Phylogenetic Analyses Alone Are Insufficient to Determine Whether Genome Duplication(s) Occurred During Early Vertebrate Evolution

AMY C. HORTON,[1] NAVIN R. MAHADEVAN,[1] ILYA RUVINSKY,[2,3] AND JEREMY J. GIBSON-BROWN[1]*

[1]*Department of Biology, Washington University, 1 Brookings Drive, St. Louis, Missouri 63130*

[2]*Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts 02114*

[3]*Department of Genetics, Harvard Medical School, Boston, Massachusetts 02114*

ABSTRACT The widely accepted notion that two whole-genome duplications occurred during early vertebrate evolution (the 2R hypothesis) stems from the fact that vertebrates often possess several genes corresponding to a single invertebrate homolog. However the number of genes predicted by the Human Genome Project is less than twice as many as in the *Drosophila melanogaster* or *Caenorhabditis elegans* genomes. This ratio could be explained by two rounds of genome duplication followed by extensive gene loss, by a single genome duplication, by sequential local duplications, or by a combination of any of the above. The traditional method used to distinguish between these possibilities is to reconstruct the phylogenetic relationships of vertebrate genes to their invertebrate orthologs; ratios of invertebrate-to-vertebrate counterparts are then used to infer the number of gene duplication events. The lancelet, amphioxus, is the closest living invertebrate relative of the vertebrates, and unlike protostomes such as flies or nematodes, is therefore the most appropriate outgroup for understanding the genomic composition of the last common ancestor of all vertebrates. We analyzed the relationships of all available amphioxus genes to their vertebrate homologs. In most cases, one to three vertebrate genes are orthologous to each amphioxus gene (median number=2). Clearly this result, and those of previous studies using this approach, cannot distinguish between alternative scenarios of chordate genome expansion. We conclude that phylogenetic analyses alone will never be sufficient to determine whether genome duplication(s) occurred during early chordate evolution, and argue that a ''phylogenomic'' approach, which compares paralogous clusters of linked genes from complete amphioxus and human genome sequences, will be required if the pattern and process of early chordate genome evolution is ever to be reconstructed. *J. Exp. Zool. (Mol. Dev. Evol.) 299B:41–53, 2003.* © 2003 Wiley-Liss, Inc.

INTRODUCTION

Whole-genome duplications (tetraploidizations) occur frequently in eukaryotes within lineages as divergent as cereals and vertebrates (Ku et al., 2000; Keller and Gerhardt, 2001; Robinson-Rechavi et al., 2001), and have been a common feature of organismal evolution throughout the history of life on Earth. Thirty years ago, Susumu Ohno proposed the influential theory, later dubbed the ''2R hypothesis,'' that vertebrate ancestors underwent two rounds of whole-genome duplication based on the observation of genome size differences between deuterostomes and apparent tetraploidization events in some fish lineages (Ohno, '70).

Support for the 2R hypothesis was bolstered by the detection of four *Hox* clusters in mammals (Krumlauf, '94), which led to the comparison of other developmental gene families in *Drosophila*

Grant sponsor: Department of Biology, Washington University to JJG-B; Grant sponsor: The Jane Coffin Childs Fund to IR.

*Correspondence to: Jeremy J. Gibson-Brown, Department of Biology, Washington University, Campus Box 1229, 1 Brookings Drive, St. Louis, MO 63130-4899.
E-mail: gibbro@wustl.edu

Received 14 May 2003; Accepted 13 August 2003
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/jez.b.00040

and vertebrates. These initial studies inferred 1:4 invertebrate-to-vertebrate gene ratios, with some loss of gene duplicates, and were interpreted as evidence in support of the 2R hypothesis (Holland et al., '94; Sidow, '96; Spring, '97). Recent studies, however, have questioned this notion. First, the human genome contains only about twice as many genes as those of invertebrates such as *Drosophila* or *Caenorhabditis* (Venter et al., 2001; Lander et al., 2001). Second, detailed phylogenetic studies of other developmental genes, not linked to the *Hox* clusters, have failed to support the original 1:4 relationships (Hughes, '99; Martin, '99; 2001; Skrabanek and Wolfe, '98; Smith et al., '99; Ruvinsky et al., 2000; Schubert et al., 2000). Third, tree topologies from analyses of non-*Hox* genes on *Hox*-bearing chromosomes are often inconsistent with two whole-genome duplications (Bailey et al., '97; Hughes et al., 2001; however see Furlong and Holland, 2002 and Larhammar et al., 2002 for critiques of these interpretations).

Previous studies contain numerous problems due to inadequate data sets. First, *Hox* clusters, and genes linked to them, represent only a tiny fraction of the genome, and the duplication history of these regions may not represent the duplication history of the genome as a whole (Spring, '97; Martin, '99; Ruvinsky et al., 2000). Second, counting genes is insufficient, since the topology of a gene tree, not pairwise alignment scores, contains the information required to determine orthology/paralogy relationships (Martin, '99; Friedman and Hughes, 2001; Venter et al., 2001).

The sequencing of complete genomes allows the large-scale comparison of individual gene histories to infer genome histories. The recent availability of the complete human genome sequence prompted us to reinvestigate the 2R hypothesis, since possessing the complete data set for humans finally allows confident assignment of a lower bound on the number of gene duplications within chordates; it eliminates the issue that missing (undiscovered) human orthologs might restructure tree topologies. Although comparisons to *Drosophila* and *Caenorhabditis* genomes are useful in setting an upper bound on the number of gene duplications in chordates, since protostomes diverged from deuterostomes very early in metazoan evolution, their genomes are highly divergent from those of stem chordates and should not therefore be used to infer duplications specific to the chordate lineage (Ruvinsky et al., 2000; Holland and Gibson-Brown, 2003; Gibson-Brown et al., 2003). The cephalochordate, amphioxus, is the closest living invertebrate relative of the vertebrates (Wada and Satoh, '94). In any phylogenetic analysis, the most appropriate outgroup is the group whose divergence predates the phenomenon studied but is as closely related as possible to the crown group. Amphioxus is therefore the most appropriate outgroup for the study of vertebrate genome duplication.

