New Evidence for Genome-Wide Duplications at the Origin of Vertebrates Using an Amphioxus Gene Set and Completed Animal Genomes

Georgia Panopoulou, Steffen Hennig, Detlef Groth, et al.

Genome Res. 2003 13: 1056-1066
Access the most recent version at doi:10.1101/gr.874803

Article

New Evidence for Genome-Wide Duplications at the Origin of Vertebrates Using an Amphioxus Gene Set and Completed Animal Genomes

Georgia Panopoulou,[1,4] Steffen Hennig,[2] Detlef Groth,[2] Antje Krause,[3] Albert J. Poustka,[1] Ralf Herwig,[2] Martin Vingron,[3] and Hans Lehrach[1,2]

[1]*Evolution and Development Group, Department Professor H. Lehrach, Max-Planck Institut für Molekulare Genetik, D-14195 Berlin, Germany;* [2]*Bioinformatics Group, Department Professor H. Lehrach, Max-Planck Institut für Molekulare Genetik, D-14195 Berlin, Germany;* [3]*Computational Molecular Biology, Department Professor M. Vingron, Max-Planck Institut für Molekulare Genetik, D-14195 Berlin, Germany*

The 2R hypothesis predicting two genome duplications at the origin of vertebrates is highly controversial. Studies published so far include limited sequence data from organisms close to the hypothesized genome duplications. Through the comparison of a gene catalog from amphioxus, the closest living invertebrate relative of vertebrates, to 3453 single-copy genes orthologous between *Caenorhabditis elegans* (C), *Drosophila melanogaster* (D), and *Saccharomyces cerevisiae* (Y), and to *Ciona intestinalis* ESTs, mouse, and human genes, we show with a large number of genes that the gene duplication activity is significantly higher after the separation of amphioxus and the vertebrate lineages, which we estimate at 650 million years (Myr). The majority of human orthologs of 195 CDY groups that could be dated by the molecular clock appear to be duplicated between 300 and 680 Myr with a mean at 488 million years ago (Mya). We detected 485 duplicated chromosomal segments in the human genome containing CDY orthologs, 331 of which are found duplicated in the mouse genome and within regions syntenic between human and mouse, indicating that these were generated earlier than the human–mouse split. Model based calculations of the codon substitution rate of the human genes included in these segments agree with the molecular clock duplication time-scale prediction. Our results favor at least one large duplication event at the origin of vertebrates, followed by smaller scale duplication closer to the bird–mammalian split.

[Supplementary material is available online at www.genome.org. The cDNA clones used in the EST sequencing are available from http://www.rzpd.de/. All ESTs are deposited in dbEST (accession nos. BI385298–BI388632, BI378370–BI381823, BI381824–BI385297, and BI376198–BI378369). The consensus of the alignments of the C, D, Y, orthologs included in the CD/CDY groups that we describe are available for similarity searches at http://www.molgen.mpg.de/~amphioxus.]

The eukaryotic genomes during almost 2 billion years of evolution have been shaped by a large number of processes. An understanding of these processes will be essential for unravelling the multitude of biological processes encoded by the genome. An important step in evolution seems to have taken place at the boundary between cephalochordates with simple body structures and the vertebrates. A plausible explanation for the increase in phenotypic complexity at this point would be a drastic increase in gene numbers, as it could be achieved through the mechanism of a whole-genome duplication, a theory introduced by Susumo Ohno (Ohno 1970). Gene duplications are now recognized as an important mechanism in shaping eukaryotic genomes. However, the hypothesis proposing two rounds of whole-genome duplication (2R hypothesis) at the origin of vertebrates is under heavy debate (Holland et al. 1994; Gibson and Spring 2000; Hughes et al. 2001). Even after having the complete sequence of a vertebrate genome, opinions vary greatly (Friedman and Hughes 2001; McLysaght et al. 2002). As common estimates of the evolutionary time that has elapsed since the speculated whole-genome duplications range from 400 to 500 Myr (Skrabanek and Wolfe 1998), it is expected that if such a phenomenon had happened, most evidence will be fragmental.

The cephalochordate amphioxus, being the closest living invertebrate relative of vertebrates (Wada and Satoh 1994) is a crucial organism for resolving the question of gene/genome duplications at the origin of vertebrates (Sidow 1996; Hughes 1999). The cloning of a number of amphioxus genes in the past (the most prominent example being that of a single Hox cluster [Garcia-Fernandez and Holland 1994]), most of them identified as single copy (Holland 1996), has helped form the idea that amplification of gene numbers in vertebrates has occurred on the vertebrate lineage after its divergence from amphioxus (for review, see Holland 1999). However, most studies published so far include a rather limited number of genes from organisms evolutionarily close to the point of the postulated two-genome duplications, and, therefore, they consider the duplication pattern of genes for which there is no proof showing that they were duplicated at the respective time period. Furthermore, most of the amphioxus genes that have been cloned and used in gene duplications studies so far

[4]Corresponding author.
E-MAIL panopoul@molgen.mpg.de; FAX 49-30-84131128.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.874803.

13:1056–1066 ©2003 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/03 $5.00; www.genome.org

are mainly developmental genes that may be under different selection constraints that could affect their duplication pattern. Therefore, it is necessary to evaluate the 2R hypothesis in light of a large and functionally broader range of genes from amphioxus, but also taking advantage of the recently completed human and mouse genomes. We have generated a nonredundant catalog of genes from amphioxus (*Branchiostoma floridae*). To focus on genes that are likely to have been duplicated during the duplication event predicted by the 2R hypothesis, we constructed a set of groups containing orthologous genes that exist as single copy at least up to the protostome–deuterostome split, namely, in the *C. elegans*, *D. melanogaster*, and *S. cerevisiae* genomes. After assigning genes from ciona, amphioxus, mouse, and human to the above orthologous groups, we tested the 2R hypothesis (1) by estimating the percentage of gene duplications that occurred after the emergence of ciona and amphioxus, and (2) by examining whether these gene duplications are due to a genome-wide event.

RESULTS

An Amphioxus Gene Catalog and Its Complexity

A total of 14,189 5′ ESTs were generated from two amphioxus cDNA libraries prenormalized by oligonucleotide fingerprinting (see Methods), which were subsequently assembled to 9173 consensus sequences (see Methods). On the basis of sample 3′ EST sequencing and the performance assessment of the oligonucleotide fingerprinting, we expect ~15% redundancy in our gene catalog. Selection of clones for sequencing on the basis of the oligonucleotide fingerprinting results does not compromise the use of the EST catalog for gene-duplication studies, as oligonucleotide fingerprinting can distinguish between members of gene families (see Methods). A total of 22% of the amphioxus sequence clusters show similarity to mammalian and 16% to zebrafish nonredundant clustered EST sets, whereas 17% and 14% show similarity to fly and worm-predicted proteins, respectively, at a BLAST E-value $<10^{-10}$. The most abundant genes of the EST catalog are enzymes (14.6%) and genes involved in cell growth and maintenance (14.4%), whereas 16.2% are intracellular and 15.7% membrane components (Suppl. Table S1).

