

# Ancient Large-Scale Genome Duplications: Phylogenetic and Linkage Analyses Shed Light on Chordate Genome Evolution

Marie-Josèphe Pébusque, François Coulier, Daniel Birnbaum, and Pierre Pontarotti

Institut de Cancérologie et d'Immunologie de Marseille, INSERM U 119, Marseille, France

Paralogous genes from several families were found in four human chromosome regions (4p16, 5q33–35, 8p12–21, and 10q24–26), suggesting that their common ancestral region underwent several rounds of large-scale duplication. Searches in the EMBL databases, followed by phylogenetic analyses, showed that cognates (orthologs) of human duplicated genes can be found in other vertebrates, including bony fishes. In contrast, within each family, only one gene showing the same high degree of similarity with all the duplicated mammalian genes was found in nonvertebrates (echinoderms, insects, nematodes). This indicates that large-scale duplications occurred after the echinoderms/chordates split and before the bony vertebrate radiation. It has been suggested that two rounds of gene duplication occurred in the vertebrate lineage after the separation of Amphioxus and craniate (vertebrates + Myxini) ancestors. Before these duplications, the genes that have led to the families of paralogous genes in vertebrates must have been physically linked in the craniate ancestor. Linkage of some of these genes can be found in the *Drosophila melanogaster* and *Caenorhabditis elegans* genomes, suggesting that they were linked in the triploblast Metazoa ancestor.

## Introduction

The total number of genes per genome is assumed to have increased during vertebrate evolution through rounds of large-scale duplications (Ohno 1970; Lundin 1993). This assumption is based on (1) the estimation of the gene number found in species from vertebrate and nonvertebrate phyla and (2) the existence and mapping of paralogous genes present in vertebrate species.

Estimation of the number of genes in a chromosomal segment or a complete genome by direct sequencing is accurate. Several prokaryote genomes have been sequenced and, to date, the complete sequence of one eukaryote, *Saccharomyces cerevisiae*, is available (Goffeau et al. 1996). Extensive sequence data are also available for one Metazoa: the nematode worm *Caenorhabditis elegans* (Blumenthal and Spieth 1996). In yeast, the gene number has been estimated to be around 6,000, whereas in *C. elegans*, it has been estimated to be around 14,000. For other Metazoa, the number of genes has been estimated in various ways. In the case of *Drosophila melanogaster* (Protostomia arthropod), the number of genes may be around 12,000, a calculation based on (1) region sampling, (2) comparison of messenger RNA lengths with their genomic counterparts, and (3) reassociation analyses (Miklos and Rubin 1996). In the case of the mouse and human genomes, the estimation, based on CpG island counting and large-scale sequencing of human cDNAs, reaches 60,000–80,000 genes (Antequera and Bird 1993; Fields et al. 1994). The same value is estimated for puffer fish *Fugu rubripes* (Brenner et al. 1993). The figure could be similar for all bony vertebrates. Estimation of the gene number for other phyla, for example, sea urchin *Strongylocentrotus purpuratus*, has often relied on reassociation analyses.

Key words: *Caenorhabditis elegans*, *Drosophila melanogaster*, evolution, gene duplication, genome, *Homo sapiens*.

Address for correspondence and reprints: Daniel Birnbaum, Institut de Cancérologie et d'Immunologie de Marseille, Université de la Méditerranée, INSERM U 119, 27 Bd Leï Roure 13009 Marseille, France. E-mail: birnbaum@marseille.inserm.fr.

*Mol. Biol. Evol.* 15(9):1145–1159. 1998

© 1998 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

These analyses suggest that this species has fewer than 25,000 genes (Galau et al. 1976). This is the only available estimate for a deuterostome which is not a bony vertebrate.

It seems, therefore, that the number of genes per genome was doubled after the separation of Protostomia and Deuterostomia and multiplied by three or four in the chordate phylum after the separation from echinoderms (fig. 1). Such an increase in gene number could be due to tetraploidization events, as first proposed by Ohno (1970).

More information on gene duplication during chordate evolution may come from comparison of the number of members present in gene families in different species. A good example of this type of study is that of the *HOX* gene clusters (Schughart, Kappen, and Ruddle 1989). By using this kind of observation together with the estimation of the number of genes per genome, several authors (Holland et al. 1994; Sidow 1996) have suggested that large-scale duplications occurred during early chordate evolution after the cephalochordate/craniata ancestors diverged (fig. 1). Indeed, several genes are found in one copy per haploid genome in Amphioxus, but are represented by families of related genes in vertebrates. For example, one cluster of *HOX* genes is found in Amphioxus, whereas four are commonly found in vertebrates (Garcia-Fernandez and Holland 1994; Holland and Garcia-Fernandez 1996; Bailey et al. 1997). Chan, Cao, and Steiner (1990) reported that Amphioxus has a single insulin-like gene (*ILP*) which resembles three genes in the mammalian genome (Insulin, *IGF1*, *IGF2*). The deduced protein sequence of *ILP* shares equal identity with each of the three human proteins. A likely explanation is that Amphioxus has a single cluster of genes representative of the *HOX* complex and one gene representative of the insulin/insulin-like growth factor genes, and that *HOX* complex and insulin gene duplications occurred in the craniata lineage after the divergence of Amphioxus and craniates (fig. 1). At least one of the duplication events occurred early in the craniata lineage, since both nonvertebrate craniata hagfish

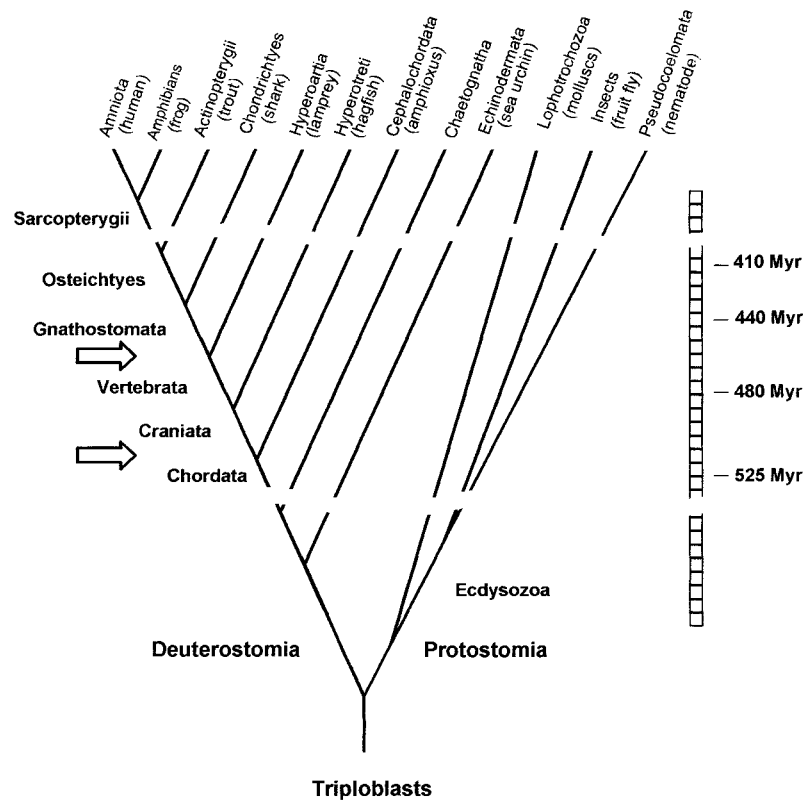


FIG. 1.—Phylogenetic tree of the Metazoa lineage according to the “tree of life” (<http://phylogeny.arizona.edu/tree/phylogeny.html>) except for the branching of Protostoma, which is according to Aguinaldo et al. (1997). Phyla are indicated at the top of the figure, as well as at specific nodes. A tentative timescale (in Myr) is shown to the right. Wavy lanes indicate a break in the timescale. Arrows indicate rounds of gene duplications that have occurred in the chordate lineage as hypothesized by Holland et al. (1994).

and the jawless vertebrate lamprey have an insulin gene and at least one *IGF* gene (Nagamatsu et al. 1991).

Potential examples of duplication found in vertebrates may be taken from genes in the region containing the major histocompatibility complex (Kandil et al. 1996). *LMP7* and *X* subunits of the proteasome complex arose from a duplication that occurred in the jawed vertebrate lineage. The two genes are present in all the jawed vertebrates (Gnathostomata) tested so far, whereas only one gene, *LMP/X*, is present in the hagfish and the lamprey. This suggests the occurrence of a second event of duplication in the jawed vertebrate lineage (fig. 1).

Paralogous chromosomal regions that are found in humans and in mice could be the remnants of these large-scale duplication or tetraploidization events proposed to explain the increase in gene number (Lundin 1993). Unfortunately, in most studies reported so far, the structural relationship between the duplicated genes found in paralogous regions is not firmly established. While it may be clear that the duplicated genes belong to the same multigenic family, it is more difficult to determine whether they are the result of a large-scale duplication. To approach the answer to such a question, it is necessary to do a phylogenetic analysis.

This type of analysis has been done by Katsanis, Fitzgibbon, and Fisher (1996) and Kasahara et al. (1996, 1997). They described gene families whose members map within the same regions of human chromosomes 1,

6, 9, and 19, respectively. The phylogenetic analysis, done for two sets of paralogous genes, supported a large-scale duplication. Unfortunately, the timing of this duplication was not estimated, and it is therefore not possible to associate the data with the hypothesized duplications that occurred during craniate evolution.

To obtain the clearest possible picture of large-scale duplication events, it is necessary to start with informative sets of genes and to simultaneously perform phylogenetic analyses and comparative genome analyses. In the present study, we hypothesize that an ancestral region has undergone duplications and is now in four copies in human chromosomes 4, 5, 8, and 10. We started from chromosome region 8p12–21, for which we have detailed information (Adélaïde et al. 1998). Each 8p12–21 gene was used to search for paralogs through bibliographic and sequence database analyses. Seven gene families were identified. The phylogenetic relationships of members of all families were established, and a mapping investigation was done, leading to the selection of four chromosomal regions which may provide information on duplication events that have led to the present-day chordate genome.

## Materials and Methods

### Databases

To assess sequence similarity, searches were performed in the EMBL and GenBank databases. Mapping

information on human genes was obtained mostly using the Online Mendelian Inheritance in Man (OMIM) from the center for Medical Genetics, John Hopkins University (Baltimore, Md.) via the National Center for Biotechnology Information, National Library of Medicine (Bethesda, Md.) (World Wide Web URL: <http://www.ncbi.nlm.nih.gov/Omim/>). The information for genes and gene products of two nonvertebrate species, *D. melanogaster* and *C. elegans*, is available through the Fly (URL: <http://astorg.u-strasbg.fr:7081/>) and ACeDB (URL: <http://probe.nal.usda.gov:8300/other/index.html>) databases, respectively.

