Extensive genomic duplication during early chordate evolution

Aoife McLysaght*, Karsten Hokamp* & Kenneth H. Wolfe

*These authors contributed equally to this work.

Published online: 28 May 2002, DOI: 10.1038/ng884

Opinions on the hypothesis[1] that ancient genome duplications contributed to the vertebrate genome range from strong skepticism[2–4] to strong credence[5–7]. Previous studies concentrated on small numbers of gene families or chromosomal regions that might not have been representative of the whole genome[4,5], or used subjective methods to identify paralogous genes and regions[5,8]. Here we report a systematic and objective analysis of the draft human genome sequence to identify paralogous chromosomal regions (paralogons) formed during chordate evolution and to estimate the ages of duplicate genes. We found that the human genome contains many more paralogons than would be expected by chance. Molecular clock analysis of all protein families in humans that have orthologs in the fly and nematode indicated that a burst of gene duplication activity took place in the period 350–650 Myr ago and that many of the duplicate genes formed at this time are located within paralogons. Our results support the contention that many of the gene families in vertebrates were formed or expanded by large-scale DNA duplications in an early chordate. Considering the incompleteness of the sequence data and the antiquity of the event, the results are compatible with at least one round of polyploidy.

We searched the draft human genome sequence[9] using an objective set of rules to detect groups of related genes at different chromosomal locations (paralogons[8]), which could potentially have been formed by degradation of the symmetry of a polyploid genome. Because the hypothesized genome duplication events were postulated to have occurred during chordate evolution[1,7], we focused on gene duplications younger than the divergence between humans and two invertebrates (*Drosophila melanogaster* and *Caenorhabditis elegans*).

We characterized the paralogons found in terms of the number of pairs of duplicated genes they contained (*sm*). The most extensive region, which paired a 41 Mb region of chromosome 1q (including the tenascin-R locus, *TNR*) with a 20 Mb region of chromosome 9q (including the tenascin-C locus, *HXB*), showed *sm* = 29. The paralogons with the next highest numbers of duplicates lay on chromosomes 7p/17q (*sm* = 28 around the *HOXA*/*HOXB* clusters), 2q/12q (*sm* = 26 around the *HOXD*/*HOXC* clusters), 15q/18q (*sm* = 23 around *NEO1* (encoding neogenin) and its homolog *DCC*), 1p/6q (*sm* = 23 around homologs *EYA3* and *EYA4*, encoding homologs of eyes-absent) and 5q/15q (*sm* = 21 around the rasGAP-related genes *IQGAP2* and *IQGAP1*). The minimal paralogons possible had *sm* = 2, and there were 1,642 paralogons with *sm* ≥ 2 (Table 1). Most chromosomes contained substantial regions of paralogy with multiple other chromosomes. If, for example, a threshold of *sm* ≥ 6 was used, parts of chromosome 17 were paired with parts of seven other chromosomes; this paralogy included extensive similarity to chromosomes 2, 7 and 12 around the Hox clusters (Fig. 1*a*). For example, a region of paralogy between 17q and 3p (Fig. 1*b*)

Table 1 • Distribution of sizes of paralogons found in the human genome

sm[a]	Number of paralogons	Number of genes[b]	Coverage[c]	Redundancy[d]
≥2	1,642	6,120	0.91	3.6
≥3	504	3,852	0.79	2.1
≥4	244	2,730	0.64	1.7
≥5	151	2,139	0.54	1.5
≥6	96	1,662	0.44	1.3
≥7	65	1,315	0.38	1.3
≥8	43	1,030	0.30	1.2
≥9	33	894	0.27	1.2
≥10	25	775	0.25	1.1
≥11	18	640	0.22	1.1
≥12	16	596	0.20	1.1
≥13	14	547	0.18	1.1
≥14	12	498	0.17	1.0
≥15	9	423	0.15	1.0
≥17	8	393	0.13	1.0
≥18	7	357	0.12	1.0
≥21	6	320	0.11	1.0
≥23	5	278	0.08	1.0
≥26	3	182	0.05	1.0
≥28	2	126	0.03	1.0
≥29	1	63	0.02	1.0

[a]Size of paralogon (number of distinct duplicated genes). [b]Number of nonredundant, duplicated genes linked within paralogons of the given size or larger. [c]Fraction of the 3.213 Gb genome that was covered by paralogons of the given size or larger. [d]Ratio of (summed lengths of paralogons)/(length of genome covered by paralogons) for paralogons of the given size or larger.

