Review

TRENDS in Genetics Vol.xx No.xx Monthxxxx

Full text provided by www.sciencedirect.com

Timing and mechanism of ancient vertebrate genome duplications – the adventure of a hypothesis

Georgia Panopoulou and Albert J. Poustka

Evolution and Development Group, Department of Vertebrate Genomics, Max-Planck Institut für Molekulare Genetik, Ihnestrasse 73, D-14195 Berlin, Germany

Complete genome doubling has long-term consequences for the genome structure and the subsequent evolution of an organism. It has been suggested that two genome duplications occurred at the origin of vertebrates (known as the 2R hypothesis). However, there has been considerable debate as to whether these were two successive duplications, or whether a single duplication occurred, followed by large-scale segmental duplications. In this article, we review and compare the evidence for the 2R duplications from vertebrate genomes with similar data from other more recent polyploids.

Introduction

For some time the differences in morphological complexity between animals have been associated directly with the number of genes. Vertebrates almost consistently have more genes than invertebrates and have unique anatomical structures that are characteristic for their phylum. Did this increasing complexity occur through more genes arising following genome duplication?

History of the 2R hypothesis

According to Ohno [1], gene and especially genome duplications are of enormous importance because they can generate large amounts of raw genetic material in a short time that can be exploited by the mutation and positive selection processes to evolve novel gene function. Based on the genome size of the cephalochordate amphioxus, which is three times as large as the genome of the urochordate (see Glossary) *Ciona*, Ohno argued in favor of a genome duplication following the divergence of urochordates. Isozyme studies, and the analysis of orthologous genes from amphioxus and *Ciona*, showed that most genes are present as single copies, whereas the genomes of jawless vertebrates, such as lamprey and hagfish, contained at least two orthologs and mammals contained three orthologs or more [2]. This evidence together with the identification of a single Hox cluster in amphioxus (the invertebrate closest to vertebrates phylogenetically) [3], compared with four clusters in mammals, enabled a refinement of the proposed time of duplication to the period following the split of the cephalochordate and vertebrate lineages and before the emergence of gnathostomes (Figure 1). Based on the apparent stepwise increase in the gene copy-number from invertebrates to jawless

Glossary

(AB)(CD) topology measure: the nodes of the phylogenetic tree of four duplicates generated from two duplication events should have the (AB)(CD) topology where the dates of duplication for the (AB) and (CD) nodes are the same. Neighbor genes within paralogons that have the same topology are assumed to have been generated through the same event.
Agnathans: jawless vertebrates.
Aneuploidy: the loss or addition of one or more specific chromosomes to the normal set of chromosomes of an organism (e.g. a form of aneuploidy is trisomy 21).
Cephalochordates: invertebrate animals that are the closest living relatives of vertebrates. They are characterized by the presence of a notochord, the non-ossified precursor of the 'vertebrate' spinal column. The only representatives are several species of the Genus *Branchiostoma* also known as amphioxus.
Diploidization: the evolutionary process whereby the gene content of a tetraploid species (after a WGD) degenerates to become a diploid with twice as many distinct chromosomes. This procedure enables the correct pairing of homologous chromosomes during meiosis/mitosis (diploid mode of inheritance).
Gnathostomes: all jawed vertebrates.
Molecular-clock-dating method: phylogenetic trees are tested for the constancy of the rate of amino acid substitution of genes. This rate is then estimated using a pre-estimated (based on fossil record or molecular data) divergence time between two species of which genes are included in the phylogenetic tree. This rate value is then used to estimate the divergence time between any other species on the same tree or the age of duplicates. A variation of the molecular-clock method is the global clock, where the time estimation is carried out on phylogenetic trees that have been reconstructed after excluding lineages that evolve significantly faster or slower than the average rate (linearized trees).
Orthologs are genes that have evolved by vertical descent from a common ancestor, whereas paralogs originate from gene duplications within a genome.
Paralogons: distinct chromosomal regions within a genome that share a set of paralogs.
Polyploidy: the situation where the normal set of chromosomes (*n*) of an organism is multiplied (2*n*, 3*n*, 4*n*) if compared with the number of chromosomes of a related species.
Allopolyploid: an organism that has a set of chromosomes that originates from different species.
Autopolyploidy: an organism that has sets of chromosomes that originate from the same species.
Relative dating: phylogenetic tree based counting of duplications relative to the divergence of major taxa.
Synonymous or silent substitutions: the replacement of a single nucleotide in a codon without the change of the amino acid encoded. Synonymous sites are preferred over non-synonymous sites for the estimation of the duplication time, because it is known that non-synonymous sites are under purifying selection, which can greatly vary among genes.
Synteny: conserved gene content and order on chromosomes.
Urochordates or Tunicates: a more distant invertebrate chordate subphylum than cephalochordates that includes the appendicularians (*Oikopleura*) and ascidians (*Ciona*).

Corresponding author: Panopoulou, G. (panopoul@molgen.mpg.de).

www.sciencedirect.com 0168-9525/$ - see front matter © 2005 Elsevier Ltd. All rights reserved. doi:10.1016/j.tig.2005.08.004