### Sequence Alignments

Sequence similarity searches were carried out using the TBLASTX algorithm (Altschul et al. 1990, 1997). The seven gene families studied encode plasminogen activators (PAs), ankyrins (ANKs), fibroblast growth factor receptors (FGFRs), adrenergic receptors (ADRs), early growth response proteins (EGRs), vesicular amine transporters (VMATs), and lipases (LPLs). In tables 1–3 are listed the accession numbers for gene sequences in Deuterostomia, *D. melanogaster*, and *C. elegans* species, respectively.

The protein sequences were aligned using the CLUSTAL W (Thompson, Higgins, and Gibson 1994) or PILEUP (Wisconsin Package, version 9.1, Genetics Computer Group [GCG], Madison, Wis.) programs, followed by manual editing. Most of the alignments were confirmed by searches in the Pfam database (<http://www.sanger.ac.uk/Pfam/>), which is a collection of protein family alignments semiautomatically constructed using hidden Markov models (HMMs).

Additional alignments were performed after partitioning protein sequences in distinct domains: kringle domain for the PA family, part or whole spectrin domain for the ANK family, and C-terminal domain for the VMAT family.

### Phylogenetic Analyses

Phylogenetic trees were inferred using neighbor-joining algorithms (Saitou and Nei 1987) of the CLUSTAL W phylogenetic package (Thompson, Higgins, and Gibson 1994) to summarize the evolutionary relationships among sequences. The results were analyzed using the bootstrap method to provide confidence levels for the tree topology (Felsenstein 1985). Reliability of clustering patterns in phylogenetic trees was assessed by bootstrapping, which involves repeatedly reconstructing the tree from a pseudosample of data (Felsenstein 1985). For each tree, 1,000 bootstrap samples were used. Trees are plotted as rectangular cladograms using Treeview (Page 1996).

Species abbreviations are as follows: Aan: *Anguilla anguilla* (European eel); Bta: *Bos taurus* (cow); Cau: *Carassius auratus* (goldfish); Cel: *C. elegans* (nematode worm); Dme: *D. melanogaster* (fruit fly); Dre: *Danio rerio* (zebrafish); Fru: *F. rubripes* (puffer fish); Gga: *Gallus gallus* (chicken); Hsa: *Homo sapiens* (human); Mga: *Meleagris gallopavo* (turkey); Mmu: *Mus musculus* (mouse); Ola: *Oryzias latipes* (japanese medaka

fish); Omo: *Oreochromis mossambicus* (tilapia); Pma: *Petromyzon marinus* (lamprey); Pwa: *Pleurodeles waltlii* (iberian ribbed newt); Rno: *Rattus norvegicus* (rat); Spu: *S. purpuratus* (sea urchin); Toc: *Torpedo ocellata* (marine ray torpedo); Xla: *Xenopus laevis* (african clawed frog).

The phylogenetic trees for the ADR (Fryxell 1995), FGFR (Coulier et al. 1997), and LPL (Hide, Chan, and Li 1992) families have been published already. The new members of these families found since the publication of these reports were compared by BLAST (Altschul et al. 1990, 1997) and FASTA analyses (Pearson 1990).

### Definitions

Throughout the study, the following definitions were 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). 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. 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 either polyploidization or megabase duplication (in *cis* or in *trans*). We used the term “local duplication” for a duplication in *cis* on a smaller scale, including tandem duplication.

Triploblastic Metazoa (animals developed from three embryonic layers) are divided into two groups: Protostomia (platyhelminthes, Annelida, Arthropoda, Mollusca) and Deuterostomia (echinoderms, chordates), which are characterized by different fates of the initial opening of the primitive digestive tract. In the former, the blastopore from the gastrula eventually develops into the mouth, whereas it develops into the anus in the latter. Invertebrata is not a natural monophyletic group; the term “nonvertebrate” is used instead.

### Results and Discussion

The 8p12–21 chromosome region contains genes that belong to multigenic families. They include (from centromere to telomere): *PLAT*, plasminogen activator tissue-type; *ANK1*, ankyrin 1; *FGFR1*, fibroblast growth factor receptor 1; *ADRB3*, adrenergic  $\beta$ 3 receptor; *ADRA1C*,  $\alpha$ -1c adrenergic receptor subtype; *EGR3*, intermediate-early transcription factor 3; *VMAT1*, vesicular monoamine transporter; and *LPL*, lipoprotein lipase.

Members of these families were found in the 4p16, 5q33–35, 8p12–21, and 10q21–26 chromosomal regions of the human genome. Searches of the EMBL databases, followed by phylogenetic analyses, showed that the same sets of paralogous genes can be found in nonmammalian vertebrates, including bony fishes, but not in nonvertebrates. These observations indicate that large-scale duplications occurred in the craniate lineage before

**Table 1**  
**Information on Studied Deuterostomian Genes**

Gene Name(s)	Full Name	Accession Number	Species	Operational Name <sup>a</sup>	Chromosomal Localization <sup>b</sup>
<i>PLAT/UROT</i> . . . . .	Plasminogen activator tissue type	K03021	Human	Hsa-PLAT	8p12
	Plasminogen activator tissue type	U31988	Chicken	Gga-PLAT	
<i>PLAU/UROK</i> . . . . .	Plasminogen activator urokinase	K03027	Human	Hsa-PLAU	10q24
	Plasminogen activator urokinase	L03546	Bovine	Bta-PLAU	
	Plasminogen activator urokinase	J05187	Chicken	Gga-PLAU	
<i>PLMN</i> . . . . .	Plasminogen	M74220	Human	Hsa-PLMN	
	Plasminogen	X79402	Bovine	Bta-PLMN	
	Plasminogen	J04766	Mouse	Mmu-PLMN	
	Plasminogen	P33574	Lamprey	Pma-PLMN	
<i>HGFA</i> . . . . .	Hepatocyte growth factor activator	D14012	Human	Hsa-HGFA	
<i>HGF</i> . . . . .	Hepatocyte growth factor	X16323	Human	Hsa-HGF	
<i>HGFL</i> . . . . .	Hepatocyte growth factor	M60718	Human	Hsa-HGFL	
<i>FXII</i> . . . . .	Coagulation factor XII	M11723	Human	Hsa-FXII	5q33–qter
<i>THRB</i> . . . . .	Prothrombin precursor	M17262	Human	Hsa-THRB	
	Prothrombin precursor	J00041	Bovine	Bta-THRB	
	Prothrombin precursor	X52835	Rat	Rno-THRB	
<i>APOA</i> . . . . .	Apolipoprotein a	X06290	Human	Hsa-APOA	
<i>ANK1</i> . . . . .	Erythrocyte ankyrin	M28880	Human	Hsa-ANK1	8p12
	Ankyrin 1	X69063	Mouse	Mmu-ANK1	
	Ankyrin 1	U50444	Chicken	Gga-ANK1	
<i>ANK2</i> . . . . .	Brain ankyrin	X56958	Human	Hsa-ANK2	4q25–27
	Ankyrin 2	U50445	Chicken	Gga-ANK2	
<i>ANK3</i> . . . . .	Node of ranvier ankyrin	U13616	Human	Hsa-ANK3	10q21
	Ankyrin 3	U89275	Mouse	Mmu-ANK3	
	Ankyrin 3	U50446	Chicken	Gga-ANK3	
<i>FGFR1</i> . . . . .	Fibroblast growth factor receptor 1	P11362	Human		8p11–8p12
	Fibroblast growth factor receptor 1	M61687	<i>Xenopus</i>		
	Fibroblast growth factor receptor 1	D13550	Medaka fish		
	Fibroblast growth factor receptor 1	X59380	<i>Pleurodeles</i>		
<i>FGFR2</i> . . . . .	Fibroblast growth factor receptor 2	P21802	Human		10q26
	Fibroblast growth factor receptor 2	X74332	<i>Pleurodeles</i>		
	Fibroblast growth factor receptor 2	X65943	<i>Xenopus</i>		
	Fibroblast growth factor receptor 2	D13551	Medaka fish		
<i>FGFR3</i> . . . . .	Fibroblast growth factor receptor 3	P22607	Human		4p16.3
	Fibroblast growth factor receptor 3	X75603	<i>Pleurodeles</i>		
	Fibroblast growth factor receptor 3	D13552	Medaka fish		
<i>FGFR4</i> . . . . .	Fibroblast growth factor receptor 4	P22455	Human		5q35.1
	Fibroblast growth factor receptor 4	X65059	<i>Pleurodeles</i>		
	Fibroblast growth factor receptor 4	D31761	<i>Xenopus</i>		
	Fibroblast growth factor receptor 4	D13553	Medaka fish		
<i>FGFR</i> . . . . .	Fibroblast growth factor receptor	U17164	Urchin		
<i>ADRB1</i> . . . . .	Adrenergic receptor B1	J03019	Human		10q25
<i>ADRB2</i> . . . . .	Adrenergic receptor B2	J02960	Human		5q33–q35
	Adrenergic receptor B2	M14379	Turkey		
<i>ADRB3</i> . . . . .	Adrenergic receptor B3	M29932	Human		8p12
<i>ADRA1A</i> . . . . .	Adrenergic receptor A1A	L31772	Human		20
<i>ADRA1B</i> . . . . .	Adrenergic receptor A1B	L31773	Human		5q23–q35
<i>ADRA1C</i> . . . . .	Adrenergic receptor A1C	U08994	Human		8p21
<i>ADRA2A</i> . . . . .	Adrenergic receptor A2A	M18415	Human		10q25
<i>ADRA2B</i> . . . . .	Adrenergic receptor A2B	M34041	Human		2
<i>ADRA2C</i> . . . . .	Adrenergic receptor A2C	J03853	Human		4p16.3
<i>D1/D1A</i> . . . . .	Dopamine receptor 1/1A	S58541	Human		5q35.1
<i>D1A</i> . . . . .	Dopamine receptor 1A	L36877	Chicken		
	Dopamine receptor 1A	U07863	<i>Xenopus</i>		
	Dopamine receptor 1A	L08602	Goldfish		
<i>D1A1</i> . . . . .	Dopamine receptor 1A1	U62918	Eel		
<i>D1A</i> . . . . .	Dopamine receptor 1A	X80174	<i>Fugu</i>		
<i>D1A2</i> . . . . .	Dopamine receptor 1A2	U62919	Eel		
<i>D5/D1B</i> . . . . .	Dopamine receptor 5/1B	M67439	Human		4p16
<i>D1B</i> . . . . .	Dopamine receptor 1B	L36878	<i>Chicken</i>		
	Dopamine receptor 1B	U07864	<i>Xenopus</i>		
	Dopamine receptor 1B	U62920	Eel		
<i>D1C</i> . . . . .	Dopamine receptor 1C	U07865	<i>Xenopus</i>		
	Dopamine receptor 1C	U62921	Eel		
	Dopamine receptor 1C	X81969	Tilapia		
	Dopamine receptor 1C	X80177	<i>Fugu</i>		
<i>D1D</i> . . . . .	Dopamine receptor 1D	L36879	Chicken		

