Eukaryote genome duplication – where's the evidence?

Lucy Skrabanek and Kenneth H Wolfe*

Several eukaryotes, including maize, yeast and *Xenopus*, are degenerate polyploids formed by relatively recent whole-genome duplications. Ohno's conjecture that more ancient genome duplications occurred in an ancestor of vertebrates is probably at least partly true but the present shortage of gene sequence and map information from vertebrates makes it difficult to either prove or disprove this hypothesis. Candidate paralogous segments in mammalian genomes have been identified but the lack of statistical rigour means that many of the proposals in the literature are probably artefacts.

Addresses
Department of Genetics, University of Dublin, Trinity College, Dublin 2, Ireland
*e-mail: khwolfe@tcd.ie

Current Opinion in Genetics & Development 1998, **8**:694–700

http://biomednet.com/elecref/0959437X00800694

Abbreviations

BLAST	basic local alignment search tool
EST	expressed sequence tag
HSA	human chromosome
Mya	million years ago
NR	nuclear receptor

Introduction

Ohno's hypothesis that multiple genome duplications occurred in an ancestor of vertebrates [1,2] has been enduringly popular with scientists even though many of the original premises behind his proposal have turned out to be incorrect. For example, his arguments about genome size differences were made before junk DNA was discovered, and the first pair of regions in the human genome proposed to be a duplicated chromosomal segment seem, ironically, to have been caused by an error in the genetic map. (Ohno [2] noted that the LDHA [lactate dehydrogenase] and GPTC [glutamate pyruvate transaminase] genes were mapped to human chromosome 11 [HSA11] whereas their homologues LDHB and GPTB were mapped to HSA12. The GPT data was in error; there is only one GPT gene, and it is on HSA8.) Despite the growing tendency to refer to whole-genome duplications in vertebrates as fact [3,4], the data remain severely limited and Ohno's idea remains an hypothesis. In this review, we examine the evidence that genome duplications have indeed occurred during the evolution of all eukaryotes, including vertebrates.

Lessons from non-vertebrates

The strongest evidence for genome duplications comes not from vertebrates but from yeast and maize. In yeast, 55 duplicated chromosomal regions account for half the genome [5–7]. Duplicated genes in these regions have conserved gene order and orientation but they are outnumbered by unique (non-duplicated) genes located between them. The unique genes must originally have been duplicated along with the rest of the genome, but one copy was subsequently deleted.

A similar degenerate tetraploid structure in the maize genome was first recognised by Helentjaris *et al.* [8] and Ahn and Tanksley [9]. The entire genome can be sorted into paired regions on the basis of the conserved order of duplicate genes [10,11•,12]. Recent results from the *Arabidopsis thaliana* genome sequencing project indicate unexpectedly that it, too, has substantial regional duplications (S Rounsley, personal communication; [13]).