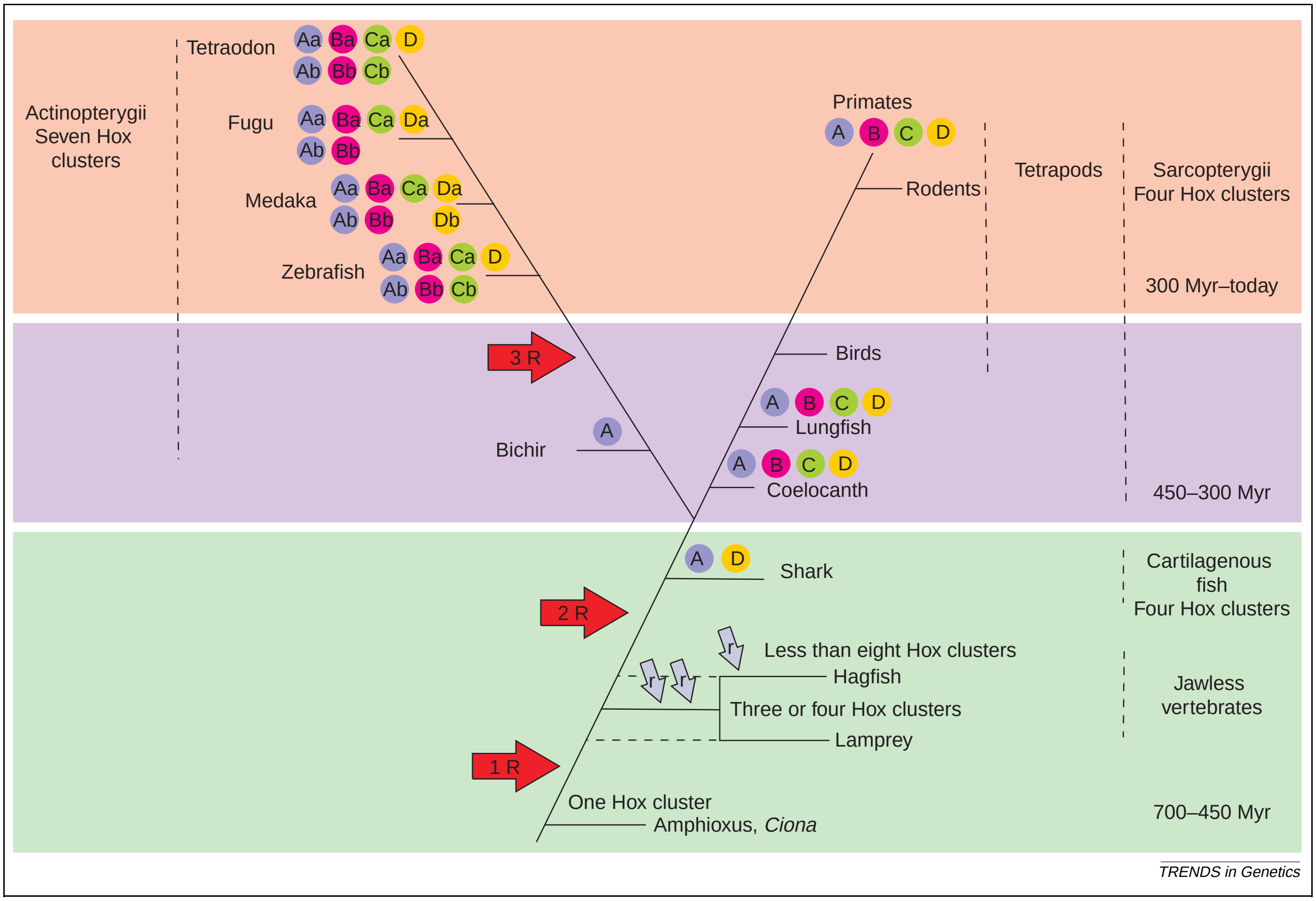

Figure 1. Whole genome duplications (WGDs) on the vertebrate lineage, based on either complete genome analysis or Hox cluster number. Each circle is equivalent to a Hox cluster with each cluster colored differently. The arrangement of circles does not represent the arrangement of clusters in the genome. Arrows indicate where WGDs have occurred. The only definitely proven WGD is the fish-specific 3R WGD (see Figure 4). Evidence for 1R or 2R WGDs is provided by numerous paralogons and by many quadruplicate regions in the human genome. Recent data from jawless vertebrates indicate that additional WGDs occurred after their divergence from the gnathostome (grey arrows). There is some controversy over the monophyly of jawless vertebrates (broken lines) [77]. The time windows given are estimates. Bichir (*Polypterus senegalus*) has one HoxA cluster, whereas all other teleosts have two, which have undergone 3R [78]. The existence of one HoxA and one HoxD cluster (which implies that HoxB and HoxC should be present) in shark (*Heterodontus francisci*) places the 2R duplication before the emergence of cartilaginous fish [62]. For information on the number of Hox clusters in other species, see the following references: lamprey (*Petromyzon marinus*) [53,54], hagfish (*Eptatretus stoutii*) [61], zebrafish (*Danio rerio*) [50], pufferfish (*S. nephalus*) [51], medaka (*Oryzia latipes*) [52], *Ciona* (*Ciona intestinalis*) [79], coelocanth (*Latimeria menadoensis*) [80], amphioxus (*Branchiostoma floridae*) [3].

vertebrates to mammals, it was suggested that two episodes of complete or whole genome duplication (WGD) occurred [2], one before and one after the jawless fish diverged, which is estimated at 500–430 million years ago (Mya) (i.e. the 2R hypothesis; see Ref. [4] for a summary of proposals for the timing of duplication events).

The identification of three 'large' quadrupled regions in the unfinished human genome, namely the major histocompatibility complex (MHC; human chromosome (Hsa) 1, 6, 9 and 19), an extended Hox (Hsa 2, 7, 12 and 17) and the fibroblast growth factor receptor (FGFR; Hsa 4, 5, 8 and 10) regions, which included genes duplicated ~530–738 Mya strongly supported tetraploidy [5–9]. These rounds of duplication could have happened in short succession within 90–106 Mya [10]. Proponents of the 2R hypothesis argued that this short interval could explain the incongruent tree topologies of neighbor genes within the described paralogons [11] (Box 1), whereas opponents quoted it as a proof that these paralogons did not arise through the duplication of an ancestral block. To explain the numerous paralogs in vertebrates, an alternative scenario of continuous mode of small-scale (tandem or segmental) gene duplications was suggested [12].

Before the completion of the human genome, gene estimates were in the range of ~70 000 for humans (±20 000) and ~20 000 for invertebrates [12–14]. This fourfold difference and the observed 1:4 relationship between many *Drosophila* and human genes (1:4 rule) [15–17] was an additional argument in favor of two rounds of WGD under the assumption that no subsequent gene loss had happened. The estimation that the human genome might contain as few as 25 000 genes [18–22] signaled that if there had been WGDs, they must have been followed by extensive gene loss; therefore, finding evidence for old duplications might not be as straightforward as originally thought.

What is the evidence for 2R duplications produced from the analysis of the complete human genome and teleost fish genomes? In this article, we will review this evidence in the light of similar data generated from the genome analysis of more recent polyploids such as *Arabidopsis* and *Saccharomyces cerevisiae*.

Box 1. Limitations of the methods for estimating the age of duplications

Absolute dating is performed using the molecular clock or an estimation of the rate of synonymous nucleotide substitutions (K_s). The molecular- clock method assumes a constant rate of amino acid or nucleotide substitution over time, which occurs in only a limited number of genes. Therefore, few duplicates can be dated using this method. Furthermore, because few taxonomically well-assigned fossils whose date is not disputed exist, the dates can be only considered as approximate (see Refs. [68–70]). Finally, genes within families evolve at constant rate but the rate between distinct gene families is different. As a result, the age of genes from distinct families that have been duplicated during the same event will differ. Therefore, the number of duplication events cannot be deduced based on the shape of the age distribution of duplicates. Synonymous sites can be saturated with changes leading to underestimation of the duplication time, thus only small K_s values (representing recent duplications) are meaningful. Relative dating can be only applied when: (i) the phylogeny is resolved; and (ii) the sequences of orthologs from a diverse range of organisms are available, which ensures correct tree resolution (without long-branch attraction).

The (AB)(CD) topology measure, used to count the order of two duplication events within gene families and used to decide whether neighbor genes within paralogons result from the same duplication event is inaccurate in several instances. For example, in two closely spaced rounds of genome duplications, where the second round occurred before the diploidization of the first round is complete [11], the topology of the duplicates will not reflect the sequence of duplication. If the two rounds of duplications are the result of autopolyploidy then the grouping of duplicates on a phylogenetic tree will be random [71]. This happens because the duplicates are not well-resolved, which translates into a similar phylogenetic distance between them. Deducing the order of duplication using the topology criteria could also give false answers when the rate with which genes evolve changes, for example, after duplications when genes enjoy a period of relaxed selective pressure with the result that they evolve at a increased rate [72]. To compensate for this a method that excludes saturated sites has been developed [73].

