FEBS Letters 400 (1997) 2–8 FEBS 17957

Hypothesis

Vertebrate evolution by interspecific hybridisation – are we polyploid?

Jürg Spring*

Institute of Zoology, University of Basel, Rheinsprung 9, CH-4051 Basel, Switzerland

Received 19 September 1996; revised version received 7 November 1996

Abstract For the growing fraction of human genes with identified functions there are often homologues known from invertebrates such as *Drosophila*. A survey of well established gene families from aldolases to zinc finger transcription factors reveals that usually a single invertebrate gene corresponds to up to four equally related vertebrate genes on different chromosomes. This pattern was before widely noticed for the *Hox* gene clusters but appears to be more general. Genome quadruplication by two rounds of hybridisation is discussed as a simple biological mechanism that could have provided the necessary raw material for the success of vertebrate evolution.

Key words: Genome duplication; Gene family; Homology; Invertebrate; Allopolyploidy

1. Introduction

It has been widely publicised that the homeobox genes corresponding to the homeotic complex *HOM-C* from *Drosophila* occur in a cluster in invertebrates from cnidarians [1] and *Caenorhabditis elegans* [2] to amphioxus [3], while they are found as four so-called paralogous *Hox* clusters on four different chromosomes in higher vertebrates [4–9]. The two rounds of duplication of the *Hox* cluster probably occurred close to the origin of vertebrates [10]. In addition, unrelated genes that code for as functionally diverse proteins such as keratins, collagens or EGF-receptor-like tyrosine kinases are linked to the *Hox* clusters and are also duplicated [11]. A similar relationship has also been shown for the *syndecan* and *myc* gene families: a single invertebrate gene is found to be equally similar to the four vertebrate genes of its group which are linked each to a member of the other group on four different mouse chromosomes [12]. In fact, several extensive paralogous genomic regions containing gene families with various functions have been reviewed for mouse and man [13,14] and Ohno [15] had already elaborated the theory of evolution by gene duplication by 1970. Polyploidy had been discussed there as one of several possibilities for vertebrate gene family complexities and appears to have become an acceptable working hypothesis [16]. The new ideas proposed here are that allopolyploidy by interspecific hybridisation would create more evolutionary potential than autopolyploidy and that an invertebrate gene and the corresponding multiple vertebrate members of gene families should be considered as a group. This allows one even to make predictions for the number and positions of homologues in the human genome from model organisms such as *Drosophila* or *C. elegans*. Of course, *Hox* genes are still special because of their transcription along the body axes according to the position of the genes within the cluster, which inspired the concept of the zootype [17]. However, the one to four relationship of invertebrate and vertebrate genes is not specific for *Hox* genes, but rather appears to be the normal case for well studied gene families.

2. Homologues, orthologues, paralogues and tetralogues

Homologous genes are all those that are derived from a common ancestor by duplication and divergence. Orthologues are equivalent genes of different species, e.g. human *HOXA4* and murine *HoxA4*. Paralogues, in contrast, are homologous genes within one species. After tandem duplication such genes could be called *cis*-paralogues. However, this is only useful as long as they stay together such as in the case of *HOXA4* and *HOXA5*. To distinguish *trans*-paralogues such as *HOXA4, HOXB4, HOXC4* and *HOXD4* from all other homologues and to make a connection to invertebrate orthologues such as *Drosophila Dfd* or amphioxus *Amphihox4*, I propose the term tetralogues. Tetralogous genes are groups of quadruplicated vertebrate genes at four different chromosomal localisations corresponding to a single invertebrate gene which are all more similar to each other than to members of other tetralogy groups (Fig. 1). The ubiquity of this one to four relationship of invertebrate and vertebrate gene subfamilies suggests two genome-wide tetraploidisation events as the source for tetralogues.

3. How many tetralogues?

While for many gene families only three and not four tetralogues are presently known in vertebrates, closer inspection of the four *Hox* gene clusters revealed that in most *Hox* gene tetralogy groups only three members are really maintained in the human genome as well. Only 2 groups consist of all four genes, 8 out of 13 groups have three and 3 groups have only two genes left (Fig. 2A). Also, corresponding linked genes coding for keratins, collagens or tyrosine kinases show a similar pattern [11]. A comparable analysis of the MHC class III region illustrates that also here an average of three vertebrate tetralogues and one invertebrate gene can be found for various gene families belonging to unrelated functional groups (Fig. 2B). The MHC class III region on human chromosome 6p21.3 is one of the best documented portions of the human genome that contains more than 30 genes located between the MHC class I and II clusters [18]. Much less is known about many of these genes than about the *Hox* clusters, and some gaps in this table might still be filled. However, the chance of finding three or four tetralogous genes or clusters of genes on different chromosomes is apparently no higher for regulatory