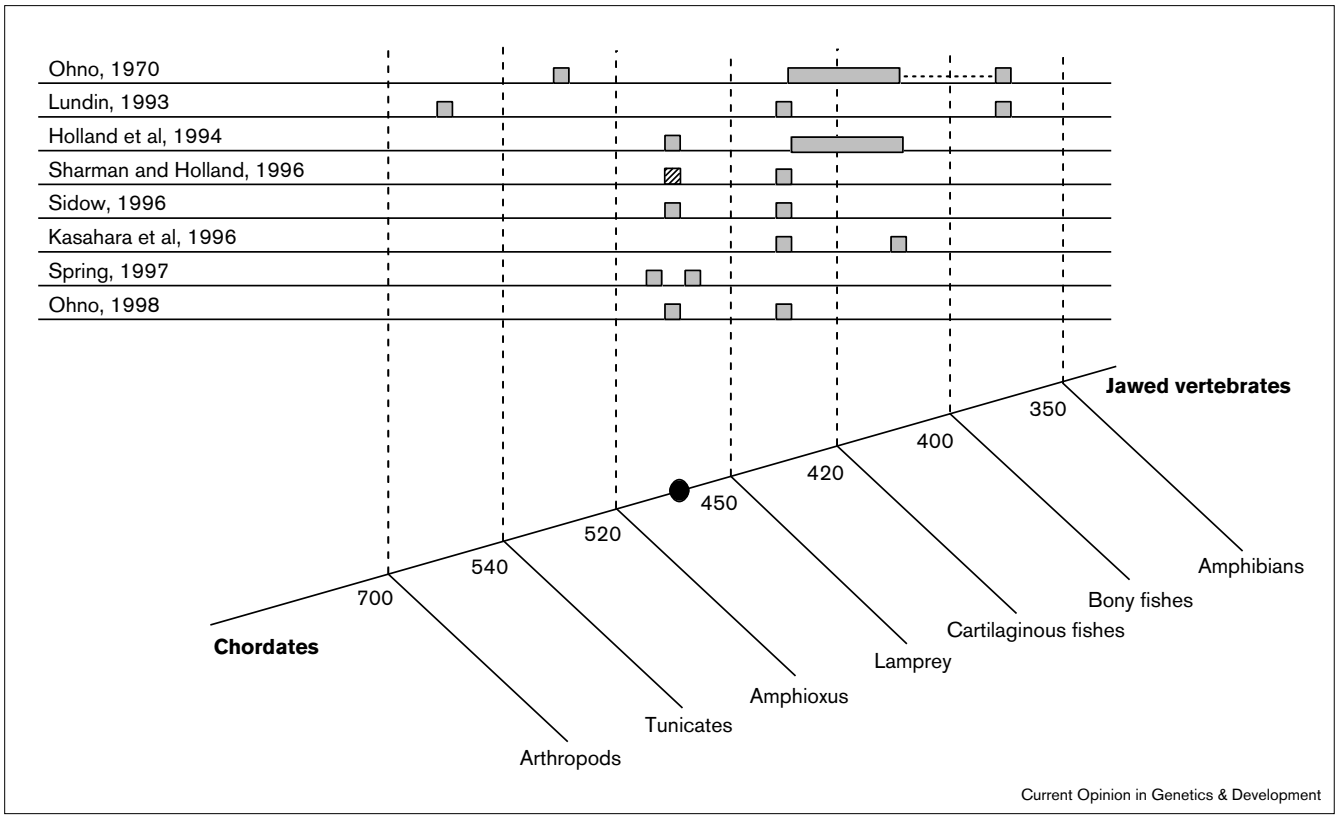
Neither yeast nor maize contains entire duplicated chromosomes. Instead, they have multiple non-overlapping duplicated regions, which indicates that numerous chromosomal rearrangements occurred after genome duplication [11•,14]. In yeast, most of the major rearrangements were reciprocal translocations. If a single duplication of the whole genome occurred, it might be expected that molecular clock analyses of different gene pairs should all converge on a single estimate of the date of the duplication; however, date estimates from both yeast [5] and maize [15••] are quite heterogeneous, which could be explained by either gene conversion or segmental allotetraploidy. (A segmental allotetraploid is an organism where parts of the genome are allotetrapoid and the rest is autotetraploid [15••].)

Yeast and maize satisfy three criteria for proof that they have duplicated genomes: conserved gene order in paired chromosomal regions; these regions are non-overlapping; and phylogenetic support for a 2:1 orthology relationship with an outgroup. In yeast, a fourth criterion is met: an outgroup species whose gene order for unique genes is similar to that of the inferred pre-duplication genome [16]. Lack of genome sequence data means that this fourth test cannot yet be applied to maize.

Fates and functions of duplicated genes

In yeast, ~8% of the original gene set were retained in duplicate and the other 92% returned to a single-copy state [14]. This is at odds with theoretical predictions that if a species has a large effective population size (as does yeast; [17]) most duplicate genes should gain new functions and not become pseudogenes [18–20]. The 8% figure also contrasts with retention estimates of 72% for maize [9] and ~50% for recent tetraploidies in fish and *Xenopus* [18,21]. Does this reflect some biological difference, or is there an ascertainment bias caused by the different methods — complete sequence analysis, cDNA hybridisation, or isozyme studies — used to produce the

Figure 1



Summary of proposals for the timing of duplication events in the vertebrate lineage. Species divergences are drawn schematically, not to scale. Shaded boxes indicate each proposed genome duplication and are drawn at the centre of possible time ranges (Mya). The hatched box indicates a proposed wave of multiple tandem gene duplications. Data from references [1,20,26,29,34,37,42,53]. Ohno [1] postulated tetraploidisation at the divergence of fish and/or amphibians, shown by the two boxes connected by a dotted line. The circle at ~500 Mya denotes the origin of vertebrates. Other proposals not shown here include suggestions that an additional (chromosomal or whole genome) duplication may have occurred in the zebrafish lineage after its divergence from the lineage leading to mammals [47••,49], or conversely that the most recent duplication in the mammalian lineage might have occurred after its divergence from bony fish [54].

estimates? One obvious biological difference is unicellularity in yeast versus multicellularity in the other species. Cooke and co-workers [22•,23•] have noted that several genes important in mammalian development have no phenotype when knocked out in mouse, proposing that there must be selective pressure to maintain redundancy of developmental genes. This has been disputed by Gibson and Spring [24•] who argue that it may simply be harder to get rid of duplicated multidomain proteins (e.g. developmental proteins) than duplicated single-domain proteins (e.g. metabolic enzymes). Gibson and Spring's hypothesis is supported by Iwabe *et al.* [25], who found that different functional classes of genes have been duplicated to differing degrees during vertebrate evolution.

