

# Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate

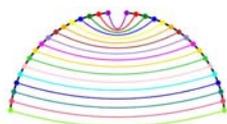
Paramvir Dehal and Jeffrey Boore

## Genomic Signature of 2R

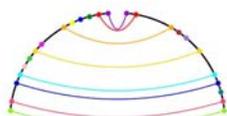
Pattern predicted for the relative locations of paralogous genes from two genome duplications



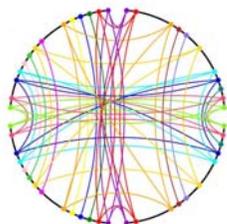
Representation of a hypothetical genome that has 22 genes shown as colored squares.



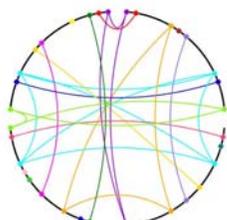
A genome duplication generates a complete set of paralogs in identical order.



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.



A second genome duplication creates 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.

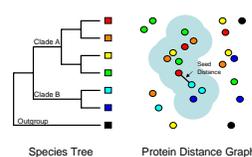


Again, many paralogous genes suffer disabling mutations, become pseudogenes, and are then lost. Of course, unrelated gene duplications and transpositions can occur. Even though this leaves only a few four-member gene families, the patterns of two- and three-fold gene families unite, in various combinations, 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.

## Abstract

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.

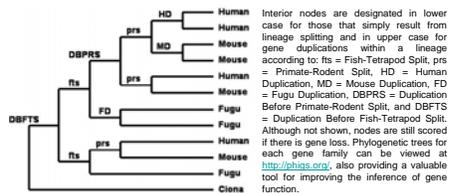
## Genes Clustered according to Species tree using PhIGs



**Illustration of the clustering method**  
The tree shown on the left side of the figure indicates the evolutionary relationships among several hypothetical organisms, four from Clade A, two from Clade B, and one that is an outgroup. The right side of the figure illustrates a protein distance graph with circles representing proteins colored to conform to each organism, with the spatial distance of the circles proportional to their sequence distance. The cluster is created by identifying a pair of sequences (a seed) that is the shortest distance from any Clade A protein to any Clade B protein. The cluster is then grown by adding all proteins that have a shorter distance than the seed until no additions can be made. The blue cloud represents one such cluster.

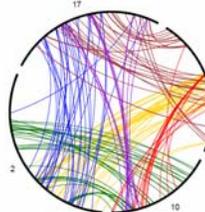
Process Step	Gene Counts				
	Ciona	Fugu	Mouse	Human	Clusters
1 Retrieve sequences	15,852	37,241	22,444	22,980	-----
2 Run BLASTP	12,448	27,090	20,918	20,718	-----
3 Make seeds	-----	-----	-----	-----	-----
4 Gene-wise cluster	7,458	11,334	10,888	10,280	6,841
5 With duplication in vertebrates	3,623	8,384	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

## Identification of Genes that Duplicated at the base of Vertebrates

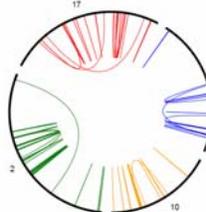


## Hypothetical phylogenetic tree showing all possible types of gene relationships and how they are most parsimoniously interpreted

## Genes duplicating at the base of vertebrates



## Recent gene duplications



## Human Chromosomes 2, 10, 12 and 17 showing tetra-paralogous relationships

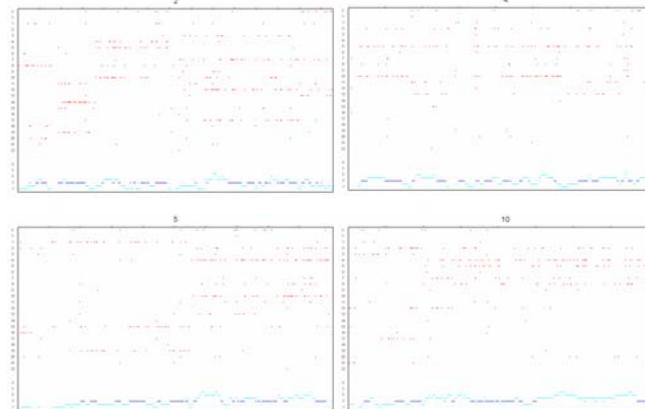
(A) This is an example of the tetra-paralogous relationships of a subset of human genes that are all inferred, by gene trees, to have duplicated prior to the split of fish from tetrapods, but after the split of Ciona from vertebrates. These genes are on four chromosomes with their identities indicated outside of the circle. (B) By contrast, genes having duplicated after the split of fish and tetrapods, do not participate in this tetra-paralogy relationship. Their pattern appears to be consistent with smaller scale tandem and segmental duplications.

## Tetra-paralogy in the Human Genome

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	612	69.6
5	1,038	675	65.8
6	1,113	598	59.7
7	1,063	536	50.4
8	788	759	96.3
9	844	788	93.4
10	826	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	945	842	89.0
17	1,222	884	72.3
18	306	25	8.2
19	1,377	789	57.3
20	636	562	88.5
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

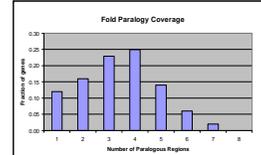
The total extent of paralogous regions participating in tetra-paralogy relationships

## 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. 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 number of pairs of paralogs detected in a sliding window analysis. 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 paralog of genes on each side of the query. The expected value of three for the 2R hypothesis is highlighted in a darker shade of blue.

## Histogram of the distribution of pairs of early-duplicating paralogous genes



This histogram shows a conservative estimate of the fold coverage for all human genes as defined by the sliding window analysis (as shown for a subset of chromosomes in the lower part of each panel of the paralogous gene plot above). The peak at four-fold coverage is consistent with the 2R hypothesis. This is an underestimate because the window analysis misses larger, sparsely defined paralogy patterns and is affected by smaller scale re-arrangement events.

## Conclusion

The mechanism of these genome duplication events, whether two separate rounds of either auto- or allo-tetraploidy or a single octoploidy, remains unresolved. 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. 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.

The broad and pervasive distribution of these tetra-paralogous segments is robust evidence that the 2R hypothesis - two rounds of whole genome duplication at the base of vertebrates - is correct. Despite the large numbers of genes that returned to single copy, many newly duplicated genes were free to adopt altered functions thought to be important for the evolutionary success of the vertebrate lineage. Questions remain about the mechanisms of WGD, the specific roles in vertebrate evolution played by those genes surviving duplication, and the effect, if any, that genome duplication may have had on rates of sequence evolution.