*Corresponding author. Fax: (41) (61) 267 34 57.
E-mail: spring@ubaclu.unibas.ch

0014-5793/97/$17.00 © 1997 Federation of European Biochemical Societies. All rights reserved.
PII S0014-5793(96)01351-8

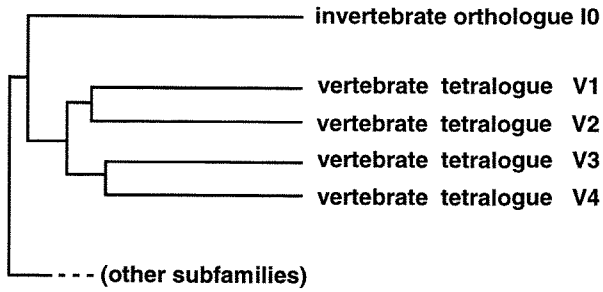

Fig. 1. Typical representation of the relationship within gene families where a single invertebrate gene corresponds to four vertebrate genes on four different chromosomes.

genes like the *Hox* or *myc* genes than for many other gene families with a wide variety of functions.

4. Tetralogues on all human chromosomes

Representatives of well studied gene families where one invertebrate gene is equally similar to three or four vertebrate tetralogues can be now found on all 23 human chromosomes (Table 1). For all these examples, linked genes that belong to independent tetralogy groups themselves are listed. When more than four related members of a gene family are known in a vertebrate, I found that they can be subdivided according to their sequence similarities, gene structures or chromosomal localisations into subgroups of up to four per corresponding invertebrate gene or into clusters of tandemly repeated genes. As an example, the recent cloning of a novel *src*-related gene from *Drosophila* [19] helped to divide this family with eight closely related human members into tetrapacks. The new *Drosophila* gene *Src41A* (*Dsrc41*) is most similar to the human subgroup with *SRC, YES1, FGR* and *FYN*. The previously known candidates for *Drosophila src* genes are *Src64B*, which might correspond to the human subfamily with *LCK, LYN, HCK* and *BLK*, while *Src29A* is clearly more similar to the group with Bruton's tyrosine kinase *BTK* which are only distantly related members of the non-receptor tyrosine kinases. 53 examples of tetralogy groups are listed in Table 1 and in a growing database on the World Wide Web, also including some interesting examples where the relationships could yet not be resolved completely or where homologous sequences are not yet available for *Drosophila*, such as for the *myc*, insulin or fibroblast growth factor families.

5. Genome quadruplication through hybridisation

The pattern of up to four vertebrate tetralogues for each invertebrate gene could provide us with new clues about the evolution of vertebrates and genomes in general. Many aspects of genome duplication had been discussed extensively [15] and even probabilities for finding existing patterns had been calculated [11,13]. However, the distribution of all these gene families in groups of two, three and apparently maximally four could simply suggest that all genes were first duplicated to the four-fold stage. Considering not only the statistics but also the biology of this problem, a single but simple mechanism that worked in many plants and invertebrates, and even in vertebrates like *Xenopus*, could explain the observed picture: allotetraploidisation. Two such rounds of interspecific hybridisations with the concomitant genome duplications of amphioxus-like animals could have created primitive vertebrates close to the Cambrian explosion 530 million years ago (Fig. 3). Hybridisation is not an efficient mode of evolution in higher vertebrates. It was therefore often generalised

A

				AbdB	abdA	Ubx	// Antp	Scr	Dfd		pb	lab	// Egfr	*Drosophila*
HOXA13		HOXA11	HOXA10	HOXA9		HOXA7	HOXA6	HOXA5	HOXA4	HOXA3	HOXA2	HOXA1	EGFR	Human 7p15-p12
				HOXB9	HOXB8	HOXB7	HOXB6	HOXB5	HOXB4	HOXB3	HOXB2	HOXB1	ERBB2	Human 17q11.2-q22
HOXC13	HOXC12	HOXC11	HOXC10	HOXC9	HOXC8		HOXC6	HOXC5	HOXC4				ERBB3	Human 12q12-q13
HOXD13	HOXD12	HOXD11	HOXD10	HOXD9	HOXD8				HOXD4	HOXD3		HOXD1	ERBB4	Human 2q31-34
				HOM/HOX-like homeoboxes									EGF-receptor-like tyrosine kinases	

