Journal of Structural and Functional Genomics **00:** 1–8, 2002.

Functional evolution in the ancestral lineage of vertebrates or when genomic complexity was wagging its morphological tail

Rami Aburomia[1], Oded Khaner[1,2] & Arend Sidow[1]*
[1]*Department of Pathology, Stanford University Medical Center, Room 248B, 300 Pasteur Drive, Stanford, CA 94305–5324, USA;*[2] *Present address: Department of Bio-Medical Sciences, Hadassah College, Jerusalem, Israel.*
* *Author for correspondence: E-mail: arend@stanford.edu*

Received 23.02.2002; accepted in final form 29.08.2002

Key words: cis-regulatory changes, early vertebrate evolution, Myb gene family, Pax 2/5/8

Abstract

Early vertebrate evolution is characterized by a significant increase of organismal complexity over a relatively short time span. We present quantitative evidence for a high rate of increase in morphological complexity during early vertebrate evolution. Possible molecular evolutionary mechanisms that underlie this increase in complexity fall into a small number of categories, one of which is gene duplication and subsequent structural or regulatory neofunctionalization. We discuss analyses of two gene families whose regulatory and structural evolution shed light on the connection between gene duplication and increases in organismal complexity.

Introduction

We are here concerned with the correlation of organismal complexity and new molecular functionality in the ancestral lineage of vertebrates. In principle, three kinds of molecular changes can give neomorphic phenotypes that could contribute to increased organismal complexity. They are, (1) cis-regulatory changes that affect timing and/or place of gene expression, (2) structural changes due to missense mutations or more severe lesions, and (3) entirely new proteins (or functional RNAs). Our analyses focus on the first two classes of changes, cis-regulatory and structural, for which duplication of an intact gene provides the facilitating raw material on which evolutionary pressures can act.

It has been clear for some time now that the ancestral lineage of vertebrates contains an excess of gene duplications in comparison to most other chordate lineages, excepting those that underwent recent genome duplications such as fish and amphibians. By contrast, the amount of morphological evolution that occurred in different lineages of chordates has not been quantified. It was therefore not known whether the increase in morphological complexity in the vertebrate ancestor was unusual or whether other vertebrate lineages underwent equally dramatic morphological changes. To add to the debate of gene duplication and its relationship with increases in organismal complexity at the origin of vertebrates, we present analyses that are intended to shed light on the interface of the two processes.

Results

Morphological Complexity in Early Vertebrate Evolution

We devised a method to estimate the amount of change in morphological complexity during all of vertebrate evolution (O. Khaner and A. Sidow, unpublished data). We first scored 21 extant higher-order chordate groups for the presence or absence of 479 morphological characters whose states for each group were obtained from the literature (Holland, 1996; Baker and Bronner-Fraser, 1997a, 1997b; Gilbert and Raunio, 1997; Kardong, 1997; Pough *et al*., 1999;

Table 1. Number of characters scored in each subgroup of organismal traits

System	No. of Characters
Early embryonic development	47
Notochord, vertebrae, skull and jaws	12
Musculature	25
Cardiovascular and respiratory systems	45
Urogenital system	35
Integument	34
Nervous system	62
Neural crest	16
Sensory organs	70
Endocrine system	26
Digestive system	48
Appendages	29

Shimeld and Holland, 2000). Table 1 shows the breakdown of the morphological characters into different organ systems. State transitions were inferred from the resulting matrix and mapped onto the currently accepted phylogenetic tree with MacClade (Maddison and Maddison, 2001). We then defined the Morphological Complexity Index, $MCI_b = (G_b - L_b)/T$, where subscript b denotes the branch in question, G_b and L_b the number of gained and lost characters on that branch, and T the total number of characters (479 in this case). We calculated MCI_b for each branch of the tree and, upon connecting the nodes of the last common ancestors of the major chordate classes, found that there were two phases with dramatically different rates of increase of the MCI (Fig. 1). The first phase (I in Fig. 1), during which jawless and jawed vertebrates evolved, had an extremely high rate of increase. The second phase (II in Fig. 1), during which the major vertebrate classes evolved, had a ten-fold slower rate of increase. Thus, the comparatively short period of time — between 50 and 100 million years — in which many gene duplications occurred and many new genes arose, coincides with the greatest rate of increase in morphological complexity.

