

Whole-Genome Duplications Spurred the Functional Diversification of the Globin Gene Superfamily in Vertebrates

Federico G. Hoffmann,^{1,2} Juan C. Opazo,³ and Jay F. Storz*¹

¹School of Biological Sciences, University of Nebraska

²Department of Biochemistry and Molecular Biology, Mississippi State University

³Instituto de Ecología y Evolución, Facultad de Ciencias, Universidad Austral de Chile, Valdivia, Chile

*Corresponding author: E-mail: jstorz2@unl.edu.

Associate editor: Adriana Briscoe

Abstract

It has been hypothesized that two successive rounds of whole-genome duplication (WGD) in the stem lineage of vertebrates provided genetic raw materials for the evolutionary innovation of many vertebrate-specific features. However, it has seldom been possible to trace such innovations to specific functional differences between paralogous gene products that derive from a WGD event. Here, we report genomic evidence for a direct link between WGD and key physiological innovations in the vertebrate oxygen transport system. Specifically, we demonstrate that key globin proteins that evolved specialized functions in different aspects of oxidative metabolism (hemoglobin, myoglobin, and cytoglobin) represent paralogous products of two WGD events in the vertebrate common ancestor. Analysis of conserved macrosynteny between the genomes of vertebrates and amphioxus (subphylum Cephalochordata) revealed that homologous chromosomal segments defined by myoglobin + globin-E, cytoglobin, and the α -globin gene cluster each descend from the same linkage group in the reconstructed proto-karyotype of the chordate common ancestor. The physiological division of labor between the oxygen transport function of hemoglobin and the oxygen storage function of myoglobin played a pivotal role in the evolution of aerobic energy metabolism, supporting the hypothesis that WGDs helped fuel key innovations in vertebrate evolution.

Key words: cytoglobin, genome duplication, globin gene family, hemoglobin, myoglobin.

Introduction

Gene duplications and whole-genome duplications (WGDs) are thought to have played major roles in promoting evolutionary innovation. It has been hypothesized that two rounds of WGD in the stem lineage of vertebrates (dubbed the “1R” and “2R” WGDs; Meyer and Schartl 1999; McLysaght et al. 2002; Dehal and Boore 2005; Hoegg and Meyer 2005; Putnam et al. 2008) helped fuel the evolutionary innovation of several vertebrate-specific features, such as the endoskeleton, the neural crest and derivative cell types, neurogenic placodes, various signaling transduction pathways, and a complex segmented brain (Ohno 1970; Holland et al. 1994; Meyer 1998; Shimeld and Holland 2000; Wada 2001; Wada and Makabe 2006; Zhang and Cohn 2008; Braasch, Volff, et al. 2009; Larhammar et al. 2009; Van de Peer et al. 2009). An additional lineage-specific WGD in the common ancestor of teleost fishes (Amores et al. 1998; Meyer and Schartl 1999; Taylor et al. 2003; Jaillon et al. 2004; Meyer and Van de Peer 2005; Kasahara et al. 2007; Kassahn et al. 2009) appears to have helped fuel the diversification of many physiological and developmental pathways in this highly speciose vertebrate group (Braasch et al. 2006, 2007; Braasch, Brunet, et al. 2009; Sato et al. 2009). However, evidence in support of a causal link between WGDs and the origins of vertebrate-specific innovations is typically indirect, as it is seldom possible to trace novel phenotypes to specific functional differences be-

tween paralogous products of a WGD event (Van de Peer et al. 2009).

Beyond a certain threshold of body size and internal complexity, the passive diffusion of oxygen is generally not sufficient to meet the metabolic demands of animal life. Early in vertebrate evolution, maintenance of an adequate cellular oxygen supply in support of aerobic metabolism was made possible by a circulatory system that actively delivers oxygen to cells throughout the body in combination with the use of specialized oxygen transport and oxygen storage proteins (Goodman et al. 1975, 1987; Dickerson and Geis 1983). Here, we report genomic evidence demonstrating that the second of these two innovations can be traced to a combination of WGD and subsequent small-scale duplications in the stem lineage of vertebrates.

Two rounds of WGD in the vertebrate common ancestor (as postulated by the 2R hypothesis) would initially produce multigene families with four members in vertebrates that are co-orthologous to a single invertebrate gene (the 4:1 rule; Meyer and Schartl 1999). However, following two rounds of WGD, only a small minority of gene families would be expected to retain all four of the resultant paralogs, and subsequent gene turnover via small-scale duplications and deletions would further obscure the signal of WGD (Pébusque et al. 1998; Abi-Rached et al. 2002; Horton et al. 2003; Dehal and Boore 2005; Braasch et al. 2006, 2007;

Braasch, Brunet, et al. 2009). For this reason, we assessed the role of WGD in the diversification of the globin gene superfamily by combining phylogenetic analyses with genomic analyses of the global physical organization of paralogous genes. Results of the combined phylogenetic and genomic analyses revealed that precursors of key globin proteins that evolved specialized functions in different aspects of oxidative metabolism and oxygen signaling pathways (hemoglobin [Hb], myoglobin [Mb], and cytoglobin [Cygb]) represent paralogous products of two WGD events in the vertebrate common ancestor.

Materials and Methods

Data Collection

We used bioinformatic searches to identify the full complement of globin genes in the genomes of 13 vertebrate species: 5 teleost fish (fugu, *Takifugu rubripes*; medaka, *Oryzias latipes*; pufferfish, *Tetraodon nigroviridis*; three-spined stickleback, *Gasterosteus aculeatus*; and zebrafish, *Danio rerio*), 1 amphibian (western clawed frog, *Xenopus tropicalis*), 1 squamate reptile (green anole lizard, *Anolis carolinensis*), 3 birds (chicken, *Gallus gallus*; turkey, *Meleagris gallopavo*; and zebra finch, *Taeniopygia guttata*), and 3 mammals (human, *Homo sapiens*; gray short-tailed opossum, *Monodelphis domestica*; and platypus, *Ornithorhynchus anatinus*). Vertebrate globin genes were annotated by comparing known coding sequences with genomic contigs using the program BLAST2 sequences version 2.2 (Tatusova and Madden 1999). In addition, we retrieved the complete repertoire of globin genes from two nonvertebrate chordates: the sea squirt, *Ciona intestinalis* (a urochordate, taken from Ebner et al. 2003) and amphioxus, *Branchiostoma floridae* (a cephalochordate).

