blocks are not always congruent, indicating that the process of gene loss and rediploidization spanned the duplication events [17].

It remains unclear to what extent such large scale genomic events have driven macroevolutionary change vs. the regular accumulation of small mutations, as is the central tenet of the classical model of evolution. We imagine that rapid and extensive evolutionary change could possibly be an emergent property of having all genes duplicated at the same time, allowing this expanded gene repertoire to evolve together, and so reaching a greater level of interaction and complexity than could evolve from cumulative single gene duplications. Whole genome duplications have occurred in many lineages, including frogs [45,46], fish [41,42,47], yeast [27-30], Arabidopsis [27-30], and corn and several other crop species [48], all of which are being studied by modern genomics techniques. We view the broad and pervasive distribution of these tetra-paralogons in the human genome, despite the remarkably small number of genes remaining in duplicate, as robust evidence that two rounds of whole genome duplication occurred at the base of Vertebrata, and anticipate that future studies will soon illuminate the roles this has played in the evolutionary success of the vertebrate lineage.

Materials and Methods

Obtaining chordate sequences. Sequences and gene annotations of the tunicate *Ciona intestinalis* and the pufferfish *Takifugu rubripes* were obtained from the DOE Joint Genome Institute website at <http://www.jgi.doe.gov>. Sequences for *Homo sapiens* (version 19.34b.2) and *Mus musculus* (version 19.32.2) were obtained from the Ensembl project website at <http://www.ensembl.org>. For genes with multiple transcripts, only the longest protein sequence was taken, resulting in 15,852 *Ciona*, 37,241 *fugu*, 22,444 mouse, and 22,980 human genes. Table 1 shows an overview of the methods along with the numbers of genes and clusters included after each step.

Clustering. The objective of the gene clustering is to reconstruct groups of genes such that each includes all (and only) the descendants of a single gene in the ancestral chordate. The underlying assumption made is that all of the vertebrate genes in such a cluster will have a higher degree of similarity to each other than they will for their ortholog in *Ciona*, since they have arisen after the *Ciona*-vertebrate divergence by either gene duplication or lineage splitting. We conceptually translated the protein sequences for all genes, then for each *Ciona* protein sequence, the best match to any vertebrate protein was found using BLASTP [49]. Likewise, for each vertebrate protein, the best *Ciona* match was found. This list of best *Ciona*-vertebrate hits was then ordered by raw score. A graph was constructed such that each protein sequence appears at a node and the raw BLASTP scores between each pair form the weight of each edge. These sequences were then grouped by using the pairs of best hits as seeds for a single linkage clustering of the graph with the minimum edge score of the seed. This recruits to each cluster any

sequence with greater similarity to the individual seed sequences than they have to each other, ensuring that genes with similarity due to a duplication before the split of Ciona and Vertebrata are properly apportioned into separate clusters. Any cluster that attempts to use a protein that has already been assigned is eliminated to reduce ambiguity and any cluster with greater than 100 members in a single species is eliminated. Figure 2 illustrates this clustering process.

Phylogenetic Analysis. A multiple sequence alignment for each cluster was created using ClustalW 1.81 [50]. This alignment was then trimmed by eliminating all positions with gap characters. If the remaining length of the multiple sequence alignment was less than 100 amino acids the entire cluster was eliminated. Phylogenetic trees were constructed by using the quartet puzzling maximum likelihood method as implemented in TREE-PUZZLE 5.1 [51] using the JTT model of amino acid substitution and a gamma distribution of rates over eight rate categories with 10,000 puzzling steps used to assess reliability. Any tree whose nodes were not strictly bifurcating was eliminated. Even with strict requirements for membership in the clusters, for reliable sequence alignment, and for confidence of evolutionary analysis, we generated 6,641 gene family clusters that include 39,136 (39.7%) of 98,517 total chordate genes (Table 1), of which 3,096 had duplicated vertebrate genes and 1,621 produced trees that are strictly bifurcating (i.e. having no polytomies).

Identification of node types. Each node of each tree was classified in comparison to the known evolutionary relationships of the animals. For example, if the gene cluster tree

contains exactly four members, and one from each animal, then the parsimonious inference is that no gene duplication occurred. In the case of a similar cluster, but where one member is missing, this is a gene loss in a single group. Gene cluster trees can show duplications specific to individual lineages by having two genes clustered together for the same animal. A gene duplication that occurred before the split of fish from tetrapods is seen as a duplicated tree of the animal relationships after their splitting from Ciona and, similarly, a gene duplication that occurred in tetrapods but before the split of the mouse and human lineages is seen as a duplication of the mouse-human group. Combinations of these, such as a duplication for tetrapods followed by a loss in one of the tetrapod lineages, are also seen and scored (Figure 3).

The sorting of orthologous and paralogous relationships for each gene cluster provides an effective tool for improving the inferences of gene function by allowing annotations from well studied genomes to be transferred to the orthologous genes of other species. Inferring function from orthology is expected to be more accurate than using sequence similarity alone, since the latter tends to incorrectly associate slowly evolving paralogs. We provide a web based resource for this sorting at <http://phigs.org/>.

Detecting concordantly duplicated genomic regions. An overview of the genomic distribution and patterns of paralogous gene arrangement was created by plotting the chromosomal location for all genes having a duplication before the fish-tetrapod divergence for each chromosome. We performed a sliding window analysis, looking upstream and downstream of each gene in turn, for 50 genes on each side, and asking whether paralogs generated prior to the fish-tetrapod split occur within 100 genes of each

other at another genomic location. This is the most conservative approach for detecting this signal. The results are visualized in Figures 5 and S2.