A Set of Invertebrate Single-Copy Orthologous Genes as a Means for Counting Gene Duplications

Single gene duplications have been proposed as an alternative to the whole-genome duplication scenario (Hughes et al. 2001). To distinguish whether the gene duplication rate was higher at the origin of vertebrates rather than earlier in evolution, we constructed two sets of groups containing orthologs between organisms that have branched off prior to the genome duplications predicted by the 2R hypothesis, for which the sequence of their complete genome was also available. One set of groups includes orthologs that exist as single copy in *C. elegans* (C), *D. melanogaster* (D), and *S. cerevisiae* (Y), or in *C. elegans* (C) and *D. melanogaster* (D) only (from now on referred to as single copy CD/CDY groups); one set includes orthologous groups containing more than a single ortholog in at least one of the compared genomes (from now on referred to as multicopy CD/CDY groups). The imposed condition of being single copy in all of the compared genomes ensures that the single-copy state is ancestral rather than due to secondary gene loss of duplicates. We use the single-copy CD/CDY groups to identify how many of the genes that exist as single copy, at least up to the protostome–deuterostome split, have single amphioxus and multiple vertebrate orthologs. This is done by attaching the amphioxus, human, and mouse orthologs to the platform of single-copy CD/CDY groups, and counting the copy number of the amphioxus and vertebrate orthologs per group.

Finally, we use the multicopy CD/CDY groups to estimate the percent of single gene duplications by counting the percent of the multicopy CD/CDY groups that contain more orthologs from one organism (e.g., *D. melanogaster*) rather than from the other two organisms involved (*C. elegans* and *S. cerevisiae*). The CD and CDY sets of single or multicopy orthologous groups do not overlap.

To define the groups of orthologous genes, we used the best hit (BeT) method (Tatusov et al. 1997), in which pairwise comparisons (using BLAST) between all of the participating complete genomes are performed, and the best hit(s) of each protein in each of the other genomes in both directions is identified as its ortholog(s). Orthologs are genes that have

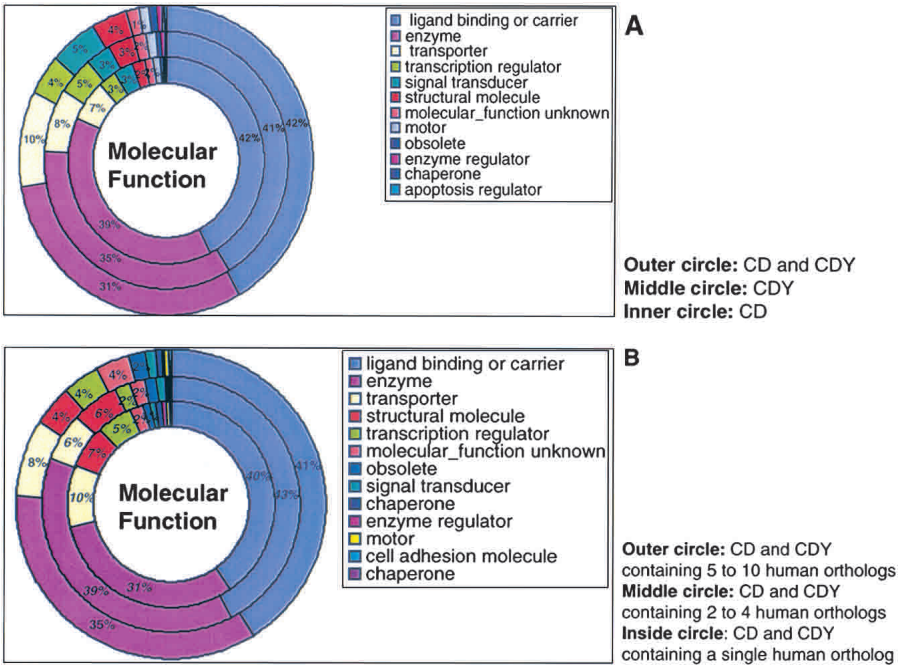

Figure 1 (*A*) Differences in the type of genes included in the CD and CDY groups on the basis of their functional classification in Gene Ontology-defined molecular function type classes. (*B*) Differences in the functional classes of human orthologs according to whether they belong to single, 2–5, and 5–10 human member CD/CDY groups.

evolved by vertical descent from a common ancestor, whereas paralogs originate from gene duplications within a genome (Fitch 1970). The simultaneous inter-relationship between proteins in the compared genomes as a condition of their assignment to the same orthologous group, a particularly stringent criteria of assigning orthology, makes the BeT method independent of similarity scores. This is the main difference from other approaches that define orthologs as those above a certain similarity threshold and/or threshold of alignment length in reciprocal whole-genome comparisons (Chervitz et al. 1998; Rubin et al. 2000). These two significant methodological differences result in the identification of a higher number of orthologous groups than reported previously, as well as to a different copy-rate distribution of orthologs within the groups.

We identified 1419 single-copy CDY groups and 2034 single-copy CD groups, and an additional 432 multicopy CD/CDY groups. At least the number of single-copy CD/CDY groups correlates to the number reported previously by Rubin et al. (2000) (1833 yeast proteins in a set of orthologous genes that also included 3303 fly and 3229 worm proteins). However, it is much higher than the number of orthologous groups reported by Friedman and Hughes (2001) (they report only 716 orthologous families between human and *S. cerevisiae*, whereas we find that 1265 CDY groups contain human orthologs, see below).

Because it has been suggested that the preservation of gene duplicates can vary according to the function that they assume, we were interested in whether the CD/CDY groups represent all functional classes of genes. We found that the most frequent classes in both CD and CDY groups are enzymes, ligand binding or carriers, and transporters if classified by Gene Ontology (Ashburner et al. 2000) molecular function classes (Fig. 1A). This distribution is similar to the distribution reported for the whole-human genome (Venter et al. 2001), which indicates that the CD/CDY groups represent fundamental genes, and are not biased toward a specific functional class.

Counting Gene Duplications at the Transition From Invertebrates to Vertebrates

To estimate the extent of gene duplications at the origin of vertebrates, we have subsequently identified the human, mouse, ciona, and amphioxus orthologs for each of the single-copy CD/CDY groups. We identified human orthologs for 3044 of the single-copy CD/CDY orthologous groups (2.6 times more than reported previously [Lander et al. 2001]), mouse orthologs for 2671 of the single-copy CD/CDY groups, and amphioxus orthologs for 773 single-copy CD/CDY orthologous groups (1248 genes; Table 1A). If one or more large-scale gene duplication event(s) have occurred at the origin of vertebrates, one would expect that the majority of the single-copy CD/CDY groups will contain a single amphioxus ortholog as opposed to multiple human and mouse orthologs. Our results show that this is the case.

We find an average of 1.6 amphioxus gene copies per group as compared with 4.2 for human in the 711 CD/CDY groups that contain orthologs from both species (Table 1A). Similar ratios are found between amphioxus and mouse. Overall, the majority of the single-copy CD/CDY orthologous groups contain human (62%) or mouse (56.2%) genes in multiple copies, whereas in 70.9% of the groups with an amphioxus ortholog, these exist as single copy (Table 1B). The number of cases in which the opposite is true, that is, the number of single-copy CD/CDY groups that contain multiple amphioxus and single human (32 groups) or mouse (40 groups) orthologs is statistically significantly lower, as indicated by the paired signs test ($P = 7.386e\text{-}9$). A detailed inspection of the groups shared between amphioxus/human and amphi-

Table 1A. Comparison of Gene Copy Number Per Single Copy CD/CDY Orthologous Group Between Amphioxus and Ciona, Mouse, or Human

	Amphioxus (A)			Ciona (C)			Mouse (M)			Human (H)		
	N	G	N/G	N	G	N/G	N	G	N/G	N	G	N/G
All	1248	773	1.6	2685	1679	1.6	7384	2671	2.8	10209	3044	3.4
(=1)	548	548	1	1133	1133	1	1173	1173	1	1156	1156	1
(>1)	700	225	3.1	1552	546	2.8	6211	1498	4.1	9053	1888	4.8
(≥1)A:(≥1)C	945	546	1.7	1094	546	2						
1A:1C	237	237	1	237	237	1						
1A:(>1)C	128	128	1	330	128	2.5						
(>1)A:1C	197	75	2.6	75	75	1						
(>1)A:(>1)C	383	106	3.6	452	106	4.2						
(≥1)A:(≥1)M	1115	676	1.6				2427	676	3.6			
1A:1M	192	192	1				192	192	1			
1A:(>1)M	277	277	1				1114	277	4			
(>)A:1M	109	46	2.4				46	46	1			
(>1)A:(>1)M	537	161	3.3				1075	161	6.7			
(≥1)A:(≥1)H	1164	711	1.6							2917	711	4.2
1A:1H	180	180	1							180	180	1
1A:(>1)H	319	319	1							1465	319	4.6
(>1)A:1H	106	46	2.3							46	46	1
(>1)A:(>1)H	559	166	3.4							1226	166	7.4

The number N of genes assigned into G groups (single copy CD/CDY) and the corresponding copy ratio N/G is shown for various cases. (All) The number of all groups containing genes from a specific organism; (>1) Groups have more than one gene from a specific organism. The fourth (light gray), the fifth (dark gray), and the sixth row (light gray) each compare various cases for only those groups in common between amphioxus (A) and ciona or amphioxus and mouse (M) or amphioxus and human (H).

