Were vertebrates octoploid?

Rebecca F. Furlong and Peter W. H. Holland*

School of Animal and Microbial Sciences, The University of Reading, Whiteknights, Reading RG6 6AĴ, UK

It has long been suggested that gene and genome duplication play important roles in the evolution of organismal complexity. For example, work by Ohno proposed that two rounds of whole genome doubling (tetraploidy) occurred during the evolution of vertebrates: the extra genes permitting an increase in physiological and anatomical complexity. Several modifications of this 'two tetraploidies' hypothesis have been proposed, taking into account accumulating data, and there is wide acceptance of the basic scheme. In the past few years, however, several authors have raised doubts, citing lack of direct support or even evidence to the contrary. Here, we review the evidence for and against the occurrence of tetraploidies in early vertebrate evolution, and present a new compilation of molecular phylogenetic data for amphioxus. We argue that evidence in favour of tetraploidy, based primarily on genome and gene family analyses, is strong. Furthermore, we show that two observations used as evidence against genome duplication are in fact compatible with the hypothesis: but only if the genome doubling occurred by two closely spaced sequential rounds of autotetraploidy. We propose that early vertebrates passed through an autoautooctoploid phase in the evolution of their genomes.

Keywords: tetraploidy; octoploidy; evolution; genome duplication; amphioxus

1. INTRODUCTION

In the formulation of the 'two tetraploidies' model by Ohno (1970), he argued for large-scale gene duplication, possibly by genome duplication, being fixed in the early chordates, specifically on the shared lineage leading to both cephalochordates (the subphylum including amphioxus) and the vertebrates (used here in the broad sense, to include lampreys, hagfish and jawed vertebrates). He suggested a second (and possibly a third) tetraploidy occurred at the 'fish or amphibian' grade (figure 1*a*). Ohno's hypothesis was based primarily on considerations of genome size and isozyme complexity; sources of data now known to be inaccurate guides to genome complexity. The first data to clearly argue against Ohno's scheme were published by Schmidtke *et al.* (1977). This report showed that amphioxus and an ascidian had similar isozyme complexity for several enzyme systems, leading the authors to conclude that the first of Ohno's proposed tetraploidies did not occur. This work, however, did not discount the possibility of later genome duplication on the vertebrate lineage. Considerable detail was added to the picture during the 1990s, through molecular cloning of numerous genes and gene families in ascidians, amphioxus and vertebrates. These studies have been reviewed extensively elsewhere (Holland 1996, 1999). In brief, they revealed that many gene families are represented by single genes in amphioxus and ascidia (and indeed in many other bilaterian invertebrates), but by several genes in each vertebrate species examined. These data support Ohno's contention that extensive gene duplication occurred early in vertebrate evolution.

Many modifications of Ohno's original model have been proposed, to take into account the emerging—and constantly updated—molecular data. For example, Holland *et al.* (1994) proposed there were two 'phases of gene duplication' on the vertebrate lineage, but suggested different dates to those of Ohno (1970). Holland *et al.* (1994) suggested that the first duplication occurred on the vertebrate lineage after divergence of the amphioxus lineage, and the second on the jawed vertebrate lineage after divergence of the jawless vertebrates (figure 1*b*). Two years later, Sharman & Holland (1996) maintained the same timings of gene duplication, but now proposed the mechanisms to be a combination of multiple tandem duplication (for the first phase) and tetraploidy (for the second). Sidow (1996) and Ohno (1998) followed these same timings, but invoked tetraploidy as the mechanism in each case. Spring (1997) was more specific; he proposed the mechanism in each case to be allotetraploidy: genome doubling by interspecific hybridization. By contrast, Kasahara *et al.* (1996) proposed both tetraploidy events to be later in vertebrate evolution, after the divergence of lampreys.

The current consensus in the literature is that extensive gene duplication occurred sometime in early vertebrate evolution. Most authors accept that it occurred in two phases, although the timings are contentious. The mechanisms are even more controversial, with tetraploidy being the most popular hypothesis, albeit debated. Furthermore, few authors distinguish between the possibilities of allotetraploidy (interspecific hybridization) and autotetraploidy (endogenous genome doubling), or indeed combinations of these (autoallooctoploidy or alloautooctoploidy). A few authors dispute that any form of tetraploidy was involved, citing either lack of direct support (Skrabanek & Wolfe 1998; Smith *et al.* 1999) or apparent counter-evidence (Hughes 1998, 1999; Martin

* Author for correspondence (p.w.h.holland@reading.ac.uk).

Figure 1. Probable phylogeny of the phylum Chordata, showing the relative timings of large-scale gene duplications, or genome duplications, as proposed by: (a) Ohno (1970); (b) Holland *et al.* (1994); and (c) this paper.

2001). We believe that much of this confusion stems from (i) not considering the totality of the relevant evidence; and (ii) incomplete consideration of the chromosomal processes that occur during and after tetraploidy. Here, we summarize the principal evidence and examine the impact of tetraploidy on patterns of molecular evolution. We conclude that two rounds of tetraploidy are the most likely explanation for the extensive gene duplication in early vertebrate evolution.

2. EVIDENCE IN FAVOUR OF TETRAPLOIDY

(a) *Gene number*

When gene number is examined in a variety of organisms, current evidence indicates that invertebrate gene number never exceeds 20 000, whereas vertebrate total gene number is generally thought to be much higher (Bird 1995). The most precise numbers, of course, come from organisms that have had their genomes completely sequenced. Thus, a nematode and a fruitfly have total gene numbers estimated at *ca.* 19 000 and 13 600, respectively (*Caenorhabditis elegans* Sequencing Consortium 1998; Adams *et al.* 2000), whereas *ca.* 31 000–39 000 genes have been estimated in the only fully sequenced vertebrate genome, that of humans (International Human Genome Sequencing Consortium 2001; Venter *et al.* 2001). Any of these numbers may be under- or overestimates due to the nature of gene prediction strategies. Complete genome sequences have not yet been determined for any animals close to the invertebrate–vertebrate transition, so it is unclear if these figures reflect a sudden increase in the vertebrate lineage. However, it is relevant that a ‘gene sampling’ method applied to the ascidian *Ciona intestinalis* yielded an estimate of only 15 500 protein-coding genes (Simmen *et al.* 1998). This figure is comparable with the fruitfly and nematode figures, even though this animal is on the chordate lineage (figure 1). Furthermore, although gene number has not been accurately determined in the lower vertebrates hagfish and lamprey, they appear to have the low-level genome-wide methylation pattern peculiar to the vertebrates (Tweedie *et al.* 1997). Thus, these data are compatible with a sudden increase in gene number early in the vertebrate lineage.

The most parsimonious way of creating a twofold or greater increase in total gene number is to duplicate all the genes in a single step. After a duplication event, loss of duplicate copies is expected to occur at a fairly high frequency due to the immediate relaxation of selective constraints on every gene. Thus, one round of tetraploidy is not expected to double the total gene number in the long run, and two tetraploidies are not expected to quadruple the number. For example, if half of the duplicated genes are lost after each tetraploidy, then two rounds of complete genome doubling will increase the total gene number by a factor of 2.25. If one multiplies the estimated *Ciona* gene number by 2.25, a number very close to the human gene total is obtained (34 000). These comparisons, therefore, are consistent with the hypothesis of genome duplication during early vertebrate evolution. They are, however, also consistent with other models invoking gene duplication in vertebrate history. For example, a high rate of single gene duplication coupled with retention of most of the duplicates, could also have caused the observed increase in gene number.

(b) *Gene families: evidence from complete genomes*

The second line of evidence often used in this debate is the existence of ‘tetralogy’, or the ‘one-to-four’ rule (Spring 1997; Meyer & Schartl 1999). Tetralogy refers to a situation in which four members of a gene family in vertebrates are homologous to a single copy of the gene in invertebrates. Two alternative approaches can be used to assess the validity of this proposed ‘rule’. The first involves comparison between complete genome sequences; the second focuses on specific gene families in key taxa. Comparisons between the complete human, *Drosophila* and *Caenorhabditis* sequences have been performed both by Venter *et al.* (2001) and by the International Human Genome Sequencing Consortium (2001). In the former, protein clusters were identified by a novel strategy termed ‘lek clustering’, which compares the list of BLAST matches obtained for every protein. These lek clusters were then plotted according to the ratio of human to fly, or human to nematode, proteins per cluster. The peak of this distribution lies over a 1 : 1 ratio. However, it is noteworthy that this plot is markedly skewed in the direction

of 'human predominance' (a left-hand skew in fig. 12 of Venter *et al.* (2001)). This indicates that 1 : 2, 1 : 3 and 1 : 4 ratios (fly : human or nematode : human) are relatively common, although none of these are the rule.