B

?a	usp	//	Vav (Ce)	//	N	//	exd	//	Ten-m	//	?d	Abl	?e	?a	*Drosophila*
"HLA@II"	RXRB		?b		INT3		PBX2		TNXB1		C4A / B		TNFA / B	"HLA@I"	Human 6p21.3
?a	RXRA		VAV2		NOTCH1		PBX3		HBX		C5	ABL1	CD30LG	?a	Human 9q33-q34
?a	RXRG		?b		?c		PBX1		TNR			ABL2	TXGP1 / APT1LG1	CD1A/B/C/D/E	Human 1q22-q31
?a			VAV1		NOTCH3						C3		CD70	?a	Human 19p13
MHC class II cluster	nuclear receptors		vav-like oncogenes		notch-like receptors		pbx-like homeoboxes		tenascin-like ECM proteins		complement components	abl-like kinases	TNF-like cytokines	MHC class I related clusters	

Fig. 2. (A) *Hox* gene organisation in *Drosophila* and on four tetralogous human chromosomes with EGF-receptor-like tyrosine kinases as examples for unrelated linked genes. Although four clusters of *Hox* genes persist in vertebrates, only three genes were maintained on average from each tetralogy group. (B) Tetralogous display of MHC class III genes, the region between MHC class I and II genes on human chromosome 6p21.3 with known vertebrate and invertebrate homologues. An average of three tetralogous genes in humans and a single orthologue from invertebrates can be found for the better studied genes from the MHC class III region. (a) Although many invertebrate members of the immunoglobulin family are known, a clear candidate corresponding to MHC class I, II or CD1 molecules is still missing. (b) *VAV2* and a *Vav* homologue from *C. elegans* (GENBANK/EMBL U23520) were cloned only recently; a *Drosophila* homologue is still missing. (c) *NOTCH2* was mapped to 1p13-p11, which could indicate a recent inversion; *INT3* was also called '*NOTCH3*' and mapped to a contig with *PBX2* and *TNXB1* (tenascin-X) which is equally related to *HBX* (tenascin-C; *TNC*) as to *TNR* (tenascin-R) [29]. (d) An invertebrate homologue of *C3, C4* and *C5* could eventually be recognised in the course of the *C. elegans* sequencing project, but it might be difficult to recognise invertebrate members of the TNF family (e) as already the known vertebrate members have very little sequence similarity; tumour necrosis factor a (*TNFA*) is only 30% identical to the CD27 ligand (*CD70*) and the Fas ligand (*APT1LG1*) and OX40 ligand (*TXGP1*) are less than 20% identical. The other gene symbols are related to the common gene names and additional information is available on the World Wide Web in the genome data bases FLYBASE, MGD, GDB or OMIM; sequences were from SWISSPROT or translated from GENBANK/EMBL and analysed with BLAST, FASTA and PILEUP in GCG.

Table 1
Gene families with multiple human tetralogues for each *Drosophila* orthologue