In an attempt to test the 2R hypothesis in light of the complete human gene dataset, we examined the phylogenetic relationships between all amphioxus genes for which sequences are available from the public databases and their vertebrate counterparts. This approach has the merit that examining large sets of unlinked genes provides a more comprehensive sampling across the genome, and might therefore provide greater insight into patterns of chordate genome evolution (Ruvinsky et al., 2000). Interestingly, we find a marked difference between the ratios of amphioxus-to-human homologs as compared to fly-to-human homologs. In very few cases do we see the 1:4 cephalochordate-to-human gene ratios predicted by a naïve interpretation of the 2R hypothesis however, the median ratio being only 1:2. We discuss the consequences of these data for various models of chordate gene duplication, but conclude that merely increasing the number of genes included in the analysis will neither prove nor reject the 2R hypothesis. We propose that a ‘‘phylogenomic’’ approach, in which paralogous clusters of genes from the complete amphioxus and human genome sequences are aligned, will be a necessary prerequisite to resolve this issue.

MATERIALS AND METHODS

523 publicly available amphioxus gene product sequences, from *Branchiostoma belcheri*, *B. californiensis*, *B. floridae*, and *B. lanceolatum*, were cataloged from GenBank. Ninety nonnuclear genes, 22 sequences which fall into gene families containing extensive lineage-specific duplications (e.g., calmodulin gene families; Karabinos and Bhattacharya, 2000), 43 fragments less than 100 amino acids long, and 22 genes with no human orthologs retrievable by BLASTP similarity searches, were excluded. For redundant sequences, the most complete entry was retained. When available, homologs in mice, chickens, frogs and/or newts, actinopterygian fish, sharks, lampreys, hagfish, tunicates, flies, and nematodes were obtained using BLASTP similarity searches (Althschul et al., '97). Human homologs were

obtained from both GenBank (http://www.ncbi.nlm.nih.gov) and the Celera Human Genome (http://www.celera.com) databases. Protein sequences were aligned using CLUSTALW (Thompson et al., '94) followed by manual adjustment. Unalignable regions were excluded from analysis. Phylogenetic trees were constructed by the neighbor joining and maximum parsimony methods as implemented in PAUP* (Swofford, 2001) with 1000 bootstrap replications. In addition, we used a maximum likelihood method as implemented in TREE-PUZZLE 5.0 with the BLOSUM62 substitution matrix (Strimmer and von Haeseler, '96). Orthologous relationships between amphioxus and mammalian genes that were supported at, or above, the 70% level by at least two of the three methods were accepted. Trees in which amphioxus lineage-specific duplications met the same criteria were categorized as containing a single cephalochordate gene, the inferred ancestral condition. Our work resolves 73 relationships (Table 1) and is consistent with an additional 61 previously published relationships that meet our criteria (Table 2).

RESULTS

We examined the phylogenetic relationships of 188 different genes from four species of cephalochordate to their vertebrate homologs. These genes perform a wide variety of functions ranging from developmental signaling to housekeeping metabolism. Of these, 134 gene families met our criteria for confident assessment of orthologous

TABLE 1. Orthologs supported by this study[1,2]

Cephalochordate gene(s)	Accession Number	Human gene(s)	Accession Number	Mammalian Orthologs
AKR1C1 (alpha-hydroxysteroid dehydrogenase)	CAB38007	CBR1	NP_001748	2
		CBR3	NP_001227	
Ald (aldolase)	BAA21101	ALDOA	NP_000025	3
		ALDOB	NP_000026	
		ALDOC	NP_005156	
BMP2/4 (bone morphogenetic protein)	AAF19841	BMP2	NP_001191	2
		BMP4	NP_001193	
Brf38	CAB05852	UCP2	NP_003346	2
		UCP3	NP_073714	
Brn (POU III)	AAL85498	POU3F1	Q03052	3
		POU3F3	NP_006227	
		POU3F4	NP_000298	
CAVP (calcium vector protein)	O01305	CAL2	NP_057450	4
		CAL3	Q9NZU6	
		CAL4	AAK83462	
		CAL5	NP_062829	
ChE1 (cholinesterase)	AAD05373	ACHE	NP_000656	2
ChE2	AAD05374	BCHE	NP_000046	
CK (creatine kinase)	AAK29780	CKB	NP_001814	2
		CKM	NP_001815	
CKSL (CDC28 protein kinase 1-like protein)	AAK91295	CKS1	NP_001817	2
		CKS2	NP_001818	
DAD1L (defender against cell death)	AAK82418	DAD1	NP_001335	1
DRP (dystrophin related protein)	CAA68069	DMD	NP_004013	2
		UTRN	NP_009055	
EF1 a (elongation factor 1-alpha)	BAB63216	HS1	AAA91835	1
Eph1 (ephrin receptor)	BAA84734	EPHA1	S44280	12
Eph2	BAA84735	EPHA2	NP_004422	
		EPHA3	NP_005224	
		EPHA4	NP_004429	
		EPHA5	P54756	
		EPHA7	NP_004431	

TABLE 1—Continued

Cephalochordate gene(s)	Accession Number	Human gene(s)	Accession Number	Mammalian Orthologs
		EPHA8	NP_065387	
		EPHB1	NP_004432	
		EPHB2	NP_059145	
		EPHB3	P54753	
		EPHB4	P54760	
		EPHB6	NP_004436	
FoxA1 (HNF3)	CAA70438	FOXA1	NP_004487	3
(HNF3-1)	CAA65368	FOXA2	NP_068556	
		FOXA3	P55318	
FoxN (whn)	CAA72307	FOXN1	NP_003584	2
		FOXN4	AAL23949	
Fspondin	CAA06854	SPON1	NP_006099	1
G10	AAK81863	G10	AAF03505	2
		EDG2	NP_003901	
Gli	CAB96572	GLI1	NP_005260	3
		GLI2	NP_084656	
		GLI3	NP_000159	
gsc (goosecoid)	AAF97935	GSC	P56915	2
		GSCL	AAC39544	
IF1 (type I keratin)	AAD23384	PRPH	NP_006253	4
		DES	AAC50680	
		GFAP	NP_002046	
		INA	NP_116116	
IFD1 (intermediate filament)	CAA11446	KRT5	AAH24292	5
IFE2	CAA09067	KRT6	AAK55109	
		KRT8	P05787	
		KRTHB4	NP_149034	
		KRTHB5	NP_002274	
IFY1	CAB75944	KRT13	NP_705694	2
		KRTH3A3	NP_004129	
INS (insulin peptide)	A38422	IGF1	P05019	3
		IGF2	NP_000603	
		INS	AAA59179	
INSR (insulin receptor)	O02466	INSR	AAA59452	3
		IGF1R	NP_000866	
		INSRR	P14616	
Islet	AAF34717	ISL1	NP_002193	2
		ISL2	NP_665804	
Krox	AAL83211	EGR1	NP_001955	3
		EGR2	NP_000390	
		EGR3	NP_058782	
lamin	CAC13104	LMNA	NP_733821	3
		LMNB1	NP_005564	
		LMNB2	NP_116126	
MIIPS (myo-inositol 1-phosphate synthase Al)	AAL02140	ISYNA1	NP_057452	1
MRDH (microsomal retinol dehydrogenase)	AAG44849	RODH4	NP_003699	5
		SDR-O	NP_683695	
		RODH	NP_003716	
		RDH5	NP_002896	
		RDHh	AAD32458	
Neuro (neurogenin)	AAF81766	NEUROG1	NP_006152	3
		NEUROG2	AAG40770	
		NEUROG3	AAK15022	
NFI (nuclear factor I)	AAC25163	NFIA	Q12857	4
		NFIB	NP_005587	