**Table 1**  
Continued

Gene Name(s)	Full Name	Accession Number	Species	Operational Name <sup>a</sup>	Chromosomal Localization <sup>b</sup>
<i>EGR1/KROX24</i> . . . . .	Early growth response 1	M80583	Human	Hsa-EGR1	5q31.1
	Early growth response 1	M20157	Mouse	Mmu-EGR1	
	Early growth response 1	U12895	Zebra fish	Dre-EGR1	
<i>EGR2/KROX20</i> . . . . .	Early growth response 2	J04076	Human	Hsa-EGR2	10q21.1
	Early growth response 2	S56884	<i>Xenopus</i>	Xla-EGR2	
	Early growth response 2	X70322	Zebra fish	Dre-EGR2	
<i>EGR3</i> . . . . .	Early growth response 3	S40832	Human	Hsa-EGR3	8p21
	Early growth response 3	P43301	Rat	Rno-EGR3	
<i>EGR4</i> . . . . .	Zinc-finger gene pAT133	X69438	Human	Hsa-EGR4	2p13
	Early growth response 4	M92433	Rat	Rno-EGR4	
<i>VMAT1/VATI</i> . . . . .	Adrenal vesicular amine transporter	U39905	Human	Hsa-VMAT1	8p21.3
<i>VMAT1</i> . . . . .	Adrenal vesicular amine transporter	M97380	Rat	Rno-VMAT1	
<i>VMAT2/SVAT</i> . . . . .	Synaptic vesicular amine transporter	L09118	Human	Hsa-VMAT2	10q25
<i>VMAT2</i> . . . . .	Synaptic amine transporter	M97381	Rat	Rno-VMAT2	
<i>VACHT</i> . . . . .	Vesicular acetylcholine transporter	U10554	Human	Hsa-VACHT	10q11.2
	Vesicular acetylcholine transporter	U05339	Ray	Toc-VACHT	
<i>LPL</i> . . . . .	Lipoprotein lipase	M15856	Human		8p22
	Lipoprotein lipase	X14670	Chicken		
	Lipoprotein lipase	U57656	Zebra fish		
	Hepatic lipase	D83548	Human		
<i>HL</i> . . . . .	Hepatic lipase	D83548	Human		15q21–23
<i>PNLIP</i> . . . . .	Pancreatic lipase	M93285	Human		10q24–25
<i>PNLIP</i> -like 1 . . . . .	Pancreatic lipase-like 1	M93283	Human		
<i>PNLIP</i> -like 2 . . . . .	Pancreatic lipase-like 2	M93284	Human		

<sup>a</sup> Names used for phylogenetic tree construction (see figs. 2, 3, 5, and 6).<sup>b</sup> For humans, when known.

the bony vertebrate radiation but after the echinoderm/chordate split. Each gene family is described successively in the following paragraphs.

#### PA Family

PLAT (plasminogen activator tissue-type), PLAU (plasminogen activator urokinase-type), HGFA (hepatocyte growth factor activator) and FXII (coagulation factor XII) are a family of serine proteases (see table 1). They are mosaic proteins composed of: (1) an EGF domain, a sequence about 30–40 amino acids long of unclear function (Davis 1990); (2) a kringle domain, which is a triple-looped disulfide cross-linked domain found in a variable number of copies in some serine proteases and plasma proteins and thought to play a role in binding mediators (Castellino and Beals 1987); (3) a trypsin domain with catalytic activity of serine proteases (Brenner 1988). Phylogenetic trees were drawn with the kringle (fig. 2) or the trypsin (not shown) domain, and similar topologies were obtained using each domain. PLAT, PLAU, FXII or HGFA are more similar one to the other than they are to other members of the kringle or trypsin domain family.

The *PLAT*, *PLAU*, and *FXII* genes map to chromosome regions 8p12, 10q24, and 5q32–34, respectively. The *HGFA* gene has not been localized so far. Potential orthologs of *PLAT* and *PLAU* are found in chickens (Leslie et al. 1990; table 1 and fig. 2). The duplication events that gave rise to the PLAT/PLAU/HGFA/FXII family therefore seem to have occurred before the amniote last common ancestor.

#### ANK Family

Ankyrins are membrane proteins that connect integral proteins with the spectrin-based membrane skel-

eton. They are also mosaic proteins formed by several so-called ankyrin repeats which bind the erythroid anion exchanger and tubulin, a 62-kDa spectrin-binding domain, and a 55-kDa regulatory domain that regulates the binding of ankyrin to spectrin and band 3 protein (Lux, John, and Bennett 1990). Three ankyrins have been described for mammals: erythrocytic ankyrin ANK1, brain ankyrin ANK2, and node-of-Ranvier ankyrin ANK3 (table 1). Alignments were based on the spectrin-binding domain. Human ANK1 shares 65% identity with ANK2 and ANK3. *ANK1*, *ANK2*, and *ANK3* genes map to chromosome regions 8p12, 4q25–27, and 10q21, respectively. The evolution of this family is more complex than that of the plasminogen activator family. One chromosome pericentric inversion (affecting chromosome 4) following the duplication events may explain the localization of the genes.

The three *ANK* genes have direct orthologs in chickens (table 1). The tree (fig. 3) was drawn from the available chicken sequences corresponding to a part of the spectrin domain (amino acid positions 118–225 of the human *ANK1* gene). The duplications therefore occurred before the amniote radiation. Ankyrin-like genes showing orthologous relationships with the *ANK* vertebrate family are also found in *D. melanogaster* (table 2) and *C. elegans* (table 3). Therefore, using *ANK* genes as markers, it appears that the duplication events occurred after the Protostomia/Deuterostomia split (see fig. 3).

#### FGFR Family

The FGFR family belongs to the superfamily of tyrosine kinase receptors. In humans, FGFRs are encoded by four distinct genes, *FGFR1*, *FGFR2*, *FGFR3*, and *FGFR4*. All four genes map to the paralogous regions

**Table 2**  
**Information on *Drosophila melanogaster* Genes**

Gene Name	Full Name	Accession Number	Vertebrate Ortholog	Chromosomal Localization	Flybase ID
<i>Yp1</i> . . . . .	Yolk protein 1	X01524/P02843	LPL	1-9A2-9A5	FBgn0004045
<i>Yp2</i> . . . . .	Yolk protein 2	X01524/P02844	LPL	1-9A2-9A5	FBgn0005391
<i>Yp3</i> . . . . .	Yolk protein 3	X04754/P06607	LPL	1-12B1-12C8	FBgn0004047
<i>5-HT1B</i> . . . . .	Serotonin receptor 1B	Z11490/P28286	ADR	2-56A1-56B7B	FBgn0004572
<i>5-HT1A</i> . . . . .	Serotonin receptor 1A	Z11489/P28285	ADR	2-56A1-56B7	FBgn0004168
<i>Ocr1</i> . . . . .	Octopamine receptor-like	Not available	ADR	2-56A3-56B5	FBgn0011268
<i>bt1</i> . . . . .	Breathless	X57746/Q09147	FGFR	3-70D2-70D6	FBgn0005592
<i>5-HT2</i> . . . . .	Serotonin receptor 2	X85407/Q24511	ADR	3-82C4-82C5	FBgn0013743
<i>htl</i> . . . . .	Heartless	X74030/Q07407	FGFR	3-90D1-90D6	FBgn0010389
<i>sr</i> . . . . .	Stripe	U42403/Q24162	EGR	3-90E1-90E1	FBgn0003499
<i>VACHT</i> . . . . .	Vesicular acetylcholine trans	Not available	VACHT	3-91C7-97D2	FBgn0015323
<i>Ocr</i> . . . . .	Octopamine receptor	X54794/P22270	ADR	3-99A10-99B1	FBgn0004514
<i>5-HT7</i> . . . . .	Serotonin receptor 7	M55533/P20905	ADR	3-100A1-100A7	FBgn0004573
<i>ANK</i> . . . . .	Ankyrin	L35601/Q24241	ANK	4-101F1-102A8	FBgn0011747

studied here, on chromosome arms 8p, 10q, 4p, and 5q, respectively. Each of the *FGFR* genes has a direct ortholog in other vertebrate orders, including bony fishes (Emori, Yasuoka, and Saigo 1992; Coulier et al. 1997).

FGFRs also exist in nonvertebrates. One FGFR sequence has been characterized in echinoderm *S. purpuratus* (see table 1). It is equally similar to all the vertebrate FGFRs (Coulier et al. 1997). This suggests that the genome duplication events that have led to four *FGFR* genes in vertebrates occurred after the separation of echinoderm and chordate ancestors. Two *FGFR* genes are found in the *D. melanogaster* laboratory fruit fly (see table 2). They are called *breathless* (*bt1*), *DFR2* or *DFGF-R1*, and *heartless* (*htl*), *DFR1* or *DFGF-R2* (Klambt, Glazer, and Shilo 1992; Shishido et al. 1993; Beiman, Shilo, and Volk 1996; Gisselbrecht et al. 1996). They are more similar to one another than they are to vertebrate or echinoderm *FGFR* genes. This suggests that the fly *FGFR* ancestor separated from the chordate ancestor and then evolved independently (this evolution might have involved a local duplication—see below). To date, only one *FGFR* sequence, *egl-15*, has been characterized for *C. elegans* (De Vore, Horvitz, and Stern 1995; table 3).

#### ADR Family

ADRB3 belongs to a large multigenic family. The phylogenetic tree (see Fryxell 1995) and mapping data

(OMIM) are helpful in approaching the putative evolution of the ADR gene family (see fig. 4). In humans, the ADR family comprises three  $\beta$ -adrenergic receptor genes, *ADRB1-3*, three  $\alpha$ -1-adrenergic receptor genes, *ADRA1A-C*, and three  $\alpha$ -2-adrenergic receptor genes, *ADRA2A-C*, two dopamine receptors, *D1* (also called *D1A*) and *D5* (also called *D1B*) (see table 1), and eight serotonin receptors (Zifa and Fillion 1992). In the fly, the family is represented by the octopamine and dopamine receptors (table 2), which share a common ancestor which is distinct from the other members of the superfamily of G-protein-coupled neurotransmitters.