Department of Genetics, Smurfit Institute, University of Dublin, Trinity College, Dublin 2, Ireland. Correspondence should be addressed to K.H.W. (khwolfe@tcd.ie).

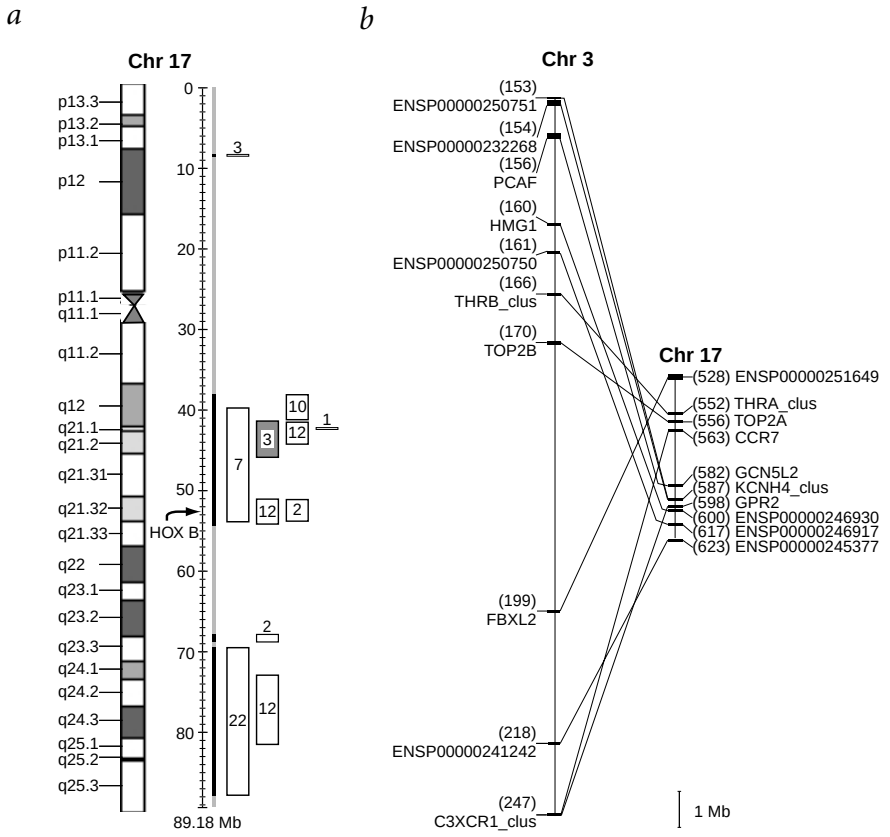

Fig. 1 Paralogons on human chromosome 17. ***a***, View of chromosome 17 showing the paralogons detected between this chromosome and the rest of the genome. Paralogons are indicated by numbered rectangles (identifying the paired chromosome) to the right of the figure. The paralogon with chromosome 3 that is shown in detail in *b* is shaded. The position of the HOXB cluster is marked. ***b***, Closer view of a paralogon containing nine different duplicated gene pairs (*sm* = 9) between chromosomes 3p22–p24 and 17q21. In counting *sm*, multiple interconnected pairs (such as the relationship seen here between C3XCR1_clus on chromosome 3 and both CCR7 and GPR2 on chromosome 17) were counted only once. The suffix '_clus' indicates that a tandem cluster of similar genes has been collapsed into a single representative (see Methods). Genes whose products have names beginning with ENSP are predicted by Ensembl; other names are from HUGO (where available through Ensembl) or Swiss-Prot. Numbers in parentheses indicate the rank order of genes along the chromosome (gene number 1 being the gene closest to the telomere of the p arm). Intervening genes that are not duplicated are not shown. THRB_clus is a cluster consisting of genes THRB (thyroid hormone receptor β) and RARB (retinoic acid receptor β), which are separated by a 1.2 Mb interval containing only one other gene on chromosome 3. THRA_clus is a three-gene tandem cluster consisting of genes THRA (thyroid hormone receptor α), RARA (retinoic acid receptor α) and NR1D1 (orphan nuclear receptor EAR-1), spanning 0.3 Mb and seven other predicted genes on chromosome 17.

included duplicated genes encoding histone acetyltransferases (*PCAF* and *GCN5L2*), topoisomerase II (*TOP2A* and *TOP2B*) and the paralogous nuclear receptor gene clusters *THRA–RARA* and *THRB–RARB*[10].