Phylogenetic Reconstruction of the Globin Gene Superfamily in Chordates

To identify the globin gene lineage in nonvertebrate chordates that exhibits the closest phylogenetic affinity to vertebrate-specific globins, we added the complete repertoire of globin genes from the sea squirt and amphioxus to an alignment of globin gene sequences from a representative set of vertebrate taxa. The alignment included amino acid sequences of *Cygb*, *globin E* (*GbE*), *globin Y* (*GbY*), α - and β -chain *Hbs*, *Mb*, *globin X* (*GbX*), and neuroglobin (*Ngb*; the full list of sequences with the corresponding accession numbers is presented in [supplementary table S1, Supplementary Material](#) online). We aligned amino acid sequences using the L-INS-i strategy from MAFFT version 6.7 (Katoh and Toh 2008) and performed maximum likelihood and Bayesian phylogenetic searches using a mixed model of amino acid substitution. Maximum likelihood searches were carried out using Treefinder, version October 2008 (Jobb et al. 2004). Support for the nodes was assessed with 1,000 bootstrap replicates. Bayesian searches were conducted with MrBayes version 3.1.2 (Ronquist and Huelsenbeck 2003), setting two independent runs of four simultaneous chains for 10,000,000 generations, sampling every 2,500 generations, and using default priors. Once convergence

was verified by tracking the split frequency, support for the nodes and parameter estimates were derived from a majority rule consensus of the last 2,500 trees. In all cases, trees were rooted with the clade that includes vertebrate *Ngb* and *GbX* and their amphioxus counterparts. To compare alternative phylogenetic hypotheses (i.e., reciprocal monophyly vs. paraphyly of vertebrate-specific globins relative to the clade containing amphioxus *Gb2*, *Gb8*, and *Gb14*), we conducted parametric bootstrapping tests (Swofford et al. 1996; Goldman et al. 2000). For each simulated data set, we calculated the difference in likelihood score, Δ , between the null hypothesis maximum likelihood topology and the alternative hypothesis maximum likelihood topology. Using an α -level of 0.01, the null hypothesis was rejected if ≥ 0.99 of the simulation-based Δ values exceeded the observed value.

Analyses of Chromosomal Homology

We searched for the existence of globin-defined paralogs by using the program CHSminer (Wang et al. 2009) to identify pairs of homologous chromosomal segments that share paralogous gene duplicates. CHSminer is designed to identify pairs of homologous chromosomal segments of size s (the number of paralogous gene duplicates that are shared between chromosomal segments) with a specified gap size g (the number of unshared genes located between shared paralogs). We delineated homologous chromosomal segments containing the *Cygb*, *Hb*, and *Mb* genes in the chicken and human genomes using assignments of gene family membership provided by Ensembl, (release 58) with $s = 10$ and $g = 50$. We used the program Genomicus (Muffato et al. 2010) to characterize patterns of conserved macrosynteny among globin-defined paralogs from each of the 13 examined genome assemblies.

Analyses of Co-duplicated Genes

We performed additional phylogenetic analyses to assess whether gene duplicates that are shared among the globin-defined paralogs showed evidence of co-duplication in the stem lineage of vertebrates. For gene families in which each of the constituent members is linked to a different vertebrate-specific globin gene, we assessed whether the set of paralogs originated via duplication events that occurred prior to the divergence between teleost fish and tetrapods (the earliest phylogenetic split among the vertebrate taxa included in our study). Sequence data for the vertebrate paralogs were obtained from Ensembl release 58, and invertebrate orthologs were identified in the EnsemblCompara database (Vilella et al. 2009). Additional putative invertebrate orthologs were identified using BLAST (Altschul et al. 1990). In each case, we aligned amino acid sequences using the L-INS-i strategy from MAFFT version 6.7 (Katoh and Toh 2008), we performed maximum likelihood searches using Treefinder, version October 2008 (Jobb et al. 2004), and we evaluated support for the nodes with 1,000 bootstrap replicates.

Assessment of Conserved Synteny with the Ancestral Chordate Proto-karyotype

If the vertebrate-specific globins are derived from two rounds of WGD in the stem lineage of vertebrates, we would expect the globin-defined paralogs to exhibit quadruple-conserved synteny relative to the genomes of nonvertebrate chordates like amphioxus. We therefore identified the genomic map positions of all vertebrate-specific globins and mapped them to the reconstructed proto-karyotype of the chordate common ancestor (Putnam et al. 2008). In human chromosomes 7, 16, 17, 19, and 22, the break points of conserved synteny with ancestral Chordate Linkage Group 15 were derived from Putnam et al. (2008), and the locations of shared gene duplicates that unite the globin-defined paralogs are based on annotations derived from Ensembl.

Results

Phylogenetic Relationships among Chordate Globins

Using an alignment that contained globin sequences from a diverse array of chordate taxa, we reconstructed phylogenetic relationships of all vertebrate-specific globins and the complete repertoire of globin genes from two nonvertebrate chordates: the sea squirt, a urochordate, and amphioxus, a cephalochordate (supplementary table S1, Supplementary Material online). The tree was rooted with vertebrate *Ngb* and *GbX* and their respective orthologs in amphioxus, as these genes originated prior to the split between proto-stomes and deuterostomes (Roesner et al. 2005; Fuchs et al. 2006; Ebner et al. 2010). The phylogenetic reconstruction revealed that the vertebrate-specific globins fall into four main clades: 1) *Cygb*, 2) *Mb* + *GbE*, 3) the α - and β -chain *Hbs* of gnathostomes, and 4) *GbY* (fig. 1). The sister relationship between gnathostome *Cygb* and the cyclostome *Hbs* indicates that progenitors of the four main vertebrate-specific globin lineages were present in the vertebrate common ancestor (Hoffmann, Opazo, et al. 2010, 2011; Storz et al. 2011). The resultant phylogeny also reveals a single clade of amphioxus globins that are sister to the four clades of vertebrate-specific globins. Parametric bootstrapping tests strongly supported the monophyly of vertebrate-specific globins over alternative phylogenetic hypotheses in which amphioxus globins *Gb2*, *Gb8*, and *Gb14* were nested within the clade of vertebrate-specific globins ($P \leq 0.001$ in all cases). Although evidence suggests that urochordates are actually sister to vertebrates (Delsuc et al. 2006; Putnam et al. 2008), the phylogeny shown in figure 1 indicates that the pro-ortholog of the vertebrate-specific globins was secondarily lost from the sea squirt genome. In summary, the phylogeny of vertebrate-specific globins appears to conform to the 4:1 rule, as there are four globin lineages in vertebrates that correspond to a single lineage of amphioxus globins.

Analyses of Chromosomal Homology

If the four main clades of vertebrate-specific globins represent products of two successive WGDs, then

representatives of the four globin gene lineages should be embedded in unlinked chromosomal regions that share similar interdigitated arrangements of paralogous genes ("paralogons"; Coulier et al. 2000; Dehal and Boore 2005; Braasch et al. 2006). The flanking tracts of paralogous duplicates may not contain identical subsets of genes, but the globin-defined paralogs should be united by 4:1 gene families and—in various combinations—by 3:1 and 2:1 gene families that trace their duplicative origins to the stem lineage of vertebrates. Moreover, the globin-defined paralogs should all derive from a single linkage group of the ancestral chordate proto-karyotype (Putnam et al. 2008).