TABLE 1—Continued

Cephalochordate gene(s)	Accession Number	Human gene(s)	Accession Number	Mammalian Orthologs
		NFIC	NP_005588	
		NFIX	NP_002492	
Nkx2.1	AAC35350	TITF1	NP_003308	2
		NKX2.4	AAG35617	
Nkx2.2	AAD01958	NKX2.2	NP_002500	1
Ntn (netrin)	CAB72422	NTN1	NP_004813	1
PAH (phenylalanine hydroxylase)	CAA04917	PAH	NP_000268	1
Pax1	AAA81364	PAX1	PI 5863	2
		PAX9	NP_006185	
Pax2	AAC12734	PAX2	NP_003981	3
		PAX5	NP_057953	
		PAX8	Q06710	
Pax3/7	AF165886	PAX3	AAH08826	2
		PAX7	NP_039236	
Pax6	CAA11368	PAX6	AAA59962	1
PC6 (proprotein convertase)	Q9NJ15	PCSK5	Q92824	2
		PACE4	JC5570.	
PSEN (presenilin)	AAL40414	PSEN1	NP_000012	2
		PSEN2	NP_036618	
Rab GDP dissociation inhibitor	CAB46230	GDI1	NP_001484	2
		GDI2	NP_001485	
Shox	AAL83210	SHOX	NP_000442	2
		SHOX2	NP_006875	
snail	AAC35351.	SNAI1	NP_005976	2
		SNAI2	NP_003059.	
SPC2 (proprotein convertase)	AAA87005	PCSK2	AAA60032	1
SPC3 (proprotein convertase)	AAA87006	PCSK1	P29120	1
SOD	P28761	MnSOD	CAA42066	1
PTPN6 (protein tyrosine phosphatase N6)	BAA95174	PTPN6	NP_536859	2
		PTPN11	Q06124	
PTP10 (protein tyrosine phophatase receptor)	BAA95168	PTPRJ	JC5290	3
		PTPRK	Q15262	
		PTPRM	NP_002836	
RAR (retinoic acid receptor)	AAM46149	RARA	NP_000955	3
		RARB	NP_000956	
		RARG	NP_000957	
S6 (40S ribosomal protein S6)	O01727	RPS6	AAH13296	1
Tbx1/10	AAG34887	TBX1	NP_542377	2
		TBX10	O75333	
Tbx2/3	AAG34888	TBX2	NP_005985	2
		TBX3	NP_005987	
Tbx4/5	AAG34889	TBX4	P57082	2
		TBX5	AAC51644	
Tbx15/18/22	AAG34891	TBX15	CAC39400	2
		TBX18	O95935	
Eomes/Tbr1/Tbx21	AAG34893	EOMES	CAB37939	3
		TBR1	NP_006584	
		TBX21	NP_037483	
Tpm (tropomyosin)	BAA96548	TPM1	P09493	5
		TPM2	NP_003280	
		TPM3	P06753	
		TPM4	NP_003281	

TABLE 1—Continued

Cephalochordate gene(s)	Accession Number	Human gene(s)	Accession Number	Mammalian Orthologs
		TRK	CAA27243	
TNNC (troponin C)	BAA13732	TNNC1	NP_003271	2
TNNCX2	JW0060	TNNC2	NP_003270	
TNNI (troponin I)	BAA96549	TNNI1	NP_003272	3
		TNNI2	NP_003273	
		TNNI3	P19429	
TOB (transducer of ERBB2)	AAB53747	TOB	BAA10971	2
		TOB1	NP_005740	
TPI (triose phosphate isomerase)	BAA22631	TPI	NP_000356	1
TR2/4 (orphan receptor)	AAM46150	TR2	NP_003288	2
		TR4	NP_003289	
twist	AAD10038	TWIST1	NP_000465	2
		TWIST2	AAH17907	
Wnt1	AAC80432	WNT1	NP_005421	1
Wnt3	AAL37555	WNT3	A47536	2
		WNT3a	NP_149122	
Wnt4	AAC80431	WNT4	NP_110388	1
Wnt5	AAL37556	WNT5a	NP_003383	2
		WNT5b	NP_110402	
Wnt6/WntB	CAA84028	WNT6	Q9Y6F9	1
Wnt7	AAC80433	WNT7a	BAA82509	1
Wnt8	AAF80559	WNT8b	NP_003384	2
		WNT8d	NP_114139	
Wnt10	AAL37558	WNT10b	NP_003385	2
		WNT10a	NP_079492	
Wnt11	AAF80555	WNT11	NP_004617	1
Zic	CAB96573	ZIC1	NP_003403	5
		ZIC2	AAC96325	
		ZIC3	NP_003404	
		ZIC4	NP_115529	
		ZIC5	NP_149123	

[1] In the interest of space constraint, access to the original publications reporting these genes may be obtained through the accession numbers provided.

[2] Whenever possible, human genes have been identified in accordance with the conventions of the Human Gene Nomenclature Committee (http://www.gene.ucl.ac.uk/nomenclature/).