From the results of phylogenetic analyses, it can be hypothesized that local duplications first gave rise to *ADRB-D1/D5-ADRA1-ADRA2* and serotonin receptor ancestor genes. At one stage during evolution, the serotonin receptor ancestor separated from the ancestor region. Local duplication of the *ADRB-D1/D5-ADRA1* ancestor gene gave rise to both *ADRB-D1/D5* and *ADRA1* ancestor genes (Fig. 4). Similarly, local duplication of the *ADRB-D1/D5* ancestor gene gave rise to the two *ADRB* and *D1/D5* ancestor genes. All these rounds of local duplication seem to have been followed by several—likely two—rounds of large-scale duplication. Indeed, in humans, *ADRB3* is linked to *ADRA1C* on chromosome arm 8p, and *ADRB2*, *ADRA1B*, and *D1* are on chromosome region 5q33-35, while *ADRB1* and

**Table 3**  
**Information on *Caenorhabditis elegans* Genes**

Cosmid Name	Locus Name	Vertebrate Ortholog	Accession Number(s)	Position P Map <sup>a</sup>	Genetic Map
C52B11.3 . . . . .		<i>ADR</i>	U41276	X (ctg 674) -1909 to -1874	X - 18.1434
MO3F4.3 . . . . .		<i>ADR</i>	U64601	X (ctg 674) 132-163	X - 6.1
F14D12.6 . . . . .		<i>ADR</i>	U41021	X (ctg 674) 445-470	X - 4.6
WO1C8.6 . . . . .		<i>VACHT/VMAT</i>	U41508	X (ctg 674) 494-517	X - 4.6
FO1E11.5 . . . . .		<i>ADR</i>	U42832	X (ctg 674) 1188-1203	X - 2.17
F58A3.2 . . . . .	<i>egl-15</i>	<i>FGFR</i>	Z81O17	X (ctg 674) 3426-3451	X 2.49549
C27C12.2 . . . . .		<i>EGR</i>	Z69883	X (ctg 674) 5572-5601	X 18.54
F59C12.2 . . . . .		<i>ADR</i>	U411038	X (ctg 674) 6624-6643	X 24
	<i>unc-17</i>	<i>VACHT</i>	L19621		IV - 3.30557
B0350.2 . . . . .	<i>unc-44</i>	<i>ANK</i>	U50071, U21734	IV (ctg 423) 1735-1706	IV 2.77831

<sup>a</sup> The chromosome, number of contig (ctg), and position in kilobases are successively listed.

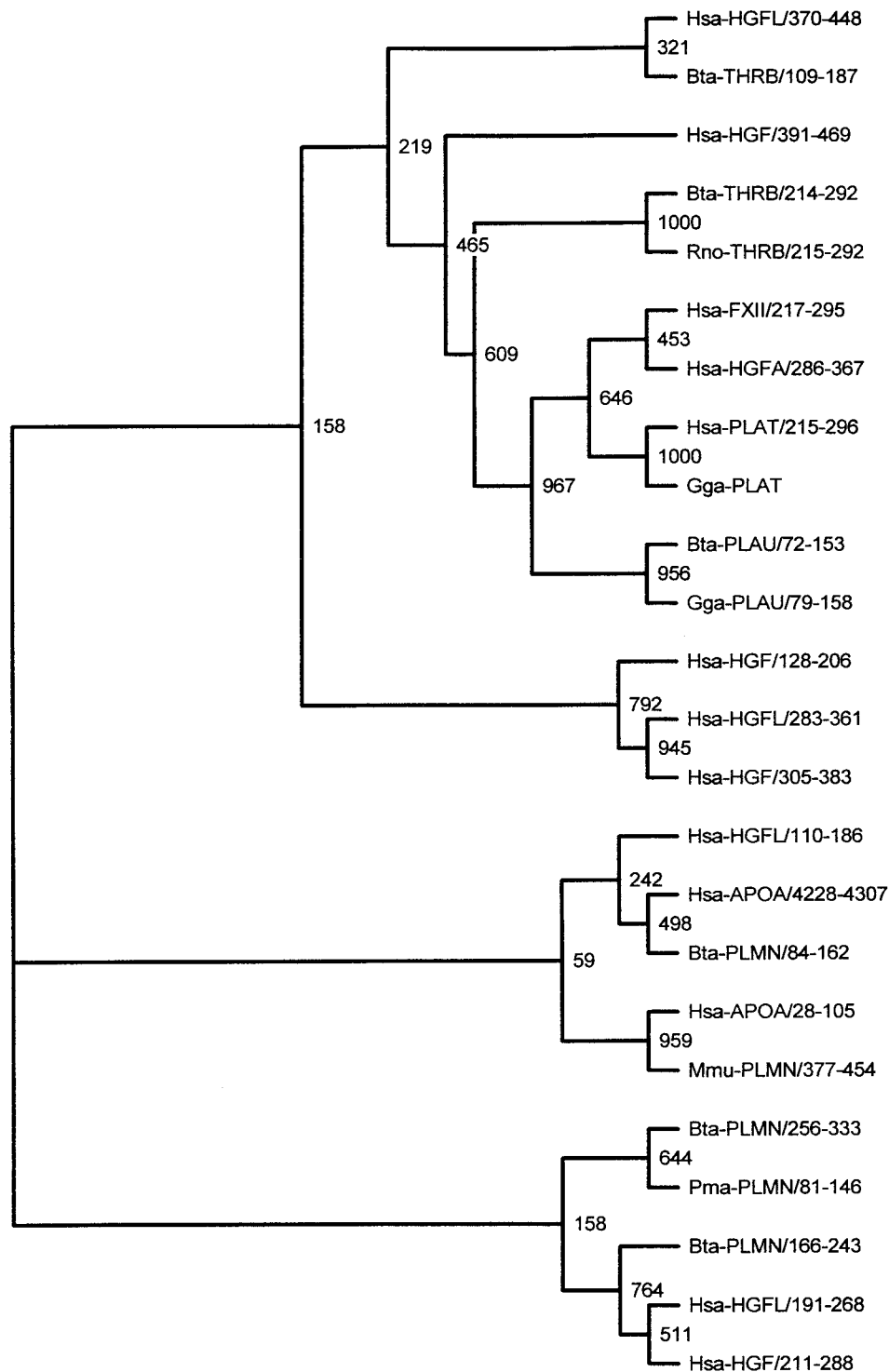


FIG. 2.—Cladogram of the plasminogen activator family. The alignment used to infer the tree was done using only the kringle domain of each protein. The vertebrate sequences were used to infer the phylogenetic tree. The tree is unrooted. The branch lengths are arbitrary and are not drawn according to the genetic distance. The number of trees with a particular node (among 1,000 bootstrap replicates) is indicated at the node.

*ADRA2A* are linked on chromosome region 10q25 (Manca et al. 1997). The *D5* and *ADRA2C* genes are on the 4p16 region. Yang-Feng et al. (1990) showed that the *ADRB2* and *ADRA1B* genes are within a 300-kb segment and the distance between *ADRB1* and *ADRA2A* is

less than 225 kb. The close proximity of these two pairs of *ADR* genes and the sequence similarity that is found among all *ADR* genes confirm their evolutionary relationship. Only two members of this family, *ADRA1A* and *ADRA2B*, are found outside the paralogous regions

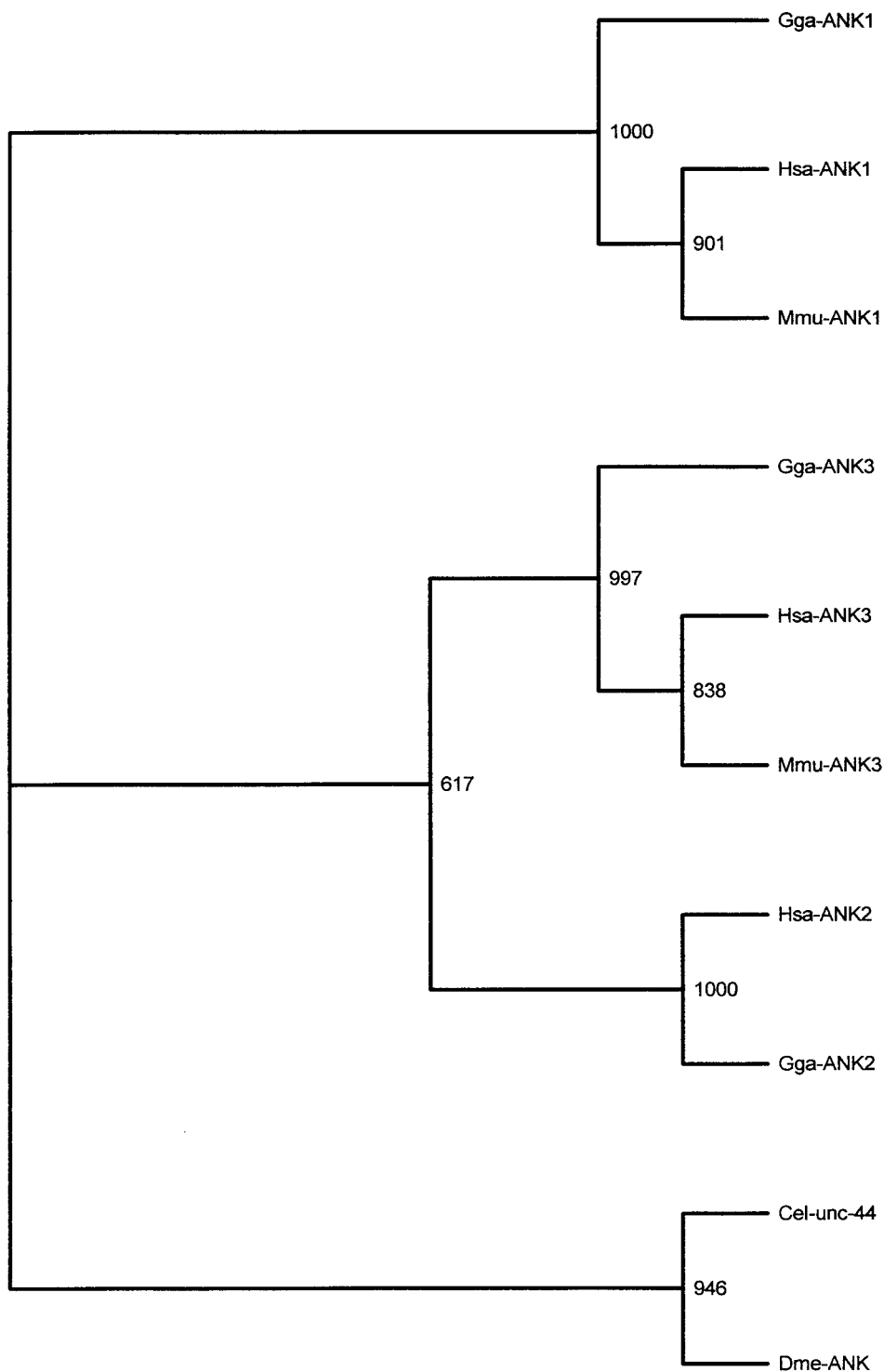


FIG. 3.—Cladogram of the ankyrin family. The alignment was done with a part of the spectrin domain (corresponding to positions 118–225 of human ANK1) of each protein, due to the fact that only partial sequences are available for chicken ankyrins. The tree is unrooted. The branch lengths are arbitrary and are not drawn according to the genetic distance. Bootstrap values are shown at nodes.

studied here, i.e., on chromosomes 20 and 2, respectively.