Tetralogy groups	*D:H*	*Drosophila*	Human (mouse)		Tetralogous neighbours	
Abl (non-receptor tyrosine kinases)	1:2	Abl	ABL1 ABL2	9q34.1 1q24-q25	PBX3 PBX1	(homeobox transcription factors)
Aldolase (glycolysis enzymes)	1:3	Ald	ALDOA ALDOB ALDOC	16q22.2 9q22.3-q31 17cen-q12	HSD17B2 HSD17B3 HSD17B1	(hydroxysteroid dehydrogenases)
Alzheimer β-amyloid (cell surface protease inhibitors)	1:3	Appl	APP APLP1 APLP2	21q21.2 19q13.1 11q23-q25	ETS2 ETS1	(Ets domain transcription factors)
Ankyrin (membrane skeleton proteins)	1:3	Ank	ANK1 ANK2 ANK3	8p12-p11.2 4q25-q27 10q21	 NFKB1 NFKB2	(Ig-fold transcription factors)
BMP/dpp (TGFb-like growth factors)	1:2	dpp	BMP2 BMP4	20p12 14	CHGB CHGA	(secretogranins)
BMP/60A (TGFb-like growth factors)	1:4	Tgfbeta-60A	BMP5 BMP6 BMP7 BMP8	6(q12-q13) 6(p23-p22) 20 ?	 ID4 ID1	(inhibitory HLH factors)
Bruton's tyrosine kinase (non-receptor tyrosine kinases)	1:3	Src29A	BTK ITK TEC/TXK	Xq21.33-q22 5q31-q32 4p12	CDX4 CDX1	(homeobox transcription factors)
Cadherin (cell adhesion molecules)	1:3	Dec	CDH1/3/14 CDH2 CDH12	16q22.1 18q12.1 5p13-p14	MT3 MTL3	(metallothioneines)
Calmodulin (calcium-binding regulators)	1:3	Cam	CALM1 CALM2 CALM3	14q32 2p21 19q13.3	CKB CKM	(creatine kinases)
Caudal (homeobox transcription factors)	1:3	cad	CDX1 CDX3 CDX4	5q31-q33 13q12.3 Xq13.2	ITK BTK	(non-receptor tyrosine kinases)
Collagen type IV (network-forming collagens)	1:3	Cg25C/viking	COL4A1/2 COL4A3/4 COL4A5/6	13q34 2q35-q37 Xq22	 GPC1 GPC3	(PI-linked proteoglycans)
Cathepsin (cysteine proteases)	1:3	CysP-1	CTSL CTSS/K CTSH	9q22.1-q22.2 1q21 15q24-q25	NTRK2 NTRK1 NTRK3	(receptor tyrosine kinases)
Dlx (homeobox transcription factors)	1:3	dll	DLX1/2 DLX4 DLX5/6	2q32 ? 7q22	EN1 EN2	(homeobox transcription factors)
E2A (bHLH transcription factors)	1:3	da	TCF3 TCF4 TCF12	19p13.3 ? 15q21	INSR IGF1R	(receptor tyrosine kinases)
E2F (Rb-binding transcription factors)	1:3	E2f	E2F2 E2F3 E2F4	1p36 6p22 16q21-q22	ID3 ID4 	(inhibitory HLH factors)
EGF (epidermal growth factors)	2:6	spi grk	EGF TGFA HGL AREG/BTC DTR TDGF1	4q25 2p13 8p21-p12 4q13-q21 5q23 3p21.3-p21.1	FGF2 FGF5 FGF1 	(fibroblast growth factors)
EGF receptor (receptor tyrosine kinases)	1:4	Egfr	EGFR ERBB2 ERBB3 ERBB4	7p12 17q11.2-q12 12q13 2q34	HOXA@ HOXB@ HOXC@ HOXD@	(homeobox transcription factors)

Table 1 (*continued*)

Tetralogy groups	*D:H*	*Drosophila*	Human (mouse)		Tetralogous neighbours	
Egr/Krox-20 (zinc finger transcription factors)	1:4	sr	EGR1	5q23-31		
			EGR2	10q21.1	PLAU	(plasminogen activators)
			EGR3	8p23-p21	PLAT	
			EGR4	2p13		
Engrailed (homeobox transcription factors)	1:2	en/inv	EN1	2q13-q21	IHH	(secreted signalling factors)
			EN2	7q36	SHH	
Emx (homeobox transcription factors)	1:2	ems	EMX1	2p14-p13	REL	(Ig-fold transcription factors)
			EMX2	10q26.1	NFKB2	
Even skipped (homeobox transcription factors)	1:2	eve	EVX1	7p15-p14	HOXA@	(homeobox transcription factors)
			EVX2	2q34.3-q31	HOXD@	
Ezrin (peripheral cytoskeletal proteins)	1:3	Moe	VIL2	6q22-q27	ESR	(steroid hormone receptors)
			RDX	11q23	PGR	
			MSN	Xq11.2-q12	AR	
FGF receptor (receptor tyrosine kinases)	2:5	Fr1	FGFR1	8p12	EGR3	(zinc finger transcription factors)
		btl	FGFR2	10q25.3-q26	EGR2	
			FGFR3	4p16.3		
			FGFR4	5q33-qter	EGR1	
			FGFR6	?		
Gli (glioblastoma family zinc fingers)	1:3	ci	GLI	12q13	HOXC@	(homeobox transcription factors)
			GLI2	2	HOXD@	
			GLI3	7p13	HOXA@	
Glypican (PI-linked proteoglycans)	1:4	dally	GPC1	2q35-q37	COL4A3/4	(network-forming collagens)
			GPC2	?		
			GPC3	Xq26	COL4A5/6	
			GPC4	?		
Hedgehog (secreted signalling factors)	1:3	hh	SHH	7q36	COL1A2	(major fibril-forming collagens)
			DHH	(12q13)	COL2A1	
			IHH	2(q35-q36)	COL3A1	
Hox gene cluster (homeobox transcription factors)	1:4	'HOM-C'	HOXA@	7p15-p14	EGFR	(receptor tyrosine kinases)
			HOXB@	17q21-q22	ERBB2	
			HOXC@	12q12-q13	ERBB3	
			HOXD@	2q31	ERBB4	
Id (inhibitory HLH factors)	1:4	emc	ID1	20q11	SDC4	(cell surface proteoglycans)
			ID2	2p25	SDC1	
			ID3	1p36.13-p36.1	SDC3	
			ID4	6p22-p21.3		
Insulin receptor (receptor tyrosine kinases)	1:3	InR	INSR	19p13.3	MEF2B	(MADS box enhancer factors)
			INSRR	1	MEF2D	
			IGF1R	15q25-qter	MEF2A	
Integrin α-chain PS2 group (extracellular matrix receptors)	1:3	if	ITGA2B	17q21.32	HOXB@	(homeobox transcription factors)
			ITGA5/7	12q11-q13	HOXC@	
			ITGA4/V	2q31-q32	HOXD@	
Integrin β-chain (extracellular matrix receptors)	2:6	mys	ITGB3/4	17q11-qter	HOXB@	(homeobox transcription factors)
		betaIntn	ITGB6	2	HOXD@	
			ITGB7	12q13.1	HOXC@	
			ITGB1	10p11.2		
			ITGB2	21q22.3		
			ITGB5/8	?		
Jak (non-receptor tyrosine kinases)	1:4	hop	JAK1	1p32.3-p31.3	JUN	(bZIP transcription factors)
			JAK2	9p24		
			JAK3	?		
			TYK2	19p13.2	JUNB/D	
Laminin α-chain (extracellular matrix proteins)	1:3	LanA	LAMA1	18p11.31	YES1	(non-receptor tyrosine kinases)
			LAMA2/4	6q21-23	FYN	
			LAMA3	18q11.2		