Even if there had been no large-scale duplications during chordate evolution, some paralogous genes would be expected to be located near one another purely by chance[11]. We performed paralogon detection on 1,000 shuffled gene maps to test the statistical significance of our results (Table 2; see Web Note A online). This analysis indicated that any paralogon with *sm* ≥ 6 was very likely to have been formed by a single duplication of a chromosomal region and that *sm* = 3 was the borderline (with our parameter set) for statistical significance of a candidate paralogy region.

Overall, 96 paralogons with *sm* ≥ 6 covered 44% of the genome with an average redundancy of magnitude 1.3, whereas 504 paralogons with *sm* ≥ 3 covered 79% of the genome with a redundancy of magnitude 2.1 (Table 1). The chromosome pairs 1/19, 1/6, 1/9, 7/17, 4/5, 2/7 and 8/20 all shared more than 50 duplicated genes in paralogons of *sm* ≥ 3. The arrangement of paralogons was generally consistent with that previously reported[12], but comparison at the gene level was not possible because of the lack of details provided in the earlier report (see Web Note B online). Our analysis identified multiply connected groups of chromosomes to a degree considerably greater than suggested by previous proposals[13–15]. These included paralogons on 8q21/14q11/16q11/20q11, where the four genes encoding the transmembrane-type subgroup of metalloproteinases[16] colocalize with four genes encoding copines, a small (five-member) family of possible membrane-trafficking proteins[17], perhaps indicating functional as well as genomic linkage.

In a second analysis, we used the molecular clock to estimate the ages of gene duplications that occurred during chordate evolution. We identified 758 gene families having two to ten human members and fly and nematode orthologs. From phylogenetic trees of these families, in which each intra-specific node represented a gene duplication event, we estimated the ages of gene duplications in humans relative to the divergence time (*D*) of the fly and human lineages (Fig. 2*a*). We analyzed only trees in which the topology was consistent with a duplication in the chordate lineage and that satisfied a molecular clock test[18].

Table 2 • Sizes of paralogons in the human genome, compared with 1,000 simulations in which the gene order was shuffled

	Number of paralogons				
	Real genome	Simulations			
sm[a]		Mean	s.d.	*Z* score[b]	Percentile[c]
2	1,138	1,051.67	29.43	2.93	99.9
3	260	159.05	12.35	8.17	100
4	93	30.10	5.62	11.20	100
5	55	6.89	2.71	17.76	100
≥6	96	2.56	1.63	57.48	100

[a]Number of duplicated genes comprising the paralogon. [b]Number of standard deviations by which the number of paralogons in the real genome exceeded the mean of simulations. [c]Percentage of simulations in which the number of paralogons found in the simulation was lower than or equal to the number of paralogons in the real genome.

