

Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates

Yoichiro Nakatani,^{1,5} Hiroyuki Takeda,² Yuji Kohara,³ and Shinichi Morishita^{1,4,5}

¹Department of Computational Biology, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa 277-0882, Japan; ²Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Tokyo 113-0033, Japan;

³Center for Genetic Resource Information, National Institute of Genetics, Mishima 411-8540, Japan; ⁴Bioinformatics Research and Development (BIRD), Japan Science and Technology Agency (JST), Tokyo 102-8666, Japan

Although several vertebrate genomes have been sequenced, little is known about the genome evolution of early vertebrates and how large-scale genomic changes such as the two rounds of whole-genome duplications (2R WGD) affected evolutionary complexity and novelty in vertebrates. Reconstructing the ancestral vertebrate genome is highly nontrivial because of the difficulty in identifying traces originating from the 2R WGD. To resolve this problem, we developed a novel method capable of pinning down remains of the 2R WGD in the human and medaka fish genomes using invertebrate tunicate and sea urchin genes to define ohnologs, i.e., paralogs produced by the 2R WGD. We validated the reconstruction using the chicken genome, which was not considered in the reconstruction step, and observed that many ancestral proto-chromosomes were retained in the chicken genome and had one-to-one correspondence to chicken microchromosomes, thereby confirming the reconstructed ancestral genomes. Our reconstruction revealed a contrast between the slow karyotype evolution after the second WGD and the rapid, lineage-specific genome reorganizations that occurred in the ancestral lineages of major taxonomic groups such as teleost fishes, amphibians, reptiles, and marsupials.

[Supplemental material is available online at www.genome.org.]

Early vertebrate genome evolution has long been in need of clarification, and it is now of particular interest because several distantly related vertebrate genomes were recently sequenced. The 2R hypothesis postulates that two rounds of whole-genome duplication (2R WGD) occurred at the base of the vertebrate lineage (Ohno 1970; Holland et al. 1994) because of the observation that invertebrates have one *Hox* gene cluster, whereas lobe-finned fishes and land vertebrates have four clusters. However, the 2R hypothesis has been quite controversial until recently (Skrabaneck and Wolfe 1998; Gibson and Spring 2000; Friedman and Hughes 2001; Wolfe 2001; Abi-Rached et al. 2002; Furlong and Holland 2002; Gu et al. 2002; McLysaght et al. 2002; Durand 2003; Panopoulou et al. 2003; Seoighe 2003; Panopoulou and Poustka 2005; Vandepoele et al. 2004) because it leaves open the possibility of one round of WGD followed by large-scale duplications such as segmental and chromosomal duplications. Recently, Dehal and Boore (2005) showed that a large part of the human genome contains four-way paralogous chromosomal regions, which are traces of the 2R WGD.

The task of reconstructing ancestral vertebrate proto-chromosomes before the 2R WGD is very different from ordinary synteny analysis using orthologs because the effect of the 2R WGD must be carefully examined. Moreover, it is necessary to determine human chromosome regions originating from the same ancestral chromosome at the second WGD and integrate these regions to rebuild the ancestral karyotype. Thus, we first

identified groups of human genes (called “ohnologs” [Wolfe 2001]) duplicated by the 2R WGD by ensuring that individual genes in a group were most similar to the identical orthologous deuterostome gene of the sea urchin (Sea Urchin Genome Sequencing Consortium 2006) or tunicate (Dehal et al. 2002; Shoguchi et al. 2006). The identification of ohnologs is a difficult task because all ohnologs and their corresponding *Ciona* genes are rarely conserved due to numerous losses and duplications of *Ciona*, human, and medaka genes. Given these difficulties, ohnologs were grouped based on the method of Dehal and Boore (2005), which is detailed in Supplemental Figure S1 and the Supplemental Document. These ohnologs typically occur consecutively in paralogous chromosomal regions in the human genome and are likely to represent a remaining block derived from a single gnathostome (jawed vertebrate, see phylogenetic tree of vertebrates in Fig. 6, below) proto-chromosome. To test if they are really remaining blocks of the gnathostome proto-chromosomes, their synteny in the medaka genome (Kasahara et al. 2007) were investigated (for details, see the Supplemental Document). The final step in our novel analysis combined qualified remaining blocks into vertebrate and gnathostome proto-chromosomes using information on ohnolog distribution among the blocks. This series of steps was newly developed for this reconstruction (see Methods).

Next, we attempted to validate the reconstructed gnathostome proto-chromosomes. If extant genomes have experienced intensive interchromosomal rearrangements, they are hardly syntenic to the reconstructed ancestral genome and are not useful in the validation. Among sequenced vertebrate genomes, we searched for a genome that was not used in the reconstruction step but preserved the proto-karyotype. The chicken genome was

⁵Corresponding authors.

E-mail nakatani@cb.k.u-tokyo.ac.jp; fax 81-47-136-3977.

E-mail moris@cb.k.u-tokyo.ac.jp; fax 81-47-136-3977.

Article published online before print. Article and publication date are at <http://www.genome.org/cgi/doi/10.1101/gr.6316407>. Freely available online through the *Genome Research* Open Access option.

most promising, given the hypothesis that avian microchromosomes represent archaic linkage groups of ancestral vertebrates (Burt 2002). Indeed, we observed that many ancestral gnathostome proto-chromosomes in our reconstruction had one-to-one correspondence with microchromosomes in the chicken genome, thus providing a strong validation of our reconstruction.

We reconstructed the gnathostome (jawed vertebrate), osteichthyan (bony vertebrate), and amniote (the group including reptiles, birds, dinosaurs, and mammals) ancestral genomes, which are located at key phylogenetic positions, leading to a novel scenario of genome evolution in early vertebrates. The two rounds of WGD events duplicated 10–13 proto-chromosomes in the vertebrate ancestor, producing the gnathostome (jawed vertebrate) ancestor ($n \approx 40$). Subsequent chromosome fusions reduced the number of chromosomes in the osteichthyan proto-karyotype ($n \approx 31$) and in the amniote proto-karyotype ($n \approx 26$). These estimates of chromosome number are considerably larger than previous estimates (Jaillon et al. 2004; Naruse et al. 2004; Woods et al. 2005; Kohn et al. 2006) and contradict the widely held hypothesis that the osteichthyan proto-karyotype was $n \approx 12$ (Postlethwait et al. 2000; Jaillon et al. 2004; Naruse et al. 2004; Woods et al. 2005; Kohn et al. 2006) and that land-vertebrate genomes were shaped by lineage-specific chromosome “fissions” (Postlethwait et al. 2000). On the contrary, our results demonstrate that many lineage-specific chromosome “fusions” shaped the ancestral karyotypes of major taxonomic groups, such as teleost fishes, amphibians, reptiles, and marsupials.

Results

Evolution of vertebrate chromosomes

Figure 1 presents our scenario of vertebrate chromosome evolution. Although the ancestral genome would have more than two proto-chromosomes before the two rounds of WGD, for simplicity we display two of the reconstructed ancestral vertebrate chromosomes, colored red and blue (Fig. 1A). The first round of WGD doubled these two chromosomes, and fission occurred in one of the duplicated copies of the red chromosome. After the second round of WGD (Fig. 1B), sister chromosomes were gradually disrupted by genome rearrangements and broken into smaller blocks of chromosomal segments during early vertebrate evolution (Fig. 1C). Eventually, these blocks were distributed over several human chromosomes (Fig. 1D) because of intensive interchromosomal rearrangements in the mammalian lineage (Burt et al. 1999). However, in the teleost fish lineage, intensive chromosome fusions and another WGD occurred after the divergence from the osteichthyan ancestor.