Table 1B. Size Distribution of the Single Copy CD/CDY Orthologous Groups According to the Number of Amphioxus, Ciona, Mouse, and Human Genes They Contain

	Amphioxus		Human		Amphioxus		Mouse		Amphioxus		Ciona	
N/G	G	% of all	G	% of all	G	% of all	G	% of all	G	% of all	G	% of all
A. All group members counted												
1	548	70.9%	1156	38%	548	70.9%	1173	43.9%	548	70.9%	1133	67.5%
2	128	16.6%	615	20.2%	128	16.6%	592	22.2%	128	16.6%	341	20.3%
3	49	6.3%	366	12%	49	6.3%	334	12.5%	49	6.3%	120	7.1%
4	17	2.2%	249	8.2%	17	2.2%	167	6.3%	17	2.2%	46	2.7%
>4	31	4.0%	658	21.6%	31	4.0%	405	15.2%	31	4.0%	39	2.3%
B. Only groups common between amphioxus-human, amphioxus-mouse, and mouse-ciona are counted												
1	499	70.2%	226	21.8%	469	69.3%	238	35.3%	365	66.8%	312	57.1%
2	121	17.0%	131	18.4%	119	17.6%	134	19.8%	103	18.7%	129	23.6%
3	45	6.3%	91	12.8%	44	6.5%	94	13.9%	36	6.6%	46	8.4%
4	16	2.3%	69	9.7%	15	2.2%	42	6.2%	15	2.7%	28	5.1%
>4	30	4.2%	194	27.3%	29	4.3%	165	24.4%	27	4.9%	31	5.7%

The top half of the table lists the groups according to the number of genes from each organism separately; the bottom half lists the respective ratios for the groups common between two species.

oxus/mouse (Table 1A) shows clearly that the copy ratios are correlated, that is, on an average higher gene-copy rates in amphioxus (>1, rate 3.4) give increased gene copy rates in human/mouse (>1, rate 7.4/6.6), which is a strong argument for a coherent duplication mechanism like whole-genome duplication. Finally, it is worth stressing that the distribution of the single-copy human orthologs in functional classes is similar to that of multicopy orthologs (Fig. 1B).

A minority of the single-copy CD/CDY groups contains multiple amphioxus genes (225 groups, 29.1%; Table 1A). This could be due to the estimated 15% redundancy in our catalog (as discussed above), or the counting of splice variants as duplicated genes, or due to misassignment of multiple 5′ ESTs representing the same transcript in the same orthologous group, or to single gene duplications that occurred either in the common ancestor of chordates or within the amphioxus lineage.

Because oligonucleotide fingerprinting recognizes splice variants as different transcripts, we expected that a high percent of the orthologous groups with multiple amphioxus orthologs represent alternative spliced products. However, sample 3′ EST sequencing showed that only 1 of 15 of the orthologous groups that we tested contained splice variants. Another misclustering effect can result from the presence of long domains or multiple copies of short domains. However, as 81% of the amphioxus consensus sequences included in the single-copy CD/CDY groups have at least 100 amino acids overlap with the rest of the genes in the same group, we expect that this has a limited effect in our analysis.

Thus, we believe that most of the multiple-copy amphioxus orthologs represent true gene duplicates. To trace when these duplications happened, we assembled a consensus EST catalog from 88,418 publicly available ESTs of *Ciona intestinalis* (Satou et al. 2001) into 17,492 sequence clusters, which we included in the single-copy CD/CDY groups (Tables 1A and 1B). A total of 237 (43%) of 546 CD/CDY groups that contain orthologs from both amphioxus and ciona include 1 ortholog from both species. A total of 75 (13%) of 546 of the CD/CDY groups common between the two species contain 1 ciona ortholog, and >1 amphioxus ortholog, whereas in 40 (7.32%) of the common CD/CDY groups, the opposite is true. Finally, 106 (19.41%) of the common groups contain more than a single ortholog from both species. On the basis of this and our estimation of the rate of single gene duplications via the multi CD/CDY groups (see below), we conclude that the amphioxus duplicates that we observe are the result of duplications that occurred at the common ancestor of ascidians and cephalochordates (this is also supported by the similarity in the duplication rates between amphioxus and ciona [Tables 1A and 1B]) and lineage-specific duplications.

The 432 groups with multiple copies in C, D, and Y show little correlation in the rates of duplicates. In 89% (Y vs. C and D) or 83% (C vs. D) of the respective common groups, extra copies are observed only in one of the species, so that lineage-specific single gene duplications are the most likely mechanism. In total, only 5%–6% of all CD/CDY groups (single and multicopy) contain >1 copy from either C, D, or Y, which is very small as compared with the 44.7% of the CD/CDY groups with a single amphioxus ortholog containing multiple human/mouse orthologs. This emphasizes that the gene duplications that occurred at the origin of vertebrates were the result of a large-scale gene duplication event rather than single gene duplications.

There are 2551 CD/CDY groups that contain orthologs from both human and mouse. A total of 20% of these are common between the two organisms' groups and contain more than four orthologs in both human and mouse (Table 1B). In 50% of those cases, the same number of extra copies occurs in both mouse and human, indicating that they are the result of duplications common to both species that predated the human–mouse split (>100 Mya).

Evidence for a Complete Genome Duplication Versus Alternative Mechanisms

Four criteria have been most commonly used to evaluate the 2R hypothesis.

The Number of Duplicates Per Gene Family

We find that the majority of duplicated genes in both human (20.2% of the single-copy CD/CDY orthologous groups) and mouse (22.2% of the single-copy CD/CDY orthologous groups) genomes are organized in orthologous groups of two members (Table 1B). Although this finding is in qualitative

agreement with recent publications (Friedman and Hughes 2001; Lander et al. 2001), the number of orthologous groups shared between human and invertebrates that contain multiple copies in human is significantly higher in our study. As an example, we find that 21.6% of the orthologous groups with members from both species contain two human orthologs and a single *S. cerevisiae* ortholog, as opposed to Friedman and Hughes (2001), who report that 11.0% of the groups shared between the two organisms have a 2:1 human-*S. cerevisiae* copy rate.