To test these predictions, we examined the genomic map positions of the vertebrate-specific globin genes and we then characterized large-scale patterns in the physical locations of paralogous gene duplicates in the flanking chromosomal regions. Comparative analysis of complete genome sequences from 13 vertebrate taxa revealed that the *Cygb* gene, the *Mb/GbE* gene pair, and the α -*Hb/GbY* genes are each embedded in clearly identifiable paralogs, indicating that they are products of large-scale duplications and/or WGDs (fig. 2 and table 1).

Analyses of Co-duplicated Genes

Phylogenetic analyses revealed that the three globin-defined paralogs are united by 3:1 and (in various pairwise combinations) 2:1 gene families that derive from duplications in the stem lineage of vertebrates (table 2 and supplementary fig. S1, Supplementary Material online). In the chicken genome, the "*Cygb*", "*Mb*", and "*Hb*" paralogs correspond to large segments of chromosomes 18, 1, and 14, respectively (fig. 2A and table 1). In the human genome, the *Cygb* paralogon corresponds to large segments of chromosome 17, the *Hb* paralogon is distributed across portions of chromosomes 7 and 16 (the location of the α -globin gene family), and the *Mb* paralogon is distributed across portions of chromosomes 7, 12, and 22 (fig. 2B, table 1, and supplementary fig. S2, Supplementary Material online). The *Hb* paralogon is defined by the α -globin gene cluster of amniotes and is defined by the tandemly linked α - and β -globin gene clusters in teleost fishes and amphibians. The *GbY* gene, which has only been found in the genomes of amphibians, squamate reptiles, and platypus (Fuchs et al. 2006; Patel et al. 2008; Hoffmann, Storz, et al. 2010), is located at the 3' end of the α -globin gene cluster, and it is therefore associated with the *Hb* paralogon. This linkage arrangement indicates that *GbY* and the proto-*Hb* gene are products of an ancient tandem gene duplication that may predate the first round of WGD.

The genomic analysis of globin-defined paralogs revealed a 3-fold pattern of conserved macrosynteny, which can be reconciled with the 4-fold pattern expected under the 2R model by invoking the secondary loss of one of four paralogous globin genes that would have been produced by two successive rounds of WGD. In principle, the "tetraparalogon" structure predicted by the 2R model could be verified by identifying a fourth set of paralogous genes on a different chromosome that co-duplicated with the

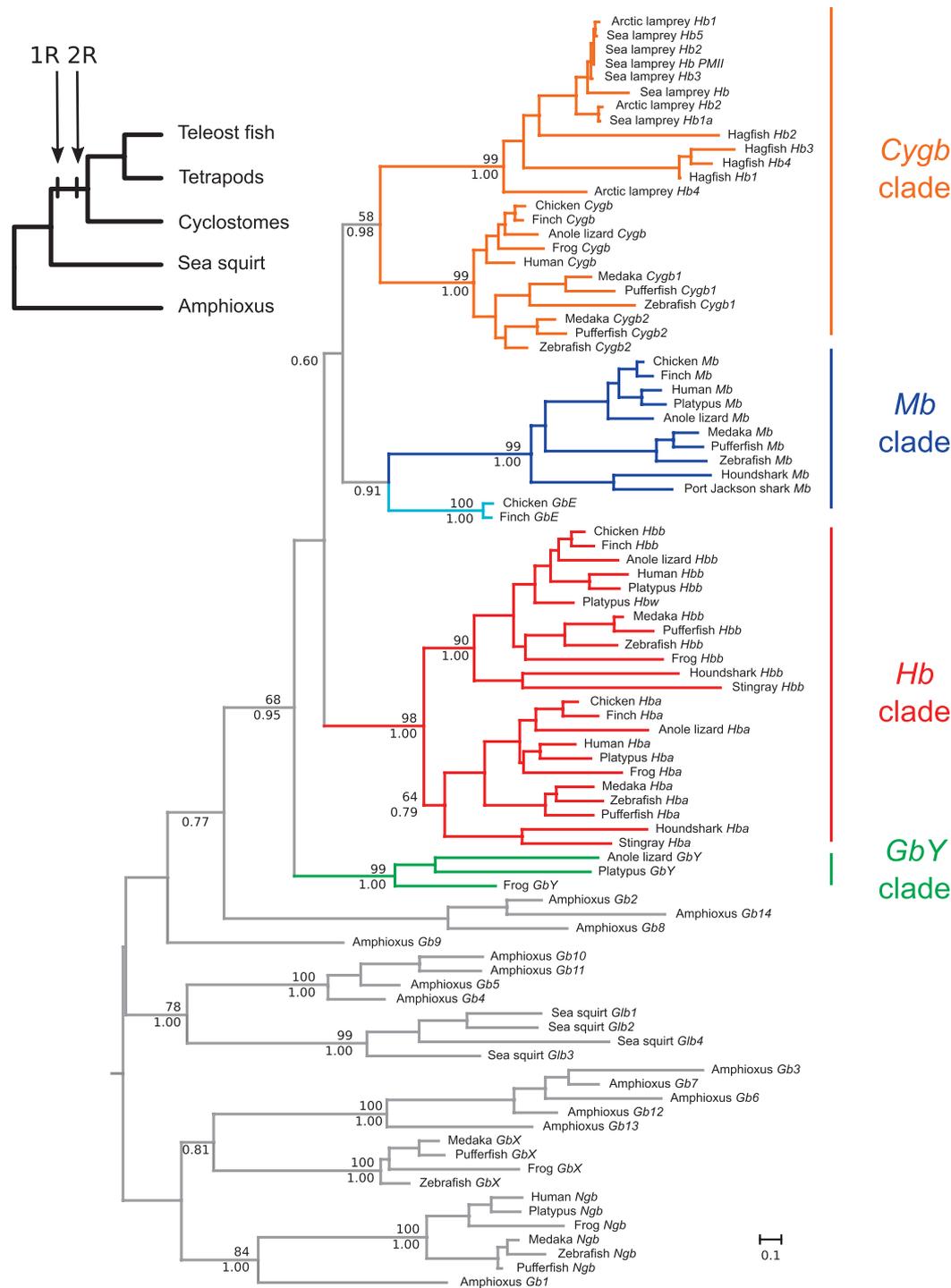


FIG. 1. Maximum likelihood phylogram describing relationships among globin genes from representative vertebrate taxa and two nonvertebrate chordates: the sea squirt, *Ciona intestinalis* (a urochordate) and amphioxus, *Branchiostoma floridae* (a cephalochordate). Numbers above the nodes correspond to maximum likelihood bootstrap support values and those below the nodes correspond to Bayesian posterior probabilities. The tree was rooted with the clade that contains the vertebrate *Ngb* and *GbX* genes and their amphioxus counterparts. The inset on top shows the organismal phylogeny with the occurrence of two successive WGDs denoted as 1R and 2R. Accession numbers for all sequences are provided in [supplementary table S1, Supplementary Material](#) online.