For phase I, we see no comparable increase in the average rate of point substitutional evolution as estimated from standard treeanalyses. In fact, some of the difficulties encountered in reconstruction of phylogenetic trees of gene families that duplicated during this critical period is due to a paucity of missense changes. We conclude that the molecular events that did happen had disproportionately large phenotypic and functional consequences. We believe that the challenge in this area lies in the identification of functionally important changes in gene families, and the mapping of those changes onto robust gene trees.

Gene Duplications in Early Vertebrate Evolution

Inferring functional evolution in a gene family requires robust gene trees; wrong trees cause misinterpretation of character evolution. Unfortunately for the inference of functional evolution of vertebrate gene families, the dissection of the exact branching patterns is fraught with difficulties that have been amply discussed in the '2R' debate (Sidow, 1996; Spring, 1997; Hughes, 1999; Meyer and Schartl, 1999; Martin, 2001). The large number of gene duplications during the short time of early vertebrate evolution suggest tetraploidization as a reasonable explanation for the majority of duplications observed. This led to the prediction that treeanalyses of families with four vertebrate paralogs should generate the branching pattern (A,B),(C,D), but many gene families do not follow this pattern (Gibson and Spring, 2000). Long branch attraction (LBA) has since been cited as one of the artefacts that may cause a failure to recover the symmetric pattern but it cannot explain all cases. Treeanalyses are therefore regarded as inconclusive at best with respect to mechanism underlying the gene duplications.

Ploidy increases are common in evolution, as evidenced by the recent tetra- or polyploidizations in the grasses (Ahn and Tanksley, 1993; Gale and Devos, 1997), the genus *Xenopus* (Kobel and Du Pasquier, 1986), and Salmonid fishes (Allendorf and Utter, 1976). Several ancient tetraploidizations are also known, including those of yeast (Wolfe and Shields, 1997), flowering plants (Doyle *et al*., 1990), and teleost fish (Amores *et al*., 1998). It is by far the most common mechanism for the duplication of many genes during a short amount of time.

Chromosomes of a genome undergo diploidization independently after an autotetraploidization (Fig. 2A). This process can take tens of millions of years, as evidenced by the Salmonid genome duplications. This is shown in Figure 2 for three hypothetical chromosomes which remained allelic for different amounts of time after the tetraploidization. As a consequence, the beginning of independent evolution of paralogs from alleles of a tetraploid will vary greatly across the genome. This is of importance for phylogenetic reconstruction because it is the beginning of