The International Human Genome Sequencing Consortium used a different method, employing all-against-all sequence comparisons to cluster deduced protein sequences into orthologue 'groups' with detectable sequence similarity (International Human Genome Sequencing Consortium 2001, pp. 903–904). A total of 1308 groups were identified with representative genes in human, fly, nematode and yeast. On average, each of these orthologue groups contains more genes in the human genome than in the other taxa (a mean of 2.4 genes per human orthologue group compared with 1.1 genes per group in fly, nematode or yeast), suggesting prevalent duplication in the lineage leading to humans. Orthologue groups with only single genes in fly and nematode (yeast not being considered) were then plotted according to the number of human genes (International Human Genome Sequencing Consortium 2001, fig. 49). In about half the cases (1195 groups), the human genome also contained just a single gene; the other cases had two, three, four or more genes in humans, although again there is no predominance of a 1 : 4 ratio. The example shown in detail, comprising nuclear hormone receptors, includes six cases of no duplication, only one case of duplication in fly, but seven cases of additional genes in humans (two of 1 : 2, four of 1 : 3 and one of 1 : 4). The same group also used the InterPro annotation protocol to cluster deduced proteins on the basis of shared motif and domain assignments. These InterPro entries equate more closely to gene 'superfamilies' than to families; for example, homeobox genes comprise just a single category, as do C2H2 zinc fingers. Even at this gross level of resolution, however, there is strong evidence for extra genes in humans. Of the 37 largest InterPro families, 32 have more proteins in the human proteome than in both fly and nematode. According to the authors, this clearly shows that 'gene duplication has been a major evolutionary force during vertebrate evolution' (International Human Genome Sequencing Consortium 2001, p. 906). In conclusion, lek clustering, orthologue group analysis and InterPro annotation studies all find strong evidence for gene duplication in the vertebrate lineage, but they do not support a strict 1 : 4 rule.

(c) *Gene families: phylogenetic analyses*

It might be thought that comparisons of complete genome sequences, such as those discussed above, would be the ultimate guide to gene family complexity and its evolution. After all, there is no bias in the choice of gene families being compared, and there is no possible recourse to 'missing' genes (genes that might exist, but have yet to be cloned). There are, however, two significant problems. First, the taxa being compared are very distant phylogenetically, having diverged well before the origin of vertebrates. Even if gene duplication is detected, it cannot be deduced whether it occurred close to vertebrate origins or earlier on the deuterostome lineage. Second, simply counting the numbers of related genes cannot distinguish between a lack of gene duplication on both lineages, and independent gene duplication on both. Consider the case of the *en* class homeobox genes. The *Drosophila* genome

contains two *en* class genes (*en* and *inv*), as does the human genome (*EN-1* and *EN-2*). This gene family, therefore, has a complexity ratio of 1 : 1. Construction of phylogenetic trees from the gene sequences, however, reveals that *en* and *inv* are the result of one gene duplication, whereas *EN-1* and *EN-2* result from a different gene duplication (Gibert *et al.* 1997; Force *et al.* 1999). This is further supported by the fact that *en* and *inv* are physically adjacent genes in *Drosophila* (suggesting origin by relatively recent tandem duplication), whereas *EN-1* and *EN-2* map to distinct human chromosomes. Inclusion of *en* class genes from other taxa allows each duplication to be dated; for example, the *en* class gene of amphioxus diverges before the duplication of *EN-1* and *EN-2*, dating this event to the vertebrate lineage (Force *et al.* 1999). Thus, despite a 1 : 1 ratio (human : fly), the *en* class gene family does show gene duplication on the vertebrate lineage.

For these reasons, detailed analysis of gene families from key taxa provides an insight into gene duplication that is complementary to comparison of whole genomes. When asking if gene duplication occurred in vertebrate evolution, the informative taxa will be invertebrates on the deuterostome lineage (echinoderms, hemichordates, tunicates and amphioxus). Furthermore, the closer the taxon is to the vertebrates, the more accurate the deduced timing of duplication will be. Amphioxus, the living sister group to the vertebrates, therefore emerges as a particularly useful test organism for this question (figure 1).

In table 1, we list numerous gene families that have been examined from amphioxus, and compare the number of genes identified in amphioxus and mammals. We stress, however, that it is not the absolute number of genes cloned that is crucial; rather, it is their phylogenetic relationships. Hence, for each gene family, we also tabulate the pattern of gene duplication on the vertebrate lineage, as deduced from molecular phylogenetic analyses. These data were compiled as follows. First, to avoid sampling bias, we surveyed all 287 amphioxus sequences on the GenBank database as of 24 August 2001. From these, we confined our attention to published sequences only, excluding random clones (98 gene families). DNA sequences for which phylogenetic analysis had not been carried out were then searched against GenBank using TBLASTN (searches all six translations of a nucleotide sequence against a translated nucleotide database). Those sequences that were unique to amphioxus or that were too short for reliable phylogenetic analysis were discarded. Of the 83 gene families remaining, phylogenetic trees were taken from the literature for 62 gene families (various methods of analysis), three were provided by personal communication, and 18 were performed for this study. For the latter, DNA sequences were retrieved from GenBank, and deduced protein sequences aligned using CLUSTALW within the sequence editor BioEDIT (Hall 1999). Regions of ambiguous alignment were removed or adjusted by eye. Both protein and DNA trees were inferred using maximum likelihood implemented in TREE-PUZZLE (Strimmer & von Haeseler 1996) and compared for discrepancies. There was rarely a difference between protein and DNA topologies, but because DNA data often failed the TREE-PUZZLE χ^2 -test for sequence bias, protein trees were used to calculate numbers of gene

Table 1. Gene number and duplication data for 98 gene families examined in amphioxus and mammals. (The majority of gene families analysed show duplication on the vertebrate lineage after divergence from amphioxus, deduced by molecular phylogenetic analysis.)

family	genes cloned in amphioxus	genes cloned in mammals	vertebrate- specific duplication?	cloning reference	phylogeny reference
aldo-keto reductase	1	1	1 to 1	Boeddrich <i>et al.</i> (1999)	this work
Brachyury (T) ^a	2	1	1 to 1	P. W. H. Holland <i>et al.</i> (1995)	as cloning
F-spondin	1	1	1 to 1	Shimeld (1998)	as cloning
Mn-SOD	1	1	1 to 1	Smith & Doolittle (1992)	as cloning
NKx2.2	1	2	1 to 1	Holland <i>et al.</i> (1998)	F. Castro, pers. comm.
PAH	1	1	1 to 1	Patton <i>et al.</i> (1998)	as cloning
Pax6 ^b	1	1	1 to 1	Gardon <i>et al.</i> (1998)	this work
SPC2	1	1	1 to 1	Oliva <i>et al.</i> (1995)	this work
SPC3	1	1	1 to 1	Oliva <i>et al.</i> (1995)	this work
Tbx20	1	1	1 to 1	Ruvinsky <i>et al.</i> (2000)	as cloning
triose phosphate isomerase	1	1	1 to 1	Nikoh <i>et al.</i> (1997)	as cloning
Wnt1	1	1	1 to 1	Holland <i>et al.</i> (2000a)	Schubert <i>et al.</i> (2000c)
Wnt11	1	1	1 to 1	Schubert <i>et al.</i> (2000b)	Schubert <i>et al.</i> (2000c)
Wnt4	1	1	1 to 1	Schubert <i>et al.</i> (2000a)	Schubert <i>et al.</i> (2000c)
actin cytoplasmic	1	2	1 to 2	Kusakabe <i>et al.</i> (1999)	as cloning
basic helix-loop-helix (myogenic) ^a	2	4	1 to 2	Araki <i>et al.</i> (1996)	as cloning
BMP2/4	1	2	1 to 2	Panopoulou <i>et al.</i> (1998)	this work
brain factor	1	2	1 to 2	Toresson <i>et al.</i> (1998)	F. Mazet, pers. comm.
cholinesterase ^a	2	2	1 to 2	Sutherland <i>et al.</i> (1997)	as cloning
COUP-TF	1	2	1 to 2	Escriva <i>et al.</i> (1997)	Langlois <i>et al.</i> (2000)
creatine kinase	1	2	1 to 2	Graber & Ellington (2001)	as cloning
Emx	2	2	1 to 2	Williams & Holland (2000), Minguillón <i>et al.</i> (2002)	as cloning
engrailed	1	2	1 to 2	L. Z. Holland <i>et al.</i> (1997)	this work
Evx ^a	2	2	1 to 2	Ferrier <i>et al.</i> (2001)	as cloning
goosecoid	1	2	1 to 2	Neidert <i>et al.</i> (2000)	as cloning
insulin-like growth factor ^c	1	2	1 to 2	Chan <i>et al.</i> (1990)	Patton <i>et al.</i> (1998)
islet	1	2	1 to 2	Jackman <i>et al.</i> (2000)	as cloning
Mnx ^d	1	2	1 to 2	Ferrier <i>et al.</i> (2001)	as cloning
netrin	1	2	1 to 2	Shimeld (2000)	as cloning
NKx2.1	1	2	1 to 2	Venkatesh <i>et al.</i> (1999)	F. Castro, pers. comm.
Otx	1	2	1 to 2	Williams & Holland (1998)	as cloning
Pax1/9	1	2	1 to 2	N. D. Holland <i>et al.</i> (1995)	Wada <i>et al.</i> (1998)
Pax3/7	1	2	1 to 2	Holland <i>et al.</i> (1999)	as cloning
PTPN3	1	2	1 to 2	Ono-Koyanagi <i>et al.</i> (2000)	as cloning
PTPN6	1	2	1 to 2	Ono-Koyanagi <i>et al.</i> (2000)	as cloning
PTPR4	3	2	1 to 2	Ono-Koyanagi <i>et al.</i> (2000)	as cloning
PTPR5	1	2	1 to 2	Ono-Koyanagi <i>et al.</i> (2000)	as cloning
Rab GDI	1	2	1 to 2	Sedlacek <i>et al.</i> (1999)	as cloning
snail	1	2	1 to 2	Langeland <i>et al.</i> (1998)	as cloning
Tbx1/10	1	2	1 to 2	Ruvinsky <i>et al.</i> (2000)	as cloning
Tbx2/3 and 4/5 ^c	2	4	1 to 2	Ruvinsky <i>et al.</i> (2000)	as cloning
Tbx6/16 ^f	1	1 or 2	1 to 2	Ruvinsky <i>et al.</i> (2000)	as cloning
Tob	1	2	1 to 2	N. D. Holland <i>et al.</i> (1997)	this work
troponin C	1	2	1 to 2	Yuasa <i>et al.</i> (1998)	as cloning
twist	1	2	1 to 2	Yasui <i>et al.</i> (1998)	as cloning
Vent ^g	1	2	1 to 2	Kozmik <i>et al.</i> (2001)	as cloning
Wnt7	1	2	1 to 2	Schubert <i>et al.</i> (2000a)	Schubert <i>et al.</i> (2000c)
Wnt8	1	2	1 to 2	Schubert <i>et al.</i> (2000d)	Schubert <i>et al.</i> (2000c)
aldolase	1	3	1 to 3	Kuba <i>et al.</i> (1997)	as cloning
distalless ^{f,h,i}	1	6	1 to 3	Holland <i>et al.</i> (1996)	this work
dystrophin	1	3	1 to 3	Roberts & Bobrow (1998)	as cloning
eomes/Tbr1/Tbx21	1	3	1 to 3	Ruvinsky <i>et al.</i> (2000)	as cloning
Gli	1	3	1 to 3	S. Shimeld, pers. comm.	as cloning
hedgehog	1	3	1 to 3	Shimeld (1999)	as cloning
HMG(sox1/2/3)	1	3	1 to 3	Holland <i>et al.</i> (2000b)	as cloning