2R genes in vertebrates and the extent of gene loss

According to the 2R hypothesis, each invertebrate gene is expected to have at least four vertebrate orthologs (in keeping with the 1:4 rule). The human genome shares 1308 gene families with the genomes of *Caenorhabditis elegans*, *D. melanogaster* and *S. cerevisiae*, 43.1% of which are single copy genes in these organisms and in humans [23–26]. If yeast is excluded from this comparison the number of families shared between the human genome and the genomes of *C. elegans* and *D. melanogaster* increases to 3044 [25]. Almost one-third of these gene families also contain a single ortholog in all three organisms.

Can the high number of single-copy human genes be explained by genome duplication (complete or segmental) followed by a high rate of gene loss? Based on the number of substitutions per silent site, *S*, of duplicate gene pairs (as a proxy for age), calculating the number of duplicates at increasing S values and assuming a constant birth and decay rate of duplicates, Lynch and Conery [27] estimated a high rate of birth (0.009 per gene per Myr) and a short life span (7.5 Myr) for human duplicate genes.

Similar high rates of gene loss have been observed for multiple eukaryotic genomes. In *S. cerevisiae*, 85% of the duplicate pairs that resulted from WGD 100 Mya are now deleted [28–30]. Similarly, only 13–18% of duplicated genes that remain in the *Arabidopsis thaliana* genome are considered to be the result of an old polyploidization or aneuploidization event at 170–300 Mya [31]. These estimates of gene loss rates obtained from yeast and plants can be extrapolated to gene loss in vertebrates. *Fugu* gene duplicates that originated from one round of fish-specific genome duplication between 250 and 450 Mya account for only 29.4% of the duplicates [32]. Similar to human, *Fugu* shares 3036 gene families with *C. elegans* and *Drosophila*, 41% of which occur as a single copy in all organisms. Massive gene loss has also been verified for the tetraodon genome through its comparison with the reference 'unduplicated' human genome [33].

Such extensive gene loss makes it impossible to apply the 1:4 rule. The majority of the duplicated genes in the human, mouse and *Fugu* genomes are organized in two member families, whereas families with four members are a minority [25,34]. How gene loss destroys the 1:4 symmetry is best illustrated with the pattern of loss of individual Hox genes within the teleost and human Hox and Parahox clusters where 50% and 25% are lost, respectively [35,36]. It is equally difficult to predict the mechanism of duplication based on the fraction of preserved duplicates. The gene-decay rate depends on the age and mechanism (complete genome or single gene) of the duplication and whether additional duplications have happened before. For example, gene loss is greater after the recent *Arabidopsis* genome duplication than after the older whole genome duplication(s) [37].

In conclusion, the high number of single-copy genes in the human genome is not evidence against genome duplications and it can be fully explained by the extensive gene loss of duplicates. Such extensive gene loss makes the comparison of gene family size between organisms an uninformative measure for deciding on the extent of duplications. Furthermore, it makes it difficult to distinguish between 1R and 2R.

How many vertebrate duplicates date at the origin of vertebrates?

The molecular-clock-based calculation (Box 1) of the age of human duplicates within 191 gene families that have a single invertebrate ortholog (i.e. genes likely to have duplicated on the vertebrate lineage) and the arthropod–chordate divergence estimate of either 833 Myr [38] or 993 Myr [10] showed that most of these human duplicates arose ~333–583 Mya or 397–695 Mya (Figure 2) [25,26]. The dating of numerous vertebrate gene families (749 vertebrate gene families, 1739 gene-duplication events) using the global-clock method similarly showed a broad distribution at 350–850 Myr with a peak at 450–750 Myr [39]. Additional evidence for an excess of duplication events at the origin of vertebrates comes from the knowledge that 70.6% of the *Fugu* duplicates originated at 500–900 Mya (Figure 2) [32,40].

However, it is impossible to conclude, based only on the distribution of ages of duplications, whether one or two closely timed, WGD or even large-scale segmental duplications occurred at the origin of vertebrates.

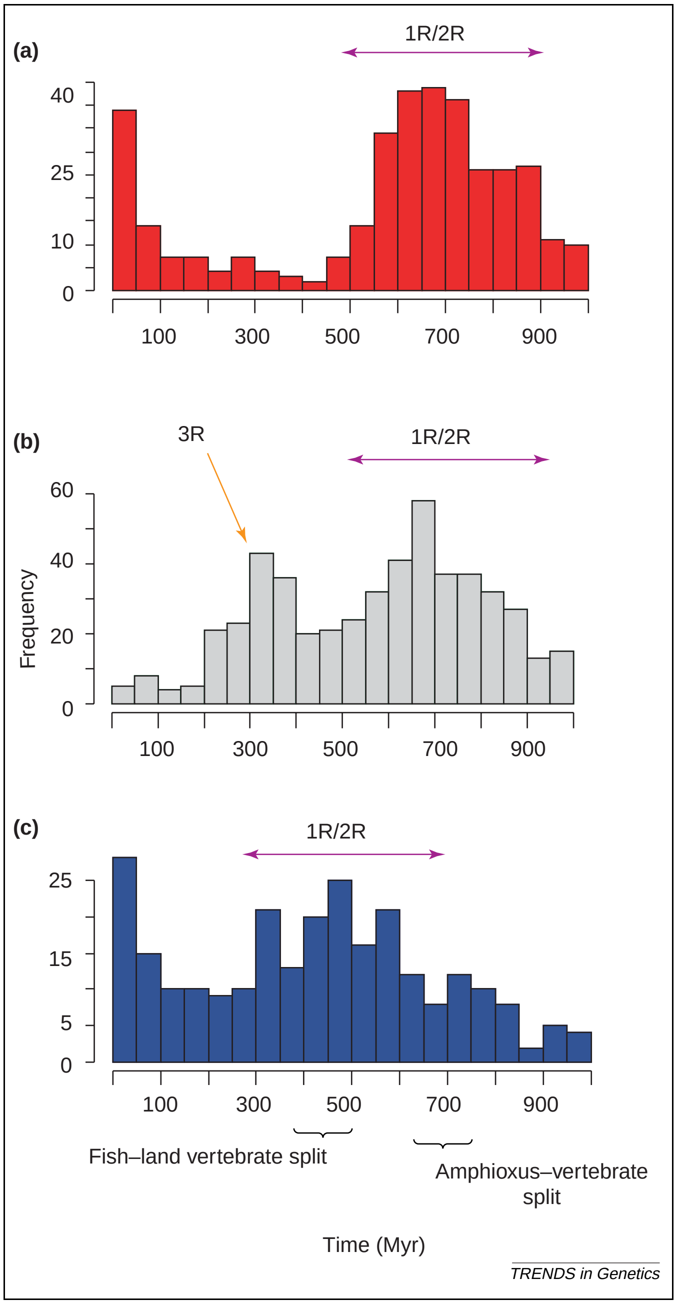

Figure 2. Distribution of the molecular-clock-based age estimates of duplicated genes in the **(a,c)** human and **(b)** *Fugu* genomes. The fish-specific duplication (marked by the orange arrow) becomes apparent when the two distributions are compared (a and b, data obtained from Vandepoele *et al.* [32]). The time estimates were obtained using the land vertebrate–fish divergence at 450 Mya as a calibration point [81,82]. (c) The dating of duplication events in the human genome, without including fish genes, and calibrating the phylogenetic trees with the chordate–arthropod divergence at 993 Mya, results in a shift of duplication ages and would overshadow the fish-specific duplications in an equivalent analysis (data from Panopoulou *et al.* [25]).