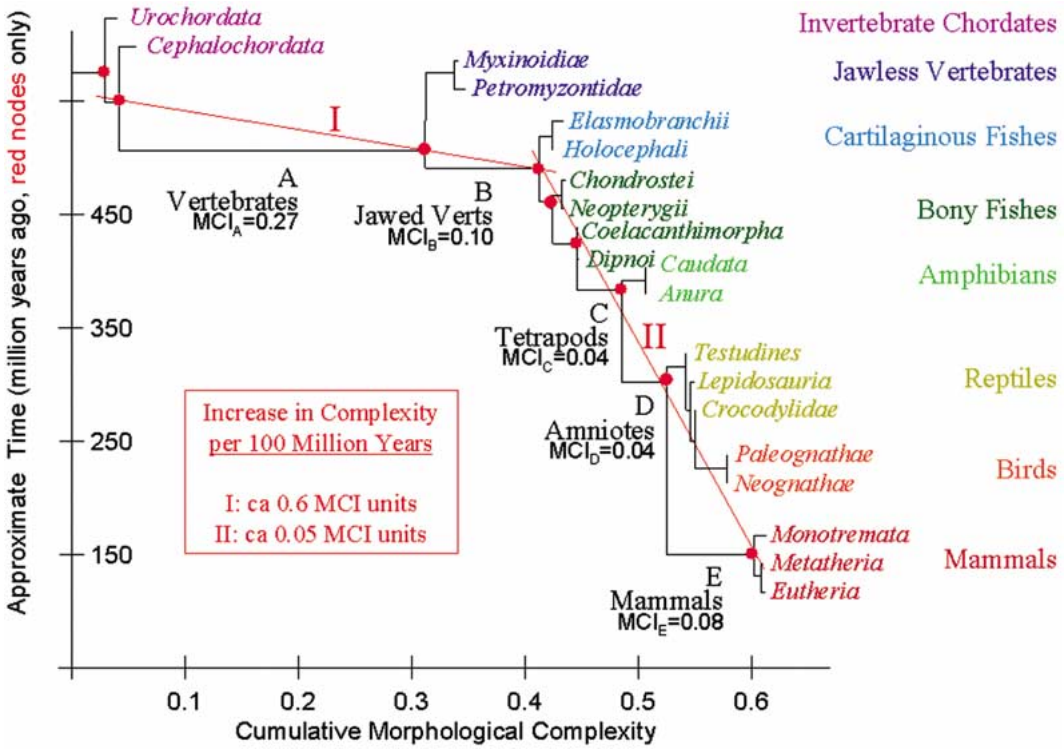

Figure 1. Phylogenetic tree of the groups analyzed in the study of morphological complexity. Branch lengths correspond to the values of the MCI for each branch and are drawn exactly to scale. The Y axis is approximate time with respect to the last common ancestors of the chordate subphyla and vertebrate classes, which are denoted by the red dots. Ancestral lineages with major transitions in morphology are labeled with their names and the values of the MCI. Red lines qualitatively illustrate the rate of increase of morphological complexity during two phases of vertebrate evolution. The rate of increase during Phase I was ten-fold higher than that of Phase II.

independent evolution of duplicates, not the autotetraploidization, that marks a gene duplication in a gene tree.

When one tetraploidization is rapidly followed by a second one (or by independent gene duplications), the time between diploidizations of different genes may vary considerably (for simplicity, Fig. 2 shows concurrent diploidizations after the second tetraploidization but this is unlikely to occur). In our example, chromosome I undergoes diploidization rapidly whereas chromosome III undergoes diploidization just before the next duplication; chromosome II is shown as intermediate. As a result, there is much phylogenetic signal that separates the duplications in the descendants of gene 1, but there is little or none for gene 3.

This situation is exacerbated when all four paralogs are retained (Fig. 2B). The times of diploidizations of the second round of paralogs, for example, 3a/3b and 3c/3d, are independent of each other. The phylogenetic signal for recovering the correct tree is small or even nonexistent if just one of the two diploidizations (for example, 3y to 3a and 3b) follows the first one (3 to 3y and 3z) rapidly. As a consequence, for many gene families in the genome, insufficient time may have elapsed between successive diploidizations to generate a phylogenetic signal that is detectable 500 million years later. A minimum of differences in evolutionary rates to avoid LBA, and sufficient time between successive diploidizations, seem to be necessary to recover the correct phylogenetic relationship. In the next section, we detail the evolution of two gene families and show how recovery of the correct tree is of paramount importance in the analysis of functional evolution of paralogs.

Functional Evolution of the Pax2/5/8 gene family

Phylogenetic trees of the Pax2/5/8 gene family have been published in several reports (Heller and Brändli, 1999; Wada *et al*., 1998; Kozmik *et al*., 1999). All of them show a rooting of the vertebrate subtree on the