(Continued)

Table 1. (Continued.)

family	genes cloned in amphioxus	genes cloned in mammals	vertebrate- specific duplication?	cloning reference	phylogeny reference
HNF ^a	2	3	1 to 3	Shimeld (1997)	as cloning
insulin receptor	1	2 or 3	1 to 3	Pashmforoush <i>et al.</i> (1996)	this work
krox ^{f,i}	1	4	1 to 3	Knight <i>et al.</i> (2000)	as cloning
lamin	1	3	1 to 3	Riemer <i>et al.</i> (2000)	as cloning
Msx ^k	1	3	1 to 3	Sharman <i>et al.</i> (1999)	this work
neurogenin	1	3	1 to 3	Holland <i>et al.</i> (2000b)	as cloning
Pax2/5/8	1	3	1 to 3	Kozmik <i>et al.</i> (1999)	as cloning
peroxidase	1	3	1 to 3	Ogasawara (2000)	this work
Pitx	1	3	1 to 3	Yasui <i>et al.</i> (2000)	as cloning
PTPR2A	1	3	1 to 3	Ono-Koyanagi <i>et al.</i> (2000)	as cloning
retinol dehydrogenase ^a	2	3	1 to 3	Dalfo <i>et al.</i> (2001)	as cloning
Tbx15/18/22	1	3	1 to 3	Ruvinsky <i>et al.</i> (2000)	as cloning
VEGFR-like	1	3	1 to 3	Suga <i>et al.</i> (1999)	as cloning
actin muscle	1	4	1 to 4	Kusakabe <i>et al.</i> (1999)	as cloning
dral ^j	1	4	1 to 4	Schubert <i>et al.</i> (1998)	this work
FGF receptor	1	4	1 to 4	Suga <i>et al.</i> (1999)	as cloning
Hox cluster ^c	14	39	1 to 4	Garcia-Fernandez & Holland (1994); Ferrier <i>et al.</i> (2000)	this work
myosin	1	4	1 to 4	L. Z. Holland <i>et al.</i> (1995)	this work
notch	1	4	1 to 4	Holland <i>et al.</i> (2001)	as cloning
nuclear factor I	1	4	1 to 4	Fletcher <i>et al.</i> (1999)	as cloning
ParaHox ^c	3	6	1 to 4	Brooke <i>et al.</i> (1998)	this work
tropomyosin	1	4	1 to 4	Suzuki & Satoh (2000)	this work
Zic	1	4	1 to 4	S. Shimeld, pers. comm.	as cloning
alcohol dehydrogenase ^l	1	7	1 to 7	Canestro <i>et al.</i> (2000)	this work
SRC-like ^a	2	8	1 to 8	Suga <i>et al.</i> (1999)	as cloning
keratin type 2 ^{m,l}	1	12	1 to 12	Luke & Holland (1999); Karabinos <i>et al.</i> (2000)	this work
Eph ^{a,m,l}	2	13	1 to 13	Suga <i>et al.</i> (1999)	as cloning
keratin type 1 ^{a,m,l}	5	19	1 to 19	Luke & Holland (1999)	this work
calcium vector ⁿ	1	—	—	Yuasa <i>et al.</i> (1999)	n/a
calmodulin ^o	4 or 5	3 or 4	—	Karabinos & Bhattacharya (2000)	as cloning
EER2 ^p	—	—	—	Escriva <i>et al.</i> (1997)	n/a
FTZ-F1 ^p	—	—	—	Escriva <i>et al.</i> (1997)	n/a
FXR ^p	—	—	—	Escriva <i>et al.</i> (1997)	n/a
LIM ^p	—	—	—	Suzuki & Satoh (2000)	n/a
lipase ^p	—	—	—	Tweedie <i>et al.</i> (1997)	n/a
PPAR ^p	—	—	—	Escriva <i>et al.</i> (1997)	n/a
proprotein convertase ^p	—	—	—	Oliva <i>et al.</i> (2000)	n/a
PTP10 ^p	—	—	—	Ono-Koyanagi <i>et al.</i> (2000)	n/a
PTPR3 ^p	—	—	—	Ono-Koyanagi <i>et al.</i> (2000)	n/a
retinoic acid receptor ^p	—	—	—	Escriva <i>et al.</i> (1997)	n/a
TR2 ^p	—	—	—	Escriva <i>et al.</i> (1997)	n/a
tubulin ^p	—	—	—	Tweedie <i>et al.</i> (1997)	n/a
Whn ^p	—	—	—	Schlake <i>et al.</i> (1997, 2000)	n/a

^a Molecular phylogenetics show independent gene duplication in the amphioxus lineage.

^b There is a possibility that Pax4 is a paralogue of Pax6, in which case this becomes 1 : 2.

^c There is a possibility that insulin is a derived IGF, in which case this ratio becomes 1 : 3.

^d As yet, no mammalian version of one of the vertebrate groups has been cloned. However, evidence from chicken genes and paralogy suggests this ratio.

^e It is likely that this family of genes duplicated as a cluster, therefore we consider it as a single locus.

^f Ambiguity in one or more nodes of the tree does not allow resolution.

^g The cloning paper uses only non-mammalian vertebrates for the tree, so this is an inferred ratio for mammals.

^h Inclusion of ascidian data suggests 1 : 3 duplication, with an additional gene yet to be cloned in amphioxus.

ⁱ Evidence from paralogous regions suggests this ratio.

^j Fourth copy only found in mouse but scanning of human genome indicates potential human copy.

^k The fourth human Msx gene, identified by Pollard & Holland (2000), is a processed pseudogene and has been renamed MSX2p by the Human Genome Organisation (HUGO).

^l Evidence from human genome map suggests that tandem duplication has contributed to these members.

^m An extra paralogue has been cloned from non-mammalian vertebrates.

ⁿ This gene is unique to amphioxus so far.

^o Because of the extreme conservation of these genes, a ratio cannot be inferred.

^p These sequences are very short or had few good BLAST matches, so were excluded from further phylogenetic study.

Figure 2. Histogram summarizing the patterns of gene duplication on the vertebrate lineage as deduced by phylogenetic analysis of 84 gene families represented in amphioxus and vertebrates; data from table 1.

duplication events. Trees and relevant accession numbers can be viewed at <http://www.rubic.rdg.ac.uk/amphioxus>.

The large majority of gene families analysed show gene duplication on the vertebrate lineage, after it had diverged from amphioxus (table 1; figure 2). The commonest pattern detected is that a single gene in the latest common ancestor of these two taxa duplicated to give two, three or four genes in mammals. Examples include transcription factors (e.g. *Pax1/9*, *Pax3/7*, *Pax2/5/8*, *Msx*, Hox gene clusters, *COUP-TF*, *Otx*, *Ptx*, *en*), secreted signalling molecules (e.g. *BMP2/4*, *hedgehog*, *IGF*, *Wnt7*, *Wnt8*), enzymes (e.g. aldolase, peroxidase) and others (e.g. dystrophin, FGF receptors, tropomyosin). In some cases, the ancestral gene duplicated independently in both the vertebrate lineage and the lineage leading to amphioxus; for example *HNF3*, cholinesterase, myogenic basic helix-loop-helix (bHLH) and actin genes. In a minority of cases, no evidence for duplication on the vertebrate lineage was detected. We conclude from these data that gene duplication was indeed very prevalent in vertebrate evolution. Tetralogy, or the 1:4 rule, does not hold up to scrutiny, although there is a very clear predominance of 1:2, 1:3 and 1:4 (comparing the extinct ancestral chordate with mammals), with few examples expanding to larger than four members (figure 2). These patterns are entirely congruent with two rounds of tetraploidy in the vertebrate lineage followed by some gene loss.