Detecting paralogs and tetra-paralogs. A paralogon was defined as two chromosomal locations in the same genome that have the same set of gene pairs, allowing for a maximum of 100 unduplicated genes between. This significantly expands the set of regions that can be detected by the sliding window analysis. These were detected for the entire human genome considering separately only those paralogs generated by duplications prior to the split between fish and tetrapods versus those arising from duplications after. Each paralogon was tested for its membership in additional paralogous-region pairs. When a segment pairs with three different paralogs, we considered all six possible pairings of the four regions. If, and only if, all six possible combinations are confirmed as paralogs, the group was defined as a tetra-paralogon (Figure 1). A minimum reconstruction of the signal of 2R remaining in the human genome is found in the extent of complete fourfold overlap and the maximum by extending these regions to include the complete extent of all contributing paralogs. Genome and chromosome coverage were calculated by summing the number of genes that are encompassed by this more expansive reconstruction and dividing by the total number of genes.

Inclusion of *Drosophila* genes and phylogenetic analysis. For each cluster that contained only a single gene from each of the four chordate species, the highest scoring BLASTP match to the *Drosophila melanogaster* gene set [15] was added. We then

performed a multiple sequence alignment for each of these 766 sets of genes, followed by phylogenetic analysis of this concatenated data set as above. The resulting tree was rooted at the midpoint with branch lengths proportional to amount of amino acid substitution as estimated by TREE-PUZZLE 5.1 [51].

Abbreviations

WGD, whole genome duplication

2R, two rounds of whole genome duplication

MY, million years

References

1. Ohno S (1970) *Evolution by Gene Duplication*. Berlin: Springer-Verlag. 160 pp.
2. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. *Nature* 409: 860-921.
3. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. *Science* 291: 1304-1351.
4. Lundin LG (1993) Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. *Genomics* 16: 1-19.
5. Meyer A, 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.
6. Spring J (1997) Vertebrate evolution by interspecific hybridisation--are we polyploid? *FEBS Lett* 400: 2-8.
7. Wang Y, Gu X (2000) Evolutionary patterns of gene families generated in the early stage of vertebrates. *J Mol Evol* 51: 88-96.
8. Larhammar D, Lundin LG, Hallbook F (2002) The human Hox-bearing chromosome regions did arise by block or chromosome (or even genome) duplications. *Genome Res* 12: 1910-1920.
9. Guigo R, Muchnik I, Smith TF (1996) Reconstruction of ancient molecular phylogeny. *Mol Phylogenet Evol* 6: 189-213.
10. McLysaght A, Hokamp K, Wolfe KH (2002) Extensive genomic duplication during early chordate evolution. *Nat Genet* 31: 200-204.

11. Gu X, Wang Y, Gu J (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.
12. Friedman R, Hughes AL (2003) The temporal distribution of gene duplication events in a set of highly conserved human gene families. *Mol Biol Evol* 20: 154-161.
13. Friedman R, Hughes AL (2001) Pattern and timing of gene duplication in animal genomes. *Genome Res* 11: 1842-1847.
14. Popovici C, Leveugle M, Birnbaum D, Coulier F (2001) Homeobox gene clusters and the human paralogy map. *FEBS Lett* 491: 237-242.
15. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, et al. (2000) The genome sequence of *Drosophila melanogaster*. *Science* 287: 2185-2195.
16. Hughes AL (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.
17. Furlong RF, Holland PW (2002) Were vertebrates octoploid? *Philos Trans R Soc Lond B Biol Sci* 357: 531-544.
18. Escriva H, Manzon L, Youson J, Laudet V (2002) Analysis of lamprey and hagfish genes reveals a complex history of gene duplications during early vertebrate evolution. *Mol Biol Evol* 19: 1440-1450.
19. Holland PW, Garcia-Fernandez J, Williams NA, Sidow A (1994) Gene duplications and the origins of vertebrate development. *Dev Suppl*: 125-133.

20. Holland PW (2003) More genes in vertebrates? *J Struct Funct Genomics* 3: 75-84.
21. Abi-Rached L, Gilles A, Shiina T, Pontarotti P, Inoko H (2002) Evidence of en bloc duplication in vertebrate genomes. *Nat Genet* 31: 100-105.
22. Panopoulou G, Hennig S, Groth D, Krause A, Poustka AJ, et al. (2003) New evidence for genome-wide duplications at the origin of vertebrates using an amphioxus gene set and completed animal genomes. *Genome Res* 13: 1056-1066.
23. Leveugle M, Prat K, Popovici C, Birnbaum D, Coulier F (2004) Phylogenetic analysis of *Ciona intestinalis* gene superfamilies supports the hypothesis of successive gene expansions. *J Mol Evol* 58: 168-181.
24. Wolfe KH (2001) Yesterday's polyploids and the mystery of diploidization. *Nat Rev Genet* 2: 333-341.
25. Lynch M, O'Hely M, Walsh B, Force A (2001) The probability of preservation of a newly arisen gene duplicate. *Genetics* 159: 1789-1804.
26. Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290: 1151-1155.
27. Wong S, Butler G, Wolfe KH (2002) Gene order evolution and paleopolyploidy in hemiascomycete yeasts. *Proc Natl Acad Sci U S A* 99: 9272-9277.
28. Vision TJ, Brown DG, Tanksley SD (2000) The origins of genomic duplications in *Arabidopsis*. *Science* 290: 2114-2117.