The search for 2R traces in the human genome

Stronger evidence for the type and number of duplication events can be obtained from the presence and arrangement of paralogons in the duplicated genome.

The Hox and MHC cluster paralogons

Did the Hox cluster regions on the human chromosomes arise at the origin of vertebrates? Were they the result of the same duplication event? Hughes *et al.* [24] found that the phylogeny of 14 non-Hox gene families with members on two or more of the human Hox-bearing chromosomes supported that they were duplicated before the vertebrate origin and even before the protostome–deuterostome split. Members of only five families were duplicated at the time expected by the 2R hypothesis (i.e. 750–528 Mya). Moreover, they argued that even these genes were not co-duplicated with the Hox clusters because their order of duplication [as indicated by the topology of the relevant phylogenetic trees (AB)(CD) measure (Box 1)] is not the same as that of the Hox cluster genes.

Larhammar *et al.* [41] stress that only genes that are ancestrally linked to the Hox clusters and not those purportedly transported on the Hox-bearing chromosomes at a later stage should be considered, because the present gene order on the Hox-bearing chromosomes is affected by the rearrangement history of the specific chromosomes. For example, Hsa 2 resulted from the fusion of two different chromosomes in the primate lineage and Hsa 12 was rearranged during primate evolution [42]. Finally, they point out that the phylogenetic-tree-topology approach is inaccurate in several cases and therefore of limited use in deciding if the members of neighbor gene families on the paralogons were generated by the same sequence of duplication events (Box 1). Larhammar *et al.* [41] concluded that 14 of the 20 families on the human Hox-cluster-bearing chromosomes were co-duplicated. These regions are significantly large and they cover 14.6%, 13.3%, 6.7% and 28.3% of the chromosomes 7, 17, 12, and 2, respectively. This makes it likely that they resulted from two rounds of WGD. These results also imply that specific chromosomal rearrangements should be considered when analyzing paralogons.

The debate over whether the human MHC cluster and its paralogous regions were generated through chromosomal duplication at the origin of vertebrates was resolved through the study of the relevant region in amphioxus. The comparison of the order of 31 amphioxus orthologs of genes in the MHC with those in its paralogous regions in the human genome [43], in addition to their fluorescence *in situ* hybridization (FISH) mapping results on a single amphioxus chromosome [44], supported the *en block* duplication of a proto-MHC region. Moreover, it showed that the MHC paralogous region on human chromosome 9 retains the ancestral organization because it contains twice as many genes derived from the ancestral genomic region as any of the three other regions on human chromosomes 1, 6 and 19.

Are there additional 2R-related paralogons in the human genome?

Paralogons that resulted from WGD as opposed to those from segmental duplications are expected to: (i) cover a significant portion of the genome; (ii) not overlap; (iii) contain duplicates that do not have a random distribution in the genome; (iv) include duplicated genes with similar duplication times; and (v) have an orientation that is similar to that of a closely related non-duplicated genome (if available).

Three studies have analyzed the human genome for 2R traces. Paralogons were characterized in terms of the number of duplicated genes they contain [i.e. smallest number of unique links (sm)] and the number of maximum unduplicated intervening genes allowed (d). Through a comparison of the distribution of duplicated genes within randomly shuffled genomes with those in the real human genome, McLysaght *et al.* [26] considered the paralogons that contain three or more duplicated genes to be statistically significant and detected 504 of these paralogons ($sm \geq 3$, $d=30$). In a similar study [25], we detected only 72 significant paralogons with an $sm=3$, but identified an additional 485 significant paralogons with an $sm=2$. This difference in the number and size of segments is because when we compared the frequency of the segment sizes from all human chromosomes with those of a randomized human chromosome set, we found that fewer intervening unduplicated genes should be allowed ($d=10$) to avoid including unrelated genes in the same paralogon.

Friedmann and Hughes [45] used a similar method to the studies above and identified fewer paralogons. This was due to the conservative threshold they imposed when defining gene families, which resulted in the selection of paralogons that contained recent duplicates. For a description of the methods developed to detect paralogons, see the review by Van de Peer [46].

Are the paralogons described in these studies the result of a WGD or even two rounds of WGD? The 2R paralogons with an average size of 0.7 Mb [25] are strikingly larger than the recently duplicated segments in the human genome (sequence similarity ≥90%, average size of 14.7–18.5 kb) [47]. Among the largest paralogons detected are a 41-Mb and 20-Mb region shared on chromosomes 1 and 9. Finally, they are comparable in size (in terms of number of duplicated genes they contain, length and percent of genome coverage) to the paralogons in the *Arabidopsis*, *S. cerevisiae* and rice genomes (Box 2). For example, they are almost the same size as the paralogons generated from a recent polyploidization event in the *Arabidopsis* genome but twice the length of the oldest paralogons in *Arabidopsis*. This could be because the old paralogons in the *Arabidopsis* genome were fragmented after the recent genome duplication.

The paralogons identified by McLysaght (with $sm \geq 3$ [26]) cover 79% of the genome (those with $sm \geq 4$ cover 64%) and they are many more than expected by chance. All of the paralogons described by us [25] do not overlap and are distributed evenly in the human genome, whereas segmental duplications in the human genome are found to be located preferentially towards telomeres and centromeres [48].

Are the paralogons in the human genome the result of the same ancient WGD(s)? McLysaght *et al.* [26] estimate that 40% of the gene pairs that duplicated at 397–695 Mya are components of paralogons with $sm \geq 3$. In the human genome, 331 of the duplicated segments are also duplicated in the mouse genome and located within known syntenic regions. Therefore, half of the paralogons are the result of a duplication event that affected the ancestor of both organisms and are at least >100 Myr old [25]. In conclusion, the analysis of paralogons strongly indicates that they are the result of a genome-wide-duplication event.

Was it a single genome-duplication event? In summary, only the quadrupled paralogons, 14 of which are identified in the human genome (Table 2 in the supplementary material online), represent evidence for two rounds of ancient WGDs. But can these 14 quadruplicate regions in the human genome statistically support the 2R hypothesis? Taking gene loss into account, and, therefore, relaxing the stringency of finding duplicates on three rather than on all four regions, the number of supporting anchor points raises dramatically to reveal a pattern of coparalogy, which indicates two rounds of duplication. Figure 3 illustrates this strategy with the example of the Hox-cluster-bearing chromosomal regions, which is often used as proof for 2R hypothesis.