As illustrated in Figure 1, conserved vertebrate linkage groups, which are groups of genes located on a single chromosome after the second round of WGD, constitute chromosomal blocks in the human genome. These blocks can also be identified by doubly conserved synteny (DCS) analysis with the medaka genome (Jaillon et al. 2004; Kellis et al. 2004). For example, although blocks 19a, 19b, and 19c are located consecutively in the human genome, both 19a and 19c are syntenic to each of the duplicated medaka chromosomes 4 and 17, whereas duplicates of 19b are found in two duplicated medaka chromosomes (1 and 8). Other neighboring blocks, i.e., 1b–1c, 3a–3b, 7a–7b, and 9a–9b, are located on distinct medaka chromosomes. This implies that deciding the boundaries of such neighboring blocks is extremely important; however, the task was difficult when genes on the

human chromosomes were compared with *Ciona* and sea urchin genes as an outgroup because of the intensive rearrangements in the mammalian lineage. We therefore needed additional information. Thus, we fully used the medaka chromosomes as an outgroup of mammals to clarify the boundaries (see details in Supplemental Fig. S2). The conserved vertebrate linkage (CVL) blocks were essential in reconstructing ancestral vertebrate proto-chromosomes.

Reconstruction of the vertebrate ancestral genome

Here, we assumed a simplified model in which no major genome rearrangements occurred between the 2R WGD events, although we will present a more elaborate model later in this paper. The expected signature of two rounds of WGD is illustrated in Figure 2, A–D. The 2R WGD quadruplicated the red and blue chromosomes, producing sister chromosomes each having the same set of genes arranged in the same order. In reality, gene losses and chromosome rearrangements must have taken place between the 2R WGD, but for simplicity, these details are omitted from Figure 2A. Wolfe (2001) proposed calling these duplicated gene pairs “ohnologs” after Susumu Ohno. One ohnolog is represented by a dot in a triangle whose X- and Y-axis coordinates represent the positions of duplicated genes in the sister chromosomes produced by the 2R WGD. Because ohnologs are produced by chromosome-wide duplication and not by smaller-scale gene duplication, they are observed between pairs of distinct sister chromosomes, as illustrated by dots in the red region. In contrast, no ohnologs would be found within one sister chromosome, which is shown by the absence of ohnolog dots in the green regions. After the 2R WGD, chromosome breaks (fissions and translocations), fusions, or inversions must have altered some of the sister chromosomes (Fig. 2B). The accumulation of rearrangements gradually disrupted the ancestral gene order and scattered the ohnolog dots (Fig. 2C). Furthermore, intensive interchromosomal rearrangements in the mammalian lineage (Burt et al. 1999) must have distributed blocks across the human genome (Fig. 2D). Reconstruction of ancestral proto-chromosomes is a task of reordering the CVL blocks in Figure 2D to estimate the orderings in Figure 2, B and C.

Figure 2E presents the actual distribution of ohnologs in the human genome, in which CVL blocks are placed in order from human chromosome 1 to X. The reconstruction step started with categorizing the blocks into groups that were derived from a single ancestral vertebrate chromosome before the 2R WGD. For this purpose, any two red CVL blocks sharing several ohnologs were placed in the same “vertebrate” group (Fig. 2F). The subsequent step divided the blocks within each group into subgroups that represented duplicated proto-chromosomes in the gnathostome ancestor. To this end, we performed an exhaustive search for the optimal subgrouping so that significantly more ohnologs were shared between distinct subgroups, as illustrated by the red regions in Figure 2G, whereas few ohnologs were observed within any single subgroup, as indicated by the green regions. The details of the reconstruction procedure are described in the Methods. Our reconstruction in Figure 2G shows that in five of 10 ancestral vertebrate chromosomes, estimating four sister chromosomes was more significant than inferring two, three, or five sister chromosomes. CVL blocks with a small number of ohnologs fail to be recognized as sister chromosomes because of their low statistical significance, thereby yielding a reconstructed vertebrate group with less than four duplicated chromosomes. In

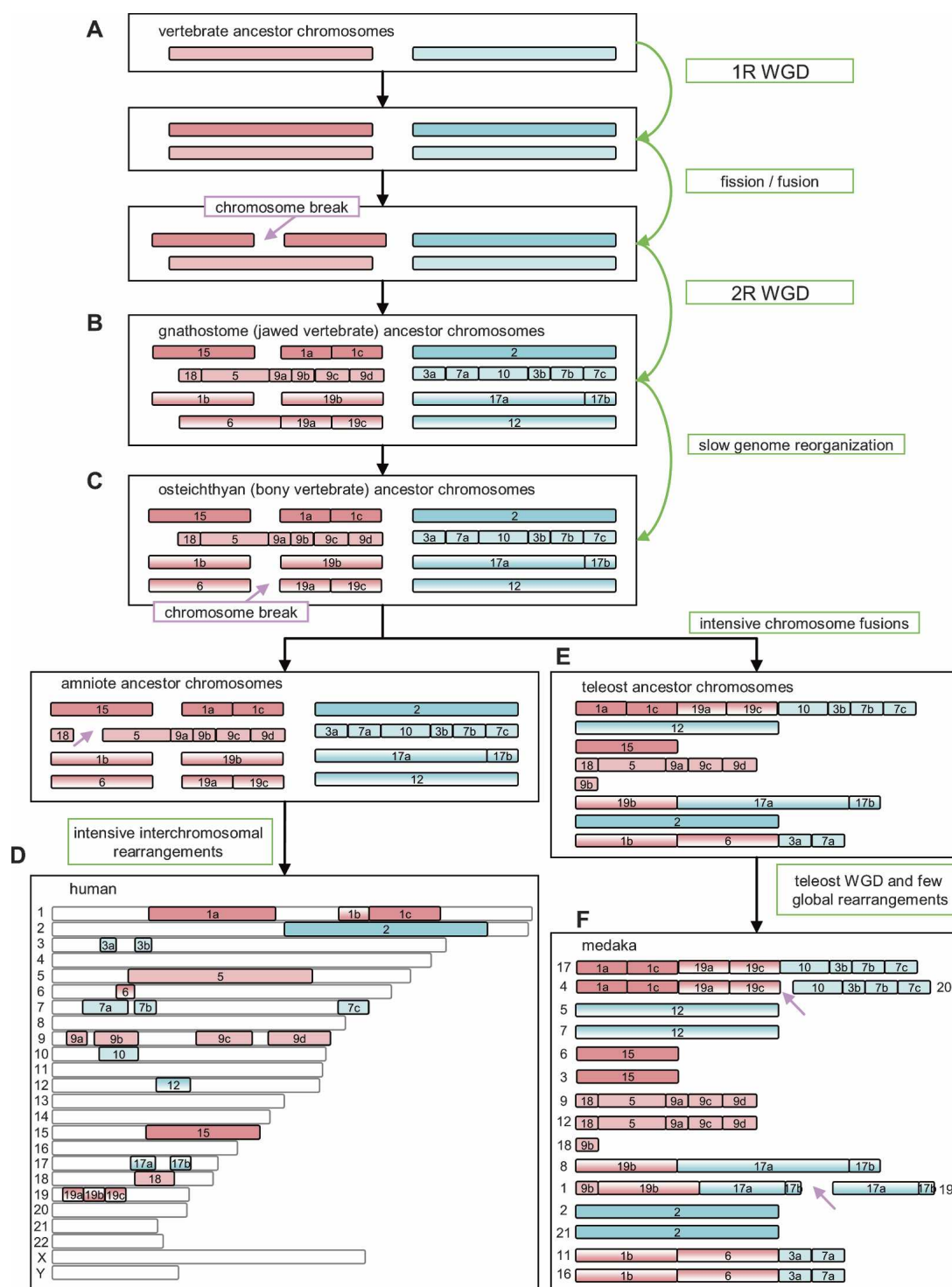
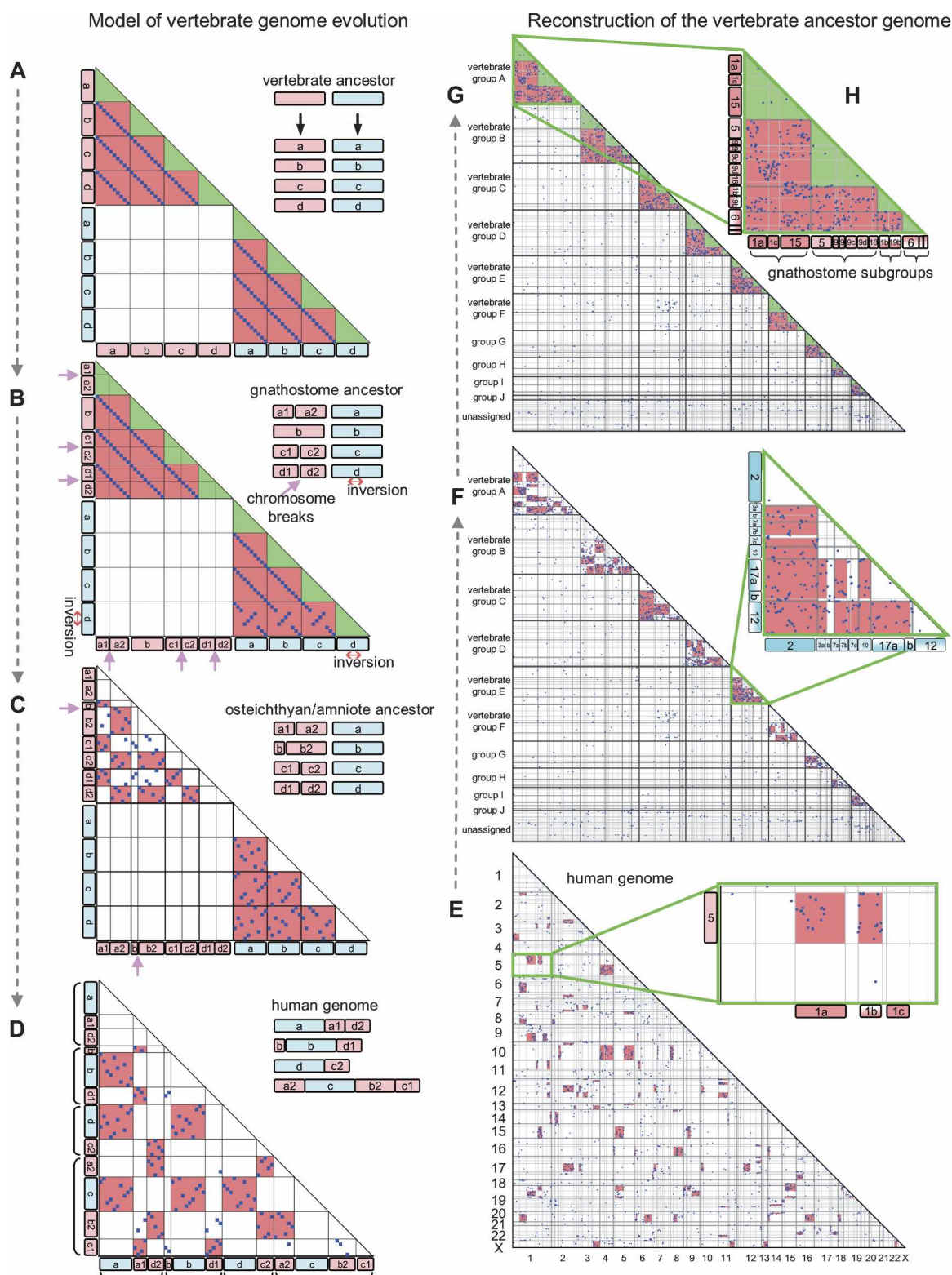