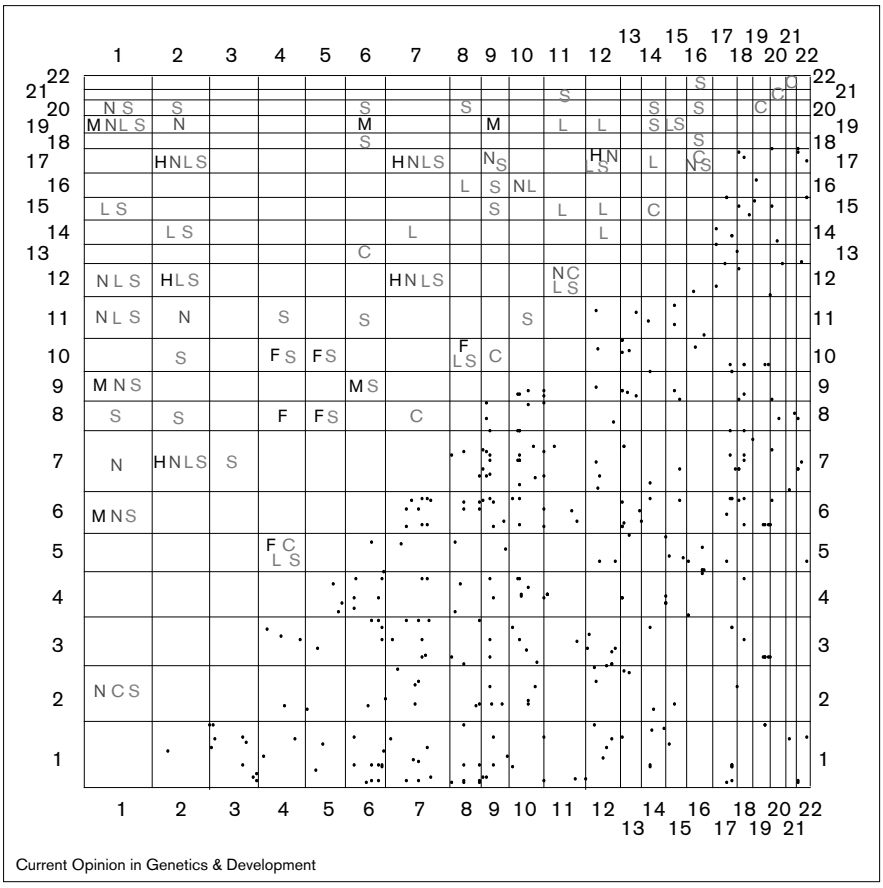
What has been proposed for vertebrates?

Different proposals have been made concerning the number and timing of genome duplications during vertebrate evolution (Figure 1). In his classic book [1], Ohno did not make an explicit hypothesis but instead proposed one duplication after the divergence of tunicates, followed by one or two others some time between the lamprey and amphibian divergences. More recent proposals, including one by Ohno himself [26], have made more precise statements and the most common model — influenced largely by studies on the Hox gene clusters — proposes two duplications: one on either side of the jawless fish divergence (~500 and 430 million years ago [Mya]; Figure 1).

Gene number arguments

The most straightforward argument for vertebrate genome duplications comes from the analysis of gene numbers in different species. *Drosophila* and *Caenorhabditis* have ~12,000 and 16,000 genes respectively. An elegant analysis [27••] estimates a similar gene number (15,000 ± 3,700) for the sea squirt *Ciona intestinalis*, a tunicate (Figure 1). Humans are estimated by expressed sequence tag (EST) analysis to have ~70,000 (± 20,000) genes [28]. The approximate fourfold ratio between these numbers is consistent with two rounds of polyploidy in vertebrates, although the human estimate is quite uncertain and the ratio between human and *Drosophila* could be as high as eightfold. This sort of arithmetic makes no allowance for gene deletion during diploidisation, however.

Figure 2



Above diagonal (bottom left to top right): human chromosome pairs that have been proposed to contain duplicated regions. H, Hox regions [43]; M, MHC regions [36,37]; F, FGFR regions [51]; N, Nadeau [55]; C, Comings [32]; L, Lundin [34]; S, Spring [29]. Below diagonal: dot-matrix summary of TBLASTX [56] search results. Human sequence pairs with significant similarity are plotted at the rank-order positions of the sequences on a physical map of chromosomes. The dataset – 13,403 mapped sequences including EST clusters, cDNAs, and genomic DNAs – was downloaded from the Genomes Division of National Center for Biotechnology Information (NCBI) Entrez Database [57] in December 1997. The criteria for similarity were a TBLASTX score of 200 (BLOSUM62 matrix), a minimum sequence overlap length of 200 bp, and an excess of synonymous nucleotide substitutions. Sequences were pre-filtered to remove repeats using DUST [58] for low-complexity regions and XBLAST [59] with the REPBASE database [60] for repetitive elements. Long sequences were broken into 20 kb fragments before searching.

To paraphrase HL Mencken, for every problem there is an explanation that is neat, plausible, and wrong.

