Evolution of vertebrate genome organisation

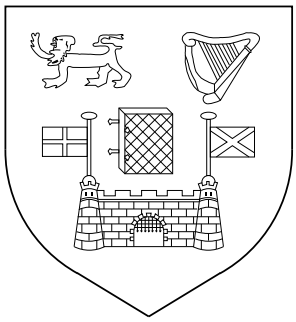
by

Aoife McLysaght

A thesis submitted to
The University of Dublin
for the degree of

Doctor of Philosophy

Department of Genetics
Trinity College
University of Dublin



October, 2001

Declaration

This thesis is submitted by the undersigned for the degree of Doctor in Philosophy at the University of Dublin. It has not been submitted as an exercise for a degree at any other university.

Apart from the advice, assistance, and joint effort mentioned in the acknowledgements and in the text, this thesis is entirely my own work.

I agree that the library may lend or copy this thesis freely upon request.

Aoife McLysaght

October 2001

Acknowledgements

Kenneth Henry - thanks for being a great supervisor and a good sport. Andrew 'brain the size of a planet' Lloyd, my mentor - Thank You For Sharing. None of it would have been the same if I wasn't in such a positive, encouraging environment with the rest of the Wolfe Cubs. Karsten working side-by-side with you for the stuff of Chapter 4 was a great experience. I hope all my future collaborations are as fruitful and good-natured. Cathal and Lucy, the predecessors, thanks for showing me that it can be done (and how to do it!). Avril, thanks for always being willing to direct your insightful intellect towards my problems. Simon thanks for your willingness to help with anything. All of you, and the new folks, Antoinette, Kevin, and Sean, thanks for being so understanding and helpful while I've been writing up. Thanks also to Denis Shields for the suggestion of the block overlap simulations in Chapter 4, and all the members of the Tuesday lunchtime group for helpful comments over the years.

My family and friends deserve special thanks for supporting me in various ways. Mum - thanks for Friday lunchtimes in Dunne & Crescenzi's and 'Friday Presents'. Dad - the running total is 2135 Brownie Points, a few more and you'll have enough for a toaster. Emer - those cinema outings helped me stay relaxed, thanks. Wendy, my adopted Scottish sister and bestest friend - thanks for always dropping everything at a moment's notice and making everything so much fun. Thanks to Gianluca for lots of emotional support[a]. Thanks to my 'roomies', Lisa and Emma Jane, for being great pals. Princess Nisa, you really are a little princess (and now it's written in the Trinity College Library, so it must be true!). The girls on the rugby team definitely helped me keep my sanity by allowing me to vent my frustrations by tipping them upside-down in the mud of College Park ... thanks!

[a]and for help with footnotes and other LaTeX oddities

Contents

List of Tables

List of Figures

Abbreviations

ANOVA	Analysis of Variance
BLAST	Basic Local Alignment Search Tool
bp	basepairs
cM	centiMorgans
cR	centiRads
E-value	Expectation value
Gb	Gigabases
HSA	Homo Sapiens
HSP	High-scoring Segment Pair
kb	kilobases
Mb	Megabases
MHC	Major Histocompatibility Complex
MSP	Maximal-scoring Segment Pair
Mya	Million years ago
N_e	Effective population size
OTU	Operational Taxonomic Unit
sm	number of duplicated genes between two paralogous regions of a genome
TDT	Tandem Duplication and Translocation model
WGD	Whole Genome Duplication

'Your talk,' I said, 'is surely the handiwork of wisdom because not one word of it do I understand.'

The Third Policeman by Flann O'Brien

Summary

The increasing availability of genomic sequences from different vertebrate organisms affords molecular biologists the opportunity to thoroughly investigate phenomena that were only hinted at by more sparse data. The work described in this thesis develops the use of inter- and intra-genomic sequence comparisons to examine genome evolution through changes in genome arrangement and content.

The vertebrate *Fugu rubrupies* (pufferfish) has a small genome with little repetitive sequence which makes it attractive as a model genome. Its genome compaction and synteny conservation relative to the human genome were studied using data from public databases. The compaction of this genome was measured by comparing lengths of orthologous *Fugu* and human introns. Analysis of orthologous introns showed an eight-fold average size reduction in *Fugu*, consistent with the ratio of total genome sizes. There was no consistent pattern relating the size reduction in individual introns or genes to gene base composition in either species. For genes that are neighbours in *Fugu*, 40-50% have conserved synteny with a human chromosome. Comparison of observed data to computer simulations suggests that 4,000-16,000 chromosomal rearrangements have occurred since *Fugu* and human shared a common ancestor, implying a faster rate of rearrangement than seen in human/mouse comparisons.

Intragenomic comparisons were used to examine the draft human genome sequence for evidence of ancient genomic duplications, by a combination

of a map-based and a phylogeny-based approach. Evidence was found for extensive paralogy regions situated throughout the genome. Statistical analyses of these regions indicated that they were formed by *en bloc* duplication events. Molecular clock analysis of 191 gene families in the human genome indicates that a burst of gene duplication activity took place approximately 333-583 Mya, spanning the estimated time of origin of vertebrates (about 500 Mya). Moreover, more gene pairs of this age are found in paralogous regions than pairs that duplicated earlier or later.

These results support the contention that many vertebrate gene families were formed by extensive duplication events, perhaps polyploidy, in an early chordate, and indicate that extensive genome rearrangement may have occurred following genome duplication.

Chapter 1

Introduction

One of the justifications for genome sequencing projects is the opportunity they provide to study the evolution of genomes and proteomes. As genome projects progress it is becoming possible to study not just the genome contents, but also the arrangement of those contents along the chromosomes, and to see how these arrangements evolve. Complete genome sequences allow us to look at the molecular evolution of chromosomes in much the same way as the first DNA sequences in the 1970s enabled us to study the evolution of individual genes.

Genomes evolve through changes in the arrangement of genes, and changes in the genome content (by gene duplication or loss). The ultimate consequence of genome evolution is speciation. Chromosomal rearrangements, and divergent resolution of duplicated genes will lead to reproductive isolation (White, 1978; Taylor *et al.*, 2001b).

1.1 A brief history of vertebrates

Sooner than initially expected (an early estimate for the completion of the Human Genome Project was 2005; Rowen *et al.*, 1997), molecular biology has delivered on its greatest promise - the sequence of the human genome

(Lander *et al.*, 2001). This leaves science in the rather surprising position that information and understanding gleaned from the human genome may help in the interpretation of other vertebrate genomes, rather than *vice versa*. The interesting task for molecular evolution now is to uncover the processes that have shaped the vertebrate genome. 'The challenge of the Human Genome Project will be to go from ordering the letters of the DNA language to understanding the words, phrases, sentences, paragraphs, and finally the story of the genome.' (Koop, 1995).

Molecular biology has significantly enriched our understanding of the evolution of vertebrate species. For example, molecular phylogenetic methods resolved the long-standing puzzle of the relationship of cetaceans (whales, dolphins, and porpoises) to other eutherian mammals, by showing them to be nested within artiodactyls (Graur and Higgins, 1994; Shimamura *et al.*, 1997).

Knowledge of the relationship of vertebrate lineages is important because it can be used to place evolutionary events on a relative timescale. Similar characteristics can be recognised as related by descent (homologous), or as independently acquired in different lineages (analogous). The fossil record has traditionally been the source of information on the timing of the origin of species, but is heavily reliant on interpretation, and on the availability of fossils. Molecular biology provides us with an objective way of measuring the relatedness of organisms through the similarity of their hereditary material (Box 1.1 overleaf), and the data are abundant.

Kumar and Hedges (1998) produced a timescale for molecular evolution based on a molecular clock analysis of 658 gene sequences from 207 vertebrate species. A summary of the divergence of major chordate lineages is shown in Figure 1.1. One important observation of this study was that at least five major mammalian lineages arose during the Cretaceous period (145-65 Mya), and that the divergence of some orders of birds was dated to the

Box 1.1: The Molecular Clock

The discovery that amino acid sequences accumulate substitutions at an approximately constant rate over time led to the realisation that distances between protein sequences of a given gene in different species could be used to resolve the history of those species both in terms of the timing of events and the relationship of species (Zuckerkandl and Pauling, 1962; Margoliash, 1963; Zuckerkandl and Pauling, 1965). There is, however, no such thing as The Universal Molecular Clock, *i.e.*, the rate of accumulation of substitutions is not the same for all genes or species (Nei, 1987; Li, 1993). Rather than eradicating the usefulness of molecular clock-based studies, this fact actually enhances the usefulness of this type of analysis because it enables the examination of both long-term and short-term evolutionary processes by selecting an appropriate dataset (Nei, 1987).

Differences in evolutionary rates are probably due to functional constraints on the protein (Zuckerkandl and Pauling, 1965), *i.e.*, some substitutions may be deleterious to the function of the gene. At sites where there is effectively no functional constraint (neutral sites) the rate of substitution should be equal to the rate of mutation (Kimura, 1983). Therefore, for proteins that are evolving in a neutral fashion, if the rate of mutation has not changed with time then the rate of evolution of that protein should be constant (Kimura, 1983; Li, 1997).

In order to relate molecular time to astronomical time the molecular clock must be calibrated. This may be achieved by reference to the evidence for lineage divergences in the fossil record or by reference to major geological events such as the isolation of populations caused by continental drift, or island (or lake) formation (Nei, 1987).

To test whether the assumption of the molecular clock holds for a particular set of sequences one must estimate the number of substitutions over time. This method is problematic because it is subject to the inaccuracies of dating divergence times from the fossil record. Sarich and Wilson (1973) devised a clever way to evaluate the adherence to the molecular clock of a set of three sequences (two sequences of interest (A and B), and an outgroup) - the relative rate test. The relative rate test specifies that the molecular distance to the outgroup sequence should be the same for A and B, *i.e.*, this method assesses the similarity of the relative rate of substitution without requiring any knowledge of species divergence times.

mid-Cretaceous. This meant that most mammalian orders, and birds, had radiated before the extinction of dinosaurs (Kumar and Hedges, 1998). Prior to these estimates of lineage divergences it had been assumed that rapid rates of morphological change had succeeded the extinction of dinosaurs because of the absence of mammalian and avian fossils dating to before the Early Tertiary period (Nei, 1987). The evidence for earlier divergence dates of these lineages removes the need to invoke rapid change to explain the origin of vertebrate diversity (Kumar and Hedges, 1998). Similarly, before the availability of molecular data, humans and chimpanzees were assumed to have diverged about 30 Mya, however, Sarich and Wilson (1967) showed that the divergence time could be as recent as 5 Mya.

1.2 Evolution by genome rearrangement

The structural divergence of genomes occurs by processes of genome rearrangement. Genome arrangement differences accumulate over time (although not necessarily in a linear fashion) so that the organisation of mouse and human genomes that separated about 96 Mya (Nei *et al.*, 2001) should be more similar than that of pufferfish (*Fugu*) and human genomes whose lineages diverged about 450 Mya (Kumar and Hedges, 1998). For biologists, it is useful to gain an understanding of the amount of conservation of genome arrangement between two species, because it enhances the usefulness of experimental organisms by facilitating positional cloning of genes in the genome of interest.

1.2.1 Mechanisms of genome rearrangement

The principal mechanisms of genome rearrangement are inversions, translocations, and transpositions of segments of chromosome, perhaps containing many genes. An inversion is the reversal of a portion of chromosome. A

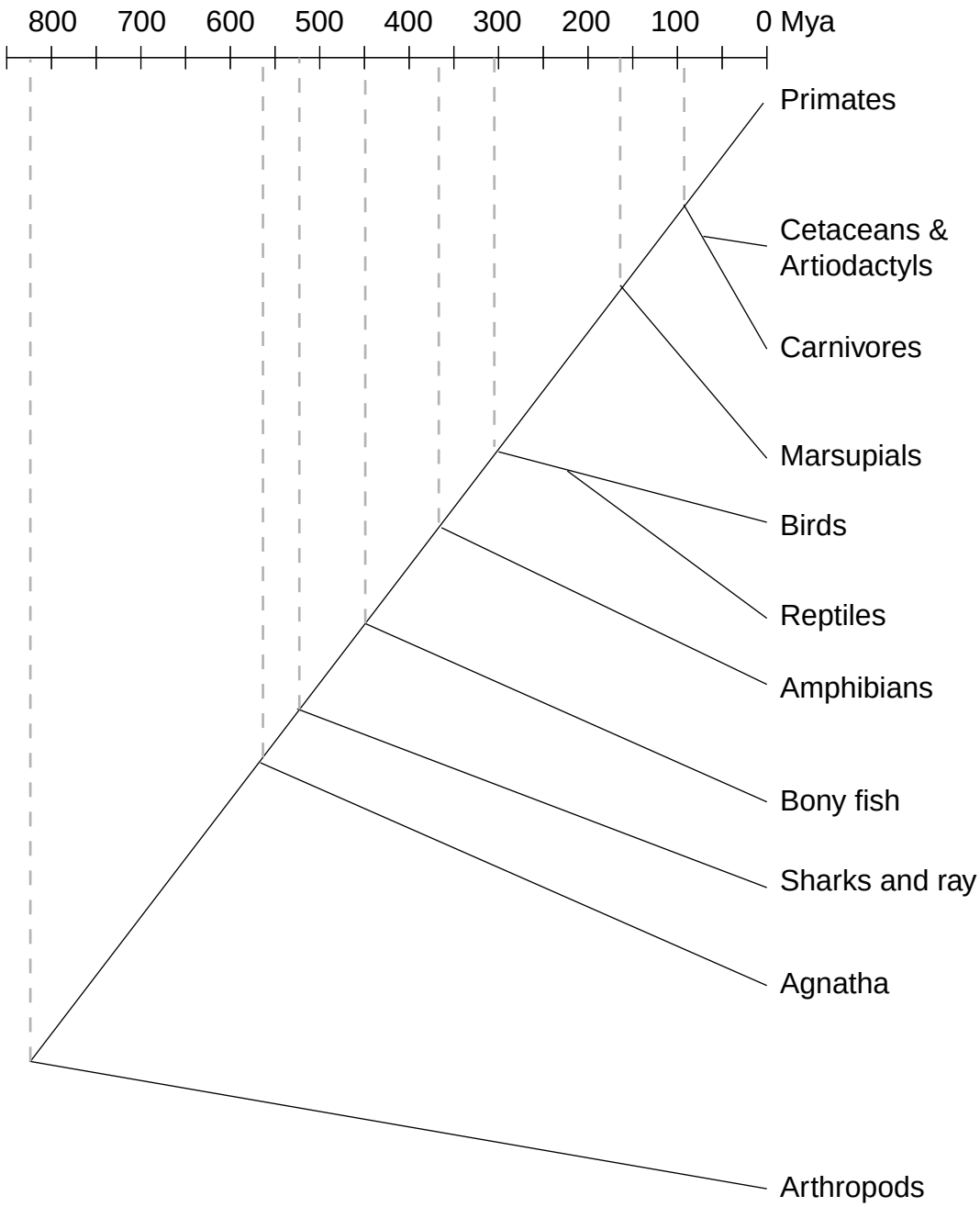


Figure 1.1: Divergence of chordate groups. Dates are taken from Kumar and Hedges (1998), except for the arthropod divergence date which is taken from Nei *et al.* (2001).

transposition is the movement of a segment of chromosome from one location to another (this may also include an inversion). A reciprocal translocation event is the result of recombination between non-homologous chromosomes. These processes are illustrated in Figure 1.2.

1.2.2 Gene order evolution

Synteny is the property of two or more genes being located on the same chromosome. Synteny conservation is the conservation of this property in different genomes. Synteny conservation is often expressed as the number of conserved segments between two genomes (*e.g.*, Nadeau and Taylor, 1984).

In the absence of complete genome sequences, gene order evolution in eukaryotes has been studied in two ways: using genome map data (marker-based studies); and using genomic sequence data (sequence-based studies). The marker-based studies are exemplified by the use of mouse and human genetic maps to identify regions of conserved synteny (*e.g.*, Nadeau and Kosowsky, 1991; Lundin, 1993). This method uses information from genetic maps, physical maps, and chromosome painting experiments. It has the advantage of looking at grand scale chromosome evolution, but lacks the resolution to detect small-scale disruptions of gene order, because when two genes are found to have conserved synteny with their orthologues in another genome they are frequently assumed to form part of an uninterrupted conserved chromosomal segment. Inversions and transpositions may disrupt the integrity of these segments without disrupting the synteny of the genes. For this reason the number of rearrangement events can be underestimated, particularly for genomes with sparse data.

Sequence-based studies compare two genomic sequences (*e.g.*, Seoighe *et al.*, 2000). This method has the advantage that it has the ultimate level of resolution - single nucleotides. The disadvantage of this method is that there is only a paltry quantity of data available (although it is increasing

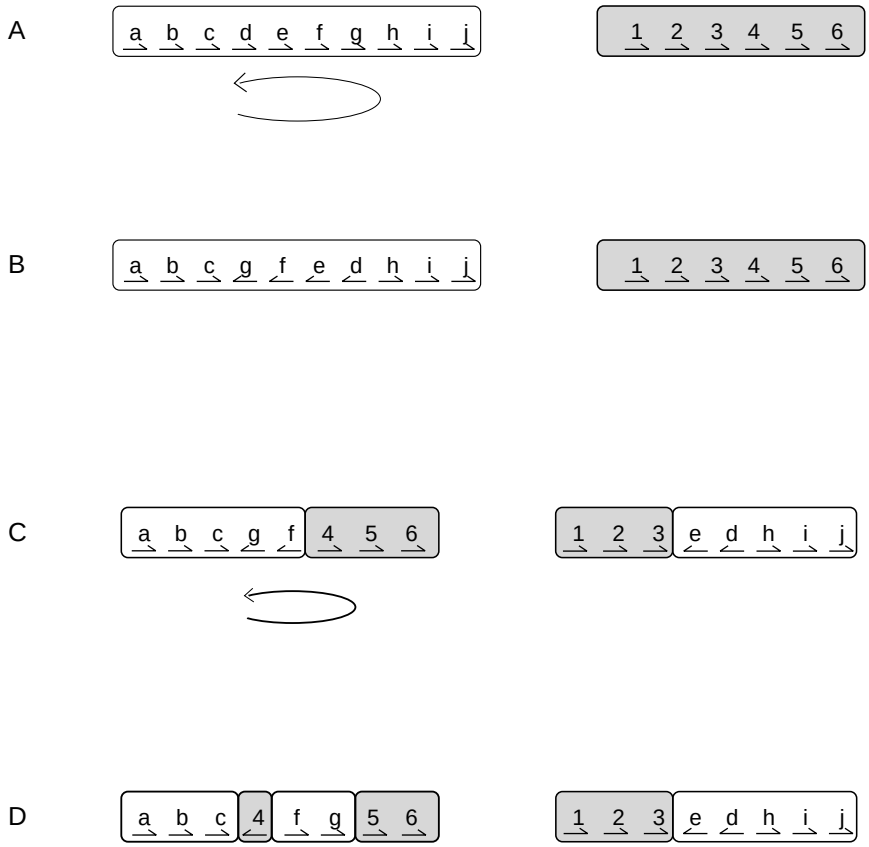


Figure 1.2: Mechanisms of genome rearrangement. Four hypothetical stages of rearrangement of a two-chromosome genome are shown. (**A**) The ancestral (or reference) genome. (**B**) An inversion of four genes (illustrated by the curved arrow in A) conserves the synteny of the genes on the white chromosome with the ancestral genome but changes gene order and the distances between some genes. (**C**) A reciprocal translocation event disrupts the synteny of the genes compared with the ancestral arrangement. (**D**) The inversion shown, subsequent to the translocation event, breaks the two conserved segments of white and grey (with reference to the ancestral genome) into four segments. Without the inference of inversion events this may cause researchers to conclude there have been two independent translocations between the grey and white chromosome.

daily), particularly from closely related organisms. One way around this difficulty is to study intragenomic duplicated segments of chromosome as has been done in yeast and *Arabidopsis* (*e.g.*, Wolfe and Shields, 1997; *Arabidopsis* Genome Initiative, 2000). Another work-around is to compare complete genome sequence from one organism to more limited data, such as cosmid-skimming data (shotgun-sequenced cosmid clones), from another.

1.2.3 Human-rodent comparative genomics

The mouse is a popular and valuable model organism for experimental analysis of mammalian diseases and the mammalian genome (Koop, 1995). In order to maximise the usefulness of mouse genome data it is important to understand the relationship of this genome to the human genome.

Currently there are very few long genomic sequences available for gene order comparisons in vertebrates. Even between human and mouse, only a few completely annotated BAC-sized sequences from mouse exist in GenBank for comparison with the human genome sequence.

In a classic study, Nadeau and Taylor (1984) investigated the lengths of conserved segments (segments of chromosome with orthologous gene contents) between the mouse and human genomes using a marker-based approach. They had only partial data (83 homologous genes from both species), and observed at least 46 (and at most 65) conserved segments, with an average length of 8.5 centiMorgans (cM). To address the incomplete nature of the available data they developed a mathematical method to scale up the estimate of mean length of conserved segments to the whole genome level which they calculated to be 8.1 ± 1.6 cM. This was problematic because when one only has partial data it is impossible to know how much further the conserved segments extend beyond the available markers. Furthermore, some conserved segments may escape detection because there is only one or no known marker, causing a bias towards detection of larger regions (Nadeau

and Taylor, 1984).

Using the estimated average length of a conserved segment Nadeau and Taylor (1984) estimated the number of rearrangement events that have occurred since the divergence of mouse and human (approximately 100 Mya) as 178 ± 39. Nadeau and Taylor's method seems to be sound because subsequent studies have not significantly changed the numbers (*e.g.*, DeBry and Seldin, 1996). In a mathematical assessment of the Nadeau-Taylor method Sankoff *et al.* (2000) concluded that the longevity of this estimate was based on the fundamental robustness of the method.

One assumption fundamental to these calculations is that linked markers form uninterrupted chromosomal segments. Any region containing at least two linked markers in both species is assumed to be part of a conserved segment. This calculation does not allow for intrachromosomal inversions which means that the estimate of the number of rearrangements is likely to be substantially lower than an estimate that incorporates inversion events, especially given the apparent high frequency of small inversion events (Seoighe *et al.*, 2000; Kumar *et al.*, 2001). The Nadeau-Taylor estimate would be better interpreted as the number of inter-chromosomal rearrangements. It is difficult to incorporate small inversions into this model when the data are still incomplete (Sankoff *et al.*, 2000).

The lengths used by Nadeau and Taylor (1984) were genetic map units (cM), which are proportional to recombination frequencies rather than physical distances. Despite the similar physical sizes of the human and mouse genome, the human genome has a total genetic length of 3,300 cM and the mouse genome has a total genetic length of 1,600 cM which indicates that the amount of rearrangement may not be equal in the two lineages (Nadeau, 1989). Moreover, the recombination rate per kb is far from uniform over the human genome (Daly *et al.*, 2001).

Watanabe *et al.* (1999) produced a genome-wide comparative map of the

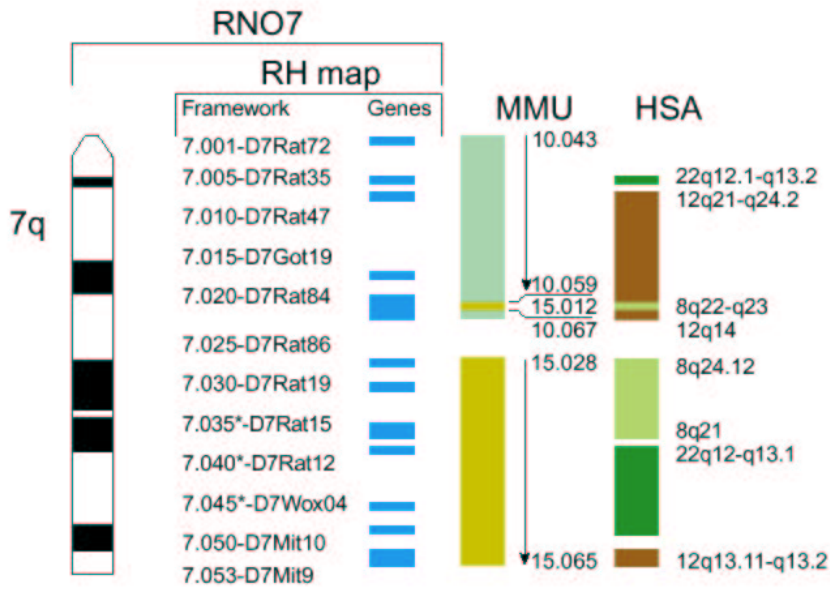


Figure 1.3: Comparative map of rat chromosome 7 (RNO7) showing the locations of homologous genes in the mouse and the human genome. Chromosomes are indicated by a colour code. This figure is part of Fig. 2 of Watanabe *et al.* (1999).

rat, mouse, and human genomes containing over 500 genes from the rat radiation hybrid map. It is evident from Figure 2 of that paper that some parts of one chromosome in one mammal are related to parts of two different chromosomes in another species in an interleaved fashion. This is indicative of small intrachromosomal rearrangement events such as inversions. An example of this is seen on the q arm of rat chromosome 7, where there is an alternating series of conserved segments with human chromosomes in the order HSA12, HSA8, HSA12, and HSA8 (Figure 1.3). Rather than inferring two independent translocation events to insert two HSA8-like segments into what might have otherwise been a region of perfect synteny with HSA12, it is more likely that there was only one translocation event and then one small inversion, as inversions are thought to occur at a higher frequency than translocations (Blanchette *et al.*, 1996; Postlethwait *et al.*, 2000).

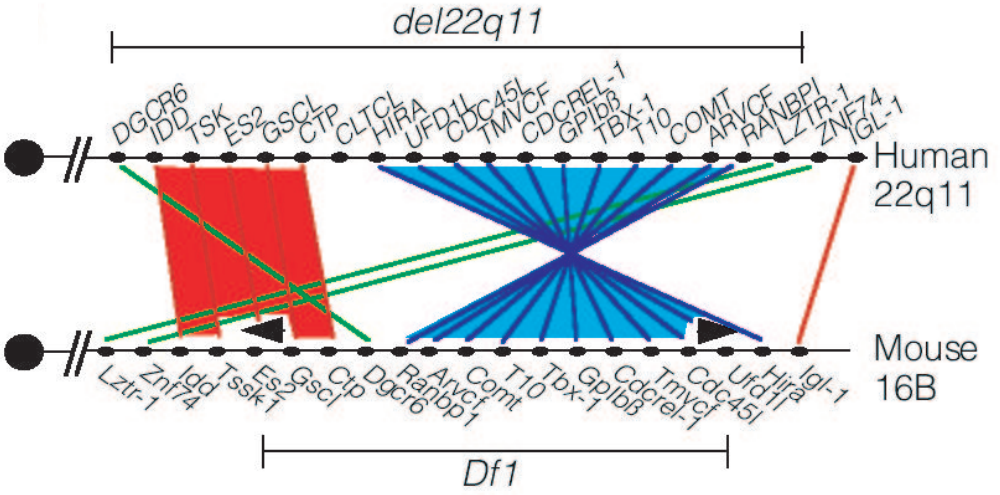


Figure 1.4: Comparison of a portion of mouse chromosome 16 with human chromosome 22. Red lines indicate genes with the same order and orientation in the two species, green lines indicate genes that have changed position, and blue lines indicate genes that have conserved their relative order but have inverted centromere-telomere orientation. Taken from Lindsay *et al.* (1999).

Other marker-based studies have revealed local genome rearrangements on a background of conserved synteny, *i.e.*, gene order was not conserved. For example, the DiGeorge Syndrome region in mouse and human have similar, but rearranged, gene content (Lindsay *et al.*, 1999, Figure 1.4) which can be reconciled by inferring two multi-gene inversions and a single gene transposition event. Similarly, a region between mouse chromosome 11 and human chromosome 5 has been disrupted by a minimum of four rearrangement events (Watkins-Chow *et al.*, 1997).

