J Mol Evol (2004) 58:168–181
DOI: 10.1007/s00239-003-2538-y

JOURNAL OF MOLECULAR EVOLUTION

Phylogenetic Analysis of *Ciona intestinalis* Gene Superfamilies Supports the Hypothesis of Successive Gene Expansions

Magalie Leveugle,[1,2,*] Karine Prat,[1,3,*] Cornel Popovici,[1] Daniel Birnbaum,[1] François Coulier[1,2,4]

[1] Département d'Oncologie Moléculaire, Unité 119 INSERM, IFR57, Marseille, France
[2] Université de la Méditerranée, Marseille, France
[3] Laboratoire de Minéralogie-Cristallographie LMCP, CNRS-UMR 7590C, Paris, France
[4] Atelier de Bioinformatique, INSERM Unité 119, Marseille, France

Received: 8 April 2003 / Accepted: 4 August 2003

Abstract. Understanding the formation of metazoan multigene families is a good approach to reconstitute the evolution of the chordate genome. In this attempt, the analysis of the genome of selected species provides valuable information. *Ciona intestinalis* belongs to the urochordates, whose lineage separated from the chordate lineage that later gave birth to vertebrates. We have searched available sequences from the small marine ascidian *C. intestinalis* for orthologs of members of five vertebrate superfamilies, including tyrosine kinase receptors, ETS, FOX and SOX transcription factors, and WNT secreted regulatory factors, and conducted phylogenetic analyses. We have found that most vertebrate subfamilies have a single *C. intestinalis* ortholog. Our results support the hypothesis of a gene expansion prior the base of chordate ancestry followed by another gene expansion during vertebrate evolution. They also indicate that *Ciona intestinalis* genome will be a very valuable tool for evolutionary analyses.

Key words: 2R hypothesis — *Ciona intestinalis* — Gene duplication — Genome — Paralogy — Orthology — Urochordates

Introduction

Gene duplication is believed to be a key force of genome evolution. This has been supported by the recent analyses of several genomes (Wolfe and Shields 1997; The Arabidopsis Genome Initiative 2000). They can occur as small-scale and large-scale duplications. Large-scale duplications, possibly in the form of polyploidizations, are thought by many authors to have molded early vertebrate evolution and, to some extent, brought about vertebrate innovations, although there is no clear relationship between complexity and gene number (Graham 2000; Shimeld and Holland 2000; Holland and Chen 2001). It is believed that the vertebrate ancestor had a single gene corresponding to a gene family of two, three, or four members in present-day tetrapods due to a rapid increase in the number of genes caused by these large-scale duplications (Ohno 1970; Schughart et al. 1989; Lundin 1993; Holland et al. 1994; Spring 1997; Pébusque et al. 1998; Popovici et al. 2001b). This is sometimes known as the "big-bang model." This model further describes two rounds of such large-scale duplication after the divergence of vertebrates from the cephalochordates; this is known as the "2R hypothesis" (for recent reviews see Wolfe 2001; Taylor and Brinkmann 2001). Additional duplications have occurred in the fish lineage (Amores et al. 1998; Aparicio 2000) and in Amphibia (see Fig. 1). Large-scale duplications may also have occurred earlier but are less well documented.

*M.L. and K.P. contributed equally to this work.
Correspondence to: F. Coulier, U.119 Inserm, 27 bd. Leï Roure, 13009 Marseille, France; *email:* coulier@marseille.inserm.fr

Another important mechanism has influenced vertebrate evolution. It is based on gene-by-gene small-scale and segmental duplications that created new paralogous genes in a continuous flux (Gu et al. 2002). Regarding vertebrate evolution, alternative explanations challenging the 2R hypothesis have been proposed that include a continuous mode of evolution made of these small-scale and segmental duplications only (Hughes et al. 2001; Friedman and Hughes 2001; Martin 2001; Page and Cotton 2002; Friedman and Hughes 2003). Finally, if the importance of gene duplications, whether common (i.e., early) or lineage-specific (i.e., late) (Popovici et al. 1999; Lespinet et al. 2002; Minguillon et al. 2002), is increasingly recognized, gene conversion (Gogarten and Olendzenski 1999) complicates our understanding of vertebrate genome history.

For many metazoan gene subfamilies, a single orthologous gene may be recognized in nonvertebrates. In vertebrates, chromosomal regions that contain paralogs (i.e., paralogous regions or paralogons; PGs) have been identified (Popovici et al. 2001b; McLysaght et al. 2002). They are likely remnants of the large-scale duplications proposed to explain the increase in gene number in vertebrates. This is the case of genes encoding proteins with homeodomains, such as HOX, ParaHOX, and MetaHOX/NKL (Brooke et al. 1998; Coulier et al. 2000a, b; Pollard and Holland 2000; Popovici et al. 2001a). Several previous works have recorded genes potentially deriving from large-scale duplications (Lundin 1993; Birnbaum et al. 1994; Spring 1997; Ollendorff et al. 1998; Gibson and Spring 1999; Wang and Gu 2000; Popovici et al. 2001b; but see also Skrabanek and Wolfe 1998). However, the analysis of the sequence of the human genome did not reveal a peak of gene families with four members and brought only partial support for the 2R hypothesis (Lander et al. 2000; Li et al. 2000; Venter et al. 2000; McLysaght et al. 2002). If the 2R hypothesis is true, it suggests that evolution of genes in vertebrates has included extensive loss of members in gene families and has erased most traces of the large-scale duplications; if gene duplication is an important phenomenon, subsequent gene loss must be their true yin-yang countereffect (Lynch and Conery 2000; Wagner 2001). Alternatively, gene losses need not be so extensive if expansion has occurred via continuous small-scale duplications. The greater number of vertebrate genes might not be due to expansion but to adaptive radiation with greater retention and fixation of gene duplicates (Friedman and Hughes 2003). Yet a strong argument in favor of large-scale duplications and of the 2R hypothesis is the existence of blocks of duplicated genes and paralogons. This has been clearly demonstrated by recent works that have taken advantage of the availability of the human genome sequence (McLysaght et al. 2002) or of comparison with a nonvertebrate chordate genome (Abi-Rached et al. 2002).

Previous genome analyses (Adams et al. 2001; The *C. elegans* Sequencing Consortium 1998) have shown a greater number of genes in vertebrates that in protostomians. The genome of chordate nonvertebrate species may be similar to those of protostomians (no expansion), similar to those of vertebrates (expansion), or intermediate (incomplete expansion). Chordate species may be found in the cephalochordates and the urochordates (Makalowski 2001; Stach and Turbeville 2002). We chose to investigate examples of gene superfamilies in the urochordate ascidian *Ciona intestinalis* since genomic sequences from this species have been released recently. We constructed phylogenetic trees of five gene superfamilies, i.e., genes encoding tyrosine kinase receptors (RTKs), FOX (forkhead box), SOX (sex determining region Y-box), and ETS transcription factors, and WNT (wingless-type integration site) secreted regulatory factors. The hypothesis of gene expansion in vertebrates predicts that a vertebrate gene subfamily will have a single ortholog (or few coorthologs in the case of a recent duplication in the ascidian lineage) in the *C. intestinalis* genome. We found that this is indeed the case: most identified vertebrate subfamilies had only one *C. intestinalis* ortholog. Our results support the hypotheses of gene expansion at the base of vertebrate ancestry after the separation from the urochordates. In addition, since many small-scale duplications have occurred in gene superfamilies, the identification of *C. intestinalis* orthologs allows us to delineate the period of these events.

Materials and Methods

Definitions

Families and superfamilies are the result of early duplications that took place before the urochordate/chordate split, while paralogs, which constitute subfamilies, derive from more recent duplications that took place after the urochordate/chordate split and before the apparition of vertebrates.

The following definitions are commonly used: two genes are orthologs if they diverged due to a speciation event; they are paralogs if they diverged due to duplication within a lineage (Fitch 1970; Koonin 2001). Therefore, when there is a speciation event followed by duplication events in both derived lineages, genes from the resulting multigenic family in one species are orthologous to any gene of the resulting family in the second species. Within each species the genes forming the multigenic family are paralogous. Although the term paralog can be used to designate genes within the same species deriving from any type of duplication, we use it in this paper only to designate the members of a vertebrate gene subfamily. Paralogs constitute subfamilies. Genes issued from ancient duplications that separated families are "metaparalogs" (i.e., different classes of RTKs). Families and superfamilies are higher orders of classification (Popovici et al. 2001b). A series of paralogous regions, within the same species, that could be recognized as

deriving from a common ancestor region is called a paralogon (PG) (Coulier et al. 2000a; Popovici et al. 2001b; McLysaght et al. 2002). Genes that belong to the same paralogon are "coparalogs," whether or not they are related by sequence similarities. When possible, one may use the term "direct orthologs" to specify pairs of genes that have a correspondence across species (for example, FGFR1 in humans and FGFR1 in the mouse).

We used the term "large-scale duplication" to indicate genome-size duplication (e.g., polyploidization in relation with the 2R hypothesis). We used the term small-scale duplication for gene-size or segmental duplication, e.g., gene duplications in relation to other theories (Hughes et al. 2001).