Cygb, *Mb*, and *Hb* paralogs. We combined phylogenetic reconstructions with an analysis of conserved synteny to identify the genomic location of this missing fourth paralog, which we henceforth refer to as the “globin minus” (*Gb⁻*) paralog (because the associated globin gene

would have been secondarily lost). In the human genome, we identified a total of seven 4:1 gene families in which one paralog maps to the *Cygb* paralog, one maps to the *Hb* paralog, one maps to the *Mb* paralog, and the fourth paralog maps to chromosome 19 ([supplementary fig. S3](#),

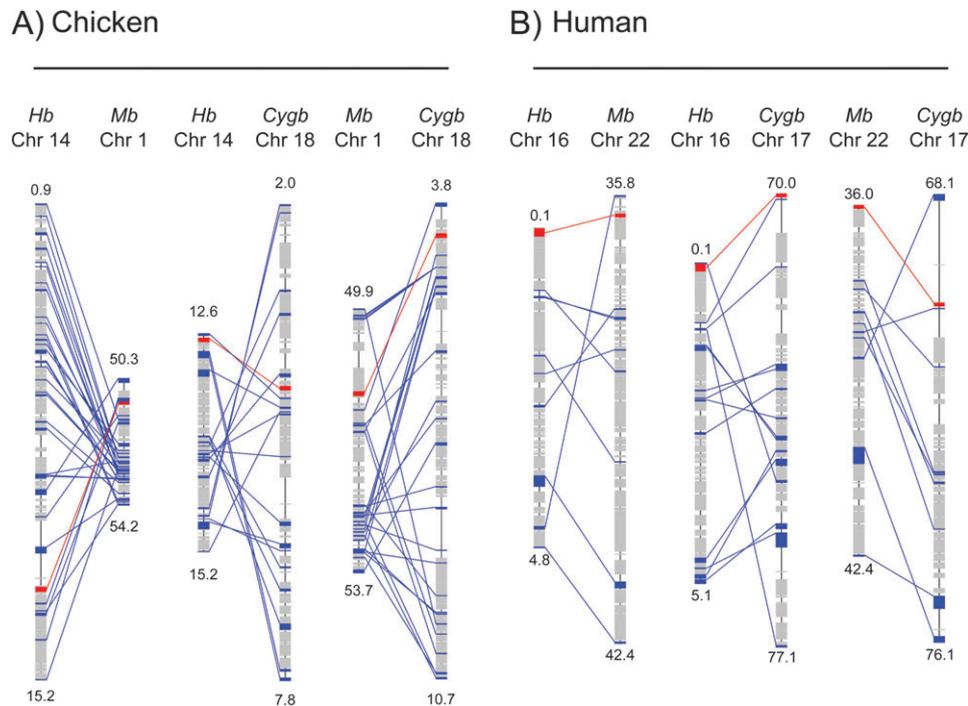


FIG. 2. Graphical depiction of shared gene duplicates among three globin-defined paralogs (*Cygb*, *Mb*, and *Hb*) in the chicken (A) and human (B) genomes. Boxes denote annotated genes on each chromosome, red lines connect paralogous globin genes between chromosomes, and blue lines connect co-duplicated paralogs in the flanking regions. Numbers indicate chromosomal positions in megabases. Statistical support for these patterns of chromosomal homology is presented in [table 1](#).

Supplementary Material online). Moreover, we used a bioinformatic approach to identify a clearly demarcated segment of human chromosome 19 that shares a total of 15 paralogous gene duplicates with the globin-defined paralogs (including seven 3:1 gene families and three 2:1 gene families; [fig. 3](#)). These results implicate human chromosome 19 as the genomic location of the fourth “*Gb*[−]” paralogon—a relictual product of WGD that previously harbored a globin gene that was co-paralogous to *Cygb*, *Mb*, and *Hb*. Thus, the identification of the *Gb*[−] paralogon on human chromosome 19 reveals the telltale pattern of 4-fold conserved macrosynteny that is predicted by the 2R model.

Four-fold Conserved Synteny with the Ancestral Chordate Proto-karyotype

As a final line of evidence documenting the role of WGD in the functional diversification of vertebrate-specific globins,

comparative analysis of the amphioxus genome revealed that the *Gb*[−] paralogon and each of the three globin-defined paralogs descend from linkage group 15 of the reconstructed proto-karyotype of the chordate common ancestor ([fig. 4](#)). In combination with the phylogenetic results and the observed linkage arrangement of paralogous genes, the fact that the globin-defined paralogs trace their duplicative origins to the same ancestral chordate linkage group clearly demonstrates that three of the four main lineages of vertebrate-specific globin genes (*Mb* + *GbE*, *Cygb*, and *Hb*) originated via WGD.

Discussion

Results of our combined phylogenetic and genomic analyses reveal the relative roles of small-scale duplication events and WGD events in fueling the functional diversification of the vertebrate globin gene superfamily ([fig. 5](#)).

Table 1. Pairwise Comparisons between Homologous Chromosomal Segments Corresponding to the *Cygb*, *Mb*, and *Hb* Paralogs Identified by CHSminer (Wang et al. 2009).

Species	Pairwise Chromosomal Homology	Homologous Segment 1	Homologous Segment 2	<i>s</i>	<i>P</i>
Chicken	<i>Hb</i> – <i>Cygb</i>	Chr. 14: 12.6–15.2 Mb	Chr. 18: 2.7–7.8 Mb	18	8.817×10^{-42}
	<i>Hb</i> – <i>Mb</i>	Chr. 14: 0.8–15.2 Mb	Chr. 1: 50.3–54.2 Mb	33	3.769×10^{-58}
	<i>Cygb</i> – <i>Mb</i>	Chr. 18: 3.7–10.7 Mb	Chr. 1: 49.9–53.7 Mb	24	2.820×10^{-44}
Human	<i>Hb</i> – <i>Cygb</i>	Chr. 16: 0.1–7.7 Mb	Chr. 17: 0.1–80.6 Mb	19	2.071×10^{-50}
	<i>Hb</i> – <i>Mb</i>	Chr. 16: 1.0–8.9 Mb	Chr. 22: 5.8–42.4 Mb	11	7.988×10^{-32}
	<i>Cygb</i> – <i>Mb</i>	Chr. 17: 2.5–80.2 Mb	Chr. 22: 6.1–42.4 Mb	18	2.755×10^{-43}

NOTES.— In the chicken genome, the *Cygb*, *Mb*, and *Hb* paralogs are located on chromosome 18 (4.3 Mb), chromosome 14 (12.7 Mb), and chromosome 1 (51 Mb), respectively. In the human genome, the identified chromosomal segments that define the *Cygb*, *Mb*, and *Hb* paralogs are located on chromosome 17 (74.5 Mb), chromosome 16 (0.2 Mb), and chromosome 22 (36 Mb), respectively. *s* = number of paralogous gene duplicates that are shared between a given pair of homologous segments. Reported *P* values indicate the probability of identifying false-positive (nonhomologous) chromosomal segments of size *s*. These patterns of chromosomal homology are graphically depicted in [figure 2](#).