29. Kellis M, Birren BW, Lander ES (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* 428: 617-624.
30. Dietrich FS, Voegeli S, Brachat S, Lerch A, Gates K, et al. (2004) The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science* 304: 304-307.
31. Katsanis N, Fitzgibbon J, Fisher EM (1996) Paralogy mapping: identification of a region in the human MHC triplicated onto human chromosomes 1 and 9 allows the prediction and isolation of novel PBX and NOTCH loci. *Genomics* 35: 101-108.
32. Pebusque MJ, Coulier F, Birnbaum D, Pontarotti P (1998) Ancient large-scale genome duplications: phylogenetic and linkage analyses shed light on chordate genome evolution. *Mol Biol Evol* 15: 1145-1159.
33. Gibson TJ, Spring J (2000) Evidence in favour of ancient octaploidy in the vertebrate genome. *Biochem Soc Trans* 28: 259-264.
34. Vienne A, Rasmussen J, Abi-Rached L, Pontarotti P, Gilles A (2003) Systematic phylogenomic evidence of en bloc duplication of the ancestral 8p11.21-8p21.3-like region. *Mol Biol Evol* 20: 1290-1298.
35. Luke GN, Castro LF, McLay K, Bird C, Coulson A, et al. (2003) Dispersal of NK homeobox gene clusters in amphioxus and humans. *Proc Natl Acad Sci USA* 100: 5292-5295.
36. Castro LF, Furlong RF, Holland PW (2004) An antecedent of the MHC-linked genomic region in amphioxus. *Immunogenetics* 55: 782-784.

37. Castro LF, Holland PW (2003) Chromosomal mapping of ANTP class homeobox genes in amphioxus: piecing together ancestral genomes. *Evol Dev* 5: 459-465.
38. Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B, et al. (2002) The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science* 298: 2157-2167.
39. Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, et al. (2002) Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* 297: 1301-1310.
40. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420: 520-562.
41. Van de Peer Y, Taylor JS, Meyer A (2003) Are all fishes ancient polyploids? *J Struct Funct Genomics* 3: 65-73.
42. Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, et al. (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* 431: 946-957.
43. Horton AC, Mahadevan NR, Ruvinsky I, Gibson-Brown JJ (2003) Phylogenetic analyses alone are insufficient to determine whether genome duplication(s) occurred during early vertebrate evolution. *J Exp Zool B Mol Dev Evol* 299: 41-53.

44. Seoighe C, Johnston CR, Shields DC (2003) Significantly different patterns of amino acid replacement after gene duplication as compared to after speciation. *Mol Biol Evol* 20: 484-490.
45. Tymowska J, Fischberg M, Tinsley RC (1977) The karyotype of the tetraploid species *Xenopus vestitus* Laurent (Anura: pipidae). *Cytogenet Cell Genet* 19: 344-354.
46. Jeffreys AJ, Wilson V, Wood D, Simons JP, Kay RM, et al. (1980) Linkage of adult alpha- and beta-globin genes in *X. laevis* and gene duplication by tetraploidization. *Cell* 21: 555-564.
47. Taylor JS, Van de Peer Y, Braasch I, Meyer A (2001) Comparative genomics provides evidence for an ancient genome duplication event in fish. *Philos Trans R Soc Lond B Biol Sci* 356: 1661-1679.
48. Blanc G, Wolfe KH (2004) Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* 16: 1667-1678.
49. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215: 403-410.
50. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22: 4673-4680.

51. Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18: 502-504.

Acknowledgements

We thank R. Baker, M. P. Francino, M. Medina, J. Schwarz, and Y. Valles for helpful comments on the manuscript and S. Rash and W. Huang for technical assistance. This work was performed under the auspices of the U.S. Department of Energy, Office of Biological and Environmental Research, by the University of California, Lawrence Berkeley National Laboratory, under contract No. DE-AC02-05CH11231.

Conflicts of interest. The authors declare that they have no conflicting interests.

Author contributions. PD and JB conceived of the design. PD wrote the code and performed the analyses. PD and JB wrote the paper.

Figure Legends

Figure 1. Pattern predicted for the relative locations of paralogous genes from two genome duplications

(A) Representation of a hypothetical genome that has 22 genes shown as colored squares.

(B) A genome duplication generates a complete set of paralogs in identical order. (C)

Many paralogous genes suffer disabling mutations, become pseudogenes, and are then lost. One could imagine this condition being evidence of a single round of genome

duplication followed by significant gene losses. (D) A second genome duplication

recreates another set of paralogs in identical order, with multigene families that retained two copies now present in four, and those that had lost a member now present in two

copies. (E) Again, many paralogous genes suffer disabling mutations, become

pseudogenes, and are then lost. Of course, unrelated gene duplications and transpositions

can occur. Even though this leaves only a few four-member gene families, the patterns of

two- and three-fold gene families unite, in various combinations, all four genomic

segments, revealing that the sequential duplications had been of very large regions, in this

case all or nearly all of this hypothetical genome.

Figure 2. Overview of the building of a gene cluster and phylogenetic tree shown by a

hypothetical example. (A) Each circle represents a gene, labeled with the source genome

according to the first letter of each taxon – C, M, H, and F for Ciona, mouse, human, and

fugu, respectively – and further differentiated by numeral. BLASTP was first used to

search all vertebrate genes for the one most similar to Ciona's C1 gene, in this case the

mouse gene M1. Then other genes are recruited to the cluster if they have a higher similarity score to M1 than that between C1 and M1, indicated here by the red lines. The six genes shown on the right side of the diagram have some sequence similarity to those in the cluster, but less than that between C1 and M1, so are not included. Since the vertebrates are more closely related to each other than any is to Ciona, each cluster will include those genes descended from a single gene in the common chordate ancestor, having arisen by either lineage splitting or gene duplication specific to one or more vertebrates. (See Materials and Methods for more details.) (B) Evolutionary tree of the genes in this cluster show separate duplications for fugu and for human. Because the maximum likelihood method does not rely solely on sequence similarity, there is no significance to the mouse gene being most similar to C1. The mouse genome simply contained the most slowly evolving vertebrate gene in this multigene family; this can be from any vertebrate taxon with approximately equal likelihood.

Figure 3. Hypothetical phylogenetic tree showing all possible types of gene relationships and how they are most parsimoniously interpreted.