Selection of Superfamilies

We selected the superfamilies to be studied based on the following criteria: (i) distribution of members in several key species including protostomians; (ii) distribution of members throughout the whole set of paralogons defined in a vertebrate genome with available sequence (here the human genome); (iii) existence of a conserved domain with a length allowing sequence comparisons and phylogenetic analyses; (iv) presence of at least six subfamilies in a superfamily; and (v) to avoid heterogeneity in selective pressure, role of members in regulatory and developmental processes (Gerhart and Kirchner, 1997).

Family Tree Construction

All protein sequences for each selected superfamily were gathered using Blast searches against the nr database of proteins and classified with phylogenetic studies. Families and subfamilies of genes were named according to previous classifications (Laudet et al. 1999; Bowles et al. 2000; Kaestner et al. 2000; Schubert et al. 2000; Grassot et al. 2003). Multiple alignments were done with Clustalx 1.81 (default parameter) for Linux (Thompson et al. 1997). Phylogeny analysis was done with PhyloWin (Galtier et al. 1996) (neighbor-joining, Poisson distance correction, and maximum parsimony methods, global gap removal, and 500 bootstrap replicates) with either complete protein sequences or family specific domains.

Maximum likelihood analyses were done on combinations of subfamilies with the Phylip 3.6 package (proml program with Jones–Taylor–Thornton amino acid change model).

Gene Search and Reconstitution of C. intestinalis *Genes*

Specific domains of each protein family were compared with Blast (default parameter) (Altschul et al. 1997) against the JGI *Ciona intestinalis* Whole Genome Shotgun reads (WGS) database (http://www.jgi.doe.gov/programs/ciona.htm) to find genomic sequences with similarity to the vertebrate proteins. All the resulting reads were collected and assembled with the Cap3 (default parameter) program (Huang and Madan 1999). After the first round of assembly, we obtained a collection of contigs (~12 per family). Each contig was scanned with GenScan (with the human matrix) to detect exons (Burge and Karlin 1997). Putative proteins were compared with the nr database to identify which family they could belong to. Only the proteins matching one of the family of interest were kept, except in the case of the SRC and CIC *Ciona intestinalis* proteins, which were used as outgroups with RTK and SOX families, respectively. Contigs that code for putative proteins of interest were extended by Blast against the *C. intestinalis* WGS database. Only alignments with an *E* value of 0 were selected to increase the probability the corresponding reads belong to the genomic contig. Moreover, the assembly program Cap3 selected only sequences with sufficient overlaping. New read sequences were assembled in the contigs which were scanned for exon sequences. All contigs were extended to obtain complete genomic sequences encoding specific domains (or in some cases complete proteins) for each subfamily. *C. intestinalis* predicted cDNAs were compared with the *C. intestinalis* cDNA database to complete our data set, or to correct possible mispredictions due to the use of human matrix in GenScan, when the cDNA was (partially or not) available.

When *Ciona intestinalis* predicted ORF sequences became available (Dehal et al. 2002), we checked that our predictions were in close agreement with the now available sequences (data not shown).

Nomenclature of Ciona intestinalis *Genes*

When they belong to a defined group or subfamily, *C. intestinalis* proteins were named according to the name of the group, except in the RTK family, where they were named according to the most popular human gene of the group.

Results

An Expansion of Genes Postdates the Separation of Urochordates from the Other Chordates

Due to its position in the tree of life (http://tolweb.org/tree/phylogeny.html), the small marine ascidian *Ciona intestinalis* is a key species to study chordate evolution (Fig. 1). It has around 16,000 genes coded by a genome of ~160 Mb, similar in size to that of *Drosophila melanogaster*. The *C. intestinalis* genome is remarkably homogeneous in base composition as well as in gene distribution, contrasting with vertebrate genomes (de Luca di Roseto et al. 2002). A genome project has made available *C. intestinalis* genomic draft sequences (http://www.jgi.doe.gov/programs/ciona.htm) and ESTs (http://ghost.zool.kyoto-u.ac.jp/indexr1.html).

We previously constructed a database dedicated to the study of paralogous relationships in vertebrates (Leveugle et al. 2003). In this database (http://abi.marseille.inserm.fr/paradb/), a large number of gene superfamilies are inventoried, paralogous relationships are established for their members, and paralogons are identified in the human genome. From this database, we selected five superfamilies of proteins: RTKs, the FOX, SOX, and ETS transcription factors, and the WNT secreted regulatory factors.

Each selected family was investigated to find orthologous genes in the *Ciona intestinalis* genome. *C. intestinalis* proteins were aligned with vertebrate and nonvertebrate proteins from representative lineages: human, mouse, chicken, zebrafish, clawed frog, and fruit fly. Other species (silkworm, trout, quail, etc.) were used when these families were not available from the representative lineages (see Table 1 for a list of all species used). We aimed to have one representative for several taxa (insects, actinopterygii,

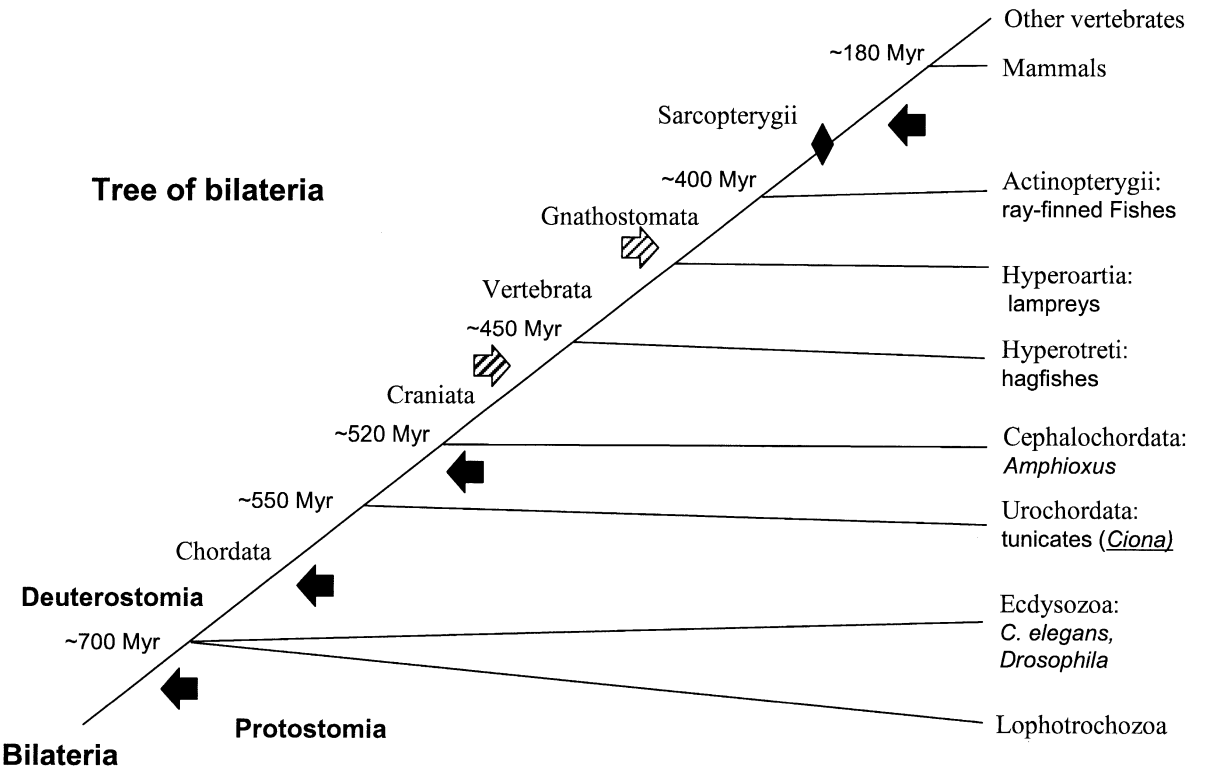


Fig. 1. Phylogenetic tree of the bilaterian lineage. Series of duplications shown by arrows are positioned along the tree and phyla are indicated after each node (http://tolweb.org/tree/phylogeny.html). Two types of duplications, large-scale (hatched arrows) and small-scale (filled arrows), that may have molded metazoan evolution are indicated. Large-scale duplications are thought to have occurred in two rounds (Ohno 1970), whereas small-scale made a continuous flux of new gene creations (Gu et al. 2002). Like small-scale duplications, gene losses (filled diamond) are probably important at every period. Time periods are tentatively indicated (Myr: megayears). The *Ciona* lineage, which is the focus of this paper, is underlined.

mammals, amphibians, and birds). Orthology and family membership of *C. intestinalis* proteins were established with phylogenetic trees. The topology of the trees was the same whether using neighbor-joining (Figs. 2–7) or parsimony methods (not shown); in some cases, maximum likelihood algorithm was used.