Table 1 (*continued*)

Tetralogy groups	*D:H*	*Drosophila*	Human (mouse)		Tetralogous neighbours	
Laminin β-chain (extracellular matrix proteins)	1:3	LanB1	LAMB1 LAMB2 LAMB3	7q22 3p21.3-p21.2 1q32	BRAF RAF1	(serine/threonine kinases)
Mef2 (MADS box enhancing factors)	1:4	Mef2	MEF2A MEF2B MEF2C MEF2D	15q26 19p12 5q14 1q12-q23	IGF1R INSR INSRR	(receptor tyrosine kinases)
MyoD (bHLH transcription factors)	1:3	nau	MYOD1 MYOG MYF5/6	11p15.1 1q31-q41 12q21	INS/IGF2 IGF1	(insulin-like growth factors)
Myosin heavy chain (smooth/ non-muscle myosins)	1:3	zip	MYH9 MYH10 MYH11	22q12.3-q13.1 17p13 16p13.1	PRKM1 PRKM3	(MAP kinases)
NFkB/Rel/dorsal (Ig-fold transcription factors)	2:5	dl Dif	NFKB1 NFKB2 REL RELA RELB	4q24 10q24 2p13-p12 11q13 ?	FGF2 FGF8 FGF3/4	(fibroblast growth factors)
NOS (nitric oxide synthases)	1:3	Nos	NOS1 NOS2A/B/C NOS3	12q24 17q11-q12 7q35-q36	COL2A1 COL1A1 COL1A2	(major fibril-forming collagens)
Notch (cell-cell interaction receptors)	1:4	N	NOTCH1 NOTCH2 NOTCH3 INT3	9q34.3 1p13-p11 19p13.2-p13.1 6p21.3	COL5A1 COL11A1 COL11A2	(minor fibril-forming collagens)
Otx (homeobox transcription factors)	1:2	oc	OTX1 OTX2	2p13 14q21-q22	CALM2 CALM1	(calcium-binding regulators)
Pbx (homeobox transcription factors)	1:3	exd	PBX1 PBX2 PBX3	1q23 6p21.3 9q33-q34	RXRG RXRB RXRA	(nuclear receptors)
Raf (serine/threonine kinases)	1:3	phl	RAF1 ARAF1 BRAF	3p25 Xp11.3-p11.23 7q34	IL5RA IL3RA	(interleukin receptors)
Ral (GTP-binding oncogenes)	1:2	Rala	RALA RALB	7p 2cen-q13	HOXA@ HOXD@	(homeobox transcription factors)
Ras (GTP-binding oncogenes)	1:3	Ras85D	HRAS KRAS2 NRAS	11p15.5 12p12.1 1p13	BDNF NTF3 NGFB	(nerve growth factors)
Retinoblastoma (tumour suppressors)	1:3	Rbf	RB1 RBL1 RBL2	13q14.3 20q11.2 16q12.2	 MMP9 MMP2	(gelatinases)
Retinoic acid receptor type X (nuclear receptors)	1:3	usp	RXRA RXRB RXRG	9q34 6p21.3 1q22-q23	PBX3 PBX2 PBX1	(homeobox transcription factors)
Src (non-receptor tyrosine kinases)	1:4	'Src41A'	SRC YES1 FGR FYN	20q11.2 18p11.31-p11.22 1p36.2-p36.1 6q21	COL9A3 COL9A2 COL9A1	(type IX collagens)
Src-related (non-receptor tyrosine kinases)	1:4	Src64B	LCK LYN HCK BLK	1p35-p34.3 8q13 20q11-q12 8p23-p22	SDC3 SDC2 SDC4	(cell surface proteoglycans)
Stat (signal transducers and activators)	1:3	mrl	STAT1/4 STAT2/6 STAT3/5A/B	(2q12-q33) (12q13-q14.1) (17q11-q22)	HOXD@ HOXC@ HOXB@	(homeobox transcription factors)