1:4 relationships

Spring [29] and Sidow [20] have observed that many single-copy *Drosophila* genes have four vertebrate orthologues and have proposed that this is consistent with two rounds of genome duplication in vertebrates. Apparent 1:3, 1:2 or 1:1 relationships between *Drosophila* and human were attributed to either gene deletion or inadequate sampling from human. Genes giving 1:5 or higher ratios were predicted to be paralogous mixtures that should resolve (given complete data) into 2:8 ratios [29]. But the difference in gene numbers between these species, and the dictum that new genes are always made by duplication, mean that almost no other result is possible. There must, on average, be about four human homologues for every *Drosophila* gene. Whereas two rounds of complete genome duplication would give 1:4 ratios for every gene, a simple computer simulation shows that even if 20,000 *Drosophila* genes were associated completely at random with 80,000 human genes, 60% of them would be in 1:2, 1:3 or 1:4 ratios.

Phylogenetic trees can be used to test the significance of the observed 1:4 relationships. If two rounds of genome duplication occurred, a tree for four vertebrate sequences and an outgroup should have the topology (outgroup [(1,2)(3,4)]), where the first genome duplication produced the common ancestors of sequences 1/2 and 3/4, and these two lineages later split simultaneously in the second genome duplication [29]. Such a tree has two testable characteristics: a '2+2' topology, and equal ages for the two later divergences, but few attempts have been made to examine these for candidate genes.

The difficulty in interpreting phylogenetic data for complex gene families is illustrated by two recent studies on a single data set. Baker [30] analysed steroid hormone receptors, which form part of the nuclear receptor (NR) superfamily analysed independently by Escriva *et al.* [31]. Both groups concluded that they had identified two rounds of gene duplication in their trees. Baker's two rounds were those giving rise to a 2+2 topology for the androgen, progesterone, glucocorticoid and mineralcorticoid receptors, both ~450 Mya. Escriva *et al.* instead regarded these events as a single 'wave' of duplications and proposed that there was also a much earlier 'wave', before the *Hydra* divergence >700 Mya, which created the six NR subfamilies (of which the steroid hormone receptors are one). Despite the differences, both groups stated that their results support Ohno's hypothesis.

Candidate duplicate or quadruplicate regions

In 1972, Comings [32] classified the human chromosomes into 11 pairs on the basis of cytological similarities but this work must be questioned because of the extent of rearrangement now known to exist between human and mouse [33]. Some of Coming's pairs, such as HSA11/HSA12 and HSA4/HSA5, are still often cited. Lundin [34] and Spring [29] listed numerous duplicated, triplicated or quadruplicated regions each containing several genes. Nadeau and Sankoff [4] analysed a set of hundreds of duplicate genes in human and mouse that they took to be derived from genome duplications but they did not identify the genes.

The sheer number and remarkable diversity of chromosome pairs that have been proposed for humans makes it unlikely that most of them can be correct. Of the 231 possible pairs that can be made from 22 autosomes, at least 66 (29%) have been proposed to contain ancient paralogous segments (Figure 2). Many of these proposals have been made on the strength of no more than two or three similar genes being located on the two chromosomes. This is clearly insufficient, given that an average human chromosome may contain 4000 genes. By performing BLAST (basic local alignment search tool) searches among all DNA sequences that appear on the physical map of human chromosomes, we find at least three pairs of similar sequences on 48 of the 231 possible autosome combinations, including 25 new ones (Figure 2). Thus it would be possible, using the criteria commonly used in the literature, to propose dozens more ancient paralogous regions. For example, HSA3 and HSA10 have not been proposed to be related to each other but they contain four hits with conserved gene order: hormone receptors (GenBank accession numbers L31785 and X68167), ribonucleoproteins (R39545 and AA088775), zinc finger proteins (U69645 and L04282), and protein kinases (L18964 and L01087).

Clearly, more stringent criteria are needed to distinguish genuine polyploidy-derived duplicate regions from artefacts. The only study so far that has adopted a statistical approach to its findings has been that of Ruddle *et al.* [35], who concluded that the HSA 2/7/12/17 relationship was highly significant. Similar approaches need to be taken with other candidate regions.

Three potentially quadruplicated regions the MHC, Hox and FGFR regions — in mammals — have been analysed in detail: these are now discussed.

The HSA 1/6/9/19 (MHC) regions

Katsanis *et al.* [36] and Kasahara *et al.* [37,38] independently identified genes near the MHC region on HSA6 that had homologues on three other chromosomes. They proposed two rounds of chromosomal duplication events and suggested that these may have been part of one or two genome-wide duplications (Figure 1). The hypothesis that the genes duplicated simultaneously was recently examined critically by Endo *et al.* [39••] and Hughes [40••]. Their phylogenetic analyses showed that, out of the 11 gene pairs that had been proposed on HSA6/HSA9, six may have had a simultaneous origin. This makes a block duplication the most parsimonious explanation of the data for these six gene pairs, even though simultaneous duplication of all 11 pairs could be resoundingly rejected. For three of the six gene pairs there is a third copy on HSA1q21-25 which in each case is slightly more closely related to the HSA9 paralogue than to the HSA6 paralogue [36,40••].