According to structural and evolutionary considerations, the RTKs can be separated into subfamilies (commonly designated classes); we named these classes according to the RTKdb nomenclature (Grassot et al., 2003; http://pbil.univ-lyon1.fr/RTKdb/). A total of 15 *C. intestinalis* genes encoding putative RTKs was identified. We did a neighbor-joining phylogenetic analysis of the RTK kinase domain, including the 15 *C. intestinalis* RTK kinase domains. The analysis was done with 183 kinase domain sequences (164 sites) and 500 bootstrap replicates. *C. intestinalis* sequences were named according to the most usual name of the class or to the class name. The tree (Fig. 2) was rooted with the SRC nonreceptor protein kinase family, which contains one *C. intestinalis* protein. Only bootstrap values over 50 are displayed in Fig. 2. Nodes with bootstraps over 70% are thought to be robust (Felsenstein 1985; Hillis and Bull 1993). For 10 classes, a single *C. intestinalis* RTK was found (FGFR, VEGFR, IGFR, DDR, ROR, RYK, EGFR, RTK VIII, RTK X, and PTK7) and the orthology relationships were clear. We found five *C. intestinalis* tyrosine kinases domains that belong to the Ephrin receptor class (RTKVI). It was not possible to identify direct orthologs of these molecules in vertebrates; it is likely that lineage-specific small-scale duplications have occurred in this class. For some of the classes (i.e., classes III and VII), no *C. intestinalis* RTKs could be found. There are several possible explanations for this. First, there were no available sequences in the database yet or we failed to retrieve them; second, the corresponding *C. intestinalis* gene was lost in the urochordate lineage; third, these classes of RTK appeared after the separation of urochordates from the other chordates. These latter two possibilities are discussed further below. We found one more RTK fragment which matches with one EST in the *C. intestinalis* cDNA database project; according to Blast results this fragment could belong to the ROS class (RTK XIII) but the kinase domain could not be extended enough to be integrated in the tree.

Very similarly, we found a single *Ciona* ortholog for most subfamilies of the other four superfamilies (Figs. 4–6). Figure 3 shows the neighbor-joining phylogeny analysis of FOX proteins based on the

Table 1. Abbreviations used for species names in the trees: Genus initial (uppercase) and first two letters of species

Abbreviation	Species name
Ame	*Ambystoma mexicanum*
Bbe	*Branchiostoma belcheri*
Bfl	*Branchiostoma floridae*
Bla	*Branchiostoma lanceolatum*
Bmo	*Bombyx mori*
Bta	*Bos taurus*
Cau	*Carassius auratus*
Cco	*Coturnix coturnix*
Cel	*Caenorhabtitis elegans*
Cin	*Ciona intestinalis*
Csa	*Ciona savignyi*
Dme	*Drosophila melanogaster*
Dre/Bre	*Danio rerio*
Ebu	*Eptatretus burgeri*
Efl	*Ephydatia fluviatilis*
Fru/Tru	*Fugu rubripes*
Gga	*Gallus gallus*
Hro	*Halocynthia roretzi*
Hsa	*Homo sapiens*
Hvu	*Hydra vulgaris*
Lre	*Lethenteron reissneri*
Lst	*Lymnaea stagnalis*
Mmu	*Mus musculus*
Ola	*Oryzias latipes*
Omy	*Oncorhyncus mykiss*
Pbu	*Petrogale burdidgei*
Pma	*Petromyzon marinus*
Pwa	*Pleurodeles waltl*
Rno	*Rattus norvegicus*
Sma	*Sminthopsis macroura*
Sman	*Schisostoma mansoni*
Smax	*Scophthalmus maximus*
Spu	*Strongylocentrotus purpuratus*
Ssc	*Sus scrofa*
Tca	*Torpedo californica*
Xla	*Xenopus laevis*
Xxi	*Xiphophorus xiphidium*

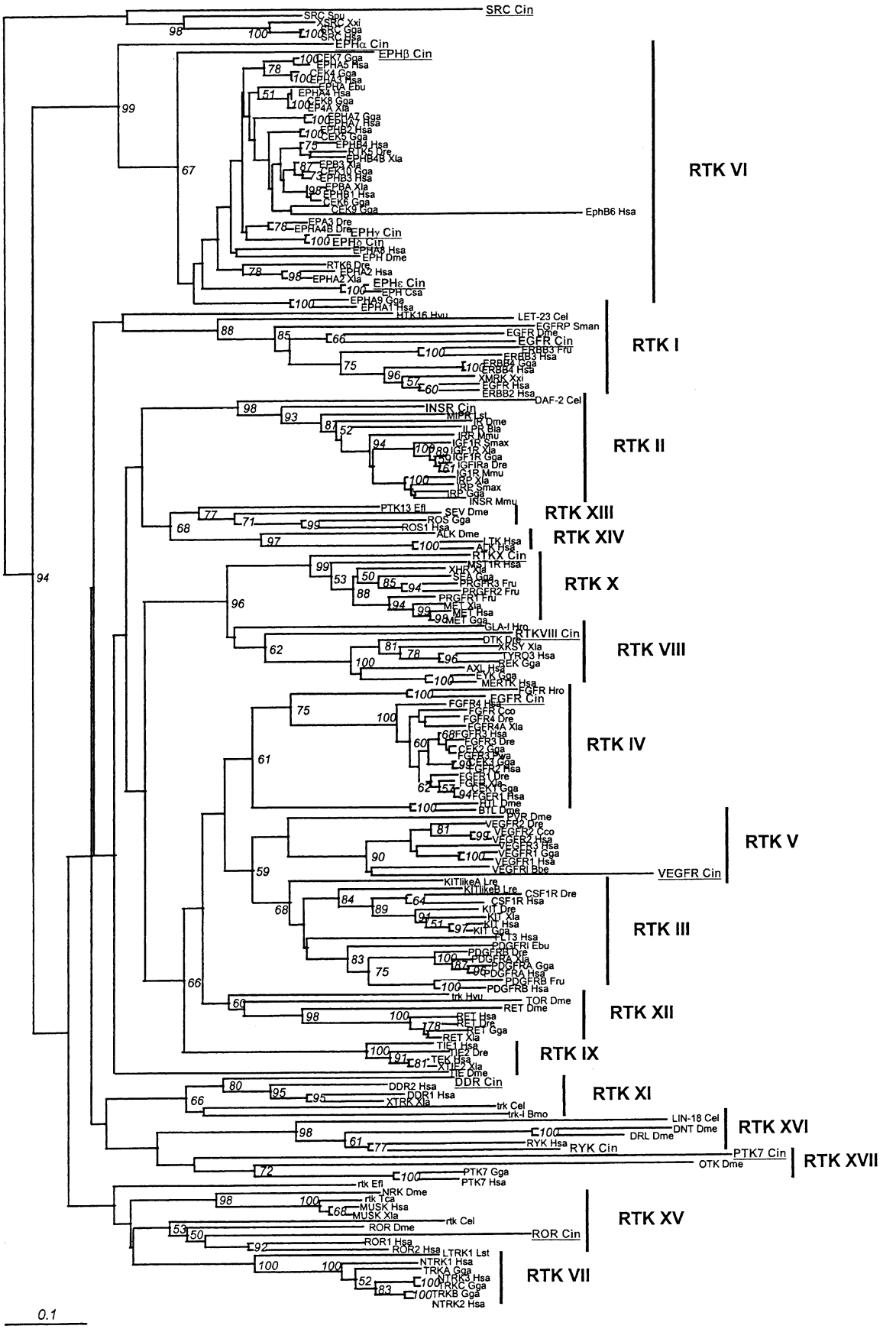
amino acid sequence of the FOX domain (about 98 amino acid residues) including 9 *C. intestinalis* FOX domains and 78 other FOX domains (87 proteins, 60 sites, 500 bootstrap replicates). The tree was rooted with the N and J FOX groups. The groups were named according to the unified FOX nomenclature (Kaestner et al. 2000). We found one *C. intestinalis* protein for each group, except groups G, H, J, and K. We identified three additional fragments, which could be putative *C. intestinalis* FOX according to Blast results, but these sequences could not be extended into a complete forkhead domain and were not included in the tree.

Figure 4 shows the neighbor-joining phylogeny of the SOX family based on the HMG (high mobility group) domain (80 domains, 71 sites, and 500 bootstrap replicates); it includes 6 *C. intestinalis* HMG domains. SOX subfamilies were named according to Bowles et al. (2000), and *Ciona* proteins according to the group they belong. The tree was rooted with the HMG domain of the CIC (capicua homolog) protein family (Lee et al. 2002); one CIC *C. intestinalis* protein was identified. We found one *C. intestinalis* protein in each major SOX group (B1, B2, C, D, E, F) except, as expected, in the SOX A group (SRY), which is known to derive from the SOX3 gene and to be specific to the mammalian chromosome Y (Katoh and Miyata 1999). In groups C, D, and E, *C. intestinalis* proteins branched as expected between fly and vertebrates. In groups F, B1, and B2, *C. intestinalis* positions were not clearly defined regarding nonvertebrate sequences, and bootstrap values were low (<50). The single-gene groups (G, H, I, and J) were not represented by *C. intestinalis* proteins.

Figure 5 shows the neighbor-joining phylogeny of ETS family based on the ETS domain (96 domains, 51 sites, 500 bootstrap replicates), including 8 *C. intestinalis* ETS domains. Groups were named according to the ETS database (http://www.biochem.missouri.edu/~martin/Etsaling.htm; Laudet et al. 1999), and *C. intestinalis* proteins according to the group name. The tree was rooted with the SPI group. Only bootstrap values over 40 are displayed. The ETS family members cluster in 13 groups (ETS, ER71, GABP, PEA3, ERG, ERF, ELK, DETS4, ELF, ESE, TEL, and YAN) (Laudet et al. 1999). Four of these groups (ER71, DETS4, TEL, YAN) are single-gene groups and are characterized by a position not well resolved in the tree. In the ERG group, there might be two *C. intestinalis* proteins. In six groups (YAN, ER71, ERF, DETS4, ELF, and TEL), we did not identify a *C. intestinalis* protein. Two other fragments of putative ETS domain were found but these domains were not complete and could not be included in the tree.