Sequence-based comparison of the human and mouse genomes has uncovered regions of high similarity. Dot-matrix DNA sequence comparisons reveal high levels of sequence similarity, even of the non-coding regions of homologous human and rodent loci (*e.g.*, the T-cell receptor α loci, and the α and β-myosin heavy chain genes; Koop, 1995: and Bruton's Tyrosine Kinase loci (BTK); Oeltjen *et al.*, 1997). Based on varying divergence patterns

in non-coding DNA comparisons between human and mouse Koop (1995) suggested a 'mosaic model' of evolution of genomes, where large portions of the genome may evolve at different rates.

1.2.4 Non-mammalian vertebrate genomes

Comparative analysis of the chicken (*Gallus gallus*) and human genomes using genetic mapping and chromosome-painting (Zoo-FISH) analysis found that the number of rearrangements between the human and chicken genomes is 72, and that human-mouse, and chicken-mouse comparisons revealed 171 and 128 rearrangement events respectively (Burt *et al.*, 1999). Another study of the chicken and human genomes using chromosome painting experiments found that there was conserved synteny between human chromosome 4 and chicken chromosome 4, but that this region was disrupted in the mouse (Chowdhary and Raudsepp, 2000). Both Burt *et al.* (1999) and Chowdhary and Raudsepp (2000) concluded that the chicken and human genomes were more alike than the mouse and human genomes, and that these genomes are probably more similar to the ancestral vertebrate genome than is the mouse genome. Burt *et al.* (1999) hypothesised three stages of chromosome evolution in birds and mammals which included an accelerated rearrangement rate in the rodent lineage analogous to what is observed for nucleotide substitution rates (Li *et al.*, 1996).

In contrast to mammals, in chicken the female is the heterogametic sex. There is evidence that the chicken sex chromosome Z (GGA Z) may also have undergone some inversion events in its history. Of the eighteen genes currently mapped to GGA Z, eleven have HSA9 orthologues (Nanda *et al.*, 1999). Comparative maps indicate that rearrangement events on either GGA Z or HSA9 (or both) have changed the order of these genes.

In a recent analysis a high-resolution human-chicken comparative map was generated for human chromosome 15 (Crooijmans *et al.*, 2001). This

analysis identified 6 inter-chromosomal and 15 intra-chromosomal rearrangement breakpoints in the HSA 15 and chicken comparison, and 3 inter-chromosomal and 15 intrachromosomal rearrangement breakpoints between HSA 15 and the mouse genome. Scaling up to the whole genome Crooijmans *et al.* (2001) estimate that there are at least 600 conserved segments between the human and chicken genomes.

A comparative analysis of human and zebrafish genomes found a high level of synteny conservation but that genes with conserved synteny did not form an uninterrupted conserved segment suggestive of intrachromosomal rearrangement events (Barbazuk *et al.*, 2000). This analysis found evidence for 247 conserved segments between the genomes, and the authors estimated that more complete data would reveal 418 conserved segments. Barbazuk *et al.* (2000) suggest that knowledge of conserved synteny arrangements may be useful for the resolution of ambiguous orthology relationships.

Analysis of the zebrafish genome led several authors to suggest a tetraploidy event in the history of teleost fishes over 100 Mya (Amores *et al.*, 1998; Gates *et al.*, 1999). An intragenomic comparison of regions of chromosomes thought to have been duplicated in the whole genome duplication event revealed that inversions have disrupted the colinearity of chromosomes in the teleost lineage (Postlethwait *et al.*, 2000).

1.2.5 Synteny conservation - A selected or a neutral trait?

Most models addressing the frequency of genome rearrangement events and the size of conserved segments between any pair of genomes assume that rearrangement breakpoints are randomly distributed in the genome, *i.e.*, assuming both that all sites in the genome are equally likely to suffer a breakage mutation, and that selection does not filter out any breakpoints.

The Nadeau and Taylor (1984) model assumed that autosomal rear-

rangements are randomly distributed within the genome. Lundin (1993) proposed that chromosomal rearrangement events are common but large, thereby conserving close linkage arrangements. If this is true then, over short evolutionary time frames, rearrangement breakpoints will effectively be distributed randomly in the genome, even if there are some regions of chromosomes that are being conserved by selection (*e.g.*, the *Hox* clusters; Ruddle *et al.*, 1994) because these regions (and other selectively neutral gene arrangements) may be preserved by stochastic processes rather than selection. Over longer evolutionary times, randomly distributed breakpoints should be present at effectively every position in the genome, and so selection will affect the observed distribution of rearrangement events.

The fact that the *Cænorhabditis* genomes, *C. elegans* and *C. briggsæ*, contain operons (Spieth *et al.*, 1993) may have the effect that rearrangement breakpoints may not be randomly distributed throughout the genome, as there will likely be selection for conservation of operon structure. The effect of operons on gene order and synteny conservation will depend on their frequency, and also on the relative size of the operon and the average size of a conserved segment. If operons are rare, or are small compared to the size of conserved segments, then they are unlikely to affect synteny conservation.

At present we have no idea what the size distribution of evolutionary rearrangements within a taxon looks like. Small inversions in eukaryotes may be tolerated by natural selection, whereas larger ones may be selected against because they may interfere with chromosome pairing at meiosis and could even lead to speciation (White, 1978). If the size distribution of inversions is affected by selection this could lead to quite different patterns of gene order evolution in prokaryotes and eukaryotes, with conservation of operon structures being an important factor in prokaryotes.

1.2.6 Introns as a tool for investigating genome evolution

Any sequence not subject to selection will be randomised over evolutionary time (Kimura, 1983) and should be useful for examining the neutral processes at work in the genome. Unfortunately, it is often impossible to identify orthologous stretches of non-coding DNA. Introns are special in this regard because although they are non-coding (with some exceptions containing transcribed genes, *e.g.*, Levinson *et al.*, 1990), their orthologous relationship can be inferred by virtue of their equivalent positions in orthologous genes. This means that introns may be useful tools for quantifying background noise in genome comparisons, such as genome compaction, and nucleotide biases.

In an analysis of vertebrate genes Duret *et al.* (1995) showed that average gene length varies with G+C content, with genes coding for long proteins being rare in G+C rich isochores (regions of base composition homogeneity Bernardi, 1989). Similarly, introns in G+C poor isochores are on average three times longer than those in G+C rich isochores (Duret *et al.*, 1995; Hurst *et al.*, 1999). Eyre-Walker (1993) observed that the recombination rate is highest in regions of high G+C content. Combining these two observations indicates that introns are, on average, smaller when the recombination rate is higher. There are at least two competing explanations for this phenomenon: mutational bias (Duret *et al.*, 1995); and selection (Carvalho and Clark, 1999). Mutational bias may arise if recombination induces deletions. Recombination is more frequent in G+C rich isochores and thus introns in these regions of the genome are more likely to be truncated by a deletion event than those in G+C poor isochores. The selectionist hypothesis is based on the premise that introns of extreme lengths (either short or long) are slightly deleterious. Selection is most efficient when the recombination rate is high (Nordborg *et al.*, 1996), so the introns in G+C rich regions will be reduced faster than those in G+C poor regions.

1.3 Genome content evolution

Genome content is not static. Gene duplication and loss are antagonistic forces changing the landscape of the genome. The number of genetic loci an organism can support is dependent on the mutational load, which itself is proportional to the mutation rate (Kimura, 1983). A virus with few genetic loci can tolerate a high mutation rate without risking extinction, whereas mammalian species with an average mutation rate per locus per generation of 10^{-5} (Kimura, 1983) may not be able to support more than 100,000 genetic loci (Ohno, 1985; Eyre-Walker and Keightley, 1999).

The most common mechanism for the origin of new genes is the duplication of existing ones. Gene duplication events may vary in extent, and frequency: small tandem duplications appear to be quite common and larger sub-genomic to whole genome duplications appear to be rarer events. Possible mechanisms for gene duplication are unequal crossing over, replicative transposition, and replicative translocation (Ohno, 1970; Nei, 1987).

1.3.1 Genome duplication

Whole genome duplication is an attractive model because it can explain increases in genome complexity which are rapid and without the interference of dosage effects of genes, as all genes will retain their relative dosages directly after a polyploidy event. Immediately following a genome duplication event the proteome is increased quantitatively, but not qualitatively. Cells in polyploids are generally larger than those in diploids (*e.g.*, Gallardo *et al.*, 1999). The redundancy provided by such an event potentially supplies raw material for the development of new functionality on a grand scale. Ohno formalised this theory in his 1970 book (Ohno, 1970), and his name is thus fundamentally associated with evolution through genome duplication (*c.f.* 'Ohnologue' for one of a homologous pair of genes generated through genome

duplication; Wolfe, 2000).

Autopolyploidy, *i.e.*, a genome doubling itself, copies every gene in the genome simultaneously. A single autotetraploidy event produces a symmetric genome with every gene present in two copies and with each gene in an identical context to its pair, including neighbouring genes and neighbouring regulatory sequences. At this point all loci will be segregating tetrasomically (*i.e.*, each locus contains four alleles). As the genome reverts to disomic inheritance (*i.e.*, contains two alleles at each locus) the former tetrasomic locus will turn into two separate disomic loci, which are free to diverge from each other. Any subsequent analysis of gene pairs resulting from autopolyploidy should find that their divergence times date to the onset of disomic inheritance (Gaut and Doebley, 1997).

An allotetraploid results from a hybridisation between two diploid species. The loci will not usually be tetrasomic, as the chromosomes are too dissimilar to form tetravalents at meiosis. The loci started diverging at the time of divergence of the two species, and so are older than the genome-doubling event. Any phylogenetic analysis of the paralogues within the genome should date the gene divergence to the time of the parental species' divergence (Gaut and Doebley, 1997).

A segmental allotetraploid is one in which the two species that underwent hybridisation had not completely diverged, so some loci will be tetrasomic, and some will be disomic. At a tetrasomic locus, the sequences (alleles) contributed by each parent genome will have started diverging at the time of speciation, but as they are now alleles the same locus, may be subject to genetic drift, and so fixation of particular alleles may occur at random (Gaut and Doebley, 1997). Depending on the speed of diploidisation of these loci and thus the time allowed for genetic drift, alleles may retain a copy of each parental allele, or may become fixed for one or other parental allele (Gaut and Doebley, 1997; Wolfe, 2001). In the former case the divergence of the

paralogous disomic loci will date to the time of parental speciation, and in the latter case the date of loci divergence will date to the time of restoration of disomic inheritance.

1.3.1.1 Reasons to be polyploid ... 1,2,3

Ideas about the advantages of a polyploid episode in the evolutionary history of a species or group of species are based on speculations concerning evolutionary novelty, fitness, and efficiency. The most widely circulated of these can be summarised as follows:

1. Produce new genes (Ohno, 1970)
2. Duplicate entire biochemical pathways (Ohno, 1970)
3. Resist inbreeding and survive population bottlenecks (Li, 1980; Allendorf and Thorgaard, 1984)

Ohno (1970) argued that the most efficient way to make new genes is by duplication of existing genes. Polyploidy is guaranteed to duplicate all genes in every biochemical pathway which may facilitate the divergence of biochemical processes. Such a collective duplication would be highly unlikely by any other means unless the genes were closely linked on a single chromosome. Furthermore, in cases where the relative quantities of gene product need to be balanced, duplication by genome doubling will preserve relative dosage relationships. These arguments are often used to explain the high frequency of polyploidy in nature.

Another advantage to polyploidy is that it buffers the genome against the effects of genetic drift (Li, 1980; Allendorf and Thorgaard, 1984). There are twice as many copies of each gene in a tetraploid with tetrasomic inheritance, which means that a tetraploid population can tolerate a higher frequency of recessive deleterious alleles. The frequency of the recessive phenotype will be q^4 in the tetraploid as compared to q^2 in the diploid, where q is the recessive

allele frequency. This means that for diploid and tetraploid populations of similar small size, the tetraploid population will have a greater chance of surviving a population bottleneck than a diploid population.

1.3.2 The fate of duplicated genes

Theoretically there are three possible fates of a duplicated gene: it may acquire a new function (neofunctionalisation); it may retain only one of the functions or expression patterns of the progenitor gene (subfunctionalisation); or it may fix a null allele and thus become a pseudogene (nonfunctionalistion) (Force *et al.*, 1999).

Intuitively, one of the primary drives to develop the theory of gene and genome duplication was to explain the origin of new gene functions by inferring reduced selective constraints on duplicated genes. This supposed freedom to mutate may not be relevant in reality because a high proportion of amino acid substitutions will have a dominant negative effect, and the only acceptable mutations will be synonymous, or recessive (Hughes, 1994; Gibson and Spring, 1998). Analysis of duplicated loci in *Xenopus lævis* and zebrafish indicated that most of the genes were subject to purifying selection (Hughes and Hughes, 1993; Van de Peer *et al.*, 2001). Although there is some evidence for positive selection on *Hox* genes (Van de Peer *et al.*, 2001), it appears that for most duplicated genes there is no acceleration of the substitution rate as might be expected if redundancy truly afforded them the freedom to accumulate previously forbidden mutations (Ohno, 1970).

Any kind of gene duplication will generate redundancy in the genome. In some cases, the resulting increased amount of gene product appears to confer enough selective advantage to preserve duplicate copies of genes (*e.g.* ribosomal proteins have been preserved in duplicate in yeast yet are often 100% identical, so cannot be functionally diverged: Seoighe and Wolfe, 1999b). Whereas, in other cases, gene silencing seems inevitable unless some

change occurs in the duplicated loci to ensure their preservation. This change could be the divergence of protein sequence between the copies, or divergence of regulatory elements of the gene. Alternatively, redundant genes which appear to be 'dispensable' may be preserved simply by virtue of stochastic processes (*e.g.* the serum albumin gene: Ohno, 1985).

If one copy of a duplicate locus becomes fixed for an advantageous mutation, then selection will confer some protection on it from subsequently fixing a null allele (Walsh, 1995). The early life of a duplicate locus can be considered as a race between fixing an advantageous mutation, or fixing a null one. Both neofunctionalisation, and subfunctionalisation (discussed below) are advantageous to the locus in question because they will give rise to selective forces that will act to preserve the locus by selecting against null or deleterious mutations.

Most models addressing the fate of duplicated genes only considered two outcomes: gene silencing, or neofunctionalisation of one copy (*e.g.*, Walsh, 1995). Force *et al.* (1999) emphasised a model which could reconcile the apparent contradictions of the unexpected high incidence of preservation of duplicate genes, the low mutation rate, and the frequency of duplicate loci having different functions. They called their model 'Duplication-Degeneration-Complementation' (DDC) or, subfunctionalisation.

In the subfunctionalisation model of Force *et al.*, degenerative mutations may actually contribute to the preservation of a duplicate locus. The key novelty of this model is that rather than attributing different expression patterns of duplicated genes to neofunctionalisation, they attribute it to a partial (complementary) loss of function of each duplicate, so that combined they retain the complete function of the pleiotropic original gene, but neither of them alone is sufficient to provide full functionality. This model is an expansion of an idea proposed by Li (1980) and is similar to a model proposed by Hughes (1994) where he talks about the partitioning of the

parent gene's functions between the daughter genes. Force *et al.*'s model emphasised subfunctionalisation of regulatory elements of genes, whereas Hughes' emphasised protein sequence divergence, but the principle is the same. This model requires that at least one mutation occurs at each duplicate locus and that they are complementary mutations (such as were observed in the *cut* locus of *Drosophila*: Force *et al.*, 1999, and references therein).

Despite the preservative influences discussed above, the most common fate of a duplicated gene appears to be silencing. This is evident from the short half-life of duplicate genes (Lynch and Conery, 2000, 2001) and from recent polyploid genomes where only a small proportion of the genes remain in two copies. For example, in yeast only 8% of the genes of the pre-duplication genome remain in duplicate (Seoighe and Wolfe, 1998). For *Arabidopsis* one estimate is that 14% of the genes of the pre-duplication genome remain in duplicate (Vision *et al.*, 2000).

The theoretical expectation for the retention of duplicate genes is dependent on the effective population size (N_e), and the ratio (ρ) of advantageous to null mutation rates (Walsh, 1995). When the effective population size is small the probability of retention of both duplicates, $P(r)$, is effectively equal to ρ. For larger N_e the probability of retaining both copies of a gene will be proportional to $N_e\rho$. For convenience $4N_es$ is written as $\mathcal{S}$, where s is the selective advantage of advantageous mutations. Walsh (1995) showed that:

$$P(r) = \left(\frac{1 - e^{-\mathcal{S}}}{\rho\mathcal{S}} + 1\right)^{-1} \tag{1.1}$$

In an analysis of duplicate genes from various taxa, Lynch and Conery (2000) estimated that there is a high turnover of duplicate genes in eukaryotic genomes. Their analysis provoked some criticism both of their data set and their methods (Long and Thornton, 2001; Zhang *et al.*, 2001) and

upon re-analysis of their data Lynch and Conery (2001) concluded that the average half-life of a duplicate gene is 23.4 million years (rather than the previous estimate of 3.2 million years). However, their conclusions remained unchanged: that the rate of origin of gene duplicates in eukaryotic genomes is high; and that most duplicates have a relatively short life.

There is some evidence that the rates of attrition will be faster if there are more than two loci with the same function. Li (1980) provides a theoretical argument in favour of this hypothesis, and evidence from the *Hox* clusters of *Fugu* and zebrafish appear to support this claim. The *Hox* clusters of these bony fish were probably duplicated in a whole genome duplication event to give rise to eight clusters (Amores *et al.*, 1998; Gates *et al.*, 1999). Despite this, these fish do not have substantially more genes in their *Hox* clusters than we see in the mammalian genome. In *Fugu* the eight clusters were resolved to four *Hox* clusters, two *HoxA* clusters, and *HoxB* and *C* (Aparicio *et al.*, 1997) containing in all 31 genes (compared to 39 in mammalian *Hox* clusters). The resolution of the eight *Hox* clusters proceeded differently in the zebrafish genome where there are seven clusters containing a total of 49 genes (Amores *et al.*, 1998).

1.3.3 Evidence for genome duplication in eukaryotes

The predictions of a genome duplication model are unclear. They are highly dependent on the unquantified rates of post-genome-duplication events such as independent gene duplications, gene losses, and genome rearrangements. Depending on the contributions of these scrambling factors, a paleopolyploid genome (degenerate polyploid genome) may look more or less like the symmetric genome that existed immediately after duplication. Due to the frequency of duplicate gene loss (Lynch and Conery, 2000, 2001), and of genome rearrangement, a paleopolyploid may actually only have a small proportion of its genes present in two copies, and of these only a subset

may be conserved in their original context of neighbouring genes. Despite these confounding influences, strong evidence for a polyploid past has been found for several eukaryotic genomes including fungi (baker's yeast; Wolfe and Shields, 1997), plants such as maize (Ahn and Tanksley, 1993; Gaut and Doebley, 1997), and *Arabidopsis* (*Arabidopsis* Genome Initiative, 2000; Blanc *et al.*, 2000; Paterson *et al.*, 2000; Vision *et al.*, 2000), and animals such as zebrafish (Amores *et al.*, 1998; Gates *et al.*, 1999), and *Xenopus* (Hughes and Hughes, 1993; Flajnik and Kasahara, 2001).

Evidence for genome duplication in eukaryotes not only illustrates the strength of the theory of evolution by genome duplication (at least for some taxa), but, by a kind of circular logic, gives us insights into the characteristics of a paleopolyploid. Details of some of these genomes are given in the next sections.

1.3.3.1 Paleopolyploidy in the *Saccharomyces cerevisiæ* genome

Shortly after the public release of the complete proteome of baker's yeast *Saccharomyces cerevisiæ* of 6000 genes (Goffeau *et al.*, 1996) this model organism, selected for sequencing because of its small genome, was shown to be a degenerate polyploid (Wolfe and Shields, 1997). This apparent paradox is resolved by the fact that only approximately 8% of the genes of the pre-duplication genome remain in two copies (Seoighe and Wolfe, 1998).

Wolfe and Shields adopted a map-based approach to analyse the duplication history of the *S. cerevisiæ* genome, and compared the relative position and orientation of paralogous genes within the yeast genome. Using this method they identified 55 blocks of similar gene content with almost conserved gene order and orientation (with some allowance for small inversions). This work was updated by Seoighe and Wolfe (1999a) who identified 52 paired regions in the genome, and a further 32 possible paired regions. The absence of overlapping blocks, and the conserved

orientation relative to the centromere, support the hypothesis of whole genome duplication.

Some authors (*e.g.*, Llorente *et al.*, 2000) propose an opposing interpretation of the duplicates in the yeast genome. They argue for an alternative model of sequential segmental duplications. The two complementary models can be distinguished because they make different predictions about the presence of overlapping blocks, and about the relative orientation of blocks. Additional evidence for a genome duplication rather than segmental duplications in yeast comes from the observation that 50 out of 55 of the blocks have conserved orientation relative to the centromere (Wolfe and Shields, 1997). This observation is only compatible with a segmental duplication model if one invokes a model of preferential insertion or survival in a particular orientation (Wolfe, 2001).

El-Mabrouk (2000) developed an algorithm to reconstruct the post-duplication genome by retracing genome rearrangements. Applying this algorithm to the yeast genome, and assuming polyploidy to be true, she concluded that a minimum of 45 rearrangement events were required to retrieve the symmetric genome (El-Mabrouk, 2000).

1.3.3.2 Paleopolyploidy in the *Arabidopsis* genome

Like yeast, *Arabidopsis* was chosen for sequencing because it is a supposedly streamlined, compact genome (Meyerowitz, 2001). Almost immediately upon sequencing it was noticed that there are large internal repeats in the genome (Lin *et al.*, 1999; Mayer *et al.*, 1999; Terryn *et al.*, 1999; *Arabidopsis* Genome Initiative, 2000) suggestive of a polyploid history of this genome which was surprising in view of its relatively small genome size of 125 Mb.

Conserved regions between chromosomes may become apparent as significant diagonals in a chromosome *vs.* chromosome dotplot. This approach was adopted at the nucleotide level by several groups of researchers (Lin

et al., 1999; Blanc *et al.*, 2000; Paterson *et al.*, 2000) and at the protein level by others (McLysaght *et al.*, 2000b; Wolfe, 2001). Both approaches revealed prominent diagonal lines relating portions of different chromosomes to each other. Close inspection of the first two *Arabidopsis* chromosomes to be sequenced and available portions of other chromosomes showed evidence for tandem gene duplications and genome rearrangement events that post-date the polyploidy event, thus disrupting the symmetry of the genome (Lin *et al.*, 1999; Wolfe, 2001).

Different analyses of this genome arrived at contrasting conclusions regarding the number and timing of genome duplications in *Arabidopsis*. One early analysis of the *Arabidopsis* genome compared the 80% complete genome with sequences from the tomato genome. The authors inferred two major duplication events, one 112 Mya, and another 180 Mya (Ku *et al.*, 2000). The *Arabidopsis* Genome Initiative (AGI) subsequently reported that there were 24 duplicated segments in the genome none of which were triplicated and supposed that these might have been formed in the more recent of the two polyploidy events proposed by Ku *et al.*. By contrast Vision *et al.* (2000) identified 103 putatively duplicated blocks with extensive overlaps and postulated several rounds of whole genome duplication in the history of this plant. Both in this study and in their previous analysis (Ku *et al.*, 2000), Vision and colleagues used the assumption of a molecular clock to date gene duplication events. The validity of this approach has been challenged for its reliance on the assumption of a constant rate of evolution at non-synonymous sites in all genes (Wolfe, 2001). Another contender for the date of the genome duplication in the plant lineage comes from an analysis of duplicate eukaryotic genes by Lynch and Conery (2000) who reported a 'conspicuous peak' at about 65 Mya in an age distribution based on substitutions at synonymous sites of duplicate *Arabidopsis* genes.

Despite these disagreements about timing and mode, the genome map

comparisons in these analyses leave a weight of evidence in favour of a polyploid past for this plant genome; the differences of opinion concern only the number and timing of polyploidy events.

1.3.4 Genome duplication in an ancient vertebrate - The 2R hypothesis

In his 1970 book Ohno proposed that there may have been one or more whole genome duplications in the vertebrate lineage. He postulated that genome duplication in the vertebrate lineage provided a platform for increasing the sophistication of the vertebrate genome and thus increasing morphological complexity. It may be particularly powerful because all genes in a biochemical pathway will be duplicated simultaneously. Ohno was not specific about how many events occurred. The most popular form of this hypothesis is that there were **2 R**ounds of genome duplication early in the vertebrate lineage, which has recently become known as the 2R hypothesis. The hypothesis in this form was proposed by Holland *et al.* (1994). There is no absolute consensus on the timing of these events, but the majority of references in the literature put one of these events immediately before, and one immediately after the divergence of agnathans from the lineage leading to tetrapods (Skrabanek and Wolfe, 1998, Figure 1.5). These timings are speculative and were probably chosen to coincide with major evolutionary transitions that they were thought to have facilitated. The lower limit on the timing of genome duplication is set by the observation of only a single *Hox* cluster in the invertebrate chordate *Amphioxus* compared to four clusters in vertebrates (Garcia-Fernandez and Holland, 1994). As an upper limit, it seems unlikely that genome duplications would be viable in the mammalian lineage. Theory predicts that a genome duplication in an organism with a chromosomal basis of sex-determination (such as that of mammals) will result in sterility of the heterogametic sex, and thus inviability (Muller, 1925). Indeed the only known tetraploid mammal,

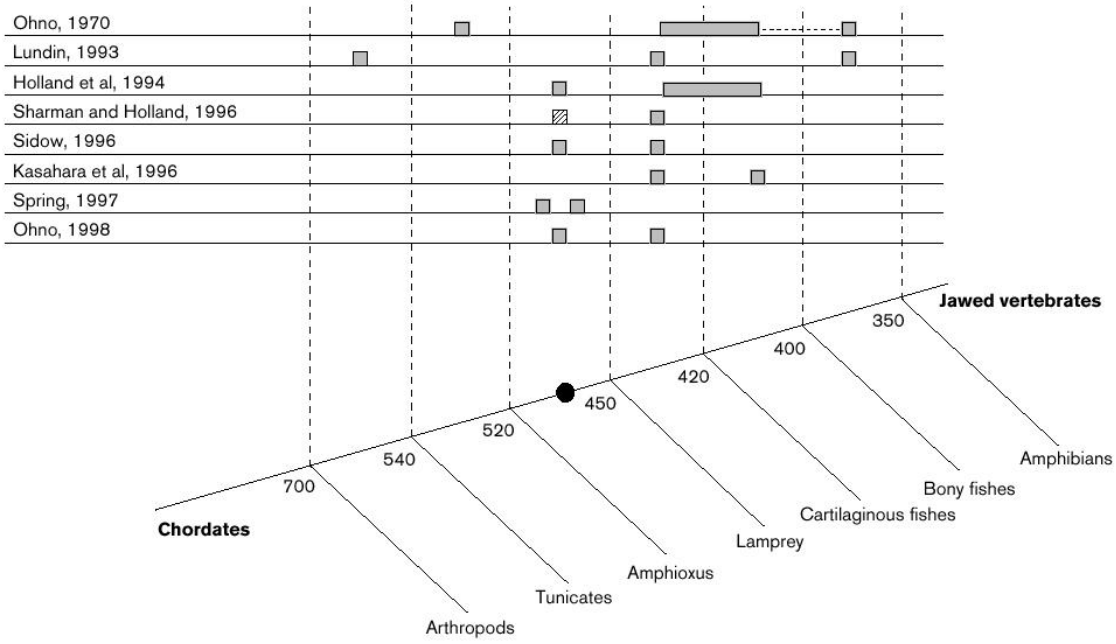


Figure 1.5: Summary of proposed timings of duplication events in the vertebrate lineage. Shaded boxes indicate proposed genome duplications. The hatched box indicates a proposed wave of gene duplications. The circle indicates the time of origin of vertebrates. The species tree is not to scale. This figure is taken from Skrabanek and Wolfe (1998).

a South-American rodent, has duplicated copies of every chromosome except the X chromosome (Gallardo *et al.*, 1999).

At the time of writing his book there was little evidence to support Ohno's claim. Very few protein sequences were known, and the hypothesis was based largely on genome size comparisons and matching patterns of cytogenetic bands. Much of the evidence which prompted Ohno to suggest a genome duplication event has lost merit in the light of our current understanding of genetics and genomes. For example, differences in genome sizes is largely due to increased amounts of non-coding DNA rather than an increased number of genes; and cytogenetic bands, whose patterns were used to list human chromosomes in pairs (Comings, 1972), are not indicative of the underlying gene content.