Box 2. Traces of WGDs in plants and yeast

A. thaliana, underwent one recent polyploidization 24–40 Mya and one-to-two older polyploidizations ~170–300 Mya [31,74,75]. Evidence for the recent polyploidy event is provided by the presence of 45 duplicated block pairs with homogeneous age that cover 70% of the genome [31]. Traces of the older duplication are provided by 63 block pairs [31]. The recent blocks (median 700 kb; 69–4632 kb) are more than twice as long as the older blocks (median 284 kb; 90–1178 kb). Recent blocks contain more duplicated genes (on average 28%±7.8% duplicated genes) than the old blocks (on average 13.5%±5% duplicated genes). Other studies, based on K_s values of the duplicated genes in the duplicated segments, detect the same recent duplication event but date it at 65–100 Mya and argue that the old event was two duplication events at 170–235 Mya and at 300 Mya [75]. The recent event involves 26 segment pairs, whereas the old event involves 29 ‘large’ segment pairs. The recent event involves 83% of the transcriptome, whereas the two older polyploidizations involved 51.6% and 20.3% of the transcriptome, respectively. The average size of paralogons remaining in the *S. cerevisiae* genome that experienced genome duplication at 100 Mya is 55 kb, and includes on average 6.9 duplicate gene pairs and covers 50% of the genome [28]. Finally, the rice genome, which is thought to have undergone genome duplication ~70 Mya, retains nine non-overlapping duplicated segments that together account for 61.9% of the transcriptome; the size of each duplicated segment corresponds to 1.8–13.8% of the transcriptome [76].

First proof for WGD in vertebrates

Additional Hox clusters have been identified in teleost fish occupying different taxonomic positions (Figure 1). The mapping of Hox clusters and many duplicated genes in zebrafish [49,50], pufferfish [51] and medaka [52] suggested an extra WGD in ray-finned fish. The analysis of the *Fugu* genome revealed 159 statistically significant paralogons that contained 544 paralogous gene pairs (3.4 anchor points per block) [32]. Seventy percent of duplicated genes in the these paralogons (that carry 406 gene families) were duplicated at the vertebrate origin (525–875 Mya). One-third of the paralogons contain genes with an origin at 320 Mya. The last peak of gene burst (3R) is absent in the human genome and hence indicates a fish-specific large-scale duplication event (Figure 2). The number of fish-specific paralogons could be even greater [40].

The definitive proof of the teleost-specific genome duplication was only recently delivered when the genome of the pufferfish *Tetraodon nigroviridis* had been sequenced with remarkably good resolution. The

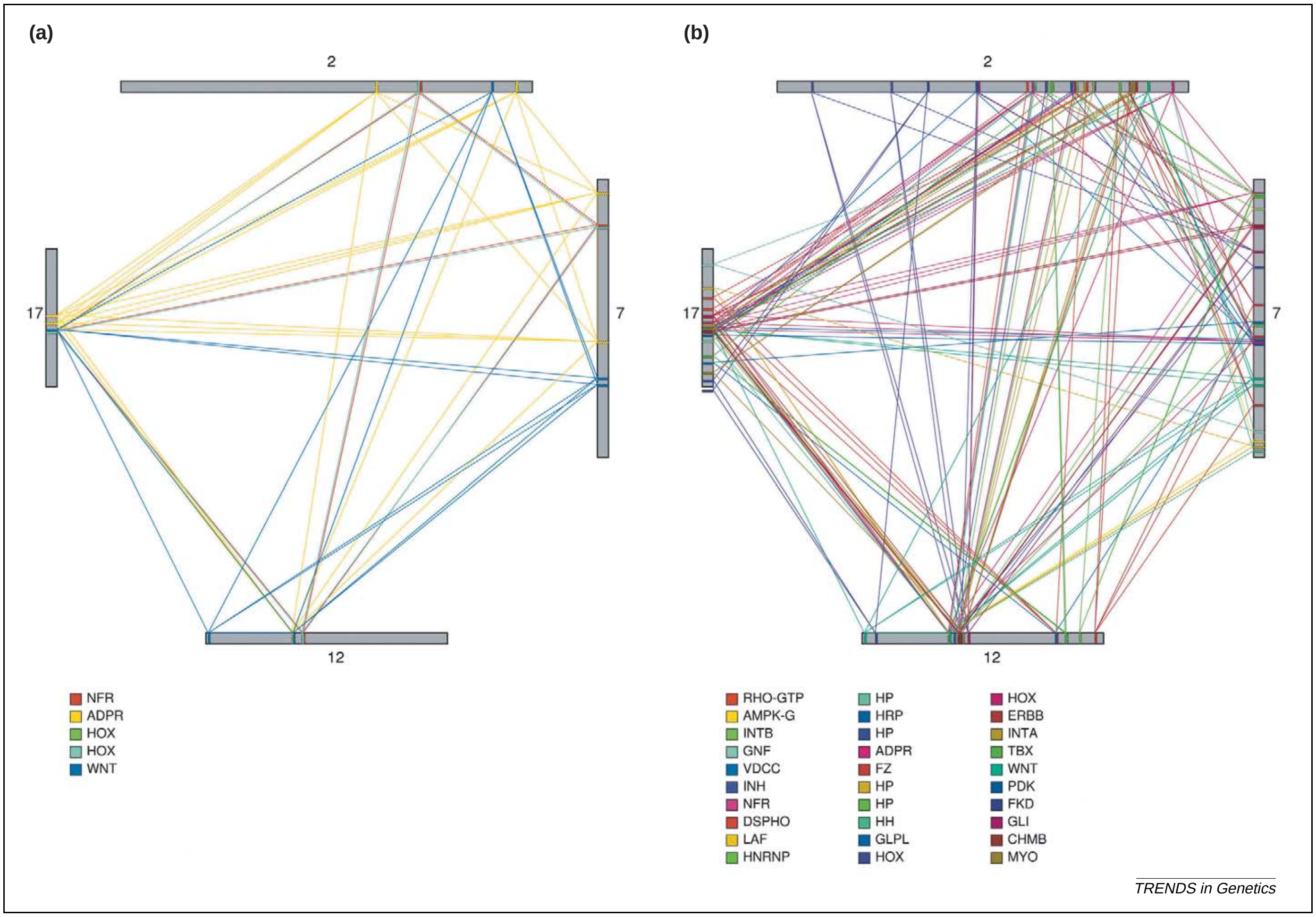

Figure 3. The example of the human Hox quadruplicated regions, the detection of which is frequently used as evidence for 2R and the effect of gene loss. The number of gene families shared between the human Hox-bearing chromosomes (Hsa 2, 7, 17 and 12) increases dramatically when the requirement of finding members of each family on three **(b)** rather than all four **(a)** chromosomes is used. To avoid including paralogons that contain genes that have been duplicated before the vertebrate origin, only gene families that have a single invertebrate ortholog were considered. The human chromosomes were searched for duplicated genes that are located next to each other on more than a single location. Each gene family is assigned a different color. The abbreviation of the name of each gene family is given at the bottom of the plot. For the complete names, see Table 1 in the supplementary material.

possibility of extensive anchoring of the sequence to chromosomes enabled Jaillon *et al.* [33] to use the comparative approach that demonstrated genome duplication in yeast [29]. Importantly, this approach is almost independent of dating methods. Using the human genome as the 'unduplicated' reference genome, they analyzed and mapped blocks of conserved syntenic regions to human chromosomes and identified 'doubly conserved synteny' (DCS). Strikingly, in most cases, syntenic blocks along human chromosomes map to two *Tetraodon* chromosomes in an interleaving pattern, consistent with WGD, and are distributed across all chromosomes (Figure 4).