TABLE 2. Published orthology groups supported by at least 2 of 3 phylogenetic methods upon reanalysis or linkage data

Amphioxus gene	Mammalian orthologs	Reference
Batl (AAM18861)	2	Abi-Rached et al., 2002
Brd2/3/4/T (AAM18883)	4	Abi-Rached et al., 2002
C2orf9 (AAM18883)	1	Abi-Rached et al., 2002
C3/C4/C5 (AAM18874)	3	Abi-Rached et al., 2002
C9orfl8 (AAM18893)	1	Abi-Rached et al., 2002
CACNA1A/B/E (AAM18875)	3	Abi-Rached et al., 2002
Cdx	3	Brooke, Garcia-Fernandez, and Holland, 1998
Dll	6 (3 tandem pairs)	Pollard and Holland, 2000
Emx	2	Williams and Holland, 2000
En	2	Pollard and Holland, 2000
Evx	2	Ferrier et al., 2001b
FGFR	4	Suga et al., 1999
Gpr54 (AAM 18884)	1	Abi-Rached et al., 2002
Gpr107/108 (AAM18888)	2	Abi-Rached et al., 2002
Gsx	2	Pollard and Holland, 2000
Hh	3	Shimeld, 1999

TABLE 2—Continued

Amphioxus gene	Mammalian orthologs	Reference
Hox1	3	Brooke, Garcia-Fernandez, and Holland, 1998
Hox2	2	Brooke, Garcia-Fernandez, and Holland, 1998
Hox3	3	Brooke, Garcia-Fernandez, and Holland, 1998
Hox4	4	Brooke, Garcia-Fernandez, and Holland, 1998
Hox5	3	Brooke, Garcia-Fernandez, and Holland, 1998
Hox6	3	Brooke, Garcia-Fernandez, and Holland, 1998
Hox7	2	Brooke, Garcia-Fernandez, and Holland, 1998
Hox8	3	Brooke, Garcia-Fernandez, and Holland, 1998
Hox9	4	Brooke, Garcia-Fernandez, and Holland, 1998
Hox10	3	Brooke, Garcia-Fernandez, and Holland, 1998
Hox11	3	Brooke, Garcia-Fernandez, and Holland, 1998
Hox12	2	Brooke, Garcia-Fernandez, and Holland, 1998
Hox13	4	Brooke, Garcia-Fernandez, and Holland, 1998
HRASLS (AAM18866)	4	Abi-Rached et al., 2002
Mdh (AAM18871)	1	Abi-Rached et al., 2002
MKI67IP (AAM18872)	1	Abi-Rached et al., 2002
Mnx	2	Ferrier et al., 200la
Mox	2	Pollard and Holland, 2000
MSL3L (AAM18870)	1	Abi-Rached et al., 2002
Msx	3	Pollard and Holland, 2000; Furlong and Holland, 2002
MTAP44 (AAM18895)	2	Abi-Rached et al., 2002
NEK6/7 (AAM18889)	2	Abi-Rached et al., 2002
NEU1 (AAM18894)	1	Abi-Rached et al., 2002
Notch	4	Abi-Rached et al., 2002
Otx	2	Williams and Holland, 1998
PBX1/2/3/4 (AAM18882)	4	Abi-Rached et al., 2002
Pitx	3	Boorman and Shimeld, 2002
PKD (AAM18864)	2	Abi-Rached et al., 2002
PRPF4 (AAM18877)	1	Abi-Rached et al., 2002
PSMB5/8 (AAM18885)	2	Abi-Rached et al., 2002
PSMB 7/10 (AAM18890)	2	Abi-Rached et al., 2002
PTGES2 (AAM18863)	1	Abi-Rached et al., 2002
PTPN3	2	Ono-Koyanagi et al., 2000
PTPR2A	3	Ono-Koyanagi et al., 2000
PTPR4(a,b,c)	2	Ono-Koyanagi et al., 2000
PTPR5	2	Ono-Koyanagi et al., 2000
RXRA/B/G	3	Abi-Rached et al., 2002
SIAT8 (AAM18873)	5	Abi-Rached et al., 2002
src	4	Suga et al., 1999
TLR (AAM18891)	4	Abi-Rached et al., 2002
TYR (AAM18867)	3	Abi-Rached et al., 2002
UGT (AAM18900)	7	Abi-Rached et al., 2002
VEGFR	3	Suga et al., 1999
WDR5	2	Abi-Rached et al., 2002
Xlox	1	Brooke, Garcia-Fernandez, and Holland, 1998

relationships ($>70\%$ support of critical nodes by at least two of three methods). These families include 73 whose results are either consistent with previously published work or provide updated ratios in light of the availability of the complete human genome dataset (Table 1), and 61 reported in previously published trees which, upon reanalysis, either met our criteria or required phylogenomic (mapping and linkage) data to determine orthology/paralogy relationships (Table 2). All sequence accession numbers, alignments, and phylogenetic trees are available from our website at http://biosgi.wustl.edu/gibsonbrown/curated/index.html.

To determine the relative time when gene duplications, if any, occurred, we included as many sequences as possible from early-diverging vertebrate species. The trees presented do not show all sequences available because intermediately-diverging sequences that are not phylogenetically informative, and early-diverging sequences that are incomplete, were excluded. Additionally, highly divergent sequences from outgroups (commonly *Caenorhabditis* and *Ciona* spp.) had to be removed, as it is frequently impossible to align sufficient sites within these