The HSA 2/7/12/17 (Hox) regions

Although the Hox regions — including nearby genes such as *Wnt* and *Dlx* — are probably the most-cited example of genes whose organisation supports the hypothesis of two rounds of genome duplication in vertebrates [41,42], in the past year it has become clear that the explanation may be more complex. Bailey *et al.* [43] have analysed sequence data from collagen genes linked to the Hox clusters, in conjunction with the Hox sequences themselves. Instead of the 2+2 topology expected for a model with 1→2→4 Hox clusters, they found strong bootstrap support for a tree where the HoxD cluster branched off first from the ancestral lineage, followed by HoxA, and finally HoxB/C. This requires three separate duplication steps. As mammals have only four Hox clusters, not eight, either some of these three steps were regional (not whole-genome) duplications, or else some Hox clusters were later deleted during mammalian evolution. An intermediate number of three clusters in lamprey seems to uphold the former view [44] and the hypothesis is testable because sequences from lamprey should contain orthologues of Hox clusters D, A and a B/C ancestor.

As with mammals, the pufferfish *Fugu* has four Hox clusters. Three of these correspond to the mammalian Hox A, B and C clusters but the fourth is so unlike mammalian HoxD that Aparicio *et al.* [45] were unable to tell if they were orthologues. They speculated that an ancestor of *Fugu* might have had more than four Hox clusters, with *Fugu* having lost HoxD completely. This is supported by the discovery that the zebrafish has at least five or six Hox clusters [46,47••]. Whether these extra clusters represent duplications that are zebrafish-specific, teleost-specific, or common to the ancestor of all vertebrates (but later deleted in mammals) is unknown [45,47••,48,49].

The recent discovery of the "ParaHox" cluster in amphioxus points to an even older duplication [50••]. Gene order, expression patterns, and phylogenetic analysis all indicate that ParaHox is a duplicate of a primordial Hox cluster, which arose in an ancestor of amphioxus and vertebrates.

The HSA 4/5/8/10 (FGFR) regions

A third possible example of a quadruplicated region was described recently by Pébusque *et al.* [51]. This is centred on four fibroblast growth receptor genes that are near adrenergic receptor genes on human chromosomes 4p16,

5q33-35, 8p12-21 and 10q24-26. Additional genes allow the region to be extended, most impressively for the HSA8/HSA10 pair which includes seven loci. Where data are available, phylogenetic analysis indicates a 1:4 relationship between invertebrate and mammalian sequences for these genes. Pébusque *et al.* argue that these duplicated genes arose before the bony fish divergence but they did not use molecular clocks to estimate dates for each gene.

Conclusions

Take four, or maybe eight, decks of 52 playing cards. Shuffle them all together and then throw some cards away. Pick 20 cards at random and drop the rest on the floor. Give the 20 cards to some evolutionary biologists and ask them to figure out what you've done. For encouragement, tell them they can have the cards on the floor in 2005 [52•].

Whole-genome duplication via polyploidy has undoubtedly occurred relatively recently in representatives of three major eukaryote kingdoms: maize, yeast, *Xenopus* and some fish. Parsimony therefore says that genome duplication probably occurred several times in the evolution of all eukaryotic lineages, including our own, but traces of these events may be hard to detect. Because of the paucity of available map and sequence data, it is premature to reach any conclusion about Ohno's original hypothesis for vertebrates [1].

A particularly acute problem is that we do not really know what we are looking for and it appears to be only too easy to adapt the hypothesis to fit any data [40••]. Polyploid genomes decay back to diploidy, eventually leaving only a larger proteome and perhaps a few fragments of conserved gene order as evidence that anything special has happened. Without knowledge of the number of genome duplication events that happened, and the relative rates of processes — such as gene deletion and transposition — that obscure the evidence for duplication, building statistical models with which to evaluate candidate duplicated regions will be a challenge to bioinformatics.

Acknowledgements

We thank Jinghui Zhang of the National Center for Biotechnology Information for help with human sequence data. Research in our laboratory is supported by the European Community 4th Framework Biotechnology Programme (BIO4-CT95-0130).

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Ohno S: *Evolution by Gene Duplication.* London: George Allen and Unwin; 1970.

2. Ohno S: **Ancient linkage groups and frozen accidents.** *Nature* 1973, **244**:259-262.

3. Henikoff S, Greene EA, Pietrokovski S, Bork P, Attwood TK, Hood L: **Gene families: the taxonomy of protein paralogs and chimeras.** *Science* 1997, **278**:609-614.

4. Nadeau JH, Sankoff D: **Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution.** *Genetics* 1997, **147**:1259-1266.

5. Wolfe KH, Shields DC: **Molecular evidence for an ancient duplication of the entire yeast genome.** *Nature* 1997, **387**:708-713.