Evidence for 2R from early vertebrates

The definitive proof that a more recent WGD occurred in teleost fish has important consequences for the 2R hypothesis because it indicates that WGD and not segmental duplication was the duplication mechanism responsible for the origin of the additional Hox clusters in this clade. Therefore, one could now accept that the Hox clusters are reliable markers of WGDs (Figure 1).

Both hagfish and lamprey genomes have been sampled so far mainly for Hox genes. Lamprey has at least four Hox clusters [53,54]. One study suggests that at least one Hox-cluster duplication occurred before the divergence of gnathostome and jawless vertebrates, whereas an independent cluster duplication occurred in the lamprey lineage, after it diverged from the gnathostome lineage [53]. Others argue for an independent origin of these clusters and suggest that the common ancestor of agnathans and gnathostomes had a single Hox cluster [55]. The phylogeny of other lamprey non-Hox gene families is also consistent with independent duplications [56–60]; therefore, it is plausible that some of the lamprey Hox clusters formed by the first duplication, before the split of agnathans from gnathostomes, were lost and the present Hox clusters are the result of recent lineage-specific WGDs.

Hagfish might have up to seven Hox clusters [61]. Two of them are homologous to mammalian Hox clusters, which also supports the hypothesis that at least one Hox-cluster duplication occurred in the ancestor of gnathostomes and agnathans. The two Hox clusters isolated so far from cartilaginous fish [62] are homologous to the mammalian *Hox*A and *Hox*D [63], placing the second '2R' duplication before the divergence of cartilagenous fish.

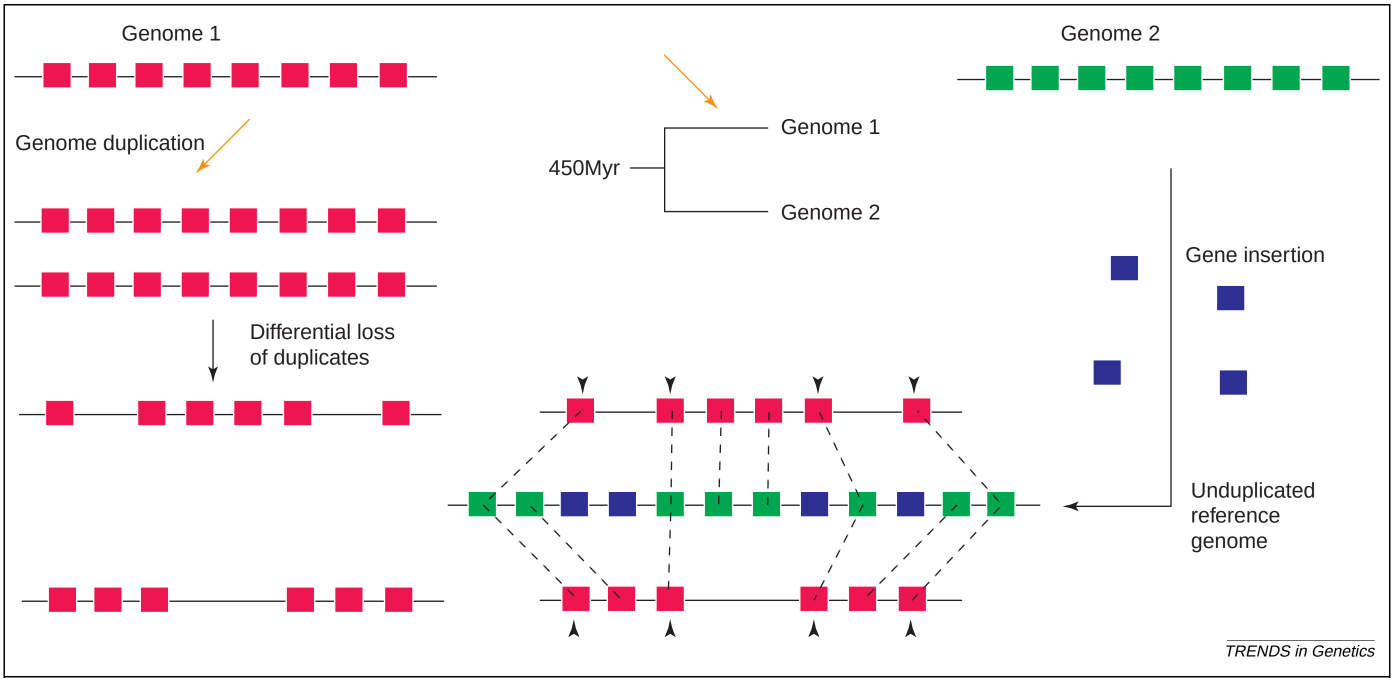

Figure 4. Illustration of the comparative approach used to prove genome duplications in yeast and *Tetraodon*. Genome 1 undergoes a Genome duplication (e.g. *Tetraodon*) creating two identical sets of chromosomes and genes followed by gene loss (left side). Genome 2 (e.g. human) experiences only some gene insertions and serves as 'unduplicated' reference genome. In most cases, large regions of 'double conserved synteny' can be identified (i.e. every chromosome of Genome 2 maps to two chromosomes of Genome 1 in an interleaving pattern; (middle lower panel). Genes that have been retained in two copies (arrowheads) would function as anchor points to identify a paralogon. The approach has been shown to be effective in detecting 'double conserved segments' in a genome that has undergone a WGD around 200–300 Mya and it has separated from its reference genome ~450 Mya.

Concluding remarks

Although polyploidy is a drastic event for a genome, it is not as rare. It is has long been known that natural polyploids are widespread in animal and plant genomes: 50% to >70% of angiosperms are thought to have experienced chromosome doubling [64]. Many amphibian [65] and fish [66] species are known for frequent recent polyploidy. Furthermore, the same amphibian species can be found with various ploidy levels [67]. Although the genome analysis of representative organisms of several of the above clades [32,33,40] has yielded solid evidence for their polyploidy, the 2R hypothesis has been exceedingly difficult to test.

Between 400 and 500 paralogons, with an average length of 700 kb that include 2511–3854 duplicated genes, cover almost 80% of the human genome. The number of paralogons is significantly more than is expected by chance and they are distributed across all chromosomes in a non-random fashion consistent with WGD. Most genes included in these paralogons were duplicated 350–650 Mya. Traces of this 'old' event have also been demonstrated in the teleost genomes. Taken together these findings provide strong support for at least one round of genome duplication early in the vertebrate lineage. This is even more compelling when these results are compared with the data from plant and vertebrate polyploids.