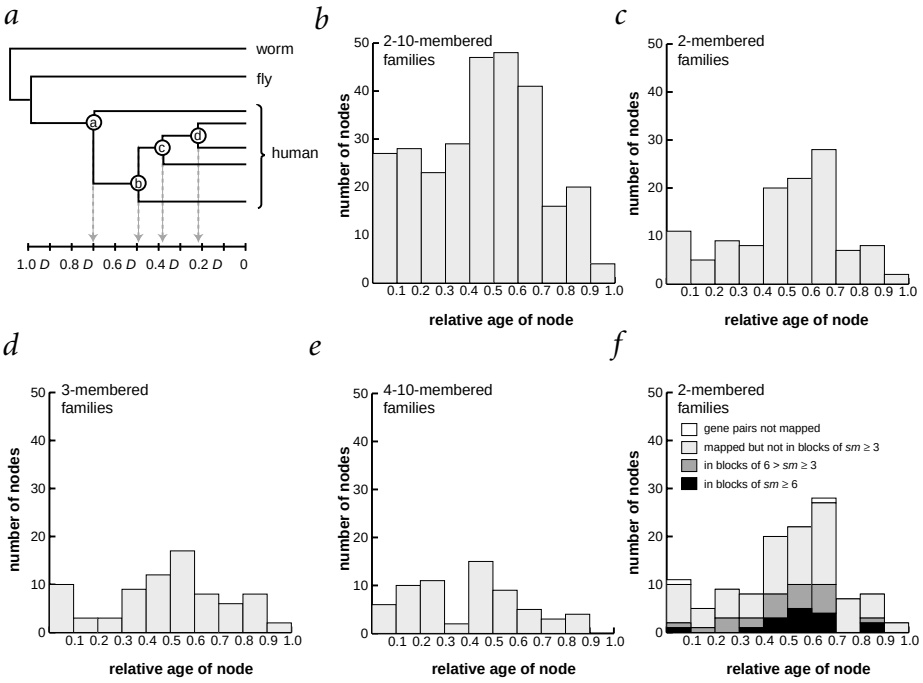

Fig. 2 Estimation of gene duplication dates using linearized trees[18] with fly and nematode outgroups. ***a***, Model linearized tree of a five-membered gene family. The time of duplication for each of the nodes (a)–(d) is indicated on the scale below the tree. Ages are expressed relative to the fly–human divergence (*D*); for example, the age of node (a) is 0.7 *D*. ***b–e***, Distribution of the estimated ages of nodes in two-to-ten-membered, two-membered, three-membered and four-to-ten-membered families, respectively. Each node represents a gene duplication event, and a family with *N* members has *N* – 1 nodes. ***f***, Breakdown of estimated duplication dates among genes mapped to paralogons for two-membered gene families. The duplicated gene pairs in the histogram in *c* were placed into four categories: pairs making up paralogons with $sm \geq 6$ (black), pairs making up paralogons with $6 > sm \geq 3$ (dark gray), pairs that appeared on the human genome map but did not comprise paralogons of $sm \geq 3$ (light gray) and pairs for which one or both genes did not appear on the gene map used in our analysis (white).

The distribution of ages of duplication events (Fig. 2*b–e*) showed an excess in the date range 0.4–0.7 *D*. This was most marked in the pooled histogram for all families with at least two members (Fig. 2*b*) and for the two-membered families alone (Fig. 2*c*). Recent estimates of $D = 833$ Myr ago[19] or $D = 993$ Myr ago[20] place the peak of duplication at 333–583 Myr ago or 397–695 Myr ago, respectively, both spanning the origin of vertebrates. The peak was more apparent in the two-membered families, for which there was only one gene duplication event per tree, than in the larger families (Fig. 2*d*,*e*). This difference was not surprising because even if one (or two) round(s) of genome duplication occurred near the origin of vertebrates, any gene family with more than two (or four) members must include nodes corresponding to gene duplications that were not part of the polyploidizations. A high number of gene duplications during the first half of chordate evolution has also been seen in other studies[21,22]. When genes from non-mammalian vertebrates are included in phylogenetic trees, almost all the resulting topologies are consistent with the gene duplication dates estimated using the molecular clock (Fig. 3).

For the two-membered families, we were able to test whether the gene pair was part of a paralogon. The majority of genes making up paralogons fall in the age class 0.4–0.7 *D* (Fig. 2*f*). Their age distribution was significantly non-uniform ($P < 0.02$ by Kolmogorov-Smirnov test for $sm \geq 3$) but not significantly different from the age distribution of all duplicated genes. Notably, more than 40% of the gene pairs in the age class 0.4–0.7 *D* were components of paralogons ($sm \geq 3$). This was consistent with the idea that many gene pairs in the 0.4–0.7 *D* age group were formed as part of large regional DNA duplications, some of which subsequently fragmented so that they are no longer recognizable as paralogons (see Web Note C online). This is also the pattern that one would expect to see if the paralogons were spurious assemblies of independently duplicated genes, but our simulations indicated that the paralogons are not spurious (Table 2).

Although not explicitly stated by Ohno in his original formulation[1], a widely held version of the genome duplication hypothesis proposes two rounds (2R) of polyploidy in an early vertebrate[5,7,23,24]. Much of the recent literature on the 2R model has invoked expectations for genome structure that are naïve or exaggerated, polarizing the debate. An emerging body of results indicates the following. First, the 'one-to-four' rule[2,6] has not been upheld by genome sequence data[3,9,12]. Second, phylogenetic trees for four-membered human gene families do not show the excess of (AB)(CD) topologies expected under a 2R model[2,3,9]. Third, the human genome contains many more paralogons than expected by chance (Table 2). Fourth, a burst of gene duplication occurred during early chordate evolution (Fig. 2; ref. 21). Fifth, if the paralogons in the human genome were formed by simultaneous large-scale DNA duplication, a widespread deletion of genes must subsequently have occurred (refs 3,8,11 and this study), as in yeast and *Arabidopsis thaliana*[23]. Extensive deletion of genes invalidates the 'one-to-four' expectation. Finally, some paralogons that have been proposed in the literature contain genes that have been duplicated at vastly different times[4,25], which shows that those paralogons (as described) could not have been formed by single duplication events, even though it leaves open the possibility that subsets of them could have been.

