Vertebrates originated in the lower Cambrian. Their diversification and morphological innovations have been attributed to large-scale gene or genome duplications at the origin of the group. These duplications are predicted to have occurred in two rounds, the “2R” hypothesis, or they may have occurred in one genome duplication plus many segmental duplications, although these hypotheses are disputed. Under such models, most genes that are duplicated in all vertebrates should have originated during the same period. Previous work has shown that indeed duplications started after the speciation between vertebrates and the closest invertebrate, amphioxus, but have not set a clear ending.

Consideration of chordate phylogeny immediately shows the key position of cartilaginous vertebrates (Chondrichthyes) to answer this question. Did gene duplications occur as frequently during the 45 Myr between the cartilaginous/bony vertebrate split and the fish/tetrapode split as in the previous approximately 100 Myr? Although the time interval is relatively short, it is crucial to understanding the events at the origin of vertebrates. By a systematic appraisal of gene phylogenies, we show that significantly more duplications occurred before than after the cartilaginous/bony vertebrate split. Our results support rounds of gene or genome duplications during a limited period of early vertebrate evolution and allow a better characterization of these events.

Introduction

Vertebrates originated in the lower Cambrian (Shu et al. 2001), and their diversification and morphological innovations have been attributed to large-scale gene or genome duplications at the origin of the group (Ohno 1970; Holland et al. 1994). These duplications are predicted to have occurred in two rounds, the “2R” hypothesis, although it may have been one genome duplication plus many segmental duplications (Gu, Wang, and Gu 2002; McLysaght, Hokamp, and Wolfe 2002; Panopoulou et al. 2003). An interesting prediction of this hypothesis is that most genes that are duplicated in all vertebrates should have originated during the same period (for a discussion of predictions of the model, see Durand [2003]). Gene phylogenies consistent with this model are predicted to contain most duplications during a given speciation interval. The comparison of gene complexes, such as hox (Holland et al. 1994; Force, Amores, and Postlethwait 2002) or MHC (Abi-Rached et al. 2002), between species chosen for their key positions in the phylogeny of chordates, thus consistently date a large number of gene duplications after the divergence between the amphioxus and vertebrates (fig. 1). The choice of complexes of linked genes limits the insight these studies bring into the evolution of the whole genome, because each group of linked genes only samples one locus. Studies of the distribution and age of duplicated genes in the whole human genome sequence have established that gene duplications were indeed a massive phenomenon at the origin of vertebrates (Gu, Wang, and Gu 2002; McLysaght, Hokamp, and Wolfe 2002). However, because of their reliance on only one complete genome from a chordate and their reliance on the molecular clock, these studies cannot be very precise with respect to the dating and to the order of events, although efforts were done to add more species to the gene trees. In a pioneering comparison of phylogenies of unlinked genes, the tree topologies obtained were inconsistent with a simple scenario of two rounds of tetraploidization (Hughes 1999), but no dating of events was proposed. Phylogenies of gene families from various chordates show similar numbers of duplications before and after the lamprey/hagfish/gnathostome split, but results are not explained simply by two tetraploidizations (Escriva et al. 2002). All of these results are consistent with periods of intensive gene duplication, rather than genome duplication (Gu, Wang, and Gu 2002), although a recent phylogenetic study challenges even this scenario (Friedman and Hughes 2003).

Overall, there is support for a large number of gene duplications after the divergence between cephalochordates and vertebrates (Panopoulou et al. 2003), both before and after the lamprey/hagfish/gnathostome split (Escriva et al. 2002). This possibility leaves an important question mark on the ending time of the duplication events, which could represent a punctual event or could have occurred gradually over a period of 160 to 300 Myr. Consideration of chordate phylogeny (fig. 1) immediately shows the key position of chondrichthians: if the massive gene duplications occurred almost exclusively before or after the chondrichthyan (cartilaginous vertebrates)/teleostome (bony vertebrates) split, this event supports “rounds” of duplications during a limited period of early vertebrate evolution. Otherwise, if gene duplications are evenly spread over the period between the cephalochordate/vertebrate split and the actinopterygian/sarcopterygian split, there is no evidence for these “rounds,” but rather for a long period during which duplication was more frequent than in sarcopterygian evolution. Most studies do not include chondrichthians, with the exception of two genes linked to the MHC, which were shown to be duplicated before the divergence of chondrichthians and teleostomes (Abi-Rached et al. 2002).