The debate on the 2R hypothesis to date has been a war of words between the phylogeneticists and the cartographers. As a general rule, phylogenetic methods come out against the genome duplication hypothesis (*e.g.*, Hughes, 1998, 1999b; Martin, 1999; Hughes *et al.*, 2001; Martin, 2001), whereas map-based studies come out in favour (*e.g.*, Lundin, 1993; Spring, 1997).

There are two main arguments used to support the theory of genome duplication in an early vertebrate: that there should be four vertebrate orthologues of each invertebrate gene, the so-called 'one-to-four rule' (Spring, 1997; Meyer and Schartl, 1999; Ohno, 1999); and that paralogous genes are clustered in a similar fashion in different regions of the genome (*e.g.*, Martin *et al.*, 1990; Lundin, 1993).

1.3.4.1 The one-to-four rule

The one-to-four rule was first proposed by Jürg Spring (1997). He listed human paralogues present on different chromosomes and their *Drosophila* orthologues, and surmised that the maximum ratio of human to *Drosophila* genes was four. These 'tetralogues' seemed to bear the hallmark of a genome-wide event because they were discovered on all 23 female human chromosomes. The observation of 2:6, and 2:5 *Drosophila*:human genes contradicts this hypothesis and Spring suggested that more complete genome sequences would provide the data that can split these families into 'tetrapacks'.

The first extensive examination of the one-to-four rule using almost complete proteomes from *D. melanogaster*, *C. elegans*, and human, showed no excess of four-membered vertebrate gene families (Lander *et al.*, 2001, Fig. 49; and Venter *et al.*, 2001, Fig. 12, reproduced here in Figure 1.6). Furthermore, the observation of gene families with five or more members directly contradicts the expectations of Spring (1997) that membership would be 'maximally four'. It appears that the one-to-four rule is an over-simplification of the history of the vertebrate genome. These data

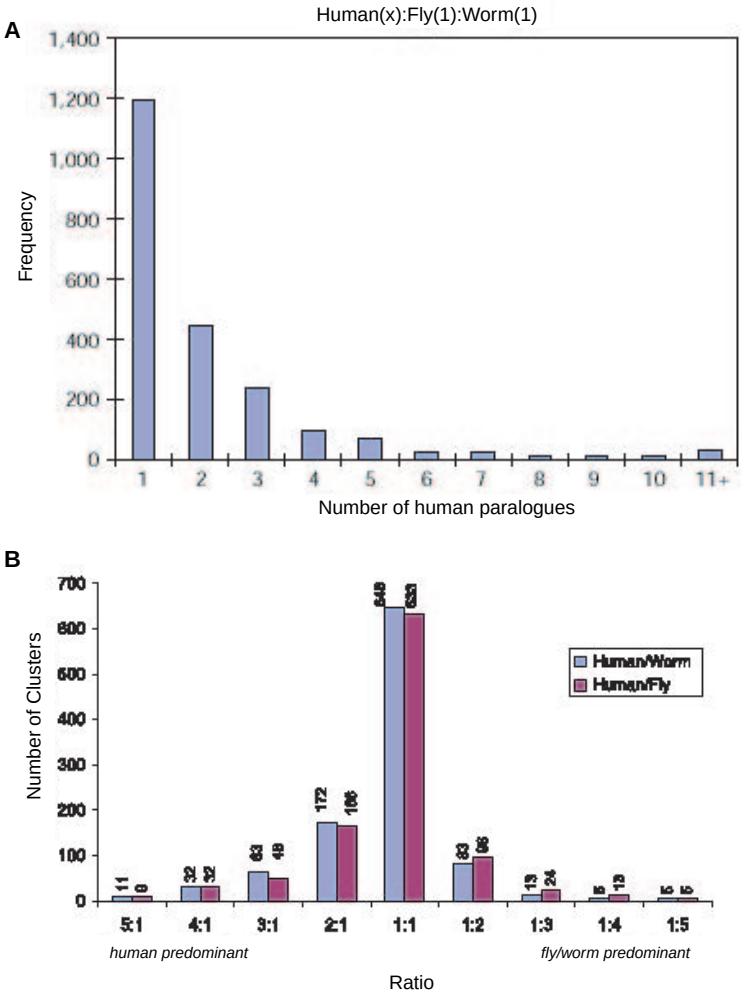


Figure 1.6: (**A**) Number of human paralogues of genes having single orthologues in fly and worm. Taken from Lander *et al.* (2001). (**B**) Numbers of gene clusters (gene families) with varying ratios of invertebrate to human family size. Taken from Venter *et al.* (2001).

can of course be explained by hypothesising two genome duplications on a background of independent gene duplication and loss. However, as it is impossible to distinguish genome duplication from gene duplication on the basis of gene family size alone, this measure is simply uninformative.

1.3.4.2 Paralogous chromosomal segments

The analysis of paralogous regions of the human genome is based on the assumption that, although it is expected that many rearrangements will have occurred in the time since the two duplication events envisaged by the 2R hypothesis, there should still be detectable remnants of the 4-way paralogy between some chromosomes, *i.e.*, some portions of some chromosomes should remain almost intact in four copies. This principle seems correct, though these studies have suffered for want of extensive genomic data. Finding as few as two genes in several linked clusters in a genome of over 30,000 is hardly overwhelming evidence for a genome duplication event (*e.g.*, Martin *et al.*, 1990). Objections that these observations can easily be explained by regional duplications of segments of chromosomes must be entertained.

MHC locus - HSA 1, 6, 9, 19 The observation of paralogous regions around the MHC locus on human chromosomes 1, 6, 9, and 19, led to the suggestion that these were duplicated by whole genome duplication events at the base of the vertebrate lineage (Kasahara *et al.*, 1996; Katsanis *et al.*, 1996; Kasahara *et al.*, 1997; Flajnik and Kasahara, 2001). This was further supported by the finding of only a single related cluster in *Amphioxus* (Flajnik and Kasahara, 2001). Ten members of particular gene families are present on chromosomes 6 and 9, and four of these are also represented on chromosome 1. The claim that this arrangement resulted from several rounds of polyploidy was refuted by Hughes (1998) using phylogenetic analysis of the nine families with sufficient data (Retinoid X receptor (RXR); α pro-collagen (COL); ATP-binding cassette (ABC) transporter;

Proteasome component β (PSMB); Notch; Pre-B-cell-leukemia transcription factor (PBX); Tenascin (TEN); C3/C4/C5 complement components; Heat shock protein 70 (HSP70)). However, Hughes' analysis did indicate that this arrangement could be partly due to block duplication. Trees of these families showed that five (RXR, COL, PBX, TEN, C3/4/5) of the nine families with sufficient phylogenetic information could have duplicated simultaneously, and that this timing was consistent with a duplication in early vertebrate history 550-700 Mya. The phylogenetic analysis indicated that the four genes on chromosome 1 probably duplicated as a block. Similarly, a phylogenetic analysis by Endo *et al.* (1997) rejected the hypothesis that the 11 gene pairs on chromsomes 6 and 9 were duplicated in a single event, but did support the simultaneous duplication of six of the pairs. However, analysis of the remaining genes showed that the ABC transporter genes diverged before the origin of eukaryotes, the PSMB and the HSP70 gene families both originated before the divergence of animals and fungi, and the Notch genes diverged before the origin of deuterostomes (Hughes, 1998). Obviously these gene families did not arise as part of a block duplication event at the base of the vertebrate lineage. However, it can still be argued that these results are consistent with block duplication of this region if one assumes that there was an ancient tandem duplication of some of these genes, and after block duplication there was differential loss of one of the tandems, so that the divergence date of paralogues on two different chromosomes is that of the tandem duplication event rather than of the block duplication event (Kasahara *et al.*, 1996; Smith *et al.*, 1999).

HSA 4, 5, 8, 10 Pebusque *et al.* (1998) reported the presence of paralogous genes on human chromosomes 4, 5, 8, and 10. In contrast to the analysis of the genes around the MHC discussed above, this study was based on a combination of phylogenetic and map-based methods. These genes are linked on the human chromosomes, with the exception that

there is one family member on each of chromosomes 2, and 20, which require genome rearrangements to be reconciled with a block duplication event. The phylogenetic analyses consistently showed that these gene family members diverged in the vertebrate lineage and so are consistent with the 2R hypothesis of genome duplication. This conclusion was criticised by Martin (1999) who pointed out that the gene trees of the ankyrin family and the EGR (Early growth response) family indicated different histories for their host chromosomes. The ankyrin gene tree groups chromosome 4 and 10 to the exclusion of chromosome 8, whereas the EGR gene tree groups 8 and 10 to the exclusion of all others. This contradicts the expectation that the family members on each chromosome have had a shared history since the block duplication event.

Other regions Some of the supposed paralogous regions of the vertebrate genome that can be found listed in the literature are based on rather sparse evidence. For example Gibson and Spring (2000) list human chromosomes X, 4, 5, and 11 as a possible paralogous quartet based only on the presence of members of two gene families (α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) and androgen/mineralocorticoid/glucocorticoid/progesterone nuclear receptors) on all of these chromosomes.

The colinearity of the four vertebrate *Hox* clusters provide the strongest map-based evidence of block duplication known. Phylogenetic tests of the relationship of the *Hox* clusters to each other are discussed in detail in section 1.3.4.3 overleaf.

The arguments for paralogous regions as remnants of block duplication events have suffered from sloppy methodology in some cases. As described above, Kasahara *et al.* (1996) included some quite ancient genes in the region they supposed to have been duplicated at the base of the vertebrate lineage. A worse failing was in the possible paralogous regions reported by Lundin

(1993) which included some genes that are not homologous (*e.g.*, the malate dehydrogenase genes on chromosomes 2 and 7 are not homologues despite sharing similar names: Hughes *et al.*, 2001). In general, the unforgiving methods of phylogeneticists save them from this disgrace because a bad homology definition becomes immediately obvious upon inspection of the tree. This does not mean, however, that map-based studies, with a carefully applied methodology, are impotent in addressing the question of block duplications.

1.3.4.3 Phylogenetic analysis of the 2R hypothesis

In its simplest form, the hypothesis of two rounds of genome duplication predicts a symmetric (A,B)(C,D) phylogenetic tree topology (where A, B, C, D, represent any four-membered gene family), with the age of the AB split the same as the age of the CD split, thus displaying the history of successive genome duplications. The alternative hypothesis, that of sequential gene duplication, will not always predict a symmetric topology. Under a sequential duplication model a four-membered family must arise from the duplication of one member of a three membered family. There is only one possible topology for three sequences, namely (A(C,D)). Duplication of gene A will result in a symmetric topology, and duplication of either C or D will result in an asymmetric topology. Assuming that all three genes are equally likely to be duplicated, sequential gene duplication will give rise to a symmetric (A,B)(C,D) topology 1/3 of the time, and an asymmetric topology (A((B,C)D)) or (A(C(B,D))) the remaining 2/3 of the time (Figure 1.7).

The quadruplication of the *Hox* cluster is the icon of the 2R hypothesis. There are four colinear *Hox* clusters in the the vertebrate genome (Kappen *et al.*, 1989), but only one in the invertebrate chordate *Amphioxus* (Garcia-Fernandez and Holland, 1994). Phylogenetic analysis of the clusters showed that they duplicated early in vertebrate history (Zhang and Nei, 1996). It

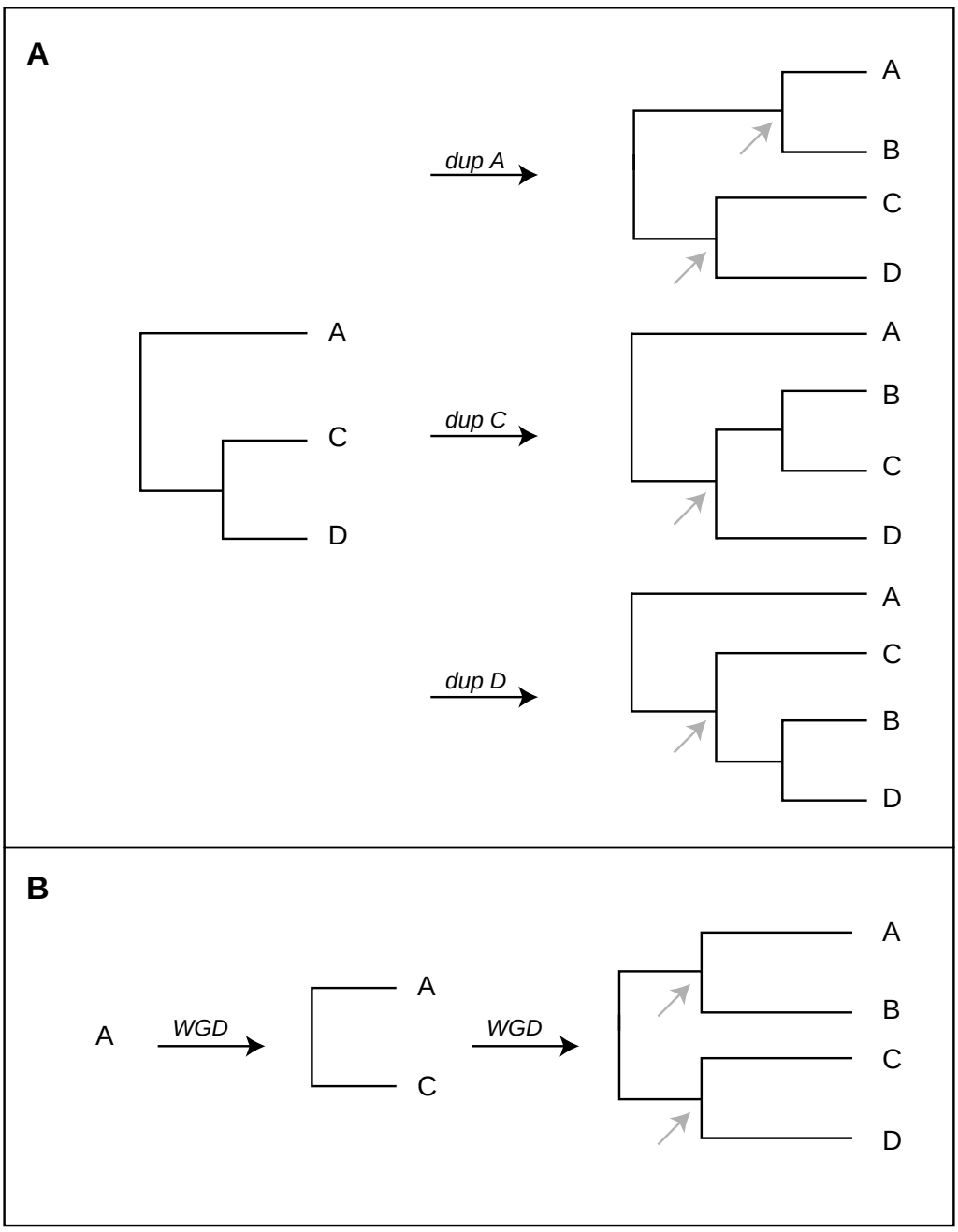


Figure 1.7: Alternative phylogenetic tree topologies of four-membered families resulting from sequential gene duplication or genome duplication. Grey arrows indicate the nodes that are critical to define the symmetry or asymmetry of the topology. (**A**) Phylogenetic tree topologies resulting from duplication of one member of a three-membered gene family. Three different trees result. The tree from the duplication of gene C and that from the duplication of gene D have asymmetric topologies. (**B**) Phylogenetic tree topologies resulting from two whole genome duplication (WGD) events. All genes are duplicated at each step, resulting in a symmetric tree topology.

seems certain that these clusters duplicated *en bloc*. The question is whether they arose by genome duplication events, or by sub-genomic duplication events, or a mixture of both. In the analysis of Zhang and Nei (1996) *Hox* clusters C and D were grouped with a high bootstrap, but there is not enough information in the alignments of the 61 amino acids of the homeodomain to resolve the phylogeny further. Instead, Bailey *et al.* (1997) analysed the relationship of the linked fibrillar-type collagen genes which presumably shared the same duplication history. Assuming the collagen genes have a shared history with the *Hox* clusters, then the results can be interpreted as a topology (outgroup(*HoxD*(*HoxA*(*HoxB*,*HoxC*)))), which contradicts the grouping of *HoxC* and *HoxD* found by Zhang and Nei (1996). This is contrary to the expectations of the 2R hypothesis which predicts a symmetric topology, but may be explained by three rounds of genome duplication with loss of 4 clusters, or by independent cluster duplications (Bailey *et al.*, 1997).

In a phylogenetic analysis of the human *Hox*-bearing chromosomes (2, 7, 12, 17) Hughes *et al.* (2001) examined 35 gene families with members on at least two of the *Hox* chromosomes. 15 of these families could be classified as either pre-vertebrate, or post-mammalian and so are inconsistent with the 2R hypothesis. For the remaining 17 gene families the tree topologies did not exclude duplication at the same time as the *Hox* clusters. There were 15 of these for which the molecular clock was not rejected and estimates for the divergence dates of these gene families were calculated. Six of the gene families were dated to within the time of divergence of the *Hox* clusters 528-750 Mya (as defined by lineage divergences), and two others had divergence estimates that were not significantly different from the time of *Hox* duplication. Phylogenies of gene families with members on at least three of the four *Hox* bearing chromosomes did not reveal a common topology for the relationship of these chromosomes.

Hughes (1999b) and Martin (2001) employed similar methodologies to

test the phylogenies of gene families listed as illustrations of the one-to-four rule (Sidow, 1996; Spring, 1997) for congruence with the 2R hypothesis (*i.e.*, whether or not they displayed a symmetric topology, and if they duplicated in the vertebrate lineage). The symmetric topology was only observed in a small minority of the cases (one out of nine trees in Hughes, 1999b; and two out of ten trees in Martin, 2001), although in Martin's analysis seven of the eight minimum-length trees that were not symmetric were not significantly shorter than a symmetric tree.

Variations on the 2R hypothesis result in different predictions for the phylogenies of vertebrate gene families. For example, if vertebrate genome doubling occurred by allopolyploidy (*i.e.*, hybridisation of two species, as has been suggested, Spring, 1997) or by segmental allopolyploidy (*i.e.*, behaving as an autopolyploid at some loci, and as an allopolyploid at others) then a single genome doubling event will produce paralogues with two different coalescence dates (Gaut and Doebley, 1997; Wolfe, 2001). Alternative models hypothesise that the two rounds of genome duplication may have occurred in short succession and thus not allowing the diploidisation procedure time to complete before the second genome duplication event. This would result in some tetrasomic loci, and some octasomic loci in the quadruplicated genome (Gibson and Spring, 2000).

1.3.5 Diploidisation

Diploidisation is a natural consequence of polyploidy. With some rare exceptions (*e.g.*, some loci of recent salmonid tetraploids; Allendorf and Thorgaard, 1984) all hypothesised paleopolyploid genomes have reverted to disomic inheritance at all loci. There is an increased incidence of non-disjunction of chromosomes when they form multivalents rather than divalents, so selection for increased fertility probably causes the reinstatement of disomic inheritance (Allendorf and Thorgaard, 1984).

Immediately after autotetraploidy all loci in the genome will be tetrasomic. These duplicated genes will not separate into two independently diverging loci until disomic inheritance is established (Ohno, 1970). This is important for our interpretation of what a paleopolyploid genome should look like because one of the properties we test in assessing genome duplication is the synchronicity of divergence of duplicated loci. Depending on the manner and speed of diploidisation this may or may not be an appropriate test for a paleopolyploid genome.

In a diploid organism, chromosomes are arranged in pairs at meiosis (*i.e.*, chromosomes are bivalent). These pairs can exchange segments of DNA by recombination, and drift and gene conversion maintain a high degree of similarity between most alleles. In a tetraploid genome, chromosomes are arranged in tetravalents, rather than pairs, at meiosis. Diploidisation can be reduced to a problem of chromosome association. By what mechanism does a genome convert from forming chromosome quartets to forming chromosome pairs, *i.e.*, from tetraploid, to diploid behaviour?

The answer to this question probably lies in a deeper understanding of the mechanisms of chromosome association. Is chromosome sequence divergence a cause or a consequence of diploidisation? If chromosome association occurs by homologous sequence attraction, then sequence divergence (by chromosome rearrangements) will cause diploidisation of chromosomes. On the other hand, if chromosome association is controlled by some other mechanism, such as attraction of homologous centromeres or telomeres, then chromosomal rearrangements may allow the independent evolution of the relocated loci and their previous partners in a tetrasomic locus, as separate loci without actually causing the diploidisation of the chromosomes in question.

The Y chromosome is unusual in the human genome because it is partially diploid (at the pseudoautosomal region), and the rest is haploid. Lahn

and Page (1999) examined the evolution of the human sex chromosomes. They identified homologous genes on the human X and Y chromosomes which would have been part of the same locus when these chromosomes behaved autosomally (the sex chromosomes are thought to have evolved from autosomes; Graves, 1996). They measured the amount of divergence at synonymous sites (K_s) between homologous gene pairs. From this they found that the homologues were in four ages classes arranged sequentially along the X chromosome. They interpreted this as the result of inversions of large sections of the Y chromosome, leaving the X intact, which had the effect of suppressing recombination between these portions of the chromosomes. These chromosomes have diverged substantially, and most of the Y chromosome loci are haploid. The X and Y chromosome still pair at meiosis (at the pseudoautosomal region), and thus behave like diploid chromosomes, yet most of the loci are haploid. It may be the case that chromosomal tetravalency and locus tetrasomy can be separated in the same way.

The wheat genome (*Triticum æstivum*) is hexaploid, the three contributory genomes being labelled *A*, *B*, and *D*. There is evidence for genetic control of chromosome association in wheat through the *Ph1* locus on chromosome V of the B genome (Riley and Kempanna, 1963). In the presence, but not the absence of a particular allele of this locus, non-homologous associated centromeres separate at the beginning of meiosis (Martinez-Perez *et al.*, 2001). The *Ph1* locus probably acts to amplify the differences between non-homologous chromosomes.

The most widely accepted hypothesis is that diploidisation proceeds by structural divergence of chromosomes. Allendorf and Thorgaard (1984) discuss a model whereby some loci may appear disomic while others apparently segregate tetrasomically. In their model they assumed that chromosome pairing occurred at the telomeres, but it can be easily modified

to assume centromere association as is indicated from the study of the wheat genome (Martinez-Perez *et al.*, 2001). The model of residual tetrasomic inheritance hypothesises that there are two stages of chromosome pairing. The first stage will allow pairing between homœologous chromosomes (partially similar chromosomes), thereby allowing recombination events between paralogous loci on different chromosomes. The second stage of pairing in this hypothesis resolves non-homologous chromosome pairing, and ensures that each gamete receives one copy of each chromosome in the normal manner. Evidence in support of this model comes from the observation of Martinez-Perez *et al.* (2001) that some non-homologous centromeres are associated just before the beginning of meiosis. This model predicts that loci closer to the point of association of the chromosomes (*i.e.*, closer to the centromere) will retain residual tetrasomic inheritance longer than others. For any locus, the likelihood that it behaves disomically rather than tetrasomically in a particular meiosis will be correlated with its distance from the centromere.

1.4 Aim

The aim of this thesis is to analyse evidence for mechanisms of genome evolution in the vertebrate lineage. In Chapter 3 genome rearrangement in vertebrate genomes is analysed through a comparative genomics study of the pufferfish (*Fugu*) and human genomes. In Chapter 4 an intragenomic comparison approach is used to examine genome content evolution in the vertebrate lineage.

Chapter 2

Methods in Genome Analysis

This chapter provides an introduction to some of the concepts and methods used throughout this thesis.

2.1 Identifying homologues

Homologous sequences are related by descent from a common ancestor. Highly similar sequences are often, but not always, homologous; and homologues are often, but not always, highly similar. Similar sequences that are not homologous, are called analogues, and arise from convergent processes. There are no degrees of homology. Sequences are either related by descent, or they are not (Reeck *et al.*, 1987; Fitch, 2000).

Orthologues are a subset of homologues where sequence divergence has occurred after a speciation event, *i.e.,* the common ancestor of the two sequences lies in the common ancestor of the species from which the sequences were obtained (Fitch, 1970). The true phylogeny of orthologues is identical to the true phylogeny of the species from which they were obtained (Fitch, 2000).

Paralogues are homologues resulting from a gene duplication event (Fitch, 1970), and may co-exist in the same genome. These genes are paralogues of

each other, but both are orthologues (sometimes called semi-orthologues or co-orthologues; Sharman, 1999; Taylor *et al.*, 2001a) of the equivalent gene in lineages that diverged prior to the gene duplication.

Because homology is defined by descent and descent only, it is not possible to define a percent identity threshold which differentiates the homologous from the similar. Nonetheless, in common practice a search for homologous sequences is a similarity search because we estimate that homologous sequences will retain a high degree of similarity that will stand out above random similarity. More divergent homologous proteins may still be recognisable at the protein 3D structure level (Bork *et al.*, 1992). The most common method for homology estimation is by using the BLAST algorithm to compare the sequence of interest to a database of other sequences.

The Basic Local Alignment Search Tool (BLAST) algorithm (Altschul *et al.*, 1990, 1997) operates on the general strategy of optimising the maximal segment pair (MSP) score, which is a general measure of similarity. It is a heuristic modification of the Needleman-Wunsch (1970) and Smith-Waterman (1981) algorithms. The MSP is said to be locally maximal if the score cannot be increased by extending or shortening the local alignment. The algorithm searches for all maxima above a specified threshold. A major advantage of the BLAST algorithm over the mathematically superior Smith-Waterman algorithm is that it takes much less computation time, a fact which has propagated its use on large datasets.

The fundamental unit of BLAST algorithm output is the High-scoring Segment Pair (HSP). The Maximal-scoring Segment Pair (MSP) is the highest-scoring of all possible segment pairs that can be produced from the two sequences. The expectation value (E-value) threshold defines the significance threshold for reporting results. The E-value reported for any pair is the expected frequency of chance occurrences of equal similarity within the database. As the E-value approaches zero it becomes equivalent to the

probability that the similarity is due to chance, and so the lower the E-value, the greater confidence one has in the reported match being a true homologue.

2.2 Using genome maps

Genome sequencing projects at different levels of completion produce varying quality genome maps. These range from shotgun-sequenced clones (*e.g.*, cosmids or BACs) where only proximity of genes identified on the same clone is known; through completely sequenced clones, where the relative position and orientation of contiguously sequenced genes are known, but the relative positions of clones is unknown; to complete chromosome or genome sequence, where the relative position and orientation of every sequence is known. Bioinformatics is as much about making good use of sparse data as it is about abstracting interesting information from large complete genomes.

2.2.1 Map units

Because genetic distances (in units of centiMorgans) will be influenced by the non-random distribution of recombination events, physical map units are generally preferable. The most precise units of physical maps are base-pair distances, but distances may also be measured in centiRads (units of a radiation hybrid map), or in terms of the number of intervening genes.

Counting the number of intervening genes as a measure of physical distance is useful because it is not influenced by uneven gene density, or by stretches of highly repetitive sequence which may inflate intergenic distances. This is a robust measurement that tolerates comparison across phyla where there can be differences in the extent of the compaction of the genomes, for example, for the purposes of a proximity conservation analysis, measuring the frequency with which genes remain close in two genomes even where there are significant differences in gene density between the species (*e.g.*, some plant

genomes have different gene densities; Keller and Feuillet, 2000).

2.2.2 Incomplete data

Incomplete genome data introduce several complications into any analysis. When searching for orthologues between species with only partial genome sequences available it is not possible to be sure that even highly similar sequences are in fact orthologous, because there always remains the chance that the true orthologue has yet to be sequenced in one species. Orthology can usually be confirmed by drawing phylogenetic trees from homologous sequences where available. Otherwise the best strategy might be to use strict criteria in a BLAST similarity search. Only accepting pairs of sequences with a low expectation value (*e.g.*, $\leq 10^{-15}$), and whose MSP covers at least 30% of the length of the longer protein may be strict enough to filter out most spurious matches. Alternatively, a mutual best hits approach (*i.e.*, where the strong similarity is reciprocated) may be effective in most cases.