All the results listed above are compatible with a single duplication of the whole genome (the 1R hypothesis), or with a single duplication of extensive parts of it (aneuploidy), or with independent large-scale duplications of parts of chromosomes, during early chordate evolution. Only the second result listed is inconsistent with the 2R hypothesis, and even this might be compatible with a modified 2R model in which two rounds of genome duplication happened in close succession without an intermediate diploid stage[23,24]. The 2R hypothesis, however, is loosely defined and essentially unfalsifiable if widespread gene deletion is permitted[23,24]. The results are also compatible with the occurrence of many individual gene duplications either in a simultaneous burst (with the broadness of the date-estimate peak in Fig. 2 being caused by imprecision of the molecular clock) or spread out over approximately 300 million years. If, however, the genes in paralogons were duplicated individually, they must have been transposed later to their current locations, and what adaptive advantage their transposition might have is not understood[11,25].

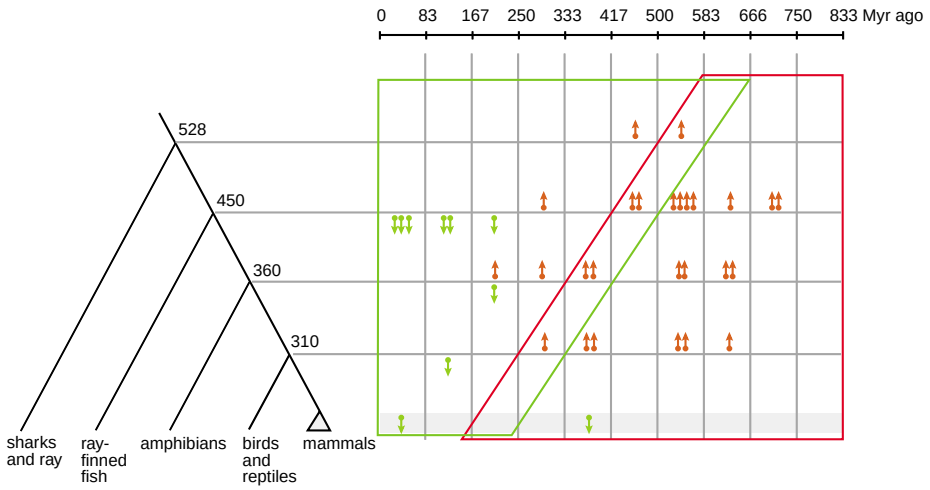

Fig. 3 Comparison of topology-based and molecular clock–based estimation of the dates of gene duplication for 36 human gene pairs. Each arrow shows the result for a pair of human genes comprising a two-membered family for which a homologous sequence from non-mammalian vertebrate species was available. Horizontally, each arrow is placed in one of ten age groups corresponding to its gene duplication date as estimated by the molecular clock, using the same methodology as in Fig. 2 (using only human, fly and nematode sequences). Vertically, each arrow is associated with a node on the phylogenetic tree that forms either a maximum (down arrows, green) or a minimum (up arrows, red) limit for the age of the gene duplication, as determined by the branching order of a phylogenetic tree that included a homologous sequence from another vertebrate. For example, each of the two rightmost red arrows in the diagram indicates a gene duplication that (according to the topology of a tree) occurred before the divergence of the ray-finned fish lineage (more than 450 Myr ago) and (according to the molecular clock) in the time range 666–750 Myr ago. When the results from the two methods are in agreement, all the green arrows should lie within the green polygon and all red arrows within the red polygon. This is true for 31 of the 36 gene pairs when Nei *et al.*'s calibration[19] (D = 833 Myr ago) is used as indicated on the scale at the top. Alternatively, if we use the calibration of Wang *et al.*[20] (D = 993 Myr ago), the clock and topology estimates are congruent for 33 of the 36 families. The timescale for speciations, indicated on the tree at the left, is from Kumar and Hedges[31]. Arrows inside the gray bar at the bottom of the figure indicate gene duplications that occurred within mammals.

It has been argued[3,4,25] (see also ref. 11) that a 'slow shuffle' (individual gene duplications followed by transpositions to form paralogons) is a more parsimonious explanation of the current structure of the human genome than is a 'big bang' (duplication of the whole genome or substantial sections of it). It can, however, be shown both empirically[26] and mathematically (data not shown) that the parsimony test[3,4] will, regardless of which model is correct, always favor the slow shuffle whenever the density of duplicated genes in a genome (or paralogon) is below 50%, as is the case in the well-documented paleopolyploids yeast (16%) and *A. thaliana* (25%) and in this study (12.9%). We believe that this is a shortcoming of the parsimony test, caused by its assumption that every gene deletion is independent, rather than a valid argument against paleopolyploidy in all three genomes. The big bang is more parsimonious than the slow shuffle if multigene deletions of the order of six genes are permitted. We conclude that paleopolyploidy of the human genome is the most parsimonious explanation of our findings, but we do not see any specific evidence for two rounds of polyploidy as opposed to one.