6. Coissac E, Maillier E, Netter P: **A comparative study of duplications in bacteria and eukaryotes: the importance of telomeres.** *Mol Biol Evol* 1997, **14**:1062-1074.

7. Mewes HW, Albermann K, Bähr M, Frishman D, Gleissner A, Hani J, Heumann K, Kleine K, Maierl A, Oliver SG *et al.*: **Overview of the yeast genome.** *Nature* 1997, **387(Suppl)**:7-65.

8. Helentjaris T, Weber D, Wright S: **Identification of the genomic locations of duplicate nucleotide sequences in maize by analysis of restriction fragment length polymorphisms.** *Genetics* 1988, **118**:353-363.

9. Ahn S, Tanksley SD: **Comparative linkage maps of the rice and maize genomes.** *Proc Natl Acad Sci USA* 1993, **90**:7980-7984.

10. Moore G, Devos KM, Wang Z, Gale MD: **Cereal genome evolution: grasses, line up and form a circle.** *Curr Biol* 1995, **5**:737-739.

11. • Gale MD, Devos KM: **Comparative genetics in the grasses.** *Proc Natl Acad Sci USA* 1998, **95**:1971-1974.
In our opinion, although there is strong evidence both that the maize genome is an ancient tetraploid and that there is substantial conservation of gene order among grasses, the 'Lego' model of Gale and co-workers is misleading and questionable. The circular representation of the aligned genomes of different species (including two putative sub-genomes from maize) implies that either there have been no chromosomal fusions or translocations during grass evolution – in which case the ancestor of the grasses must have had just a single, giant, chromosome; [12] – or else that chromosomal fusions occur but each chromosome is only permitted to fuse with a particular designated partner. Neither of these seems plausible.

12. Moore G, Foote T, Helentjaris T, Devos K, Kurata N, Gale M: **Was there a single ancestral cereal chromosome?** *Trends Genet* 1995, **11**:81-82.

13. Kowalski SP, Lan TH, Feldmann KA, Paterson AH: **Comparative mapping of *Arabidopsis thaliana* and *Brassica oleracea* chromosomes reveals islands of conserved organization.** *Genetics* 1994, **138**:499-510.

14. Seoighe C, Wolfe KH: **Extent of genomic rearrangement after genome duplication in yeast.** *Proc Natl Acad Sci USA* 1998, **95**:4447-4452.

15. •• Gaut BS, Doebley JF: **DNA sequence evidence for the segmental allotetraploid origin of maize.** *Proc Natl Acad Sci USA* 1997, **94**:6809-6814.
The authors propose that the duplicated regions of the maize genome arose from segmental allotetraploidy and they provide a very clear and reasoned explanation of their model. Molecular clock analysis of 14 maize gene pairs yielded two non-overlapping groups of date estimates centred on 11.4 and 20.5 Mya. An allotetraploid ancestor is proposed to have gone through a phase of tetrasomic inheritance – each locus has four alleles – before becoming 'diploidised' (disomic). At each locus, genetic drift during the tetrasomic phase might result in the loss of alleles inherited from one of the progenitor species, or alleles from both progenitors might be retained. Consequently, after diploidisation, the divergence time between the two sequences at a duplicated locus could correspond either to the speciation time between the two progenitors (20.5 Mya), or to the time of establishment of disomy (11.4 Mya). It would be impossible to predict the outcome for any particular locus.

16. Keogh RS, Seoighe C, Wolfe KH: **Evolution of gene order and chromosome number in *Saccharomyces*, *Kluyveromyces* and related fungi.** *Yeast* 1998, **14**:443-457.

17. Sharp PM, Li WH: **An evolutionary perspective on synonymous codon usage in unicellular organisms.** *J Mol Evol* 1986, **24**:28-38.

18. Bailey GS, Poulter RT, Stockwell PA: **Gene duplication in tetraploid fish: model for gene silencing at unlinked duplicated loci.** *Proc Natl Acad Sci USA* 1978, **75**:5575-5579.

19. Walsh JB: **How often do duplicated genes evolve new functions?** *Genetics* 1995, **139**:421-428.

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

21. Hughes MK, Hughes AL: **Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*.** *Mol Biol Evol* 1993, **10**:1360-1369.

22. Nowak MA, Boerlijst MC, Cooke J, Smith JM: **Evolution of genetic**
• **redundancy.** *Nature* 1997, **388**:167-171.
Four models are proposed here for how genetic redundancy could be evolutionarily stable but these require different mutation rates in different genes – which seems unlikely for genes with, initially, identical DNA sequences – or are only applicable to multicellular organisms.

23. Cooke J, Nowak MA, Boerlijst M, Maynard-Smith J: **Evolutionary**
• **origins and maintenance of redundant gene expression during metazoan development.** *Trends Genet* 1997, **13**:360-364.
Building on their earlier results [22•], the authors here point out that developmental genes are much more likely to acquire new functions than are housekeeping genes with metabolic functions. Redundancy generated by gene duplication may increase fitness, or it may be retained if the duplicated genes have different efficiencies and different mutation rates which interact in such a way as to keep the two copies in a dynamic equilibrium.