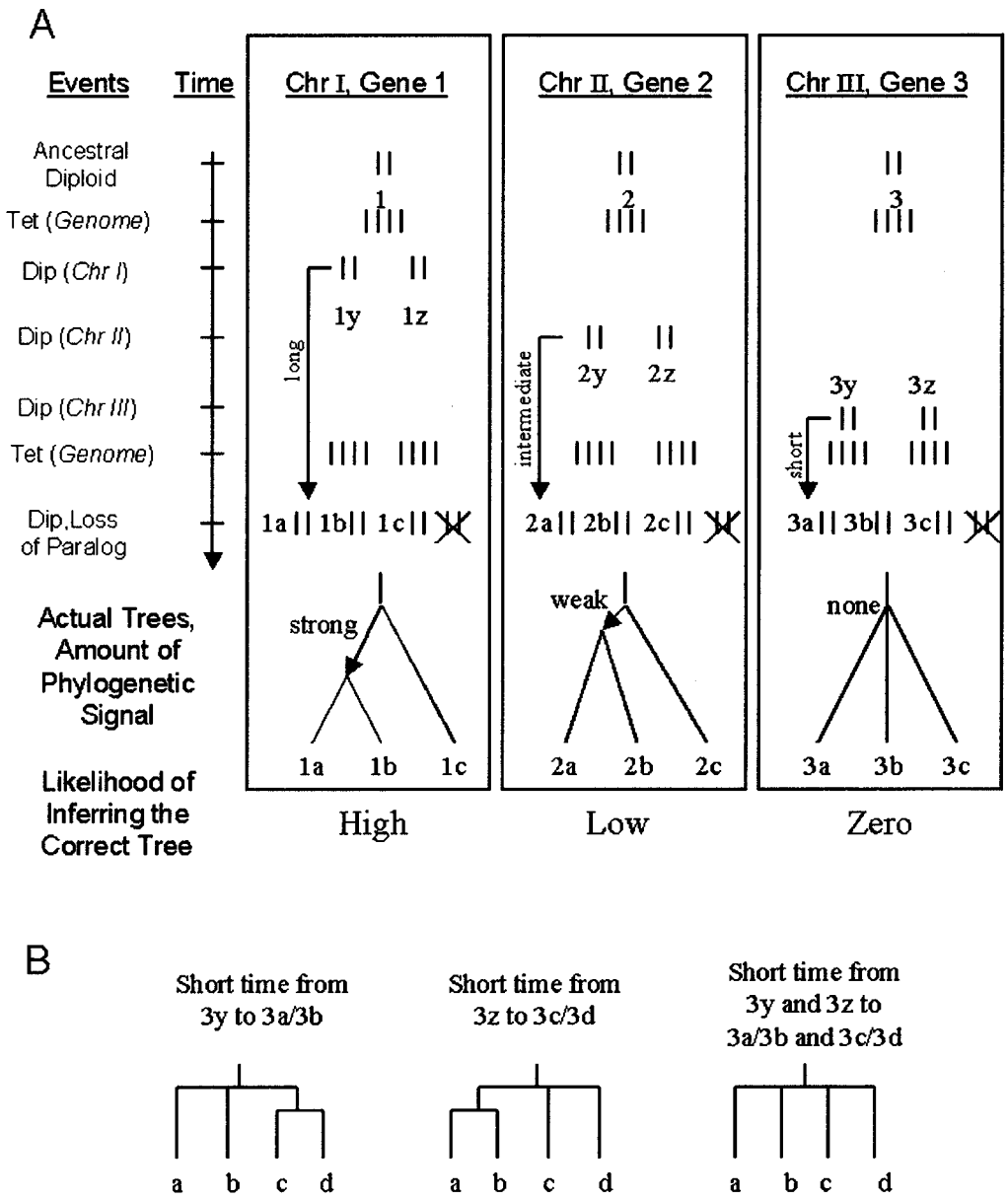

Figure 2. The relationship between timing of tetraploidization and chromosomal diploidization, and their effect on tree reconstruction. (A). Events are shown on the left and time progresses downward. For three chromosomes of a hypothetical genome, the tetraploidizations (Tet) happen at the same time, but each chromosome varies in the amount of time during which all four homologs remain allelic. Dip (diploidization) indicates the time at which the two pairs of homologs begin to evolve independently. Long, intermediate, and short is the amount of time between a chromosome's first and second diploidization. (B). Four-paralog cases analogous to Chr III, Gene 3. When all paralogs are retained, the phylogenetic signal may tend toward zero when the time between diploidizations is short for either pair of paralogs.

ancestral lineage of Pax8, with the Pax258 ancestor duplicating once to yield Pax8 and Pax25, and Pax25 duplicating again to give extant Pax2 and Pax5. We here suggest that this arrangement is erroneous, and that it is most likely due to LBA between invertebrate Pax258 and vertebrate Pax8: Pax8 evolves more than two-fold faster than Pax5 and nearly three-fold faster than Pax2 (Fig. 3A).

It is interesting to note that the rates of evolution of Pax2, Pax5, and Pax8 inversely correlate with the duration of the genes' expression and the importance of their functions as inferred from mouse knockout

AUTHOR!
Please notice that this figure will be printed in colour!