Table 2. Multigene Families in the Human and Chicken Genomes That Share the Same Duplicative Origins as the *Cygb*, *Mb*, and *Hb* Paralogons.

Intragenomic Redundancy Relative to Nonvertebrate Chordates	No. of Gene Families	Connections among Globin-Defined Paralogons
3:1	25	<i>Cygb</i> – <i>Hb</i> – <i>Mb</i>
2:1	39	(9 <i>Cygb</i> – <i>Mb</i> , 19 <i>Cygb</i> – <i>Hb</i> , and 11 <i>Mb</i> – <i>Hb</i>)

NOTES.—In each gene family, genomic analysis revealed that each of the constituent paralogs map to a different globin-defined paralogon (*Cygb*, *Mb*, or *Hb*), and phylogenetic analyses revealed that the set of globin-linked paralogs originated via duplication events that occurred prior to the divergence between teleost fish and tetrapods (the earliest phylogenetic split among the vertebrate taxa included in our study). Phylogenetic reconstructions of representative 3:1 and 2:1 gene families are shown in [supplementary fig. S1, Supplementary Material online](#).

The first round of WGD produced a proto “*Mb*/*GbE*/*Cygb*” paralogon and a proto “ α/β -*Hb* + *GbY*” paralogon ([fig. 5](#) and [supplementary fig. S4, Supplementary Material online](#)). Reduplication of the former paralogon in the subsequent round of WGD gave rise to the progenitors of the *Cygb* and *Mb*/*GbE* gene lineages, and reduplication of the latter paralogon gave rise to the progenitor of the α - and β -globin gene families as well as a fourth paralogous globin gene lineage that does not appear to have been retained in any extant vertebrate taxa ([fig. 5](#) and [supplementary fig. S4, Supplementary Material online](#)). Interestingly, the progenitors of the vertebrate *GbX* and *Ngb* genes, which diverged early in animal evolution (Roesner et al. 2005; Fuchs et al. 2006; Ebner et al. 2010), were also present in the vertebrate

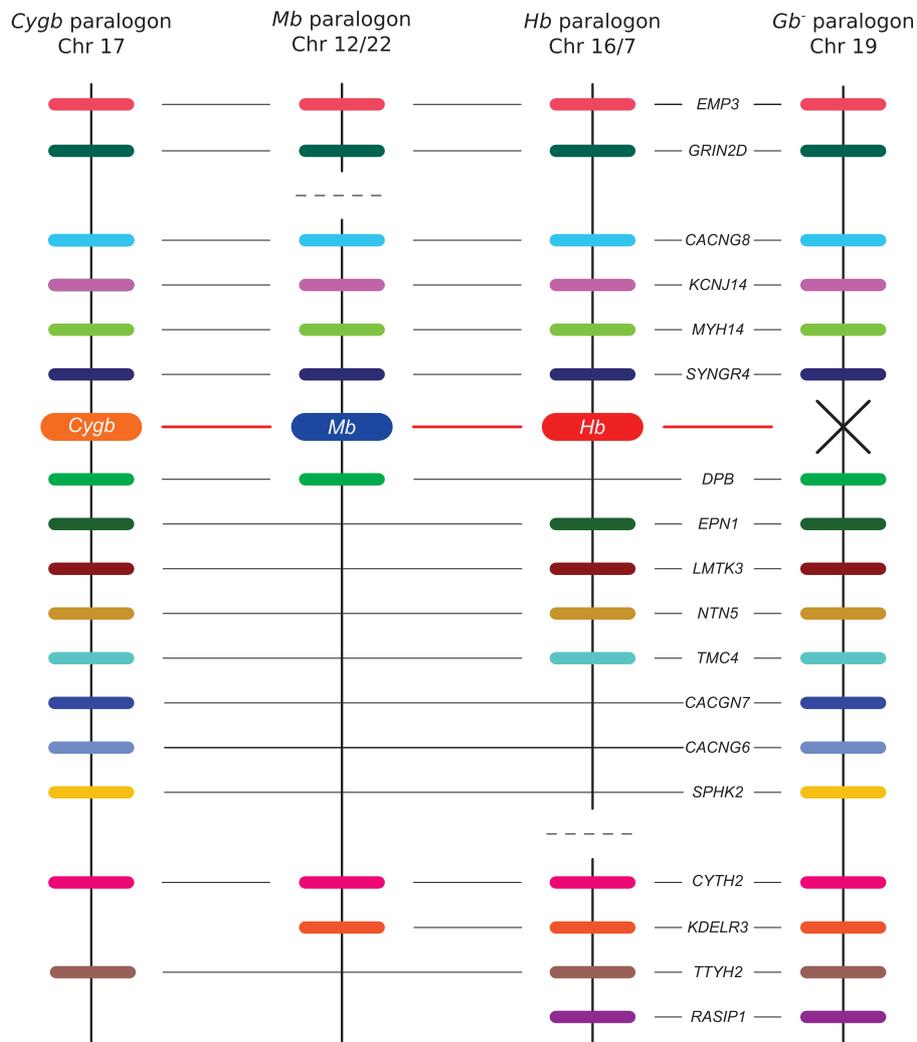


FIG. 3. Graphical depiction of gene duplicates that are shared between the Gb^- paralogon and the remaining three globin-defined paralogons (*Cygb*, *Mb*, and *Hb*) in the human genome. There are seven 4:1 gene families that unite the Gb^- paralogon with the *Cygb*, *Mb*, and *Hb* paralogons (phylogenies for these gene families are shown in [supplementary fig. S3, Supplementary Material online](#)); there are seven 3:1 gene families that unite the Gb^- paralogon with two of the three globin-defined paralogons; and there are four 2:1 gene families that unite the Gb^- paralogon with a single globin-defined paralogon. On each chromosome, annotated genes are depicted as colored bars. The “missing” globin gene on the Gb^- paralogon is denoted by an “X”. The shared paralogs are depicted in colinear arrays for display purposes only, as there is substantial variation in gene order among the four paralogons. For clarity of presentation, genes that are not shared between the Gb^- paralogon and any of the three globin-defined paralogons are not shown. In the human genome, the Gb^- paralogon on chromosome 19 shares multiple gene duplicates with fragments of the *Hb* paralogon on chromosomes 16 and 7, and fragments of the *Mb* paralogon on chromosomes 12 and 22. Members of the *EPN1*, *LMTK3*, and *KCNJ14* gene families that map to the *Hb* paralogon have been secondarily translocated from chromosome 16.