Figure 1. Vertebrate chromosome evolution scenario. (A) For simplicity, we illustrate two proto-chromosomes (red and blue bars) duplicated by the first round of WGD. Subsequently, fission divided one of the duplicated chromosomes. (B) The second round of WGD doubled the proto-chromosomes. Blocks in chromosomes are labeled with their respective chromosome positions in the human genome. (C) After the second WGD, early vertebrates underwent slow changes in karyotype over a long evolutionary process. (D) In the ancestral mammalian lineage, intensive interchromosomal rearrangements occurred and the ancestral chromosomes were broken into smaller segments that were distributed across many human chromosomes. (E) In the ancient ray-finned fish lineage, intensive chromosome fusions merged the ancestral chromosomal segments into ancestral teleost chromosomes. (F) Another round of WGD in the ancestral teleost doubled proto-chromosomes, but afterward, few global rearrangements shaped the present medaka genome.



contrast, five ancestral vertebrate chromosomes (A–E) had fairly large groups of CVL blocks with >250 ohnolog gene pairs, providing strong support for the 2R hypothesis.

Assuming fusions and fissions between the 2R WGD

For simplicity, we have thus far assumed no major chromosome rearrangements between the 2R WGD. Here, we extend the simplified model to consider the possibility of chromosome fusions and fissions between the 2R WGD events, though how to detect remains of these events is a nontrivial question. Figure 3A presents vertebrate groups A, B, and F in Figure 2F, and it is remarkable to observe pairs of light blue boxes with few ohnologs in individual triangles because these boxes are expected to contain many ohnologs according to the simplified model assumed so far. We think that these phenomena provide clues to the problem of inferring genome rearrangements between the 2R WGD; here, we show three alternative scenarios that may produce these phenomena.

Figure 3B shows the first scenario in which duplicates of two ancestral chromosomes are fused, whereas Figure 3C shows the second scenario, in which one duplicate of a single ancestral chromosome is divided by a fission event. Both scenarios generate the same ohnolog distribution pattern among the six descendant chromosomes, as illustrated in Figure 3E; note that pairs A21–B22 and B21–A22 have no ohnologs. The third scenario in Figure 3D indicates that neither fissions nor fusions take place between the 2R WGD events, but independent fission events occur for duplicated sister chromosomes after the 2R WGD. As illustrated in Figure 3F, the pair A21a–A22b is free of ohnologs; however, some ohnologs are observed in the pair A21b–A22a. Both pairs would have no ohnologs if the two independent fission events took place at the identical chromosomal positions, but this case is less likely. Among the three scenarios, the former two scenarios of one fusion or fission event are more parsimonious than the third scenario of two independent fissions, and they better explain pairs of light blue boxes with few ohnologs in Figure 3A. Thus, we estimated that among 10 vertebrate groups, A, B, and F had fusions or fissions, whereas the other seven groups were free of these chromosome rearrangements.

However, it remains difficult to judge whether a fusion or fission happened between the 2R WGD; each of the vertebrate groups A, B, and F may have had two ancestral chromosomes before the 2R WGD according to the fusion scenario, or one proto-chromosome if fission took place. This argument may be settled by using the genomes of some invertebrates as an out-group. However, the draft genome of the sea urchin (Sea Urchin Genome Sequencing Consortium 2006) consists of very small genomic fragments, and the *Ciona* draft genome (Shoguchi et al. 2006), with a nearly complete chromosome map, fails to preserve

proto-chromosomes before the divergence from vertebrates, presumably because of numerous chromosome rearrangements. Therefore, at this stage, we estimated that the ancestral vertebrate had 10–13 proto-chromosomes before the 2R WGD. We assert that the ancestral gnathostome had 40 proto-chromosomes, and was not affected by the selection of a fusion or fission scenario.