It is impossible to be certain, using current methods and based on the human or *Fugu* data, whether two rounds of duplication occurred and if they were in close succession because the duplication event is 'old'. Criteria such as the 1:4 rule or (AB)(CD) topology, which have been used to address this issue have led to false assumptions in several cases. The only strong evidence indicating that two duplications occurred is the existence of multiple quadruplicated regions in the human genome.

It has been suggested that because there are twice as many 'old' duplicates in the *Fugu* genome compared with the duplicates generated from a fish-specific duplication, and assuming an equal rate of gene loss after each duplication event, the 'old' duplication could be two rounds of duplication. This suggestion should be treated with caution because fewer recent *Fugu* duplicates might reflect the lower rate of gene retention following the second duplication event.

Will we be able to identify the type of 2R duplications that occurred? Some have suggested allopolyploidy might be responsible because it creates greater evolutionary potential than autopolyploidy [17]. Allopolyploids are assumed to be more viable because they are more prevalent in nature perhaps because allopolyploid genomes are 'dynamic' at the molecular level, generating an array of novel genomic instabilities during the initial stages after their formation. However, autopolyploidy is a process that results in asymmetric trees, which is what is observed in the majority of phylogenetic trees of genes duplicated at the origin of vertebrates. Currently, the data and methods available make it impossible to decide between allopolyploidy and autopolyploidy.

What next?

The complete genome sequence of lamprey or hagfish will help to resolve the timing of the duplications. The definitive answer to whether there were one or two rounds of ancient vertebrate genome duplications primarily rests in the upcoming amphioxus genome, which will serve as an unduplicated reference genome. Importantly, in addition to the complete sequence of these genomes, high-resolution genomic maps that will enable genes to be anchored to the chromosomes are required to tackle the problem if we are to employ the approach used in the *Tetraodon* genome.

Acknowledgements
We thank Steffen Hennig, Detlef Groth, James Adjaye and especially Hans Lehrach for stimulating discussions. This work was supported by the Max-Planck Gesellschaft zur Förderung der Wissenschaften e.v.

Supplementary data

Supplementary data associated with this article can be found at doi:10.1016/j.tig.2005.08.004

References

1 Ohno, S. (1970) *Evolution by Gene Duplication*, Springer-Verlag, Heidelberg
2 Holland, P.W. *et al.* (1994) Gene duplications and the origins of vertebrate development. *Dev. Suppl.* 43, 125–133
3 Garcia-Fernandez, J. and Holland, P.W. (1994) Archetypal organization of the amphioxus Hox gene cluster. *Nature* 370, 563–566
4 Skrabanek, L. and Wolfe, K.H. (1998) Eukaryote genome duplication – where"s the evidence? *Curr. Opin. Genet. Dev.* 8, 694–700
5 Ruddle, F.H. *et al.* (1994) Gene loss and gain in the evolution of vertebrates. *Dev.* (Suppl.) 155–161
6 Kasahara, M. (1996) Ancient chromosomal duplication involving the major histocompatibility complex. *Seikagaku* 68, 1717–1721
7 Katsanis, N. *et al.* (1996) Paralogy mapping: identification of a region in the human MHC triplicated onto human chromosomes 1 and 9 allows the prediction and isolation of novel PBX and NOTCH loci. *Genomics* 35, 101–108
8 Lundin, L.G. (1993) Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. *Genomics* 16, 1–19
9 Pebusque, M.J. *et al.* (1998) Ancient large-scale genome duplications: phylogenetic and linkage analyses shed light on chordate genome evolution. *Mol. Biol. Evol.* 15, 1145–1159
10 Wang, Y. and Gu, X. (2000) Evolutionary patterns of gene families generated in the early stage of vertebrates. *J. Mol. Evol.* 51, 88–96
11 Gibson, T.J. and Spring, J. (2000) Evidence in favour of ancient octaploidy in the vertebrate genome. *Biochem. Soc. Trans.* 28, 259–264
12 Hughes, A.L. (1998) Phylogenetic tests of the hypothesis of block duplication of homologous genes on human chromosomes 6, 9, and 1. *Mol. Biol. Evol.* 15, 854–870
13 Antequera, F. and Bird, A. (1993) CpG islands. *EXS* 64, 169–185
14 Fields, C. *et al.* (1994) How many genes in the human genome? *Nat. Genet.* 7, 345–346
15 Meyer, A. and Schartl, M. (1999) Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. *Curr. Opin. Cell Biol.* 11, 699–704
16 Sidow, A. (1996) Gen(om)e duplications in the evolution of early vertebrates. *Curr. Opin. Genet. Dev.* 6, 715–722
17 Spring, J. (1997) Vertebrate evolution by interspecific hybridization – are we polyploid? *FEBS Lett.* 400, 2–8
18 IHGSC. (2004) Finishing the euchromatic sequence of the human genome. *Nature* 431, 931–945
19 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351
20 Roest Crollius, H. *et al.* (2000) Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nat. Genet.* 25, 235–238
21 Dunham, I. *et al.* (1999) The DNA sequence of human chromosome 22. *Nature* 402, 489–495
22 Bork, P. and Copley, R. (2001) The draft sequences. Filling in the gaps. *Nature* 409, 818–820
23 Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
24 Hughes, A.L. *et al.* (2001) Ancient genome duplications did not structure the human Hox-bearing chromosomes. *Genome Res.* 11, 771–780
25 Panopoulou, G. *et al.* (2003) New evidence for genome-wide duplications at the origin of vertebrates using an amphioxus gene set and completed animal genomes. *Genome Res.* 13, 1056–1066
26 McLysaght, A. *et al.* (2002) Extensive genomic duplication during early chordate evolution. *Nat. Genet.* 31, 200–204
27 Lynch, M. and Conery, J.S. (2003) The evolutionary demography of duplicate genes. *J. Struct. Funct. Genomics* 3, 35–44
28 Wolfe, K.H. and Shields, D.C. (1997) Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387, 708–713
29 Kellis, M. *et al.* (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature* 428, 617–624
30 Dietrich, F.S. *et al.* (2004) The *Ashbya gossypii* genome as a tool for mapping the ancient Saccharomyces cerevisiae genome. *Science* 304, 304–307
31 Blanc, G. *et al.* (2003) A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Res.* 13, 137–144
32 Vandepoele, K. *et al.* (2004) Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proc. Natl. Acad. Sci. U. S. A.* 101, 1638–1643
33 Jaillon, O. *et al.* (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* 431, 946–957
34 Friedman, R. and Hughes, A.L. (2001) Pattern and timing of gene duplication in animal genomes. *Genome Res.* 11, 1842–1847
35 Holland, P.W. (2003) More genes in vertebrates? *J. Struct. Funct. Genomics* 3, 75–84
36 Pollard, S.L. and Holland, P.W. (2000) Evidence for 14 homeobox gene clusters in human genome ancestry. *Curr. Biol.* 10, 1059–1062
37 Maere, S. *et al.* (2005) Modeling gene and genome duplications in eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* 102, 5454–5459
38 Nei, M. *et al.* (2001) Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. *Proc. Natl. Acad. Sci. U. S. A.* 98, 2497–2502
39 Gu, X. *et al.* (2002) Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. *Nat. Genet.* 31, 205–209
40 Christoffels, A. *et al.* (2004) Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. *Mol. Biol. Evol.* 21, 1146–1151
41 Larhammar, D. *et al.* (2002) The human Hox-bearing chromosome regions did arise by block or chromosome (or even genome) duplications. *Genome Res.* 12, 1910–1920
42 Murphy, W.J. *et al.* (2001) Evolution of mammalian genome organization inferred from comparative gene mapping. *Genome Biol.*, 2. doi:10.1186/gb-2001-2-6-reviews0005 (http://genomebiology.com/2001/2/6/reviews/0005)
43 Abi-Rached, L. *et al.* (2002) Evidence of *en bloc* duplication in vertebrate genomes. *Nat. Genet.* 31, 100–105
44 Castro, L.F. *et al.* (2004) An antecedent of the MHC-linked genomic region in amphioxus. *Immunogenetics* 55, 782–784
45 Friedman, R. and Hughes, A.L. (2003) The temporal distribution of gene duplication events in a set of highly conserved human gene families. *Mol. Biol. Evol.* 20, 154–161
46 Van de Peer, Y. (2004) Computational approaches to unveiling ancient genome duplications. *Nat. Rev. Genet.* 5, 752–763
47 Zhang, L. *et al.* (2005) Patterns of segmental duplication in the human genome. *Mol. Biol. Evol.* 22, 135–141
48 Bailey, J.A. *et al.* (2001) Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.* 11, 1005–1017
49 Woods, I.G. *et al.* (2000) A comparative map of the zebrafish genome. *Genome Res.* 10, 1903–1914
50 Amores, A. *et al.* (1998) Zebrafish hox clusters and vertebrate genome evolution. *Science* 282, 1711–1714
51 Amores, A. *et al.* (2004) Developmental roles of pufferfish Hox clusters and genome evolution in ray-fin fish. *Genome Res.* 14, 1–10