When only partial data is available for a species (*e.g.*, shotgun sequencing reads) the relative positions of most genes in the genome are unknown. Comparative genomics studies including this type of data will be limited to the question of whether genes are on the same cosmid, or other clone contig.

2.2.3 Genome maps

In order to study genome rearrangement, extensive genome sequences and/or map data are needed from a variety of species. Genome sequencing is ongoing in the form of concerted international efforts (*e.g.*, Lander *et al.*, 2001) and individual lab efforts, most of which release their data to the EMBL/GenBank/DDBJ centralised database.

Dedicated mapping projects for human (Deloukas *et al.*, 1998), mouse (Avner *et al.*, 2001; Hudson *et al.*, 2001), rat (Watanabe *et al.*, 1999) and

other vertebrate genomes are producing radiation hybrid maps. Radiation hybrid mapping is a technique for generating physical maps using irradiation to break DNA at random locations and then allowing the fragments to fuse with the DNA of a recipient rodent cell line distinguishable from the DNA of interest (McCarthy, 1996). The hybrid DNA is then analysed for the co-retention of markers on the same fragment (Newell *et al.*, 1998). The frequency of this event will be dependent on their physical distance. Maps generated by this method are in units of centiRads (cR). Radiation hybrid maps are useful because it is relatively easy to produce a map with a high density of markers but, unlike genetic maps, the units are directly proportional to physical distances.

2.3 Dating gene duplications

There are two types of information in a molecular phylogenetic tree: topology and branch lengths. The topology is a qualitative trait which indicates the relationship of the sequences to each other. Branch length is quantitative and illustrates the evolutionary distance from one sequence to any other. Both of these can be used to estimate the timing of events (nodes) in the tree. Nodes correspond to either speciation events or gene duplication events. A phylogenetic tree that includes some duplicated sequences can be recognised upon comparison with a known species tree for the source organisms. Gene duplication events can be dated by at least two methods, topology-based methods, and molecular clock based methods.

2.3.1 Topology-based methods

Comparison of a gene family tree topology with the species tree topology shows which species share any gene duplication events. Tree topologies with strong bootstrap support (>80% support is often taken as 'strong support')

are likely to be accurate, and so a gene duplication event can be dated to before or after species divergence events with high confidence. This method is attractive because it is immune to molecular clock criticisms. However, it is limited by available data (one must identify orthologues in related species); and is imprecise in that duplication dates estimated by this method have low resolution. Often with this method it is possible only to say that a gene duplication pre-dates (or post-dates) some particular speciation node. Nevertheless this approach has been useful in 2R studies, allowing Hughes and colleagues to discount some gene duplications as far too old or young to be consistent with the hypothesis (Hughes, 1998; Hughes *et al.*, 2001).

2.3.2 Molecular clock methods

The principle of a molecular clock is based on the assumption of a constant evolutionary rate for a gene. If this assumption is true then the amount of sequence divergence between two sequences should be proportional to the time since they last shared an ancestor (Figure 1.1 on page 3). In the case of gene duplication (and in the absence of gene conversion), the time of the most recent common ancestor is the time of the duplication event.

Because there is no such thing as a Universal Molecular Clock (Li, 1993), the best one can hope for is that a local molecular clock applies, *i.e.,* that the gene of interest and its close relatives have experienced an almost constant rate of evolution in he species under study. When this condition is met recent events in the history of a gene can be dated relative to each other.

Takezaki *et al.* (1995) developed a method to test whether a set of sequences has evolved in a clock-like manner, and if so to estimate divergence dates. Their method converts a phylogenetic tree into a linearised tree, where sequence divergence dates are represented on a linear time scale under the assumption of a constant evolutionary rate (Figure 4.5A). In order to test the molecular clock a phylogenetic tree is inferred from the sequence alignment

without the assumption of rate-constancy. These sequences are then tested for evolutionary rate heterogeneity by the two-cluster test. The two-cluster test is an expanded and modified version of a relative rate test (Sarich and Wilson, 1973) which compares two sequences relative to an outgroup sequence. The two-cluster test examines each node in the tree asking if there is relative rate heterogeneity between the clusters of sequences bifurcating at that node. The two-cluster test reports a χ^2 value which can be converted to a P-value by comparison with a table of χ^2 critical values at $n-1$ degrees of freedom, where n is the number of sequences in the tree. The P-value gives the significance at which the molecular clock can be rejected. Rather than only rejecting extreme abberations of the molecular clock, it is good practice to reject with 1% or even 5% significance (Takezaki *et al.*, 1995).

If the two-cluster test fails to reject the molecular clock then the data are appropriate for constructing a linearised tree. This is done by re-estimating the branch lengths for the given topology under the assumption of rate constancy.

This method has the advantage that there is no need to search for related sequences from other species as with the topology method. This technique may be applied to as few as two sequences with an outgroup.

2.3.3 Correction for multiple hits

Not only do evolutionary rates vary between proteins, but they also vary within proteins, so that not all sites are experiencing the same substitution rate. This means that distance estimation methods assuming a Poisson distribution of substitutions (which forms the basis of many models for correction of multiple hits), where the probability of substitution events is rare but equal over the whole protein, may be misleading. Only if all substitutions are selectively neutral, and if there are a fixed number of cell divisions per year in the germ lines of the species sampled, will

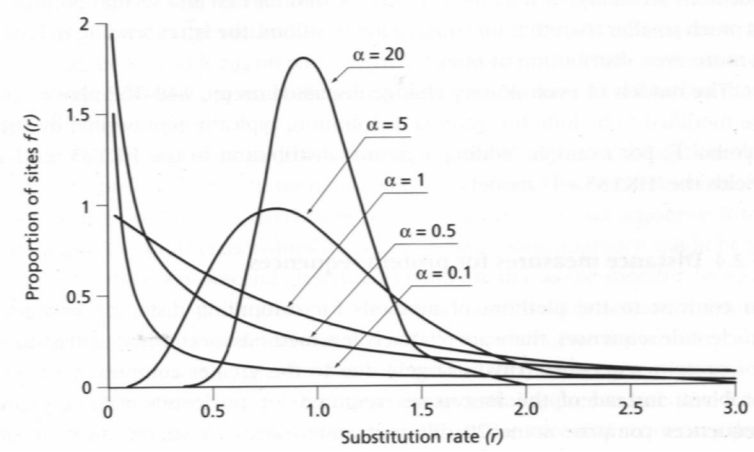


Figure 2.1: The effect of the α parameter on the shape of the gamma distribution. Low α corresponds to high rate variation. This figure is taken from Page and Holmes (1998)

the assumptions underlying the Poisson distribution be met (Uzzell and Corbin, 1971). Selection acting at sites in a protein sequence will violate the assumptions of the Poisson distribution.

Mathematically, rates of evolution can be modelled as a negative binomial distribution (Uzzell and Corbin, 1971) which assumes that substitutions at each site follow a Poisson process, and that the substitution rate varies according to the gamma distribution. This allows for different rates of substitution at different sites. The shape of the gamma distribution is determined by the alpha parameter (Fig. 2.1) which must be estimated from the data. Maximum parsimony methods for alpha estimation tend to overestimate alpha in all cases (Gu and Zhang, 1997). Maximum likelihood methods exist (Yang, 1997) but are prohibitively time-consuming for use on large data sets. A new method for alpha estimation was developed by Gu and Zhang (1997) which is computationally fast, and has comparable accuracy to maximum likelihood methods. This method has two main steps: first the expected number of substitutions corrected for multiple hits is estimated for each site by a likelihood method; then the estimate of alpha is obtained from

a negative binomial distribution using the expected number of substitutions.

Gu and Zhang's **GAMMA** program, and Takezaki *et al.*'s **lintre** program were downloaded from Masatoshi Nei's lab website: `http://www.bio.psu.edu/People/`

Chapter 3

Estimation of synteny conservation and genome compaction between pufferfish (*Fugu*) and Human

3.1 Introduction

Sydney Brenner and colleagues (Brenner *et al.*, 1993; Elgar, 1996) proposed the pufferfish *Fugu rubripes* as a model genome for use in dissecting the human genome. As a vertebrate, *Fugu* is expected to have a similar gene repertoire to human. However, its genome, at ∼400 Mb, is approximately 7.5 times smaller than that of human. The reduced amount of repetitive sequence and high gene density make this small genome attractive to molecular biologists.

There are two main factors that will determine whether *Fugu* will be genuinely useful as a model vertebrate for reference to other genomes. First, *Fugu* genes must show sufficient similarity to their human orthologues so as to enable the isolation of a *Fugu* gene with a human (or other mammalian)

DNA probe and vice versa. Furthermore, knowledge of the extent of linkage conservation between the two genomes will advise as to the feasibility of positional cloning using map information extrapolated from one species to the other (Elgar, 1996). Several regions of conserved synteny (but not necessarily conserved gene order) have already been reported between these two genomes (*e.g.*, Baxendale *et al.*, 1995; Trower *et al.*, 1996; Elgar *et al.*, 1999, and references in Table 3.2). The academic utility of studying this genome extends beyond this. The compacted *Fugu* genome may be useful to highlight potentially functional non-coding regions (by virtue of their preservation). Also, the presence and absence of genes may give some insights as to the constitution of the core or minimum vertebrate genome.

Exploring the relationship between the human and pufferfish genomes in terms of the extent of synteny conservation and patterns of genome compaction could give insights into the evolution of vertebrate genomes, and could also provide more information on the usefulness of *Fugu* as a model genome. However, at present it is not known how large the syntenic regions are, or how well the gene order is conserved between *Fugu* and human. Recent research on zebrafish (*Danio rerio*) indicated that for some groups of genes synteny is conserved in human but the order of the genes along the syntenic chromosome is different in the two species (Postlethwait *et al.*, 1998). Moreover, many mammalian genes have two zebrafish orthologues, and this is probably due to whole genome or chromosomal duplications that occurred in bony fish (including zebrafish and *Fugu*) after their divergence from the tetrapod lineage (Amores *et al.*, 1998; Gates *et al.*, 1999). It is also not known whether the compaction of the *Fugu* genome relative to human is uniform throughout the genome, particularly in view of the uneven distribution of genes in the human genome (Ikemura and Wada, 1991; Duret *et al.*, 1995; Deloukas *et al.*, 1998).

Here we have made a comparative genomics study of *Fugu* and human to

investigate the phenomenon of genome compaction and to estimate the level of synteny conservation. There is no genetic map for *Fugu* (it is not possible to breed this fish in the laboratory), so gene linkage is only discernible at the level of genes that were sequenced on the same cosmid or other clone contig. We used two sources of *Fugu* sequence data: large contiguous genomic sequences determined by a variety of laboratories and obtained from GenBank; and "cosmid skimming" data from the *Fugu* Landmark Mapping Project at the UK MRC HGMP-RC (Elgar *et al.*, 1996, 1999). The human map data was obtained from two sources: the Online Mendelian Inheritance in Man database (OMIM 1999); and the physical map of about 30,000 genes (GeneMap '98) constructed from radiation hybrid data by Deloukas *et al.* (1998). Initial work on the skimmed *Fugu* cosmids was done by Anton J. Enright with help from Lucy Skrabanek. The work described in this chapter has been published (McLysaght *et al.*, 2000a,b).

3.2 Materials and Methods

3.2.1 Analysis of homologous introns from *Fugu* and human

The 22 genes included in this analysis were: RPS3, RPS24, DLST, STK9, PAX6, RPS7, APP, (low GC3 group); SURF3, SMC1, RPL41, ARF3, CFOS, XLRS1, PCOLCE, (medium GC3 group); CSFR1, GH, TSC2, HMOX1, WNT1/INT1, PKD1, G6PD, IT (high GC3 group). All sequences were obtained from GenBank.

3.2.2 *Fugu* sequence data

SwissProt version 37 (27 July 1999) contains 5406 human proteins. These were compared to the database of *Fugu* skimmed cosmids using TBLASTN

(Altschul *et al.*, 1990) using the BLOSUM62 scoring matrix and the SEG filter (Wootton and Federhen, 1996). To remove obvious paralogous hits, only the top hit for each query was retained (provided that it had $P \leq 10^{-15}$) as well as weaker hits that were within a factor of 10^5 of the top hit. The results of this BLAST search including human map information are available at `http://biotech.bio.tcd.ie/~amclysag/skimmed.html`.

A "skimmed" cosmid was deemed to contain two genes if two non-overlapping subclones hit different mapped human proteins that are $< 40\%$ identical in sequence and had $P \leq 10^{-15}$ in a BLASTP search. Overlapping *Fugu* cosmids were identified manually and reduced to one entry in Table 3.1.

Fugu proteins from completely sequenced cosmids were compared to the database of human sequences from GeneMap '98 by the TBLASTN program applying the SEG filter. Only hits with a significance of $\leq 10^{-15}$ and that were no more than 10^5 less likely than the top hit were accepted. Only the best hit per chromosome was included in further analysis.

Some of the limitations on the analysis of the skimmed cosmids become apparent when the results are compared with the fully sequenced cosmids. Cosmid 168J21 has been fully sequenced under accession number AJ010348 (Cottage *et al.*, 1999). The full sequence has three annotated proteins, all of which had human homologues on chromosome 3. In the analysis of the skimmed cosmid sequence only one gene was found. As all three human orthologues are in the SwissProt database, it must be the case that the cosmid subclones do not include the coding sequences of the other two genes.

In a similar analysis the *Fugu* cosmid sequences were compared to the predicted protein set (Solovyev and Salamov, 1999) from the first completely sequenced human chromosome, chromosome 22 (Dunham *et al.*, 1999), using the same protocol.

3.2.3 Human GeneMap '98 sequences

Deloukas *et al.* (1998) compiled a map (GeneMap '98) of human gene-based markers by radiation hybrid mapping. This includes approximately 30,000 genes. By electronic PCR (Schuler, 1997) they found the corresponding genomic sequence, mRNA and/or EST from the public databases. These results are updated weekly and were downloaded from the NCBI FTP site on 21 December 1998.

A BLAST database of human sequences represented on this radiation hybrid map was created. In order to have comparable map units only the data from the GeneBridge4 panel (Gyapay *et al.*, 1996) were included. Some parts of the genome are represented more than once in the ePCR output because they have been sequenced more than once as genomic sequence, mRNA and/or EST. Redundancies of this kind were removed, preferentially keeping genomic sequences over mRNA over unfinished sequences over ESTs. The final database had 28,133 entries totalling 226,506,753 nucleotides.

Some markers in GeneMap '98 are listed with several allocated map positions. In these cases the same position found from several independent experiments or the position with the highest confidence value as determined by Deloukas *et al.* (1998) was used. Distances within the genome were estimated by counting the number of intervening genes in GeneMap '98. We then adjusted these values for missing data by multiplying this number by 80000/30000 (assuming that the human genome contains 80,000 genes (Antequera and Bird, 1993) and the map contains 30,000 genes).

3.2.4 Computer simulation of genomic rearrangement

In order to make this simulation as realistic as possible paralogues were assigned at the frequencies observed in the real data. Of the 91 *Fugu* proteins analysed, 78 had hits in the database of mapped human sequences. The distribution of hits is as follows: 47 hit one human sequence, 14 hit two, eight hit three, two hit four, and families of seven, 11, 12, 15, 39, 42, and

59 human proteins were observed once each. More extensive human protein family size data from an intra-genome comparison (Imanishi *et al.*, 1997) was used to confirm these results in an independent simulation.

3.3 Results

3.3.1 Compaction of *Fugu* introns

The *Fugu* genome is much smaller than the human genome, but by virtue of being vertebrate is presumed to have a similar gene repertoire (Brenner *et al.*, 1993). The difference in size must therefore be primarily due to differences in non-coding DNA, including both intergenic and intronic DNA. In vertebrate genomes there is a correlation between gene length and G+C content, with long genes being rare in G+C-rich isochores (Duret *et al.*, 1995). This suggests that there might be a correlation between base composition and the size difference between a human gene and its *Fugu* homologue.

Orthologous *Fugu* and human introns were identified by finding orthologous genomic sequences in GenBank, aligning the protein sequences using the Gap program (with default settings) of the GCG package, and mapping intron locations onto the protein alignment. Introns were designated orthologous if they were in the same phase and occurred at precisely the same position in the protein alignment produced by Gap. No allowance was made for possible intron sliding during evolution. Using this method, 199 pairs of orthologous introns from 22 genes were found. There were only six cases where we could say with confidence that an intron had been gained or lost after the divergence of these two species. These were all cases where there was an unambiguous alignment of the two protein sequences and where an intron was present in one sequence but there was no equivalent intron nearby or out of phase in the other organism. Non-coincident introns and introns in ambiguous alignments were excluded from further analysis. Recent

research by Hurst *et al.* (1999) tentatively suggests that there may be a dichotomy in the relationship of synonymous G+C content and intron size, with homothermic vertebrates showing a negative correlation, as previously observed, and heterothermic vertebrates (including *Fugu*) showing a positive correlation. However, this is not borne out here. In our dataset there is no correlation between intron size and GC3 content of the genes that house them.

Genes were assigned into three equal-sized groups according to their G+C content at codon third positions (GC3) in human, and the lengths of equivalent introns were compared (Fig. 3.1A). The sum of the lengths of all 199 introns in *Fugu* was 59,392 bp, just over eight times smaller than the sum of the lengths of all the human introns (488,726 bp). The large introns of GC3-poor genes are seen to be severely compacted. The compaction averages are 2.9, 6.0, and 14.6 respectively, for the high-, medium-, and low-GC3 groups of genes (Fig. 3.1A), which is broadly consistent with expectations. One fifth of the *Fugu* introns (41 of the 199) are actually larger than their human counterparts (many only marginally so), and most of these are high-GC3 genes in human (Fig. 3.1B). However, for the majority of introns (Fig. 3.1B) there does not appear to be any consistent relationship between intron lengths in the two species, or between these and GC3 in their host genes.

The compaction of individual genes, instead of individual introns, was also calculated (Fig. 3.1 C, D). Compaction was calculated by dividing the sum of the lengths of introns in a human gene by the sum of the lengths of their *Fugu* orthologues (excluding any non-coincident introns). The compaction values range from 46 (in the APP gene; Villard *et al.*, 1998) down to values of less than 1 in two genes (growth hormone and int1/wnt1) where the *Fugu* gene is larger than the human one. If the GC3 content of a gene and the compaction of its introns are related, then one would expect the greatest compaction to be

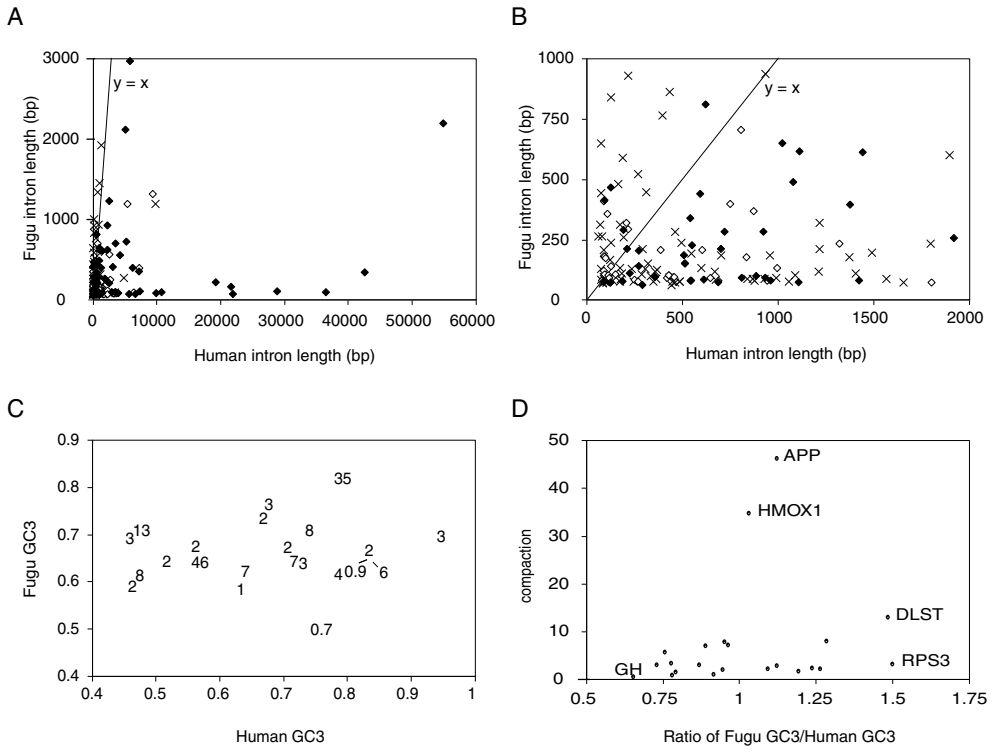


Figure 3.1: (**A**) Lengths of 199 orthologous introns from *Fugu* and human. A line of slope 1 is indicated. The symbols for the points represent different GC3 content categories in the human gene where the black diamond denotes low GC3 (<63.5%), the white diamond denotes medium GC3 (63.5%-76%), and the cross denotes high GC3 content (>76%). The categories were designed in such a way as to have equal numbers of genes in each group. The 22 genes from which the introns are derived are named in Materials and Methods. (**B**) Inset of (A) showing only the smaller introns. (**C**) GC3 content of the 22 orthologous genes whose introns were analysed. The points are replaced by values indicating relative gene compaction. Compaction was calculated by dividing the sum of the lengths of introns of a human gene by the sum of the lengths of their *Fugu* orthologues, ignoring non-conserved introns. (**D**) Compaction of 22 genes versus the ratio of GC3 in *Fugu* to that in human. Outlying genes are labelled: APP, amyloid precursor protein; GH, growth hormone; RPS3, ribosomal protein S3; HMOX1, heme oxygenase; DLST, dihydrolipoamide succinyltransferase.

between human genes with low GC3 and *Fugu* genes with high GC3. Rather surprisingly there does not appear to be any relationship between the degree of compaction and the base composition in either species (Fig. 3.1C), or the amount of interspecies difference in base composition (Fig. 3.1D). The two most severely compacted genes have similar GC3 content in *Fugu* and human (Fig. 3.1D).

3.3.2 Synteny conservation between *Fugu* and human

Synteny conservation between two species can be measured in two directions. We can ask "what proportion of genes that are syntenic in species A are also syntenic in species B?", or conversely, "what proportion of genes that are syntenic in B are also syntenic in A?". These are two distinct quantities, as becomes obvious if one considers a hypothetical case where one of the species has only a single chromosome. The only syntenic genes that are known in *Fugu* are those that have been sequenced on the same clone; there are no large-scale maps of chromosomes. Therefore, we measured *Fugu*/human synteny conservation in terms of the proportion of neighbouring genes (from the same clone or GenBank entry) in *Fugu* that are syntenic in human. We also applied various limits to the physical distance permitted between the syntenic genes in human. Two separate datasets were analysed, as described below.

3.3.2.1 Synteny conservation - "cosmid skimming" data

The HGMP-RC *Fugu* landmark mapping project (Elgar *et al.*, 1996; Elgar, 1996; Elgar *et al.*, 1999) surveyed the *Fugu* genome by limited sequencing ("skimming") of a large number of genomic cosmid clones. Sets of shotgun sequence reads for 850 randomly chosen cosmids are publicly available from their website (`http://fugu.hgmp.mrc.ac.uk/`). The data consist of 40,303 sequence reads with an average of 47 reads per cosmid and 486 bp per read.

Each read is assumed to contain no more than one gene.

Because these sequences are short and largely unannotated, we compared them to human data from SwissProt rather than GeneMap '98 (which contains a large number of EST sequences). Cytogenetic map positions for 3963 of the 5406 human proteins in SwissProt were obtained by following links to OMIM. All 5406 proteins were searched against the *Fugu* cosmid database using TBLASTN (Altschul *et al.*, 1990). Putative orthologous relationships were identified as described in Materials and Methods.

A *Fugu* cosmid was considered "informative" (*i.e.,* it appeared to contain more than one gene, and so contained linkage information) if two different sequence reads hit two different mapped human sequences which do not themselves show significant sequence identity to one another. We identified 48 informative cosmids, containing 58 links between nearby *Fugu* genes (Table 3.1). For 26 of these links (45%), the human homologues are on the same chromosome (*i.e.* synteny was conserved).