Validation of reconstructed proto-chromosomes

To check the validity of the reconstructed proto-chromosomes, we searched for a genome that preserved the ancestral chromosomes among available deuterostome, chordate, and vertebrate genomes. The sea urchin draft genome (Sea Urchin Genome Sequencing Consortium 2006) is largely fragmented into very small pieces, whereas the *Ciona* genome (Shoguchi et al. 2006) experienced numerous genome rearrangements after the divergence from vertebrates >500 million yr ago; for example, even *Hox* genes are separated onto multiple chromosomes (Ikuta et al. 2004). Thus, these two genomes fail to preserve proto-chromosomes. We also examined vertebrate genomes and found that rat, mouse, and dog genomes did not preserve the proto-karyotype because of several interchromosomal rearrangements in the mammalian lineage (Burt et al. 1999). Among tested genomes, the chicken genome significantly preserved the reconstructed ancestral gnathostome chromosomes. More precisely, a series of CVL blocks that occur both in one ancestral gnathostome chromosome and in one chicken chromosome presents strong evidence for the reconstruction; in contrast, occurrences of CVL blocks of one chicken chromosome in different sister chromosomes of the gnathostome ancestor suggest intensive interchromosomal rearrangements in the avian lineage or possible mistakes in the reconstruction. We observed that CVL blocks located in the same chicken microchromosomes belonged primarily to the same ancestral gnathostome chromosomes (Fig. 4). To be precise, chicken chromosomes 7, 8, 11, 15, 19, 20, 21, 22, 23, 24, 27, and 28 had one-to-one correspondence to E0, A2, B4, H0, H1, B1, F0, C3, B3, J0, E2, and A1, respectively (Fig. 5). Thus, this clear correspondence between chicken and gnathostome ancestor chromosomes provides a firm grounding for our reconstruction.

Reconstruction of ancestral osteichthyan and amniote proto-chromosomes

The reconstruction of proto-chromosomes in the gnathostome ancestor allowed us to infer the karyotypes of evolutionarily important ancestors such as the bony vertebrate (osteichthyan) ancestor and the amniote ancestor, leading to novel insights into genome evolution after the 2R WGD. Previous studies have suggested that the osteichthyan proto-karyotype was $n \approx 12$ by treat-

Figure 2. Model of vertebrate genome evolution and reconstruction of the ancestral genome. (A) For simplicity, suppose that the ancestral chromosome had 10 genes. The 2R WGD produced ohnologs (blue dots along the diagonal line in the triangular dot plot) in the duplicated chromosomes. (B) Chromosome breaks and inversions may have altered the order of ohnologs on the sister chromosomes. (C) In the course of early vertebrate genome evolution, the ancestral gene order was disrupted by many inversions, resulting in scattered ohnolog dots. (D) Eventually, CVL blocks were distributed across several human chromosomes by intensive interchromosomal rearrangements. (a–d) A typical model of genome evolution involving the 2R WGD. In the next step, we handle real human genome data. (E) This is a real instance of the dot plot in D. CVL blocks were ordered from the human chromosomes 1 to X, and ohnologs shared among these CVL blocks were plotted. (Red) Regions representing pairs of paralogous CVL blocks with a great number of ohnologs ($P < 10^{-4}$, see Methods). (F) This corresponds to the state in C. CVL blocks were reordered in such a way that paralogous CVL blocks were grouped so that each group represented one ancestral vertebrate chromosome (see Methods). (G) This state corresponds to that in B. CVL blocks within individual vertebrate groups were further reordered to obtain ancestral gnathostome subgroups (namely, chromosomes), which were duplicated from a single ancestral vertebrate chromosome by the 2R WGD events. The partition of subgroups that optimizes the significance defined in the Methods. Rectangles and triangles are colored in accordance with those in B to make the correspondence clear. (H) The vertebrate group A was decomposed into four gnathostome subgroups by statistical analysis, indicating that the ancestral chromosome underwent 2R WGD.

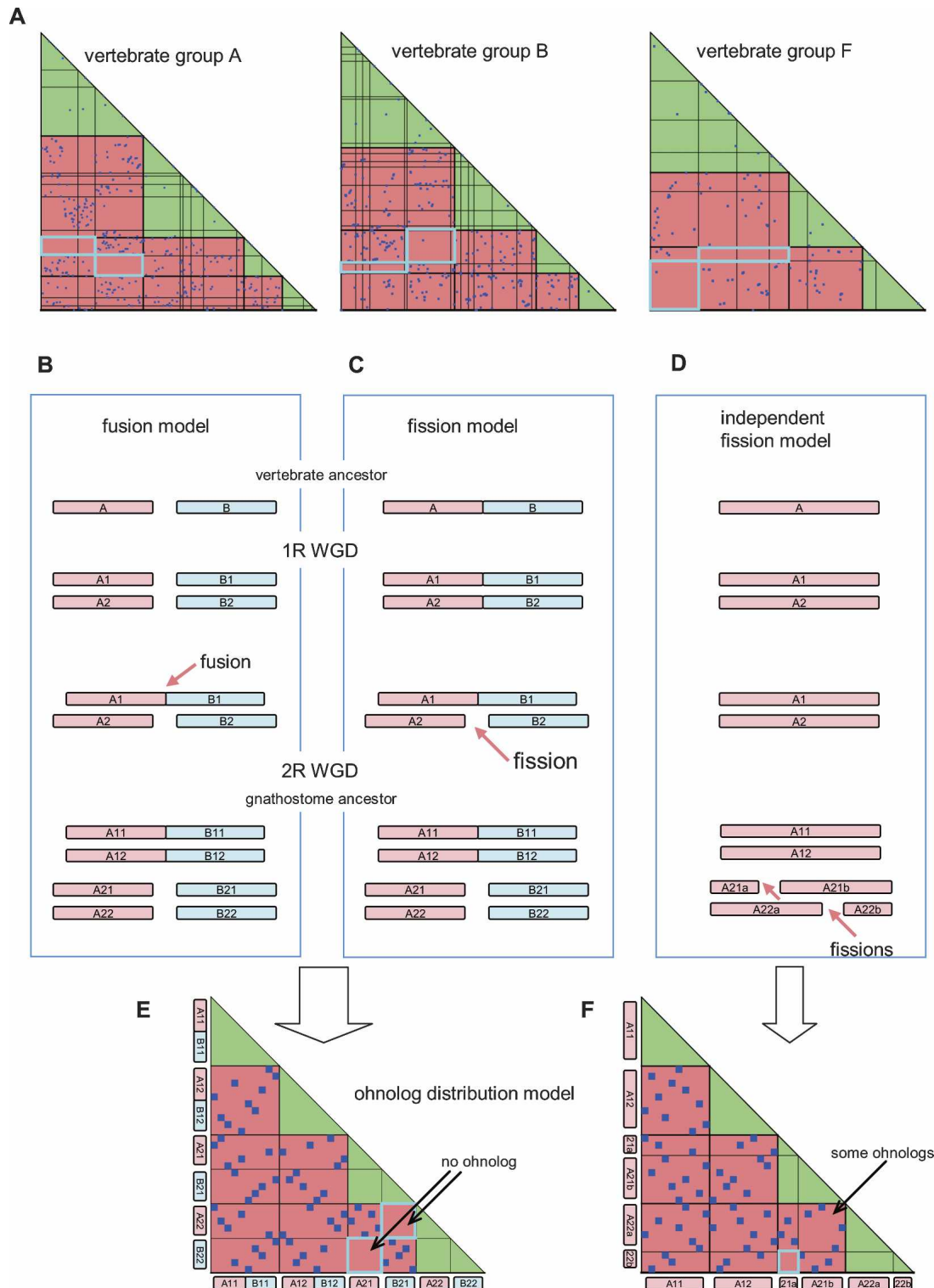


Figure 3. Models of chromosome evolution during 2R WGD. (A) Vertebrate groups A, B, and E in Fig. 2F. Pairs of light blue boxes with few ohnologs may be the remains of genome rearrangements between the two WGD events. (B) Two distinct ancestor chromosomes (A and B) were duplicated by the first WGD, and duplicated chromosomes A1 and B1 underwent a chromosome fusion. (C) One ancestor chromosome consisting of two parts (A and B) was duplicated by the first WGD, and a copy of the duplicated chromosomes was split into two chromosomes (A1 and B1) by a chromosome fission event. (D) After the second round of WGD, two of the four sister chromosomes underwent two independent chromosome fissions. Because the two fissions split two chromosomes (A11 and A12) at different chromosome positions, blocks A21b and A22a have some ohnologs in common. (E) Both of the fusion and fission models in B and C produce the same distribution of ohnologs, especially pairs of light blue boxes without ohnologs. (F) The distribution originating from independent fission events in D is distinguishable from that in E as indicated by one light blue box.