Methods

Sequences. We downloaded the human sequence data set, comprising 27,615 human proteins representing 24,046 genes, from Ensembl version 1.0. We downloaded the *Drosophila melanogaster* proteome (14,335 proteins) from GenBank release 123 (April 2001), and 19,835 *Caenorhabditis elegans* proteins from Wormpep 49. We carried out BLASTP (version 2.1.3) searches of 27,572 human proteins (length 7 residues or greater) against a set of sequences containing all human (except alternative splicing isoforms), fly and nematode proteins (58,216 in total), using a 20-processor Linux cluster and the following parameters: BLOSUM45 matrix, SEG filtering switched on and expectation cutoff of 1. We sorted the resulting 1.7 million query/hit pairs into a MySQL relational database table. Of the queries, 510 produced no hits and 3,002 hit only themselves.

Map. In Ensembl version 1.0, 23,664 genes have been mapped to the reference human genome sequence, which is the Golden Path of December 2000. In cases of alternative splicing, we chose the longest protein to represent a gene. The set was reduced to 20,842 proteins after replacing potential tandem duplicates with their longest representative. We defined a pair of tandem duplicates by a protein that has a BLASTP hit with another protein within a distance of ≤30 genes and an expectation (E) value ≤ 10^{-15}. We identified a further 12 cases in which individual exons appeared to have been incorrectly annotated as complete genes. These were detected by looking for annotated genes ≤30 positions apart, dissimilar in sequence ($E \geq 10^{-5}$), that both hit the same remote protein with $E \leq 10^{-15}$ and aligned with an overlap of <20 amino acids. For these cases, we retained only the longest peptide of each group. This resulted in a final set of proteins representing 20,830 mapped genes.

Paralogon detection. Sequences from nematode and fly were included in the BLAST database to serve as an approximate natural orthology threshold: for each human query protein, we skipped any human BLASTP hits having less similarity than the best invertebrate ortholog, thus distinguishing gene-family expansions that occurred in chordates from older paralogy relationships. This approach is heuristic but is preferable to using only an absolute cutoff for sequence similarity because it recognizes that different proteins evolve at different rates. In large chordate-specific families, we used filters as described below to include only the most similar members, or to exclude the whole family if this were not possible.

For each protein on the map, we compared the BLASTP hits with those of the neighboring proteins, scanning them for matches within the same remote chromosomal location. Our algorithm searched for both intra- and interchromosomal duplications. We combined the resulting pairs within certain limits into paralogons and stored them in MySQL tables. The number of paralogons identified, and the amount of the genome they occupied, varied according to the parameters chosen for the analysis, which were as follows. (i) Alignment length (*al*): the minimum fraction of the length of the longer sequence that is covered by the alignment. (ii) Hit limit: the number of BLASTP hits taken into consideration for each protein was limited by whichever was the lower of either the E-value of the best invertebrate hit or a user-defined E-value limit (*e*). Only hits within a certain range of E-values (*er*) from the top were considered. If the number of these exceeded a threshold (*h*), the whole family was skipped. (iii) Gap size: the maximum number (*d*) of unduplicated genes allowed between two duplicated genes in each paralogon.

We explored various parameter combinations extensively before deciding on those used here, which are all within a stable range (that is, small changes in parameter values do not significantly affect the results). The following parameters were chosen: *al* = 30%, $e = 10^{-7}$, $er = 10^{20}$, h = 20, d = 30. The same parameter-exploration procedure, applied to the genomes of yeast and *Arabidopsis*, produced paralogon maps very similar to those of previous reports on these species[27,28] (data not shown).

Of the 20,830 proteins on the map, 6,281 did not produce hits with other proteins that aligned over at least 30% of the longer sequence length. Of the remaining 14,549 proteins, we excluded 3,911 because their top hit was an invertebrate sequence. We discarded a further 915 proteins because their best hits did not reach the *E*-value threshold (10^{-7}), 334 because they had more than 20 hits within a factor of 10^{20} of the top hit, and 615 because none of the hits could be mapped. This left a set of 8,774 query proteins, of which 329 were mapped only to collapsed tandem repeats, whose BLASTP results were used in the paralogon detection process. In some cases, human proteins that had been eliminated because their top hit was an invertebrate sequence were restored to the data set because they were hit (more strongly than an invertebrate sequence) by another human protein. This made the total number of human proteins used in the paralogon detection process 9,519.