Table 3.1: *Fugu* skimmed cosmids containing homologues of at least two mapped human SwissProt sequences

Cosmid	Syntenic links[a]		Subclone	Swissprot name[b]	Description	OMIM location
	+	-				
002I16	0	1	bB8	CGB1	G2/mitotic specific cyclin B1	5q12
			bC1	UBCG	Ubiquitin-conjugating enzyme E2 G1	1q42
003A22	0	1	aD2	LCFD	Long-chain fatty-acid CoA ligase 4	Xq22.3
			aE9	AP19	Clathrin coat assembly protein	Chr.7
018N05	0	1	cB3	COMT	Catechol O-methyltransferase	22q11.2
			cB7	RYK	Tyr-protein kinase RYK	3q22
020M06	1	1	bF2	F16P	Fructose-1-6-bisphosphate	9q22.2-q22.3
			bG9	GAS1	Growth arrest specific protein	9q21.3-q22.1
			aE1	LMG2	Laminin γ-2 chain	1q25-q31
030J22	2	0	aF4	TRFE	Serotransferrin	3q21
			aF7	IF4G	Translation initiation factor 4 G	3q27
			aG1	CLC2	Chloride channel protein 2	3q26-qter
032I12	0	1	aD1	PA2Y	Cytosolic phospholipase A2	1q25

Cosmid	Syntenic links[a] +	Syntenic links[a] -	Subclone	Swissprot name[b]	Description	OMIM location
			aE3	TSP1	Thrombospondin 1	15q15
035P08	1	0	aC2	KPT1	Ser/Thr protein kinase PCTAIRE-1	Xp11.3-p11.23
			aD5	HFC1	Host cell factor C1	Xq28
042H13	1	2	aE6	PIGF	Phosphatidylinositol-glycan synthase F	2p21-p16
			bA4	GCH1	GTP cyclohydrolase I	14q22.1-q22.2
			bD10	MSH2	DNA mismatch repair protein	2p22-p21
			bF8	CIKA	Voltage gated K channel KV21	20q13.2
050M16	1	0	bC5	CYCH	Cyclin H	5q13.3-q14
			bG2	GTPA	GTPase-activating protein (GAP)	5q13.3
055I13	0	1	bD9	A2MG	α-2-macroglobulin	12p13.3-p12.3
			bE2	ECH1	δ3,5-δ2,4-dienoyl-CoA isomerase	19q13
057B20	0	1	aC11	SC14	sec-14-like	17q25.1-q25.2
			aH1	GNT5	Glucoseaminyltransferase V	2q21
059A13	0	1	aD6	VLCS	Very long-chain acyl-CoA synthetase	15q21.2
			aE6	AMBP	AMBP protein	9q32-q33
060I09	0	1	aF1	ITA1	Integrin α-1	Chr.5
			aG3	ROK	Het. nuclear ribonucleoprotein K	9q21.32-q21.33
063J19	1	1	aA5	AGAL	α galactosidase A	Xq22
			aD12	RL44	60S rpL44	Chr.14
			aH4	DDP	Dystonia protein	Xq22
068B10	1	0	aA9	MET	Hepatocyte growth factor receptor	7q31
			aC8	MGR8	Metabotrophic glutamate receptor 8	7q31.3-q32.1
077E20	1	1	bB7	COGT	Matrix metalloproteinase-14	14q11-q12
			cC4	PKD2	Polycystin 2	4q21-q23
			cC5	AF4	AF-4 protein	4q21
081G09	1	0	aD12	CIK4	Voltage gated K channel protein	11q13.4-q14.1
			aF6	EAT2	Excitatory amino acid transporter 2	11p13-p12
082H05	0	1	aG5	KMLS	Myosin light-chain kinase	3cen-q21
			aH4	NED4	NEDD-4 protein	15q
082L03	1	0	aD12	MPCP	Mitochondrial $P0_4$ carrier	12q23
			aF10	THPA	Thymopoietin α	12q22
086H03	1	0	bC4	DOC2	Differentially expressed protein 2	5p13
			cE8	CO9	Complement component C9	5p13
096F11	0	1	aA7	WN11	WNT-11	11q13.5

Cosmid	Syntenic links[a] +	Syntenic links[a] -	Subclone	Swissprot name[b]	Description	OMIM location
			bC7	ACHD	Acetylcholine receptor δ chain	2q33-q34
103N12	1	0	aB9	RO52	Ro protein, 52 kD	11p15.5
			bA6	COGM	Macrophage meltalloelastase	11q22.2-q22.3
104N10	1	0	aD3	FER	FER Tyr protein kinase	5q21-q22
			bA12	MAN2	α-mannosidase II	5q21-q22/20q11.2
107H09	0	1	aF11	RS12	40S rpS12	6q
			aG6	EYA1	eyes absent homologue 1	8q13.3
107N05	0	2	aG10	BCAM	branched-chain aminotransferase	19q13
			aH4	GRN	Granulins	Chr.17
			aF6	EAT2	Excitatory amino acid receptor 2	11p13-p12
110I12	0	1	dA4	PAK1	Ser/Thr protein kinase PAK-α	11q13-q14
			dD3	PET1	Oligopeptide transporter	13q33-q34
114M17	1	0	bB8	IHBA	Inhibin β a chain	7p15-p13
			bC3	EGFR	Epidermal growth factor receptor	7p12.3-p12.1
116E05	0	1	aB3	GNT2	Acetylglucoseaminyl-transferase	14q21
			aE6	HSA9	Heat shock protein 90-a	1q21.2-q22
118A15	0	1	cC8	PERT	Thyroid peroxidase	2p25
			cG3	VMD2	Bestrophin	11q13
122O20	0	1	cA4	CASR	Extracellular Ca-sensing receptor	3q13.3-q21
			cD1	CTR2	Low affinity cationic amino acid transporter	8p22
123I02	0	1	aC11	BTG1	B-cell translocation 1	12q22
			aE5	TEF	Thyrotroph embryonic factor	22q13
128G19	1	0	aD4	FMO1	Dimethylaniline monooxygenase	1q23-q25
			aF3	TRK3	Receptor protein Tyr kinase TKT	1q12-qter
137O18	0	1	aE10	CYA1	Adenylate cyclase, type I	7p13-p12
			bA4	BNA1	Amiloride-sensing brain Na^+ channel	17q11.2-q12
141H19	1	0	aH10	LDHH	L-lactate dehydrogenase H chain	12p12.2-p12.1
			aH9	UGS2	Glycogen synthetase	12p12.2
143P11	1	0	aB6	ANK1	Ankyrin R	8p11.2
			aD6	NFM	Neurofilament triplet M protein	8p21
145K17	0	1	bF3	RHM1	Rhombotin-1	11p15
			cB1	AHR	AH receptor	7p15
147P16	2	0	aD1	DDP	Deafness dystonia protein	Xq22
			aF9	BTK	Tyr-protein kinase BTK	Xq21.3-q22
			aG7	GRA2	Gly receptor α-2 chain	Xp22.1-p21.2

Cosmid	Syntenic links[a] +	-	Subclone	Swissprot name[b]	Description	OMIM location
155N11	1	1	bE7	SYB2	Synaptobrevin 2	17pter-p12
			bH3	MPP2	Maguk P55 subfamily member 2	17q12-q21
			aD7	UTY	Ubiquitously transcribed TPR on Y	Yq11
156P04	1	0	aH2	RO52	Ro protein, 52 kD	11p15.5
			hC8	Z195	Zinc finger protein 195	11p15.5
157C15	0	1	aA3	RIR2	Ribonucleoside reductase M2	2p25-p24
			aD10	RL30	60S rpL30	Chr.8
159J19	2	0	aB1	MPK4	MAP kinase kinase 4	17p11.2
			aD11	MYSP	Myosin H perinatal skeletal muscle	17p13.1
164B03	0	1	aD11	ISL1	Insulin gene enhancer protein ISL-1	5q
			aD4	ETFA	Electron transfer flavoprotein α	15q23-q25
165O08	0	1	aH7	UBA1	Ubiquitin-activating enzyme E1	Xp11.23
			bD10	DPOE	DNA polymerase epsilon, subunit A	12q24.3
171K15	0	1	bB10	DMK	Myotonin protein kinase	19q13.2-q13.3
			bB6	BMAL	Brain and muscle ARNT-like 1	11p15
174C18	1	1	aD11	G6PD	G6PD	Xq28
			bA1	CCB3	Ca^{2+} channel β-3	12q13
			bB11	CYA6	Adenylate cyclase type VI	12q12-q13
176J15	1	0	aA8	DESM	Desmin	2q35
			aC5	PTPN	Protein-Tyr phosphatase N	2q35-q36.1
192G14	0	1	aA2	ADG	γ-adaptin	16q23
			aA7	RFP	Zinc finger protein RFP	Chr.6
222J11	1	0	bE3	NTTA	Taurine transporter	3p25-q24
			bC4	ACTQ	Ca^{2+} transporting ATPase	3p62-p25
Totals:	26	32				

[a]The '+' column refers to conserved linkages between *Fugu* and human, and the '-' column refers to non-conserved linkages
[b]All Swissprot IDs are truncated, omitting '_HUMAN' from each one

The same *Fugu* Landmark Mapping Project data were recently analysed by Elgar *et al.* (1999). They reported that "three-quarters" of informative cosmids showed synteny to human. However, it is difficult to account for the differences between our results and theirs as they do not specify what stringency they imposed on the definition of orthology, nor do they indicate which cosmids displayed an orthologous relationship with which human

sequences. Perhaps the greatest discrepancy between these analyses is in the number of informative cosmids found (349 by Elgar *et al.* compared to 48 in this study). We expect that this difference is due to a greater stringency employed by us in the designation of orthologues (as described in Materials and Methods).

Figure 3.2 shows a comparison of the *Fugu* skimmed cosmids with the predicted proteins of the complete sequence of human chromosome 22 (Dunham *et al.*, 1999; Solovyev and Salamov, 1999). All *Fugu* cosmids from which two sequence reads had TBLASTN hits to proteins from human chromosome 22 are indicated. This small study is interesting because it includes physical positions of the human genes rather than the lower resolution cytogenetic positions used in the rest of this study. This analysis identified three possibly colinear regions of these two genomes (cosmids 147D08, 123I02, and 013A01), though the order and orientation of the sequence reads within each *Fugu* cosmid is not known. We also observed in the case of three other cosmids (104N10, 156P04, and 159J19) that genes that are apparently close or adjacent in *Fugu* (*i.e.,* are from the same cosmid) are separated by long distances in the human genome; for example, the human homologues of genes on *Fugu* cosmid 159J19 are separated by 5.6 Mb. A similar phenomenon was observed in the analysis of *Fugu* complete genomic sequences (see below).

3.3.2.2 Synteny conservation - complete *Fugu* genomic sequences

We examined the GenBank annotation of all *Fugu* sequences greater than 5 kb long to look for sequences that coded for two or more proteins. The 21 GenBank entries that fit this criterion (Table 3.2) total just under 0.9 Mb and encode 91 annotated proteins (some putative). Genes from the same GenBank entry have a known linkage relationship in the *Fugu* genome because they were sequenced contiguously.

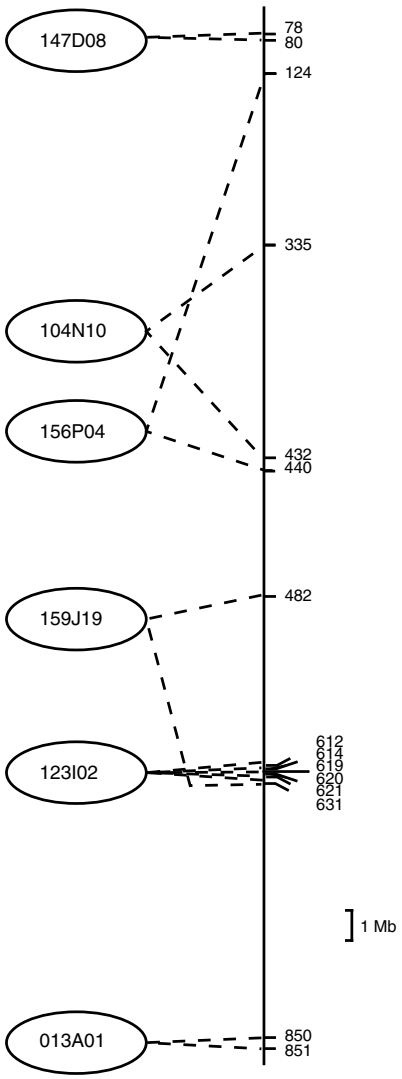


Figure 3.2: *Fugu* cosmids showing conserved synteny with human chromosome 22 predicted proteins. The vertical line represents HSA22 with genes numbered according to their order on the chromosome and distances scaled to kb position. *Fugu* cosmid names (Elgar *et al.*, 1999) are written in the ovals on the left. The order of genes within each *Fugu* cosmid is unknown because the data come from a random subclone sequencing (skimming) approach. Dashed lines indicate sequence similarities.

Table 3.2: Details of completely sequences *Fugu* cosmids used in this analysis

Accession No.	Base pairs	Genes included	Reference
af056116	148,640	ACVR1B, ALR, fhh, Ikaros-like, wnt1, wnt10b, ARF3, erbB3, PAS1, rpL41, LRP1	Gellner and Brenner, 1999
af094327	69,056	SCML2, STK9, XLRS1, PPEF-1, KELCH2, KELCH1, PHKA2, AP19, U2AF1-RS2	Brunner *et al.*, 1999
u90880	61,901	RNA-H, CAB3B, Adenyl Cyclase-VI, G6PD, LG3P, Na+ channel 2	Riboldi Tunnicliffe, G.R., *et al.* unpublished
af016494	66,729	GABRB, P55, VAMP-1, PCOLCE, GRMP	Riboldi Tunnicliffe, G.R., *et al.* unpublished
af026198	63,155	L1-CAM, SMC1, CCA1	Riboldi Tunnicliffe, G.R., *et al.* unpublished
af083221	43,373	Neurotransmitter receptors, YDR140w homologue, glycinamide ribonucleotide transformylase	Reboul *et al.*, 1999
aj010317	39,410	GRM-7, TRIP, Sand, PRGFR3	Cottage *et al.*, 1999
y15170	10,753	EST00098 homologue, SURF2, SURF4, ASS	Armes *et al.*, 1997
aj010348	39,850	UBE1-like, PRGFR2, calmodulin binding protein kinase	Cottage *et al.*, 1999
al021880	37,170	IGFII, TH, NAP2	Chen, E., *et al.* unpublished
al021531	45,565	WT, Reticulocalbin, PAX6	Miles *et al.*, 1998
z93780	34,807	CPS3, MLC, MAP2	Schofield *et al.*, 1997
u92572	20,919	HOXC-9, HOXC-8, HOXC-6	Aparicio *et al.*, 1997
y15171	8,902	rpL7a (SURF3), SURF1, SURF6	Armes *et al.*, 1997
af013614	55,892	TSC2, PKD1	Sandford *et al.*, 1997
af022814	37,400	Zinc finger transcription factor, HMOX1	Gottgens *et al.*, 1998
af030881	5,645	gag, pol	Poulter and Butler, 1998
aj010316	10,959	Cav-2, Cav-1	Cottage *et al.*, 1999
u63926	23,196	PDGFR-beta, CSF1R	How *et al.*, 1996
u92573	13,583	HOXA-10, HOXA-9	Aparicio *et al.*, 1997

Cosmids are listed in order of decreasing number of annotated proteins. The list of annotated proteins for each cosmid does not include putative proteins with no known human homologues at the time of submission to the sequence database

The proteins encoded by these *Fugu* sequences were compared using TBLASTN to the database of human nucleotide sequences whose map positions are known in GeneMap '98 (Deloukas *et al.*, 1998). For some of the *Fugu* sequences our results confirm previously published analyses (Sandford *et al.*, 1996; Aparicio *et al.*, 1997; Armes *et al.*, 1997; Schofield *et al.*, 1997; Miles *et al.*, 1998; Brunner *et al.*, 1999; Gellner and Brenner, 1999; Reboul *et al.*, 1999).

The results were examined to look for candidate conserved syntenous regions between human and *Fugu*. This was facilitated by a new method for displaying the relative positions of the homologues in the two species. In many cases, such as in the example shown in Figure 3.3, there was more than one candidate human chromosomal region for conserved synteny. In Figure 3.3 the *Fugu* sequence (AF056116) appears to have conserved synteny with human chromosome 12 by virtue of having several top scoring BLAST hits to human genes that map close together on that chromosome, largely as described by Gellner and Brenner (1999). What is interesting is that regions on chromosomes 7, 17, and 2 also show synteny with this *Fugu* sequence (including matches to *Fugu* proteins not having homologues on chromosome 12: genes 3, 4, 6, and 14; Fig. 3.3). These are the human chromosomes that contain the HOX clusters and this indicates that the similarity of these human chromosomes to each other extends beyond those clusters, as has been suggested by others (Ruddle *et al.*, 1994).

To examine synteny conservation in a quantitative way, instead of simply the presence or absence of genes on the same chromosome, we calculated the proportion of *Fugu* close neighbours (genes from the same GenBank entry) whose homologues were within a specified distance x of each other in human. We use the term "proximity conservation" to denote this property of genes remaining within a specified distance of each other (regardless of gene order). To allow for the uneven distribution of genes in the human

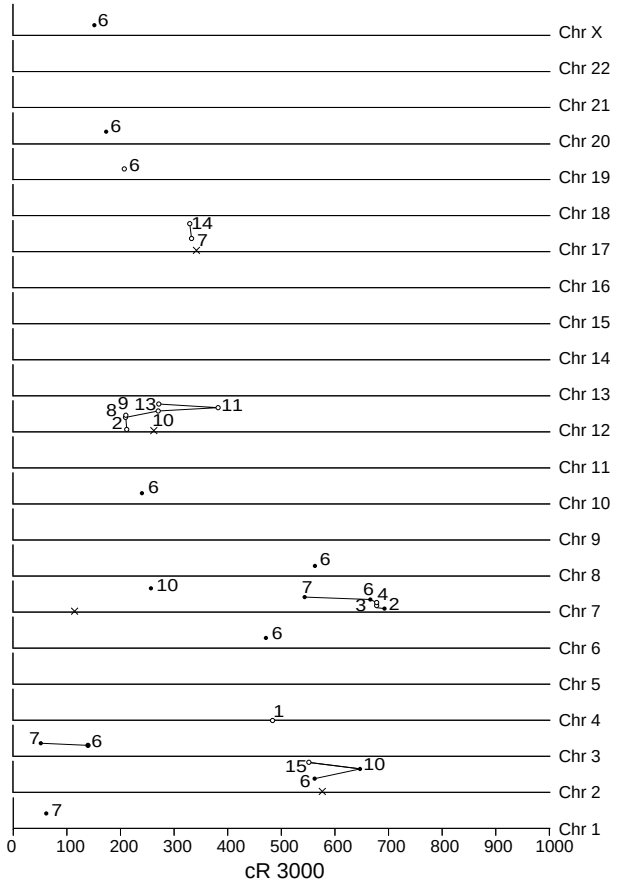


Figure 3.3: Graphical representation of the results of the TBLASTN search of the proteins from *Fugu* sequence AF056116 (Gellner and Brenner, 1999) against a database of mapped human sequences (GeneMap '98). The relative positions of the best hits of each of the 15 annotated *Fugu* proteins from this cosmid are shown for each chromosome in turn. The horizontal axis represents position (measured in centiRads) on the human chromosome in question, and each vertical axis represents the relative order (1-15) of the *Fugu* genes on the *Fugu* cosmid. White dots designate the top-scoring TBLASTN hit for each *Fugu* protein; black dots indicate weaker hits (that are within 10^5 of the strongest hit). The genes are in the following order in Fugu: 1, ACVR1B; 2, ALR; 3, fhh; 4, R05D3.2-like protein; 5, 138E3.2-like protein; 6 Ikaros-like; 7, wnt1; 8, wnt10b; 9, ARF3; 10, erbB3; 11, PAS1; 12, rpl41; 13, 178O23.1-like protein; 14, diaphonous-like protein; 15, LRP1. In addition to the matches shown here (based on data in GeneMap '98), genes 1, 7, 12, and 15 also have homologues on chromosome 12q13 (Kenmochi *et al.*, 1998; Gellner and Brenner, 1999). The positions of Hox clusters A, B, C, D are represented by crosses on chromosomes 7, 17, 12 and 2 respectively.

genome, the distance x was expressed in terms of the estimated number of intervening genes instead of in the physical map units (cR) that were used in GeneMap '98 (Deloukas *et al.*, 1998). The number of intervening genes was estimated from GeneMap '98 by counting the number of intervening genes appearing on the map between the genes of interest and scaling by a factor of 80000/30000 to allow for unsequenced genes. This allows for gene density variation within and between chromosomes. Where more than one human sequence had been assigned to the same map position by Deloukas *et al.* (1998), these sequences were arbitrarily assigned an order.

The results are summarised in Table 3.3. Only 18% of *Fugu* neighbours have sequenced human homologues that are within ten genes of one another. This increases to 39% within a limit of 200 intervening genes, and to a maximum of 47% within a limit of 4000 intervening genes (this is effectively no limit, because it is approximately the size of a chromosome). The last value is similar to the synteny estimate from Table 3.1 (which has no limit on the intervening distance).

3.3.3 Computer simulation of genomic rearrangement

We used computer simulations to try to relate the observed level of proximity conservation to the number of genomic rearrangements that have occurred since the divergence of *Fugu* and human. The simulation started with a linear array of 80,000 genes, representing the current gene order in *Fugu*. Varying numbers of rearrangements were made in a copy of this genome (representing human) by randomly choosing two endpoints in the genome and inverting the segment in-between. To reflect the missing data in the human map, randomly chosen genes were marked 'unmapped' until only 30,000 remained (the number of genes in Deloukas *et al.*, 1998). Pairs of genes that are neighbours in *Fugu* were then examined to see if they are neighbours in human, similar to the method of analysis in Tables 3.1 and 3.3.

Table 3.3: Observed levels of synteny conservation between completely sequenced *Fugu* cosmids and human

Fugu Accession No.	Annotated proteins	Proteins with human BLAST hits P<10^{-15}	Maximum possible links	Human Chromosome[a]	Number of links on human chromosome, at different value of x intervening genes[b]						
					5	10	20	50	200	1000	4000
AF056116	15	13	12	2	0	1	1	1	1	1	2
				3	0	0	0	0	0	1	1
				7	0	0	0	1	2	4	5
				12*	1	1	2	2	3	4	5
				17	0	0	0	0	1	1	1
AF094327	9	9	8	5	0	0	0	0	0	1	1
				X*	1	2	4	4	4	4	4
U90880	9	6	5	2	0	0	0	0	0	0	1
				20*	0	0	0	0	1	2	2
U72484	6	6	5	12	0	0	0	0	2	2	2
AF016494	5	4	3		0	0	0	0	0	0	0
AF026198	5	3	2		0	0	0	0	0	0	0
AF083221	4	3	2		0	0	0	0	0	0	0
AJ010317	4	3	2	3	0	0	0	0	2	2	2
Y15170	4	2	1	9	1	1	1	1	1	1	1
AJ010348	3	3	2	3	0	0	0	0	0	1	1
AL021880	3	3	2	11*	0	0	0	0	1	1	1
				12	0	0	0	0	0	0	1
AL021531	3	3	2	11	0	0	0	1	1	1	1
Z93780	3	3	2	2	0	0	0	1	1	1	1
U92572	3	3	2	2	0	1	1	1	1	1	1
Y15171	3	3	2	9	1	2	2	2	2	2	2
AF013614	2	2	1	16	0	0	0	0	0	1	1
AF030881	2	1	0		0	0	0	0	0	0	0
AJ010316	2	2	1	7	1	1	1	1	1	1	1
U63926	2	2	1	4	0	1	1	1	1	1	1
U92573	2	2	1	7	1	1	1	1	1	1	1
Totals:			57		6	10	13	15	22	26	27
Proximity conservation (%)					11	18	23	26	39	46	47

[a] In cases where there is more than one candidate human chromosome, * marks the human chromosome with the highest number of top scoring BLAST hits, which was used in the calculation of the totals at the bottom. Some of these relationships to human chromosomes have previously been described by the original authors (Sandford *et al.*, 1996; Aparicio *et al.*, 1997; Armes *et al.*, 1997; Schofield *et al.*, 1997; Miles *et al.*, 1998; Brunner *et al.*, 1999; Gellner and Brenner, 1999; Reboul *et al.*, 1999).

[b] The quantity x is the largest allowed distance (in genes) between one of the human homologues and its nearest neighbour in the syntenous group. For the parts of the genome studied here the intervals of $x = 5$, 10, 20, 50, 200, 1000 and 4000 genes correspond to average physical distances of 0.50, 1.27, 2.81, 7.05, 29.14, 144.01, and 494.04 cR respectively.

To make the simulation more realistic, we modelled the presence of gene families. Because more than half of all human genes are still not included in the human gene map, there is a real possibility that if the human orthologue of a *Fugu* gene is not mapped, the *Fugu* gene would mistakenly be paired with a mapped human paralogue instead. This could reduce the estimated level of synteny conservation. Simulating this problem requires knowledge of the distribution of gene family sizes, which we addressed in two ways. First, we used the distribution of the numbers of human BLAST hits to the *Fugu* proteins considered in Table 3.2 (plus annotated putative proteins, totalling 91) as an approximation of the distribution of family sizes. Second, we used the distribution calculated by Imanishi *et al.* (1997) from an all-against-all FASTA comparison of human proteins translated from mapped entries in DDBJ/EMBL/GenBank. In both cases the family sizes were scaled by a factor of 8/3 to account for unsequenced and unmapped genes. The latter (within-genome) method has the advantage that all the hits to a protein represent paralogues, whereas with between-genome comparisons the orthologues must be identified and removed before gene families can be examined. The results from the two methods were similar and only those using the *Fugu* data are presented here.

Paralogous gene families were randomly assigned among the 80,000 genes in the simulated genome, according to the distributions described above. This process resulted in each simulated *Fugu* gene having one human orthologue, and possibly also a list of human paralogues, analogous to a list of BLAST hits. Some of the orthologues and paralogues could be 'unmapped'. Linkage conservation was measured by looking for the human homologues of 1000 pairs of adjacent *Fugu* genes, chosen at random. If the human orthologue of one (or both) of the *Fugu* genes in the pair was 'unmapped', a mapped paralogue from the list was used instead where possible. The extent of linkage conservation in human was then calculated, allowing various intervals

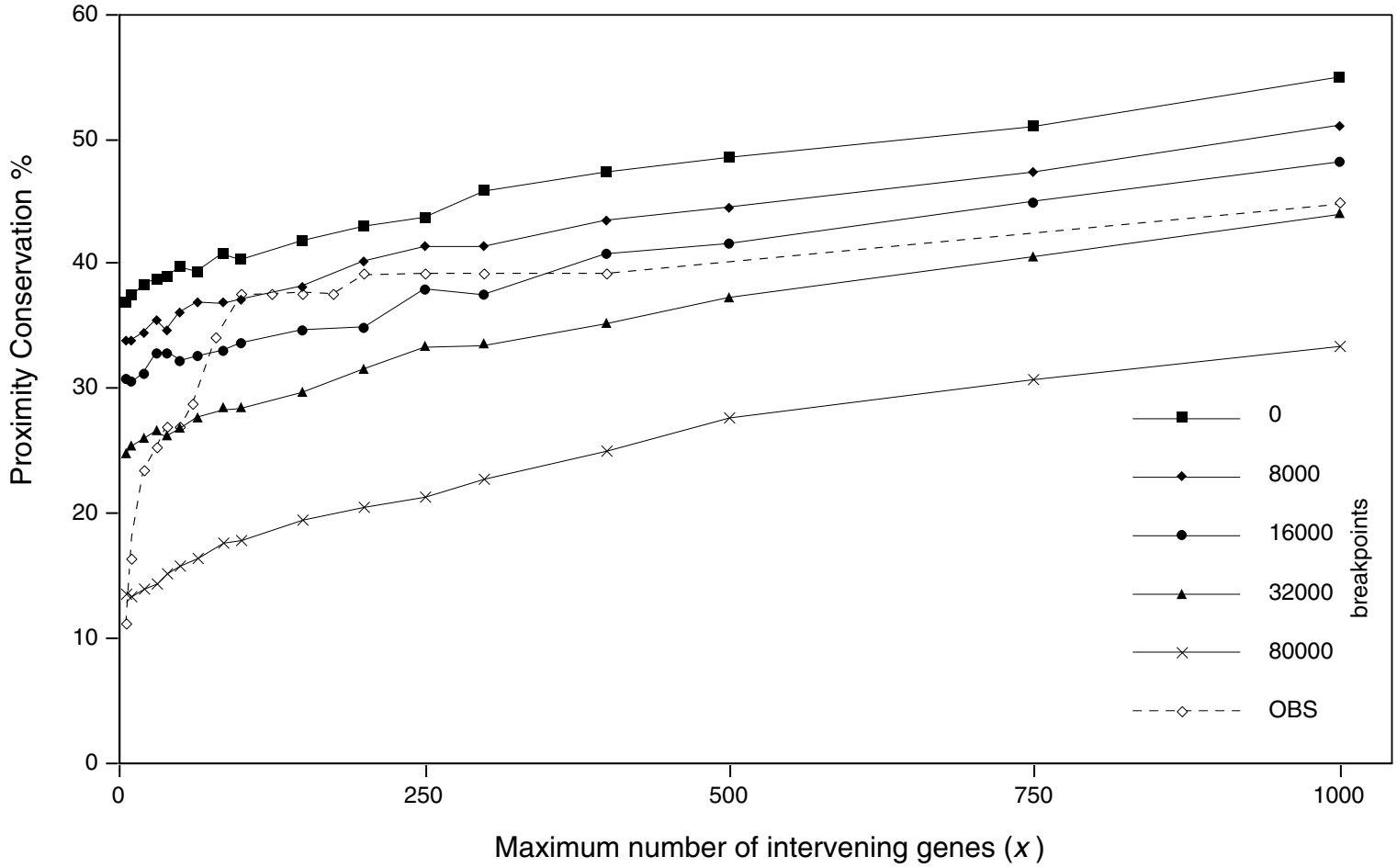


Figure 3.4: Extent of proximity conservation in real and simulated datasets. Proximity was measured allowing different gene distances between the human homologues of pairs of linked *Fugu* genes. The line marked "OBS" graphs the observed data (Table 3.3). Average results for 30 computer simulations with 0, 8000, 16000, 32,000 and 80,000 rearrangement breakpoints are shown (8000 breakpoints = 4000 rearrangements). The x-axis is the limit used for the distance permitted between two human genes that are homologues of two *Fugu* neighbours, expressed in terms of the estimated number of intervening genes on the chromosome.

between the human homologues. The simulation was run 30 times, looking at 1000 pairs of genes each time, with the average results shown in Figure 3.4.

The most striking feature in the simulation results is that the presence of paralogues in an incompletely-sequenced genome has a substantial effect on the measured extent of linkage conservation. If there have been no genomic rearrangements (top curve in Fig. 3.4), gene order conservation (and thus proximity conservation) should be 100%. However, the measured level is only 37%, because for many gene pairs one or both of the human orthologues is 'unmapped' and a mapped paralogue at some other location in the genome has been used instead. This makes many linkages appear broken artefactually. Our measures of proximity conservation in the real data may also be underestimated to a similar degree (see Discussion). When the observed data from the fully-sequenced *Fugu* cosmids (Table 3.3) is plotted on the same axes, its initial slope is much greater than for the simulations (Fig. 3.4). Possible reasons for this are discussed below. At large window sizes the line is approximately the same as the simulations with 8,000-32,000 breakpoints.

3.4 Discussion

Although we confirmed the compaction of *Fugu* genes with respect to their human orthologues, we did not observe any strong relationship between gene compaction and the synonymous G+C content of the gene in either species. This may be an artefact of the sample analysed, or it may indicate true randomness in the compaction of the *Fugu* genome. There is an inverse relationship between the average compaction of the genes in each GC3 content category and their average GC3 content, which is consistent with expectations based on the lengths of genes in G+C rich isochores in vertebrate genomes (Duret *et al.*, 1995). However, this relationship is not strong enough to be predictive for individual genes.