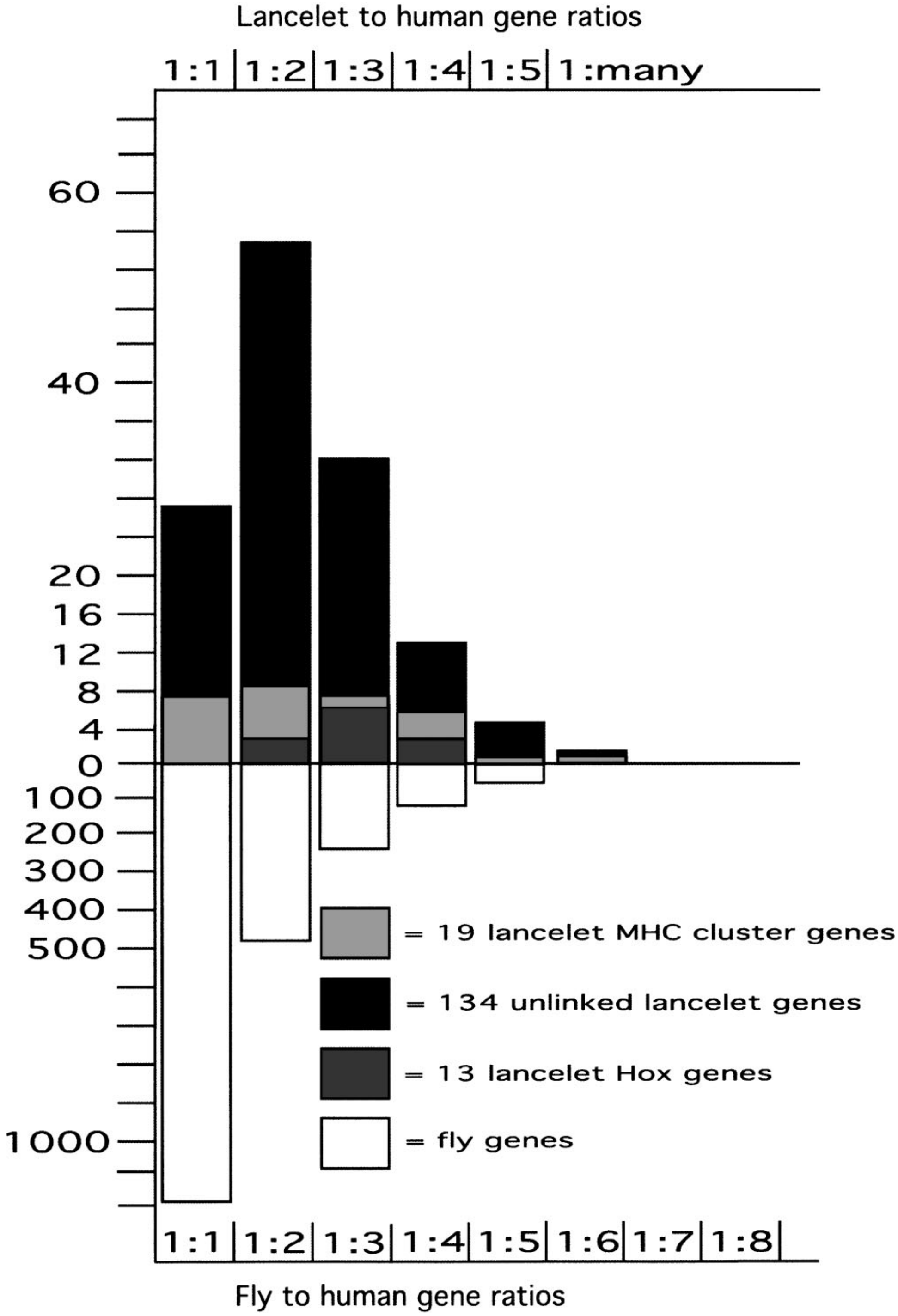

Fig. 1. Summary of gene ratios. Vertical bars represent number of occurrences for each gene ratio. Black, cephalochordate-to-mammalian gene ratios for 134 unlinked amphioxus genes. Dark grey, Hox genes, and light grey, MHC-homologous region genes, form a subset of the genes included in this study. White, Drosophila-to-human gene ratios as reported by Friedman and Hughes (2001).

sequences with any degree of confidence. We determined orthology assignments between amphioxus and the inferred common ancestors of gnathostomes, tetrapods, or mammals by accepting only nodes well supported by at least two of three phylogenetic methods. Computer-predicted genes, which often contain splicing errors (resulting in missing or incorrect exons), as well as short PCR-generated sequences from hagfish, zebrafish, frog and chicken, either reduce the amount of sequence usable in an alignment, or do not support nodes predicted by the established species tree. Such sequences were excluded from final tree reconstructions. In certain gene families orthology assignments remained unchanged from those reported in previous publications for several reasons. These groups, their citations, and the inferred amphioxus-to-human gene ratios, are listed in Table 2. In some cases, such as the *engrailed* and *hedgehog* families, we found no new human orthologs in the databases, and newer sequences from other organisms did not affect amphioxus-to-human orthology interpretations. Other gene families, such as *Hox*, contained insufficient alignable sequence to provide reliable support for any informative tree topology. In such cases, mapping and linkage (i.e., phylogenomic) data were also used in these publications to determine gene relationships.

Our orthology assignments generate a very different distribution of invertebrate-to-vertebrate gene ratios than comparative studies of either amphioxus versus vertebrate *Hox* genes or large numbers of unlinked genes in protostomes versus humans (Fig. 1). The median ratio of cephalochordate-to-vertebrate gene ratios equals 1:2. Although this result is consistent with two whole-genome duplications followed by extensive gene loss, a 1:2 median ratio is also consistent with a single whole-genome duplication, multiple local duplications, or a combination of any of the above. Despite recently collecting a very similar data set to ours, Furlong and Holland (2002) have interpreted their results as strong evidence in support of the 2R hypothesis, a conclusion we believe to be neither substantiated nor refuted by current data.

Within the subset of trees showing a 1:4 relationship, a further analysis is possible (Hughes, '99, Larhammar et al., 2002). Trees with a topology of chordate genes in the form ((AB)(CD)) support two sequential whole-genome duplications, whereas trees with topologies which are some variant of (A(B(CD))) do not. The former should be significantly more abundant than the latter if the 2R hypothesis is correct, although Furlong and Holland (2002) have raised a plausible objection to this prediction if the proposed genome duplications occurred in relatively rapid succession. Our dataset did not include a sufficient number of gene families with four vertebrate members (N=13, of which 3 are *Hox* cluster genes and 5 are linked within the MHC-homologous region) to test this hypothesis. However, 57 phylogenies using protostome genes as outgroups only support the former topology over the latter approximately 25% of the time, a frequency not significantly different from a random distribution (Lander et al., 2001).