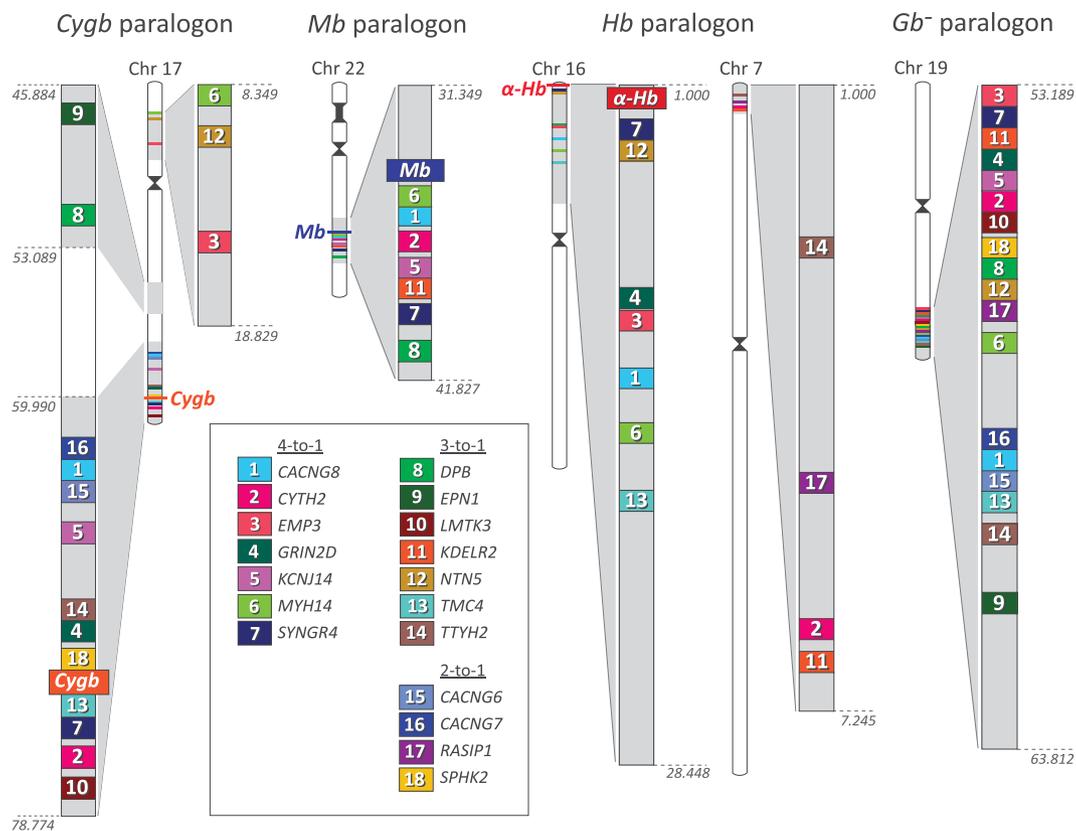


FIG. 4. Four-fold pattern of conserved macrosynteny between the four globin-defined paralogs in the human genome (including the Gb^- paralogon) and “linkage group 15” of the reconstructed proto-karyotype of the chordate common ancestor (Putnam et al. 2008; shaded regions). This pattern of conserved macrosynteny demonstrates that the *Cygb*, *Mb*, *Hb*, and Gb^- paralogs trace their duplicative origins to the same proto-chromosome of the chordate common ancestor and provides conclusive evidence that each of the four paralogs are products of a genome quadruplication in the stem lineage of vertebrates. Shared gene duplicates that map to secondarily translocated segments of the *Mb* paralogon (on chromosome 12) and the *Hb* paralogon (on chromosomes 7 and 17) are not pictured.

common ancestor and were therefore duplicated and re-duplicated through two rounds of WGD. However, unlike the proto *Mb/GbE/Cygb* gene, the *GbX* and *Ngb* genes subsequently reverted to the ancestral single-copy state.

These findings reveal direct links between vertebrate-specific WGDs and key physiological innovations in the vertebrate oxygen transport system. Specifically, our results demonstrate that the physiological division of labor between the oxygen transport function of the proto-Hb protein and the oxygen storage function of the proto-Mb protein likely evolved after the first round of WGD. In the ancestor of modern gnathostomes, a single member of one duplicated gene pair (the proto-*Mb* gene) evolved an oxygen storage function and tissue-specific expression mainly restricted to cardiac and striated muscle, and a single member of the paralogous gene pair (the proto-*Hb* gene) evolved an oxygen transport function and tissue-specific expression mainly restricted to erythroid cells (Goodman et al. 1975, 1987; Dickerson and Geis 1983; Hardison 1998).

The first round of WGD may have initially set the stage for the physiological division of labor between the evolutionary forerunners of *Mb* and *Hb* by permitting divergence in the tissue specificity of gene expression. The subsequent tandem duplication that produced the proto α - and β -globin genes

then promoted a further specialization of function with respect to blood-gas transport. In the common ancestor of gnathostomes, functional divergence of the proto α - and β -globin genes permitted the formation of multimeric Hbs composed of unlike subunits such as the $\alpha_2\beta_2$ heterotetramers of most tetrapods and teleost fish (Goodman et al. 1975, 1987; Dickerson and Geis 1983). The evolution of this heteromeric quaternary structure was central to the emergence of Hb as a specialized oxygen transport protein because it provided a mechanism for cooperative oxygen-binding and allosteric regulation of oxygen affinity. Both these features require a coupling between the effects of ligand binding at individual subunits and the interactions between subunits in the quaternary structure.

The third WGD-derived paralog, *Cygb*, evolved very distinct specializations of function in the gnathostomes and cyclostomes (jawless fishes, represented by lampreys and hagfish). In cyclostomes, which have secondarily lost orthologs of the *Mb* and *Hb* genes, the *Cygb* protein was co-opted to serve an oxygen transport function (Hoffmann, Opazo, et al. 2010). In gnathostomes, by contrast, the hexacoordinate *Cygb* protein is expressed in the cytoplasm of fibroblasts and related cell types that are actively engaged in the production of extracellular matrix components in visceral organs, and it is also expressed in neuronal cells

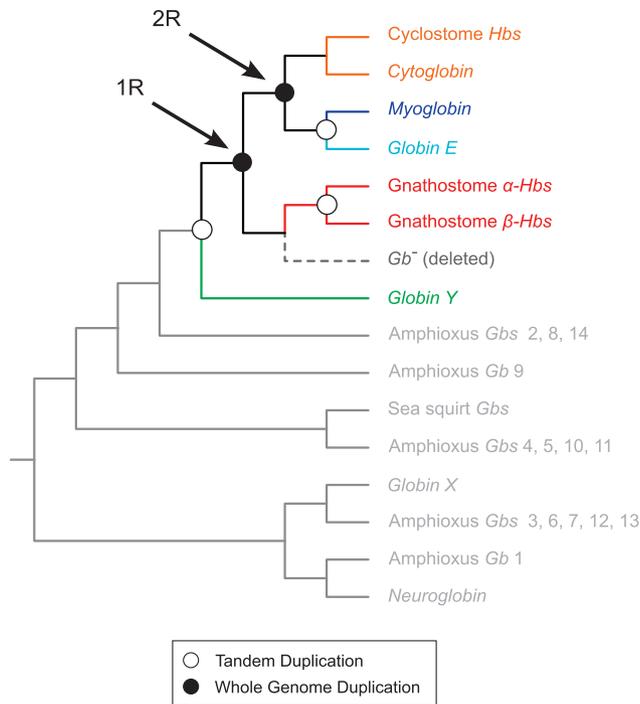


FIG. 5. Cladogram describing phylogenetic relationships among vertebrate-specific globins, with the inferred mode of duplication indicated at each node.