Interior nodes are designated in lower case for those that simply result from lineage splitting and in upper case for gene duplications within a lineage according to: fts = Fish-Tetrapod Split, prs = Primate-Rodent Split, HD = Human Duplication, MD = Mouse Duplication, FD = Fugu Duplication, DBPRS = Duplication Before Primate-Rodent Split, and DBFTS = Duplication Before Fish-Tetrapod Split. Although not shown, nodes are still scored if there is gene loss. Phylogenetic trees for each gene family can be viewed at

<http://phigs.org/>, also providing a valuable tool for improving the inference of gene function.

Figure 4. Phylogenetic analysis of the four chordates with *Drosophila* as an outgroup

This phylogenetic tree is based on 766 concatenated single copy protein sequences totaling 313,797 amino acid positions with branches proportional to the amount of change. Numerals in bold above the branches indicate the number of gene duplications occurring in each lineage; numerals below indicate branch lengths.

Figure 5. Plot of the genomic positions of paralogous pairs of human genes that arose from duplications pre-dating the fish-tetrapod split

The queries shown here use chromosomes 2, 4, 5, and 10, as indicated for the four panels. (The complete set can be seen in Figure S2.) On the X-axis is each chromosome arranged from p- to q-telomeres. On the Y-axis is each of the 22 human autosomes plus the X and Y chromosomes. For each query gene on the X-axis a “hit” is scored if the subject chromosome contains a paralog generated by a gene duplication prior to the fish-tetrapod split. The lower portion of each panel plots the n-fold redundancy along the query chromosome as defined by pairs of paralogs detected in a sliding window analysis. See the Material and Methods section for details, but briefly, for every human query gene, a window was considered of 50 genes to the left and 50 genes to the right, with a “hit” obtained for the subject chromosome if it includes the early-duplicated paralogs of genes on each side of the query. Four-fold (i.e., including the query) matching, as expected by the 2R hypothesis, is highlighted in a darker shade of blue.

Figure 6. Histogram showing the lower bound estimate of N-fold redundancy using the analysis reported in Figure 5

This histogram is generated by counting the depth of paralogon redundancy across all human chromosomes as shown in the lower part of Figure S2 (and subsampled for Figure 5). The peak at four-fold coverage is consistent with the 2R hypothesis, and constitutes a lower bound estimate, since the sliding window examines only a small span of flanking genes and would be highly subject to effects of local gene rearrangements.

Figure 7. An arbitrarily selected subset of the human genome showing the physical relationships among paralogous genes

(A) This is an example of the tetra-paralogous relationships of a subset of human genes that are all inferred, by gene trees, to have duplicated prior to the split of fish from tetrapods, but after the split of Ciona from vertebrates. These genes are on four chromosomes with their identities indicated outside of the circle. The complete set of tetra-paralogons can be viewed in Figure S3. **(B)** In contrast, paralogous human genes generated by duplications after the split of fish and tetrapods, as shown for this sample of the same four human chromosomes, do not form such tetra-paralogons. Their pattern appears to result from smaller scale tandem duplications of individual genes or segments, followed by slow rearrangements. In addition to these apparently functional gene pairs shown in the figure for this portion of the human genome, we have identified eight pseudogene pairs that occur on different chromosomes; it is not clear whether these pseudogenes are the result of random retrotransposition (or other rearrangement

mechanisms) versus gene conversion events between older duplicates, making it appear as though these had duplicated later than they actually did, as has been observed in yeast [29].

Supporting Online Material:

Figure S1. Histogram of gene cluster membership

The numbers of genes per cluster are shown for each of the three vertebrates individually as well as for all three grouped together. There is no peak at four for any species, or at 12 as the total for all (or 16 for all, considering that there may have been a further genome duplication for fish), indicating that gene losses have been common and have eradicated this signal of genome duplications.

Figure S2. Plot of the genomic positions of paralogous pairs of human genes that arose from duplications pre-dating the fish-tetrapod split

On the X-axis is each chromosome arranged from p- to q-telomeres. On the Y-axis is each of the 22 human autosomes plus the X and Y chromosomes. For each query gene on the X-axis a “hit” is scored if the subject chromosome contains a paralog generated by a gene duplication prior to the fish-tetrapod split. The lower portion of each panel plots the n-fold redundancy along the query chromosome as defined by pairs of paralogs detected in a sliding window analysis. See the Material and Methods section for details, but briefly, for every human query gene, a window was considered of 50 genes to the left and 50 genes to the right, with a “hit” obtained for the subject chromosome if it includes the early-duplicated paralogs of genes on each side of the query. The expected value of four for the 2R hypothesis is highlighted in a darker shade of blue.

Figure S3. Illustration of whole genome four-fold paralogy

The lines connect paralogous genes in the human genome that originated in duplications that occurred after the tunicate-vertebrate split but before the fish-tetrapod split.

Numerals around the outside of the figure refer to human chromosome numbers.

Table S1. Paralogons in the human genome as defined by having two or more pairs of paralogous genes separated by no more than 100 intervening genes (see Materials and Methods)

A and B in the header refer to the first and second chromosome in considered pairs. The columns labeled “Start” and “End” define the extent of each paralogon by numbered genes. The number of paralogous gene pairs defining the paralogon and the total number of genes encompassed by the region are indicated.

Table 1. Overview of the process for analyzing the complete gene sets with number of genes included at each step

Process Step	Gene Counts				Clusters
	Ciona	Fugu	Mouse	Human	
1 Retrieve sequences	15,852	37,241	22,444	22,980	-----
2 Run BLASTP	12,448	27,090	20,918	20,718	-----
3 Make seeds	-----	-----	-----	-----	-----
4 Generate clusters	7,438	11,339	10,069	10,290	6,641
5 With duplication in vertebrates	3,623	8,394	7,131	7,235	3,096
6 At least one fugu and one tetrapod gene, <100 copies in each taxon	3,402	7,885	6,907	7,015	2,951
7 Multiple sequence alignment	-----	-----	-----	-----	-----
8 Trimming gaps, alignability	2,565	5,618	4,924	4,987	2,340
9 Phylogenetic analysis	-----	-----	-----	-----	-----
10 Strictly bifurcating	1,776	3,770	3,118	3,190	1,621

Table 2. Distribution of the human genome’s tetra-paralogons by chromosome under the most permissive model for signal detection

Working from the assumption that two rounds of genome duplication occurred at the base of the Vertebrata, this shows the maximum detectable extent of residual signal. Coverage here is determined by expanding the subset of paralogous segments shown in Figure 5 that are tetra-paralogous to include the complete span of each member paralogon.