Lack of chondrichthyan genome data has led us to use...
the gene phylogeny approach to solve the question of when vertebrate-specific gene duplications did happen, by constructing phylogenetic trees of many protein-coding genes sequenced in Chondrichthyes. 

As mentioned above, if there were two major rounds of duplication, whether of genes or genomes, we would expect most gene families to show similar relative timing of speciation and duplication events. It should be noted that we are only interested in vertebrate-specific duplications here. Duplications that predate the chordate/arthropod/nematode split (approximately the origin of bilaterian animals), or more recent duplications such as frequently observed in actinopterygian fishes (Robinson-Rechavi et al. 2001), are outside the scope of this study.

Materials and Methods

Data Set

A first selection of gene families was done on Hovergen (Duret, Mouchiroud, and Gouy 1994) version 42 (April 2002), with the following criteria: at least one Chondrichthyes sequence, sequences from at least two Teleostome classes (to distinguish vertebrate specific and class specific duplications), and exclusion of mitochondrion-encoded genes. These criteria selected 149 gene families, as defined in Hovergen, including 415 chondrichthyan protein sequences. Protein alignments corresponding to the selected families were saved from Hovergen and checked using Seaview (Galtier, Gouy, and Gautier 1996). Outgroup sequences were added by Blast (Altschul et al. 1990) searches on Swissprot+TrEMBL (Boeckmann et al. 2003), excluding results from Vertebrata and from viruses, as implemented at PBIL (Perrière et al. 2001), and by Blast searches on the genome sequences of Drosophila melanogaster (Adams et al. 2000), Caenorhabditis elegans (The C. elegans Sequencing Consortium 1998), Ciona intestinalis (Dehal et al. 2002), and Anopheles gambiae (Holt et al. 2002). Twelve gene families for which no outgroup sequence could be reliably identified were excluded.

Gene families with duplications predating the arthropod/nematode/chordate divergence (fig. 2A) were split into subfamilies, which were then evaluated separately for vertebrate-specific duplications. In cases of a vertebrate gene without any known mammalian ortholog, additional Blast searches were done on the human genome (International Human Genome Sequencing Consortium 2001). In all Blast searches, an expect value of 0.01 and the default filter for repeated sequences were used, and potential new genes were assessed for relevance to our study by a phylogenetic analysis. Once gene trees were built (see below), 86 gene families were found to yield phylogenies that could not be interpreted for dating of events at the origin of vertebrates (see Results). Notably, insufficient phylogenetic resolution was diagnosed when the gene tree was strongly inconsistent with the expected species phylogeny (for example, lamprey grouping with chicken and mammals not monophyletic [NPY gene family]) with very low bootstrap support (i.e., under 50%).