(d) Paralogy regions

'Paralogy regions' in vertebrate genomes have been considered a strong line of evidence in favour of tetraploidy during vertebrate evolution (Lundin 1993). Paralogy regions consist of a series of linked (but usually unrelated) genes on one chromosome, many of which have linked homologues on at least one other chromosome. The simplest explanation for such regions is that they are historical remnants of ancient genome or chromosomal duplications. For example, *IGF-2*, the myogenic bHLH gene *myoD*, two LDH genes, two aromatic amino-acid hydroxylase (*AAAH*) genes, a *ras* gene and parathyroid hormone map to human chromosome 11 at band 11p15. Paralogues (duplicates) of all these genes map to human chromosome 12: *IGF-1*, two myogenic bHLH genes, *LDH-B*, an *AAAH* gene, a *ras* gene and parathyroid

hormone-like hormone (see Patton *et al.* 1998). In this case, the genes are divided between two regions of chromosome 12 (both arms); this could be explained by duplication followed by inversion or translocation in one of the descendent chromosomes.

Although the existence of paralogy regions is very striking, three questions must be answered before they can be used as evidence in favour of tetraploidy. First, it must be shown that they are indeed historical remnants of duplication and not the result of convergent evolution bringing similar sets of genes together at disparate genomic locations. Second, if they are historical remnants, their ages must be determined. Any hypothesis invoking whole genome duplication in vertebrate ancestry will predict that different paralogy regions should have identical ages. Third, how much of the human genome is contained within paralogy regions? Tetraploidy predicts the proportion should be high.

We do not, as yet, have definitive answers to each question. The first question—history or convergence—has stimulated heated debate. Hughes (1998, 1999) has argued that phylogenetic analyses of gene families within paralogy regions do not support the 'historical remnants' hypothesis; hence, paralogy regions must have been assembled convergently, for adaptive reasons. We dissect this argument later in this paper, arguing that it is not valid. In addition, there is no direct evidence in favour of adaptive assembly of these regions; that paralogy regions are historical remnants is a simpler hypothesis. If duplication did produce paralogy regions, the genes within them must have already been adjacent (or neighbouring) prior to any duplication. This prediction can be tested by comparison with non-vertebrate genomes. Thus, Ruddle *et al.* (1994) pointed out that many of the gene families mapping close to the Hox gene clusters in paralogy regions on human chromosomes 2, 7, 12 and 17 are also in the vicinity of the Hox clusters of *Drosophila* and *C. elegans*. A similar conclusion was reached by Coulier *et al.* (2000) for the genes mapping close to the NK class homeobox genes on human chromosomes 2, 4, 5, 8 and 10 and *Drosophila* chromosome 3R, region 81–100. These analyses need extending to other paralogy regions, but already they provide an independent line of evidence that paralogy regions are indeed historical remnants of duplication. For maximal resolution of these analyses, genomic organization in amphioxus will be a much better comparator than flies or nematodes, as it is the closest invertebrate relative of the vertebrates. This test awaits extensive genomic sequence from amphioxus.

The second question asks whether each paralogy region is the same age. The maximum (but not the minimum) age for each paralogy region duplication can be deduced by dating the duplication event for each of the constituent gene families. In the example introduced above, three of the gene families have been studied in amphioxus and the phylogenetic trees deduced (Araki *et al.* 1996; Patton *et al.* 1998). The IGF and myogenic bHLH gene families were both found to have duplicated on the vertebrate lineage, to give the descendent genes on chromosomes 11 and 12. This sets a maximal age for duplication of this chromosomal region as the base of the vertebrate lineage, after it had diverged from amphioxus. Analysis of the AAAH genes gives an older maximal age, but this possi-

bility is excluded by the IGF and bHLH dates (Patton *et al.* 1998). Similar analyses have been undertaken for the Hox-containing paralogy regions (chromosomes 2, 7, 12, 17; Garcia-Fernandez & Holland 1994; Pollard & Holland 2000) and the NK paralogy region (chromosomes 2, 4, 5, 8, 10; Coulier *et al.* 2000; Pollard & Holland 2000). In all cases, the maximal date of duplication is found to be the same: the base of the vertebrate lineage. Additional taxa need to be included in these comparisons to enable dates to be resolved further.

The third question concerns the proportion of a vertebrate genome encompassed by paralogy regions. In the analytical review of Lundin (1993), and elsewhere (Lundin 1995; Katsanis *et al.* 1996; Lundin & Larhammar 1998; Skrabanek & Wolfe 1998; Kasahara 1999), impressive lists of paralogy regions are presented. Nonetheless, their length is unlikely to total more than a small percentage of the human genome. This list is continually growing, however, as previously undescribed paralogy regions are recognized. For example, Pollard & Holland (2000) analysed the chromosomal positions of 'ANTP superclass' homeobox genes, and found that almost all these genes are in paralogy regions. In addition to the well described Hox regions (fourfold paralogy), they describe a ParaHox paralogy region (also fourfold), an NK paralogy region (also fourfold) and an en paralogy region (twofold). All are supported by numerous linked gene families. A less detailed, but maximally comprehensive, analysis is provided by Venter *et al.* (2001). In their analyses of the whole human genome sequence, Venter *et al.* applied a multiple alignment algorithm to concatenated protein sequences representing every human chromosome. This revealed 1077 duplicated blocks spread throughout the genome. Some of these are astonishingly large; for example, 70% of chromosome 14 has detectable paralogy with a region of chromosome 2. It is not possible, from the data presented in Venter *et al.* (2001), to calculate the proportion of the genome encompassed by these duplicated blocks, but certainly a large proportion of protein-coding genes in the human genome are involved. Furthermore, if substantial gene loss had occurred after duplication, this would have led to underestimation of the proportion of the genome affected by duplication. The authors suggest that large-scale ancient segmental duplications provide the most likely explanation, although the data are also consistent with whole genome duplication.

In summary, paralogy regions are abundant and widely distributed in human (and other vertebrate) genomes. They are the historical remnants of duplicated genomic regions; the progenitor regions of which existed prior to the chordate lineage. The descendent paralogy regions have similar or identical ages, the duplications dating (where examined) to no earlier than the base of the vertebrate lineage. However, these data do not yet resolve whether they arose by multiple, independent, regional duplications (over a short period of time), serial duplication of a portion of the genome, tetraploidy, or a combination of these processes.

3. EVIDENCE AGAINST TETRAPLOIDY

The multiple lines of evidence outlined above all clearly indicate that gene duplication was very extensive in early

vertebrate evolution. This conclusion can hardly be doubted. Indeed, it is an insight of considerable relevance to our understanding of 'what makes a vertebrate?' (Shimeld & Holland 2000). Much more controversial are the mechanism and precise timing of these gene duplication events. In particular, did whole genome duplication (tetraploidy) occur close to the base of the vertebrate lineage? Three of the lines of evidence (total gene number, gene families in complete genomes and gene family phylogenetics) discussed above cannot be used to definitively answer this question. Each can be construed as evidence in favour of tetraploidy only if one accepts that tetraploidy is the most parsimonious explanation for the very widespread gene duplication detected. Strangely, they have occasionally been taken as evidence against tetraploidy, but only when some erroneous assumptions have been made. For example, examination of table 1 and figure 2 indicates that relatively few gene families show true tetralogy (1:4): or even a one to four pattern of duplication as inferred from a phylogenetic tree. Such a pattern, however, is not a sound prediction of two rounds of tetraploidy, as gene loss is expected to be prevalent after duplication. Hence, the amphioxus data are consistent with two rounds of tetraploidy. They are also consistent with other mechanisms of gene duplication. The line of evidence that has the most direct bearing on the mechanism of gene duplication is the fourth of those outlined above: paralogy regions. There are essentially three competing hypotheses for the origin of paralogy regions: they are historical remnants of tetraploidy events; they are historical remnants of segmental duplications; or they are adaptive assemblages of genes, evolved convergently on different chromosomes. Three tests have been used to distinguish between these hypotheses, or more specifically to distinguish between the adaptive hypothesis on the one hand, and the two historical hypotheses on the other. We examine each of these tests in turn.

(a) Ancient duplications in paralogy regions

If paralogy regions were created by duplication of the genome, or by regional duplications, it seems reasonable to expect that all the gene families shared by the chromosomal regions should have the same date of duplication. Using the chromosome 11/12 example again, one might expect that the IGF, AAAH, myogenic bHLH, LDH, PTH and ras gene families would all show coincident duplication, yielding the related genes on chromosomes 11 and 12. However, this is not the case. In this example, phylogenetic trees indicate that *IGF-1* and *IGF-2* evolved by duplication on the vertebrate lineage, after divergence from amphioxus (Patton *et al.* 1998); the same is true for their neighbours, the myogenic bHLH genes (Araki *et al.* 1996). However, the duplication that created the AAAH genes on chromosomes 11 and 12 predates the divergence of arthropods, nematodes and vertebrates (Patton *et al.* 1998). A similar situation is seen for the *Wnt1*, *Wnt2* and *Wnt3* genes linked to the Hox gene clusters on human chromosomes 7, 12 and 17. Here the Wnt duplications greatly predate the Hox gene cluster duplications (Sidow 1992; Garcia-Fernandez & Holland 1994).

At first sight, such findings seem inconsistent with the historical explanations for paralogy regions. Consequently, Hughes (1998) suggested that paralogy regions must have

been assembled independently on different chromosomes for adaptive reasons. For example, perhaps there is a selective advantage for particular gene types to be neighbours in the genome, leading to convergent evolution of gene groupings at disparate loci. There is an alternative explanation, however, that is completely consistent with the historical models. Let us assume *Wnt1*, *Wnt2* and *Wnt3* were neighbouring genes in an invertebrate, near the single ancestral Hox gene cluster. Similarly, assume that the three AAAH genes were neighbours, adjacent to the single ancestral bHLH and IGF genes. Neither assumption is outrageous, as these are related genes that presumably arose by tandem gene duplication; furthermore, it is testable by comparison with outgroups. After duplication of these chromosomal regions in the vertebrate lineage, a high incidence of gene loss in the Wnt and AAAH gene families could restore the original complement of these genes, but now they have been split between different paralogy regions (see Patton *et al.* 1998). As a consequence, the fact that a few genes in paralogy regions are older than their neighbours is not sufficient to disprove the historical explanations for paralogy regions. We contend that duplication of the whole genome, or of regions of a genome, are simpler and more parsimonious explanations for the origin of paralogy regions than is adaptive assembly.