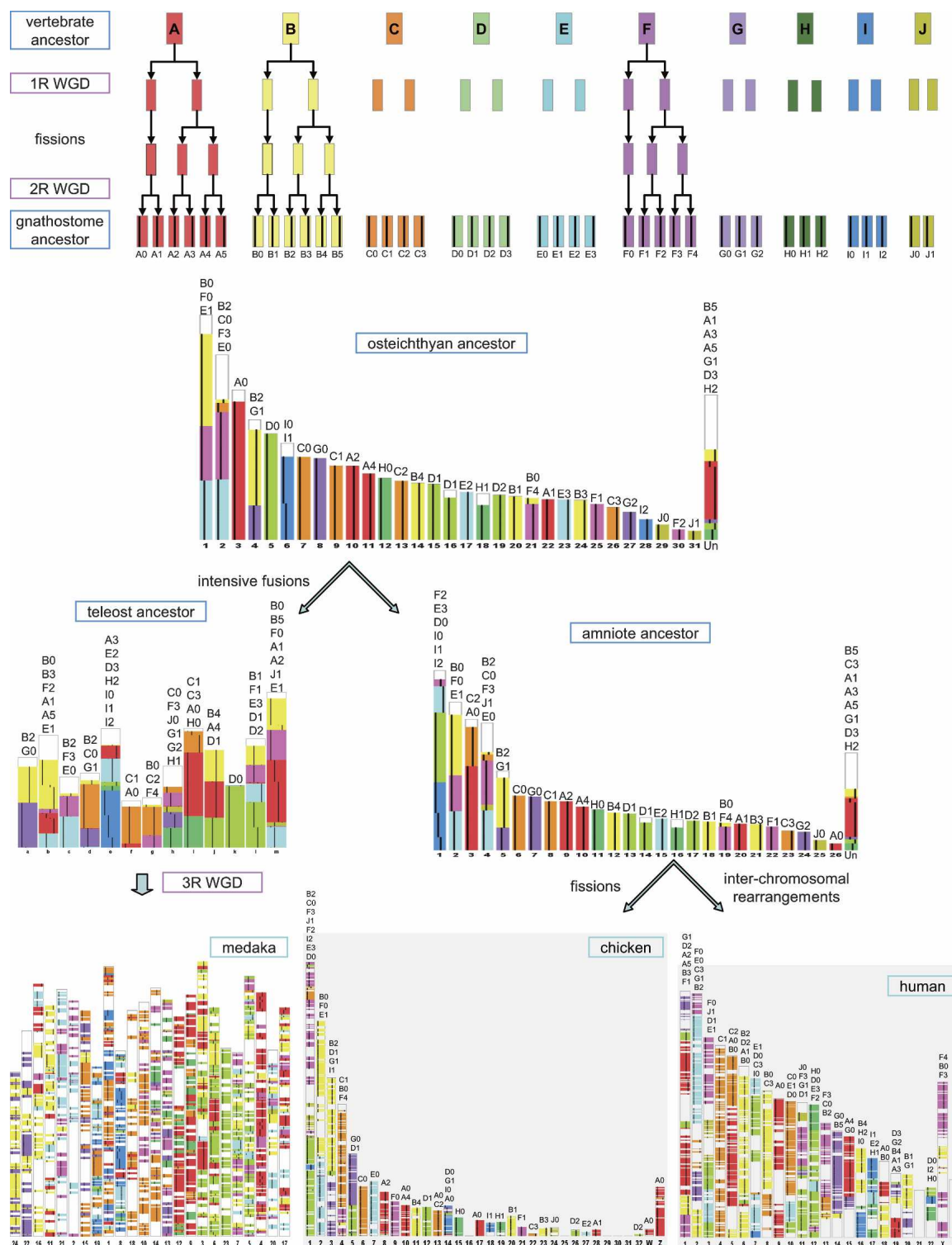


Figure 4. Reconstructed ancestral chromosomes. Ancestral vertebrate chromosomes A, B, and F had two alternative scenarios, fusions or fissions, between the 2R WGD events, as shown in Fig. 3. Thus, the number of proto-chromosomes ranges from 10 to 13 depending on the choice of two alternatives. The figure illustrates the scenario in which only fissions took place. Ten reconstructed proto-chromosomes in the vertebrate ancestor shown at the *top* are assigned distinct colors, and their daughter chromosomes in the gnathostome ancestor are distinguished by their respective vertical bars. In the genomes of the osteichthyan, teleost, and amniote ancestors, and human, chicken, and medaka genomes, genomic regions are assigned colors and vertical bars that represent correspondences of individual regions to the proto-chromosomes in the gnathostome ancestor from which respective regions originated. Unassigned blocks are shown in the *rightmost* chromosome (Un) in the osteichthyan and amniote ancestors.

Figure 5. Syntenic chromosome correspondence between the chicken and gnathostome ancestor chromosomes. In the table, the number in a cell indicates the number of orthologous genes in syntenic regions (see Supplemental materials) between the chicken chromosome in the column and the reconstructed gnathostome proto-chromosome in the row. To emphasize chicken chromosomes that were primarily derived from a single ancestral gnathostome chromosome, cells with the maximum number of orthologs are colored yellow or red. In particular, red cells imply chicken chromosomes that have one-to-one correspondence to their ancestral gnathostome chromosomes.

The reconstructed ancestral karyotypes in Figure 4 provide a scenario for karyotype evolution in vertebrates, as depicted in Figure