of the retina, brain, and peripheral nervous system (Kawada et al. 2001; Burmester et al. 2002, 2004; Pesce et al. 2002; Trent and Hargrove 2002; Fago et al. 2004; Hankeln et al. 2005; Hankeln and Burmester 2008). Although the specific functions of gnathostome Cygb have yet to be fully elucidated (Kakar et al. 2010), the protein is characterized by higher levels of amino acid sequence conservation relative to both Mb and Hb (Burmester et al. 2002; Blank et al. 2011; Hoffmann et al. 2011), suggesting that it is performing an essential physiological role. In gnathostomes, the three WGD-derived globin proteins (the monomeric Mb, the homodimeric Cygb, and the heterotetrameric Hb) each evolved highly distinct specializations of function involving different heme-coordination chemistries and ligand affinities, and they also evolved equally distinct expression domains that are specific to completely different cell types.

In summary, results of our phylogenetic and genomic analyses reveal that two rounds of WGD fueled the key innovations in the oxygen transport system that were central to the evolution of aerobic energy metabolism in early vertebrates. This discovery supports the hypothesis that WGDs have played a pivotal role in the evolution of phenotypic novelty and also illuminates the surprising evolutionary origins of Hb and Mb, two of the most intensively studied proteins in existence.

Supplementary Material

Supplementary figures 1–4 and supplementary table 1 are available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>).

Acknowledgments

This paper is dedicated to the memory of Morris Goodman (1925–2010), a pioneer in the study of globin gene family evolution. We thank Z. Cheviron, J. Good, M. Goodman, E. Lessa, J. Mower, S. Smith, S. Vinogradov, and two anonymous reviewers for their helpful comments and suggestions. This work was funded by grants to J.F.S. from the National Science Foundation (NSF; IOS-0949931) and the National Institutes of Health/National Heart, Lung, and Blood Institute (R01 HL087216 and HL087216-S1), to F.G.H. from NSF (EPS-0903787), and to J.C.O. from the Fondo Nacional de Desarrollo Científico y Tecnológico (FONDECYT 11080181), the Programa Bicentenario en Ciencia y Tecnología (PSD89), and the Concurso Estadía Jóvenes Investigadores en el Extranjero from the Universidad Austral de Chile.

References

- Abi-Rached L, Gilles A, Shiina T, Pontarotti P, Inoko H. 2002. Evidence of *en bloc* duplication in vertebrate genomes. *Nat Genet.* 31:100–105.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol.* 215:403–410.
- Amores A, Force A, Yan YL, et al. (13 co-authors). 1998. Zebrafish hox clusters and vertebrate genome evolution. *Science* 282:1711–1714.
- Blank M, Kiger L, Thielebein A, Gerlach F, Hankeln T, Marden MC, Burmester T. 2011. Oxygen supply from the bird's eye perspective: globin E is a respiratory protein in the chicken retina. *J Biol Chem.* 286:26507–26515.
- Braasch I, Brunet F, Volff J, Scharl M. 2009. Pigmentation pathway evolution after whole-genome duplication in fish. *Genome Biol Evol.* 1:479–493.
- Braasch I, Salzburger W, Meyer A. 2006. Asymmetric evolution in two fish-specifically duplicated receptor tyrosine kinase paralogons involved in teleost coloration. *Mol Biol Evol.* 23:1192–1202.
- Braasch I, Scharl M, Volff J. 2007. Evolution of pigment synthesis pathways by gene and genome duplication in fish. *BMC Evol Biol.* 7:74.
- Braasch I, Volff J, Scharl M. 2009. The endothelin system: evolution of vertebrate-specific ligand-receptor interactions by three rounds of genome duplication. *Mol Biol Evol.* 26:783–799.
- Burmester T, Ebner B, Weich B, Hankeln T. 2002. Cytoglobin: a novel globin type ubiquitously expressed in vertebrate tissues. *Mol Biol Evol.* 19:416–421.
- Burmester T, Haberkamp M, Mitz S, Roesner A, Schmidt M, Ebner B, Gerlach F, Fuchs C, Hankeln T. 2004. Neuroglobin and cytoglobin: genes, proteins and evolution. *IUBMB Life.* 56: 703–707.
- Coulier F, Popovici C, Villet R, Birnbaum D. 2000. Metahox gene clusters. *J Exp Zool.* 288:345–351.
- Dehal P, Boore JL. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.* 3:e314.
- Delsuc F, Brinkmann H, Chourrout D, Philippe H. 2006. Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature* 439:965–968.
- Dickerson RE, Geis I. 1983. Hemoglobin: structure, function and evolution, and pathology. Menlo Park (CA): Benjamin/Cummings Publishing Co.
- Ebner B, Burmester T, Hankeln T. 2003. Globin genes are present in *Ciona intestinalis*. *Mol Biol Evol.* 20:1521–1525.