The timing of the large-scale duplication event(s) can be estimated by looking for these paralogous genes in other species. The duplications must have occurred before the appearance of the bony vertebrates. *DIA* and

*D5/DIB* have direct orthologs in the bony fish *A. anguilla*, for example (Cardinaud et al. 1997; table 1). Furthermore, a *DIC* receptor gene is found in chickens, *X. laevis*, and bony fishes (table 1). This gene belongs to the *D1* subfamily and is a paralog of *DIA* and *DIB* genes. Therefore, *DIC* emerged before the bony verte-

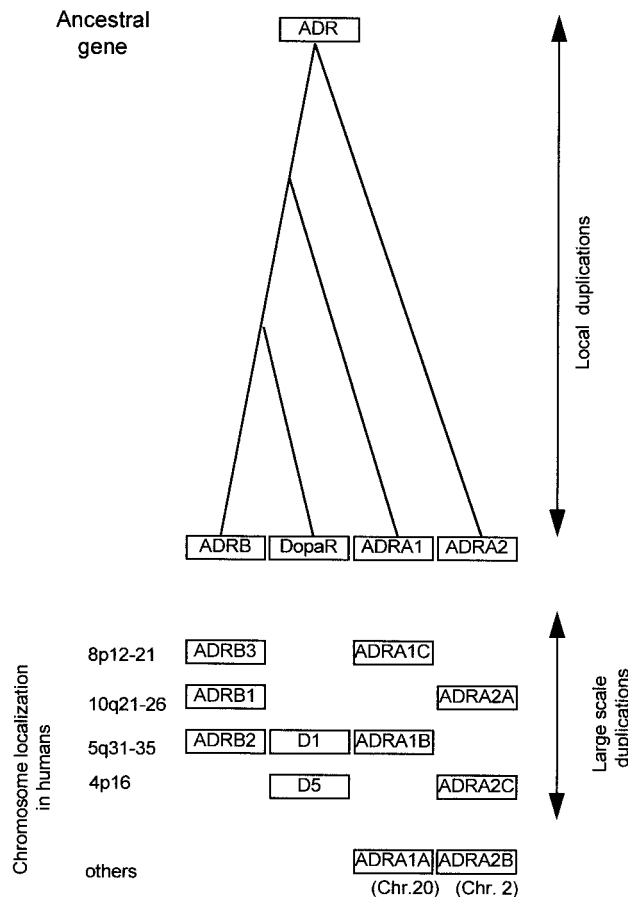


FIG. 4.—Tentative scenario for the evolution of the *ADR* gene family found in four paralogous chromosomal regions in humans. Human gene mapping information was obtained using OMIM database (see *Materials and Methods*).

brate radiation. This gene has not yet been found in mammals. A fourth *DI* paralog, *DID*, is found in *X. laevis* and chickens (table 1). It is tempting to speculate that *DIC* and *DID*, if they have not been eliminated, exist in the human genome, possibly on chromosome arms 8p and 10q. With respect to the other subfamilies, information is available only for *ADRB2*, which has been characterized in the turkey, *M. gallopavo* (Fryxell 1995; table 1). Therefore, the large-scale duplication involving the *ADRB* genes could have occurred before the amniote radiation, which is consistent with what is found for *DIA* and *D5/DIB*. The phylogenetic tree described by Fryxell (1995) also hypothesizes that the large-scale duplications occurred after the separation of the fly/vertebrate ancestor.

Database searches revealed the presence of several protein sequences in *C. elegans* that are more related to this family (see below) than to other families of G-coupled receptors: C52B11.3, MO3F4.3, F14D12.6, F01E11.5, and F59C12.2 (Table 3).

#### EGR Family

EGR proteins have zinc-finger motifs and display early induction kinetics in fibroblasts, epithelial cells, and lymphocytes following mitogenic stimulation. Four

human *EGR* genes have been identified to date: *EGR1* (*KROX24*), *EGR2* (*KROX20*), *EGR3*, and *EGR4* (table 1). The first three genes share about 85% identity. *EGR1*, *EGR3*, and *EGR2* genes map to chromosome bands 5q33.1, 8p21, and 10q21.1, respectively.

*EGR1* has been identified in the zebrafish, *D. rerio*, and in *X. laevis*, and *EGR2* has been identified in *X. laevis*. It seems that the duplications leading to the present-day family occurred before the bony vertebrate radiation. An *EGR* gene showing an orthologous relationship with the vertebrate *EGR* family is found in *D. melanogaster* (*stripe*, table 2) and in *C. elegans* (table 3).

The phylogenetic tree of the *EGR* family, derived from the most conserved domain corresponding to amino acid positions 257–883 of human *EGR2* (fig. 5), suggests that the duplication giving rise to some members of the family (i.e., *EGR1*, *EGR2*, and *EGR3*) occurred before the vertebrate radiation. In humans, *EGR4*, which is much less conserved, is located on chromosome 2p13. It is also found in other mammalian species (table 1). *EGR4* has evolved independently and has diverged from the other members of the family, possibly before the Protostomia/Deuterostomia split.

#### VMAT Family

VMAT1 (also named VAT1 or SLC18A1) is the adrenal vesicular amine transporter. VMAT2 (also named synaptic vesicle amine transporter, SVAT) is highly similar to VMAT1 and has a brain-specific expression. *VMAT1* and *VMAT2* genes map to chromosome regions 8p21.3 and 10q25, respectively. In this family, there are only two paralogs of the type studied here; however, VMAT1 and VMAT2 proteins share similarities with the vesicular acetylcholine transporter (*VACHT*) (table 1). The *VACHT* gene maps to 10q11.2.

*VACHT* is present in cartilaginous fish *Torpedo ocellata* (Varoqui et al. 1994; table 1). It shows an orthologous relationship with human *VACHT*. A vesicular acetylcholine transporter has also been described in *C. elegans* (*unc-17*) (Alfonso et al. 1993; table 3). It shows an orthologous relationship with human and torpedo *VACHT*. A *VMAT*-like sequence is found in a *C. elegans* cosmid and shares a last common ancestor with *VACHT* sequences (fig. 6). A *D. melanogaster VACHT* gene has also been described, but no sequence is available (see table 2).

The duplication giving rise to *VMAT* and *VACHT* occurred before the separation of the vertebrate and nematode ancestors, and the *VMAT1/VMAT2* large-scale duplication occurred after the separation of these two phyla. It is possible that, as for the *FGFR* genes, a duplication event created four *VMAT* paralogs, two of which were lost or have not been characterized yet.

#### LPL Family

The LPL family currently includes five related vertebrate proteins (table 1) and three less-related *D. melanogaster* proteins (table 2). The vertebrate proteins are lipoprotein lipase (LPL), hepatic lipase (HL), pancreatic lipase (PNLIP), and two pancreatic lipase-related proteins (PNLIP-like 1 and 2) (Giller et al. 1992). These

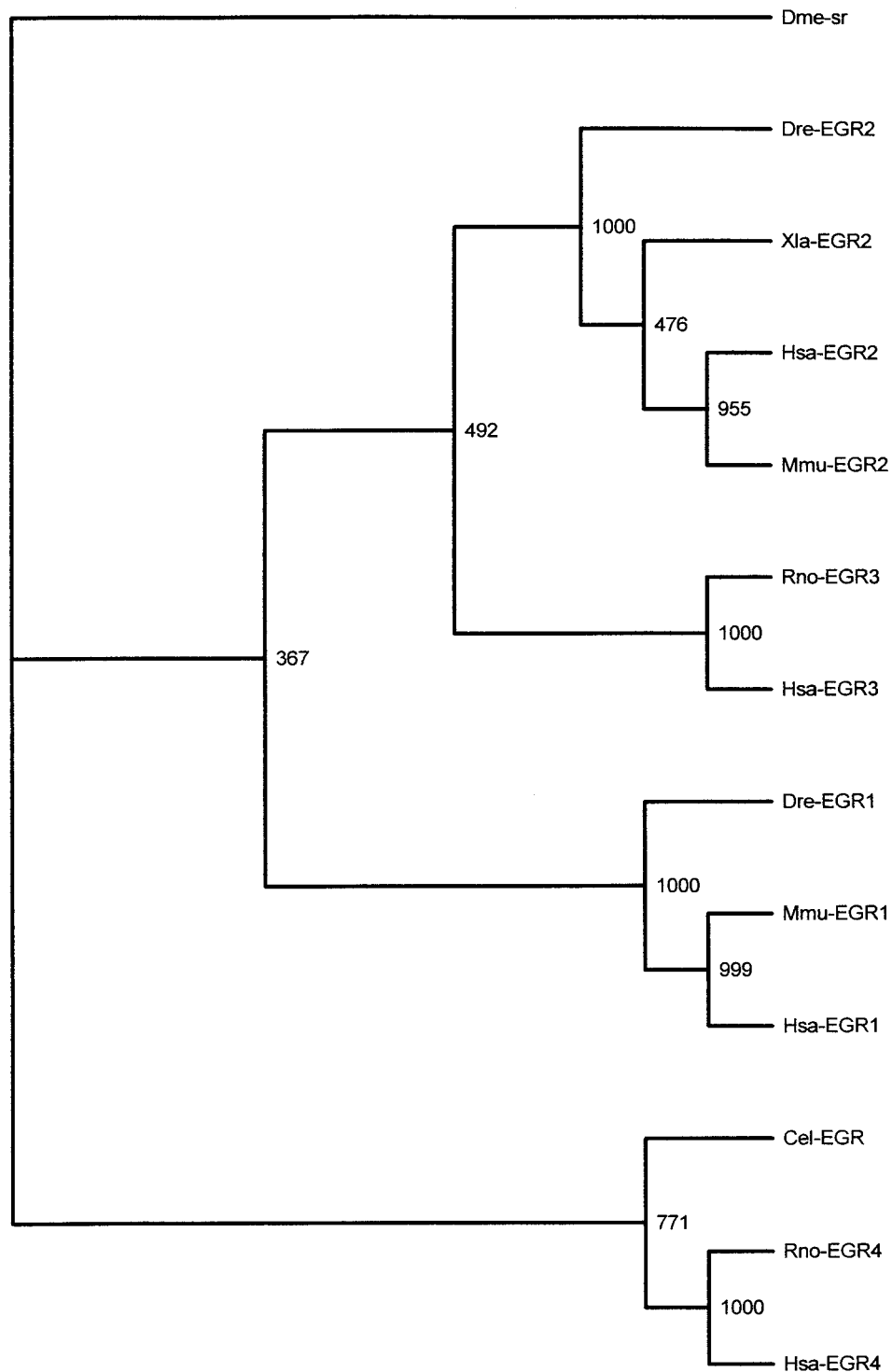


FIG. 5.—Cladogram of the EGR family. The alignment was done with the most conserved domain of each protein, corresponding to positions 257–883 of human EGR2. The tree is unrooted. The branch lengths are arbitrary and are not drawn according to the genetic distance. Bootstrap values are shown at nodes.