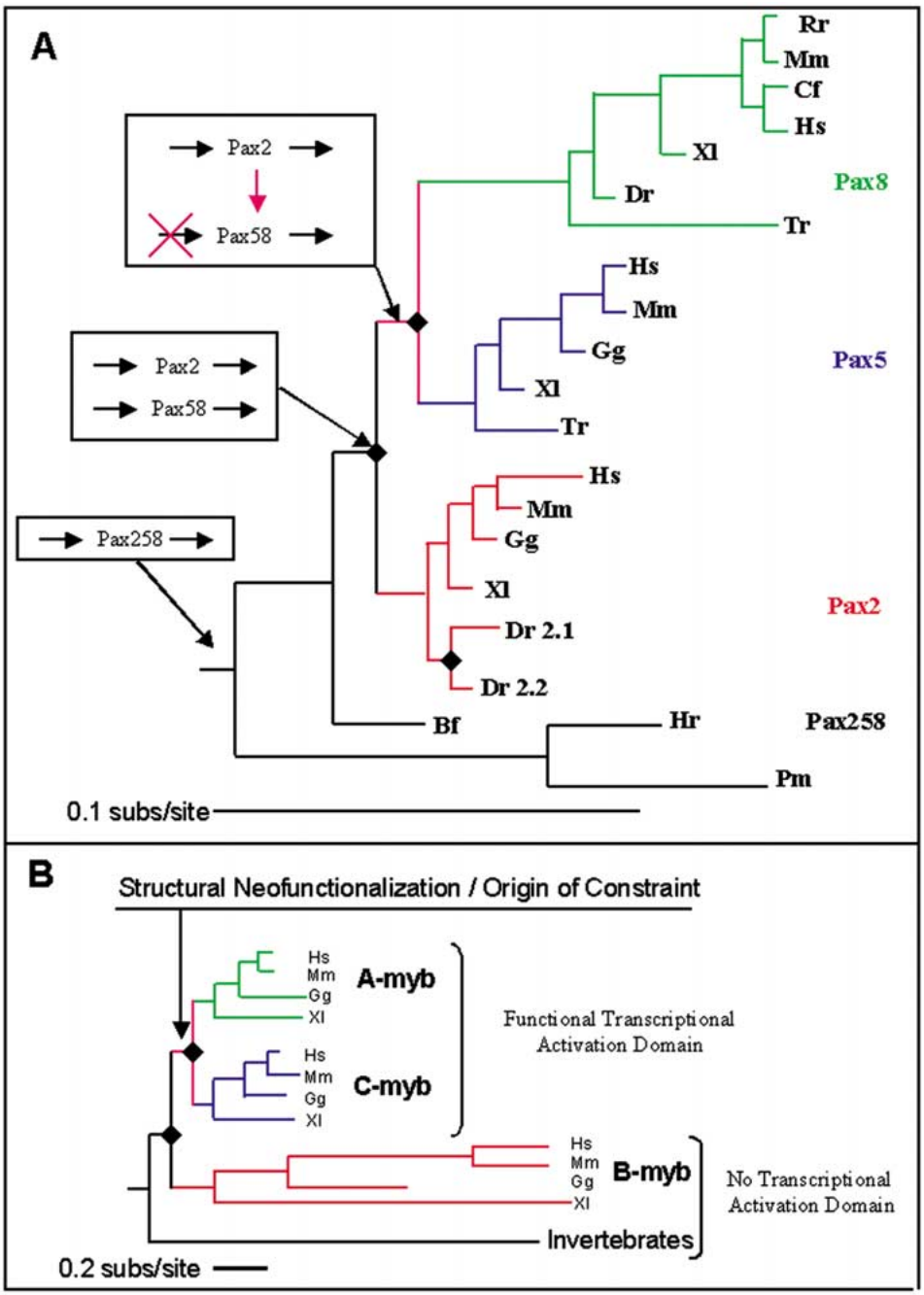

Figure 3. Regulatory and structural evolution in the Pax2/5/8 and Myb gene families. Diamonds are gene duplications. (A). Regulatory evolution of the Pax2/5/8 gene family. Arrows in boxes denote interactions in a simplified regulatory hierarchy. Arrows to the left of the Pax genes signify the initial signal that 'turns on' Pax expression. Each state is mapped onto the appropriate branch of the tree. (B). The Myb case study for illustration of protein neofunctionalization after gene duplication. Branch lengths of each Myb paralog are calculated from the positions corresponding to the transcriptional activation region only, and do not represent evolution of the entire protein. The transcriptional activation domain is constrained equally in A and C, but much less so in B. The arrow denotes the position of the origin of the constraint, and of the transcriptional activation function, which is shared between A and C, but not present in B and invertebrates. Maximum likelihood using PROTML was used to build the trees (Adachi and Hasegawa, 1992). Abbreviations: Bf, *Branchiostoma floridae*; Cf, *Canis familiaris*; Dr, *Danio rerio*; Gg, *Gallus gallus*; Hr, *Halocynthia roretzi*; Hs, *Homo sapiens*; Mm, *Mus musculus*; Rr, *Rattus rattus*; Pm, *Phallusia mammilata*; Tr, *Takifugu rubripes*; Xl, *Xenopus laevis*.