Table 1 (*continued*)

Tetralogy groups	*D:H*	*Drosophila*	Human (mouse)		Tetralogous neighbours	
Syndecan (cell surface proteoglycans)	1:4	Syd	SDC1	2p(24-p23)	MYCN	(bHLH transcription factors)
			SDC2	8q22-q23	MYC	
			SDC3	(1p36-p32)	MYCL1	
			SDC4	20q12-q13		
Tenascin (extracellular matrix proteins)	1:3	Ten-m	HXB	9q32-q34	PBX3	(homeobox transcription factors)
			TNXB1	6p21.3	PBX2	
			TNR	1q25-q31	PBX1	
Wnt (wingless/int-1 signalling factors)	1:3	wg	WNT1	12q13	COL2A1	(major fibril-forming collagens)
			WNT2	7q31	COL1A2	
			WNT3	17q21-q22	COL1A1	

53 representative gene families are listed that include all 22 human autosomes and the X chromosome and a wide variety of functions. Only one example of linked tetralogous genes is shown per group. Most tetralogy groups are subfamilies of larger gene families. Additional members belong to an independent tetralogy group if duplication occurred before the divergence of the lineages leading to *Drosophila* and man such as in the case of *Src41A* [19], *Src64B* and *Src29A*. For the ratio *D:H* the number of *Drosophila* and human gene clusters rather than individual genes was used. Lineage specific tandem duplications appear to be common in vertebrates while *en/inv* is the only example in *Drosophila* listed here. Gene families with a ratio of 2:5 or 2:6 could not yet be resolved into tetralogy groups. Localisations shown in parentheses were predicted from mapping data in the mouse. Data were collected and analysed as described in Fig. 2, especially from FLYBASE and the human genome data base GDB. Additional information can be found in TetraBase, a continuously upgraded data base at the URL: http://www.unibas.ch/dib/zoologie/research/spring.html.

that in contrast to plants, hybridisation is not important for animals. However, in many invertebrates and even lower vertebrates such as fish and amphibians hybridisations are widespread. Immediately after speciation, hybridisation leading to allopolyploidy is not much different from autopolyploidisation and probably has few advantages, since one of the gene copies would continue to function while the other should accumulate mutations and disappear quickly [15]. Hybridisation in modern, highly adapted species has probably few advantages too and became rare in animals, possibly also due to the involvement of behaviour in species specific fertilisation mechanisms. Exceptions like *Xenopus*, salmon, trout or goldfish show that vertebrates can still undergo further polyploidisation, but additional constraints such as the increasing chromosome number might then become limiting. There could have been a very narrow hybridisation window when allopolyploidy really permitted evolutionary jumps through the combination of advantageous traits that had evolved previously in separate lineages. Candidates that resemble putative amphioxus-like founder species already lived in the Cambrian, for example *Pikaia gracilens* and *Yunnanozoon lividum* [20]. Modern hagfish and lampreys could be descendants of the proposed intermediate allotetraploids. As hagfish are so different from lampreys and all other, extinct jawless fish [21], they could be independently derived allotetraploids AB and AC or even CD (cf. Fig. 3). Alternatively, they might also be allohexaploids ABC or ABD, i.e. hybrids between an allotetraploid AB and a diploid C or D.