Chromosome	Total Genes	Coverage	% Coverage
1	2,165	1,624	75.0
2	1,455	1,063	73.1
3	1,138	810	71.2
4	849	812	95.6
5	1,008	875	86.8
6	1,113	998	89.7
7	1,063	536	50.4
8	788	759	96.3
9	844	788	93.4
10	839	786	93.7
11	1,415	280	19.8
12	1,088	597	54.9
13	377	338	89.7
14	709	658	92.8
15	679	618	91.0

16	946	842	89.0
17	1,222	884	72.3
18	306	25	8.2
19	1,377	789	57.3
20	636	582	91.5
21	261	0	0.0
22	528	321	60.8
X	869	700	80.6
Y	110	0	0.0
Genome	21,785	15,685	72.0

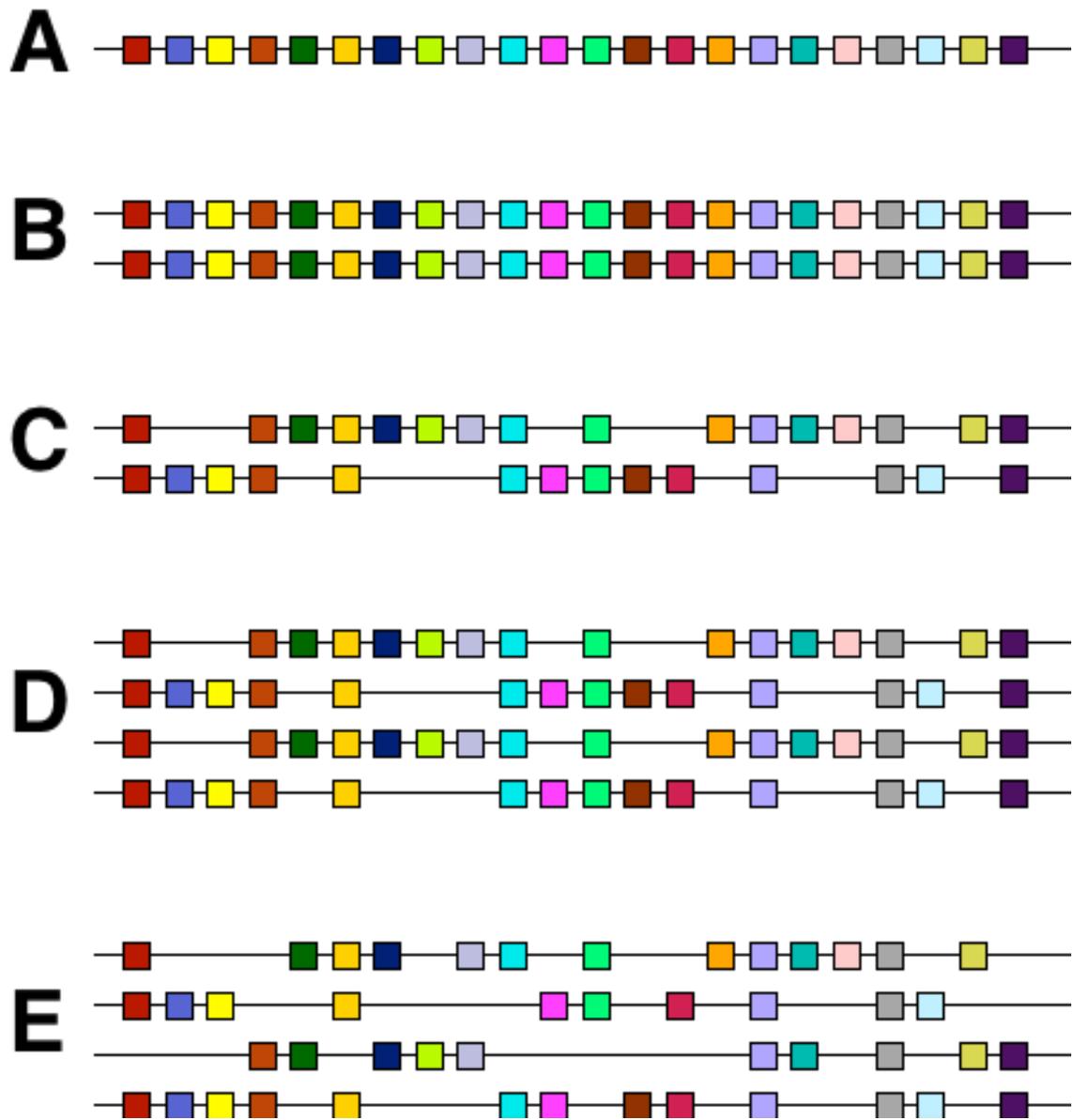


Figure 1-Dehal and Boore

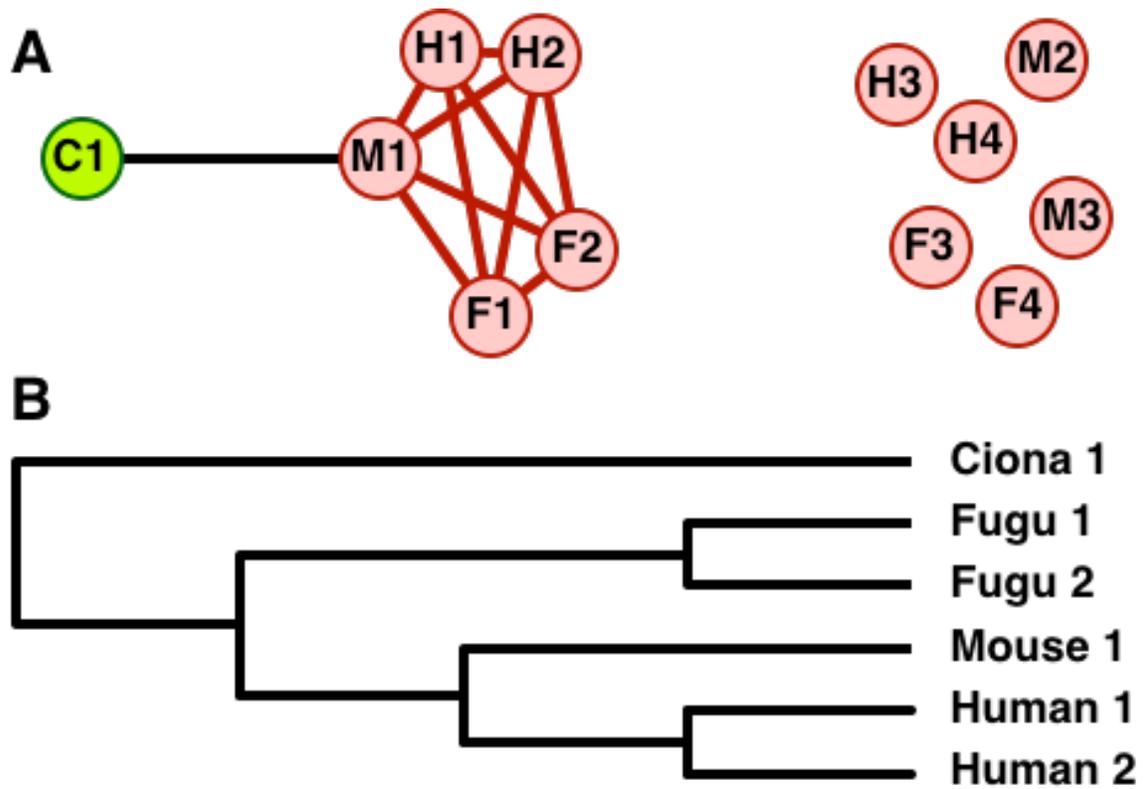


Figure 2 – Dehal and Boore

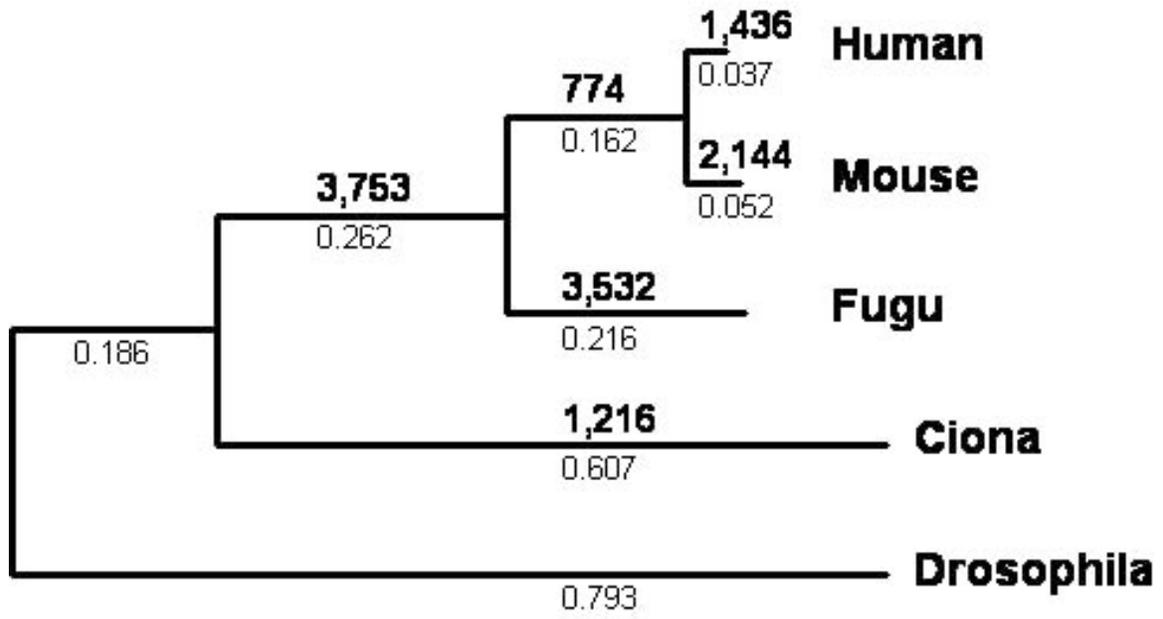


Figure 4 – Dehal and Boore

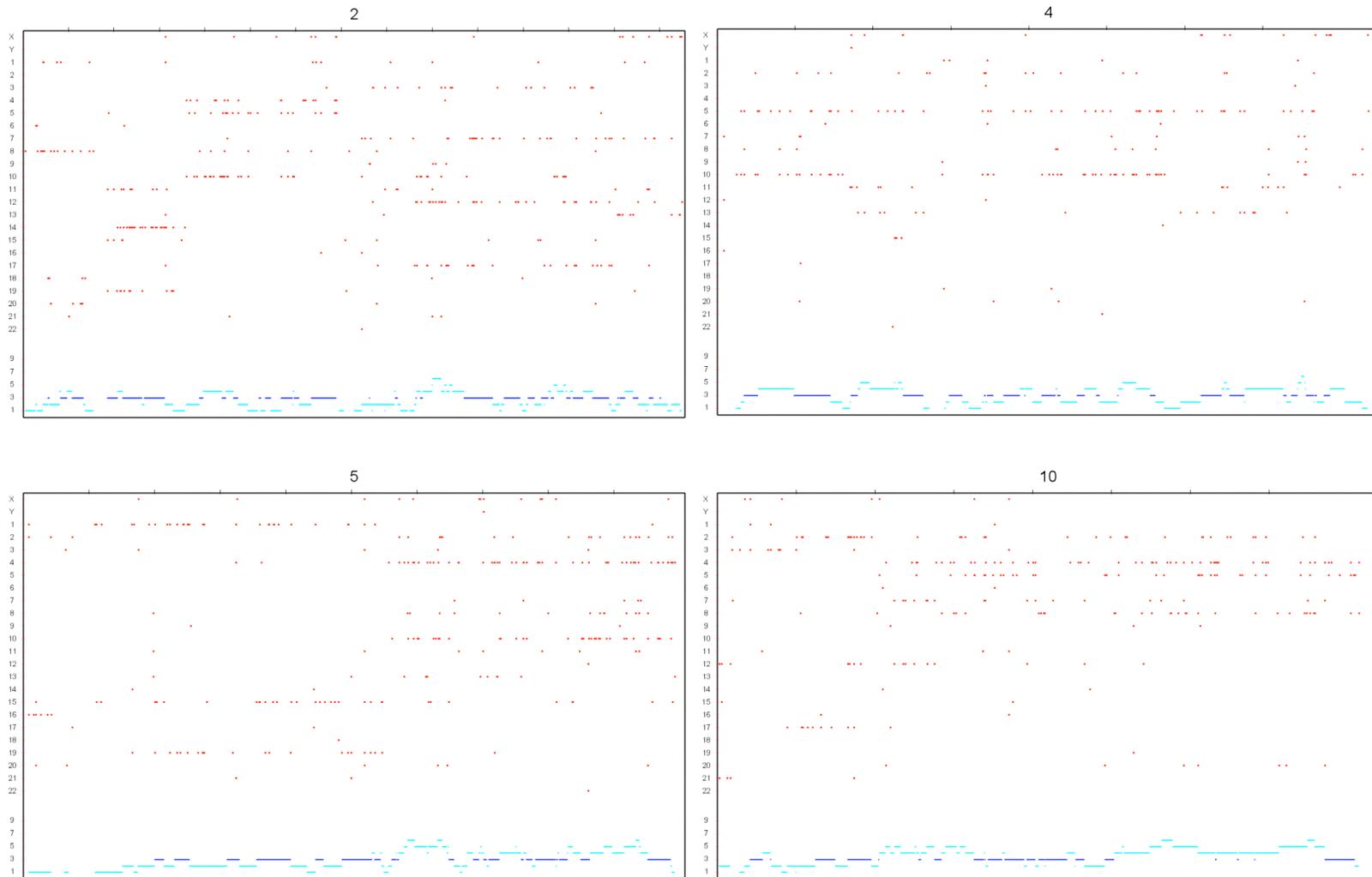