Neighbor-joining phylogeny of WNT proteins is shown in Fig. 6. It is based on the full-length sequences (77 proteins, 70 sites, 500 bootstrap replicates) and includes 4 *C. intestinalis* WNT proteins. The tree was rooted using complete WNT1 group as an outgroup. Only bootstrap values over 40 are displayed. We found *C. intestinalis* proteins in the WNT 1, 3, 5, and 8 groups. Two more WNT fragments were found and Blast results suggested that they are good candidates for the WNT2 and WNT4 orthologs. However, we could not extend these fragments to the complete protein sequence and they were not

Fig. 2. Neighbor-joining phylogeny of the RTK proteins based on the tyrosine kinase domain of the 17 subfamilies. The tree includes 15 *Ciona intestinalis* RTK kinase domains in 11 RTK families and 1 *C. intestinalis* SRC kinase domain in the SRC family used as outgroup. All named proteins are appended with the species designation (one letter for genus and two for species, for example, Hsa—*Homo sapiens*; see Table 1) and *C. intestinalis* proteins are in boldface and underlined. Group names are based on the RTKdb nomenclature (Grassot et al. 2003).


SRC Cin
SRC Spu
SRC Xxi
SRC Gga
SRC Hsa
EPHα Cin
EPHβ Cin
CEK7 Gga
EPHA5 Hsa
CEK4 Gga
EPHA3 Hsa
EPHA Ebu
EPHA4 Hsa
CEK8 Gga
EPHA4 Xla
EPHA7 Gga
EPHA7 Hsa
EPHB2 Hsa
CEK5 Gga
EPHB4 Hsa
EPHB4 Dre
EPHB4B Xla
EPHB3 Xla
CEK10 Gga
EPHB3 Hsa
EPHB Xla
EPHB1 Hsa
CEK6 Gga
CEK9 Gga
EphB6 Hsa
EPHA3 Dre
EPHA4B Dre
EPHγ Cin
EPHδ Cin
EPHA8 Hsa
EPH Dme
RTK6 Dre
EPHA2 Hsa
EPHA2 Xla
EPHε Cin
EPH Csa
EPHA9 Gga
EPHA1 Hsa
RTK VI
HTK16 Hvu
LET-23 Cel
EGFRP Sman
EGFR Dme
EGFR Cin
ERBB3 Fru
ERBB3 Hsa
ERBB4 Gga
ERBB4 Hsa
XMRK Xxi
EGFR Hsa
ERBB2 Hsa
RTK I
DAF-2 Cel
INSR Cin
MIPR Lst
IR Dme
ILPR Bla
IRR Mmu
IGF1R Smax
IGF1R Xla
IGF1R Gga
IGFIRa Dre
IGF1R Mmu
IRC Xla
IRC Smax
IRE Gga
INSR Mmu
RTK II
PTK13 Efl
SEV Dme
ROS Gga
ROS1 Hsa
RTK XIII
ALK Dme
LTK Hsa
ALK Hsa
RTK XIV
RTKX Cin
MST1R Hsa
XHR Xla
SEA Gga
PRGFM3 Fru
PRGFR2 Fru
PRGFR1 Fru
MET Xla
MET Hsa
MET Gga
RTK X
GLA-I Hro
RTKVIII Cin
DTK Dre
XRKSY Xla
TYRO3 Hsa
BEK Gga
AXL Hsa
EVK Gga
MERTK Hsa
RTK VIII
EGFR Hro
EGFR Cin
FGFR4 Hsa
FGFR Cco
FGFR4 Dre
FGFR4A Xla
FGFR3 Hsa
FGFR3 Dre
CEK2 Gga
FGFR3 Xla
FGFR1 Gga
FGFR2 Hsa
FGFR Dre
FGFR Xla
CEK1 Gga
FGFR1 Hsa
BTL Dme
RTK IV
BTL Dme
PVR Dme
VEGFR2 Dre
VEGFR2 Cco
VEGFR2 Hsa
VEGFR3 Hsa
VEGFR1 Gga
VEGFR1 Hsa
VEGFR Bge
VEGFR Cin
RTK V
KITlikeA Lre
KITlikeB Lre
CSF1R Dre
CSF1R Hsa
KIT Dre
KIT Xla
KIT Hsa
KIT Gga
FLT3 Hsa
PDGFRI Ebu
PDGFRB Dre
PDGFRA Xla
PDGFRA Gga
PDGFRA Hsa
PDGFRB Fru
PDGFRB Hsa
RTK III
trk Hvu
TOR Dme
RET Dme
RET Hsa
RET Dre
RET Gga
RET Xla
RTK XII
TIE1 Hsa
TIE2 Dre
TEK Hsa
XTIE2 Xla
TIE Dme
RTK IX
DDR Cin
DDR2 Hsa
DDR1 Hsa
XTRK Xla
RTK XI
trk Cel
trk-I Bmo
LIN-18 Cel
DNT Dme
DRL Dme
RYK Hsa
RYK Cin
RTK XVI
PTK7 Cin
OTK Dme
PTK7 Gga
PTK7 Hsa
RTK XVII
rtk Efl
NRK Dme
rtk Tca
MUSK Hsa
MUSK Xla
rtk Cel
ROR Dme
ROR Cin
ROR1 Hsa
ROR2 Hsa
RTK XV
LTRK1 Lst
NTRK1 Hsa
TRKA Gga
NTRK3 Hsa
TRKC Gga
TRKB Gga
NTRK2 Hsa
RTK VII
0.1

included in the tree. Identification of *C. intestinalis* WNTs was difficult because these proteins have a moderately conserved domain and the GenScan matrix used was a human matrix.

Discussion

Gene duplications seem to have a singular importance in evolution. Depending on their extent, they provide a suddenly enlarged repertoire of genes and possibilities from which new regulatory interactions can create novel developmental strategies and characters before most of the duplicated genes undergo detrimental substitutions and loss (Graham 2000; Shimeld and Holland 2000; Holland and Chen 2001). After a potential initial beneficial or deleterious gene dosage effect (Kondrashov et al. 2002), the evolution of gene duplicates (Page and Cotton 2002) between gene fixation or gene extinction (inactivation or loss) is thought to depend on the rapidity of acquisition of a new function (neofunctionalization) (Hughes 2002) or partitioning of an ancestral one (subfunctionalization) (Force et al. 1999; Mazet and Shimeld 2002), with cooption (True and Carroll 2002) and redundancy (Cooke et al. 1997) being results of the overlap of the new functions or territories of expression. The race between fixation and extinction exists for duplicates whatever the mechanism of their generation. Gene duplicates that can be recognized in a genome represent a biased subset of the ones that have occurred (Otto and Yong 2002). Large-scale duplications may also facilitate (Sidow 1996) punctuated equilibria (Eldredge and Gould 1972; Gould and Eldredge 1993).

The exact importance of gene and genome duplications in evolutionary innovation remains to be further delineated and the mechanism of the duplications is still debated. With respect to vertebrate evolution, one mechanism involving two rounds of large-scale duplications (known as the "2R hypothesis") has been proposed by Ohno (1970) and modified by Holland (1994). Alternatively, continuous small-scale duplications only may have led to gene expansion in the vertebrate lineage (Friedman and Hughes 2001). Comparison of the genomes of several key species provides information on the importance, mechanisms, and period of occurrence of the duplications. Most informative are comparisons made with lineages that diverged closer to the origin of vertebrates, such as ascidians and amphioxus. The marine tunicate *Ciona intestinalis* is a urochordate whose genome has been deciphered. If most of the vertebrate gene subfamilies were to be represented by a single *C. intestinalis* ortholog (or coortholog in the case of independent duplications in the urochordate lineage), it would prove that the ancestor of the urochordates separated from the vertebrate ancestor prior to the occurrence of vertebrate gene expansion. To test this prediction, we reconstituted *C. intestinalis* genes and genes families from available databases. We found that most vertebrate subfamilies are represented by a single *C. intestinalis* ortholog.

Several *Ciona* gene families have been studied recently. Ferrier and Holland (2002) have reported the genomic organization of the ParaHox genes of *C. intestinalis*. ParaHox genes, although not clustered as in vertebrates, are present in single copies, as in amphioxus. Similarly, FGF genes from *Ciona intestinalis* belong to separate subfamilies (Satou et al. 2002). To test the value of our approach we did a phylogenetic analysis of the FGF superfamily (data not shown) and found the same results as Satou et al. (2002). More recently, the Satoh laboratory has issued a series of analyses of *Ciona* superfamilies, including the five studied here (Hino et al. 2003; Satou et al. 2003; Yagi et al. 2003; Yamada et al. 2003). Their data are in perfect agreement with our results and hypotheses.

Thus, taken together with our results, the data on *Ciona* superfamilies show that vertebrate genome expansion postdated the separation of the urochordates from the other chordates. They show that expansion has occurred in all gene families studied so far.

Small-Scale Duplications Have Contributed to Vertebrate Gene Expansion

Vertebrate genome expansion occurred either through a sudden "big bang' due to large-scale duplications or continuously upon extensive small-scale duplications. In the former case, extensive gene loss should have occurred. In the latter proposition, adaptive radiation and gene fixation should explain the greater number of genes (Hughes 2002).