6. Partial redundancy of allooctoploids

If genome duplication was the result of hybridisation of rather different species, by allotetraploidisation, the faster evolving genes would already be quite different at the time of hybridisation and thus could serve as an only partially redundant pool for further divergent evolution of gene families. According to this idea, highly conserved genes are more likely to be perfectly redundant at the time of such hybridisation and are therefore more likely to be reduced to a single copy than rapidly diverging genes. Regulatory regions of genes can mutate even faster and with less constraints than coding regions and can thus lead to at least partial tissue specificity of expression of functionally still redundant genes. We have, for example, three calmodulin genes on three different chromosomes coding for identical proteins [22]. Could their survival be due to differences in their regulatory sequences, as suggested for three otherwise redundant paired-box containing genes in *Drosophila* [23]? Similarly, the homeobox gene *En-2* can rescue *En-1* knock-out mice when the *En-2* coding sequence is brought under the control of the regulatory sequences of its tetralogue *En-1* [24]. The close relationship, not only of the coding sequences, but also of

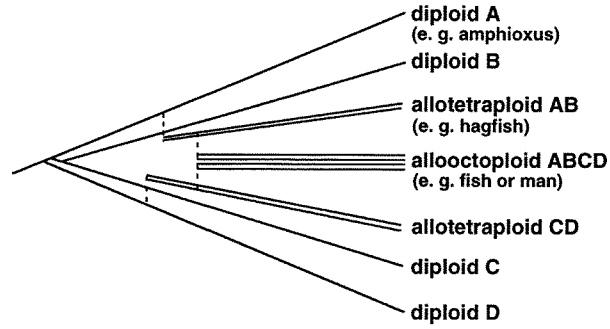

Fig. 3. A phylogenetic view of quadruplicated genome parts in vertebrate evolution. Hybridisation events are indicated by stippled lines connecting the involved lineages. Other scenarios can be imagined such as the formation of an allooctoploid ABA'B' from two diverged allotetraploids AB and A'B', respectively. Amphioxus is a good candidate for a direct descendant of a diploid ancestor. The jawless hagfish and lampreys might be allotetraploids while jawed vertebrates from fish to man would be allooctoploids. Around the so-called Cambrian explosion, 530 million years ago, hybridisation might have been common in little diverged ancestors of vertebrates. But immediately after speciation, hybridisation leading to allopolyploidy is not much different from autopolyploidisation and has few advantages. With increasing differentiation, the chance of diverged species producing successful hybrids is declining. Allopolyploidisation of closely related modern species or autopolyploidisation might still be possible but should have little evolutionary impact; tetraploid *Xenopus* still look like diploid or octoploid *Xenopus*. During a narrow hybridisation window allopolyploidy of rather primitive animals could have been more advantageous: allotetraploid lineages evolved and gave rise to an allooctoploid combining in a short period of time the advantages from previously separated lineages.

the regulatory sequences of tetralogous genes could also help to explain why so many of the knock-out mice have much milder phenotypes than expected from the expression patterns of the individually investigated genes. Therefore, tetralogues should be investigated simultaneously whenever possible.

7. Concluding remarks

Random gene, chromosome or genome duplications would be expected to result in complicated patterns of genome complexities. The simple one to four relationship observed for many invertebrate and vertebrate genes, developmental control genes as well as household enzymes or structural proteins, argues for unspecific quadruplication. A set of roughly 10 000 primitive metazoan genes is only slightly varied by tandem duplications or deletions within invertebrate genomes from worms to amphioxus; e.g. *C. elegans* has fewer genes than amphioxus in the *Hox* cluster and probably also in the whole genome. This set of primitive metazoan genes is represented up to four times on different vertebrate chromosomes or chromosomal regions, often with additional gene copies due to higher numbers of tandem duplications in vertebrates. More than 100 chromosomal rearrangements have visibly scrambled the genomes of the mouse and man since the divergence of their lineages about 70 million years ago [25]. In vertebrate evolution, this rate of rearrangements could still have left some genes next to each other purely by chance, without any functional implications. Conservation of gene linkage in *Drosophila* or *C. elegans* and vertebrates, however, could indeed point towards functional constraints [11]. Analysis of the genome of amphioxus, or even more conveniently an urochordate with a smaller genome, might combine the advantages of close relationship to vertebrates and a four-fold reduction of complexity as compared to vertebrate genomes. Similarly, the pufferfish (fugu) was chosen as a model vertebrate simply based on its small genome size of only 400 Mb [26], which is just four times the size of the *C. elegans* genome. Comparison of characteristic regions of model genomes from urochordates or amphioxus, jawless fish and vertebrates from pufferfish to mouse and man could further clarify the phylogeny of tetralogous genome parts and the time points of duplications [27]. Changes in genome complexities are also associated with other major evolutionary transitions such as from prokaryotes to eukaryotes or from protozoa to metazoa [28], which, therefore, should be compared to the transition from invertebrates to vertebrates. Short-term benefits of the recognition of the four-fold complexity of vertebrate genomes might include a unified and phylogenetic nomenclature for invertebrate and vertebrate gene families and immediate help in sorting our roughly 80 000 genes into $4 \times 20\,000$ groups on the quadruplicated parts of the human genome.