Figure 5 – Dehal and Boore

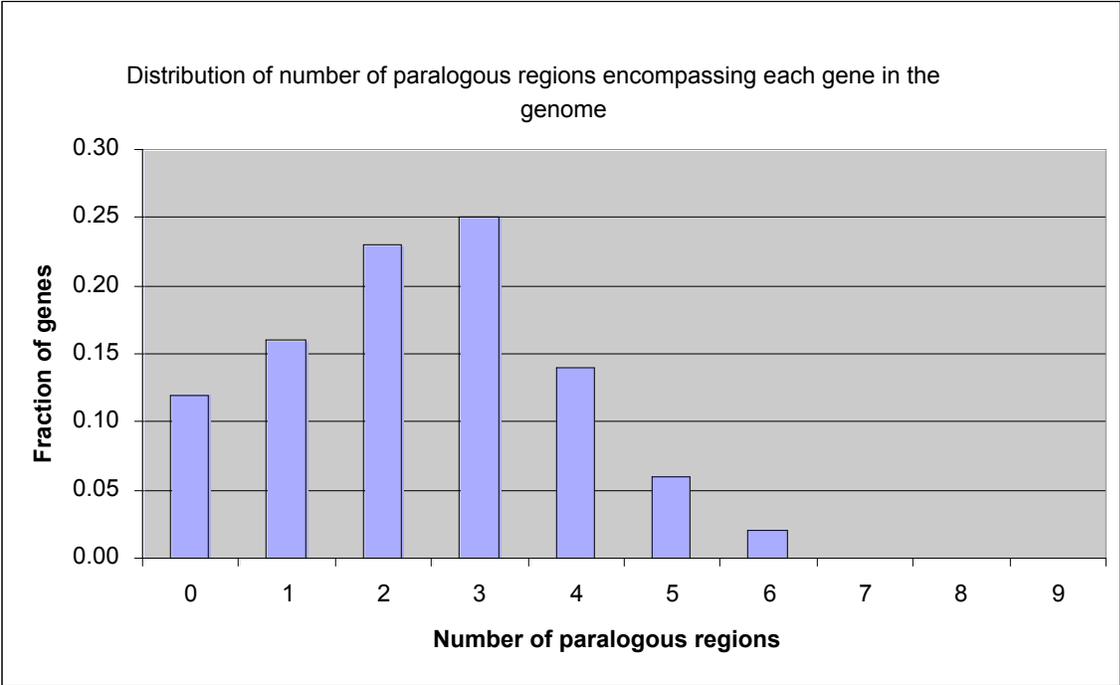


Figure 6 – Dehal and Boore

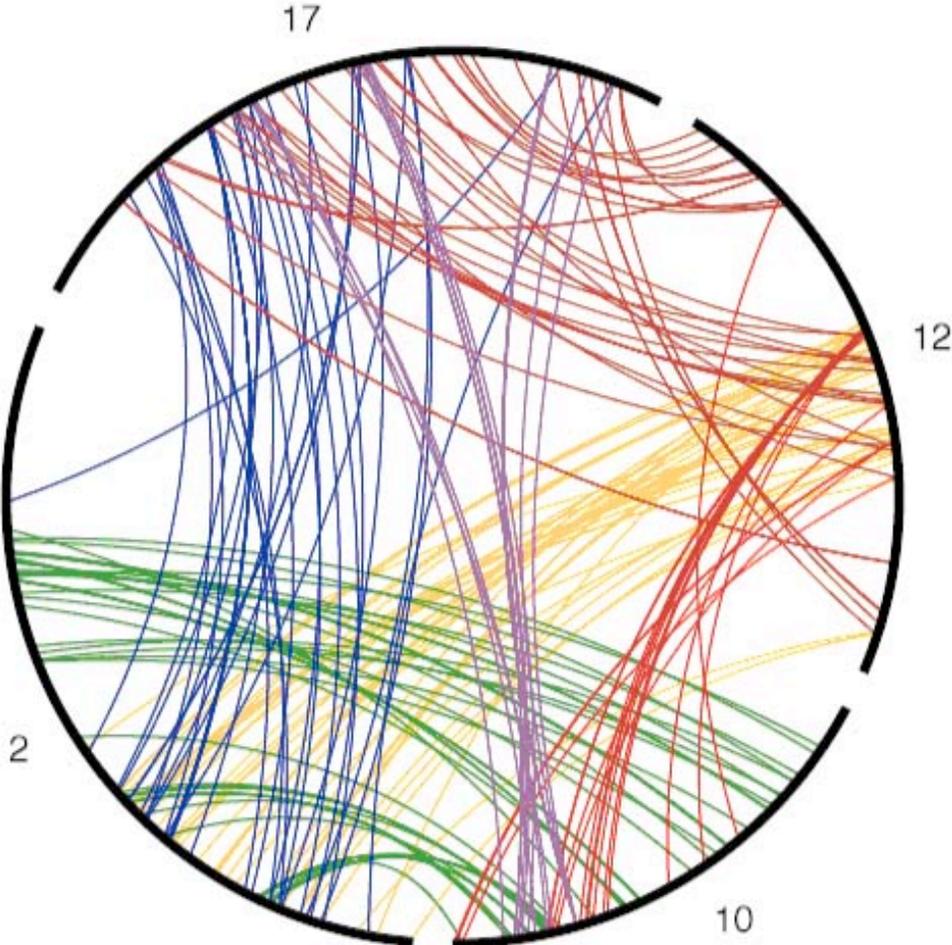


Figure 7A – Dehal and Boore

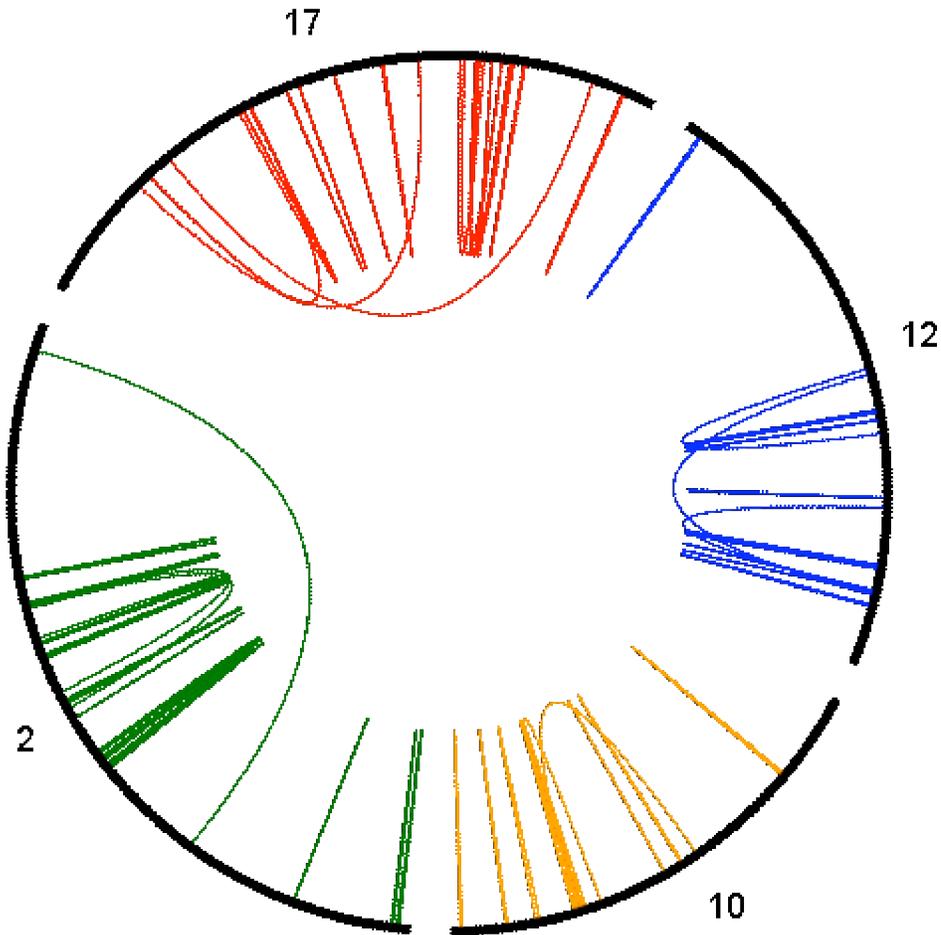


Figure 7B – Dehal and Boore