Phylogeny

All analyses were done using only complete sites (no gap, no X). When the inclusion of partial sequences led to less than 50 complete sites in the alignment, these sequences were excluded manually in Phylo_win (Galtier, Gouy, and Gautier 1996), taking care to keep representatives of each taxonomic group (i.e., actinopterygians, sarcopterygians, chondrichthyans, and outgroup) and of each paralog, as much as possible. Sequences that did not pass a $\chi^2$ test for homogeneity of amino acid composition (as implemented in Tree-Puzzle [Schmidt et al. 2002])
were excluded. This exclusion meant that six gene families no longer fulfilled the conditions set in terms of species sampling and were thus excluded from the data set. Trees were constructed using Neighbor-Joining (Saitou and Nei 1987) with distances corrected for multiple substitutions under a gamma model of rate heterogeneity (Yang 1996); the alpha parameter of the gamma model was estimated for each alignment by Tree-Puzzle version 5.1 (Schmidt et al. 2002) with eight rate categories, using default parameters. The following topologies were systematically compared by an SH likelihood test (Shimodaira and Hasegawa 1999), under the VT substitution model (Muller and Vingron 2000) with a c model of rate heterogeneity, as implemented in Tree-Puzzle 5.1 (Schmidt et al. 2002): (1) species tree with no duplication (fig. 2A), (2) duplication after the chondrichthyan/teleostome split (fig. 2B), (3) duplication before the chondrichthyan/teleostome split (fig. 2C). When there were more than two vertebrate paralogs, all relative positions of the chondrichthyan/teleostome split and the duplications were compared; for example (Chondr (Teleos-α (Teleos-β, Teleos-γ))) versus (Teleos-α (Chondr (Teleos-β, Teleos-γ))) versus (Teleos-α (Teleos-β (Chondr, Teleos-γ)))) versus ((Teleos-α, Chondr), (Teleos-β, Teleos-γ)) and so on. Results were considered supported if the likelihood of the favored topology was significantly higher than that of the best alternative topology (SH test; \( P < 0.05 \)). Other results are classified as “not supported.” It should be noted that we are only interested in the relative order of events of gene duplication and the chondrichthyan/teleostome split. Thus, teleost fish-specific duplications, as well as contradictions between gene phylogeny and teleostome phylogeny, as long as the latter were not statistically supported (they never were), were not taken into consideration to classify phylogenetic results, as far as they do not hamper interpretation of the trees. Moreover, when there were inaccuracies in teleostome phylogeny in the Neighbor-Joining tree, likelihood tests were performed under both the Neighbor-Joining and the species topology; significance of results was robust to the change.

Results

We selected gene families for the study in three steps: (1) selection on taxonomic criteria (sampling of cartilaginous and bony vertebrates, outgroup sequence); (2) manual consideration of phylogenetic trees, to assess whether the gene families are appropriate to the question being asked; and (3) evaluation of phylogenetic robustness. Notably, a total of 86 gene families were eliminated in step 2. The main causes limiting interpretation were (1) after splitting into vertebrate-specific subfamilies, some genes no longer fulfill the conditions set in terms of species sampling (typically the chondrichthyan sequence fell in a subtree with mammalian sequences and no other taxa); (2) very short sequences (NPY genes for example) with no
phylogenetic resolution; (3) extremely conserved sequences with no phylogenetic resolution (histones for example); (4) clustered multigene families for which conversion and recombination are well documented, typically from the immunological system; and (5) other genes with no phylogenetic resolution, such as hox genes, which include a very conserved homeodomain, with little information, the rest of the sequence being very divergent and with little information also (see a discussion in Force, Amores, and Postlethwait [2002]). It may be noted that while this selection mostly reduced the number of gene families used, splitting families with duplications predating the arthropod/nematode/chordate divergence increased the number of phylogenies analyzed (two additional “families” of pro teasea beta subunit genes and one additional “family” of tyrosine phosphatase genes [see table 1 in Supplementary Material online]). Overall, the three steps of selection lead us from 149 gene families with cartilaginous and bony vertebrate homologous sequences to 48 gene families whose evolutionary history can be used to date duplication events at the origin of vertebrates (eliminated gene families in table 3 of Supplementary Material online), a figure very similar to the numbers of genes analyzed in recent studies using the same approach in other organisms (i.e., Langkjaer et al. 2003; Taylor et al. 2003).