- Ebner B, Panopoulou G, Vinogradov SN, Kiger L, Marden MC, Burmester T, Hankeln T. 2010. The globin gene family of the cephalochordate amphioxus: implications for chordate globin evolution. *BMC Evol Biol.* 10:370.
- Fago A, Hundahl C, Malte H, Weber RE. 2004. Functional properties of neuroglobin and cytoglobin. Insights into the ancestral physiological roles of globins. *IUBMB Life.* 56: 689–696.
- Fuchs C, Burmester T, Hankeln T. 2006. The amphibian globin gene repertoire as revealed by the *Xenopus* genome. *Cytogenet Genome Res.* 112:296–306.
- Goldman N, Anderson JP, Rodrigo AG. 2000. Likelihood-based tests of topologies in phylogenetics. *Syst Biol.* 49:652–670.
- Goodman M, Czelusniak J, Koop BF, Tagle DA, Slightom JL. 1987. Globins: A case study in molecular phylogeny. *Cold Spring Harb Symp Quant Biol.* 52:875–890.
- Goodman M, Moore GW, Matsuda G. 1975. Darwinian evolution in the genealogy of haemoglobin. *Nature* 253:603–608.
- Hankeln T, Burmester T. 2008. Neuroglobin and cytoglobin. In: Gosh A, editor. *The smallest biomolecules: diatomics and their interactions with heme proteins.* Amsterdam (The Netherlands): Elsevier. p. 203–218.
- Hankeln T, Ebner B, Fuchs C, et al. (23 co-authors) 2005. Neuroglobin and cytoglobin: in search of their role in the vertebrate globin family. *J Inorg Biochem.* 99:110–119.
- Hardison R. 1998. Hemoglobins from bacteria to man: evolution of different patterns of gene expression. *J Exp Biol.* 201: 1099–1117.
- Hoegg S, Meyer A. 2005. Hox clusters as models for vertebrate genome evolution. *Trends Genet.* 21:421–424.
- Hoffmann FG, Opazo JC, Storz JF. 2010. Gene cooption and convergent evolution of oxygen transport hemoglobins in jawed and jawless vertebrates. *Proc Natl Acad Sci U S A.* 107:14274–14279.
- Hoffmann FG, Opazo JC, Storz JF. 2011. Differential loss and retention of cytoglobin, myoglobin, and globin-E during the radiation of vertebrates. *Genome Biol Evol.* 3:588–600.
- Hoffmann FG, Storz JF, Gorr TA, Opazo JC. 2010. Lineage-specific patterns of functional diversification in the α - and β -globin gene families of tetrapod vertebrates. *Mol Biol Evol.* 27: 1126–1138.
- Holland PW, Garcia-Fernández J, Williams NA, Sidow A. 1994. Gene duplications and the origins of vertebrate development. *Dev Suppl.* 1994:125–133.
- Horton AC, Mahadevan NR, Ruvinsky I, Gibson-Brown JJ. 2003. Phylogenetic analyses alone are insufficient to determine whether genome duplication(s) occurred during early vertebrate evolution. *J Exp Zool B Mol Dev Evol.* 299:41–53.
- Jaillon O, Aury J, Brunet F, et al. (61 co-authors) 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* 431:946–957.
- Jobb G, von Haeseler A, Strimmer K. 2004. Treefinder: a powerful graphical analysis environment for molecular phylogenetics. *BMC Evol Biol.* 4:18.
- Kakar S, Hoffman FG, Storz JF, Fabian M, Hargrove MS. 2010. Structure and reactivity of hexacoordinate hemoglobins. *Biophys Chem.* 152:1–14.
- Kasahara M, Naruse K, Sasaki S, et al. (37 co-authors) 2007. The medaka draft genome and insights into vertebrate genome evolution. *Nature* 447:714–719.
- Kassahn KS, Dang VT, Wilkins SJ, Perkins AC, Ragan MA. 2009. Evolution of gene function and regulatory control after whole-genome duplication: comparative analyses in vertebrates. *Genome Res.* 19:1404–1418.
- Katoh K, Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform.* 9:286–298.
- Kawada N, Kristensen DB, Asahina K, Nakatani K, Minamiyama Y, Seki S, Yoshizato K. 2001. Characterization of a stellate cell activation-associated protein (STAP) with peroxidase activity found in rat hepatic stellate cells. *J Biol Chem.* 276:25318–25323.
- Larhammar D, Nordström K, Larsson TA. 2009. Evolution of vertebrate rod and cone phototransduction genes. *Philos Trans R Soc Lond B Biol Sci.* 364:2867–2880.
- McLysaght A, Hokamp K, Wolfe KH. 2002. Extensive genomic duplication during early chordate evolution. *Nat Genet.* 31:200–204.
- Meyer A. 1998. Hox gene variation and evolution. *Nature* 391:225,227–228.
- Meyer A, 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.
- Meyer A, Van de Peer Y. 2005. From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *Bioessays* 27:937–945.
- Muffato M, Louis A, Poinsel C, Crollius HR. 2010. Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. *Bioinformatics* 26:1119–1121.
- Ohno S. 1970. *Evolution by gene duplication.* New York: Springer-Verlag.
- Patel VS, Cooper SJB, Deakin JE, Fulton B, Graves T, Warren WC, Wilson RK, Graves JAM. 2008. Platypus globin genes and flanking loci suggest a new insertional model for beta-globin evolution in birds and mammals. *BMC Biol.* 6:34.
- Pébusque MJ, Coulier F, Birnbaum D, Pontarotti P. 1998. Ancient large-scale genome duplications: phylogenetic and linkage analyses shed light on chordate genome evolution. *Mol Biol Evol.* 15:1145–1159.
- Pesce A, Bolognesi M, Bocedi A, Ascenzi P, Dewilde S, Moens L, Hankeln T, Burmester T. 2002. Neuroglobin and cytoglobin. Fresh blood for the vertebrate globin family. *EMBO Rep.* 3:1146–1151.
- Putnam NH, Butts T, Ferrier DEK, et al. (36 co-authors) 2008. The amphioxus genome and the evolution of the chordate karyotype. *Nature* 453:1064–1071.
- Roesner A, Fuchs C, Hankeln T, Burmester T. 2005. A globin gene of ancient evolutionary origin in lower vertebrates: evidence for two distinct globin families in animals. *Mol Biol Evol.* 22:12–20.
- Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–1574.
- Sato Y, Hashiguchi Y, Nishida M. 2009. Temporal pattern of loss/persistence of duplicate genes involved in signal transduction and metabolic pathways after teleost-specific genome duplication. *BMC Evol Biol.* 9:127.
- Shimeld SM, Holland PW. 2000. Vertebrate innovations. *Proc Natl Acad Sci U S A.* 97:4449–4452.
- Storz JF, Opazo JC, Hoffmann FG. 2011. Phylogenetic diversification of the globin gene superfamily in chordates. *IUBMB Life.* 63:313–322.
- Swofford DL, Olsen GJ, Waddell PJ, Hillis DM. 1996. Phylogenetic inference. In: Hillis DM, Moritz C, Mable B, editors. *Molecular systematics*, 2nd ed. Sunderland (MA): Sinauer Associates. p. 407–514.
- Tatusova TA, Madden TL. 1999. BLAST 2 sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett.* 174:247–250.
- Taylor JS, Braasch I, Frickey T, Meyer A, Van de Peer Y. 2003. Genome duplication, a trait shared by 22,000 species of ray-finned fish. *Genome Res.* 13:382–390.
- Trent JT III, Hargrove MS. 2002. A ubiquitously expressed human hexacoordinate hemoglobin. *J Biol Chem.* 277:19538–19545.

- Van de Peer Y, Maere S, Meyer A. 2009. The evolutionary significance of ancient genome duplications. *Nat Rev Genet.* 10:725–732.
- Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2009. EnsemblCompara genetrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* 19:327–335.
- Wada H. 2001. Origin and evolution of the neural crest: a hypothetical reconstruction of its evolutionary history. *Dev Growth Differ.* 43:509–520.
- Wada H, Makabe K. 2006. Genome duplications of early vertebrates as a possible chronicle of the evolutionary history of the neural crest. *Int J Biol Sci.* 2:133–141.
- Wang Z, Ding G, Yu Z, Liu L, Li Y. 2009. CHSminer: a GUI tool to identify chromosomal homologous segments. *Algorithms Mol Biol.* 4:2.
- Zhang G, Cohn MJ. 2008. Genome duplication and the origin of the vertebrate skeleton. *Curr Opin Genet Dev.* 18:387–393.