The incomplete nature of the human genome data, and the uncertainty regarding whether homologues found in BLAST searches are orthologues or paralogues, reduces our power to examine synteny conservation between *Fugu* and human. The measured proximity conservation depends not only on whether the genes remain close or not, but also on whether they are mapped and sequenced, and if there are paralogues of these genes in the mapped data. The simulations (Fig. 3.4) suggest that the combination of incomplete sequence sets and the presence of gene families may cause the level of proximity conservation to be underestimated substantially, perhaps twofold.

There is an obvious discrepancy between the slope of the graph of proximity conservation in real data from fully sequenced cosmids, as compared to the results from computer simulations (Fig. 3.4). The observed proximity conservation rises steeply to 37% at a window size of 100 intervening genes, and then plateaus to a shape more like the simulated data. This suggests that the assumptions underlying the simulation are incorrect in some way.

The steep rise may be attributable to three primary factors. One possibility is that the real data is not a random sample of genes from the two organisms. A bias may result from *Fugu's* role as a model vertebrate genome inevitably influencing the selection of cosmids for complete sequencing. Cosmids with hypothesised synteny conservation with mammalian genomes may have been chosen preferentially. At least five of the *Fugu* complete sequences used had known synteny conservation with human chromosomes prior to sequencing (Aparicio *et al.*, 1997; Armes *et al.*, 1997; Sandford *et al.*, 1997).

Second, lack of resolution and incomplete data in GeneMap '98 data may affect the results. The arbitrary ordering of human genes that lie in the same radiation hybrid map interval could inflate apparent distances in human,

though this effect is unlikely to be significant because the average number of genes per interval in the GeneMap '98 data used here is only 1.98. At least one distance in Table 3.3 has been overestimated due to missing data in GeneMap '98. This occurs with the genes TSC2 and PKD1 (*Fugu* accession number AF013614) which are neighbours in both species (Sandford *et al.*, 1996). However, PKD1 is not present in the map and instead our method identified a PKD1-like sequence elsewhere on chromosome 16 (Loftus *et al.*, 1999).

A third factor may be that our model of rearrangements is too simple. Our model assumed a random distribution of breakpoints throughout the genome, but comparative analysis of the human and mouse maps has shown that, although inter-chromosomal rearrangements seem to have random endpoints, the number of intra-chromosomal rearrangements is more than expected at random (Ehrlich *et al.*, 1997; Nadeau and Sankoff, 1998). The steep incline at the beginning of the graph may indicate a high frequency of small inversions or other small intra-chromosomal rearrangements as has been observed in yeast species (Seoighe *et al.*, 2000). Inversions of small segments of chromosome would disrupt gene adjacencies while preserving gene vicinities. This has been proposed by Gilley and Fried (1999) who noticed that some genes that are adjacent in *Fugu* are 2-4 Mb apart in human. Further examples from our study include wnt10b, ARF3, and erbB3. These genes are adjacent in *Fugu* (Gellner and Brenner, 1999). In human wnt10b and ARF3 are adjacent but erbB3 is separated from them by an estimated distance of 603 genes (226 mapped GenBank sequences scaled by 8/3 to allow for missing data) or 7.5 Mb (estimated from the map distance of 31 cR; chromosome 12 has an average of 234 kb/cR (estimated from the map distance of 31 cR, chromosome 12 has an average of 234 kb/cR; Gyapay *et al.*, 1996).

It is likely that the initial portions of the simulations in Figure 3.4 are

not directly comparable with the observed data. However, as the window size gets larger the graph lines are approximately parallel to the plot of the observed data. From these an estimate of the extent of rearrangement since the divergence of these two lineages 400 million years ago is 8,000-32,000 breakpoints (i.e., 4,000-16,000 reciprocal translocations or inversions). This is higher than expected from comparisons of the human and mouse genomes which diverged 100 million years ago and have only an estimated 180 rearrangements (Nadeau and Taylor, 1984; Nadeau and Sankoff, 1998). Adjusting our simulations to incorporate a bias towards small rearrangements would only increase the estimated number of rearrangements since the *Fugu*/human divergence, making the discrepancy in rates even greater.

Another possible shortcoming of our analysis is the presence of short ESTs (which are not necessarily coding sequence) in the human DNA database used here, resulting in an overestimate of the frequency with which we can expect to find orthologues in this dataset from an amino acid level search. However, this is unlikely to have a great effect on the results because we found that 78% of a random sample of over 500 human proteins submitted to TREMBL after we downloaded GeneMap '98 were represented in the database. The gene family data is also likely to be oversimplified, as it is based on results from only 91 *Fugu* proteins. The Imanishi *et al.* (1997) data is from a larger set of proteins but is not as easy to relate to the human dataset used in this analysis.

Because we have approached the question of synteny conservation from the perspective of known gene adjacencies in *Fugu*, the proposed genome duplication in the bony fish lineage (Amores *et al.*, 1998) followed by differential gene loss should not influence the results. If genes in the ancestral genome were ordered $ABCD$ and this was duplicated in the fish lineage, differential gene loss could result in paralogous chromosomes, one bearing AC and another bearing BD. If synteny of these genes had not been

disturbed then the human genome would still contain the four genes arranged $ABCD$. If we were counting conservation of human linkages in *Fugu* then we might plausibly have selected genes A and B for analysis and found that they are not syntenous in *Fugu*, an artefact of gene loss, rather than genome rearrangement. However, as we are starting from the complementary viewpoint (given known relationships in *Fugu*), the only possible questions are "Are A and C syntenous in the human genome?" and "Are B and D syntenous in the human genome?", which is true in both cases. It is, however, possible that differential gene loss (after the genome duplication) in the *Fugu* lineage has contributed to the reduction of some intergenic distances as compared to human (*e.g.*, the distance from A to C in the hypothetical example). This may also contribute to the steep initial slope seen in Figure 3.4. One example of apparent differential gene loss may already have been discovered in the case of the genes IGF2 and TH (insulin-like growth factor and tyrosine hydroxylase) which are adjacent in *Fugu* but separated by one intervening gene (insulin) in human (E. Chen *et al.* unpublished, GenBank accession number AL021880; Lucassen *et al.*, 1993). Patterns of gene loss and gene order evolution should become clearer when more long homologous sequences from these species become available.

Chapter 4

Extensive genomic duplication during early vertebrate evolution

4.1 Introduction

The recent arrival of the draft human genome sequence in a database near you was expected to open the door to a(nother) new age of molecular biology. In terms of molecular evolution, the burden of expectation lay on the origins of vertebrate complexity. One theory proposes that this complexity originated by genome duplication at the base of the vertebrate lineage (Ohno, 1970). Opinions on the contribution of ancient genome duplication(s) to the vertebrate genome range from highly skeptical (Hughes, 1999b; Martin, 1999; Hughes *et al.*, 2001) to highly credent (Spring, 1997; Holland, 1999; Wang and Gu, 2000).

Here we examine the data from the International Human Genome Sequencing Consortium (Lander *et al.*, 2001) for evidence of genome duplication in an early vertebrate genome, by a combination of map-based and phylogeny-based methods. This work was done in collaboration with Karsten

Hokamp in our laboratory.

4.1.1 Formalising the problem

It could been argued that it is easy (and tempting) to fit any data to the genome duplication hypothesis as it can accommodate a seemingly arbitrary amount of gene loss and genome rearrangement, thus rendering the modern genome unrecognisable as a paleopolyploid. This may come, in large part, from the fact that there is little agreement on what a paleopolyploid genome should look like. In fact, a genome duplication is not just a genome duplication, it is genome duplication imposed on a background of other kinds of duplication, be they chromosomal, segmental, or single gene duplications. Not enough is known about the frequency and extent of sub-genomic duplications to be able to effectively exclude them from any analysis, so we can only hope that any evidence for a whole genome duplication would be louder than the background noise.

The first step in analysing any problem is to define the null hypothesis, that which you believe unless an alternative hypothesis is shown to be true. Hughes argued that the null hypothesis should be the hypothesis of no effect, that of no genome duplication (Hughes, 1999a). The null hypothesis proposed by Hughes explains the presence of segments of chromosome with some paralogous gene content in the genome (paralogous regions) by selection for preferred translocation events (Hughes, 1998; Hughes *et al.*, 2001). There is no empirical evidence to support this hypothesis. By contrast there is strong evidence that paralogous regions may arise through block duplication events (*e.g.*, Wolfe and Shields, 1997). The discovery in recent years of polyploid origins of organisms from within the three eukaryotic kingdoms, fungi (yeast: Wolfe and Shields, 1997), plants (*Arabidopsis*: Vision *et al.*, 2000; *Arabidopsis* Genome Initiative, 2000), and animals (zebrafish: Amores *et al.*, 1998; Gates *et al.*, 1999) is indicative of a previously unknown ubiquity

of genome duplication, and thus of its credentials as a hypothesis to explain the presence of paralogous regions in the genome.

The whole genome duplication hypothesis predicts that gene duplication is spatially and temporally concerted, *i.e.,* including all genes simultaneously. The proportion of the genome that is related to some other part of the genome can be examined by map-based analysis of intra-genomic paralogues. Phylogeny-based methods including molecular clock analyses can be employed to ascertain the timing of gene duplications.

Various alternative hypotheses predict the existence of blocks of paralogous gene content within a genome (*e.g.,* segmental duplication, or selection for clustering of interacting genes), but only a genome duplication hypothesis predicts concerted timing of these events. The coalescence dates of duplicated loci will depend on the mechanism of tetraploidy (be it autotetraploidy, allotetraploidy, or segmental allotetraploidy) and the manner of restoration of disomic inheritance of loci (Gaut and Doebley, 1997, and as discussed in Chapter 1)

4.2 Materials and Methods

4.2.1 Sequences

The human sequence dataset was obtained from Ensembl version 1.00 (`http://www.ensembl.org`), and comprised 27,615 proteins encoded by 24,046 genes. Ensembl is a project based in the European Bioinformatics Institute (EBI) which aims to develop a system for automatic annotation of any genome sequence although it currently only provides protein predictions for the human genome. Ensembl supplies two classes of protein annotation, 'predicted' and 'confirmed'. Predicted proteins are based on GenScan predictions alone (Burge and Karlin, 1997), and confirmed proteins have supporting evidence from homology with proteins in other databases. Only

the confirmed protein dataset was used here.

Proteomes of the two invertebrate species *Drosophila melanogaster* (Adams *et al.*, 2000, 14,335 proteins) and *Cænorhabditis elegans* (*C. elegans* Sequencing Consortium, 1998, 19,835 proteins) were retrieved from GenBank release 123 (April 2001) and WormPep49 respectively. Alternative splice variants were removed from these datasets (retaining the longest isoform), leaving 13,473 fly proteins and 18,685 worm proteins.

4.2.2 Detection of paralogous regions in the human genome

A brute-force algorithm was developed by Karsten Hokamp which was used to detect paralogous regions within the human genome (described in detail in Hokamp, 2001).

BLASTP (Altschul *et al.*, 1997) was run on an all-against-all dataset of the complete set of Ensembl proteins using the BLOSUM45 substitution matrix, applying the SEG filter, and with an Expectation value (E-value) threshold of 1. Recent tandem repeats were excluded from further analysis by reducing all proteins within 30 genes distance and with BLASTP E-values less than 10^{-15} to a single entry by keeping only the longest protein. The resulting dataset is referred to as the 'collapsed' dataset. Adjacent predicted proteins with non-overlapping hits to the same protein on another chromosome were considered to be exons of the same gene.

Paralogous relationships of human genes were identified through implementing several post-processing steps on the BLASTP report: an upper limit on the BLAST E-value of 10^{-7} was imposed; the alignment of the maximal-scoring segment pair (MSP) was required to be at least 30% of the length of the longer protein; and only hits with an E-value within a range of 10^{20} of the top hit were considered. In addition *Drosophila* and *Cænorhabditis* proteins were included in the similarity search database to act

as a natural orthology threshold (any human proteins that are less similar than an invertebrate protein to the human query protein probably duplicated before the invertebrate-vertebrate lineage divergence and so are not relevant to the 2R hypothesis). High copy genes with greater than 20 hits passing all these criteria were excluded from further analysis.

The relative distribution of the hits of the collapsed protein set were examined to find blocks of similar gene content in diverse genomic locations. A block was built starting from an anchor of a pair of genes at different chromosomal locations. This was extended by including protein pairs on these chromosomes that were positioned no further than 30 genes distance from another protein included in the block. Paralogous regions detected by this algorithm can be browsed at the website `http://www.gen.tcd.ie/dup`.

4.2.3 Gene family construction

Mutual best hits between fly and worm were found using BLASTP (Altschul *et al.*, 1997) with the following parameters: BLOSUM45 matrix; SEG filter to mask simple repetitive sequences; and an E-value threshold of 10^{-20}. Additionally, to exclude similarity based only on short domains only protein pairs where the alignment of the maximal-scoring segment pair (MSP) in the BLAST analysis covered at least 30% of the longer protein were accepted. This search retrieved 2,802 fly/worm mutual best hit protein pairs. Tandem and other sorts of duplication of genes is known to be frequent within each of these invertebrate lineages (*C. elegans* Sequencing Consortium, 1998; Semple and Wolfe, 1999; Ashburner *et al.*, 1999) with the result that for some human genes there will be several invertebrate orthologues (as illustrated in Figure 1.6B reproduced from Venter *et al.*, 2001). Gene families where the invertebrate:human family size ratio is many:many, or many:one will not be excluded by the mutual best hits criterion, rather, this criterion reduces the chances of selecting a paralogue instead of the orthologue when

the orthologue has been lost. The fly sequences from this set were used as queries against the human protein set with alternative splice variants removed (retaining the longest isoform), and using the same BLAST protocol as for the fly *vs* worm comparison and enforcing the minimum alignment parameter of 30%.

Gene families are highly dependent on the method employed to define them. Some methods, such as single link clustering, where groups of related proteins can be joined together as a single family by the existence of a single relationship between a member of each group, may result in large networks of gene families which can be difficult to interpret and are prone to annotation artefacts. In this study, human gene families were defined conservatively as mutually exclusive sets of BLASTP hits, so that no protein can be a member of more than one family. Where two lists of hits were not mutually exclusive, both sets of hits were excluded from further analysis. This procedure retrieved 1808 human gene families and their invertebrate orthologues. 758 families had at least two members (see Table 4.4 on page 96). The BLASTP E-value threshold (10^{-20}) used in these searches was selected because it maximised the number of human gene families obtained (Figure 4.1). Less stringent cutoffs recovered fewer families because of the requirement that they should be non-overlapping.

4.2.4 Duplication date estimation

The 758 two-to-ten membered human gene families defined by this method were aligned with their fly and worm orthologues using T_COFFEE (Notredame *et al.*, 2000) with its default parameters. These alignments, and initial tree topologies generated by the PHYLIP program **protdist** with default parameters, were used to estimate the alpha parameter for a gamma distribution using the program **GAMMA** (Gu and Zhang, 1997). In the gamma distribution of evolutionary rate, the variance of the number

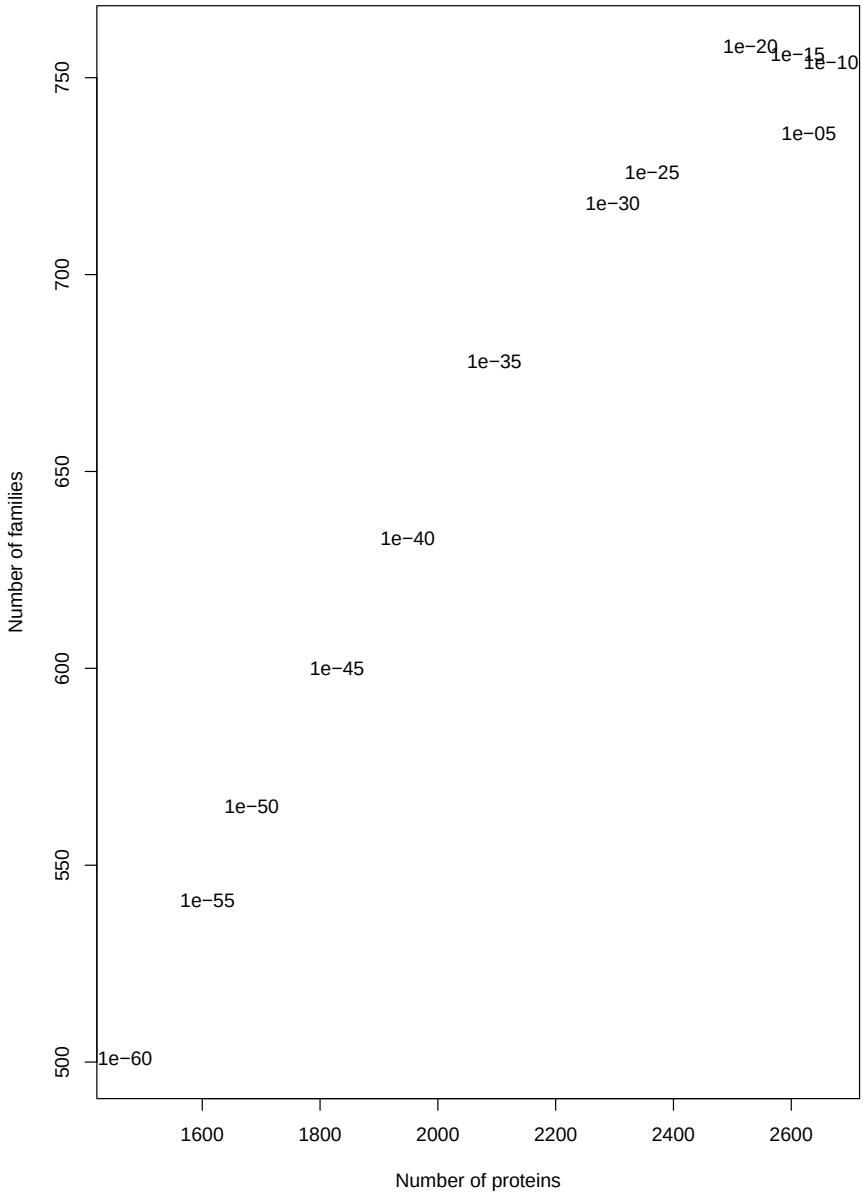


Figure 4.1: Effect of E-value threshold on the number of gene families recovered by the method described in section 4.2.3 on page 80. For a given E-value cutoff, the number of families with two or more members (y axis) and the number of proteins making up those families (x axis) are shown. The points on the graph are replaced by the E-value parameter used in each case.

of substitutions among sites should be greater than the mean. If this condition is not satisfied, alpha is not defined (infinity). This was the case for 154 of these gene families. Neighbour-joining trees (Saitou and Nei, 1987; Takezaki *et al.*, 1995) were drawn for the remaining 602 families using Gamma-corrected distances. Two families returned an unexplained 'format error'. 121 families where the fly and worm sequences did not group were excluded from further analysis because these family expansions must predate the arthropod-chordate divergence and we are specifically interested in vertebrate lineage gene family expansions. The two-cluster test (Takezaki *et al.*, 1995) for rate heterogeneity was applied to the remaining families to test for deviations from the molecular clock at 5% significance. Linearised trees (Takezaki *et al.*, 1995) were drawn for the 191 families that passed all these criteria. Gene duplication dates were calculated from each of the 191 linearised trees of 2-10 membered families by the method shown in Figure 4.5 on page 97. Nodes where the age was calculated to be zero were excluded from further analysis.

4.3 Results

4.3.1 Analysis of paralogous regions

In order to test the theory of genome duplication(s) at the base of the vertebrate lineage we examined the draft human genome sequence (Lander *et al.*, 2001) for the presence of blocks of homologous genes at different chromosomal locations, such as are predicted to result from the degeneration of the symmetry of a post-polyploidy genome. An algorithm was developed by K. Hokamp to detect paralogous regions within the human genome (see Materials and Methods). Paralogy regions were characterised in terms of the number of different pairs of genes used to define them (*i.e.,* homologous pairs present on the two paired chromosomal segments, as distinct from intervening

unduplicated genes) which we term *sm* (smallest number of unique links). Where two genes on one chromosome are paired with a single gene on another chromosome, this is counted as a single unique pairing between these chromosomes. This method found 1138 blocks with $sm \geq 2$ (the minimum possible size of a block) covering 91% of the genome, and 96 blocks with at least 6 defining pairs ($sm \geq 6$) covering 44% of the total genome (just over 3 Gb). The 20 largest paralogous regions detected are listed in Table 4.1. K. Hokamp's database of paralogous regions in the human genome can be browsed at the website `http://www.gen.tcd.ie/dup`.

The most extensive region which was found (Figure 4.3) includes 29 duplicated genes and pairs a 41Mb region of chromosome 1 (including the tenascin-R locus) with a 20Mb region of chromosome 9 (including tenascin-C). The pairs of chromosomes with the highest numbers of duplicated genes forming paralogy regions with $sm \geq 3$ are, in decreasing order, 1/19, 1/6, 1/9, 7/17, 4/5, 2/7, 8/20, 2/12, 1/12, 5/15, and 12/17; these chromosomes all share at least 40 duplicate genes (Table 4.2). All chromosomes contain paralogy regions with at least one other chromosome, and most contain paralogy regions with multiple chromosomes. For example, parts of chromosome 17 are paired with parts of seven other chromosomes (Figure 4.2A) including extensive similarity to chromosomes 2, 7, and 12 around the *Hox* clusters (Ruddle *et al.*, 1994). The organisation of one paralogy region, with $sm = 9$, between chromosomes 17 and 3 is shown in detail in Figure 4.2B. This includes duplicated genes for the histone acetlytransferases PCAF and GCN5L2 (Xu *et al.*, 2000), and for the alpha and beta forms of topoisomerase II (TOP2A and TOP2B; Lang *et al.*, 1998).

4.3.1.1 Comparison with Celera results

A conceptually similar analysis was performed by Venter *et al.* (2001) in their analysis of the sequence of the human genome. The results are presented as a

Table 4.1: Details of the 20 largest paralogous regions identified in the human genome by K. Hokamp's algorithm. Blocks are listed in decreasing order of size in terms of number of duplicated genes.

Block id	chr A	chr B	kb A[a]	kb B[b]	density[c]	sm[d]
0109147702450	1	9	40.9	20.3	13.48	29
0717007601360	7	17	25.5	14.1	18.17	28
0212082302010	2	12	41.3	9.3	15.70	26
1518024002430	15	18	27.9	35.7	15.19	23
0106034601910	1	6	14.9	39.1	12.69	23
0515031102270	5	15	48.2	21.1	11.19	21
0410044701460	4	10	36.0	26.1	12.51	18
1722088301950	17	22	18.9	7.8	11.99	17
0214027500850	2	14	18.4	26.5	13.79	15
0207091200810	2	7	12.0	12.0	18.84	14
0119086901630	1	19	31.5	6.4	9.99	14
0111185902030	1	11	30.8	5.0	10.24	14
0112172000710	1	12	5.5	29.2	14.23	13
0109086100960	1	9	17.2	9.9	11.05	13
0606016000490	6	6	3.0	2.0	19.56	12
0405067601020	4	5	37.7	29.9	10.19	12
1112008400990	11	12	15.2	29.9	10.60	11
0119062800910	1	19	9.4	7.3	10.70	11
1114048500800	11	14	2.5	22.2	10.69	10
0820048800740	8	20	22.7	1.4	17.17	10

[a] length of block in kb on chromosome a
[b] length of block in kb on chromosome b
[c] proportion of genes covered by the block that are present in duplicate
[d] number of duplicated genes

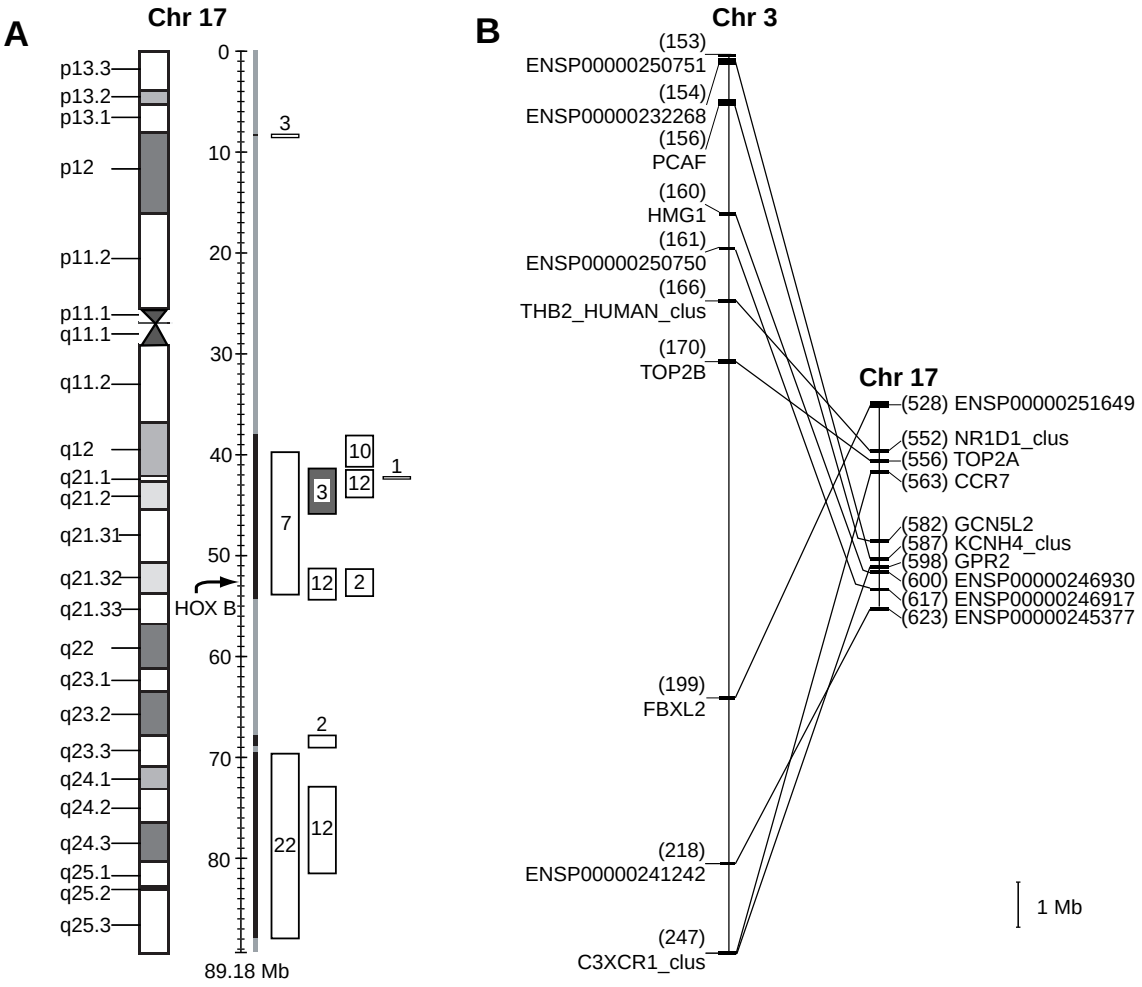


Figure 4.2: Paralogous regions on human chromosome 17. These blocks were detected by K. Hokamp's algorithm (**A**) View of chromosome 17 showing the blocks detected between this chromosome and the rest of the genome. Only blocks with $sm \geq 6$ are shown. Blocks are indicated by numbered rectangles (identifying the paired chromosome) to the right of the figure. The block with chromosome 3 that is shown in detail in (B) is shaded. The position of the *HoxB* cluster is marked. (**B**) Closer view of a nine-membered block detected between chromosomes 17 and 3. Genes whose products have names beginning with ENSP are predicted by Ensembl; other names are from HUGO or SwissProt. Numbers in parentheses indicate the rank order of genes along the chromosome (gene number 1 is the closest to the telomere of the p arm). Intervening genes that are not duplicated are not shown. Clusters of tandemly duplicated genes that have been reduced to a single representative on the map are indicated by 'clus' following the gene name. The relative duplication date of the PCAF/GCN5L2 pair from this block was calculable, and is 0.43D. This contributes to the peak in Fig 4.5B.