In addition, a high proportion of the single-copy orthologous groups contain a single human (38%, 1156 groups) or mouse (43.9%, 1173 groups) ortholog. This family-size distribution is not altered if single exon genes (which comprise 7.86% of all the predicted human genes [ENSEMBL build 28] assigned to the CD/CDY groups) that might represent processed pseudogenes that do not result from gene duplications are excluded (Suppl. Table S2). The presence of a large proportion of human genes as single copy or in two-member families is not in contradiction with the 2R hypothesis, as extensive gene loss is known to happen after duplication (Lynch and Conery 2000; Lynch 2002). In addition, we expect that the number of single-copy genes is lower than anticipated at present, as silenced genes have not been included. As an example, we found additional GENSCAN (Burge and Karlin 1997) predicted human orthologs of single-copy CD/CDY groups that do not exist in the ENSEMBL set (422 of 616 single-copy CD/CDY groups that have no ortholog in the ENSEMBL set, match 3001 human GENSCAN predicted proteins [April 2001, freeze]), which might represent pseudogenes (Suppl. Table S3). The addition of these genes can increase the number of multimember orthologous groups at the expense of the single-copy groups (Suppl. Fig. S1).

The Branching Pattern of Phylogenetic Trees of the Duplicated Genes

Although the presence of the (AB)(CD) topology has been used as a means for evaluating the 2R theory (Burge and Karlin 1997; Hughes 1999; Wang and Gu 2000; Hughes et al. 2001; Lander et al. 2001) or even to resolve the timing of gene duplications (Friedman and Hughes 2001), this criterion is highly disputed as the topology of, for example, closely timed duplications cannot be easily resolved with the current phylogenetic methods (Gibson and Spring 2000). We find that even including the amphioxus orthologs, which, because they are evolutionarily much closer to vertebrate than the fly or worm genes, offer better resolution in the phylogenetic tree and avoid potential artifacts (e.g., long branch attraction), results in very few trees (one of nine of the four-member and three of five of the five-member groups) with the topology expected by the 2R hypothesis. The above observation is based on the phylogenetic trees of 62 CD/CDY groups (including 34 amphioxus genes available from GenBank) (Suppl. Table S4 and Fig. S4). In conclusion, the phylogenetic analysis has resulted in findings similar to other studies, that is, although a higher number of amphioxus genes was used, almost all of the tree topologies could agree with one round of genome duplication, followed by additional duplications or two rounds of genome duplications followed by gene loss.

Identification of Duplicated Segments in the Human Genome

The identification of duplicated segments in the human genome that could have been derived from a duplication event at the origin of vertebrates has been addressed for either specific regions in the human genome such as the Hox cluster-bearing chromosomes (Hughes et al. 2001; Larhammar et al. 2002) or the entire human genome (McLysaght et al. 2002). A central issue in the above studies that is debatable is whether all genes that are included in the segments are the result of the same duplication event, and furthermore, whether this event was the one predicted by the 2R hypothesis. The classification of human genes to orthologous groups, which are known to have invertebrate genes only in single copy, at least up to the protostome–deuterostome split prior to finding whether they are organized in duplicated segments in the human genome is an important difference of our approach to the method of McLysaght et. al. (2002). We identified duplicated segments by using each chromosome as query to investigate whether the human orthologs of a series of single-copy CD/CDY groups on the query chromosome are also found in a specific order and neighborhood on another chromosome (target). To compensate for genomic rearrangements, for example, chromosome inversions, translocations, tandem duplications, and gene loss, we started from the smallest unit (pairs of CD/CDY gene groups) that could be conserved, which we then extended to larger segments, while we also allowed for unduplicated genes to intervene between the duplicated CD/CDY pairs (Fig. 2A). We found that only specific chromosomes share CD/CDY pairs that are significant (e.g., Chromosomes 2 and 12, 1 and 6, 1 and 9, 1 and 4, etc.; Fig. 2B). Furthermore, by comparing the frequency of the segment sizes from all human chromosomes with those of a randomized human chromosome set (genes positioned randomly), we found clear significance for only those CD/CDY gene pairs that are not separated by more than four unrelated CD/CDY groups' genes (Suppl. Fig. S3). Therefore, we excluded from further analysis all pairs with more than four intervening genes that belong to other than the query CD/CDY pairs. As each CD/CDY group gene is separated, on average, by 1.3 genes that are not related to a CD/CDY group, this in extent means that the maximum number of genes allowed between CD/CDY pairs is around 10 genes (4 CD/CDY group genes plus $1.3 \times 5 = 6.5$ non-CD/CDY genes). Thus, as opposed to McLysaght et al. (2002), who allow for 30 unduplicated genes between 2 duplicated genes in each paralogon, we found that much less intervening unrelated genes can be allowed if one wants to avoid inclusion of segments that contain duplicated genes found as neighbor by chance. We finally extended the identified pairs of genes by joining neighbor pairs that had one CD/CDY group in common. Overall, we detected 1872 CD/CDY group gene pairs (961 of which had zero intervening unduplicated CD/CDY group genes) containing 2511 genes, when 0–4 unrelated CD/CDY group genes are allowed. The above 1872 CD/CDY human gene pairs (or pair combinations) are included in 656 CD/CDY groups. A total of 192 of these CD/CDY groups with a human ortholog contained within a detected duplicated segment also include an amphioxus ortholog. Neighbor minimal size units (pairs) were subsequently joined in 485 larger segments (extended segments) containing 1611 genes (Table 2). As an example, 21 segments are shared between chromosomes 2 and 12, two of the most frequently quoted chromosomes (Hox-containing chromosomes) in gene duplication studies (Fig. 4). The largest duplicated segment containing 5 genes is located on 1 (46.119 Mb), with its duplicon on chromosome 12 containing 6 genes (5.519Mb). However, the majority of the extended segments contain 2–3 CD/CDY group genes (Table 2).

A first rough estimate of the age of these duplicated segments can be performed by using the known synteny rela-