proteins hydrolyze circulating and dietary triglycerides, allowing assimilation and distribution to tissues. *LPL* and *HL* are more similar to one another than they are to *PNLIP* (Hide, Chan, and Li 1992). The *LPL* gene maps to chromosome band 8p22, while *PNLIP* maps to 10q24–25. *HL* is located on chromosome 15. No map-

ping information is available for the two pancreatic lipase-related protein genes. It is possible that the duplication giving rise to *HL* and the pancreatic lipase-related genes occurred after the large-scale duplication.

In chickens and zebrafish, only the *LPL* gene (85% and 67% identity with the human counterpart,

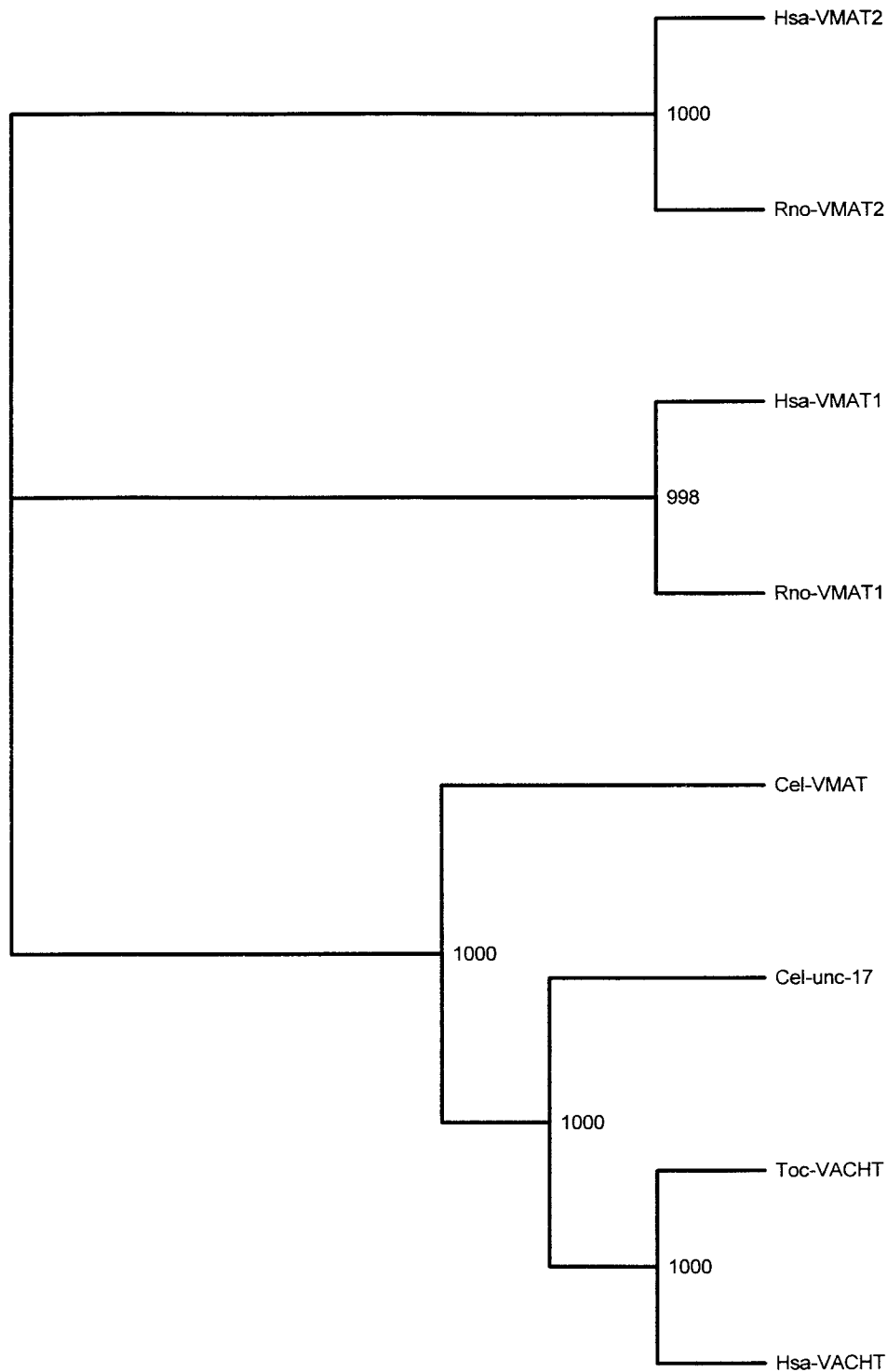


FIG. 6.—Cladogram of the VMAT family. The alignment was done with the C-terminal domain of each protein (corresponding to positions 288–478 of human VMAT1). The tree is unrooted. The branch lengths are arbitrary and are not drawn according to the genetic distance. Bootstrap values are shown at nodes.

respectively) has been identified (Cooper et al. 1989; Arnaud et al. 1996). *PNLIP* and *HL* could be present in these genomes, but they have not been identified yet.

The three *D. melanogaster* yolk proteins are more similar to one another than they are to the LPL/HL/

PNLIP family (see Hide, Chan, and Li 1992). Therefore, as hypothesized for the *FGFR* genes, the two groups of corresponding genes probably evolved independently. The events of duplication occurred independently after the separation of the Protostomia and Deuterostomia ancestors.

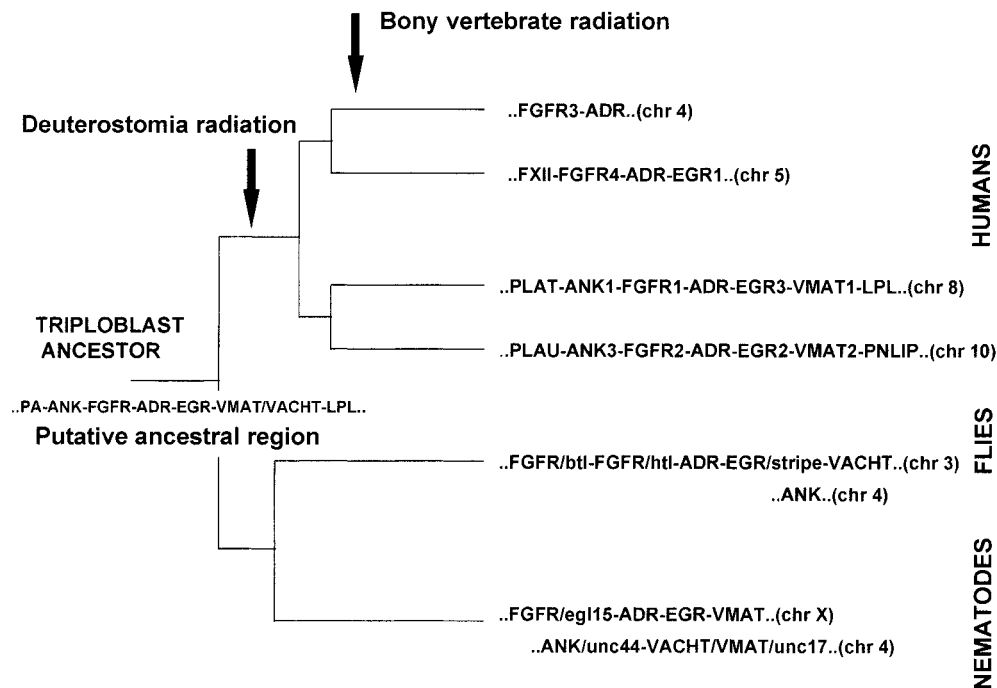


FIG. 7.—A schematic and hypothetical reconstitution of the organization and evolution of a segment of the triploblastic animal ancestral genome and its present-day homologs (shown along the branches) in key species, as deduced by analyses of paralogous and orthologous genes. Branch lengths are approximatively drawn according to timescale. The branching of flies and nematodes is according to Aguinaldo et al. (1997). The duplications that led to the four regions in humans occurred after the Deuterostomia radiation and before the bony fish radiation. However, the number and order of rounds are arbitrary. The order and sizes of the genes are not defined and they are designated either by specific names (i.e., *FGFR/egl15*, *FGFR/btl*, *EGR/stripe*, *VACHT/unc17*, or *ANK/unc44*) or by generic names of families. For humans, the size (40–50 Mb) of the region and the order of genes are only known with relative precision for 8p12–21. Chromosomal localization for each region is indicated in parentheses. For *D. melanogaster* and *C. elegans*, tentative gene order is as shown in tables 2 and 3. It is speculated that the evolution of this ancestral region led to a different organization in various species. The *ANK* and *VACHT* genes (and possibly also the *PA* and *LPL*-like genes) became separated from the other genes in humans, *D. melanogaster*, and *C. elegans*. With respect to the *ANK* gene, it is possible that a unique genetic event led to the separation of the *ANK* gene from the other genes considered here for the hypothetical (Aguinaldo et al. 1997) ancestor of flies and nematodes.

#### Integrative View of the Evolution of the Ancestral Chromosome-8-like Region

We have hypothesized that [PLAU, PLAT, HGFA, FXII], [ANK], [FGFR], [ADRB3, D1/D5, ADRA1, ADRA2], [EGR], [VMAT], and [LPL, HL, PNLIP] ancestors were on the same chromosomal region (fig. 7). This ancestral genomic organization existed before the emergence of the bony vertebrate common ancestor. Indeed, we show that members of each family are linked on paralogous chromosomal regions and that the duplications occurred before the radiation of bony vertebrates. We found that genes with similarities to human genes described here are found in nonvertebrates—echinoderms, protostomes, and pseudocoelomates—but are not the direct orthologs of individual human genes. Thus duplications of the ancestral chromosome are likely to have occurred in the Deuterostomia lineage after the separation of echinoderms and chordates (fig. 1).