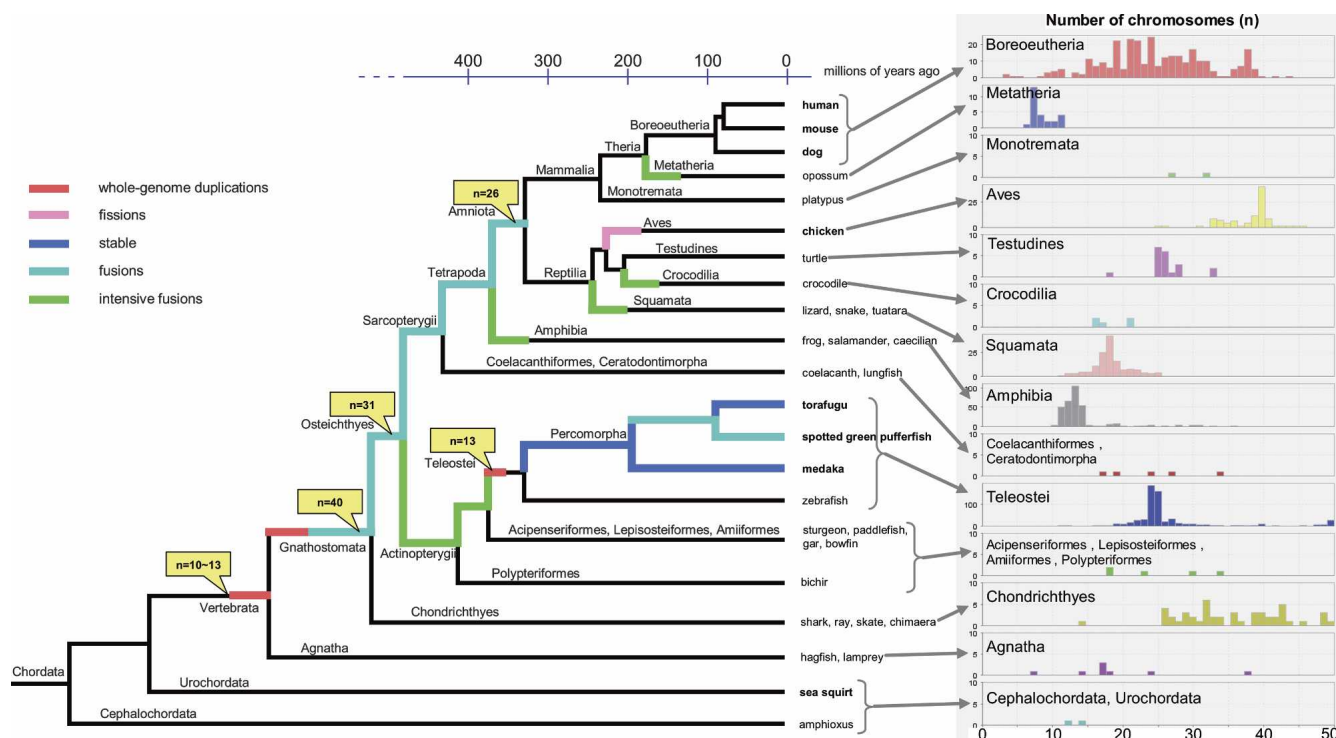


Figure 6. Changes in chromosome number during vertebrate karyotype evolution. The reconstructed proto-chromosomes in Fig. 4 allow us to discuss how the number of chromosomes has changed in individual vertebrate lineages. (Left) The phylogenetic tree of vertebrates, (right) the distribution of chromosome number in individual lineages (Gregory et al. 2007; Animal Genome Size Database, <http://www.genomesize.com>). Considering the numbers of proto-chromosomes in Fig. 4, two ancient whole-genome duplication events almost quadrupled the number of chromosomes; subsequently, chromosome numbers in individual lineages tended to decrease in many lineages, although not in the avian lineage.

6. Before the first round of WGD, the vertebrate ancestor karyotype was $n \approx 10-13$, and the subsequent 2R WGD and some genome rearrangements yielded the gnathostome ancestor of $n \approx 40$. After the divergence of bony vertebrates (Osteichthyes) and cartilaginous fishes (Chondrichthyes), genome rearrangements reduced the number of chromosomes in the osteichthyan ancestor to $n \approx 31$. In the meantime, modern cartilaginous fishes have genomes comprising a variety of karyotypes; the number of chromosomes in the haploid set, n , ranges from 25 to 50. We speculate that some of extant cartilaginous fishes may retain the gnathostome ancestral karyotype ($n \approx 40$), assuming no genome duplication events took place in the cartilaginous lineage. After the divergence of ray-finned and lobe-finned fishes, in the lineage of ray-finned fishes (Actinopterygii), chromosome fusions reduced the number of chromosomes and produced the teleost ancestor with $n \approx 13$. Our fusion model made it necessary to revise the $n \approx 12$ ancestral osteichthyan proto-karyotype hypothesis, which treats the teleost ancestor of ~ 350 Mya as the representative of the osteichthyan ancestor of ~ 450 Mya (Postlethwait et al. 2000; Jaillon et al. 2004; Naruse et al. 2004; Woods et al. 2005; Kohn et al. 2006). Subsequently, the whole-genome duplication in the teleost ancestor doubled the number of chromosomes to $n \approx 26$. The number of chromosomes in the teleost lineage has remained nearly unchanged during evolution, and the chromosome numbers of extant teleost species peak at $n = 24$ or 25.

In the lineage of lobe-finned fishes (Sarcopterygii), the distribution of chromosome number differs largely across vertebrate taxa (Fig. 6). In particular, an interesting observation in our

model is that the ancestors of major vertebrate groups such as teleosts, amphibians, squamates, and metatherians underwent numerous chromosome fusions (except for birds), leading to slow changes in karyotype, as indicated by the distribution of chromosome number. Similar speculations have been made regarding amphibians and reptiles; Morescalchi postulated that the number of chromosomes decreased in the ancestral amphibian based on the large number of chromosomes in some primitive amphibians (Green and Sessions 1991). With respect to reptiles, Olmo (2005) indicated that ancestral reptiles had a higher rate of karyotypic changes. Our karyotype evolution model is consistent with these hypotheses.

In contrast, a large increase in the number of chromosomes was found in the avian lineage. One explanation is that fewer repeat sequences in avian genomes were likely to result in chromosome fissions rather than fusions and interchromosomal rearrangements (Burt et al. 1999; Burt 2002). Another interesting finding involved microchromosomes in the avian lineage. Burt argued that many avian microchromosomes were already present in the common ancestor of birds and other terrestrial vertebrates and were conserved for >400 million yr (Burt 2002). Our analysis extends this idea considerably by showing that many of the microchromosomes had one-to-one correspondence with proto-chromosomes in the gnathostome ancestor.

We conclude that in early vertebrate genome evolution, two ancient whole-genome duplication events increased the number of chromosomes, but subsequent chromosome fusions reduced the number of chromosomes in individual lineages, with the exception of the avian lineage.

Conclusions

We present the first study to reconstruct the vertebrate proto-karyotype before the 2R-WGD events. The reconstruction was made possible by the newly developed computational method outlined in Figures 2 and 3. The reconstructed genome organization of the vertebrate ancestor provides new insights into early vertebrate genome evolution. First, some rearrangements occurred during the 2R WGD in the ancestral vertebrate. Second, after the second WGD, the gnathostome, osteichthyan, and amniote ancestors underwent slow karyotype evolution. Third, rapid, lineage-specific chromosome fusions shaped the ancestral genomes of major vertebrate taxa such as teleosts, amphibians, reptiles, and marsupials.

Our reconstruction will also integrate newly sequenced genomes such as the pipid frog (*Xenopus tropicalis*), gray short-tailed opossum (*Monodelphis domestica*), and amphioxus (*Branchiostoma floridae*) into the global picture of vertebrate and chordate genome evolution. Our presumption is that the *X. tropicalis* genome evolved from the osteichthyan ancestor genome after undergoing numerous chromosome fusions. Because those fusions took place specifically in the ancestral amphibian lineage, they would be largely different from fusions that took place from the osteichthyan ancestor to the teleost ancestor, which will make it difficult to identify one-to-one correspondence between chromosomes of *X. tropicalis* and the teleost ancestor.

Methods

Protein sequences for genes and identification of ohnologs

Human, mouse, dog, chicken, and *Tetraodon* protein sequences were obtained from Ensembl (Birney et al. 2006) (v.36–Dec2005); *Takifugu* (version 2) from Ensembl; *Ciona intestinalis* (version 1) from JGI (<http://www.jgi.doe.gov>); and *Strongylocentrotus purpuratus* (version 2.1) from NCBI. If multiple sequences overlapped in the genome, they were represented by the longest sequence. We conducted all-against-all BLASTP searches (Altschul et al. 1997) ($E\text{-value} < 1 \times 10^{-10}$). Special treatments were necessary to identify ohnologs produced by the 2R WGD (see details in the Supplemental materials).