52 Naruse, K. *et al*. (2004) A medaka gene map: the trace of ancestral vertebrate proto-chromosomes revealed by comparative gene mapping. *Genome Res.* 14, 820–828
53 Force, A. *et al*. (2002) Hox cluster organization in the jawless vertebrate Petromyzon marinus. *J. Exp. Zool.* 294, 30–46
54 Irvine, S.Q. *et al*. (2002) Genomic analysis of Hox clusters in the sea lamprey Petromyzon marinus. *J. Exp. Zool.* 294, 47–62
55 Fried, C. *et al*. (2003) Independent Hox-cluster duplications in lampreys. *J. Exp. Zoolog. B. Mol. Dev. Evol.* 299, 18–25
56 Neidert, A.H. *et al*. (2001) Lamprey Dlx genes and early vertebrate evolution. *Proc. Natl. Acad. Sci. U. S. A.* 98, 1665–1670
57 Uchida, K. *et al*. (2003) Development of the adenohypophysis in the lamprey: evolution of epigenetic patterning programs in organogenesis. *J. Exp. Zoolog. B. Mol. Dev. Evol.* 300, 32–47
58 Tomsa, J.M. and Langeland, J.A. (1999) Otx expression during lamprey embryogenesis provides insights into the evolution of the vertebrate head and jaw. *Dev. Biol.* 207, 26–37
59 McCauley, D.W. and Bronner-Fraser, M. (2004) Conservation and divergence of BMP2/4 genes in the lamprey: expression and phylogenetic analysis suggest a single ancestral vertebrate gene. *Evol. Dev.* 6, 411–422
60 Escriva, H. *et al*. (2002) Analysis of lamprey and hagfish genes reveals a complex history of gene duplications during early vertebrate evolution. *Mol. Biol. Evol.* 19, 1440–1450
61 Stadler, P.F. *et al*. (2004) Evidence for independent Hox gene duplications in the hagfish lineage: a PCR-based gene inventory of Eptatretus stoutii. *Mol. Phylogenet. Evol.* 32, 686–694
62 Kim, C.B. *et al*. (2000) Hox cluster genomics in the horn shark, *Heterodontus francisci*. *Proc. Natl. Acad. Sci. U. S. A.* 97, 1655–1660
63 Prohaska, S.J. *et al*. (2004) The shark HoxN cluster is homologous to the human HoxD cluster. *J. Mol. Evol.* 58, 212–217
64 Wendel, J.F. (2000) Genome evolution in polyploids. *Plant Mol. Biol.* 42, 225–249
65 Kawamura, T. (1984) Polyploidy in amphibians. *Zool Sci* 1, 1–5
66 Volff, J.N. (2005) Genome evolution and biodiversity in teleost fish. *Heredity* 94, 280–294
67 Becak, M.L. and Kobashi, L.S. (2004) Evolution by polyploidy and gene regulation in *Anura*. *Genet. Mol. Res.* 3, 195–212
68 Graur, D. and Martin, W. (2004) Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. *Trends Genet.* 20, 80–86
69 Hedges, S.B. and Kumar, S. (2004) Precision of molecular time estimates. *Trends Genet.* 20, 242–247
70 Reisz, R.R. and Muller, J. (2004) Molecular timescales and the fossil record: a paleontological perspective. *Trends Genet.* 20, 237–241
71 Furlong, R.F. and Holland, P.W. (2002) Were vertebrates octoploid? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 357, 531–544
72 Nembaware, V. *et al*. (2002) Impact of the presence of paralogs on sequence divergence in a set of mouse–human orthologs. *Genome Res.* 12, 1370–1376
73 Van de Peer, Y. *et al*. (2002) Dealing with saturation at the amino acid level: a case study based on anciently duplicated zebrafish genes. *Gene* 295, 205–211
74 Simillion, C. *et al*. (2004) Building genomic profiles for uncovering segmental homology in the twilight zone. *Genome Res.* 14, 1095–1106
75 Bowers, J.E. *et al*. (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422, 433–438
76 Paterson, A.H. *et al*. (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc. Natl. Acad. Sci. U. S. A.* 101, 9903–9908
77 Takezaki, N. *et al*. (2003) Molecular phylogeny of early vertebrates: monophyly of the agnathans as revealed by sequences of 35 genes. *Mol. Biol. Evol.* 20, 287–292
78 Chiu, C.H. *et al*. (2004) Bichir HoxA cluster sequence reveals surprising trends in ray-finned fish genomic evolution. *Genome Res.* 14, 11–17
79 Ikuta, T. *et al*. (2004) *Ciona intestinalis* Hox gene cluster: Its dispersed structure and residual colinear expression in development. *Proc. Natl. Acad. Sci. U. S. A.* 101, 15118–15123
80 Koh, E.G. *et al*. (2003) Hox gene clusters in the Indonesian coelacanth, *Latimeria menadoensis*. *Proc. Natl. Acad. Sci. U. S. A.* 100, 1084–1088
81 Zhu, M. and Yu, X. (2002) A primitive fish close to the common ancestor of tetrapods and lungfish. *Nature* 418, 767–770
82 Blair Hedges, S. and Kumar, S. (2003) *Genomic clocks and evolutionary timescales* 19, 200–206