Finally, we point out that the historical explanations and the adaptive assembly model make quite different predictions for the arrangement of genes in prevertebrate genomes. If paralogy regions are historical remnants of duplications, their progenitor genes will also be linked in at least some relatives of vertebrates. Thus, some (or all) of the AAAH, IGF, myogenic bHLH, LDH, PTH and ras genes should be linked in the amphioxus genome. By contrast, none of these genes should be linked in amphioxus if human paralogy regions have been assembled independently on chromosomes 11 and 12 (which, incidentally, begs the question as to what could possibly be the selective advantage of these genes being chromosomal neighbours in vertebrates, but not in their relatives).

(b) *Asymmetrical tree topologies*

If paralogy regions are historical remnants of two whole genome duplications, perhaps this will be reflected in the shape of phylogenetic trees drawn from their constituent genes. It has been argued that a gene family with paralogues at four distinct chromosomal regions should show a symmetrical phylogenetic tree of the form ((A,B)(C,D)), if two tetraploidy events had occurred (Hughes 1998; Smith *et al.* 1999). Examples that have been used to test this prediction include the major histocompatibility complex paralogy region, containing at least 10 distinct gene families shared by chromosomes 1, 6, 9 and possibly 19 (Hughes 1998; Katsanis *et al.* 1996), seven tightly linked gene families shared by chromosomes 1, 2, 8 and 20 (Gibson & Spring 2000) and the four Hox gene clusters and neighbouring collagen genes (Bailey *et al.* 1997). In each of these cases, however, topologies of the form ((A,B)(C,D)) are rare; instead, the general rule is for asymmetrical or sequential tree topologies of the form (A,(B,(C,D))). For example, Hughes (1998) found an asymmetrical tree (with good support) for four out of five tetralogous gene families. Similarly, Martin (2001) did not

find a strongly supported symmetrical tree in 10 (unlinked) gene families analysed. These authors conclude that two rounds of tetraploidy is disproved by these findings. They argue that one must invoke sequential regional duplications, or adaptive assembly. Other authors have used asymmetrical trees as evidence that there were in fact three tetraploidy events in vertebrate evolution, rather than two (Bailey *et al.* 1997).

Our own analyses of tetralogous gene families also recover asymmetrical gene trees. Indeed, at first we took these as evidence against a simple model invoking two tetraploidy events, just as almost all other authors have done. This is mistaken, however. To realize why, it must be remembered that there are two distinct mechanisms of tetraploidy: allotetraploidy (interspecies hybridization) and autotetraploidy (endogenous genome doubling). Two sequential rounds of interspecies hybridization would, indeed, be expected to produce symmetrical trees of the form ((A,B)(C,D)). The fact that such trees are rarely found effectively disproves the occurrence of two rounds of allotetraploidy in vertebrate history (contrary to the suggestion of Spring (1997)). However, we argue here that two close rounds of autotetraploidy does not predict trees of the form ((A,B)(C,D)); instead, it predicts asymmetrical trees, exactly as is found. We explain the logic behind this deduction later (see § 4).

(c) *Non-congruent linked tree topologies*

If paralogy regions are the result of any form of duplication (genome or regional), then it seems reasonable to expect that the constituent gene families will show the same phylogenetic tree shapes. Thus, even if we accept that asymmetrical gene trees are compatible with duplication models, surely if one gene family shows the topology (A,(B,(C,D))) then its neighbour cannot show (C,(D,(B,A)))? Martin (2001) examined pairs of gene families found on the same chromosomes and found their tree topologies to be incongruent (although the genes he examined were not all tightly linked, and may not have been recent paralogues in all cases). Martin asserts that this evidence favours rejection of the hypothesis of successive genome duplications at the base of vertebrate lineage. Gibson & Spring (2000) were more rigorous in their definition of paralogous regions, but they still found incongruent topologies for seven gene families analysed. These authors do not reject the genome duplication hypotheses, but state that their tree topologies are unstable, as might be predicted if two genome duplications had occurred in rapid succession.

We argue that incongruent tree topologies are not evidence against sequential genome duplication. In fact, incongruent tree topologies for some linked genes are actually a prediction of two rounds of genome duplication, if both occurred by autotetraploidy (not allotetraploidy) and if the two events occurred in relatively quick succession. We explain the logic behind this deduction below.

4. THE EFFECT OF OCTOPOIDY

Doubling of the genome may occur by two routes. The first, allotetraploidy, occurs when two closely related species hybridize, as is common in the plant kingdom (Wendel 2000). In this case, the two genomes coexist in a single

Figure 3. (a) Allotetraploidization at one or both genome doubling events will lead to a symmetrical phylogenetic tree topology. (b) Autotetraploidization may also provide a symmetrical topology, but only if sufficient time has elapsed between the duplications that the chromosomes have diploidized fully.

nucleus, and thus the chromosomes segregate normally at meiosis, each pairing with its own homologue in the usual way. If two tetraploidizations had occurred during the evolution of the vertebrates, and one or both of these were by allotetraploidy, we would predict that gene trees should show symmetrical topologies (figure 3a). However, genome doubling may also occur within a single species via autotetraploidy, which may occur after non-disjunction of sister chromatids at meiosis, or by uncoupling of mitosis and cell division early in germline development. Immediately after such an event, there will be two identical copies of every chromosome and thus four identical homologues at meiosis I. Meiotic pairing may occur between any two of these four, meaning that occasionally a chromosome may form a bivalent with its duplicate 'homeologue' rather than the homologue created prior to the first meiotic division. Furthermore, a 'pairing switch' may occur between homeologues such that a quadrivalent is formed. Non-homologous bivalents and quadrivalents have been well described in autotetraploid fishes, molluscs and plants (Hauber *et al.* 1999; Allendorf & Thorngard 1984; Guo & Allen 1997). The effect of this process is that each gene has four alleles that are able to interchange freely. The DNA sequences of these alleles are able to diverge so that they are distinguishable, and crossovers can be observed empirically (Allendorf & Thorngard 1984). This has been referred to as 'tetrasomic inheritance'.

After a period of time, the duplicate chromosomes are expected to diverge so they are no longer able to form

homeologous pairs (Li 1980). Divergence of different alleles is then predicted to accelerate such that distinct alleles become fixed for each chromosome pair. This diploidization process will be gradual and random, however, so there is an intermediate stage in which some chromosome sets have diploid inheritance, and some have fully tetraploid inheritance. Still others have begun to diploidize but exhibit an interesting phenomenon termed 'residual tetraploidy', in which quadrivalents or homeologous bivalents form only in occasional meioses. This situation is still observed in salmonid fishes, descended from a lineage-specific autotetraploidization 25–50 million years ago (Allendorf & Thorngard 1984).

We now extrapolate to a situation in which two autotetraploidizations take place in succession. If diploidization after the first tetraploidization was substantially complete by the occurrence of the second, we suggest that a symmetrical gene tree would be expected in most, if not all, cases (figure 3b). By contrast, if the homeologues were still able to pair at the time of the second tetraploidization, this duplication would create four essentially identical chromosomes, or eight at meiosis. These eight chromosomes would be able to pair as random bivalents or as multivalents, as is observed in living octoploid amphibians (Schmid *et al.* 1985). Because diploidization is a random process, it will probably occur in a sequential manner, with a mutational event separating one chromosome of the four such that it is no longer likely to form homeologous structures. The remaining three sets of homologues may

Figure 4. (a) If genome duplications occur in close succession, diploidization will be sequential from an octoploid or pseudo-octoploid state. Gene trees will then reflect the order of diploidization of chromosomes, rather than the order of chromosome duplication. Furthermore, topology will be sequential (asymmetrical). (b) Crossover is still possible during the gradual diploidization process, so that genes distal to the crossover event may show a topology incongruent with those proximal to the crossover.

continue to pair randomly until another mutation occurs. We argue that the result of this process is that gene trees deduced from descendent species may show a sequential topology, as is indeed observed in tetralogous gene families (figure 4a). Additionally, as the time of divergence of the genes is not the same as the time of duplication, and different chromosomes may diploidize at different rates, it may not be possible to calculate a consensus date for the duplications using a molecular clock.

The peculiarities of meiosis after autotetraploidization have further implications for linked genes. As has been noted above, diploidization is a random and gradual process with intermediate stages during which homeologous pairs and quadrivalents are still able to form. 'Homeologous crossover' between such chromosome pairs has been shown to occur in autotetraploid salmonid fishes. Note that if two autotetraploidy events occurred in relatively quick succession, such homeologous crossover could occur between any of the related chromosomes. In effect, this would be an octoploid (or pseudo-octoploid) phase. Such crossover is still likely to occur (to a lesser extent) after diploidization has commenced, in the phase of 'residual octoploidy'. Consequently, genes that were linked in the common ancestor of all vertebrates, and have several linked copies in extant vertebrates, may not have congruent tree topologies due to crossovers between the duplicate chromosomes during the diploidization process (figure 4b).