Identification of paralogous CVL blocks for grouping CVL blocks

We evaluated the significance of the number of ohnologs, denoted by n , that two CVL blocks share in common by evaluating the probability that the two CVL blocks have $\geq n$ ohnologs in common under the condition that all ohnologs are distributed at random in all CVL blocks. The probability is defined as follows: Let N be the total number of genes in all CVL blocks, m be the total number of ohnologs among all CVL blocks, and x and y be the respective number of genes in CVL blocks X and Y . Assuming a random distribution of ohnologs among CVL blocks, the probability of observing $\geq n$ ohnologs is approximated by

$$p = \sum_{i=n}^m \binom{m}{i} q^i (1-q)^{m-i}, \text{ where } q = \frac{2xy}{N(N-1)}.$$

Here, p is the summation of the probability that $i (\geq n)$ among m ohnologs are members of X and Y with the probability q that an arbitrary pair of genes selected from all $N(N-1)/2$ pairs becomes an ohnolog shared by X and Y .

If the probability was at most a reasonable threshold, we rejected the null hypothesis that ohnologs between the two CVL

blocks were produced by independent gene duplications and concluded that the two CVL blocks were paralogous. The crux is how to select a meaningful value as the threshold. We identified 118 CVL blocks (for details, see the Supplemental Document). Because the total number of all pairs of CVL blocks amounted to 6903, the maximum threshold of the probability was set to 10^{-4} so that the expected number of false-positive paralogous pairs was < 1 . In Figure 2, E and F, regions representing paralogous CVL blocks are colored red if the P -value $< 10^{-4}$.

Reconstruction of sister chromosomes in the gnathostome ancestor by grouping paralogous CVL blocks

We began by constructing an undirected graph in which the nodes are CVL blocks and edges are drawn between two paralogous CVL blocks. The graph was then divided into connected components so that there was a path between any two nodes in a connected component, as illustrated in Supplemental Figure S7. CVL blocks represented by nodes in one connected component were thought to be derived from one vertebrate proto-chromosome before the 2R WGD, as illustrated in Figure 2H.

Subsequently, to estimate proto-chromosomes in the gnathostome ancestor that were produced by WGD events, nodes in one connected component were further partitioned into several subgroups of nodes so that each subgroup represented one proto-chromosome in the gnathostome ancestor. The number of gnathostome proto-chromosomes originating from one vertebrate proto-chromosome was expected to be four assuming that the 2R WGD events took place (Dehal and Boore 2005), but we enumerated patterns of partitioning CVL blocks into two, three, four, or five subgroups to search for the most significant pattern. Specifically, for each connected component, all combinations of CVL blocks were enumerated with the constraint that any paralogous CVL blocks could not be combined into the same subgroup. To define the significance of one partition, CVL blocks in one subgroup (one gnathostome proto-chromosome) are unlikely to share ohnologs, as illustrated in the green regions of Figure 2; however, in contrast, many pairs of CVL blocks in distinct gnathostome proto-chromosomes are likely to share a significant number of ohnologs, as shown by red boxes in Figure 2. We therefore measured the significance of one partition with n ohnologs among different subgroups (i.e., n ohnolog dots in the red regions in Figure 2H) by the probability that $\geq n$ ohnologs are shared by different subgroups under the random distribution of ohnologs in the whole triangular region. More precisely, the probability, denoted by p , is defined by assuming that c is the total number of pairs of genes in the focal connected component, m is the total number of ohnologs in the connected component, and r is the total number of pairs of genes in different subgroups of the partition. Intuitively, c is the area of the whole triangular region in Figure 2H, and r is the total area of red regions. The probability p is the total number of cases when each pair of paralogous CVL blocks in different subgroups of the partition shares $\geq n$ ohnologs divided by the total number of cases when m ohnologs appear in pairs of two arbitrary CVL blocks:

$$p = \sum_{i=n}^m \binom{r}{i} \binom{c-r}{m-i} / \binom{c}{m}.$$

The partition with the minimum probability is the most significant one, and the subgroups in the partition are treated as gnathostome proto-chromosomes.

In the literature, some algorithms have been proposed for reconstructing pre-duplicated genomes (El-Mabrouk and Sankoff 2003; Kellis et al. 2004). The idea of doubly conserved synteny

analysis (Kellis et al. 2004) was used to compute CVL blocks (see details in the Supplemental Document). However, our reconstruction procedure is largely different from the approach of El-Mabrouk and Sankoff, which assumes rigorous constraints that are too restrictive to be applicable to the analysis of the vertebrate genome evolution with two rounds of whole-genome evolution; for example, two copies of individual genes must occur in the genome. In reality, many duplicated genes are lost, and studying the number of whole-genome duplication events in early vertebrates is a major research objective, thereby making it meaningless to assume that the number of WGD events is exactly two.

The 2-of-3 rule for ancestral karyotype reconstruction

The bony vertebrate (osteichthyan) proto-chromosomes were reconstructed using the gnathostome ancestor together with the chicken and teleost ancestor by applying the 2-of-3 rule that two CVL blocks are in the same bony vertebrate proto-chromosome if the two appear in an identical chromosome of at least two of the three following karyotypes: chicken, teleost ancestor, and gnathostome ancestor. The 2-of-3 rule indicates three scenarios of genome evolution. The first scenario is the presence of the two CVL blocks in one chromosome throughout the evolution of the genome. The second is that the two occurred in the same chromosome in the chicken and teleost ancestor genomes, which implies the presence of the two in the same proto-chromosome of the bony vertebrate ancestor; however, the two appeared in distinct proto-chromosomes in the gnathostome ancestor, suggesting that a chromosomal fusion merged the different proto-chromosomes into one in the bony vertebrate ancestor. The third scenario is that the two CVL blocks appeared in the same chromosome in the gnathostome ancestor and teleost ancestor (or alternatively in the chicken), but the two were located in distinct chromosomes in the chicken (or in the teleost ancestor) by some genome rearrangements after the divergence of the chicken and the teleost ancestor. The 2-of-3 rule was also applied to reconstruction of the amniote ancestor proto-chromosomes of the human and chicken genomes by using the osteichthyan ancestor. The identification of syntenic regions in the chicken genome is described in the Supplemental Information.

One might be concerned that when the proto-karyotype of the bony vertebrate ancestor was reconstructed, the karyotypes of the gnathostome ancestor, teleost ancestor, and chicken were not independent data because the gnathostome ancestor was reconstructed using only part of the medaka genome information. Here, we describe how we avoided such circular arguments. We separated two types of information, namely CVL blocks with conserved ohnologs in the human genome and chromosomes on which individual blocks are located. We used the former information in the reconstruction of the gnathostome ancestor genome and the latter in estimating proto-chromosomes of osteichthyan and amniote ancestors. Thus, when the gnathostome ancestor genome was reconstructed, CVL blocks in the human genome were identified by using the medaka genome to clarify the boundaries of blocks; subsequently, blocks were clustered into proto-chromosomes in the gnathostome ancestor. The critical point was that blocks were combined into gnathostome proto-chromosomes without using any information on the human and medaka chromosomes in which individual blocks were located. Stated another way, even if human and medaka chromosome numbers were blacked out, the same gnathostome proto-chromosomes could be reconstructed. We grouped these blocks into the chromosomes of the gnathostome ancestor solely using information on ohnologs that were shared among blocks. In contrast, human chromosome numbers of individual blocks

in the gnathostome proto-chromosomes were used to find counterpart blocks in the medaka and chicken chromosomes in the step of reconstructing proto-chromosomes in osteichthyan and amniote ancestors.

Acknowledgments

This work was supported by a Grant-in-Aid for Scientific Research on Priority Areas "Genome" from the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT), and the grant program for Bioinformatics Research and Development (BIRD), Japan Science and Technology Agency (JST). We thank Asao Fujiyama, Shinichi Hashimoto, Kiyoshi Naruse, Masahiro Kasahara, Shin Sasaki, Wei Qu, Budrul Ahsan, and Tomoyuki Yamada. Computation time was provided by the Super Computer System, Human Genome Center, Institute of Medical Science, University of Tokyo.