DISCUSSION

In fly-to-human gene comparisons, the median orthologous gene ratio equals 1:1 (Friedman and Hughes, 2001; Fig. 1). In contrast, our analysis of cephalochordate-to-human gene comparisons reveals a median gene ratio of 1:2 (Fig. 1). One possible reason for the difference between the fruit fly and cephalochordate results may be due to the way in which the genes were sampled. In the case of amphioxus, most genes were isolated and sequenced because of their involvement in the development of other animals. It has been suggested that developmental genes, with their complex spatiotemporal regulation, are more likely to have separable regulatory modules, and are therefore more likely to be fixed by subfunctionalization following duplication (Force et al., '99). In our study only 30 amphioxus genes could be denoted ‘‘metabolic.'' Of these, 13 possess only one human ortholog. Eleven possess two human orthologs. While the number of genes in this category is too small to state definitively that different rates of fixation following duplication occur for metabolic and developmental signaling genes, this observation may support such a trend and warrants further study. The cephalochordate-to-vertebrate gene ratio distribution may therefore be biased toward overestimating high gene ratios, and a more complete sampling of the amphioxus genome could reveal a distribution with an even stronger trend toward low gene ratios.

Our results with a large number of unlinked amphioxus genes are also different from the *Hox* gene ratios (median ratio=1:3, Fig. 1). We suspect this is because *Hox* genes are unlikely to be representative of gene families as a whole. In addition to being linked, and thus revealing the history of only one small region of the genome, the coordinate regulation of these genes is likely to cause unusual selection pressure such that the rate of *Hox* gene loss following cluster duplication may be lower than that for unlinked, independently regulated genes. We suspect that a larger sampling of amphioxus genes might lead to a general distribution of gene ratios even more skewed to the left from that of the *Hox* cluster and genes linked to it (Fig. 1).

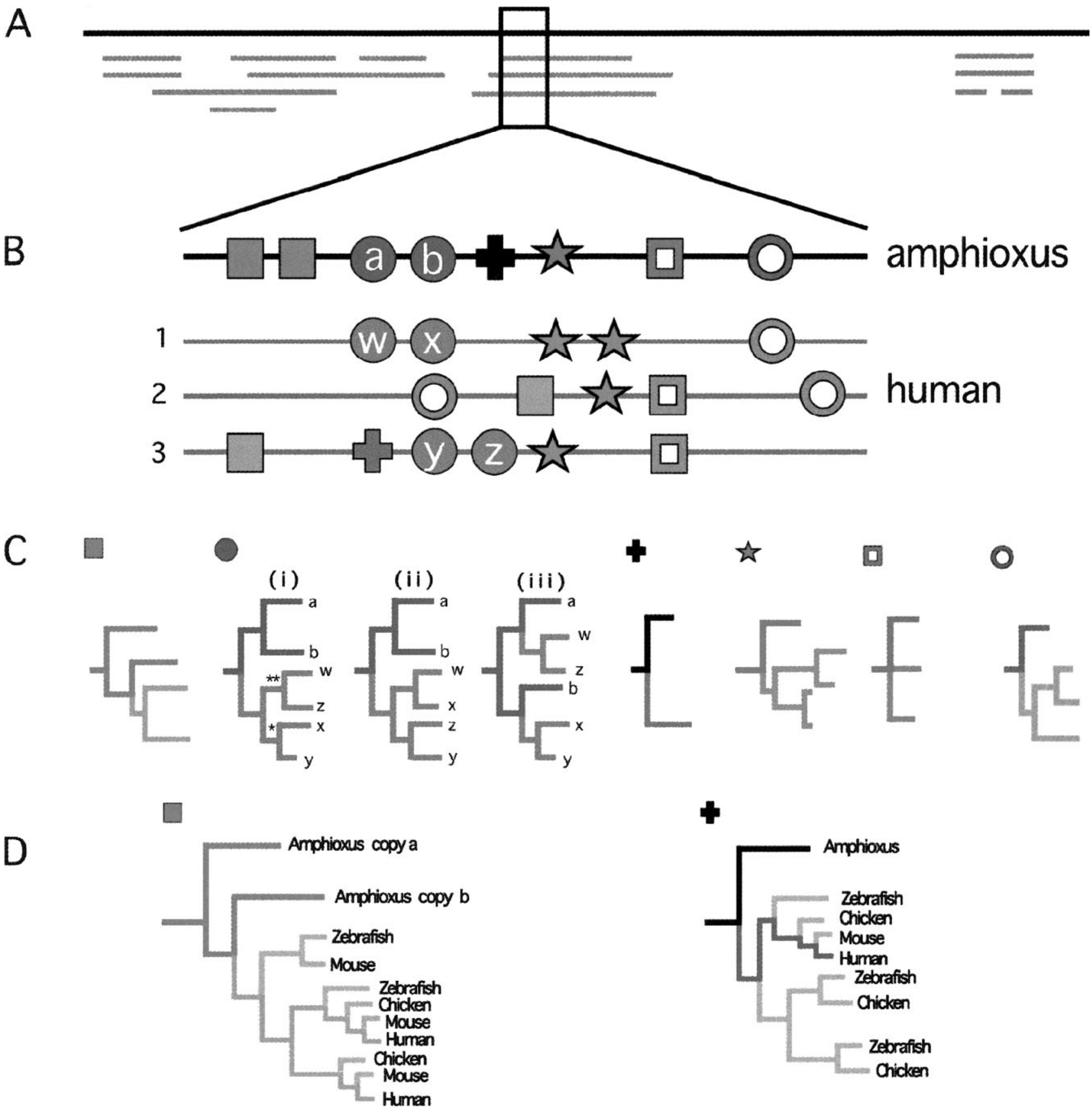

Fig. 2. Schematic diagram representing the phylogenomic approach we believe will be required if the issue of whole-genome duplication within chordates is to be resolved. **A**, A hypothetical amphioxus (black) – human (grey) synteny map reconstructed by aligning clusters of paralogy groups as shown in **B**. **C**, Phylogenetic trees of paralog groups. Dark colored terminal branches represent amphioxus genes; lighter colored branches represent human genes. Trees support different duplication histories for different paralog groups. **D**, Phylogenetic trees containing orthologs from *additional* vertebrate species; grey lines, not only serve to date duplications, but also allow the detection of lineage-specific gene death that might otherwise result in misleading evolutionary inferences.