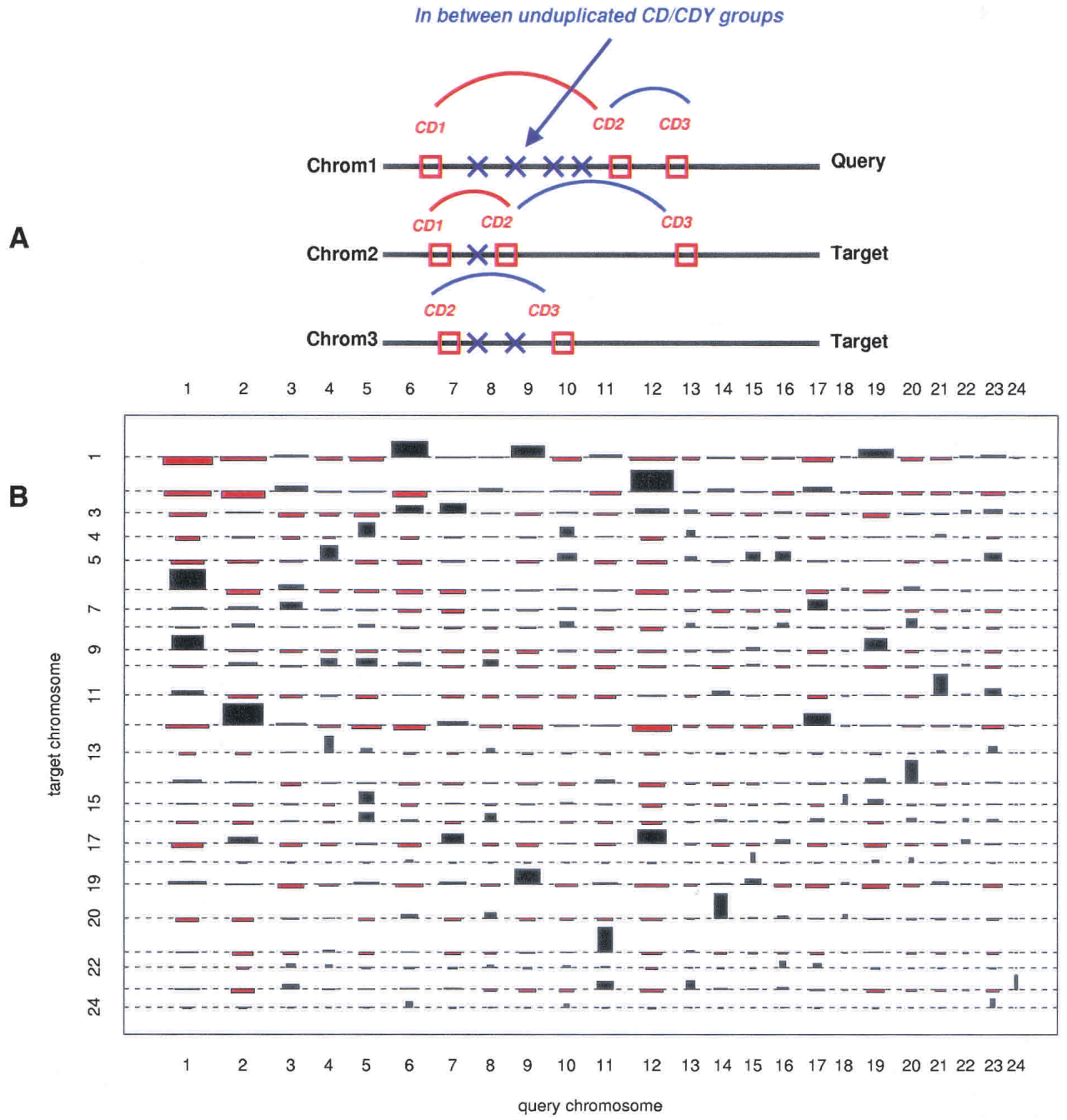

Figure 2 (*A*) A pair of orthologs, each included in one of the CD1 and CD2 orthologous groups located on Chromosome 1 have their orthologs arranged in the same order on Chromosome 2. The CD1 and CD2 are identified as a duplicated pair between Chromosomes 1 and 2. Another duplicated pair are the CD2–CD3 on Chrom1, Chrom2, and Chrom3. Because pairs CD1–CD2 and CD2–CD3 on Chrom1 have the CD2 group in common, they have been joint (named extended segments in our analysis). (*B*) Cohen-Friendly association plot of all-vs-all human chromosomes on the basis of detected duplicated segments. The width of each block indicates the number of pairs of CD/CDY groups shared between the respective pair of chromosomes, whereas the block height indicates the significance (as compared with a normal distribution). Black bars denote highly significant association; red bars denote cases in which the normal distribution shows higher association levels.

tionships between human and mouse genes. We found that the majority (331) of the duplicated extended segments in the human genome are also duplicated in the mouse genome and within described syntenic regions between human and mouse. A total of 142 extended segments and their paralogous region were both found intact (they are located within the same syntenic region) in the mouse genome; in another 189, only one of the paralogous segments is intact in mouse genome, and in 58 cases, both the segment and its duplicon are broken (i.e., they are located in different syntenic regions) in the mouse genome. Thus, we conclude that the traces of those duplications point to a time period before the human–mouse split (>100 Mya).

Table 2. Number and Size of the Duplicated Extended Segments Detected in the Human Genome Ranked According to the Number of CD/CDY Group Genes They Contain

No. of CD/CDY genes included in a segment	No. of extended segments	Mean size of extended segments (Mb)
2	485	0.717
3	72	1.554
4	37	3.438
5	8	6.918
6	3	1.879
8	1	0.715
9	1	1.499
10	1	5.573

A more precise estimate of the duplication time of these segments can be achieved via the calculation of the duplication times of the genes included in the segments. We find that the majority (91% of 195 orthologous groups) of the human orthologs included in CD/CDY groups that could be dated via the molecular clock method (see Methods) were duplicated between 300 and 680 Myr, with a mean at 488 Myr. Furthermore, on the basis of a large number of amphioxus genes (28), we estimate that the cephalochordate–vertebrate split occurred around 647–654 Mya (Fig. 3; Suppl. Fig. S2 and Table S5). This result shows that a high percentage of human duplicates originated from an event after the cephalochordate–vertebrate split. Similar results for the estimation of the duplication times of human orthologs were obtained by Gu et al. (2002). However, due to the inherent disadvantage of the molecular clock approach, which allows the dating of genes that evolve at constant rate, very few genes can be dated by this method. This, in extent, limits the number of genes included in the detected segments that we can link to a date. Only 90 genes of 1611 human orthologs included in the duplicated segments are among those dated by the molecular clock method. To be able to estimate the duplication time of a larger number of human genes, we estimated the codon substitution rate of pairs of human orthologs included in CD/CDY groups, which are contained within the detected duplicated segments.

Calculation of substitution rates can be misleading in the case of genes that have been duplicated a long time ago as silent substitutions that can be saturated, thereby not reflecting the real number of substitutions. Most programs deal with this problem by using various models of sequence evolution. Despite this, we believe that even if some genes may not be reliable indicators of age due to saturation, using a large sample can overcome this bias. We calculated the rate of codon substitutions for 1414 gene pairs included in the segments, and we found that the time calculated on the basis of this rate reproduces the duplication time curve based on the molecular clock (Fig. 3).

DISCUSSION

Due to the paucity of data from organisms that are phylogenetically close to the origin of vertebrates, most studies addressing whether extensive gene duplications have happened early in chordate evolution, and, furthermore, whether these duplications were due to complete genome duplication(s) as opposed to segmental or even sequential single gene duplications, base their results on one-to-one comparisons of human-to-fly or human-to-nematode genomes. Furthermore, they include either very limited or no data from organisms that are phylogenetically very close to the origin of vertebrates. Through a large catalog of genes from amphioxus, the closest invertebrate relative of vertebrates, we provide the first evidence for a large-scale gene duplication immediately after the separation of cephalochordates and vertebrates at 650 Mya. The main difference in our study from those published recently on the 2R hypothesis, is that we use a more stringent method to define orthology prior to comparing gene copy rates, while we also include in our analysis a large novel gene dataset from amphioxus. In light of these differences, we also search for traces of duplications that were potentially created around the origin of vertebrates in the human and mouse genomes. Our conclusions can be summarized as follows:

1. On the basis of orthologous groups of genes that exist as single copy in the genomes of two invertebrate protostomes (*C. elegans* and *D. melanogaster*) and in yeast (*S. cerevisiae*) and in one or more copies in amphioxus, ciona, mouse, and human, we could derive detailed duplication rates showing that, on an average, the number of duplicates has increased more than twofold after the emergence of amphioxus. The number of gene duplications that we observe in the amphioxus catalog are significantly lower, and they represent either small-scale gene duplications that occurred prior to the urochordate–cephalochordate split or lineage-specific duplications.
2. The pattern of gene duplications as determined via a large number of phylogenetic trees that also included amphioxus genes was not conclusive of the mechanism of duplication, that is, under the assumption of extensive gene loss and gain by single gene duplications subsequent to genome wide duplication, both models of one or two rounds of whole-genome duplication would fit with our results.
3. A significant number of human orthologs included in the CD/CDY groups (21.6% of all of the CD/CDY groups that contain a human ortholog and 24.5% of all human orthologs included in the CD/CDY groups) are organized in segments found to be duplicated within the human genome. The majority of these duplicated segments are also duplicated in the mouse genome and within regions syntenic between the two species. The latter is evidence that these segments resulted from a common duplication mechanism.
4. The majority of human orthologs are found to be duplicated around 488 Mya, when the estimation is carried out either by the molecular clock or the rate of codon substitutions. Both methods give very consistent results and show no evidence for a bimodal distribution as expected by the 2R hypothesis.
5. A significant number of human and mouse orthologs appears in high-copy number (>4) simultaneously in both species, which insinuates that these are due to duplications common to both organisms (perhaps segmental duplications), which are additional to the duplications at the origin of vertebrates.