studies. Given that the physicochemical constraints on these closely related proteins are essentially equivalent (indeed, the protein coding regions can substitute for one another in knock-in experiments), the differences in their rates of evolution are a direct reflection of their differences in expression domains and timings.

We built the Pax2/5/8 tree in two steps in order to maximize the amount of usable sequence data. First, by excluding invertebrate chordates we were able to use 238 unambiguously homologous positions to build the best vertebrate-only tree. We then tested alternative rootings of the vertebrate tree, whose topology we fixed, with the invertebrate sequences. Only 136 positions, which are mostly in the highly conserved Paired domain, could be used for this analysis. The tree shown in Fig. 3A, in which the rooting is on the Pax2 lineage, has a log-likelihood value that is 3.8 points greater than the previously published arrangement in which the root falls on the Pax8 lineage.

In order to understand whether the duplications in this gene family facilitated the evolution of more complex regulatory hierarchies, we then asked whether any regulatory changes could be mapped to the periods after the gene duplications. We surveyed the extensive literature on the function and expression domains of each paralog and used the parsimony principle to map the regulatory changes onto the tree.

Pax258 of invertebrate chordates is expressed in a thin strip of cells in a region of the neural tube which may be homologous to the vertebrate mid-hindbrain (Wada *et al*., 1998; Kozmik *et al*., 1999). In vertebrates, the paralogs are expressed at the mid-hindbrain boundary (MHB) in spatial and temporal domains which are partially overlapping (Aasano and Gruss, 1992; Millet *et al*., 1996; Murphy and Hill, 1991). In almost all vertebrate organs where these genes are expressed, Pax2 expression is initiated prior to that of Pax5 and Pax8 (Pfeffer *et al*., 1998, 2002; Heller and Brändli, 1999; Bouchard *et al*., 2000). It was also shown that Pax2 protein is required for maintaining Pax5 and Pax8 transcription, and that the mouse Pax2 knockout has a much more severe brain phenotype than those of Pax8 and Pax5.

Right after the duplication of Pax258 into Pax2 and Pax58, their regulation must have been equivalent. Then, on the ancestral lineage of Pax58, two events occurred that converted this equivalence into a hierarchical relationship: subfunctionalization, when Pax58 lost the ability to respond to the ancestral signal that initiated Pax258 gene expression; and neofunctionalization, when Pax58 gained the susceptibility to be controlled by Pax2. This combination of subfunctionalization and neofunctionalization is a likely hallmark of the increase in regulatory complexity that occurred in the vertebrate ancestor, and provides one molecular link to the increase in organismal complexity.

Functional Evolution of the Myb gene family

Our analysis of the Pax2/5/8 gene family suggests how regulatory evolution may confer developmental, and therefore morphological, complexity. In this section, we use the Myb gene family to illustrate the evolution of a new biochemical function after an early vertebrate gene duplication.

Metazoan Myb genes have three highly conserved Myb repeats that function as DNA binding domains. Invertebrate Mybs and B-Myb function in the regulation of the cell cycle. A- and C-Myb are also known to have an additional independent transcriptional activation function that is missing in invertebrate Mybs and B-Myb. In A- and C-Myb, the second most-constrained region (after the Myb repeats) is the acidic domain, which has been shown to carry this transcriptional activation function.

We built the best tree of the Myb gene family, which agreed with previously published reports (Ganter and Lipsick, 1999). The vertebrate Myb tree is rooted on the B lineage, with A and C sharing a more recent common ancestor. LBA is not a problem because the average rate of evolution in those regions in which A-Myb, B-Myb, and C-Myb can be reliably aligned is very similar among the paralogs (Simon *et al*., 2002). However, an analysis of the regions that correspond to the transcriptional activation domain revealed vastly different rates of evolution. B-myb evolves at a 6.6-fold faster rate than the average of the B-Myb protein (Fig. 3B; Simon *et al*., 2002). In contrast, the corresponding region in both A-myb and C-myb evolves only 1.4 times faster than their average. Using the phylogeny of the Myb gene family, we can map the origin of this constraint onto the ancestral lineage of A- and C-myb, after the B-myb lineage diverged. This is also the most parsimonious placement of the origin of the transcriptional activation function.