The implication is that paralogy regions, asymmetrical trees and non-congruent linked trees are all compatible with two sequential rounds of autotetraploidy. All have been detected in vertebrate genomes.

5. PROPOSAL

We propose that two rounds of genome duplication did indeed occur in the early evolution of the vertebrates. Our proposal differs from all previous ones, however, in respect of the mechanism. We have argued that only one mechanism is compatible with the data: that genome duplication occurred through two sequential rounds of autotetraploidy. This gave an autoauto-octoploid phase (or pseudo-octoploid phase), followed by gradual and sequential diploidization of the constituent chromosomes, homeologous crossover events and chromosome translocations.

The evidence in favour of this proposal, including examination of total gene numbers in genomes and phylogenetic analysis of gene families, provides overwhelming evidence that gene duplication was very widespread at the base of the vertebrate lineage. These data sources do not reveal the mechanism. The principal reasons for invoking octoploidy are the widespread occurrence of paralogy regions and asymmetrical gene trees, and incongruence between tree topologies of linked genes (explained by homeologous crossover). The octoploidy hypothesis

Table 2. Predictions made by the octoploidy hypothesis in comparison with some competing theories.

	no extraordinary duplication events	multiple independent gene duplications	single tetra ^a	two tetra ^a , at least one allo	two widely spaced autotetra ^b	two closely spaced autotetra ^b
(i) Total gene number will be higher for all vertebrate genomes than their closest invertebrate relatives. It will not necessarily be fourfold higher, but may be two- to threefold higher.		✓	✓	✓	✓	✓
(ii) Many single genes in invertebrates will have two, three or four homologous copies in vertebrates.		✓		✓	✓	✓
(iii) Phylogenetic trees for these gene families may show a sequential (asymmetrical) topology.		✓	✓			✓
(iv) For gene families with a higher copy in invertebrates, this will always be due to independent duplication.		✓	✓	✓	✓	✓
(v) Paralogy regions will have single-copy homologous regions in the genomes of basal chordates.			✓	✓	✓	✓
(vi) Much of the human genome will be made of severalfold paralogy regions.				✓	✓	✓
(vii) Gene families within linked paralogy regions will sometimes have congruent tree topologies, and sometimes not, due to crossovers during diploidization.						✓
(viii) The syntenic paralogy regions from divergent vertebrate taxa may have slightly different gene constitutions, due to exchange of gene positions by crossover.						✓

^a tetra, tetraploidization; ^b autotetra, autotetraploidization.

presented here also makes a series of specific predictions that can be tested by incorporation and analysis of additional data as shown in table 2.

As for timing, we have not discussed this issue in detail in this paper, although the gene family analyses point to much duplication occurring in the 'stem lineage' of vertebrates (i.e. after the lineage leading to amphioxus had diverged, and before divergence of the lineage, or lineages, leading to hagfish and lampreys). To allow for homeologous crossover, we propose that the two events must have occurred in relatively quick succession. They could either immediately 'straddle' the divergence point of lampreys and hagfish, or both could have occurred before this date (figure 1c).

We thank Mark Pagel, Seb Shimeld and Andrew C. R. Martin for advice and discussions, Françoise Mazet, Seb Shimeld and Filipe Castro for permission to cite results of their analyses,

and anonymous referees for constructive comments. The authors' work is funded by the BBSRC.

REFERENCES

- Adams, M. D. (and 194 others) 2000 The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195.
- Allendorf, F. W. & Thorngard, G. H. 1984 Tetraploidy and the evolution of salmonid fishes. In *Evolutionary genetics of fishes* (ed. B. J. Turner), pp. 1–48. New York: Plenum.
- Araki, I., Terazawa, K. & Satoh, N. 1996 Duplication of an amphioxus myogenic bHLH gene is independent of vertebrate myogenic bHLH gene duplication. *Gene* **171**, 231–236.
- Bailey, W. J., Kim, J., Wagner, G. P. & Ruddle, F. H. 1997 Phylogenetic reconstruction of vertebrate Hox cluster duplications. *Mol. Biol. Evol.* **14**, 843–853.
- Bird, A. P. 1995 Gene number, noise-reduction and biological complexity. *Trends Genet.* **11**, 94–100.