Although none of the results (1–5), if considered alone, provides full support for the 2R hypothesis, when combined, they provide strong evidence for a duplication event at 300–

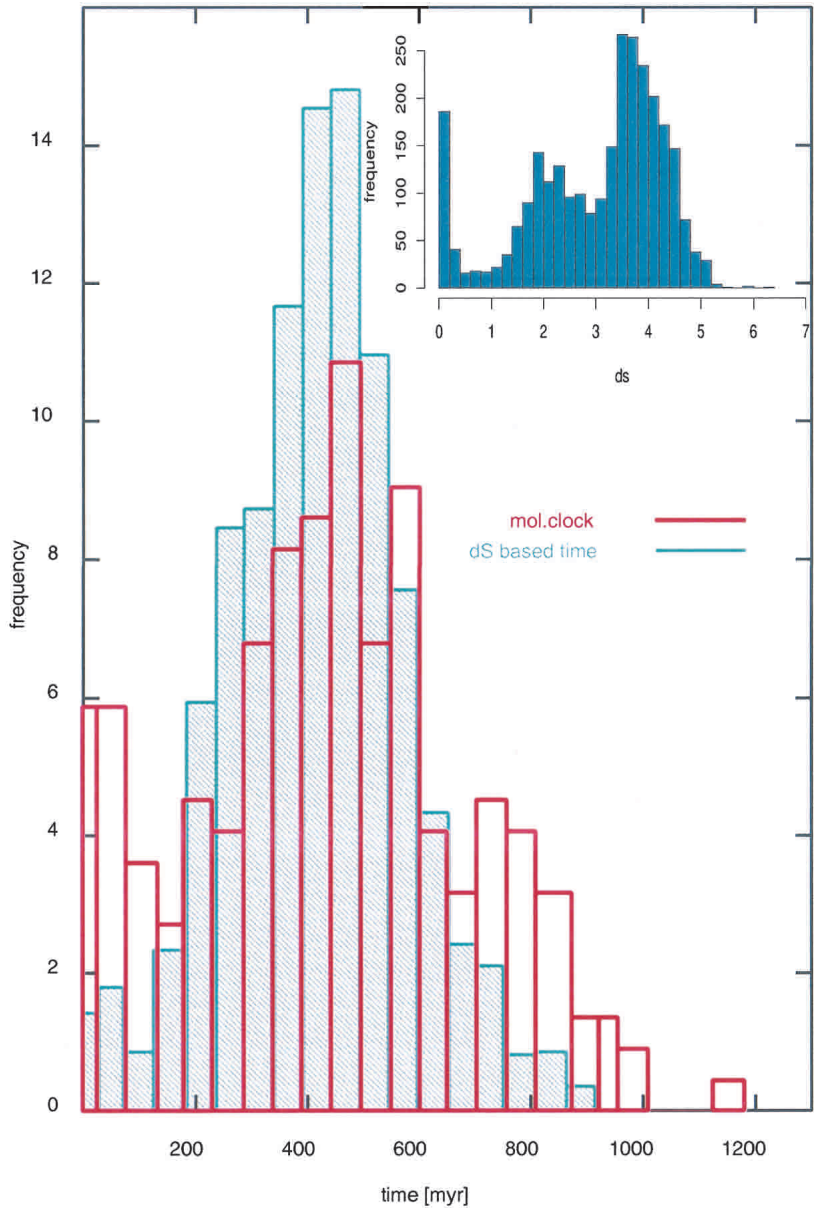

Figure 3 Distribution of the duplication times of the human orthologs of 195 CDY/CD groups calculated by the molecular clock approach (red line). The times were averaged over intervals of 50 Myr. The mean value of 91% of all duplication times is at 488 Myr. Cephalochordates have diverged from vertebrates around 650 Mya (see Suppl. Fig. S2 and Table S5). The duplication time measured as the expected number of nucleotide substitutions per codon (blue line) follows the distribution of the molecular clock duplication times. The distribution of synonymous substitutions per synonymouse site (the *top right* corner of the plot) shows 3 peaks, at 4 (the majority of the values of 1414 pairs of human genes), at 2 and at 0–0.2.

680 Mya that affected either a complete or significant parts of an ancestral genome.

The large number of gene duplicates and segments present in the human and mouse genomes containing genes that we found to be preferentially duplicated after the cephalochordate–vertebrate split is a strong argument for a genome-wide duplication. Single gene duplications are estimated to occur at a rate of 1% per gene per Myr (Lynch and Conery 2000). Some fraction of our duplicates might be due to single gene duplications. However, they cannot explain our principal results in 1,3–5 shown above. Segmental duplications as another mechanism of gene duplications have been discussed recently as one of the major driving forces of evolution (Lynch 2002), and the number of extra duplicates that are common to both human and mouse indicate that this is the case. According to Bailey et al. (2002), segmental duplications are expected to happen at a rate of 0.4% per gene per Myr in human with a duplicon size of 100–200 kb. At the moment, it is unclear whether these features are unique to mammalians. However, if a similar rate of segmental duplications were assumed for the vertebrate lineage, it would be sufficient to destroy the original gene order, so that only traces of an ancient genome duplication would be detectable. This is in qualitative agreement with our results. We find duplicons that range in size from 0.7–5.5 Mb (Table 2) and, therefore, they cannot be regarded as standard segmental duplicates, as they exceed the expected sizes for those by far, and they originate mainly from an ancient time period. Thus, the most likely explanation is that they are remainders of at least one round of a genome-wide duplication event.

Further insights into the mechanisms of genome duplication may be gained from the study of genomes of species phylogenetically closer to cephalochordate–vertebrate split than human and mouse, for example, zebrafish and fugu, which are also believed to have undergone more recent genome duplications.

METHODS

Normalization of the cDNA Libraries

The 5′ EST sequencing was based on two amphioxus oligodT primed embryonic cDNA libraries that were normalized by oligonucleotide fingerprinting (OF). OF is based on the hybridization of 200 10 mers on cDNA arrays, and the subsequent clustering of the hybridization fingerprints (Clark et al. 1999). Clones sharing the same oligo hybridization pattern (fingerprint) and therefore, sequence, are grouped to clusters, whereas unique clones remain unclustered (singletons; Herwig et al. 1999). One clone per fingerprinting cluster was selected for 5′ EST sequencing. Oligonucleotide fingerprinting can distinguish with the same efficiency between members of the same gene family, but also splice variants. We found that in silico clustering using 100–200 of the oligonucleotides that have been used for the experimental normalization of the amphioxus cDNA libraries on sequences of human genes resulted in the accurate assignment of members of gene families or splice variants in different clusters (we tested 190 members of 10