Discussion

These case studies underline the potential pitfalls (wrong trees and their causes) in the interpretation of organismal and genomic evolution. Had we not taken special care to infer the best trees, our tree-based interpretation of the evolutionary events in the two gene families during this critical time period would have been misled. In our experience, building the best tree from the ingroup sequences first, and then rooting this tree with the outgroup sequences, consistently maximizes the sequence information used for building the tree.

The main purpose of this study was to explore whether connections could be made between the increase in genomic, developmental, and organismal complexity at the origin of vertebrates. Both regulatory evolution and protein structural evolution are likely to have contributed to an increase in complexity at the origin of vertebrates. We inferred potentially important changes in the regulation of the Pax gene family and in the function of the Myb proteins that occurred *between* consecutive duplication events.

The Pax gene family in particular illustrates the relationship between organismal and molecular complexity well, as it is involved in the patterning of the brain, one of the most important organs that underwent dramatic increases in morphological complexity. One of the characeristics that sets apart the brain of primitive chordates from that of vertebrates is the cerebellum, which forms at the MHB. Pax2 is necessary for its formation, whereas Pax5 and Pax8 have milder cerebellar phenotypes, as knockouts of the gene in mice and zebrafish show. Because the cerebellum is present in jawless vertebrates, but not in invertebrate chordates, the origin of the cerebellum appears to be broadly coincident with the molecular events that differentiated the function of these three Pax genes in brain patterning.

With the ever-increasing amount of genomic and developmental data from a broad range of chordate model organisms, the study of early vertebrate evolution is entering a new phase that uses robust gene trees as a basis for interpretation of functional evolution. Two such case studies are presented here, and we are looking forward to more.

References

1. Ahn, S. and Tanksley, S.D. (1993) Linkage maps of the rice and maize genomes. *Proc. Natl. Acad. Sci. USA*, **90,** 7980−7984.
2. Adachi, J. and Hasegawa, M. (1992) *MOLPHY, Programs for Molecular Phylogenetics, I. PROTML, Maximum Likelihood Inference of Protein Phylogeny* (Computer Science Monographs, Vol. 27). Institute of Statistical Mathematics, Tokyo, Japan
3. Allendorf, F.W. and Utter, F.M., (1976) Gene duplication in the family Salmonidae. III. Linkage between two duplicated loci coding for aspartate aminotransferase in the cutthroat trout (*Salmo clarki*). *Hereditas*, **82,** 19–24.
4. Amores, A., Force, A., Yan, Y.-L., Joly, L., Amemiya, C., Fritz, A., Ho, R.K., Langeland, J., Prince, V., Wang, Y.-L. *et al*. (1998) Zebrafish *hox* clusters and vertebrate genome evolution. *Science*, **282,** 1711–1714.
5. Asano, M. and Gruss, P. (1992) *Pax−5* is expressed at the midbrain-hindbrain boundary during mouse development. *Mech. Dev*., **39,** 29−39.
6. Baker C.V.H. and Bronner-Fraser M. (1997a) The origins of the neural crest. Part I: Embryonic induction. *Mech. Dev*., **69,** 13−29.
7. Baker, C.V.H. and Bronner-Fraser, M. (1997b) The origins of the neural crest. Part II: An evolutionary perspective. *Mech. Dev*., **69,** 3−11.
8. Bouchard, M., Pfeffer, P. and Busslinger, M. (2000) Functional equivalence of the transcription factors Pax2 and Pax5 in mouse development. *Development*, **127,** 3703−3713.
9. Doyle, J.J., Doyle, J.L., Brown, A.H. and Grace, J.P. (1990) Multiple origins of polyploids in the *Glycine tabacina* complex inferred from chloroplast DNA polymorphism. *Proc. Natl. Acad. Sci. USA*, **87,** 714−717.
10. Gale, M.D. and Devos, K.M. (1997) Comparative genetics in the grasses *Proc. Natl. Acad. Sci. USA*, **95,** 1971−1974.
11. Ganter, B. and Lipsick, J.S. (1999) Myb and oncogenesis. *Adv. Cancer Res*., **76,** 21−60.
12. Gibson, T.J. and Spring, J. (2000) Evidence in favour of ancient octaploidy in the vertebrate genome. *Biochem. Soc. Trans*., **28,** 259−264.
13. Gilbert, S.F. and Raunio, A.M. (1997) *Embryology*, Sinauer, Sunderland, MA.
14. Heller, N. and Brändli, A.W. (1999) *Xenopus Pax−2/5/8* orthologues: novel insights into *Pax* gene evolution and identification of Pax−8 as the earliest marker for otic and pronephric cell lineages. *Dev. Genet*., **24,** 208−219.
15. Holland P.W.H. (1996) Molecular biology of lancelets: insights into development and evolution. *Israel J. Zool*., **42,** 247−272.
16. Hughes, A.L. (1999) Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history. *J. Mol. Evol*., **48,** 565−576.
17. Kardong, K.V. (1997) *Vertebrates*, 2nd ed., McGraw-Hill, New York, NY.
18. Kobel, H.R. and Du Pasquier, L. (1986) Genetics of polyploid *Xenopus*. *Trends Genet*., **2,** 310−315.
19. Kozmik, Z., Holland, N.D., Kalousova, A., Paces, J., Schubert, M. and Holland, L.Z. (1999) Characterization of an am-