Independently of the problem of vertebrate ancestry, it is evident that a continuous flux of small-scale duplications occurred at all stages of metazoan evolution. Some small-scale cis-duplications have been described (Smith et al. 1999; Nusse 2001; Popovici et al. 2001b). They have created gene families, including the *HOX*, Para*HOX*, and Meta-*HOX* gene clusters. The analysis of the *Ciona intestinalis* genome is particularly helpful in determining the time of occurrence of these small-scale duplications. We have therefore taken this opportunity to speculate on the evolution of the selected superfamilies. Figure 7 illustrates examples of small-scale duplications that expanded the RTK gene superfamily.

For the RTK superfamily, the presence of a single *C. intestinalis* ortholog was found in most classes. However, no ortholog of class III was found. Receptors of classes III, IV, and V have

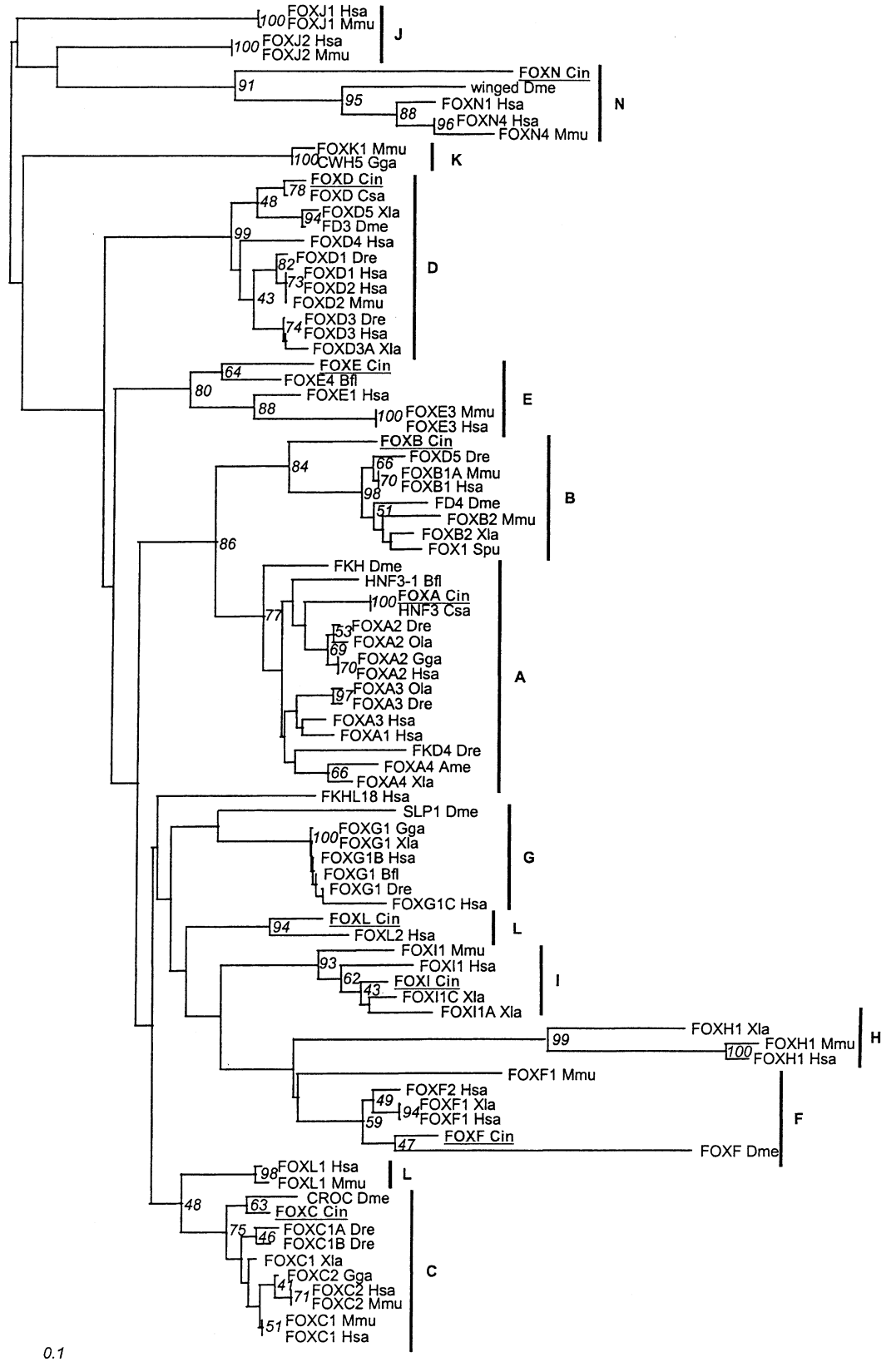


Fig. 3. Neighbor-joining phylogeny of FOX proteins based on the amino acid sequence of the FOX domain (about 98 amino acid residues) including 9 *Ciona intestinalis* FOX domains. The tree was rooted using FOX N and J groups as outgroup. For each protein we indicated the organism, the name, and the proposed name for *C. intestinalis* FOX. The groups were formed with the unified FOX nomenclature (Kaestner et al. 2000). *C. intestinalis* proteins are in boldface and underlined.

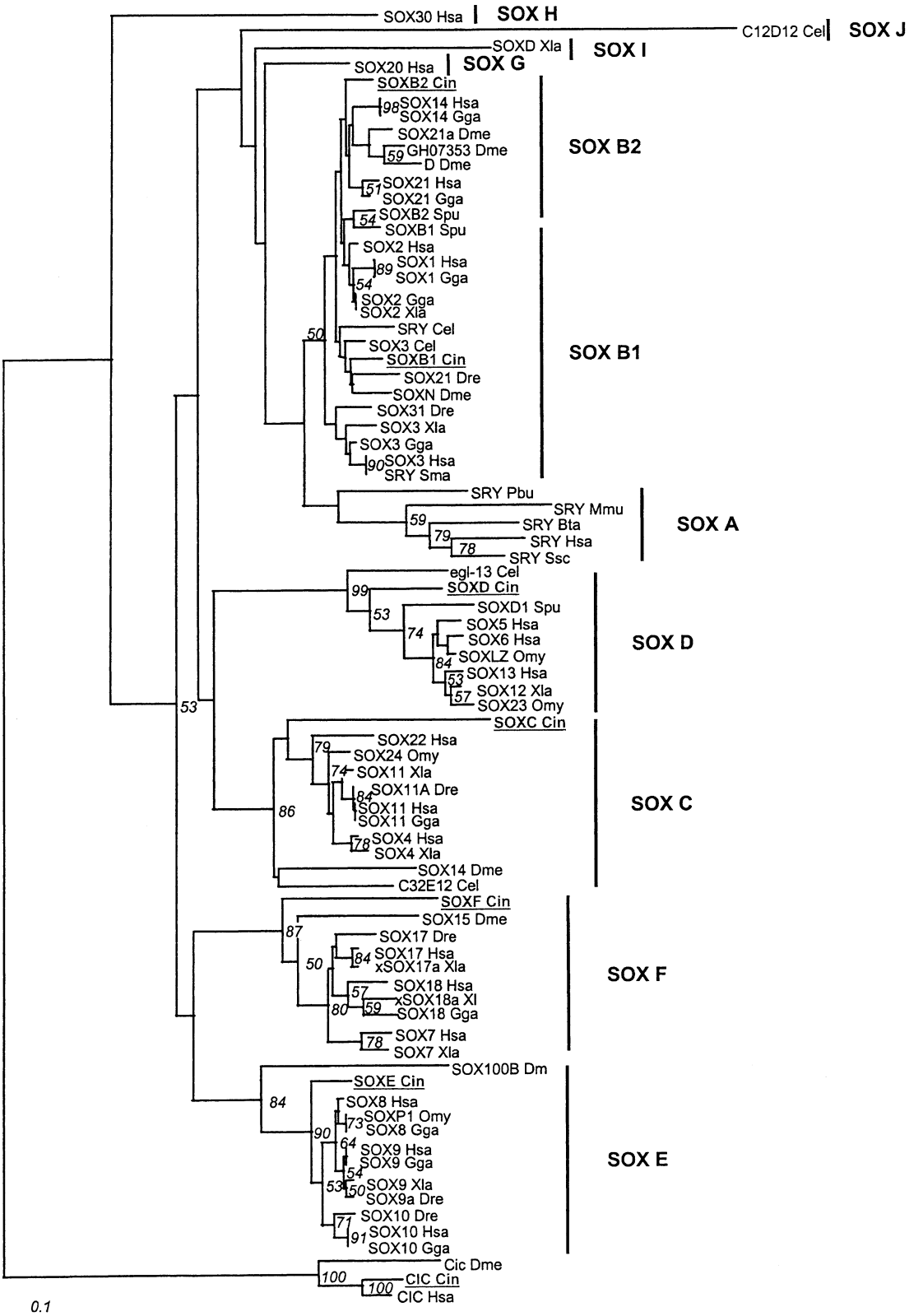


Fig. 4. Neighbor-joining phylogeny of SOX proteins based on the HMG domain sequences (80 amino acids), including six *Ciona intestinalis* SOX HMG domains and one *C. intestinalis* Capicua HMG domain in the Capicua HMG protein family which was used as outgroup. *C. intestinalis* proteins were named according to the SOX group they belong. *C. intestinalis* proteins are underlined.

three, five, and seven immunoglobulin domains, respectively, and group together in a phylogenetic tree (see Fig. 3). In humans, class III and class V genes are located in clusters in paralogous chromosomal