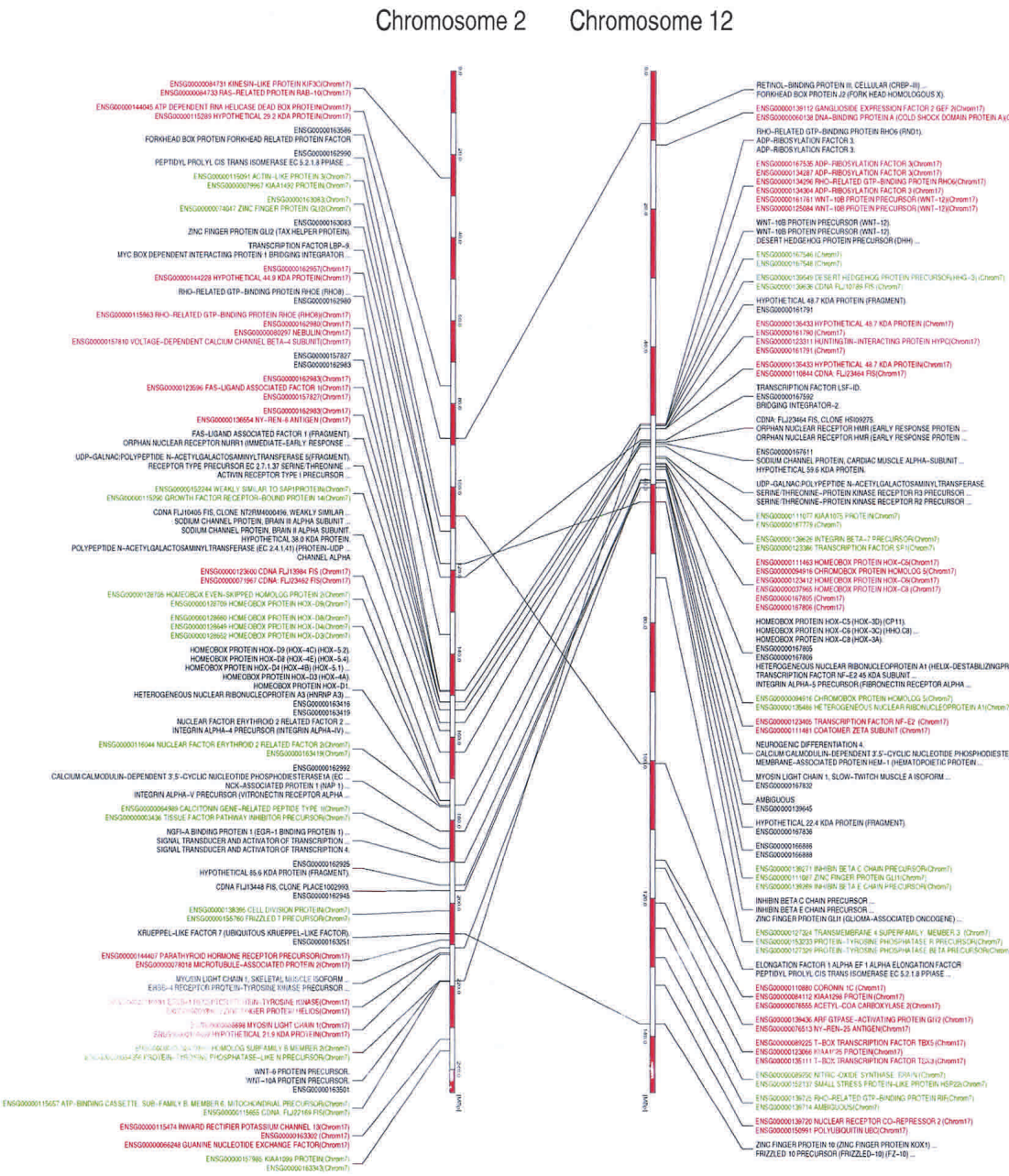

Figure 4 Segments shared between two of the four Hox-bearing chromosomes (Chrom2 and Chrom12). Each connecting line represents one segment. Red and Green-colored segments are those shared between the marked chromosome and the chromosomes 17 and 7, respectively.

human gene families (e.g., paired box containing homeobox genes like *Ptx*, *Sox*, *zinc finger*, *Notch*), as well as 200 splice variants of human genes and 200 randomly selected human genes.

Processing of ESTs

Consensus sequences of 5′ ESTs are built by a procedure described in Haas et al. (2000). After a preclustering step based on pairwise all-versus-all sequence comparisons using BLAST (Altschul et al. 1997), the ESTs are grouped into the final clusters by sequence assembly. The amphioxus consensus sequences are available for similarity searches at http://www.molgen.mpg.de/~amphioxus.

Building Procedure of Orthologous Groups

The CD/CDY groups were constructed by reciprocal searches of the whole genomes of the *C. elegans* (19,099 predicted genes, Wormpep20, Sanger Centre), *D. melanogaster* (13,675 genes, GadFly release 1), and *S. cerevisiae* (6,358 genes, MIPS) based on the principle of the BeT method using gapped BLASTP and imposing the condition of each ortholog being single copy in the respective organism.

To preserve the sensitivity (that might be compromised by the large evolutionary distance between the CD/CDY and chordate genes) while avoiding false positives, we introduced a modification in the BeT method for appending the chor-

dates genes. The reciprocal searches of the CD/CDY genes versus the human (29,076 proteins ENSEMBL build 28), mouse (20,652 proteins ENSEMBL, January 2002 release), and amphioxus (9173 consensus of 14,000 5′ESTs) and ascidian (17,492 sequence clusters of 88,418 publicly available 5′ and 3′ ESTs) genes were carried out using consensus sequences of Hidden Markov Model (HMM) profiles (HMMBUILD [-f option for the local model configuration] and HMMCALIBRATE programs [HMMER2.1.1 package, S. Eddy, http://hmmer.wustl.edu]) of the ClustalW (Higgins et al. 1996) alignments of the *C. elegans*, *Drosophila*, and *S. cerevisiae* predicted proteins included in each CD/CDY group. The consensus sequences of the profiles were extracted with HMMEMIT (HMMER2.1.1 package, S. Eddy, unpubl.). A list of the accession nos. of the C, D, Y genes assigned to the CD/CDY groups can be retrieved at http://www.molgen.mpg.de/~amphioxus. The CD/CDY groups were classified in Gene Ontology-defined functional classes through the INTERPRO (Apweiler et al. 2001) domains contained in the HMM consensus of their HMMprofiles. Functional classification of the CD/CDY groups was also performed through the already existing annotation of the *C. elegans*, *D. melangaster*, and *S. cerevisiae* genes that are included in the CD/CDY groups. The functional classification of the amphioxus ESTs was performed via homology searches against the GO-annotated entries in SWISS-PROT/TrEMBL.

Tree Building and Estimation of Gene Duplication and Species Divergence Time

We have constructed a ClustalW alignment for each of the orthologous groups. Translation of the amphioxus ESTs prior to their alignment with the orthologs from the rest of the organisms included in the same CD/CDY group was carried out using the Framefinder program (G. Slater, http://www.hgmp.mrc.ac.uk/~gslater/estateman/framefinder.html). This program uses log odds hexamer frequency statistics and multiple dynamic programming frameshift models to predict the extend and the reading frame of coding sequence from EST-derived consensus sequences. Because the translation of sequences depends on codon frequency, the program, when translating, takes into account a reference database that is specified by the user. We found that using a human dataset (14,000 human full-length mRNAs from RefSeq) as reference database gives the best (i.e., the longer predicted and correct as judged by the strigency of their BLAST result frames) amphioxus EST translation results. Subsequently, the divergent parts of the alignments were eliminated using the Gblocks program (Castresana 2000).

Phylogenetic Tree Analysis

For the phylogenetic tree analysis, we used the alignments of groups that included an amphioxus ortholog that were at least 100 amino acids in length. For assessing the branching pattern and in order to make sure that the above alignments were long enough to recover the right branching order, we thought of the following test. We prepared the alignments for the above groups without including the amphioxus sequence, but cut the length that they had when the amphioxus sequence was included (reduced alignment). Furthermore, we prepared the alignment for the same groups without including the amphioxus sequences and without cutting the alignment at the length that was imposed by the amphioxus sequence (full-length alignment). Cases in which the four-cluster test (*Phyltest*program [Kumar 1996]) was supporting a different topology for the tree, generated by the full-length alignment rather than for the reduced length tree were not used in our study for counting branching topologies (Suppl. Table S4).