We reemphasize the importance of the cephalochordate, amphioxus, for understanding chordate gene function and evolution (Ruvinsky et al., 2000; Gibson-Brown et al., 2003). Protostome or urochordate genomes do not serve as the best outgroups for determining the basal state of the vertebrate genome because the former are very early diverging, and the latter are highly derived (Dehal et al., 2002; Holland and Gibson-Brown, 2003; Gibson-Brown et al., 2003). In contrast, amphioxus represents the sister group to vertebrates within the deuterostomes. It is more closely related to vertebrates than echinoderms, hemichordates, or urochordates, yet its separation from vertebrates predates the proliferation of genes early in the history of the chordate lineage. Cyclostomes (hagfish and lampreys) and chondricthyans (sharks, rays, and chimeras) are also of great interest for the study of chordate genome evolution, but divergence of these groups appears to postdate much of the early chordate gene proliferation, so these groups do not reveal the basal condition of the chordate genome (Kim et al., 2000; Neidert et al., 2001; Escriva et al., 2002). The complete sequence of an amphioxus genome will not only provide gene sequences for phylogenetic and developmental studies, but also the linkage information vital to the phylogenomic approach discussed below. Without this genome sequence, it will be impossible to reconstruct the early genomic events underlying the rapid genetic changes that led to the evolutionary successes of vertebrates.

We conclude that future attempts to resolve the issue of genome duplication(s) during chordate evolution will require the use of methods complementary to phylogenetics, despite continued attempts to rely on this method alone (e.g., Furlong and Holland, 2002; Gu et al., 2002). One possibility is to take advantage of positional information contained in completely sequenced genomes, as recently undertaken (McLysaght et al., 2002, Friedman and Hughes, 2003). Paralogous clusters of genes (i.e., regions of conserved synteny or ‘‘paralogons’’) can be used to determine the history of chromosomal regions as shown in Figure 2A. By building phylogenetic trees for each paralog group within a cluster, a tree for the entire cluster can be inferred from nodes supported in a statistically significant majority of individual trees (i.e., generating Fig. 2A from the trees in Fig. 2C.) Such an analysis of one small region of the amphioxus genome was recently reported (Abi-Rached et al., 2002).

We certainly do not intend to imply that such a task will be trivial, far from it. Numerous difficulties will impede an accurate reconstruction of the state of the ancestral vertebrate genome. The use of multiple paralogy groups is a necessary part of future analyses however, because it can reveal instances of gene death within a cluster which would otherwise be masked and lead to incorrect inferences regarding the number of gene duplication events (Fig. 2B); up to 80% of the duplicate genes may have been lost following a teleost-specific genome duplication (Postlethwaite et al., 2000), and based on the examination of gene duplicates in nine divergent taxa, most of this loss is predicted to happen quite rapidly, within 10 million years of the duplication (Lynch and Conery, 2000). The inclusion of additional species in phylogenetic reconstructions can reveal duplicate losses which have occurred after longer intervals, as well as providing a more reliable tool than molecular clocks for inferring duplication dates (Fig. 2D). Transpositions and intrachromosomal inversions are also very common, complicating paralogon reconstructions (Ruvinsky and Silver, 1998, Postlethwaite et al., 2000.) Gene loss or gain (by tandem duplications) in one or more lineages further confuses orthology assignments. For example, the genes depicted as blue circles in Fig. 2B may have resulted from one of several different duplication histories as depicted in Fig. 2C (i) – (iii). Naturally, weak support for key nodes will make distinguishing between duplication scenarios difficult; for example, weak support for nodes * and ** (Fig. 2C) will make any choice between scenarios (i) and (ii) questionable. Amphioxus lineage-specific duplications demonstrate the particular importance of positional data in determining gene histories when support from phylogenetic analyses is weak (Holland et al., 1995; Minguillon et al., 2002).

Additionally, more than 600 million years of independent evolution separate the genomes of humans and amphioxus. The possibility that sites in protein sequences have changed multiple times in this long interval may result in misleading homoplasies, just as third-position saturation in DNA sequences does on much shorter timescales. Also, different selection pressures may have caused very different patterns of sequence evolution in one or another lineage. Moreover, uniform selection pressures between gene duplicates do not necessarily imply uniform selection pressures in different protein domains; different rates or types (balancing versus directional) of evolution in different protein domains lower the reliability of nodes in phylogenetic trees based upon complete protein sequences. Additionally, after duplication, sequence similarity between paralogs may allow for gene conversion events, homogenizing gene sequences, and causing misleadingly recent divergence times to appear in tree reconstructions. If such an error occurs in a large fraction of gene families within a syntenic region, a plausible event given the molecular mechanism of gene conversion, an inaccurate estimate of cluster age will result.

In summary, it is quite possible, due to the large number of potential complications, that even a ''phylogenomic'' approach will fail to support one single model over other possible chromosomal and/or genomic duplication scenarios, but we conclude that the question of pattern and process in early chordate genome evolution will most certainly *not* be resolved without incorporating such an approach.

ACKNOWLEDGMENTS

We thank K. Nicholson for insightful comments into phylogenetic methods and statistical tests during the tree-reconstruction portion of this project. We also thank M. Bezanilla and the two anonymous reviewers for helpful comments that greatly improved the manuscript.

LITERATURE CITED

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

Altschul SF, Madden TL, Schaeffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acids Res 25:3389–3402.

Bailey WJ, Kim J, Wagner GP, Ruddle FH. 1997. Phylogenetic reconstruction of vertebrate Hox cluster duplications. Mol Biol Evol 14:843–853.

Boorman CJ, Shimeld SM. 2002. *Pitx* homeobox genes in *Ciona* and amphioxus show left-right asymmetry is a conserved chordate character and define the ascidian adenohypophysis. Evol Dev 4:354–365.

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

Dehal P, Satou Y, Campbell RK, et al. (87 co-authors). 2002. The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. Science 298:2157–2167.

Escriva H, Mazon L, Youson J, Laudet V. 2002. Analysis of lamprey and hagfish genes reveals a complex history of gene duplications during early vertebrate evolution. Mol Biol Evol 19:1440–1450.

Ferrier DEK, Brooke NM, Panopoulou G, Holland PWH. 2001a. The *Mnx* homeobox gene class defined by *HB9*, *MNR2* and amphioxus *AmphiMnx*. Dev Genes Evol 211:103–107.