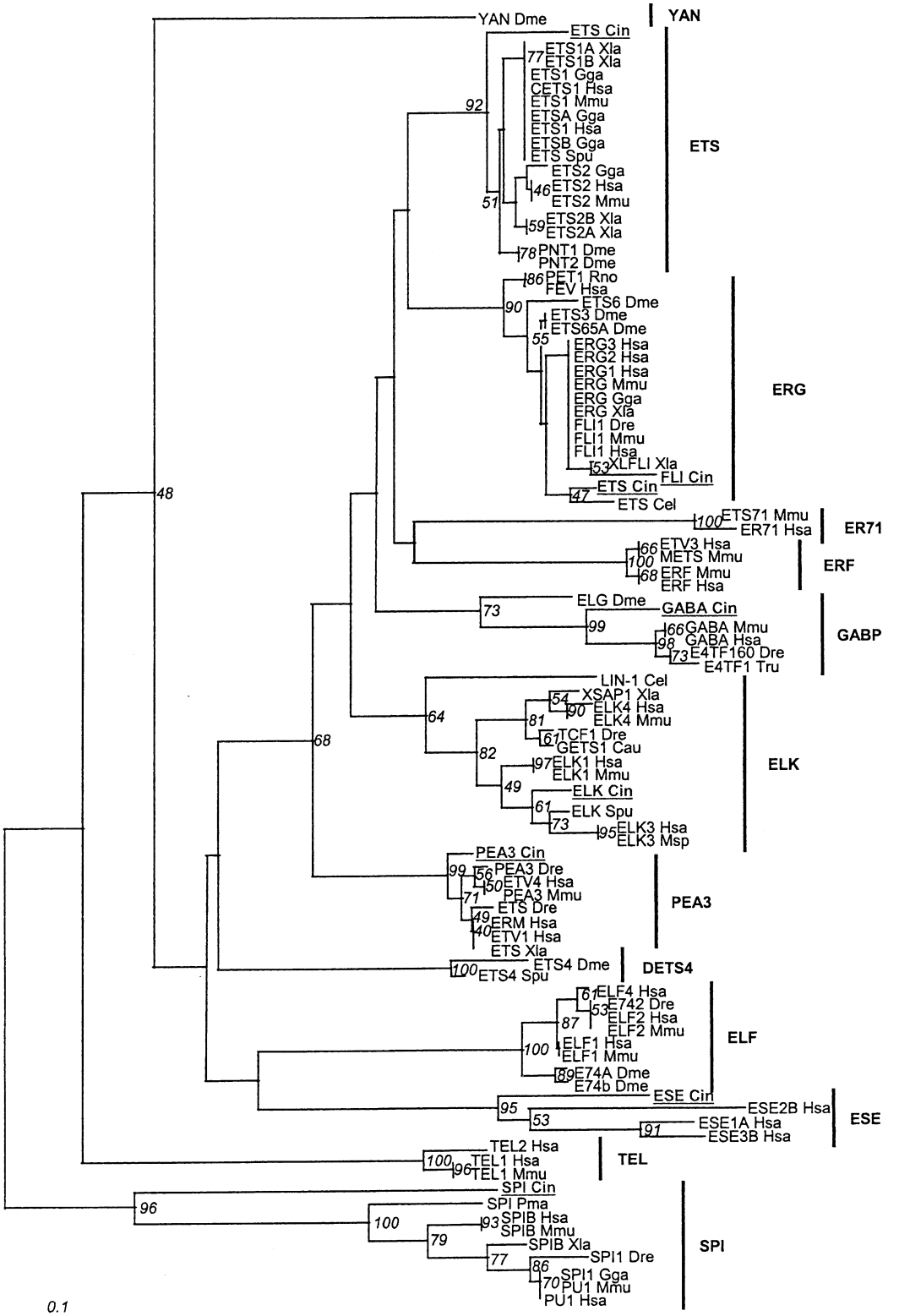


Fig. 5. Neighbor-joining phylogeny of ETS proteins based on the amino acid sequence of the ETS domain (about 100 amino acids) including eight *Ciona intestinalis* ETS domains. The tree was rooted using the SPI group as the outgroup. Groups are named according to Laudet et al. (1999). *C. intestinalis* proteins are in boldface and underlined.

regions (Rosnet et al. 1993; Agnès et al. 1997). There are several possibilities to explain the absence of *C. intestinalis* class III RTK: the duplication class III-class V might have occurred after the separation of urochordates from the chordate branch or the *C. intestinalis* class III ancestor might have been lost.

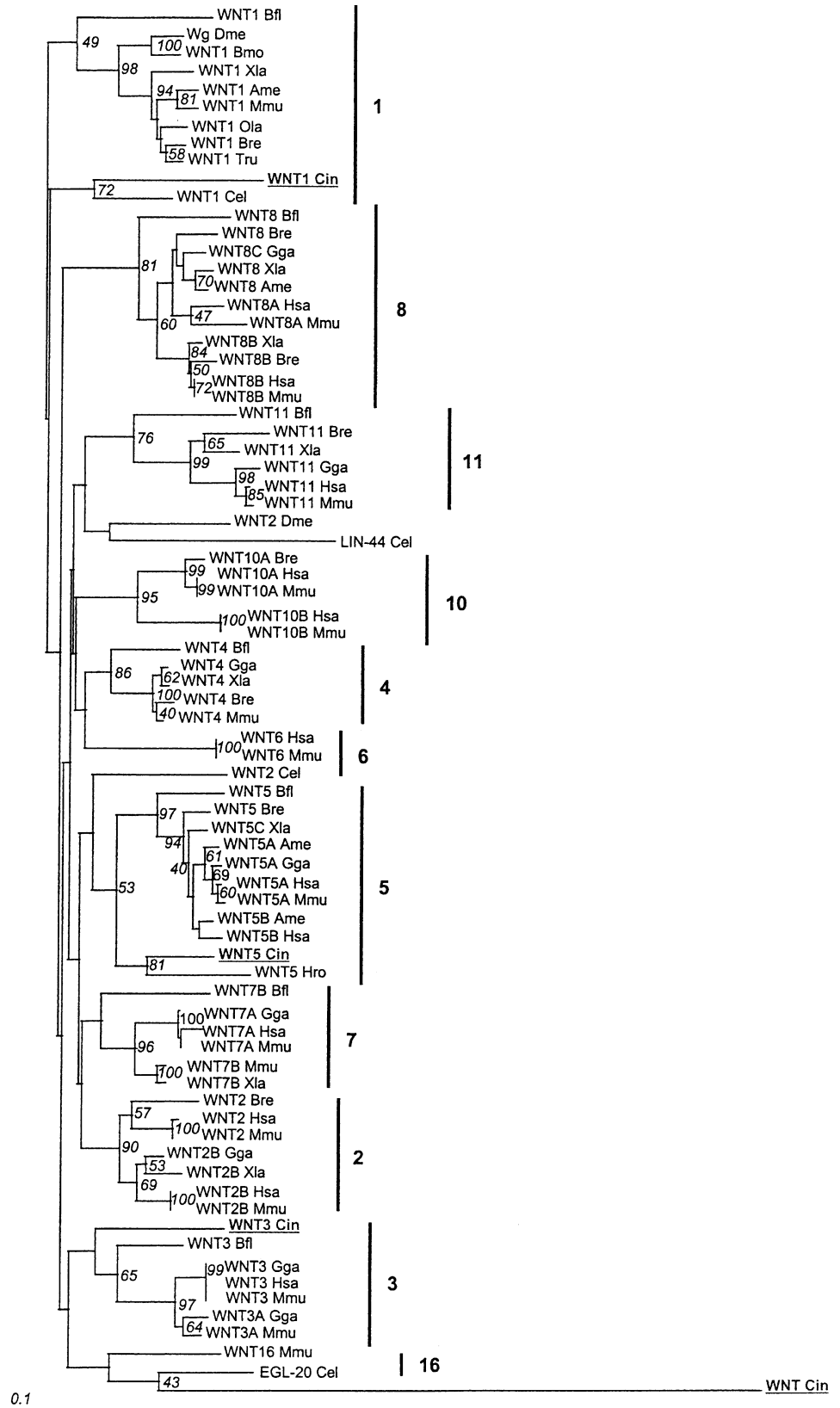


Fig. 6. Neighbor-joining phylogeny of WNT proteins based on the full-length sequences including four *Ciona intestinalis* WNT sequences. The tree was rooted using WNT1 group as outgroup. For each protein we have indicated the organism, the name, and the proposed WNT name for *Ciona intestinalis*. The groups were formed on the model of Schubert et al. (2000). *C. intestinalis* proteins are in boldface and underlined.