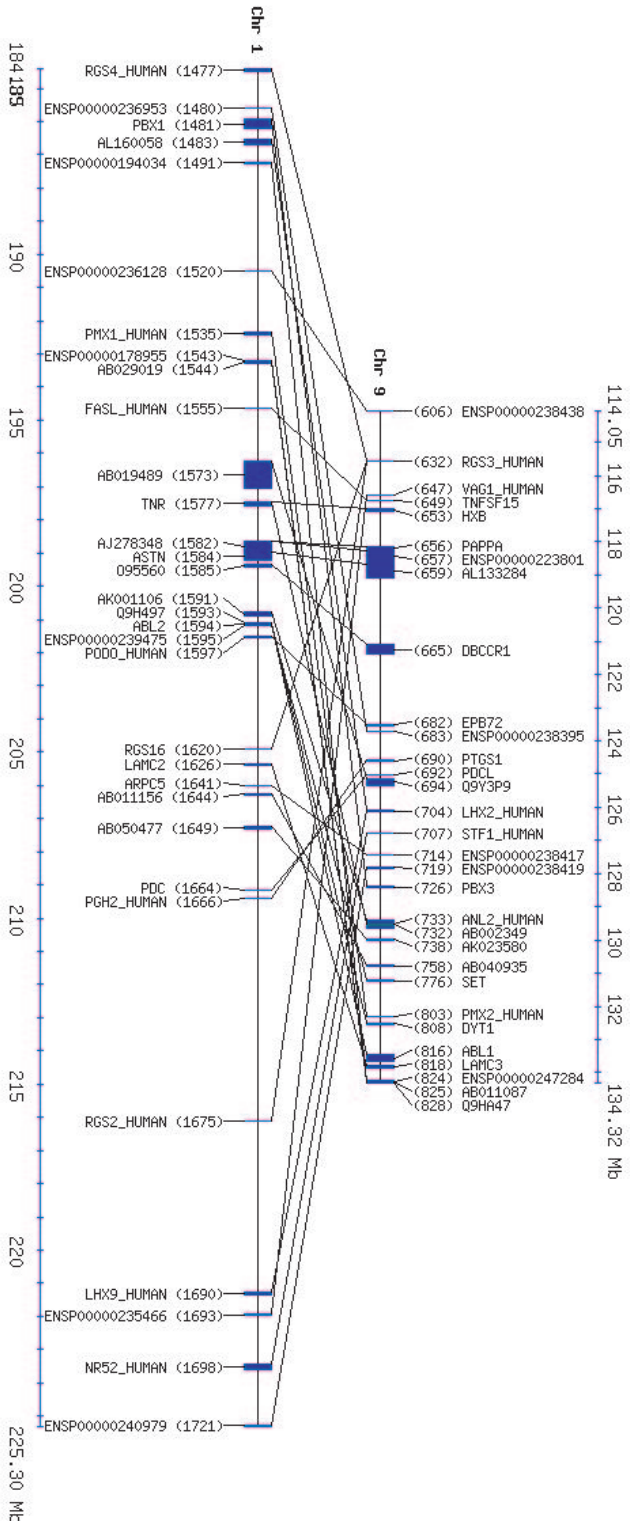


Figure 4.3: Paralogous block between human chromosomes 1 and 9. This is the largest paralogous block in the human genome detected by K. Hokamp's algorithm including 29 duplicated genes.

Table 4.2: Summary of chromosome relationships and comparison with Venter *et al.* results.

Chr	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	X	Y
1		**3**	26	0	15	65	16	16	57	3	31	44	**3**	26	11	**9**	**21**	**3**	66	**6**	**3**	0	**12**	0
2			12	**7**	**9**	0	52	10	**4**	25	**16**	46	**10**	31	12	0	27	3	12	**6**	0	**4**	**10**	0
3				12	4	7	34	**4**	0	12	12	20	7	0	0	**6**	17	0	6	0	0	**3**	39	4
4					54	0	3	**10**	0	39	9	0	13	0	0	0	0	0	0	0	0	0	**5**	0
5						3	11	**8**	0	29	**4**	0	3	0	42	16	0	**3**	7	0	0	0	3	0
6							0	**7**	**20**	0	11	3	6	11	0	**3**	**3**	**4**	0	9	9	7	12	0
7								3	0	6	6	33	**3**	0	**3**	**3**	57	0	3	0	0	**9**	**15**	0
8									0	28	**6**	0	0	**9**	0	**10**	0	0	6	52	0	0	**3**	0
9										0	9	3	**3**	0	16	0	0	0	34	0	0	0	**4**	0
10											0	3	0	0	**3**	0	15	0	3	3	**3**	0	3	0
11												36	0	16	0	**6**	7	**6**	28	**3**	18	13	13	0
12													0	0	0	0	42	0	9	0	0	0	**11**	0
13														0	0	0	**6**	0	0	0	4	0	29	0
14															0	**14**	0	0	14	16	0	0	0	0
15																**3**	3	23	**12**	0	0	0	0	0
16																	17	0	0	**17**	0	**17**	0	0
17																		0	12	0	0	36	0	0
18																			12	28	0	**3**	0	0
19																				4	**4**	**3**	0	0
20																					0	**3**	0	0
21																						0	0	0
22																							0	0
X																								11
Y																								

The number in each pairwise comparison is the sum total sm between those chromosomes (*i.e.* number of duplicated genes in blocks of $sm \geq 3$. Values in boldface indicate chromosome pairs where we identified a relationship between the chromosomes but where no pairing was found in the Celera analysis, and the converse is indicated by an underlined zero.

large figure (Fig. 13 of Venter *et al.*) illustrating chromosomal relationships. A comparison with our results is not entirely straightforward because the information provided on both the method and the results is minimal. The most robust way of considering their results is by simply recording the presence or absence of any relationship between all pairwise chromosome comparisons in this figure. It is not possible to determine which genes, or how many are involved in Venter *et al.*'s pairings.

The results of this comparison are summarised in Table 4.2 where the number in each pairwise comparison is the sum total of the the sizes of blocks between those chromosomes found by K. Hokamp's algorithm. Because the same gene may be involved in several blocks between the same pair of chromosomes, this is not necessarily the same as the number of gene pairs in blocks shared by the chromosomes, but is an upper limit. Of the 276 possible chromosome pairs, our method detected 151. We detected 55 regions that were not found in the analysis by Venter *et al.*, and we did not detect any relationship between 21 pairs of chromosomes for which they illustrated pairings.

A recent comparison of the Ensembl and Celera predicted gene sets revealed that the novel genes predicted by each group are largely non-overlapping (*i.e.*, unique to each dataset Hogenesch *et al.*, 2001). This may, at least partially, explain the differences in the results of the genome-wide search for paralogous regions.

4.3.1.2 Paralogous regions - block duplication or artefacts?

Even if no ploypoidies or block duplications had ever occurred during the evolution of the human genome, some apparent duplicated blocks would probably exist purely by chance. The existence of intervening non-duplicated genes within a block and gene order differences between these putatively paralogous regions means that the significance of the regions we discovered

Box 4.1: Computer simulations with shuffled genome map

The genome map was shuffled in 1000 computer simulations. The block detecting algorithm was applied to each shuffled genome exactly as was done for the real genome map. The blocks found in the shuffled genomes were compared to those of the real genome in terms of frequency of blocks of different sizes. This used the block-detecting algorithm by K. Hokamp, and the database of paralogous regions generated by the algorithm.

is not intuitively obvious. Are these the rearranged remnants of extensive genomic (or whole genome) duplication events? Or, are they simply artefacts of the lenience of this algorithm, combined with the frequency of multi-gene families in the human genome? We used computer simulations with randomised genome maps (Box 4.1) to estimate the background level of block detection by this algorithm. Any blocks detected in the randomised genome must be artefacts.

The results of the block detection algorithm from 1000 simulations are summarised in Table 4.3. The number of blocks of $sm = 2$ was similar in the simulations and the real data. All blocks with $sm \geq 3$ are highly significant by two statistical tests, a Z-score test, and a percentile rank test. The Z-score (normal deviate) test assumes that the number of blocks of a particular size found in the simulations is normally distributed about the mean (μ). In a normal distribution 99.9% of the measurements lie within $\mu \pm 3.29\sigma$, where σ is the standard deviation of the distribution. The Z-score is the number of standard deviations by which the observed value exceeds the mean of the simulations. The percentile rank is a non-parametric test (*i.e.*, a test that does not require the estimation of the population variance or mean). The percentile rank indicates the proportion of the simulations that had equal or fewer results that were less than or equal to the observed value. Therefore if a value is in the 99th percentile, it has a statistical significance of 1% as

Table 4.3: Sizes of duplicated regions in the human genome, compared to simulations where gene order was shuffled

	Number of blocks				
sm[a]	**Real genome**	**Simulations**		**Z Score**[b]	**Percentile**[c]
		mean	SD		
2	1138	1051.67	29.43	2.93	99.9
3	260	159.05	12.35	8.17	100.0
4	93	30.10	5.62	11.20	100.0
5	55	6.89	2.71	17.76	100.0
≥6	96	2.56	1.63	57.48	100.0

[a]Number of duplicated genes comprising the block
[b]Number of standard deviations by which the number of blocks in the real genome exceeds the mean of simulations
[c]Proportion of simulation results that were less than or equal to the observed value

estimated by this non-parametric test.

The deviation in terms of the Z-Score is more marked for the larger blocks, with blocks defined by at least six duplicated genes being in excess of 50 standard deviations more frequent than expected from the simulations. This analysis indicates that any paralogous region with $sm \geq 6$ has almost certainly been formed by a regional duplication, and that blocks of $sm = 3$ are on the borderline of statistical significance for this dataset. The mean number of blocks with $sm \geq 6$ in the simulations was 2.56, compared with the observed value of 96. The highest number of blocks with $sm \geq 6$ in any of the 1000 simulations was nine. The only alternative hypothesis that could fit these data is selection for clustering of these genes on a chromosome as has been suggested for the mammalian MHC gene complex and the *surfeit* locus (Hughes, 1999a).

Box 4.2: Computer simulations with randomised block distribution

Paralogous regions of the genome were defined randomly in a computer simulation. The number and sizes of the blocks found in the real genome were preserved, as were the number and sizes of chromosomes. Blocks were randomly assigned a position in the genome with the only restrictions being that they must not run over the end of a chromosome and that the physical positions of paired chromosomal segments could not overlap (*i.e.,* a region of chromosome could not form a block with itself). This was repeated 10,000 times each for all blocks with $sm \geq 3$, ≥ 4, ≥ 5, and ≥ 6.

Genome coverage was measured for all genes in each of the simulations and in the real genome (K. Hokamp's database of paralogous regions). The coverage was defined as the number of blocks that cover the location of the gene regardless of whether or not that gene was involved in the pairing between the two chromosomal segments. A gene that is not under any block has a coverage of zero, a gene that is covered by a single block has a coverage of one, and so on.

4.3.1.3 Extent of paralogous block overlap

In the case of a single round of whole genome duplication, the whole genome should be duplicated once, and once only. Paralogy regions that result from this event should not overlap each other, *i.e.,* each portion of genome should only be paired with one other. After a second genome duplication event each region will be present in four copies, *i.e.,* each region should be paired with three other regions. Regions with more or fewer overlaps do not constitute falsification of this hypothesis because they could arise from extra segmental duplications, or gene loss respectively. However there should be an excess of one-fold block coverage in the case of a single genome duplication, or three-fold block coverage in the case two rounds of genome duplication.

Starting with the assumption that the paralogy regions defined by K. Hokamp's algorithm did duplicate *en bloc*, we used computer simulations (Box 4.2) to investigate the degree of overlapping blocks compared to a

random distribution of block positions. These results were compared with the amount of coverage seen in the real block distribution. The significance of the difference between the observed amount of overlap in the genome and that seen in the 10,000 simulations is expressed as a percentile rank. The percentile rank indicates the proportion of the simulations that had equal or fewer genes with the same amount of overlap. The results for different thresholds of block size are shown in Figure 4.4.

Disappointingly there is no clear pattern in these results. Some levels of coverage appear to approach statistical significance for some block size thresholds, and then appear insignificant for another. For example, $sm \geq 4$ has low 1× coverage and high 3× coverage which taken alone could be interpreted as evidence in support of the 2R hypothesis (3× coverage indicates that paralogy regions are present in four copies), but the 3× coverage is insignificant for other *sm* thresholds. One could easily choose data that support or contradict the genome duplication hypothesis with selective representation of results, so I am forced to conclude that this metric is simply uninformative.

4.3.2 Estimating dates of gene duplications in the vertebrate lineage using *Cænorhabditis* and *Drosophila* outgroups

In testing genome duplication, analysis of the relative map position of paralogues is complemented by analysis of the timing of the duplication events giving rise to these paralogues. We sought to estimate the ages of gene duplications in the vertebrate lineage by employing the molecular clock where applicable. The human, *Drosophila*, and *Cænorhabditis* proteomes were compared to identify gene family expansions in the chordate lineage, *i.e.,* gene duplications that postdate the arthropod-chordate divergence. Family definition was conservative with no gene permitted to be a member of more

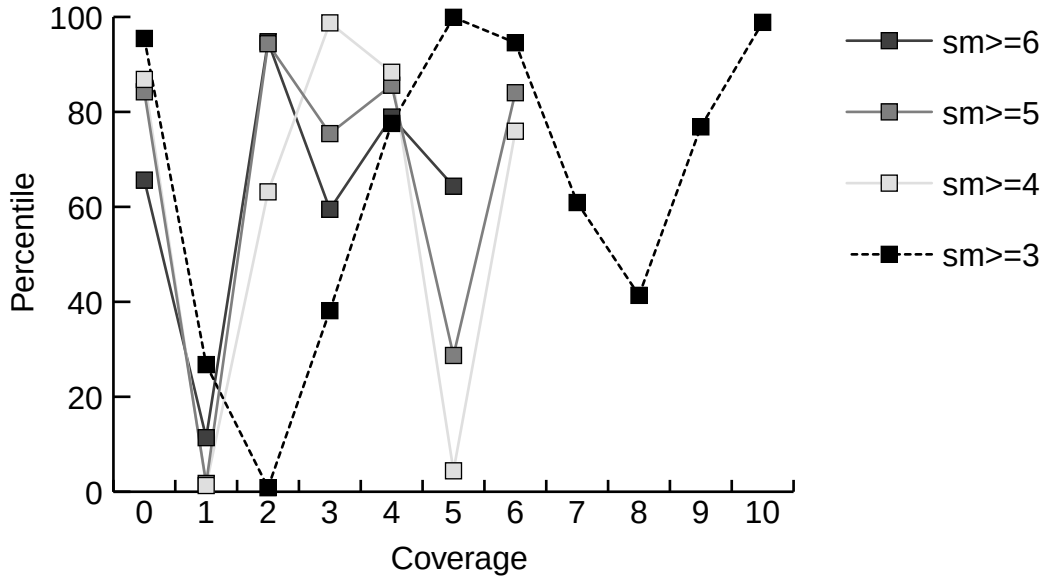


Figure 4.4: Comparison of genome coverage overlap by the paralogous regions found in the real genome with randomly distributed paralogous regions. Coverage indicates the number of blocks that cover a single gene. If the whole genome had a coverage of 1, then everything in the genome could be said to have duplicated once, and once only. Triplicated regions have a coverage of 2 because one region is similar to two others. Quadruplicated regions are seen as one region matched with three others, and so should have a coverage of 3. The percentile rank indicates the proportion of the 10,000 simulations that had equal or fewer genes with the same amount of coverage. The different lines indicate the different thresholds of minimum size block used in independent simulations.

that one family. This strategy was adopted to avoid the ambiguities arising from networks of gene families which may only be related by a homologous domain.

Of the 1808 families identified, 758 had between two and ten members (Table 4.4). One-membered families have either not been duplicated or have lost all paralogues, and are not informative in this analysis. The 758 two-to-ten membered families were aligned and phylogenetic trees were drawn as described in Materials and Methods. 191 of these families were informative, in that the molecular clock was not rejected by the two-cluster statistical test (Takezaki *et al.*, 1995), and the tree topology was consistent with vertebrate lineage gene duplication (Table 4.4).

In a tree of a gene family, each intra-specific node represents a gene duplication event. A bifurcating tree of an N-membered family contains $N-1$ intra-specific nodes. The ages of these nodes were calculated relative to arthropod-chordate divergence (termed D) as illustrated in Figure 4.5 A. Branch lengths were estimated using a Gamma correction for multiple hits, and linearised trees were drawn under the assumption of a molecular clock. The relative ages of nodes in these trees were estimated by expressing the branch length from the node to the tip of the tree as a proportion of the total length from the outgroup divergence node to the tip (Figure 4.5 A). Because the definition of gene families for this analysis excluded families where the duplication was also present in the outgroups, all node ages must be younger than D.

The distributions of ages of duplication events (Figure 4.5 B-F) all show an excess of nodes originating between 0.4-0.7 D. This is most marked in the pooled histogram for all families with at least two members (Figure 4.5 B) and for the two-membered families alone (Figure 4.5 C). The timing of the arthropod-chordate divergence is uncertain, but one recent estimate of D=833 Mya (Nei *et al.*, 2001) would place this peak between 333-583 Mya,

Table 4.4: Analysis of vertebrate gene families

Human gene family size	**Number of families**	**Alpha undefined**[a]	**Failed two-cluster test**[b]	**Outgroups not together**[c]	**Linearised trees**[d]
2	377	124	93	32	128
3	179	17	79	36	46
4	66	2	43	15	6
5	52	2	26	18	6
6	14	2	4	5	2
7	13	1	7	4	1
8	10	0	7	3	0
9	9	0	7	3	0
10	38	6	24	6	2
Totals:	758	154	290	121	191

[a]Shape parameter for gamma distribution not calculable
[b]Failed to reject substitution rate heterogeneity at 5% significance
[c]Some gene duplications predated the arthropod-chordate divergence
[d]Number of phylogenetic trees with branch lengths calculated under the assumption of a molecular clock

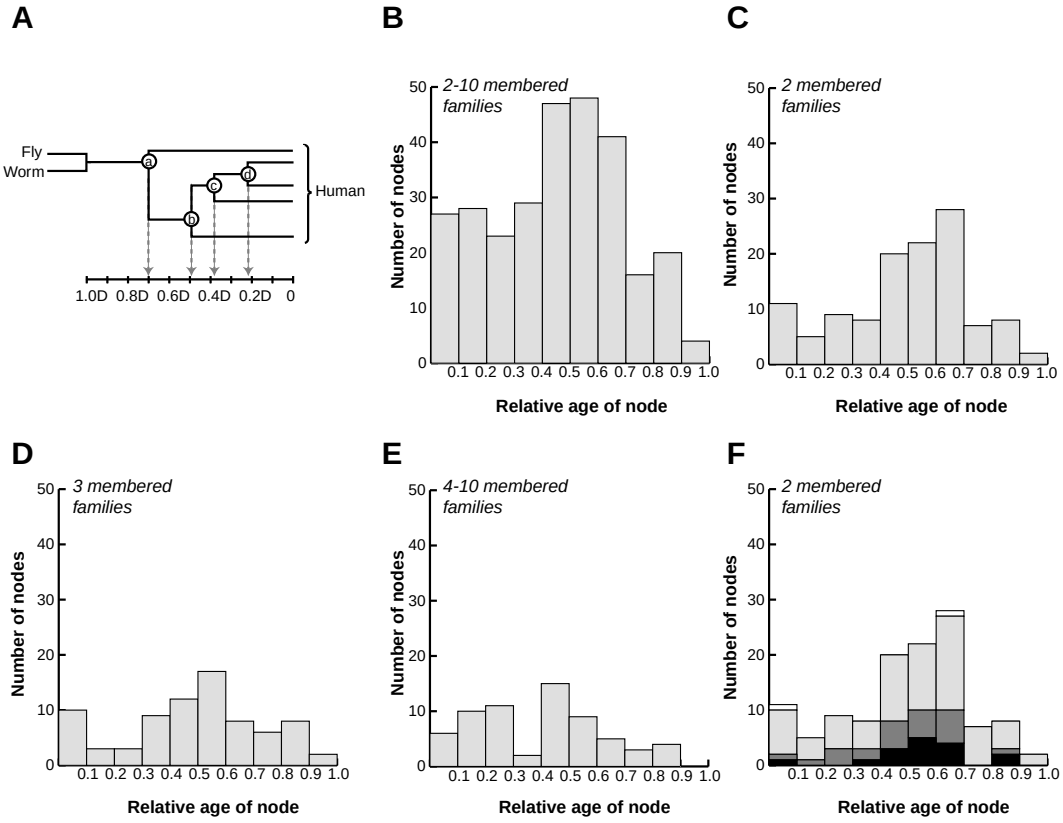


Figure 4.5: Estimation of gene duplication dates using linearised trees with fly and worm outgroups. **(A)** Model linearised tree of a five-membered gene family. The time of duplication for each of the nodes $a-d$ is indicated on the scale below the tree. Ages are expressed relative to the arthropod-chordate divergence (D); for example, the age of node a is $0.7D$. **(B-E)**Distribution of the estimated ages of vertebrate specific nodes of 2-10 membered, 2 membered, 3 membered, and 4-10 membered families respectively. Each node represents a duplication event, and a family with N members has $N-1$ nodes. **(F)** Breakdown of estimated duplication dates among gene pairs mapped to paralogy blocks for two-membered gene families. The duplicated gene pairs in the histogram in (C) were placed into four categories: pairs making up paralogy regions with $\geq$6 duplicated genes (black); pairs making up paralogy regions with $\geq$3 duplicated genes (dark grey); pairs that appear on the map but are not present in blocks with at least 3 duplicated genes (light grey); and pairs for which one or both of the genes did not appear on the condensed gene map used for our analysis (white).

which includes the origin of vertebrates. Two membered families have only one gene duplication event per tree. The effect is weaker for larger families (Figure 4.5 D and E) which may be attributable to the fact that there are fewer data and also more noise from the many gene duplication events that shaped these families.

4.3.3 Estimation of duplication dates using a topology approach with sequences from additional vertebrates

The ages of duplications in the two-membered families were also calculated by a topology-based method. Trees were drawn from alignments with available non-mammalian vertebrate orthologues and the duplications were dated according to whether they pre- or post-dated major lineage divergences within vertebrata (Box 4.3 overleaf; trees are shown in Figures 4.10 on page 106 to 4.18 on page 113). For example, the human PCAF and GCN5L2 genes (Xu *et al.*, 2000) were estimated by the method described in the preceding section to have a duplication date of 0.43 D or 358 Mya. When other vertebrate sequences are included in the tree (Figure 4.6), the branching order confirms that the gene duplication is at least older than the chicken divergence (310 Mya; Kumar and Hedges, 1998). Similar topological support was found for 31 of the 36 families that could be examined in this way (Figure 4.7 on page 101). If the estimate of Wang *et al.* (1999) for the time of arthropod-chordate divergence (993 Mya) is used instead of Nei *et al.*'s estimate (833 Mya), then only 3 families give incongruent results between the two methods. Overall, these phylogenetic analyses using other vertebrate sequences provide support for the approximate time scale used, and for the consistency of the molecular clock method.

Box 4.3: Topology-based method of duplication date estimation

To test the congruence of the molecular-clock based method with topology-based methods, human proteins from 2-membered families were compared to a database of 105,860 non-human vertebrate sequences from SWALL (SwissProt plus daily updates, 19th Sept. 2001) using the same BLASTP protocol as in Section 4.2.3 on page 80, and enforcing a minimum alignment length of 30% of the longer sequence. Neighbour-joining trees with gamma-corrected distances were drawn for each family, and the trees were examined to determine whether the gene duplication pre- or post-dated the divergence of the lineages leading to bony fish, amphibians, or birds and reptiles.

4.3.4 Placing duplication date estimates on the paralogy map

For the two-membered gene families, it is possible to look up whether any gene pair appears on K. Hokamp's map of paralogous regions described earlier. This analysis shows a non-uniform distribution for blocks of $sm \geq 3$ (P=0.02 by Kolmogorov-Smirnov test), with many of duplicated genes in the age range 0.4-0.7 D also being components of paralogous regions (Figure 4.5 on page 97 F). The age distribution of pairs in blocks with $sm \geq 6$ is not significantly different from a uniform distribution (P < 0.5 by Kolmogorov-Smirnov), which is not surprising considering there is little statistical power left in such a small sample size. The number of pairs of each age group on the map of paralogous regions of $sm \geq 3$ is not significantly different from a random subset of the total age distribution of pairs with known map position (P=0.5 by Kolmogorov-Smirnov)

A similar excess of pairs in blocks aged between 0.4-0.7 D is seen in larger gene families (Figure 4.8 on page 102). As there are more pairwise ages ($\frac{N(N-1)}{2}$) than there are duplication events ($N-1$) for families with $N \geq 3$ (where N is the size of the gene family), this is not simply a subset

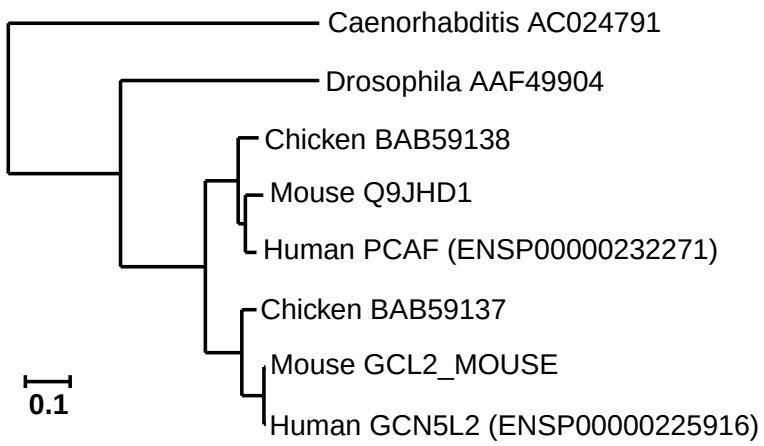


Figure 4.6: Phylogenetic tree calculated from protein sequences in the PCAF/GCN5L2 family. GenBank, Swissprot, or Ensembl identifiers are given beside species names.

of mapped genes as it was in the case of two membered families. Therefore statistical interpretation of this result is difficult. A single duplication event (one node on a tree) will contribute to several pairwise ages if there are more than two operational taxonomic units (OTUs) below that node.

Only 14 blocks contained two or more pairs of genes for which a duplication date was calculable (Figure 4.9 on page 103). No more than three age estimates were calculable from any one block. In order for these data to be appropriate for an ANOVA test the standard deviation of each group (block) should be the same. It is obvious from inspection of Figure 4.9 that this assumption is violated. For ten of the 14 blocks, all of the calculable dates are within the range 0.4-0.7 D, and the ages from three of other blocks overlap with this age range. This result is weak because there are so few data, but it further indicates that paralogous regions tended to originate around 0.4-0.7 D.