Estimation of the Divergence Time of Cephalochordates and Vertebrates and the Duplication Times of Human Orthologs

The estimation of the divergence time of cephalochordates from vertebrates was based on neighbor-joining trees (with a γ distribution [a = 2] using the MEGA software [Kumar et al. 1994]) of alignments of C, D, Y, or C, D and single-copy human and amphioxus orthologs that were longer than 100 amino acids. The trees were tested for their rate constancy (*Phyltest* program) and the estimation of the divergence time was carried out using the average distance method (Kumar and Hedges 1998). We used as calibration points the chordate-arthropod (993 Myr) and chordate-nematode (1173 Myr) molecular data-based estimates (Wang et al. 1999).

The estimation of the duplication times of human genes was based on CD/CDY groups that include from 2–4 human orthologs. Multimember families tend not to evolve at a constant rate and, therefore, cannot be used for the estimation via the molecular clock approach. The codon substitution rates of the human genes included in the detected segments was carried by the yn00 programme that is included in the PAML package (Yang 1997).

ACKNOWLEDGMENTS

We thank Linda Holland and Nick Holland for the amphioxus RNA, Elizabeth Brundke, Pierre Emmesberger, and Dirk Schudde for their technical assistance, and Thomas Kreitler for plotting the chromosome maps.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D., et al. 2001. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* **29:** 37–40.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. 2000. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25:** 25–29.

Bailey, J.A., Gu, Z., Clark, R.A., Reinert, K., Samonte, R.V., Schwartz, S., Adams, M.D., Myers, E.W., Li, P.W., and Eichler, E.E. 2002. Recent segmental duplications in the human genome. *Science* **297:** 1003–1007.

Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268:** 78–94.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17:** 540–552.

Chervitz, S.A., Aravind, L., Sherlock, G., Ball, C.A., Koonin, E.V., Dwight, S.S., Harris, M.A., Dolinski, K., Mohr, S., Smith, T., et al. 1998. Comparison of the complete protein sets of worm and yeast: Orthology and divergence. *Science* **282:** 2022–2028.

Clark, M.D., Panopoulou, G.D., Cahill, D.J., Bussow, K., and Lehrach, H. 1999. Construction and analysis of arrayed cDNA libraries. *Methods Enzymol.* **303:** 205–233.

Fitch, W.M. 1970. Distinguishing homologous from analogous proteins. *Syst. Zool.* **19:** 99–113.

Friedman, R. and Hughes, A.L. 2001. Pattern and timing of gene duplication in animal genomes. *Genome Res.* **11:** 1842–1847.

Garcia-Fernandez, J. and Holland, P.W. 1994. Archetypal organization of the amphioxus Hox gene cluster. *Nature* **370:** 563–566.

Gibson, T.J. and Spring, J. 2000. Evidence in favour of ancient octaploidy in the vertebrate genome. *Biochem. Soc.Trans.* **28:** 259–264.

Gu, X., Wang, Y., and 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.
Haas, S., Beissbarth, T., Rivals, E., Krause, A., and Vingron, M. 2000. GeneNest: Automated generation and visualization of gene indices. *Trends Genet.* **16:** 521–523.
Herwig, R., Poustka, A.J., Muller, C., Bull, C., Lehrach, H., and O'Brien, J. 1999. Large-scale clustering of cDNA-fingerprinting data. *Genome Res.* **9:** 1093–1105.
Higgins, D.G., Thompson, J.D., and Gibson, T.J. 1996. Using CLUSTAL for multiple sequence alignments. *Methods Enzymol.* **266:** 383–402.
Holland, P.W. 1999. Gene duplication: Past, present and future. *Semin. Cell. Dev. Biol.* **10:** 541–547.
Holland, P.W., Garcia-Fernandez, J., Williams, N.A., and Sidow, A. 1994. Gene duplications and the origins of vertebrate development. *Dev. Suppl.* 125–133.
Holland, P.W.H. 1996. Molecular biology of lancelets: Insights into development and evolution. *Israel J. Zool.* **42:** S247–S272.
Hughes, A.L. 1999. Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history. *J. Mol. Evol.* **48:** 565–576.
Hughes, A.L., da Silva, J., and Friedman, R. 2001. Ancient genome duplications did not structure the human Hox-bearing chromosomes. *Genome Res.* **11:** 771–780.
Kumar, S. 1996. *Phyltest: A program for testing phylogenetic hypotheses*, Institute of Molecular Evolutionary Genetics, Pennsylvania State University, University Park, PA.
Kumar, S. and Hedges, S.B. 1998. A molecular timescale for vertebrate evolution. *Nature* **392:** 917–920.
Kumar, S., Tamura, K., and Nei, M. 1994. MEGA: Molecular evolutionary genetics analysis software for microcomputers. *Comput. Appl. Biosci.* **10:** 189–191.
Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.
Larhammar, D., Lundin, L.G., and Hallbook, F. 2002. The human Hox-bearing chromosome regions did arise by block or chromosome (or even genome) duplications. *Genome Res.* **12:** 1910–1920.
Lynch, M. 2002. Genomics. Gene duplication and evolution. *Science* **297:** 945–947.
Lynch, M. and Conery, J.S. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290:** 1151–1155.
McLysaght, A., Hokamp, K., and Wolfe, K.H. 2002. Extensive genomic duplication during early chordate evolution. *Nat. Genet.* **31:** 200–204.
Ohno, S. 1970. Evolution by gene duplication. Springer-Verlag, New York.
Rubin, G.M., Yandell, M.D., Wortman, J.R., Gabor Miklos, G.L., Nelson, C.R., Hariharan, I.K., Fortini, M.E., Li, P.W., Apweiler, R., Fleischmann, W., et al. 2000. Comparative genomics of the eukaryotes. *Science* **287:** 2204–2215.
Satou, Y., Takatori, N., Yamada, L., Mochizuki, Y., Hamaguchi, M., Ishikawa, H., Chiba, S., Imai, K., Kano, S., Murakami, S.D., et al. 2001. Gene expression profiles in Ciona intestinalis tailbud embryos. *Development* **128:** 2893–2904.
Sidow, A. 1996. Gen(om)e duplications in the evolution of early vertebrates. *Curr. Opin. Genet. Dev.* **6:** 715–722.
Skrabanek, L. and Wolfe, K.H. 1998. Eukaryote genome duplication—Where's the evidence? *Curr. Opin. Genet. Dev.* **8:** 694–700.
Tatusov, R.L., Koonin, E.V., and Lipman, D.J. 1997. A genomic perspective on protein families. *Science* **278:** 631–637.
Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.
Wada, H. and Satoh, N. 1994. Details of the evolutionary history from invertebrates to vertebrates, as deduced from the sequences of 18S rDNA. *Proc. Natl. Acad. Sci.* **91:** 1801–1804.
Wang, D.Y., Kumar, S., and Hedges, S.B. 1999. Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. *Proc. R. Soc. Lond. B. Biol. Sci.* **266:** 163–171.
Wang, Y. and Gu, X. 2000. Evolutionary patterns of gene families generated in the early stage of vertebrates. *J. Mol. Evol.* **51:** 88–96.
Yang, Z. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13:** 555–556.

WEB SITE REFERENCES

http://hmmer.wustl.edu; Computational tools for building Hidden Markov Models from multiple alignments.
http://www.hgmp.mrc.ac.uk/~gslater/estateman/framefinder.html; EST framefinding program.
http://www.molgen.mpg.de/~amphioxus; Amphioxus project page at Max-Planck Institute for Molecular Genetics. Facility for similarity searches versus the amphioxus gene catalogue and overview of ongoing amphioxus research in the group are provided.
http://www.rzpd.de/; Resource Center/Primary Database, Deutsches Ressourcezentrum für Genomforschung. Genomic resources for amphioxus (*Branchiostoma floridae*) and other organisms.

Received October 7, 2002; accepted in revised form March 24, 2003.