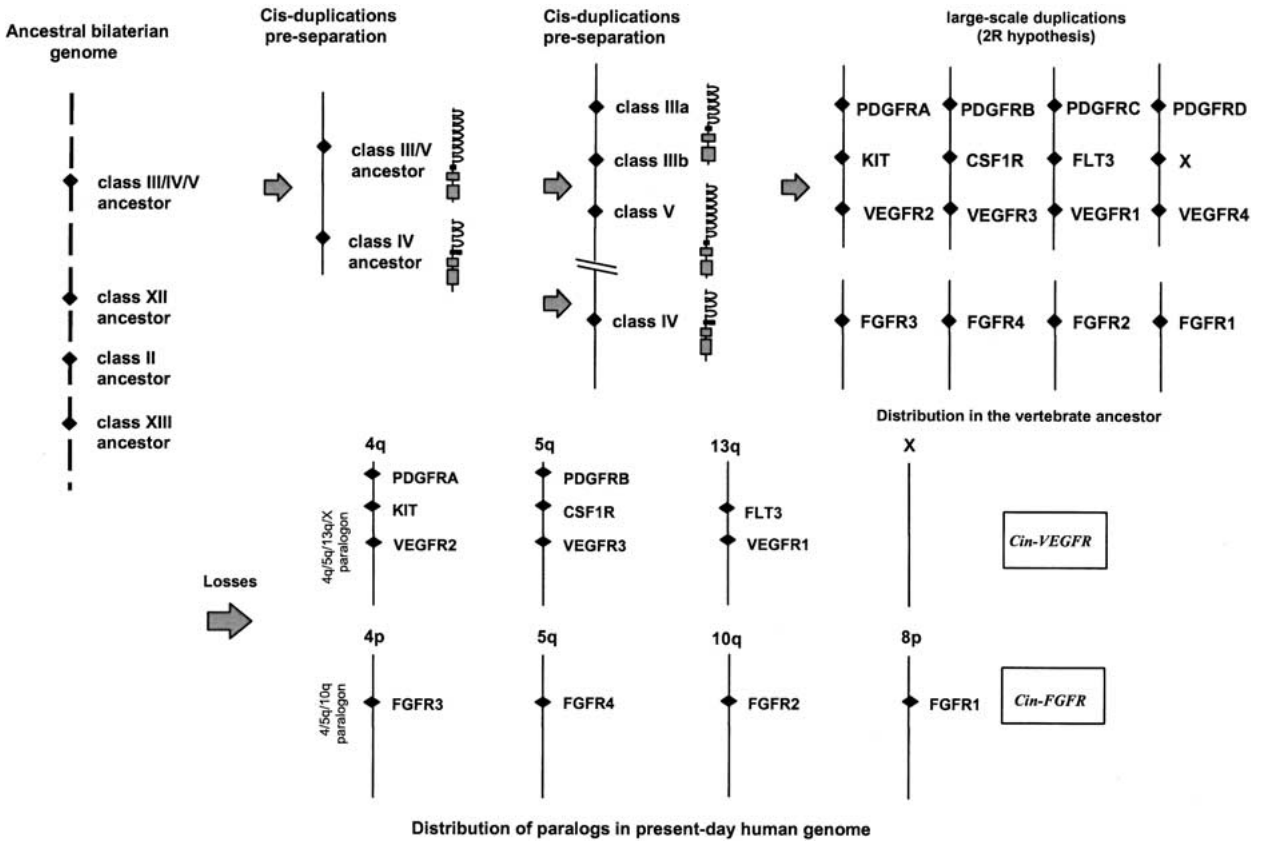


Fig. 7. Examples of gene subfamilies submitted to events of small- and large-scale duplication and loss. Tyrosine kinase receptors (RTK) of classes III (five Ig domains) and V (seven Ig domains) are shown at the right. *Drosophila melanogaster* has one such RTK, named PVR (PDGFR/VEGFR-like receptor [Ducek et al. 2001; Heino et al. 2001]); we found one in *Ciona intestinalis*; there are three class V, two class IIIa, and three class IIIb RTKs in *Homo sapiens*. The corresponding gene clusters (Popovici et al. 2001b), thought to derive from both large-scale ("2R") and small-scale duplications (prior to the separation urochordates/other chordates) are shown on the human chromosomes (paralogon 4q/5q/13q/X of Popovici et al. [2001a]; there were probably genes on a fourth location, possibly the equivalent of the X chromosome, which were lost). The absence of a *C. intestinalis* class III ortholog may be due to the loss of a class III ancestor gene in the urochordate lineage.

In the first case, this is reminiscent of what exists in *Drosophila*, where a single gene is the ortholog of both class III (Hematopoietic and PDGF receptors) and class V (VEGF receptors) genes, and has been accordingly named PVR (PDGF/VEGF receptor) (Duchek et al. 2001). However, the second scenario fits better the topology of the tree. The same reasoning may be applied to the group comprising classes VII (TRK), XI (DDR), and XV (MUSK, ROR) and encoding RTKs that all play a role in the nervous system and whose genes all map in the same paralogon (not shown). A similar analysis, applied to the FOX, SOX, ETS, and WNT superfamilies showed the importance of small-scale (cis-) duplications (not shown).

Thus, by combining phylogenetic analyses in which the *Ciona* lineage is considered and paralogy information at both the phylogenetic and the chromosome level, it is possible to reconstitute the history of gene superfamilies. The availability of sequences from the amphioxus genome and other key species should soon bring additional information to delineate the exact period of the small-scale duplications.

In conclusion, our work shows that *Ciona intestinalis* will be a good model for evolutionary analyses. So far, the comparison vertebrates versus nonvertebrates used mainly the completely sequenced *Drosophila melanogaster* and *Caenorhabditis elegans* protostomian genomes. It is now widely thought that the *C. elegans* genome is not adapted for this type of study due to a high frequency of lineage-specific duplications (Gu et al. 2002). Due to its location in the tree of life, *Ciona intestinalis* is an interesting subject for evolution analysis; we show here that it is all the more true due to the fact that massive lineage-specific duplications or losses did not seem to obscure the picture.

Acknowledgments. This work has been supported by Inserm, Institut Paoli-Calmettes, CNRS, and grants from the Ligue Nationale pour la Recherche centre le Cancer (label). M.L. is the recipient of a fellowship from the Ministère de la Recherche.

References

Abi-Rached L, Gilles A, Shiina T, Pontarotti P, Inoko H (2002) Evidence for en bloc duplication in vertebrate genomes. Nat Genet 31:100–105

Adams MD, Celniker SE, Holt RA, et al. (2001) The genome sequence of *Drosophila melanogaster*. Science 287:2185–2195

Agnès F, Toux MM, André C, Galibert F (1997) Genomic organization of the extracellular coding region of the human FGFR4 and FLT4 genes: Evolution of the genes encoding receptor tyrosine kinases with immunoglobulin-like domains. J Mol Evol 45:43–49

Altschul SF, Madden TL, Schäffer AA, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Amores A, Force A, Yan YL, Joly L, Amemiya C, Fritz A, Ho R, Langeland J, Prince V, Wang YL, Westerfield M, Ekker M, Postlethwait JM (1998) Zebrafish hox clusters and vertebrate genome evolution. Science 282:1711–1714

Aparicio S (2000) Vertebrate evolution: recent perspectives from fish. Trends Genet 16:54–56

Birnbaum D, Pébusque MJ, Imbert J, Dib A, deLapeyrière O, Coulier F (1994) Oncogenesis and genome duplication maps. Oncology Rep 1:477–480

Bowles J, Schepers G, Koopman P (2000) Phylogeny of the SOX family of developmental transcription factors based on sequence and structural indicators. Dev Biol 227:239–255

Brooke NM, Garcia-Fernandez J, Holland PW (1998) The ParaHox gene cluster is an evolutionary sister of the Hox cluster. Nature 392:920–922

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol 268:78–94

Cooke J, Nowak M, Boerlijst M, Maynard-Smith J (1997) Evolutionary origins and maintenance of redundant gene expression during metaozoan development. Trends Genet 13:360–364

Coulier F, Burtey S, Chaffanet M, Birg F, Birnbaum D (2000a) Ancestrally-duplicated paraHOX gene clusters in humans. Int J Oncol 17:439–444

Coulier F, Popovici C, Villet R, Birnbaum D (2000b) MetaHOX gene clusters. J Exp Zool 288:345–351

Dehal P et al. (2002) The draft genome of *Ciona intestinalis*: Insights into chordate and vertebrate origins. Science 298:2157–2167

de Luca di Roseto G, Bucciarelli G, Bernardi G (2002) An analysis of the genome of *Ciona intestinalis*. Gene 295:311–316

Duchek P, Somogyi K, Jekely G, Beccari S, Rorth P (2001) Guidance of cell migration by the *Drosophila* PDGF/VEGF receptor. Cell 107:17–26

Eldredge N, Gould SJ (1972) Punctuated equilibria: An alternative to phyletic gradualism. In: Schopf TJM (eds) Models of paleobiology. Freeman & Cooper, San Francisco, pp 82–115

Felsenstein J (1985) Confidence-limits on phylogenies—An approach using the bootstrap. Evolution 39:783–791

Ferrier DEK, Holland PWH (2002) *Ciona intestinalis* ParaHox genes: Evolution of Hox/ParaHox cluster integrity, developmental mode, and temporal colinearity. Mol Phylogenet Evol 24:412–417

Fitch W (1970) Distinguishing homologous from analogous proteins. Syst Zool 19:99–113

Force A, Lynch M, Pickett FB, Amores A, Van YL, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1534

Friedman R, Hughes AL (2001) Pattern and timing of gene duplication in animal genomes. Genome Res 11:1842–1847

Friedman R, Hughes AL (2003) The temporal distribution of gene duplication events in a set of highly conserved human gene families. Mol Biol Evol 20:154–161

Galtier N, Gouy M, Gautier C (1996) SEAVIEW and PHYLO_WIN: Two graphic tools for sequence alignment and molecular phylogeny. Comp Appl Biosci 12:543–545

Gerhart J, Kirschner M (1997) Cells, embryos and evolution. Blackwell Science, Malden