Duplication date estimation using fly and nematode outgroups. We removed alternative splice variants from the fly and nematode data sets (retaining the longest isoform), leaving 13,473 and 18,685 proteins, respectively. We found mutual best hits between fly and nematode proteins with BLASTP (SEG filter, BLOSUM45 matrix), with a maximum *E*-value of 10^{-20} allowed and enforcing a minimum alignment length of 30% of the longer sequence's length. This search retrieved 2,802 mutual best-hit protein pairs. We then used the same protocol to search the fly sequences from this set against the human protein set with alternative splice variants removed. Human gene families were conservatively defined as mutually exclusive BLASTP hits, so that no protein could be a member of more than one family. Where two lists of hits were not mutually exclusive, we excluded both lists from further analysis. This procedure found 1,808 sequence sets containing one fly sequence, one nematode sequence and one to ten human sequences; the fly and nematode genes in these sets were not necessarily single-copy in their genomes, but only one sequence from each invertebrate was used. The family size distribution was similar to those reported elsewhere[3,9,12]. The BLASTP *E*-value threshold (10^{-20}) used in all these searches was chosen because it maximized the number of human gene families obtained (less stringent cutoffs recovered fewer families because of the requirement that they be non-overlapping).

We aligned the 758 two-to-ten-membered human gene families defined by this method with their fly and nematode orthologs using T-COFFEE with its default parameters[29]. We then used these alignments, and initial tree topologies generated by the PHYLIP program protdist with default parameters, to estimate the α parameter for a γ distribution using the program GAMMA[30]. In the γ distribution of evolutionary rates, the variance of the number of substitutions among sites should be greater than the mean. This condition was not satisfied for 154 gene families, and the program returned an unexplained 'format error' for two others, so these families were excluded. We drew neighbor-joining trees for the remaining 602 families using γ-corrected distances[18]. Because we were studying only gene duplications that occurred during chordate evolution, we excluded another 121 families in which the fly and nematode sequences did not group together. The two-cluster test for rate heterogeneity[18] was applied to the 481 remaining families to test for deviations from the molecular clock at 5% significance, and linearized trees were drawn for the 191 families that passed all these criteria. We estimated gene duplication dates for each node of the 191 linearized trees of two-to-ten-membered families by the method shown in Fig. 2*a*. Nodes at which the age was calculated to be zero were excluded from further analysis.

To test the congruence between this molecular clock–based method and the topologies of trees that included sequences from other vertebrates as well as humans (Fig. 3), we compared human proteins from two-membered families with a database of 105,860 non-human vertebrate sequences from SWALL (SwissProt/TrEMBL plus daily updates, 19 September 2001) using the same BLASTP and alignment-length protocol described above. We drew neighbor-joining trees with γ-corrected distances for each family and examined the trees to determine whether the gene duplication pre- or postdated the divergence of ray-finned fish, amphibians, or birds and reptiles.

URLs. The paralogons reported here can be viewed interactively at http://wolfe.gen.tcd.ie/dup. The Ensembl database can be accessed at http://www.ensembl.org and the reference genome sequence at http://genome.ucsc.edu.

Supplementary information is available on the Nature Genetics website.

Acknowledgments

We thank D.C. Shields, A. Coghlan, A.T. Lloyd and other members of the Wednesday lunch group for discussion. This work was supported by the Health Research Board (Ireland), a Trinity College Dublin High Performance Computing studentship award and Science Foundation Ireland.

Competing interests statement

The authors declare that they have no competing financial interests.