These large-scale duplications could be the consequence of the two rounds of tetraploidization hypothesized to have occurred in the craniate lineage (Ohno 1970; Holland et al. 1994). It is believed that the first round occurred just after the separation of the Cephalochordata and Craniata ancestors. Therefore, the ancestral genomic organization could be present in the craniate last common ancestor.

To investigate the origin of this linkage, we looked for the localization of orthologs of these gene families in the two nonvertebrate species for which mapping information is available, *D. melanogaster* and *C. elegans*.

#### Gene Mapping Information for *Drosophila*

The two *Drosophila* *FGFR* genes, *breathless* and *heartless*, are both on chromosome 3, at positions 70C6–70D and 90C–D, respectively (this suggests that a local duplication has occurred in the *Drosophila* lineage). *Drosophila melanogaster* octopamine (ODM), D1, and serotonin receptors, as well as vertebrate [ADRB, D1/D5, ADRA1, ADRA2] protein families (Fryxell 1995), belong to the same superfamily. *ODM* maps to bands 99a–100B1 on chromosome 3 (Arakawa et al. 1990). *Serotonin receptor 2* maps to 82C–82D on chromosome 3. *Serotonin receptor 7* is at position 100A, also on chromosome 3. *Serotonin receptor 1B* and *serotonin receptor 1A* are on chromosome 2, at position 56A–B. The *Drosophila* *EGR* (*stripe*) maps to 90E, and *VACHT* maps to position 91–97DC, on chromosome 3.

The *D. melanogaster* *ankyrin* gene maps to chromosome 4, at positions 101F–102A. A rearrangement thus separated the *ANK* gene from the other genes of the region. This rearrangement may have taken place before the radiation of the *Drosophila* genus, since, in

most cases, the element homologous to the *D. melanogaster* fourth chromosome is a separate element, except for *Drosophila busckii* (Ashburner 1989)

The size of the *D. melanogaster* genome is 104 cytogenetic bands. *FGFR heartless*, *ODM*, *Serotonin receptor 7*, *EGR stripe*, and *VACHT* are all located between bands 90 and 100 on chromosome 3, representing about one tenth of the *Drosophila* genome. The probability of finding these genes linked in such a limited distance is low unless they were ancestrally linked. The linkage of the ancestors of these gene families is likely to have predated the radiation of triploblastic animals.

#### Gene Mapping in *C. elegans*

Genes showing similarity with the gene families we discussed have been found in *C. elegans*. Some of the genes have been described functionally, such as *egl-15*, which encodes an FGF receptor (De Vore, Horvitz, and Stern 1995), and *VACHT* (Alfonso et al. 1993). Several sequences showing similarity with the [*ADRB*, *D1/D5*, *ADRA1*, *ADRA2*, *serotonin receptor*] family are found in *C. elegans*. The more similar ones are in cosmids C52B113, MO3F4.3, F14D12.6, F01E11.5, and F59C12.2 (see table 3). *Ankyrin* (*unc-44*) (Otsuka et al. 1995) and some other genes, such as an *EGR* sequence showing 90% identity with the mammalian *EGR* genes, have been defined upon analysis of the *C. elegans* genome.

The *C. elegans* genome is 100 Mb, or 300 cM, long. *FGFR*, the five *ADR*-like genes, *EGR*, and *VACHT/VMAT* are found on chromosome X within an 8-Mb segment, while *ANK* and *VACHT* are linked on chromosome 4, separated by a genetic distance of 6 cM (the physical length is unknown). The probability of finding *FGFR*, *ADR*-like, *EGR*, and *VACHT/VMAT* genes within an 8-Mb fragment would be  $5 \times 10^{-4}$  if they were randomly distributed. The probability of finding *ANK* and *VACHT* genes on a 6 cM fragment would be 0.02 under the same assumptions. The probability of having these genes linked in two groups by chance alone is  $10^{-5}$ .

One likely explanation for the clustering of these genes is that they were linked in the triploblast common ancestor. The *ANK/unc-44* and *VACHT/unc-17* genes likely became separated from the other genes of the region in the *C. elegans* lineage.

#### Conclusions and Speculations

We propose two important hypotheses. First, a unique chromosomal region in the triploblastic last common ancestor contained single-copy linked genes, and its duplication led to four copies in bony vertebrates. Synteny (but not order) has been conserved. Second, the duplication events occurred before the bony vertebrate radiation and, more probably, after the Protostomia/Deuterostomia split. The robustness of all phylogenetic trees does not enable a particular duplication scheme (either two rounds of large-scale duplication occurred at different periods of time, as discussed by Holland et al. [1994], or successive duplications occurred in a short period of time) to be clearly identified.

This kind of study could be extended to other paralogous regions of the human genome, or of any vertebrate genome. Several paralogous regions have already been recognized in humans (Lundin 1993; Rosnet et al. 1993; Birnbaum et al. 1994; Dib et al. 1994; Ruddle et al. 1994; Kasahara et al. 1996; Katsanis, Fitzgibbon, and Fisher 1996; Spring 1997) and conserved synteny between mammals and nonmammalian vertebrates (Postlethwait et al., 1998) and between mammals and non-vertebrates (Trachtulec et al. 1997) has been also noted. To find more similar observations, we suggest that a good starting point could be to search for syntenies conserved between the *C. elegans* and human genomes.

This type of investigation may help with both the reconstitution of the theoretical ancestral genome and the understanding of the organization of our present-day genes. It constitutes an important area of postgenome projects in search of the events that have molded animal evolution. Here we propose a little piece of this fascinating puzzle: we suggest the organization of a series of genes in the ancestral genome as depicted in figure 7. One important question is whether this ancestral linkage occurred fortuitously or was associated with a common function or regulation. From the known functions of the present-day genes, it is hard to choose between the two possibilities, although plasminogen activators, *FGFR*, and *EGR* are all involved at some stage in the FGF stimulatory pathway.

Phylogenetic analyses rely much on morphological and sequence analyses. The kind of "reconstitution" we propose in figure 7 may be an additional useful method. Although they remain speculative, more "reconstitution studies" may bring a wealth of information on possible evolutionary relationships between phyla.

#### Acknowledgments

We thank F. Birg, C. Mawas, J. Thiery-Mieg, and R. Roubin for their encouragement and critical reading of the manuscript. This work was supported by IN-SERM.

#### LITERATURE CITED

- ADÉLAÏDE, J., M. CHAFFANET, A. IMBERT et al. (13 co-authors). 1998. Chromosome region 8p11-p21: refined mapping and molecular alterations in breast cancer. *Genes Chromosom. Cancer* **22**:186-199.
- AGUINALDO, A. M., J. M. TURBEVILLE, L. S. LINFORD, M. C. RIVERA, J. R. GAREY, R. A. RAFF, and J. A. LAKE. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* **387**:489-493.
- ALFONSO, A., K. GRUNDAHL, J. S. DUERR, H. P. HAN, and J. B. RAND. 1993. The *Caenorhabditis elegans unc-17* gene: a putative vesicular acetylcholine transporter. *Science* **261**: 617-619.
- ALTSCHUL, S. F., W. GISH, W. MILLER, E. W. MYERS, and D. J. LIPMAN. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**:403-410.
- ALTSCHUL, S. F., T. L. MADDEN, A. A. SCHAFFER, J. ZHANG, Z. ZHANG, W. MILLER, and D. J. LIPMAN. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**:3389-3402.