Gibson TJ, Spring J (1999) Evidence in favour of ancient octaploidy in the vertebrate genome. Biochem Soc Trans 28:259–264

Gogarten JP, Olendzenski L (1999) Orthologs, paralogs and genome comparisons. Curr Opin Genet Dev 9:630–636

Graham A (2000) The evolution of the vertebrates—Genes and development. Curr Opin Genet Dev 10:624–628

Grassot J, Mouchiroud G, Perrière G (2003) RTKdb: Database of receptor tyrosine kinase. Nucleic Acids Res 31:353–358

Gould SJ, Eldredge N (1993) Punctuated equilibrium comes of age. Nature 366:223–227

Gu X, Wang Y, Gu J (2002) Age distribution of both human gene families shows significant roles of both large-scale and small-scale duplications in vertebrate evolution. Nat Genet 31:205–209

Gu Z, Cavalcanti A, Chen FC, Bouman P, Li WH (2002) Extent of gene duplication in the genomes of *Drosophila*, nematode, and yeast. Mol Biol Evol 19:256–262

Heino T, Karpanen T, Wahlstrom G, Pulkinen M, Eriksson U, Alitalo K, Roos C (2001) The *Drosophila* VEGF receptor homolog is expressed in hemocytes. Mech Dev 109:69–77

Hillis DM, Bull JJ (1993) An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst Biol 42:182–192

Hino K, Satou Y, Yagi K, Satoh N (2003) A genomewide survey of developmentally relevant genes in *Ciona intestinalis* VI. Genes for Wnt, TGFbeta, hedgehog and JAK/STAT signaling pathways. Dev Genes Evol 213:264–272

Holland N, Chen J (2001) Origin and early evolution of the vertebrates: new insights from advances in molecular biology, anatomy, and palaeontology. BioEssays 23:142–151

Holland PWH, Garcia-Fernandez J, Williams NA, Sidow A (1994) Gene duplications and the origins of vertebrate development. Development Suppl:125–133

Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. Genome Res 9:868–877

Hughes AL (2002) Adaptive evolution after gene duplication. Trends Genet 18:433–434

Hughes AL, da Silva J, Friedman R (2001) Ancient genome duplications did not structure the human HOX-bearing chromosomes. Genome Res 11:771–780

Kaestner KH, Knochel W, Martinez DE (2000) A unified nomenclature for the winged helix/forkhead transcription factors. Genes Dev 14:142–146

Katoh K, Miyata T (1999) A heuristic approach of maximum likelihood method for inferring phylogenetic tree and an application to the mammalian SOX-3 origin of the testis-determining gene SRY. FEBS Lett 463:129–132

Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV (2002) Selection in the evolution of gene duplications. Genome Biol 3:0008

Koonin EV (2001) An apology for orthologs—or brave new memes. Genome Biol 2:1005.1–1005.2

Lander ES, International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Laudet V, Hanni C, Stehelin D, Duterque-Coquillaud M (1999) Molecular phylogeny of the ETS family. Oncogene 18:1351–1359

Lee CJ, Chan WI, Cheung M, Cheng YC, Appleby VJ, Orme AT, Scotting PJ (2002) CIC, a member of a novel subfamily of the HMG-box superfamily, is transiently expressed in developing granule neurons. Brain Res Mol Brain Res 106:151

Lespinet O, Wolf YI, Koonin EV, Aravind L (2002) The role of lineage-specific gene family expansion in the evolution of eukaryotes. Genome Res 12:1048–1059

Leveugle M, Prat K, Perrier N, Birnbaum D, Coulier F (2003) ParaDB: A tool for paralogy mapping in vertebrate genomes. Nucleic Acids Res 31:63–67

Li WH, Gu Z, Wang H, Nekrutenko A (2001) Evolutionary analysis of the human genome. Nature 409:847–849

Lundin LG (1993) Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. Genomics 16:1–19

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155

Makalowski W (2001) Are we polyploids? A brief history of one hypothesis. Genome Res 11:667–670

Martin A (2001) Is tetralogy true? Lack of support for the ''one-to-four'' rule. Mol Biol Evol 18:89–93

Mazet F, Shimeld SM (2002) Gene duplication and divergence in the early evolution of vertebrates. Curr Opin Genet Dev 12:393–396

McLysaght A, Hokamp K, Wolfe KH (2002) Extensive genomic duplication during early chordate evolution. Nat Genet 31:200–204

Minguillon C, Ferrier CE, Cebrian C, Garcia-Fernandez J (2002) Gene duplications in the prototypical cephalochordate amphioxus. Gene 287:121–128

Nusse R (2001) An ancient cluster of Wnt paralogues. Trends Genet 17:443

Ohno S (1970) Evolution by gene duplication. Springer Verlag, Berlin/Heidelberg/New York

Ollendorff V, Mattei MG, Fournier E, Adélaïde J, Lopez M, Rosnet O, Birnbaum D (1998) A third human CBL gene is on chromosome 19. Int J Oncol 13:1159–1161

Otto SP, Yong P (2002) The evolution of gene duplicates. Adv Genet 46:451–483

Page RD, Cotton JA (2002) Vertebrate phylogenomics: Reconciled trees and gene duplications. Pac Symp Biocomput, pp 536–547

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

Pollard S, Holland PWH (2000) Evidence for 14 homeobox gene clusters in human genome ancestry. Curr Biol 10:1059–1062

Popovici C, Roubin R, Coulier F, Pontarotti P, Birnbaum D (1999) The family of *Caenorhabditis elegans* tyrosine kinase receptors: Similarities and differences with mammalian receptors. Genome Res 9:1026–1039

Popovici C, Leveugle M, Birnbaum D, Coulier F (2001a) Homeobox gene clusters and the human paralogy map. FEBS Lett 491:237–242

Popovici C, Leveugle M, Birnbaum D, Coulier F (2001b) Coparalogy; Physical and functional clusterings in the human genome. Biochem Biophys Res Commun 288:362–370

Rosnet O, Stephenson D, Mattei MG, Marchetto S, Shibuya M, Chapman VM, Birnbaum D (1993) Close physical linkage of the FLT1 and FLT3 genes on chromosome 13 in man and chromosome 5 in mouse. Oncogene 8:173–179

Satou Y, Imai K, Satoh N (2002) Fgf genes in the basal chordate *Ciona intestinalis*. Dev Genes Evol 212:432–438

Satou Y, Sasakura Y, Yamada L, Imai KS, Satoh N, Degnan B (2003) A genomewide survey of developmentally relevant genes in *Ciona intestinalis*. V. Genes for receptor tyrosine kinase pathway and Notch signaling pathway. Dev Genes Evol 213:254–263

Schubert M, Holland LZ, Panopoulou GD, Lehrach H, Holland ND (2000) Characterization of amphioxus AmphiWnt8: Insights into the evolution of patterning of the embryonic dorsoventral axis. Evol Dev 2:85–92

Schughart K, Kappen C, Ruddle FH (1989) Duplication of large genomic regions during the evolution of vertebrate homeobox genes. Proc Natl Acad Sci USA 86:7067–7071

Shimeld SM, Holland PWH (2000) Vertebrate innovations. Proc Natl Acad Sci USA 97:4449–4452

Sidow A (1996) Gen(om)e duplications in the evolution of early vertebrates. Curr Opin Genet Dev 6:715–722

Skrabanek L, Wolfe KH (1998) Eukaryote genome duplication—Where's the evidence? Curr Opin Genet Dev 8:694–700

Smith NG, Knight R, Hurst LD (1999) Vertebrate genome evolution: A slow shuffle or a big bang? Bioessays 21:697–703

Spring J (1997) Vertebrate evolution by interspecific hybridisation—Are we polyploid? FEES Lett 400:2–8

Stach T, Turbeville JM (2002) Phylogeny of Tunicata inferred from molecular and morphological characters. Mol Phylogenet Evol 25:408–428

Taylor JS, Brinkmann H (2001) 2R or not 2R? Trends Genet 17:48–489

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815

The *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: A platform for investigating biology. Science 282:2012–2018

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The ClustalX windows interface: Flexible strategies or multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 24:4876–4882

True JR, Carroll SB (2002) Gene co-option in physiological and morphological evolution. Annu Rev Cell Dev Biol 18:53–80

Venter JC, et al. (2001) The sequence of the human genome. Science 291:1304–1351

Wagner A (2001) Birth and death of duplicated genes in completely sequenced eukaryotes. Trends Genet 17:237–239

Wang Y, Gu X (2000) Evolutionary patterns of gene families generated in the early stage of vertebrates. J Mol Evol 51:88–96

Wolfe KH (2001) Yesterday's polyploids and the mystery of diploidization. Nat Rev Genet 2:333–341

Wolfe KH, Shields DC (1997) Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387:708–713

Yagi K, Satou Y, Mazet F, Shimeld SM, Degnan B, Rokhsar D, Levine M, Kohara Y, Satoh N (2003) A genomewide survey of developmentally relevant genes in *Ciona intestinalis*. III. Genes for Fox, ETS, nuclear receptors and NFkappaB. Dev Genes Evol 213:235–244

Yamada L, Kobayashi K, Degnan B, Satoh N, Satou Y (2003) A genomewide survey of developmentally relevant genes in *Ciona intestinalis*. IV. Genes for HMG transcriptional regulators, bZip and GATA/Gli/Zic/Snail. Dev Genes Evol 213:245–253