Acknowledgements: I would like to thank V. Schmid, P. Flook and M. Šuša for constructive comments. Additional information and the hundreds of references for gene localisations, sequences and methods may be obtained on the World Wide Web or from the author.

References

[1] Miller, D.J. and Miles, A. (1993) Nature 365, 215–216.
[2] Bürglin, T.R., Ruvkun, G., Coulson, A., Hawkins, N.C., McGhee, J.D. Schaller, D., Wittmann, C., Müller, F. and Waterston, R.H. (1991) Nature 351, 703.
[3] Garcia-Fernàndez, J. and Holland, P.W.H. (1994) Nature 370, 563–566.
[4] Graham, A., Papalopulu, N. and Krumlauf, R. (1989) Cell 57, 367–378.
[5] Duboule, D. and Dollé, P. (1989) EMBO J. 8, 1497–1505.
[6] Scott, M.P. (1992) Cell 71, 551–553.
[7] Gehring, W.J., Affolter, M. and Bürglin, (1994) Annu. Rev. Biochem. 63, 487–526.
[8] Carrol, S.B. (1995) Nature 376, 479–485.
[9] Maconchie, M., Nonchev, S., Morrison, A. and Krumlauf, R. (1996) Annu. Rev. Genet. 30 (in press).
[10] Holland, P.W.H. and Garcia-Fernàndez, J. (1996) Dev. Biol. 173, 382–395.
[11] Ruddle, F.H., Bartels, J.L., Bentley, K.L., Kappen, C., Murtha, M.T. and Pendleton, J.W. (1994) Annu. Rev. Genet. 28, 423–442.
[12] Spring, J., Goldberger, O.A., Jenkins, N.A., Gilbert, D. J, Copeland, N.G. and Bernfield, M. (1994) Genomics 21, 597–601.
[13] Nadeau, J.H. (1991) in: Advanced Techniques in Chromosome Research (Adolph, K. ed.), pp. 269–296, Dekker, New York.
[14] Lundin, L.G. (1993) Genomics 16, 1–19.
[15] Ohno, S. (1970) Evolution by Gene Duplication, Springer, Berlin.
[16] Miklos, G.L.G. and Rubin, G.M. (1996) Cell 86, 521–529.
[17] Slack, J.M.W., Holland, P.W.H. and Graham, C.F. (1993) Nature 361, 490–492.
[18] Trowsdale, J. (1993) Trends Genet. 9, 117–122.
[19] Takahashi, F., Endo, S., Kojima, T. and Saigo, K. (1996) Genes Dev. 10, 1645–1656.
[20] Chen, J.-Y., Dzik, J., Edgecombe, G.D., Ramsköld, L. and Zhou, G.-Q. (1995) Nature 377, 720–722.
[21] Forey, P. and Janvier, P. (1993) Nature 361, 129–134.
[22] Berchtold, M.W., Egli, R., Rhyner, J.A., Hameister, H. and Strehler, E.E. (1993) Genomics 16, 461–465.
[23] Li, X. and Noll, M. (1994) Nature 367, 83–87.
[24] Hanks, M., Wurst, W., Anson-Cartwright, L., Auerbach, A.B. and Joyner, A.L. (1995) Science 269, 679–682.
[25] Eppig, J.T. and Nadeau, J.H. (1995) Curr. Opin. Genet. Dev. 5, 709–716.
[26] Brenner, S., Elgar, G. Sandford, R., Macrae, A., Venkatesh, B. and Aparicio, S. (1993) Nature 366, 265–268.
[27] Wray, G.A., Levinton, J.S. and Shapiro, L.H. (1996) Science 274, 568–573.
[28] Szathmàry, E. and Maynard Smith, J. (1995) Nature 374, 227–232.
[29] Sugaya, K., Fukagawa, T., Matsumoto, K.-I., Mita, K., Takahashi, E.-I., Ando, A., Inoko, H. and Ikemura, T. (1994) Genomics 23, 408–419.