Ferrier DEK, Minguillon C, Cebrian C, Garcia-Fernandez J. 2001b. Amphioxus *Evx* genes: implications for the evolution of the midbrain-hindbrain boundary and the chordate tailbud. Dev Biol 237:270–281.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwaite J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545.

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.

Furlong RF, Holland PWH. 2002. Were vertebrates octaploid? Phil Trans Roy Soc Lond B 357:531–544.

Gibson-Brown JJ, Osoegawa K, de Jong PT, McPherson JB, Waterston RH, Holland LZ. 2003. Proposal to sequence the amphioxus genome. J Exp Zool (Mol Dev Evol): in press.

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

Holland LZ, Gibson-Brown JJ. 2003. The *Ciona intestinalis* genome: when the constraints are off. BioEssays 25:529–532.

Holland PWH, Garcia-Fernandez J, Williams NA, Sidow A. 1994. Gene duplications and the origins of vertebrate development. Development (Suppl) S125–133.

Holland PWH, Korschoz B, Holland LZ, Herrmann BG. 1995. Conservation of *Brachyury* (*T*) genes in amphioxus and vertebrates: developmental and evolutionary implications. Development 121:4283–4291.

Hughes AL. 1999. Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history. J Mol Evol 48:565–576.

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

Karabinos A, Bhattacharya D. 2000. Molecular evolution of calmodulin and calmodulin-like genes in the cephalochordate *Branchiostoma*. J Mol Evol 51:141–148.

Keller MJ, Gerhardt HC. 2001. Polyploidy alters advertisement call structure in gray treefrogs. Proc Roy Soc Lond B 268:341–345.

Kim CB, Amemiya C, Bailey W, Kawasaki K, Mezey J, Miller W, Minoshima S, Shimizu N, Wagner G, Ruddle F. 2000. Hox cluster genomics in the horn shark, *Heterodontus francisci*. Proc Natl Acad Sci USA 97:1655–1660.

Krumlauf R. 1994. Hox genes in vertebrate development. Cell 78:191–201.

Ku H, Vision T, Liu J, Tanksley SD. 2000. Comparing sequenced segments of the tomato and *Arabidopsis* genomes: large-scale duplication followed by selective gene loss creates a network of synteny. Proc Natl Acad Sci USA 97:9121–9126.

Lander ES, Linton LM, Birren B, et al. (243 co-authors) 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.

Larhammar D, Lundin L, Hallböök F. 2002. The human Hox-bearing chromosomes regions did arise by block or chromosome (or even genome) duplications. Genome Res. 12: 1910–1920.

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

Martin AP. 1999. Increasing genomic complexity by gene duplication and the origin of vertebrates. Am Nat 154: 111–128.

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

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

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

Neidert AH, Virupannavar V, Hooker GW, Langeland JA. 2001. Lamprey *Dlx* genes and early vertebrate evolution. Proc Natl Acad Sci USA 98:1665–1670.

Ohno S. 1970. Evolution by gene duplication. Springer-Verlag, New York.

Ono-Koyanagi K, Suga H, Katoh K, Miyata T. 2000. Protein tyrosine phosphatases from amphioxus, hagfish, and ray: Divergence of tissue-specific isoform genes in the early evolution of vertebrates. J Mol Evol 50:302–311.

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

Postlethwaite JH, Woods IG, Ngo-Hazelett P, Yan Y, Kelly PD, Chu F, Huang H, Hill-Force A, Talbot WS. 2000. Zebrafish comparative genomics and the origins of vertebrate chromosomes. Genome Res 10:1890–1902.

Robinson-Rechavi M, Marchand O, Escriva H, Bardet PL, Zelus D, Hughes S, Laudet V. 2001. Euteleost fish genomes are characterized by expansion of gene familes. Genome Res 11:781–788.

Ruvinsky I, Silver LM. 1997. Newly identified paralogous groups on mouse chromosomes 5 and 11 reveal the age of a T-box cluster duplication. Genomics 40:262–266.

Ruvinsky I, Silver LM, Gibson-Brown JJ. 2000. Phylogenetic analysis of T-box genes demonstrates the importance of amphioxus for understanding evolution of the vertebrate genome. Genetics 156:1249–1257.

Schubert M, Holland LZ, Holland ND, Jacobs DK. 2000. A phylogenetic tree of the *Wnt* genes based on all available full-length sequences, including five from the cephalochordate amphioxus. Mol Biol Evol 17:1896–1903.

Shimeld SM. 1999. The evolution of the hedgehog gene family in chordates: insights from amphioxus hedgehog. Dev Genes Evol 209:40–47.

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 CR, Hurst LD. 1999. Vertebrate genome evolution: a slow shuffle or a big bang? BioEssays 21: 697–703.

Spring J. 1997. Vertebrate evolution by interspecific hybridization: are we polyploid? FEBS Lett 400:2–8.

Strimmer K, von Haeseler A. 1996. Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. Mol Biol Evol 13:964–969.

Suga H, Hoshiyama D, Kuraku S, Katoh K, Kubokawa K, Miyata T. 1999. Protein tyrosine kinase cDNAs from amphioxus, hagfish, and lamprey: isoform duplications around the divergence of cyclostomes and gnathostomes. J Mol Evol 49:601–608.

Swofford DL. 2001. PAUP* 4.0 beta 5: Phylogenetic Analysis Using Parsimony and Other Methods. Sinauer Associates, Sunderland, MA.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl Acids Res 22: 4673–4680.

Venter JC, Adams MD, Myers EW, et al. (274 co-authors) 2001. The sequence of the human genome. Science 291:1304–1351.

Wada H, Satoh N. 1994. Details of the evolutionary history from invertebrates to vertebrates as deduced from the sequences of 18S rDNA. Proc Natl Acad Sci USA 91: 1801–1804.

Williams NA, Holland PWH. 1998. Gene and domain duplication in the chordate *Otx* gene family: insights from amphioxus *Otx*. Mol Biol Evol 15:600–607.

Williams NA, Holland PWH. 2000. An amphioxus *Emx* homeobox gene reveals duplication during vertebrate evolution. Mol Biol Evol 17:1520–1528.