Received 30 November 2001; accepted 10 April 2002

1. Ohno, S. *Evolution by Gene Duplication* (George Allen and Unwin, London, 1970).
2. Martin, A. Is tetralogy true? Lack of support for the 'one-to-four' rule. *Mol. Biol. Evol.* **18**, 89–93 (2001).
3. Friedman, R. & Hughes, A.L. Pattern and timing of gene duplication in animal genomes. *Genome Res.* **11**, 1842–1847 (2001).
4. Hughes, A.L., da Silva, J. & Friedman, R. Ancient genome duplications did not structure the human Hox-bearing chromosomes. *Genome Res.* **11**, 771–780 (2001).
5. Thornton, J.W. Evolution of vertebrate steroid receptors from an ancestral estrogen receptor by ligand exploitation and serial genome expansions. *Proc. Natl Acad. Sci. USA* **98**, 5671–5676 (2001).
6. Spring, J. Vertebrate evolution by interspecific hybridisation—are we polyploid? *FEBS Lett.* **400**, 2–8 (1997).
7. Holland, P.W.H., Garcia-Fernandez, J., Williams, N.A. & Sidow, A. Gene duplications and the origins of vertebrate development. *Development Suppl.*, 125–133 (1994).
8. Popovici, C., Leveugle, M., Birnbaum, D. & Coulier, F. Coparalogy: physical and functional clusterings in the human genome. *Biochem. Biophys. Res. Commun.* **288**, 362–370 (2001).
9. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
10. Koh, Y.S. & Moore, D.D. Linkage of the nuclear hormone receptor genes NR1D2, THRB, and RARB: evidence for an ancient, large-scale duplication. *Genomics* **57**, 289–292 (1999).
11. Smith, N.G.C., Knight, R. & Hurst, L.D. Vertebrate genome evolution: a slow shuffle or a big bang? *Bioessays* **21**, 697–703 (1999).
12. Venter, J.C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
13. Ruddle, F.H., Bentley, K.L., Murtha, M.T. & Risch, N. Gene loss and gain in the evolution of the vertebrates. *Development Suppl.*, 155–161 (1994).
14. Flajnik, M.F. & Kasahara, M. Comparative genomics of the MHC: glimpses into the evolution of the adaptive immune system. *Immunity* **15**, 351–362 (2001).
15. Pébusque, M.-J., Coulier, F., Birnbaum, D. & Pontarotti, P. Ancient large scale genome duplications: phylogenetic and linkage analyses shed light on chordate genome evolution. *Mol. Biol. Evol.* **15**, 1145–1159 (1998).
16. Kojima, S., Itoh, Y., Matsumoto, S., Masuho, Y. & Seiki, M. Membrane-type 6 matrix metalloproteinase (MT6-MMP, MMP-25) is the second glycosyl-phosphatidyl inositol (GPI)-anchored MMP. *FEBS Lett.* **480**, 142–146 (2000).
17. Tomsig, J.L. & Creutz, C.E. Biochemical characterization of copine: a ubiquitous Ca^{2+}-dependent, phospholipid-binding protein. *Biochemistry* **39**, 16163–16175 (2000).
18. Takezaki, N., Rzhetsky, A. & Nei, M. Phylogenetic test of the molecular clock and linearized trees. *Mol. Biol. Evol.* **12**, 823–833 (1995).
19. Nei, M., Xu, P. & Glazko, G. 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 (2001).
20. Wang, D.Y., Kumar, S. & Hedges, S.B. Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. *Proc. R. Soc. Lond. B* **266**, 163–171 (1999).
21. Miyata, T. & Suga, H. Divergence pattern of animal gene families and relationship with the Cambrian explosion. *Bioessays* **23**, 1018–1027 (2001).
22. Gu, X., Wang, Y. & Gu, J. Age distribution of human gene families showing significant roles of both large- and small-scale duplications in vertebrate evolution. *Nature Genet.* **31**, 205–209 (2002); advance online publication 28 May 2002 (DOI: 10.1038/ng902).
23. Wolfe, K.H. Yesterday's polyploids and the mystery of diploidization. *Nature Reviews Genet.* **2**, 333–341 (2001).
24. Makalowski, W. Are we polyploids? A brief history of one hypothesis. *Genome Res.* **11**, 667–670 (2001).
25. Hughes, A.L. Phylogenetic tests of the hypothesis of block duplication of homologous genes on human chromosomes 6, 9, and 1. *Mol. Biol. Evol.* **15**, 854–870 (1998).
26. Gu, X. & Huang, W. Testing the parsimony test of genome duplications: a counterexample. *Genome Res.* **12**, 1–2 (2002).
27. Wolfe, K.H. & Shields, D.C. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**, 708–713 (1997).
28. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
29. Notredame, C., Higgins, D.G. & Heringa, J. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217 (2000).
30. Gu, X. & Zhang, J. A simple method for estimating the parameter of substitution rate variation among sites. *Mol. Biol. Evol.* **14**, 1106–1113 (1997).
31. Kumar, S. & Hedges, S.B. A molecular timescale for vertebrate evolution. *Nature* **392**, 917–920 (1998).