Results for each gene family are detailed in the first table and the figures in the Supplementary Material online at www.mbe.oupjournals.org. Gene families with a duplication before the chondrichthyan/teleostome split (fig. 2C) clearly represent the majority of gene families we analyzed, including all 19 genes with significant phylogenetic resolution (table 1). Among the other 29 gene families, phylogenetic resolution is not significant at the chondrichthyan/teleostome divergence level (table 1). These include only the two gene families indicating a duplication after the chondrichthyan/teleostome split: a mannos e-binding lectin, or tetranectin (HBG008208), and the PTP1D tyrosine phosphatase (“tyrosines phosphatases (1)” in the Supplementary Material online). Of note, a different result was found for PTP1D in a previous study that did not include all available mammalian sequences (Ono-Koyanagi et al. 2000). Finally, 15 genes show no evidence for any vertebrate-specific duplication. Our classification of these trees as “not supported” means that the species tree was not significantly more likely than other positions of chondrichthyan. This is consistent with a previous study in which individual nuclear genes had low power in solving the phylogenetic position of chondrichthyanas (Martin 2001). The low phylogenetic resolution for the position of chondrichthyanas among vertebrates is also consistent with the small divergence time between chondrichthyan and teleostomes reported in the fossil record (fig. 1). By contrast, the good phylogenetic resolution for the position of vertebrate-specific gene duplications may imply that the divergence time between these duplications and the chondrichthyan/teleostome split was important and that the duplications occurred early in vertebrate evolution.

It is possible that the observed distribution of gene duplications simply reflects the difference between the time intervals considered as “before the chondrichthyan/teleostome split” and “after the chondrichthyan/teleostome split.” To test this, let us consider only the 27 gene families for which we have a vertebrate-specific duplication and a chordate outgroup (table 1: 16 + 9 + 2 = 27), since they allow a more precise dating of events. If we use paleontological datings (fig. 1), the interval between chordate diversification and the chondrichthyan/teleostome split is 98 Myr, whereas the interval between this and the sarcopterygian/actinopterygian split is 45 Myr. Then we expect 31% (45/[98 + 45]) of vertebrate-specific gene duplications to be after the chondrichthyan/teleostome (C/T) split, under the assumption of a constant rate of gene duplication; the 95% confidence interval of this estimate is 14% to 49% (f ± 1.96 √var = f(1 − f)/N; f = 0.31; N = 27). If we use molecular clock estimates of divergence dates (fig. 1), we expect 26% of gene duplications after C/T (confidence interval = 9.5% to 43%). The observed proportion of 7.4% (2/27) is significantly lower than expected by chance in either dating system (outside of the 95% confidence intervals). This conclusion holds true if we only use the 16 significantly supported phylogenies with a chordate outgroup (table 1): the observed proportion of duplications after the C/T split is 0%, whereas the expected value’s confidence interval is either 8.7% to 54% (paleontological dates), or 4.5% to 47% (molecular clock dates). Thus, gene duplications are significantly less frequent after than before the chondrichthyan/teleostome split, taking into account evolutionary time.

Although our data set is not meant for detailed testing of duplication hypotheses in other branches of the tree, it is interesting to compare duplications that appear specific to either of the two major branches of teleostomes: out of 48 gene families, there are three with sarcopterygian-specific duplications and eight with actinopterygian-specific

### Table 1

#### Distribution of Duplication Histories of Gene Families

<table>
<thead>
<tr>
<th>Outgroup</th>
<th>None</th>
<th>Before C/T Split</th>
<th>After C/T Split</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Chordate</td>
<td>Other</td>
<td>Chordate</td>
</tr>
<tr>
<td>Significant</td>
<td>0</td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>Not significant</td>
<td>12</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>Total</td>
<td>15 (31%)</td>
<td>31 (65%)</td>
<td>2 (4%)</td>
</tr>
</tbody>
</table>

Note.—Numbers of gene families supporting each evolutionary history; no gene is counted twice. Phylogenetic support is noted as “significant” if the position of the chondrichthyan gene(s) is supported by a likelihood test (P < 0.05). The “C/T” split is the divergence between Chondrichthyes and Teleostomi. Chordate outgroups include ascidians, amphioxuses, lampreys, and hagfishes.
duplications (see the second table in the Supplementary Material online at www.mbe.oupjournals.org), consistent with previous observations (Robinson-Rechavi et al. 2001). Interestingly, these more recent duplications concern 28% of the 32 gene families for which we have observed gene duplications ancestral to vertebrates but only 12.5% of the 16 gene families without vertebrate specific duplications.