24. Gibson TJ, Spring J: **Genetic redundancy in vertebrates: polyploidy**
• **and persistence of genes encoding multidomain proteins.** *Trends Genet* 1998, **14**:46-49.
Point mutations in developmental genes often have dominant deleterious phenotypes, whereas complete deletion of these genes often has no phenotype. Gibson and Spring argue that this is to be expected for genes encoding multidomain proteins and that this may prevent these genes from decaying into pseudogenes.

25. Iwabe N, Kuma K, Miyata T: **Evolution of gene families and relationship with organismal evolution: rapid divergence of tissue-specific genes in the early evolution of chordates.** *Mol Biol Evol* 1996, **13**:483-493.

26. Ohno S: **The notion of the Cambrian pananimalia genome and a genomic difference that separated vertebrates from invertebrates.** In *Molecular Evolution: Evidence for Monophyly of Metazoa (Progress in Molecular and Subcellular Biology, vol. 19)*. Edited by Müller WE, Kuchino Y, Jeanteur P, Paine PL. New York: Springer-Verlag; 1998:in press.

27. Simmen MW, Leitgeb S, Clark VH, Jones SJ, Bird A: **Gene number in**
•• **an invertebrate chordate, *Ciona intestinalis*.** *Proc Natl Acad Sci USA* 1998, **95**:4437-4440.
A method for estimating the number of genes in any eukaryote on the basis of BLAST searches with random genomic and cDNA sequences. The method was tested using subsets of the data from the *Caenorhabditis elegans* genome project, and was then applied to the tunicate *Ciona intestinalis*. Its apparent accuracy, simplicity, and low cost – only 76 EST and 1487 genomic single-pass sequencing runs were made – make it attractive to apply to other organisms such as amphioxus and lamprey.

28. Fields C, Adams MD, White O, Venter JC: **How many genes in the human genome?** *Nat Genet* 1994, **7**:345-346.

29. Spring J: **Vertebrate evolution by interspecific hybridisation – are we polyploid?** *FEBS Lett* 1997, **400**:2-8.

30. Baker ME: **Steroid receptor phylogeny and vertebrate origins.** *Mol Cell Endocrinol* 1997, **135**:101-107.

31. Escriva H, Safi R, Hanni C, Langlois MC, Saumitou-Laprade P, Stehelin D, Capron A, Pierce R, Laudet V: **Ligand binding was acquired during evolution of nuclear receptors.** *Proc Natl Acad Sci USA* 1997, **94**:6803-6808.

32. Comings DE: **Evidence for ancient tetraploidy and conservation of linkage groups in mammalian chromosomes.** *Nature* 1972, **238**:455-457.

33. Carver EA, Stubbs L: **Zooming in on the human-mouse comparative map: genome conservation re-examined on a high-resolution scale.** *Genome Res* 1997, **7**:1123-1137.

34. Lundin LG: **Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse.** *Genomics* 1993, **16**:1-19.

35. Ruddle FH, Bentley KL, Murtha MT, Risch N: **Gene loss and gain in the evolution of the vertebrates.** *Development* 1994, **Suppl**:155-161.

36. Katsanis N, Fitzgibbon J, Fisher EMC: **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* 1996, **35**:101-108.

37. Kasahara M, Hayashi M, Tanaka K, Inoko H, Sugaya K, Ikemura T, Ishibashi T: **Chromosomal localization of the proteasome Z subunit gene reveals an ancient chromosomal duplication involving the major histocompatibility complex.** *Proc Natl Acad Sci USA* 1996, **93**:9096-9101.

38. Kasahara M: **New insights into the genomic organization and origin of the major histocompatibility complex: role of chromosomal (genome) duplication in the emergence of the adaptive immune system.** *Hereditas* 1997, **127**:59-65.

39. Endo T, Imanishi T, Gojobori T, Inoko H: **Evolutionary significance of**
•• **intra-genome duplications on human chromosomes.** *Gene* 1997, **205**:19-27.
See annotation [40••].

40. Hughes AL: **Phylogenetic tests of the hypothesis of block**
•• **duplication of homologous genes on human chromosomes 6, 9, and 1.** *Mol Biol Evol* 1998, **15**:854-870.
The combined results from this study and that of Endo *et al.* [39••] show that, of the 11 gene pairs on HSA6/HSA9 that have previously been proposed, a simultaneous origin seems possible for six: RXRA/RXRB, COL5A1/COL11A2, ORFX/RING3, PBX3/PBX2, C5/C4A and TNC/TNX. Gene order for these pairs is conserved except for one inversion of C5 and TNC. Other gene pairs are either much older (ABC2/TAP1, NOTCH1/INT3, PSMB7/PSMB8 and HSPA5/HSPA1A) or much younger (VARS1/VARS2). Despite their similar conclusions, the use of different molecular clock calibrations in the two studies causes them to disagree on the absolute date of the block duplication: 579–696 Mya in Hughes' study, but 161–580 Mya in Endo *et al.*'s. The latter group also put forward a convoluted hypothesis involving two rounds of duplication to explain the presence of older gene pairs in the region, whereas Hughes proposes that there may be some sort of selective constraint causing clustering of the ancient paralogues.