- ANTEQUERA, F., and A. BIRD. 1993. Number of CpG islands and genes in human and mouse. *Proc. Natl. Acad. Sci. USA* **90**:11995–11999.
- ARAKAWA, S., J. D. GOCAYNE, W. R. MCCOMBIE, D. A. URQUHART, L. M. HALL, C. M. FRASER, and J. C. VENTER. 1990. Cloning, localization, and permanent expression of a *Drosophila* octopamine receptor. *Neuron* **4**:343–354.
- ARNAUD, F., J. ETIENNE, L. NOE, A. RAISONNIER, D. BRAULT, J. HARNEY, M. BERRY, C. FROMENTAL-RAMAIN, J. HAMELIN, and F. GALIBERT. 1996. Human lipoprotein lipase last exon is not translated in contrast to lower vertebrates. *J. Mol. Evol.* **43**:109–115.
- ASHBURNER, M. 1989. *Drosophila*: a laboratory handbook and manual. Two volumes. Cold Spring Harbor Laboratory Press, New York.
- BAILEY, W. J., J. KIM, G. P. WAGNER, and F. H. RUDDLE. 1997. Phylogenetic reconstruction of vertebrate Hox cluster duplications. *Mol. Biol. Evol.* **14**:843–853.
- BEIMAN, M., B. Z. SHILO, and T. VOLK. 1996. Heartless, a *Drosophila* FGF receptor homolog, is essential for cell migration and establishment of several mesodermal lineages. *Genes Dev.* **10**:2993–3002.
- BIRNBAUM, D., M. J. PÉBUSQUE, A. IMBERT, A. DIB, O. DELAPEYRIERE, and F. COULIER. 1994. Oncogenesis and genome duplication maps. *Oncol. Rep.* **1**:477–480.
- BLUMENTHAL, T., and J. SPIETH. 1996. Gene structure and organization in *Caenorhabditis elegans*. *Curr. Opin. Genet. Dev.* **6**:692–698.
- BRENNER, S. 1988. The molecular evolution of genes and proteins: a tale of two serines. *Nature* **334**:528–530.
- BRENNER, S., G. ELGAR, R. SANDFORD, A. MACRAE, B. VENKATESH, and S. APARICIO. 1993. Characterization of the pufferfish (*Fugu*) genome as a compact model vertebrate genome. *Nature* **366**:265–268.
- CARDINAUD, B., K. S. SUGAMAORI, S. COUDOUEL, J. D. VINCENT, B. NIZNIK, and P. VERNIER. 1997. Early emergence of three dopamine D1 receptor subtypes in vertebrates. Molecular phylogenetic, pharmacological, and functional criteria defining D1A, D1B, and D1C receptors in European eel *Anguilla anguilla*. *J. Biol. Chem.* **272**:2778–2787.
- CASTELLINO, F. J., and J. M. BEALS. 1987. The genetic relationships between the kringle domains of human plasminogen, prothrombin, tissue plasminogen activator, urokinase, and coagulation factor XII. *J. Mol. Evol.* **26**:358–369.
- CHAN, S. J., Q. P. CAO, and D. F. STEINER. 1990. Evolution of the insulin superfamily: cloning of a hybrid insulin/insulin-like growth factor cDNA from *Amphioxus*. *Proc. Natl. Acad. Sci. USA* **87**:9319–9323.
- COOPER, D. A., J. C. DESTAIN, J. STRIELEMAN, and A. BENASADOU. 1989. Avian adipose lipoprotein lipase: cDNA sequence and reciprocal regulation of mRNA levels in adipose and heart. *Biochim. Biophys. Acta* **1008**:92–101.
- COULIER, F., P. PONTAROTTI, R. ROUBIN, H. HARTUNG, M. GOLDFARB, and D. BIRNBAUM. 1997. Of worms and men: an evolutionary perspective on the fibroblast growth factor (FGF) and FGF receptor families. *J. Mol. Evol.* **44**:43–56.
- DAVIS, C. G. 1990. The many faces of epidermal growth factor repeats. *New Biol.* **2**:410–419.
- DE VORE, D. L., H. R. HORVITZ, and M. J. STERN. 1995. An FGF receptor signaling pathway is required for the normal cell migrations of the sex myoblasts in *C. elegans* hermaphrodites. *Cell* **83**:611–620.
- DIB, A., J. ADELAIDE, F. COURJAL, A. COURSEAU, J. JACQUEMIER, P. GAUDRAY, C. THEILLET, M. J. PÉBUSQUE, and D. BIRNBAUM. 1994. Co-amplification in human breast tumors and physical linkage at chromosomal band 12p13, of *CCND2* and *FGF6* genes. *Int. J. Oncol.* **5**:1375–1378.
- EMORI, Y., A. YASUOKA, and K. SAIGO. 1992. Identification of four FGF receptor genes in Medaka fish (*Oryzias latipes*). *FEBS Lett.* **314**:176–178.
- FELSENSTEIN, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**:783–791.
- FIELDS, C., M. D. ADAMS, O. WHITE, and J. C. VENTER. 1994. How many genes in the human genome? *Nat. Genet.* **7**:345–346.
- FITCH, W. M. 1970. Distinguishing homologous from analogous proteins. *Syst. Zool.* **19**:99–113.
- FRYXELL, K. J. 1995. The evolutionary divergence of neurotransmitter receptors and second-messenger pathways. *J. Mol. Evol.* **41**:85–97.
- GALAU, G. A., W. H. KLEIN, M. M. DAVIS, B. J. WOLD, R. J. BRITTEN, and E. H. DAVIDSON. 1976. Structural gene sets active in embryos and adult tissues of the sea urchin. *Cell* **74**:487–505.
- GARCIA-FERNANDEZ, J., and P. W. HOLLAND. 1994. Archetypal organization of the *Amphioxus* Hox gene cluster. *Nature* **18**:563–566.
- GILLER, T., P. BUCHWALD, D. BLUM-KAELIN, and W. HUNZIKER. 1992. Two novel human pancreatic lipase related proteins, hPLRP1 and hPLRP2. Differences in colipase dependence and in lipase activity. *J. Biol. Chem.* **267**:16509–16516.
- GISSELBRECHT, S., J. B. SKEATH, C. DOE, and A. M. MICHELSON. 1996. *Heartless* encodes a fibroblast growth factor receptor (DFR1/DFGF-R2) involved in the directional migration of early mesodermal cells in the *Drosophila* embryo. *Genes Dev.* **10**:3003–3017.
- GOFFEAU, A., B. G. BARRELL, H. BUSSEY et al. (13 co-authors). 1996. Life with 6000 genes. *Science* **274**:546.
- HIDE, W. A., L. CHAN, and W. H. LI. 1992. Structure and evolution of the lipase superfamily. *J. Lipid Res.* **33**:167–178.
- HOLLAND, P. W., and J. GARCIA-FERNANDEZ. 1996. Hox genes and chordate evolution. *Dev. Biol.* **173**:382–395.
- HOLLAND, P. W., J. GARCIA-FERNANDEZ, N. A. WILLIAMS, and A. SIDOW. 1994. Gene duplications and the origin of vertebrate development. Pp. 125–133 in M. AKAM, P. HOLLAND, P. INGHAM, and G. WRAY, eds. *The evolution of developmental mechanisms*. Development supplement. The Company of Biologists, Cambridge, England.
- KANDIL, E., C. NAMIKAWA, M. NONAKA, A. S. GREENBERG, M. F. FLAJNIK, T. ISHIBASHI, and M. KASAHARA. 1996. Isolation of low molecular mass polypeptide complementary DNA clones from primitive vertebrates. Implications for the origin of MHC class I-restricted antigen presentation. *J. Immunol.* **156**:4245–4253.
- KASAHARA, M., M. HAYASHI, K. TANAKA, H. INOKO, K. SUGAYA, T. IKEMURA, and T. ISHIBASHI. 1996. Chromosomal localization of the proteasome Z subunit gene reveals an ancient chromosomal duplication involving the major histocompatibility complex. *Proc. Natl. Acad. Sci. USA* **93**:9096–9101.
- KASAHARA, M., J. NAKAYA, Y. SATTI, and N. TAKAHATA. 1997. Chromosomal duplication and the emergence of the adaptive immune system. *Trends Genet.* **13**:90–92.
- KATSANIS, N., J. FITZGIBBON, and E. M. FISHER. 1996. Paralogy mapping: identification of a region in the human MHC triplicated onto human chromosomes 1 and 9 allows the prediction and isolation of novel PBX and NOTCH loci. *Genomics* **35**:101–108.
- KLAMBT, C., L. GLAZER, and B. Z. SHILO. 1992. Breathless, a *Drosophila* FGF receptor homolog, is essential for migration of tracheal and specific midline glial cells. *Genes Dev.* **6**:1668–1678.

- LESLIE, N. D., C. A. KESSLER, S. M. BELL, and J. L. DEGEN. 1990. The chicken urokinase-type plasminogen activator gene. *J. Biol. Chem.* **265**:1339–1344.
- LUNDIN, L. G. 1993. Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. *Genomics* **16**:1–19.
- LUX, S. E., K. M. JOHN, and V. BENNETT. 1990. Analysis of cDNA for human erythrocyte ankyrin indicates a repeated structure with homology to tissue-differentiation and cell-cycle control proteins. *Nature* **344**:36–42.
- MANCA, A., E. V. VOLPI, F. LAFICARA, R. MURESU, I. C. GRAY, N. K. SPURR, and C. NOBILE. 1997. Detailed physical analysis of a 1.5-megabase YAC contig containing the MXII and ADRA2A genes. *Genomics* **45**:407–411.
- MIKLOS, G. L., and G. M. RUBIN. 1996. The role of the genome project in determining gene function: insights from model organisms. *Cell* **86**:521–529.
- NAGAMATSU, S., S. J. CHAN, S. FALKMER, and D. F. STEINER. 1991. Evolution of the insulin gene superfamily. Sequence of a preproinsulin-like growth factor cDNA from the Atlantic hagfish. *J. Biol. Chem.* **266**:2397–2402.
- OHNO, S. 1970. *Evolution by gene duplication*. Springer Verlag, Berlin/Heidelberg/New York.
- OTSUKA, A. J., R. FRANCO, B. YANG et al. (11 co-authors). 1995. An ankyrin-related gene (*unc-44*) is necessary for proper axonal guidance in *Caenorhabditis elegans*. *J. Cell. Biol.* **129**:1081–1092.
- PAGE, R. D. 1996. TREEVIEW: an application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci.* **12**:357–358.
- PEARSON, W. R. 1990. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol.* **183**:63–98.
- POSTLETHWAIT, J. H., Y. L. YAN, M. A. GATES et al. (26 co-authors). 1998. Vertebrate genome evolution and the zebrafish gene map. *Nat. Genet.* **18**:345–349.
- ROSNET, O., D. STEPHENSON, M. G. MATTEI, S. MARCHETTO, M. SHIBUYA, V. CHAPMAN, and D. BIRNBAUM. 1993. Close physical linkage of the *FLT1* and *FLT3* genes on chromosome 13 in man and chromosome 5 in mouse. *Oncogene* **8**:173–179.
- RUDDLE, F. H., K. L. BENTLEY, M. T. MURTHA, and N. RISCH. 1994. Gene loss and gain in the evolution of the vertebrates. Pp. 155–161 in M. AKAM, P. HOLLAND, P. INGHAM, and G. WRAY, eds. *The evolution of developmental mechanisms*. Development supplement. The Company of Biologists, Cambridge, England.
- SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**:406–425.
- SCHUGHART, K., C. KAPPEN, and F. H. RUDDLE. 1989. Duplication of large genomic regions during the evolution of vertebrate homeobox genes. *Proc. Natl. Acad. Sci. USA* **86**:7067–7071.
- SHISHIDO, E., S. I. HIGASHIJIMA, Y. EMORI, and K. SAIGO. 1993. Two FGF-receptor homologues of *Drosophila*: one is expressed in mesodermal primordium in early embryos. *Development* **117**:751–761.
- SIDOW, A. 1996. Gen(om)e duplications in the evolution of early vertebrates. *Curr. Opin. Genet. Dev.* **6**:715–722.
- SPRING, J. 1997. Vertebrate evolution by interspecific hybridization—are we polyploid? *FEBS Lett.* **400**:2–8.
- THOMPSON, J. D., D. G. HIGGINS, and T. J. GIBSON. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**:4673–4680.
- TRACHTULEC, Z., R. M. HAMVAS, J. FOREJT, H. R. LEHRACH, V. VINCEK, and J. KLEIN. 1997. Linkage of TATA-binding protein and proteasome subunit C5 genes in mice and humans reveals synteny conserved between mammals and invertebrates. *Genomics* **44**:1–7.
- VAROQUI, H., M. F. DIEBLER, F. M. MEUNIER, J. B. RAND, T. B. USDIN, T. I. BONNER, L. E. EIDEN, and J. D. ERICKSON. 1994. Cloning and expression of the vesamicol binding protein from the marine ray *Torpedo*. Homology with the putative vesicular acetylcholine transporter UNC-17 from *Caenorhabditis elegans*. *FEBS Lett.* **342**:97–102.
- YANG-FENG, T. L., F. Y. XUE, W. W. ZHONG, S. COTECCHIA, T. FRIELLE, M. G. CARON, R. J. LEFKOWITZ, and U. FRANCKE. 1990. Chromosomal organization of adrenergic receptor genes. *Proc. Natl. Acad. Sci. USA* **87**:1516–1520.
- ZIFA, E., and G. FILLION. 1992. 5-Hydroxytryptamine receptors. *Pharmacol. Rev.* **44**:401–458.

SIMON EASTEAL, reviewing editor

Accepted May 20, 1998