References

- Abi-Rached, L., Gilles, A., Shiina, T., Pontarotti, P., and Inoko, H. 2002. Evidence of en bloc duplication in vertebrate genomes. *Nat. Genet.* **31**: 100–105.
- 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.
- Birney, E., Andrews, D., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cox, T., Cunningham, F., Curwen, V., Cutts, T., et al. 2006. Ensembl 2006. *Nucleic Acids Res.* **34**: D556–D561.
- Burt, D.W. 2002. Origin and evolution of avian microchromosomes. *Cytogenet. Genome Res.* **96**: 97–112.
- Burt, D.W., Bruley, C., Dunn, I.C., Jones, C.T., Ramage, A., Law, A.S., Morrice, D.R., Paton, I.R., Smith, J., Windsor, D., et al. 1999. The dynamics of chromosome evolution in birds and mammals. *Nature* **402**: 411–413.
- Dehal, P. and Boore, J.L. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.* **3**: e314. doi: 10.1371/journal.pbio.0030314.
- Dehal, P., Satou, Y., Campbell, R.K., Chapman, J., Degnan, B., De Tomaso, A., Davidson, B., Di Gregorio, A., Gelpke, M., Goodstein, D.M., et al. 2002. The draft genome of *Ciona intestinalis*: Insights into chordate and vertebrate origins. *Science* **298**: 2157–2167.
- Durand, D. 2003. Vertebrate evolution: Doubling and shuffling with a full deck. *Trends Genet.* **19**: 2–5.
- El-Mabrouk, N. and Sankoff, D. 2003. The reconstruction of doubled genomes. *SIAM J. Comput.* **32**: 754–792.
- Friedman, R. and Hughes, A.L. 2001. Pattern and timing of gene duplication in animal genomes. *Genome Res.* **11**: 1842–1847.
- Furlong, R.F. and Holland, P.W. 2002. Were vertebrates octoploid? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **357**: 531–544.
- Gibson, T.J. and Spring, J. 2000. Evidence in favour of ancient octaploidy in the vertebrate genome. *Biochem. Soc. Trans.* **28**: 259–264.
- Green, D.M. and Sessions, S.K. 1991. *Amphibian cytogenetics and evolution*. Academic Press, San Diego.
- Gregory, T.R., Nicol, J.A., Tamm, H., Kullman, B., Kullman, K., Leitch, I.J., Murray, B.G., Kapraun, D.F., Greilhuber, J., and Bennett, M.D. 2007. Eukaryotic genome size databases. *Nucleic Acids Res.* **35**: D332–D338.
- Gu, X., Wang, Y.F., and Gu, J.Y. 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.
- 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.
- Ikuta, T., Yoshida, N., Satoh, N., and Saiga, H. 2004. *Ciona intestinalis* Hox gene cluster: Its dispersed structure and residual colinear expression in development. *Proc. Natl. Acad. Sci.* **101**: 15118–15123.
- Jaillon, O., Aury, J.M., Brunet, F., Petit, J.L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., et al. 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* **431**: 946–957.
- Kasahara, M., Naruse, K., Sasaki, S., Nakatani, Y., Qu, W., Ahsan, B., Yamada, T., Nagayasu, Y., Doi, K., Kasai, Y., et al. 2007. The medaka

- draft genome and insights into vertebrate genome evolution. *Nature* **447**: 714–719.
- Kellis, M., Birren, B.W., and Lander, E.S. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428**: 617–624.
- Kohn, M., Hogel, J., Vogel, W., Minich, P., Kehrer-Sawatzki, H., Graves, J.A., and Hameister, H. 2006. Reconstruction of a 450-My-old ancestral vertebrate protokaryotype. *Trends Genet.* **22**: 203–210.
- McLysaght, A., Hokamp, K., and Wolfe, K.H. 2002. Extensive genomic duplication during early chordate evolution. *Nat. Genet.* **31**: 200–204.
- Naruse, K., Tanaka, M., Mita, K., Shima, A., Postlethwait, J., and Mitani, H. 2004. A medaka gene map: The trace of ancestral vertebrate proto-chromosomes revealed by comparative gene mapping. *Genome Res.* **14**: 820–828.
- Ohno, S. 1970. *Evolution by gene duplication*. Springer-Verlag, New York.
- Olmo, E. 2005. Rate of chromosome changes and speciation in reptiles. *Genetica* **125**: 185–203.
- Panopoulou, G. and Poustka, A.J. 2005. Timing and mechanism of ancient vertebrate genome duplications—The adventure of a hypothesis. *Trends Genet.* **21**: 559–567.
- Panopoulou, G., Hennig, S., Groth, D., Krause, A., Poustka, A.J., Herwig, R., Vingron, M., and Lehrach, H. 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.
- Postlethwait, J.H., Woods, I.G., Ngo-Hazelett, P., Yan, Y.L., Kelly, P.D., Chu, F., Huang, H., Hill-Force, A., and Talbot, W.S. 2000. Zebrafish comparative genomics and the origins of vertebrate chromosomes. *Genome Res.* **10**: 1890–1902.
- Sea Urchin Genome Sequencing Consortium. 2006. The genome of the sea urchin *Strongylocentrotus purpuratus*. *Science* **314**: 941–952.
- Seoighe, C. 2003. Turning the clock back on ancient genome duplication. *Curr. Opin. Genet. Dev.* **13**: 636–643.
- Shoguchi, E., Kawashima, T., Satou, Y., Hamaguchi, M., Sin-I, T., Kohara, Y., Putnam, N., Rokhsar, D.S., and Satoh, N. 2006. Chromosomal mapping of 170 BAC clones in the ascidian *Ciona intestinalis*. *Genome Res.* **16**: 297–303.
- Skrabanek, L. and Wolfe, K.H. 1998. Eukaryote genome duplication—Where's the evidence? *Curr. Opin. Genet. Dev.* **8**: 694–700.
- Vandepoele, K., De Vos, W., Taylor, J.S., Meyer, A., and Van de Peer, Y. 2004. Major events in the genome evolution of vertebrates: Paraneome age and size differ considerably between ray-finned fishes and land vertebrates. *Proc. Natl. Acad. Sci.* **101**: 1638–1643.
- Wolfe, K.H. 2001. Yesterday's polyploids and the mystery of diploidization. *Nat. Rev. Genet.* **2**: 333–341.
- Woods, I.G., Wilson, C., Friedlander, B., Chang, P., Reyes, D.K., Nix, R., Kelly, P.D., Chu, F., Postlethwait, J.H., and Talbot, W.S. 2005. The zebrafish gene map defines ancestral vertebrate chromosomes. *Genome Res.* **15**: 1307–1314.

Received January 23, 2007; accepted in revised form June 11, 2007.



Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates

Yoichiro Nakatani, Hiroyuki Takeda, Yuji Kohara, et al.

Genome Res. 2007 17: 1254-1265 originally published online July 25, 2007

Access the most recent version at doi:[10.1101/gr.6316407](https://doi.org/10.1101/gr.6316407)

Supplemental Material

<http://genome.cshlp.org/content/suppl/2007/09/05/gr.6316407.DC1>

References

This article cites 32 articles, 11 of which can be accessed free at:
<http://genome.cshlp.org/content/17/9/1254.full.html#ref-list-1>

Open Access

Freely available online through the *Genome Research* Open Access option.

License

Freely available online through the Genome Research Open Access option.

Email Alerting Service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).

ThruPLEX[®] HV
failproof DNA-seq of FFPE & cfDNA



Chromtech **TAKARA** cellartis

To subscribe to *Genome Research* go to:
<http://genome.cshlp.org/subscriptions>