phioxus paired box gene, *AmphiPax2/5/8*: developmental expression patterns in optic support cells, nephridium, thyroid-like structures and pharyngeal gill slits, but not in the midbrain-hindbrain boundary region. *Development*, **126,** 1295−1304.

20. Maddison, D.R. and Maddison, W.P. (2001) *MacClade Version 4.0*, Sinauer, Sunderland, MA.
21. Martin, A. (2001) Is tetralogy true? Lack of support for the 'one-to-four rule'. *Mol. Biol. Evol*., **18,** 89−93.
22. 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.
23. Millet, S., Bloch-Gallego, E., Simeone, A. and Alvarado-Mallart, R.M. (1996) The caudal limit of *Otx2* gene expression as a marker of the midbrain/hindbrain boundary: a study using in situ hybridisation and chick/quail homotopic grafts. *Development*, **122,** 3785−3797.
24. Murphy, P. and Hill, R.E. (1991) Expression of the mouse labial-like homeobox-containing genes, *Hox 2.9* and *Hox 1.6*, during segmentation of the hindbrain. *Development*, **111,** 61−74.
25. Pfeffer, P.L., Gerster, T., Lun, K., Brand, M. and Busslinger, M. (1998) Characterization of three novel members of the zebrafish *Pax2/5/8* family: dependency of *Pax5* and *Pax8* expression on the *Pax2.1* (*noi*) function. *Development*, **125,** 3063−3074.
26. Pfeffer, P.L., Payer, B., Reim, G., di Magliano, M.P. and Busslinger, M. (2002) The activation and maintenance of *Pax2* expression at the mid-hindbrain boundary is controlled by separate enhancers. *Development*, **129,** 307−318.
27. Pough, F.H., Janis, C.M. and Heiser J.B. (1999) *Vertebrate Life*, 5th ed., Prentice Hall, Upper Saddle River, NJ.
28. Shimeld, S.M. and Holland, P.W.H. (2000) Vertebrate innovations. *Proc. Natl. Acad. Sci. USA*, **97,** 4449−4452.
29. Sidow, A. (1996) Gen(om)e duplications in the evolution of early vertebrates. *Curr. Opin. Genet. Dev*., **6,** 715−22.
30. Simon, A.L., Stone, E.A. and Sidow, A. (2002) Inference of functional regions in proteins by quantification of evolutionary constraints. *Proc. Natl. Acad. Sci. USA*, **99,** 2912−2917.
31. Spring, J. (1997) Vertebrate evolution by interspecific hybridisation--are we polyploid? *FEBS Lett*., **400,** 2−8.
32. Wada, H., Saiga, H., Satoh, N. and Holland, P.W. (1998) Tripartite organization of the ancestral chordate brain and the antiquity of placodes: insights from ascidian *Pax−2/5/8*, *Hox* and *Otx* genes. *Development*, **125,** 1113−1122.
33. Wolfe, K.H. and Shields, D.C. (1997) Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, **387,** 708−713.