Discussion

The “2R” hypothesis, modified from Ohno (1970), can be summarized by the idea that major duplication events occurred specifically in chordate genomes before the emergence of bony vertebrates. This hypothesis predicts that duplications should have occurred over a short period of time, in much greater numbers than in the previous or following periods. This prediction is shared by more recent hypotheses that there may have been one genome duplication and one major wave of segmental duplications (Gu, Wang, and Gu 2002; McLysaght, Hokamp, and Wolfe 2002; Panopoulou et al. 2003). The beginning time has been relatively well established, with studies showing that gene duplications occurred after the cephalochordate/vertebrate split and both before and after the gnathostome/jawless vertebrate split (Pennisi 2001; Wolfe 2001; Abi-Rached et al. 2002; Escriva et al. 2002; Gu, Wang, and Gu 2002; McLysaght, Hokamp, and Wolfe 2002; Panopoulou et al. 2003), but these studies did not set an ending time to these events. Given the prevalence of gene duplications in actinopterygian fishes (Wittbrodt, Meyer, and Schartl 1998; Robinson-Rechavi et al. 2001; Taylor et al. 2001), this raises the question of whether something specific really happened at the origin of vertebrates or whether gene duplications have been a common phenomenon throughout chordate evolutionary history, with the exception of sarcopterygians.

It is indeed noticeable that there has been no report of genome duplications ancestral to sarcopterygians (Pennisi 2001; Wolfe 2001; Durand 2003) or to any of the well-studied groups therein (e.g., tetrapods, mammals, or sauropsids). Our own data are consistent with previous observations (Robinson-Rechavi et al. 2001; Taylor et al. 2001) that duplicate genes are significantly less abundant in sarcopterygians than in actinopterygians. Analysis of invertebrate chordate data also indicates that gene duplications are not abundant in these lineages (Dehal et al. 2002; Panopoulou et al. 2003).

Comparison of MHC-associated genes gave limited evidence for duplications before the chondrichthyan/teleostome split from two genes (Abi-Rached et al. 2002). Our results show that this pattern is general, with almost all vertebrate-specific gene duplications occurring before the chondrichthyan/teleostome split (table 1). This, added to all the previously published evidence, implies three waves of gene or genome duplications, two between the cephalochordate split and the chondrichthyan split and the other in actinopterygian fishes, separated by a period of “duplication calm” of about 45 Myr (which continued for 400 Myr in tetrapodes), which, although short, is significant. A major prediction of Ohno’s (1970) original hypothesis, that of intense gene or genome duplication activity before the origin of vertebrates, is thus confirmed by the study.

Moreover, our results show that these gene duplications characterize all the jawed vertebrates and predict similar genetic complexity in sharks and rays as in tetrapodes. Consistent results are found for the evolution of hox clusters, which allow a direct connection between block duplications and morphological adaptations. Although hox genes are very poor phylogenetic markers, as illustrated by the difficulty in resolving the events that led to the different clusters of gnathostomes and lampreys (Force, Amores, and Postlethwait 2002; Irvine et al. 2002), partial sequences from the horn shark indicate that the duplications that led to four hox clusters in teleostomes occurred before the chondrichthyan/teleostome divergence (Kim et al. 2000). Moreover, horn shark and human hoxA clusters are remarkably conserved (Chiu et al. 2002). Thus, hox cluster analysis and our phylogenetic results are consistent in establishing no relation between gene duplications and the larger diversity of bony vertebrates than of cartilaginous vertebrates.