- Boeddrich, A., Burgtorf, C., Francis, F., Hennig, S., Panopoulou, G., Steffens, C., Borzym, K. & Lehrach, H. 1999 Sequence analysis of an amphioxus cosmid containing a gene homologous to members of the aldo-keto reductase gene superfamily. *Gene* **230**, 207–214.
- Brooke, N. M., Garcia-Fernandez, J. & Holland, P. W. H. 1998 The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster. *Nature* **392**, 920–922.
- Caenorhabditis elegans Sequencing Consortium 1998 Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018.
- Canestro, C., Hjelmqvist, L., Albalat, R., Garcia-Fernandez, J., Gonzalez-Duarte, R. & Jornvall, H. 2000 Amphioxus alcohol dehydrogenase is a class 3 form of single type and of structural conservation but with unique developmental expression. *Eur. J. Biochem.* **267**, 6511–6518.
- Chan, S. J., Cao, Q. P. & Steiner, D. F. 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.
- Coulier, F., Popovici, C., Villet, R. & Birnbaum, D. 2000 MetaHox gene clusters. *J. Exp. Zool.* **288**, 345–351.
- Dalfo, D., Canestro, C., Albalat, R. & Gonzalez-Duarte, R. 2001 Characterization of a microsomal retinol dehydrogenase gene from amphioxus: retinoid metabolism before vertebrates. *Chem. Biol. Interact.* **130–132**, 359–370.
- Escriva, H., Safi, R., Hanni, C., Langlois, M. C., Saumitou-Laprade, P., Stehelin, D., Capron, A., Pierce, R. & Laudet, V. 1997 Ligand binding was acquired during evolution of nuclear receptors. *Proc. Natl Acad. Sci. USA* **94**, 6803–6808.
- Ferrier, D. E. K., Minguillon, C., Holland, P. W. H. & Garcia-Fernandez, J. 2000 The amphioxus Hox cluster: deuterostome posterior flexibility and Hox14. *Evol. Dev.* **2**, 284–293.
- Ferrier, D. E. K., Minguillon, C., Cebrian, C. & Garcia-Fernandez, J. 2001 Amphioxus *Evx* genes: implications for the evolution of the midbrain–hindbrain boundary and the chordate tailbud. *Dev. Biol.* **237**, 270–281.
- Fletcher, C. F., Jenkins, N. A., Copeland, N. G., Chaudhry, A. Z. & Gronostajski, R. M. 1999 Exon structure of the nuclear factor I DNA-binding domain from *C. elegans* to mammals. *Mamm. Genome* **10**, 390–396.
- Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L. & Postlethwait, J. 1999 Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–1545.
- Garcia-Fernandez, J. & Holland, P. W. H. 1994 Archetypal organization of the amphioxus Hox gene-cluster. *Nature* **370**, 563–566.
- Gibert, J. M., Mouchel-Vielh, E. & Deutsch, J. S. 1997 *engrailed* duplication events during the evolution of barnacles. *J. Mol. Evol.* **44**, 585–594.
- Gibson, T. J. & Spring, J. 2000 Evidence in favour of ancient octoploidy in the vertebrate genome. *Biochem. Soc. Trans.* **28**, 259–264.
- Girardon, S., Holland, L. Z., Gehring, W. J. & Holland, N. D. 1998 Isolation and developmental expression of the amphioxus Pax-6 gene (*AmphiPax-6*): insights into eye and photoreceptor evolution. *Development* **125**, 2701–2710.
- Graber, N. A. & Ellington, W. R. 2001 Gene duplication events producing muscle (M) and brain (B) isoforms of cytoplasmic creatine kinase: cDNA and deduced amino acid sequences from two lower chordates. *Mol. Biol. Evol.* **18**, 1305–1314.
- Guo, X. M. & Allen, S. K. 1997 Sex and meiosis in autotetraploid Pacific oyster, *Crassostrea gigas* (Thunberg). *Genome* **40**, 397–405.
- Hall, T. A. 1999 BioEDIT: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl. Acids Symp. Ser.* **41**, 95–98.
- Hauber, D. P., Reeves, A. & Stack, S. M. 1999 Synapsis in a natural autotetraploid. *Genome* **42**, 936–949.
- Holland, L. Z., Pace, D. A., Blink, M. L., Kene, M. & Holland, N. D. 1995 Sequence and expression of amphioxus alkali myosin light-chain (*AmphiMLC-alk*) throughout development: implications for vertebrate myogenesis. *Dev. Biol.* **171**, 665–676.
- Holland, L. Z., Kene, M., Williams, N. A. & Holland, N. D. 1997 Sequence and embryonic expression of the amphioxus engrailed gene (*AmphiEn*): the metameric pattern of transcription resembles that of its segment-polarity homolog in *Drosophila*. *Development* **124**, 1723–1732.
- Holland, L. Z., Venkatesh, T. V., Gorlin, A., Bodmer, R. & Holland, N. D. 1998 Characterization and developmental expression of *AmphiNk2-2*, an NK2 class homeobox gene from amphioxus (Phylum Chordata; Subphylum Cephalochordata). *Dev. Genes Evol.* **208**, 100–105.
- Holland, L. Z., Schubert, M., Kozmik, Z. & Holland, N. D. 1999 *AmphiPax3/7*, an amphioxus paired box gene: insights into chordate myogenesis, neurogenesis, and the possible evolutionary precursor of definitive vertebrate neural crest. *Evol. Dev.* **1**, 153–165.
- Holland, L. Z., Holland, N. D. & Schubert, M. 2000a Developmental expression of *AmphiWnt1*, an amphioxus gene in the Wnt1/wingless subfamily. *Dev. Genes Evol.* **210**, 522–524.
- Holland, L. Z., Schubert, M., Holland, N. D. & Neuman, T. 2000b Evolutionary conservation of the presumptive neural plate markers *Amphisox1/2/3* and *AmphiNeurogenin* in the invertebrate chordate amphioxus. *Dev. Biol.* **226**, 18–33.
- Holland, L. Z., Rached, L. A., Tammé, R., Holland, N. D., Inoko, H., Shiina, T., Burgtorf, C. & Lardelli, M. 2001 Characterization and developmental expression of the amphioxus homolog of Notch (*AmphiNotch*): evolutionary conservation of multiple expression domains in amphioxus and vertebrates. *Dev. Biol.* **232**, 493–507.
- Holland, N. D., Holland, L. Z. & Kozmik, Z. 1995 An amphioxus Pax gene, *AmphiPax-1*, expressed in embryonic endoderm, but not in mesoderm: implications for the evolution of class-I paired box genes. *Mol. Mar. Biol. Biotechnol.* **4**, 206–214.
- Holland, N. D., Panganiban, G., Henyey, E. L. & Holland, L. Z. 1996 Sequence and developmental expression of *AmphiDll*, an amphioxus Distal-less gene transcribed in the ectoderm, epidermis and nervous system: insights into evolution of craniate forebrain and neural crest. *Development* **122**, 2911–2920.
- Holland, N. D., Zhang, S. C., Clark, M. & Panopoulou, G. 1997 Sequence and developmental expression of *AmphiTob*, an amphioxus homolog of vertebrate Tob in the PC3/BTG1/Tob family of tumor suppressor genes. *Dev. Dyn.* **210**, 11–18.
- Holland, P. W. H. 1996 Molecular biology of lancelets: insights into development and evolution. *Israel J. Zool.* **42**, S247–S272.
- Holland, P. W. H. 1999 Gene duplication: past, present and future. *Semin. Cell Dev. Biol.* **10**, 541–547.
- Holland, P. W. H., Garcia-Fernandez, J., Williams, N. A. & Sidow, A. 1994 Gene duplications and the origins of vertebrate development. *Development* (Suppl.), 125–133.
- Holland, P. W. H., Koschorz, B., Holland, L. Z. & Herrmann, B. G. 1995 Conservation of Brachyury (T) genes in amphioxus and vertebrates: developmental and evolutionary implications. *Development* **121**, 4283–4291.
- Hughes, A. L. 1998 Phylogenetic tests of the hypothesis of block duplication of homologous genes on human chromosomes 6, 9 and 1. *Mol. Biol. Evol.* **15**, 854–870.
- Hughes, A. L. 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.
- International Human Genome Sequencing Consortium 2001 Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.
- Jackman, W. R., Langeland, J. A. & Kimmel, C. B. 2000 *islet* reveals segmentation in the amphioxus hindbrain homolog. *Dev. Biol.* **220**, 16–26.
- Karabinos, A. & Bhattacharya, D. 2000 Molecular evolution of calmodulin and calmodulin-like genes in the cephalochordate *Branchiostoma*. *J. Mol. Evol.* **51**, 141–148.
- Karabinos, A., Riemer, D., Panopoulou, G., Lehrach, H. & Weber, K. 2000 Characterisation and tissue-specific expression of the two keratin subfamilies of intermediate filament proteins in the cephalochordate *Branchiostoma*. *Eur. J. Cell Biol.* **79**, 17–26.
- Kasahara, M. 1999 The chromosomal duplication model of the major histocompatibility complex. *Immunol. Rev.* **167**, 17–32.
- Kasahara, M., Hayashi, M., Tanaka, K., Inoko, H., Sugaya, K., Ikemura, T. & Ishibashi, T. 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.
- Katsanis, N., Fitzgibbon, J. & Fisher, E. M. C. 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.
- Knight, R. D., Panopoulou, G. D., Holland, P. W. H. & Shimeld, S. M. 2000 An amphioxus Krox gene: insights into vertebrate hindbrain evolution. *Dev. Genes Evol.* **210**, 518–521.
- Kozmik, Z., Holland, N. D., Kalousova, A., Paces, J., Schubert, M. & Holland, L. Z. 1999 Characterization of an amphioxus paired box gene, *AmphiPax2/5/8*: developmental expression patterns in optic support cells, nephridium, thyroid-like structures and pharyngeal gill slits, but not in the midbrain–hindbrain boundary region. *Development* **126**, 1295–1304.
- Kozmik, Z., Holland, L. Z., Schubert, M., Lacalli, T. C., Kreslova, J., Vlcek, C. & Holland, N. D. 2001 Characterization of amphioxus *amphivent*, an evolutionarily conserved marker for chordate ventral mesoderm. *Genesis* **29**, 172–179.
- Kuba, M., Yatsuki, H., Kusakabe, T., Takasaki, Y., Nikoh, N., Miyata, T., Yamaguchi, T. & Hori, K. 1997 Molecular evolution of amphioxus fructose-1,6-bisphosphate aldolase. *Arch. Biochem. Biophys.* **348**, 329–336.
- Kusakabe, R., Satoh, N., Holland, L. Z. & Kusakabe, T. 1999 Genomic organization and evolution of actin genes in the amphioxus *Branchiostoma belcheri* and *Branchiostoma floridae*. *Gene* **227**, 1–10.
- Langeland, J. A., Tomsa, J. M., Jackman, W. R. & Kimmel, C. B. 1998 An amphioxus *snail* gene: expression in paraxial mesoderm and neural plate suggests a conserved role in patterning the chordate embryo. *Dev. Genes Evol.* **208**, 569–577.
- Langlois, M. C., Vanacker, J. M., Holland, N. D., Escriva, H., Queva, C., Laudet, V. & Holland, L. Z. 2000 *Amphicouptf*, a nuclear orphan receptor of the lancelet *Branchiostoma floridae*, is implicated in retinoic acid signalling pathways. *Dev. Genes Evol.* **210**, 471–482.
- Li, W.-H. 1980 Rate of gene silencing at duplicate loci: a theoretical study and interpretation of data from tetraploid fishes. *Genetics* **95**, 237–258.
- Luke, G. N. & Holland, P. W. H. 1999 Amphioxus type I keratin cDNA and the evolution of intermediate filament genes. *J. Exp. Zool.* **285**, 50–56.
- 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.
- Lundin, L. G. 1995 Paralogous genes and metazoan macroevolution. In *New frontiers in theoretical biology* (ed. C. A. C. Dreisman), pp. 3–51. Palm Harbor, FL: Hadronic.
- Lundin, L. G. & Larhammar, D. 1998 Paralogous genes and nervous systems. In *Genetics and psychiatric disorders* (ed. J. Wahlstrom), pp. 27–56. Oxford, UK: Pergamon.
- Martin, A. P. 2001 Is tetralogy real: testing the assumptions of the ‘one to four’ rule. *Mol. Biol. Evol.* **18**, 89–93.
- Meyer, A. & Schartl, M. 1999 Gene and genome duplications in vertebrates: the one-to-four (–to-eight in fish) rule and the evolution of novel gene functions. *Curr. Opin. Cell Biol.* **11**, 699–704.
- Minguillón, C., Ferrier, D. E. K., Cebrian, C. & Garcia-Fernandez, J. 2002 Gene duplications in the prototypical cephalochordate amphioxus. *Gene*. (In the press.)
- Neider, A. H., Panopoulou, G. & Langeland, J. A. 2000 Amphioxus *gooseoid* and the evolution of the head organizer and prechordal plate. *Evol. Dev.* **2**, 303–310.
- Nikoh, N. (and 10 others) 1997 An estimate of divergence time of Parazoa and Eumetazoa and that of Cephalochordata and Vertebrata by aldolase and triose phosphate isomerase clocks. *J. Mol. Evol.* **45**, 97–106.
- Ogasawara, M. 2000 Overlapping expression of amphioxus homologs of the thyroid transcription factor-1 gene and thyroid peroxidase gene in the endostyle: insight into evolution of the thyroid gland. *Dev. Genes Evol.* **210**, 231–242.
- Ohno, S. 1970 *Evolution by gene duplication*. Berlin: Springer.
- Ohno, S. 1998 The notion of the Cambrian pananimalia genome and a genomic difference that separated vertebrates from invertebrates. In *Molecular evolution: towards the origin of metazoa*, vol. 21 (ed. W. E. G. Müller), Progress in molecular and subcellular biology series, pp. 97–117. Berlin: Springer.
- Oliva, A. A., Steiner, D. F. & Chan, S. J. 1995 Proprotein convertases in amphioxus: predicted structure and expression of proteases Spcb and Spc3. *Proc. Natl Acad. Sci. USA* **92**, 3591–3595.
- Oliva, A. A., Chan, S. J. & Steiner, D. F. 2000 Evolution of the prohormone convertases: identification of a homologue of PC6 in the protochordate amphioxus. *Biochem. Biophys. Acta* **1477**, 338–348.
- 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.
- Panopoulou, G. D., Clark, M. D., Holland, L. Z., Lehrach, H. & Holland, N. D. 1998 *AmphiBMP2/4*, an amphioxus bone morphogenetic protein closely related to *Drosophila* decapentaplegic and vertebrate BMP2 and BMP4: insights into evolution of dorsoventral axis specification. *Dev. Dyn.* **213**, 130–139.
- Pashmforoush, M., Chan, S. J. & Steiner, D. F. 1996 Structure and expression of the insulin-like peptide receptor from amphioxus. *Mol. Endocrinol.* **10**, 857–866.
- Patton, S. J., Luke, G. N. & Holland, P. W. H. 1998 Complex history of a chromosomal paralogy region: insights from amphioxus aromatic amino acid hydroxylase genes and insulin-related genes. *Mol. Biol. Evol.* **15**, 1373–1380.
- Pollard, S. L. & Holland, P. W. H. 2000 Evidence for 14 homeobox gene clusters in human genome ancestry. *Curr. Biol.* **10**, 1059–1062.
- Riemer, D., Wang, J., Zimek, A., Swalla, B. J. & Weber, K. 2000 Tunicates have unusual nuclear lamins with a large deletion in the carboxyterminal tail domain. *Gene* **255**, 317–325.
- Roberts, R. G. & Bobrow, M. 1998 Dystrophins in vertebrates and invertebrates. *Hum. Mol. Genet.* **7**, 589–595.
- Ruddle, F. H., Bentley, K. L., Murtha, M. T. & Risch, N.