41. Kappen C, Schughart K, Ruddle FH: **Two steps in the evolution of Antennapedia-class vertebrate homeobox genes.** *Proc Natl Acad Sci USA* 1989, **86**:5459-5463.

42. Holland PW, Garcia-Fernandez J, Williams NA, Sidow A: **Gene duplications and the origins of vertebrate development.** *Development* 1994, **Suppl**:125-133.

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

44. Sharman AC, Holland PW: **Estimation of Hox gene cluster number in lampreys.** *Int J Dev Biol* 1998, **42**:617-620.

45. Aparicio S, Hawker K, Cottage A, Mikawa Y, Zuo L, Venkatesh B, Chen E, Krumlauf R, Brenner S: **Organization of the *Fugu rubripes* Hox clusters: evidence for continuing evolution of vertebrate Hox complexes.** *Nat Genet* 1997, **16**:79-83.

46. Prince VE, Joly L, Ekker M, Ho RK: **Zebrafish hox genes: genomic organization and modified colinear expression patterns in the trunk.** *Development* 1998, **125**:407-420.

47. Postlethwait JH, Yan YL, Gates MA, Horne S, Amores A, Brownlie A,
•• Donovan A, Egan ES, Force A, Gong Z *et al.*: **Vertebrate genome evolution and the zebrafish gene map.** *Nat Genet* 1998, **18**:345-349.
Duplicated genes linked to the four Hox clusters in mammals – the HSA 2/7/12/17 region – are also linked to them in zebrafish, and parts of the HSA 1/6/9/19 region are conserved. This implies that the duplications producing these regions occurred prior to the bony fish/tetrapod divergence, contrary to Lundin's proposal ([34]; Figure 1). As noted in the commentary by Aparicio [49], although zebrafish and mammals show conservation of synteny, gene order is often rearranged in these examples. Postlethwait *et al.* also report some examples where a pair of linked genes in a mammal seems to correspond to two linked pairs in zebrafish and propose that additional duplications (either of chromosomal fragments or of the whole genome) may have occurred in this species.

48. Meyer A: **Hox gene variation and evolution.** *Nature* 1998, **391**:225-228.

49. Aparicio S: **Exploding vertebrate genomes.** *Nat Genet* 1998, **18**:301-303.

50. Brooke NM, Garcia-Fernandez J, Holland PW: **The ParaHox gene**
•• **cluster is an evolutionary sister of the Hox gene cluster.** *Nature* 1998, **392**:920-922.
Three homeobox genes, *Gsx*, *Xlox* (*Pdx*) and *Cdx*, are clustered in the amphioxus genome and may also be clustered in mammals. These genes are similar to three of the Hox paralogy groups in terms of their sequence, expression, and order on the chromosome. ParaHox and Hox arose by duplication >520 Mya, before Hox duplicated further to produce the four clusters found in mammals.

51. Pébusque M-J, Coulier F, Birnbaum D, Pontarotti P: **Ancient large scale genome duplications: phylogenetic and linkage analyses shed light on chordate genome evolution.** *Mol Biol Evol* 1998, **15**:1145-1159.

52. Rowen L, Mahairas G, Hood L: **Sequencing the human genome.**
• *Science* 1997, **278**:605-607.
The human genome sequence should be completed by 2005.

53. Sharman AC, Holland PWH: **Conservation, duplication, and divergence of developmental genes during chordate evolution.** *Neth J Zool* 1996, **46**:47-67.

54. Koop BF, Nadeau JH: **Pufferfish and new paradigm for comparative genome analysis.** *Proc Natl Acad Sci USA* 1996, **93**:1363-1365.

55. Nadeau JH: **Genome duplication and comparative gene mapping.** In *Advanced Techniques in Chromosome Research*. Edited by Adolph KW. New York: Marcel Dekker; 1991:269-296.

56. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403-410.

57. NCBI Entrez Database on World Wide Web URL: http://www.ncbi.nlm.nih.gov

58. Tatusov R: **DUST sequence filter** on the internet at ftp://ncbi.nlm.nih.gov/pub/tatusov/dust

59. Claverie JM, States DJ: **Information enhancement methods for large scale sequence analysis.** *Computers Chem* 1993, **17**:191-201.

60. Jurka J: **REPBASE database of human repetitive DNA** on the internet at ftp://ncbi.nlm.nih.gov/repository/repbase/REF/humrep.ref