Although the basal branching of chondrichthyans among jawed vertebrates is considered extremely well supported by morphological and paleontological data (Janvier 1996), the analysis of complete mitochondrial sequences suggests a very different phylogeny, with chondrichthyans branching among bony ray-finned fishes (Actinopterygii) (Rasmussen and Arnnson 1999). This surprising result has not been confirmed by any other source of data, and molecular phylogenies based on nuclear-encoded genes either are not informative (Martin 2001; this study) or strongly support the conventional branching position of chondrichthyans (Takezaki et al. 2003). In any case, our results show that vertebrate-specific gene duplications occurred before the divergence between chondrichthyans, actinopterygians, and sarcopterygians, whatever the order of these latter events.

Our results are at odds with a recent study that used a similar approach, dating gene duplications by their phylogenetic position relative to speciation events (Friedman and Hughes 2003). There are several differences between our methodology and that of Friedman and Hughes, but the main difference is the criterion for classifying gene duplications within speciation intervals. We consider genes to be duplicated within a given interval (i.e., between chordate diversification and the chondrichthyan/teleostome split) only if all relevant taxonomic groups (and thus speciations) are represented in the gene tree (i.e., a urochordate or a cephalochordate, a chondrichthyan, and a teleostome). Friedman and Hughes (2003) classify duplications as soon as they can be dated before or after one speciation. Moreover they used very distant dating points (i.e. the primate/rodent, amniote/amphibian, and deuterostome/protostome splits). It is unclear why they did not date duplications relative to the actinopterygian/sarcopterygian split, because this speciation would have been more relevant to the “2R” controversy, while taking advantage of genome data. As amphibians are the only lineage involved for which a genome sequence is not available, this may lead them
to include in the “before primate/rodent” category gene duplications that occurred before the amphibian/amniote split but for which they do not have amphibian sequences in the tree. This in turn may introduce a bias in their argument that the abundance of “before primate/rodent” versus “before amniote/amphibian” duplications is evidence against a peak of gene duplications at the origin of vertebrates. We believe that in our study, the division of the sequences into major taxonomic units, and our separation of the results according to the outgroup sequences used (table 1), preserve our results from such biases. Thus, differences in the conclusions between that study (Friedman and Hughes 2003) and ours probably reflect different sampling strategies.

An interesting side observation from our data set is that observations of gene duplications at the origin of vertebrates, and more recently in either the actinopterygian or sarcopterygian lineage, appear correlated. This may be the result of sampling; for example, better detection of duplications in more studied genes. Alternatively, it may indicate that the function of certain genes makes them more prone to persisting as duplicate copies. Such a tendency has indeed been recently shown in yeasts, where certain genes are retained independently as duplicates in different species (Hughes and Friedman 2003).

This study and other recent studies draw an increasingly precise picture of gene or genome duplication waves in chordates (fig. 3), although questions remain. Among the six branches of the chordate tree for which sufficient data are available, three are characterized by abundant preservation of duplicate genes, all of them in vertebrates. It has also been suggested on the basis of chromosome counts that polyploidy played an important part in lamprey evolution (Potter and Rothwell 1970). Of course it is probable that small-scale duplications have been continuous on all branches of the tree (Lynch and Conery 2000; Gu, Wang, and Gu 2002). However, large-scale duplications seem to have been frequent in vertebrate evolution, and the branches where they are absent, such as the origin of bony vertebrates, appear as the exception rather than the rule.

Acknowledgments

We thank Hector Escriva and Manolo Gouy for critical reading. This work was supported by the CNRS and the ENS Lyon.

Literature Cited


Perrière, G., C. Combet, S. Penel et al. (11 co-authors). 2003. Integrated databanks access and sequence/structure analysis services at the PBIL. Nucleic Acids Res. 31:3393–3399.


Robinson-Rechavi et al. Accepted October 25, 2003


Hervé Phillippe, Associate Editor