- 1994 Gene loss and gain in the evolution of the vertebrates. *Development* (Suppl.), 155–161.
- Ruvinsky, I., Silver, L. M. & Gibson-Brown, J. J. 2000 Phylogenetic analysis of T-Box genes demonstrates the importance of amphioxus for understanding evolution of the vertebrate genome. *Genetics* **156**, 1249–1257.
- Schlake, T., Schorpp, M., Nehls, M. & Boehm, T. 1997 The *nude* gene encodes a sequence-specific DNA binding protein with homologs in organisms that lack an anticipatory immune system. *Proc. Natl Acad. Sci. USA* **94**, 3842–3847.
- Schlake, T., Schorpp, M. & Boehm, T. 2000 Formation of regulator/target gene relationships during evolution. *Gene* **256**, 29–34.
- Schmid, M., Haaf, T. & Schempp, W. 1985 Chromosome-banding in amphibia. 9. The polyploid karyotypes of *Odontophrynus americanus* and *Ceratophrys ornata* (Anura, Leptodactylidae). *Chromosoma* **91**, 172–184.
- Schmidtke, J., Weiler, C., Kunz, B. & Engel, W. 1977 Isozymes of a tunicate and a cephalochordate as a test of polyploidisation in chordate evolution. *Nature* **266**, 532–533.
- Schubert, M., Holland, N. D. & Holland, L. Z. 1998 Amphioxus *AmphiDRAL* encoding a LIM-domain protein: expression in the epidermis but not in the presumptive neuroectoderm. *Mech. Dev.* **76**, 203–205.
- Schubert, M., Holland, L. Z. & Holland, N. D. 2000a Characterization of two amphioxus Wnt genes (*AmphiWnt4* and *AmphiWnt7b*) with early expression in the developing central nervous system. *Dev. Dyn.* **217**, 205–215.
- Schubert, M., Holland, L. Z. & Holland, N. D. 2000b Characterization of an amphioxus Wnt gene, *AmphiWnt11*, with possible roles in myogenesis and tail outgrowth. *Genesis* **27**, 1–5.
- Schubert, M., Holland, L. Z., Holland, N. D. & Jacobs, D. K. 2000c 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.
- Schubert, M., Holland, L. Z., Panopoulou, G. D., Lehrach, H. & Holland, N. D. 2000d Characterization of amphioxus *AmphiWnt8*: insights into the evolution of patterning of the embryonic dorsoventral axis. *Evol. Dev.* **2**, 85–92.
- Sedlacek, Z., Shimeld, S. M., Munstermann, E. & Poustka, A. 1999 The amphioxus rab GDP-dissociation inhibitor (GDI) gene is neural-specific: implications for the evolution of chordate rab GDI genes. *Mol. Biol. Evol.* **16**, 1231–1237.
- Sharman, A. C. & Holland, P. W. H. 1996 Conservation, duplication, and divergence of developmental genes during chordate evolution. *Neth. J. Zool.* **46**, 47–67.
- Sharman, A. C., Shimeld, S. M. & Holland, P. W. H. 1999 An amphioxus *Msx* gene expressed predominantly in the dorsal neural tube. *Dev. Genes Evol.* **209**, 260–263.
- Shimeld, S. M. 1997 Characterisation of amphioxus HNF-3 genes: conserved expression in the notochord and floor plate. *Dev. Biol.* **183**, 74–85.
- Shimeld, S. M. 1998 Characterization of *AmphiF-spondin* reveals the modular evolution of chordate F-spondin genes. *Mol. Biol. Evol.* **15**, 1218–1223.
- Shimeld, S. M. 1999 The evolution of the hedgehog gene family in chordates: insights from amphioxus hedgehog. *Dev. Genes Evol.* **209**, 40–47.
- Shimeld, S. M. 2000 An amphioxus netrin gene is expressed in midline structures during embryonic and larval development. *Dev. Genes Evol.* **210**, 337–344.
- Shimeld, S. M. & Holland, P. W. H. 2000 Vertebrate innovations. *Proc. Natl Acad. Sci. USA* **97**, 4449–4452.
- Sidow, A. 1992 Diversification of the Wnt gene family on the ancestral lineage of vertebrates. *Proc. Natl Acad. Sci. USA* **89**, 5098–5102.
- Sidow, A. 1996 Gen(om)e duplications in the evolution of early vertebrates. *Curr. Opin. Genet. Dev.* **6**, 715–722.
- Simmen, M. W., Leitgeb, S., Clark, V. H., Jones, S. J. M. & Bird, A. 1998 Gene number in an invertebrate chordate, *Ciona intestinalis*. *Proc. Natl Acad. Sci. USA* **95**, 4437–4440.
- Skrabanek, L. & Wolfe, K. H. 1998 Eukaryote genome duplication: where's the evidence? *Curr. Opin. Genet. Dev.* **8**, 694–700.
- Smith, M. W. & Doolittle, R. F. 1992 A comparison of evolutionary rates of the two major kinds of superoxide-dismutase. *J. Mol. Evol.* **34**, 175–184.
- Smith, N. G. C., Knight, R. & Hurst, L. D. 1999 Vertebrate genome evolution: a slow shuffle or a big bang? *BioEssays* **21**, 697–703.
- Spring, J. 1997 Vertebrate evolution by interspecific hybridisation: are we polyploid? *Fed. Eur. Biol. Soc. 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.
- Sutherland, D. (and 10 others) 1997 Two cholinesterase activities and genes are present in amphioxus. *J. Exp. Zool.* **277**, 213–229.
- Suzuki, M. M. & Satoh, N. 2000 Genes expressed in the amphioxus notochord revealed by EST analysis. *Dev. Biol.* **224**, 168–177.
- Toresson, H., Martinez-Barbera, J. P., Bardsley, A., Caubit, X. & Krauss, S. 1998 Conservation of BF-1 expression in amphioxus and zebrafish suggests evolutionary ancestry of anterior cell types that contribute to the vertebrate telencephalon. *Dev. Genes Evol.* **208**, 431–439.
- Tweedie, S., Charlton, J., Clark, V. & Bird, A. 1997 Methylation of genomes and genes at the invertebrate–vertebrate boundary. *Mol. Cell. Biol.* **17**, 1469–1475.
- Venkatesh, T. V., Holland, N. D., Holland, L. Z., Su, M. T. & Bodmer, R. 1999 Sequence and developmental expression of amphioxus *AmphiNk2-1*: insights into the evolutionary origin of the vertebrate thyroid gland and forebrain. *Dev. Genes Evol.* **209**, 254–259.
- Venter, J. C. (and 273 others) 2001 The sequence of the human genome. *Science* **291**, 1304–1351.
- Wada, H., Saiga, H., Satoh, N. & Holland, P. W. H. 1998 Tripartite organisation of the ancestral chordate brain and the antiquity of placodes: insights from ascidian *Pax-2/5/8*, *Hox* and *Otx* genes. *Development* **125**, 1113–1122.
- Wendel, J. F. 2000 Genome evolution in polyploids. *Plant Mol. Biol.* **42**, 225–249.
- Williams, N. A. & Holland, P. W. H. 1998 Gene and domain duplication in the chordate *Otx* gene family: insights from amphioxus *Otx*. *Mol. Biol. Evol.* **15**, 600–607.
- Williams, N. A. & Holland, P. W. H. 2000 An amphioxus *Emx* homeobox gene reveals duplication during vertebrate evolution. *Mol. Biol. Evol.* **17**, 1520–1528.
- Yasui, K., Zhang, S. C., Uemura, M., Aizawa, S. & Ueki, T. 1998 Expression of a twist-related gene, *Bbtwist*, during the development of a lancelet species and its relation to cephalochordate anterior structures. *Dev. Biol.* **195**, 49–59.
- Yasui, K., Zhang, S. C., Uemura, M. & Saiga, H. 2000 Left-right asymmetric expression of *BbPtx* a *Ptx*-related gene, in a lancelet species and the developmental left-sidedness in deuterostomes. *Development* **127**, 187–195.
- Yuasa, H. J., Cox, J. A. & Takagi, T. 1998 Diversity of the troponin C genes during chordate evolution. *J. Biochem.* **123**, 1180–1190.
- Yuasa, H. J., Cox, J. A. & Takagi, T. 1999 Genomic structure of the amphioxus calcium vector protein. *J. Biochem.* **126**, 572–577.