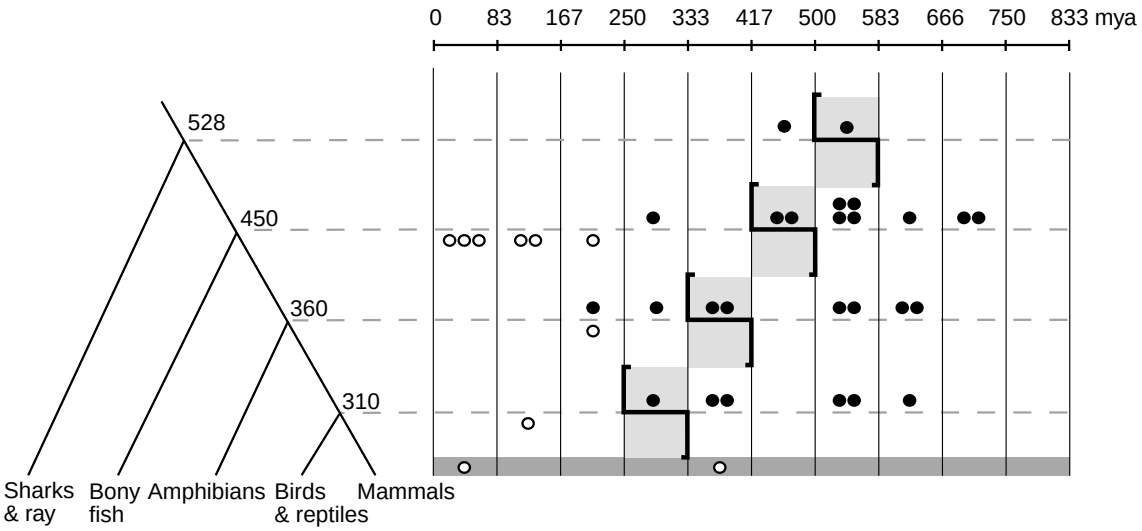


Figure 4.7: Comparison of topology-based and clock-based estimation of the dates of gene duplication. Each dot represents a pair of human genes for which a homologous sequence from non-mammalian vertebrate species was available. The vertical position of a dot indicates the minimum (black dots) or maximum (white dots) age of its gene duplication, as determined by the branching order of a phylogenetic tree. For example, the leftmost dot in the diagram indicates a gene duplication that was found to have occurred more recently than the divergence of the bony fish lineage, as shown by the topology of a tree that included the two human sequences, a bony fish sequence, and fly and worm outgroups. The horizontal position of a dot indicates the gene duplication date estimated by the molecular clock, using the same methodology and age groups as in Figure 4.5 (*i.e.,* using only human, fly and worm sequences). When the two methods give congruent results, all black dots should lie to the right of the thick black lines, and all white dots should lie to their left. This is true for 31 of the 36 dots. Any dots inside the grey bar at the bottom of the figure indicate a post-mammalian gene duplication. The thick lines are zigzags due to the use of binned age classes for the molecular clock date estimates. The timescale for speciations, indicated on the tree at the left, is from Kumar and Hedges (1998).

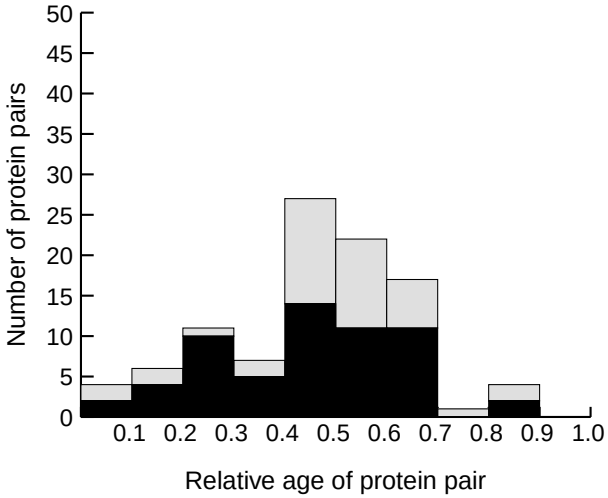


Figure 4.8: Ages of pairs of genes from gene families with more than two members that are in blocks with $sm \geq 6$ (black) or $sm \geq 3$ (grey).

4.3.5 Phylogenetic test of (AB)(CD) topology in human four-membered families

A four-membered family must fulfil two criteria in order to be consistent with the 2R hypothesis. The four genes must have duplicated after the origin of vertebrates, and they must exhibit a symmetrical (AB)(CD) topology (Skrabanek and Wolfe, 1998). When this topology test was applied to the four-membered families defined in this analysis 47% of the trees were of the form (AB)(CD) (Table 4.5 on page 104). Sequential gene duplication will give rise to a symmetrical (AB)(CD) topology 1/3 of the time, and an asymmetrical topology (A(B(CD))) the remaining 2/3 of the time (Figure 1.7 on page 34). The observed frequency of the symmetrical topology in four-membered families defined here is not significantly greater by χ^2 test (P=0.25) than is expected from sequential gene duplications, and is consistent with results from other analyses (*e.g.*, Hughes, 1999b; Lander *et al.*, 2001; Martin, 2001).

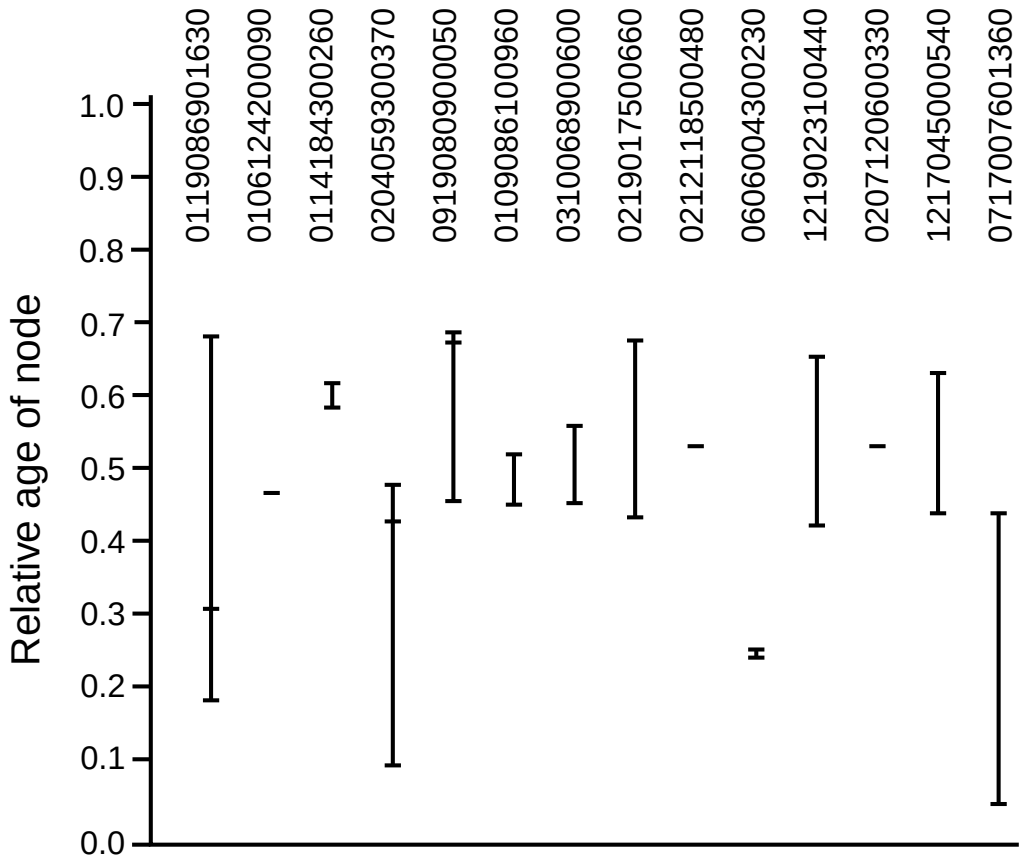


Figure 4.9: Relative ages of pairs of genes from the same block. 14 blocks are shown where there were at least two duplication date estimates for members of the block. No more than three age estimates were calculable from any one block. Horizontal bars indicate the age of a duplicated pair of genes. Vertical bars connect age estimates from the same block. Block names are indicated above.

Table 4.5: Tree topologies of four-membered vertebrate gene families.

Method[a]	Number of families with topology	
	(AB)(CD)	(A(B(CD)))
NJ/Gamma	25 (49%)	26 (51%)
NJ/Gamma/bootstrap	15 (47%)	17 (53%)

[a]Phylogenetic trees were constructed by the Neighbour-joining method using a Gamma correction for multiple hits (see Materials and Methods). Topologies were examined with and without a requirement for 80% bootstrap support for the topology determining branches. There is only one topology determining internal branch for the (A(B(CD))) tree, but two for the (AB)(CD) tree (Figure 1.7).

4.4 Discussion

We have shown that the human genome contains more large paralogous segments than expected by chance, that an unexpectedly large number of duplicated genes are in the approximate age range 333-583 Mya, and that many of the gene pairs of this age are located in paralogous regions. The results are even more striking considering that the human genome is not yet fully sequenced or annotated, which means that the detection of paralogous segments may have been hindered by many genes remaining unidentified or assigned to the wrong location. Errors of this kind will almost certainly make paralogy regions harder, not easier, to detect. Some of the paralogous segments may indicate functional links between genes (Hughes, 1998). For example, genes encoding the four members of the transmembrane-type subgroup of metalloproteinases (Kojima *et al.*, 2000) are each closely linked to genes for four copines, a small (five-member) family of proteins suggested to be involved in membrane trafficking (Tomsig and Creutz, 2000), on chromosomes 8/14/16/20.

The symmetric topology expected by the 2R hypothesis may not always be easy to retrieve by phylogenetic methods, even when it is the true topology.

An analysis by Zharkikh and Li (1993) showed that tree drawing methods have more success in recovering asymmetric trees than symmetric ones. Furthermore, when requiring bootstrap support for a phylogeny, there are two nodes that are critical to the (A,B)(C,D) topology, but only one critical node of the alternative (A(B,C,D)) topology (the branching order of B, C, and D is irrelevant to the asymmetry) as illustrated in Figure 1.7 on page 34. This may mean that more symmetric trees will fail in terms of bootstrap significance.

Although the result in Table 4.5 appears to stand against two rounds of polyploidy, proponents of the 2R hypothesis argue that it is still compatible with a variant of the 2R hypothesis where the two rounds of genome duplication occurred in close succession to form a species with partial octosomic inheritance that subsequently became diploidised (Gibson and Spring, 2000; Wolfe, 2001). It should be noted that critics of the 2R hypothesis have declared that it is 'difficult to devise ways to discriminate between this hypothesis [genome duplication] and alternatives' (Hughes, 1999b, p.575) because it can be modified to accommodate almost any observation.

Our findings are consistent with the 2R hypothesis but do not constitute proof of it. They are also consistent with other possible scenarios (Smith *et al.*, 1999; Martin, 1999), including aneuploidy (chromosomal duplication) or an increased rate of production (or fixation) of duplicated chromosomal segments in an early chordate. Polyploidy is probably the most parsimonious explanation, particularly if deletion affects several genes in a single event (see Conclusions), but we do not see any specific evidence for two rounds of genome duplication as opposed to one. Genome sequencing in invertebrate chordates and basal vertebrates such as *Ciona*, *Amphioxus*, or lamprey should throw light on the mechanism by which paralogous regions originated in chordate genomes.

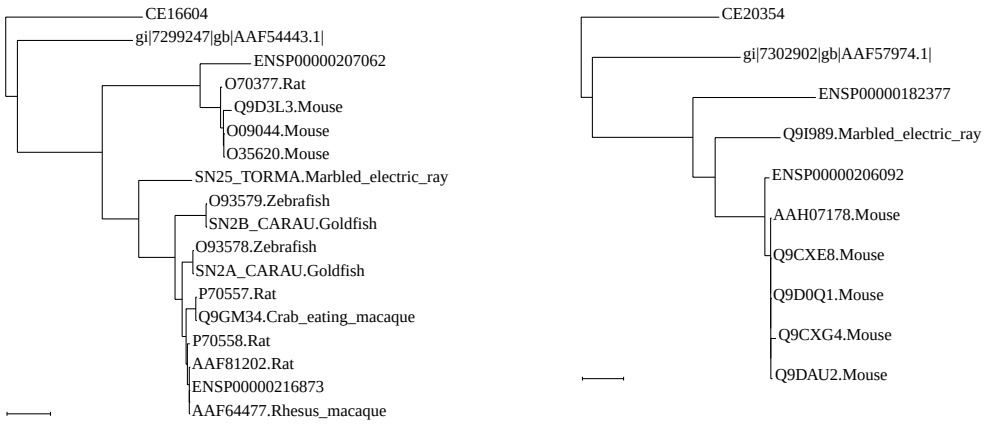


Figure 4.10: Phylogenetic tree topologies indicating duplication of the human genes prior to divergence of the cartilaginous fish lineage. Human genes are listed by their Ensembl accession number (beginning 'ENSP'), worm sequences are listed by their WormPep accession number (beginning 'CE'), fly sequences are listed with the GenBank identifier and accession number (beginning 'gi'), for other species GenBank, Swissprot, or Ensembl identifiers are given beside species names. The scale bar for each tree indicates a distance of 0.1 substitutions per site.

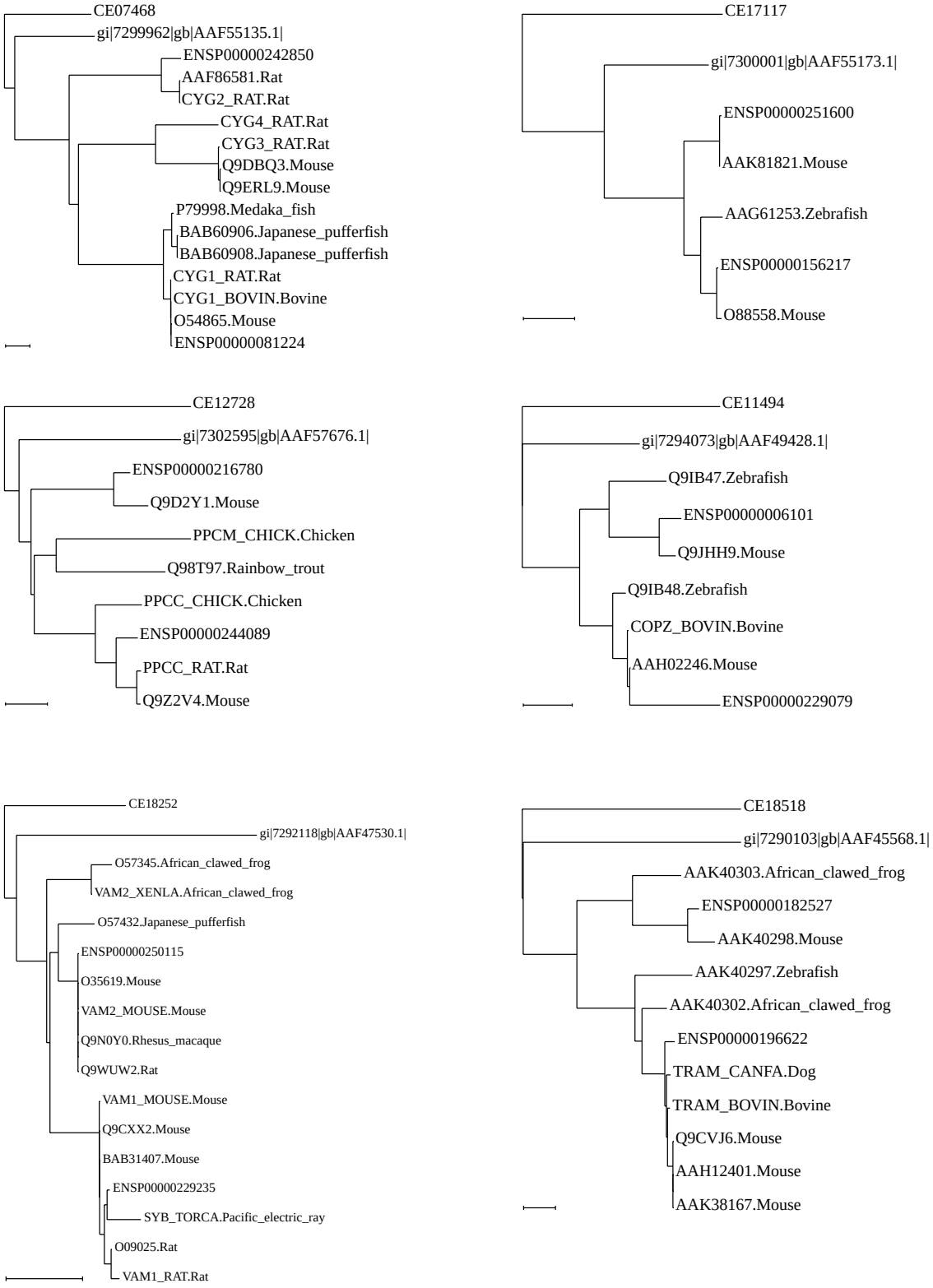


Figure 4.11: Phylogenetic tree topologies indicating duplication of the human genes prior to divergence of the bony fish lineage. Human genes are listed by their Ensembl accession number (beginning ENSP), for other species GenBank, Swissprot, or Ensembl identifiers are given beside species names. The scale bar for each tree indicates a distance of 0.1 substitutions per site.

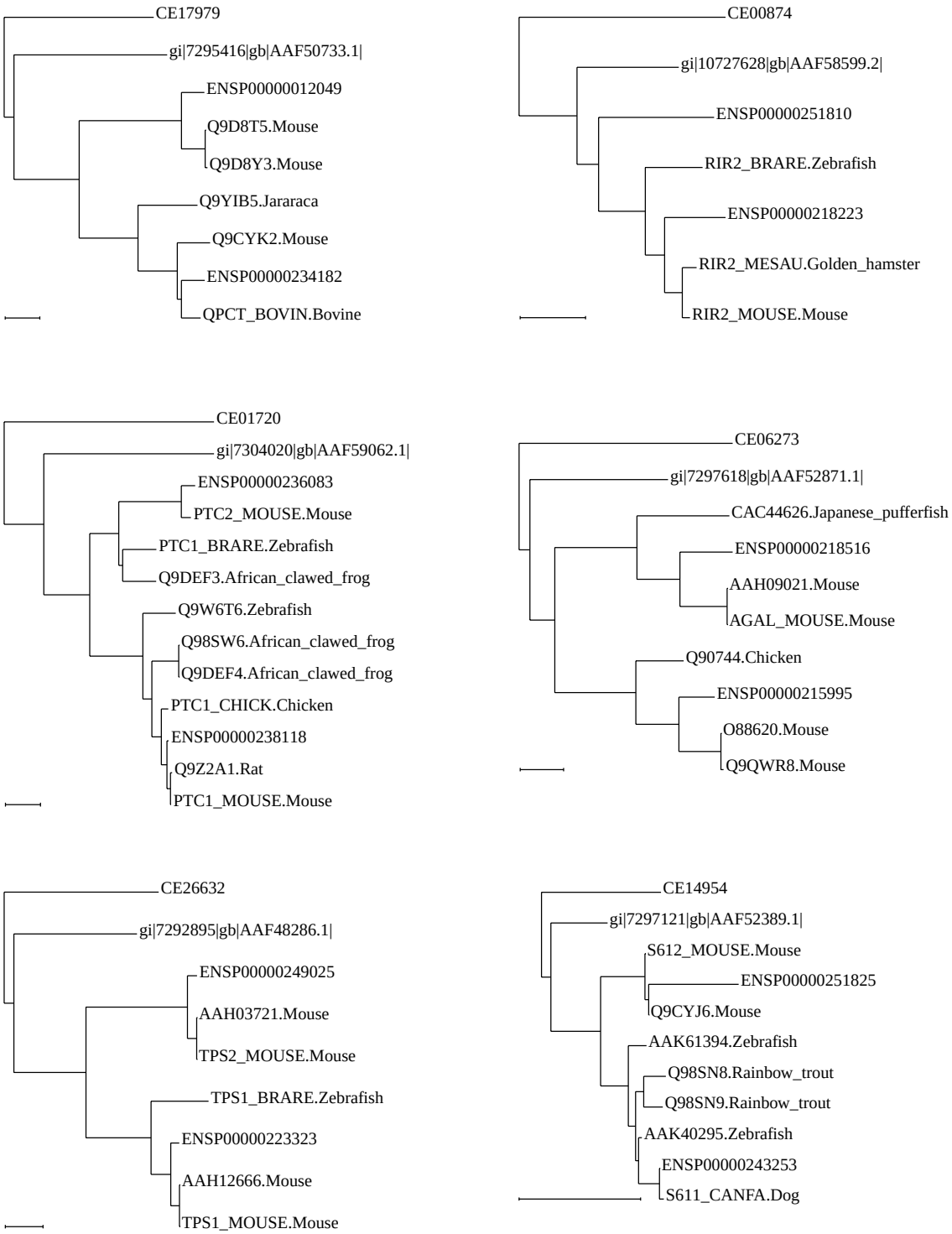


Figure 4.12: Continuation of Figure 4.11 on the page before

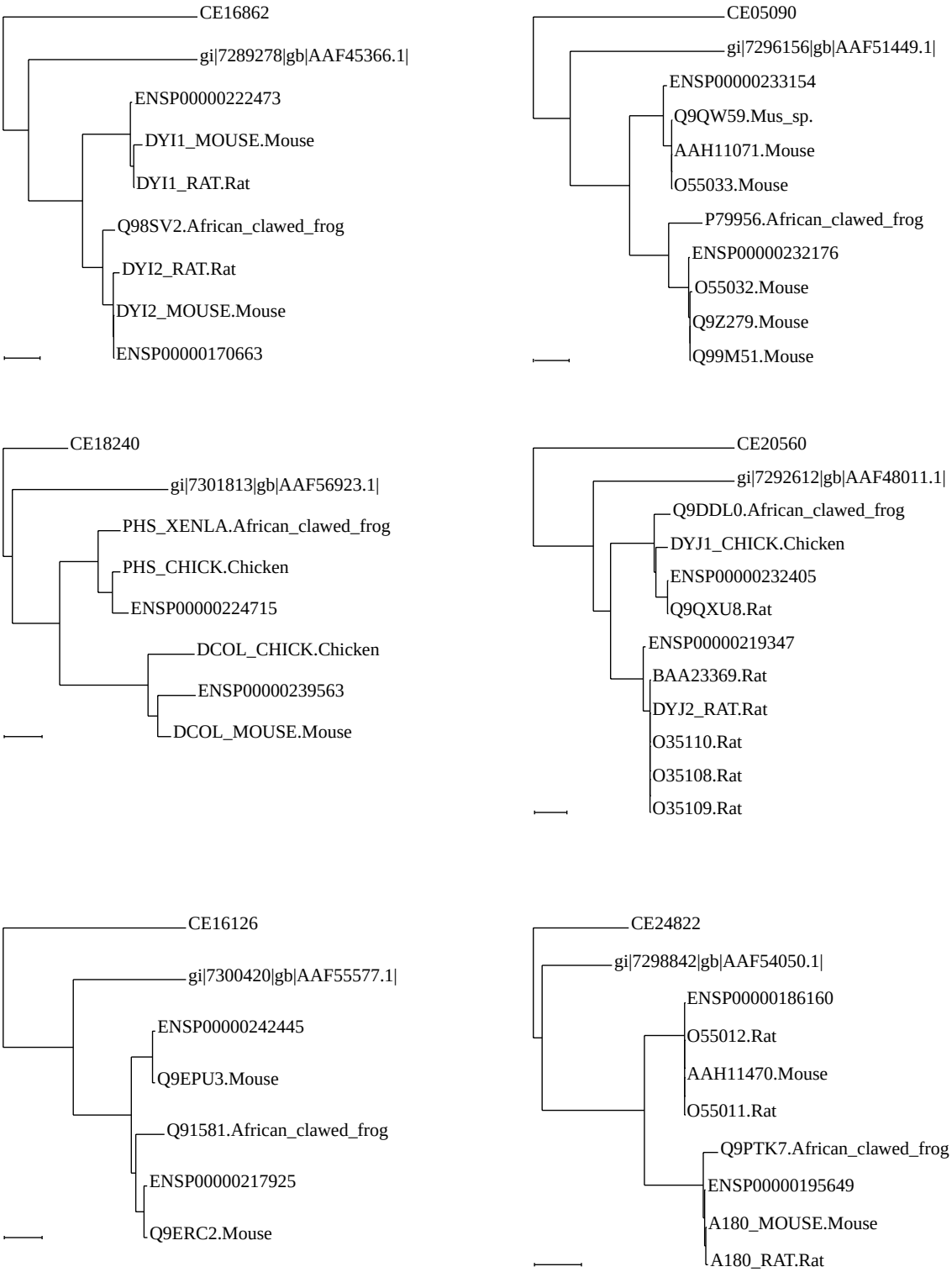


Figure 4.13: Phylogenetic tree topologies indicating duplication of the human genes prior to divergence of the amphibian lineage. Human genes are listed by their Ensembl accession number (beginning ENSP), for other species GenBank, Swissprot, or Ensembl identifiers are given beside species names. The scale bar for each tree indicates a distance of 0.1 substitutions per site.

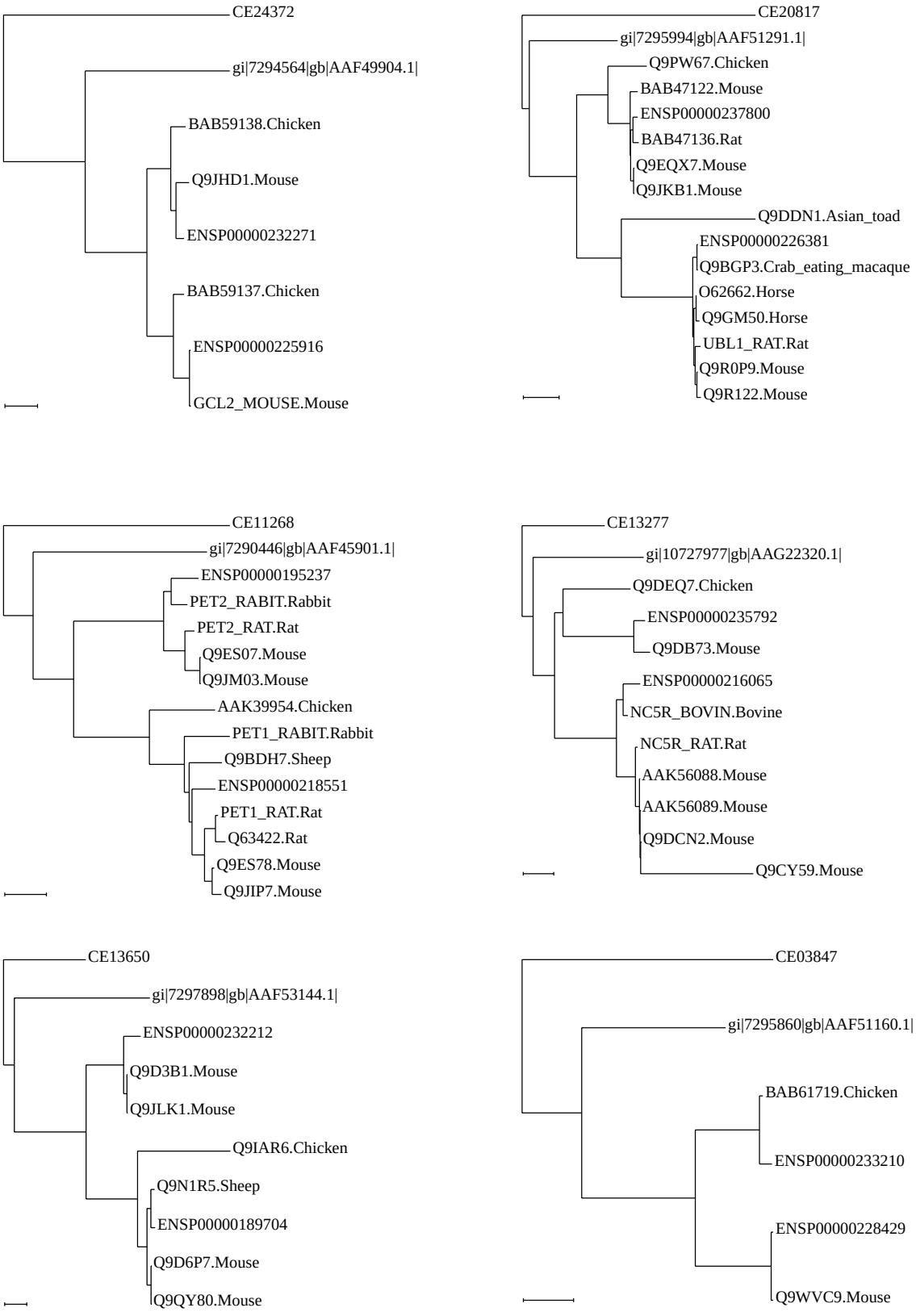


Figure 4.14: Phylogenetic tree topologies indicating duplication of the human genes prior to divergence of the lineage leading to birds and reptiles. Human genes are listed by their Ensembl accession number (beginning ENSP), for other species GenBank, Swissprot, or Ensembl identifiers are given beside species names. The scale bar for each tree indicates a distance of 0.1 substitutions per site.

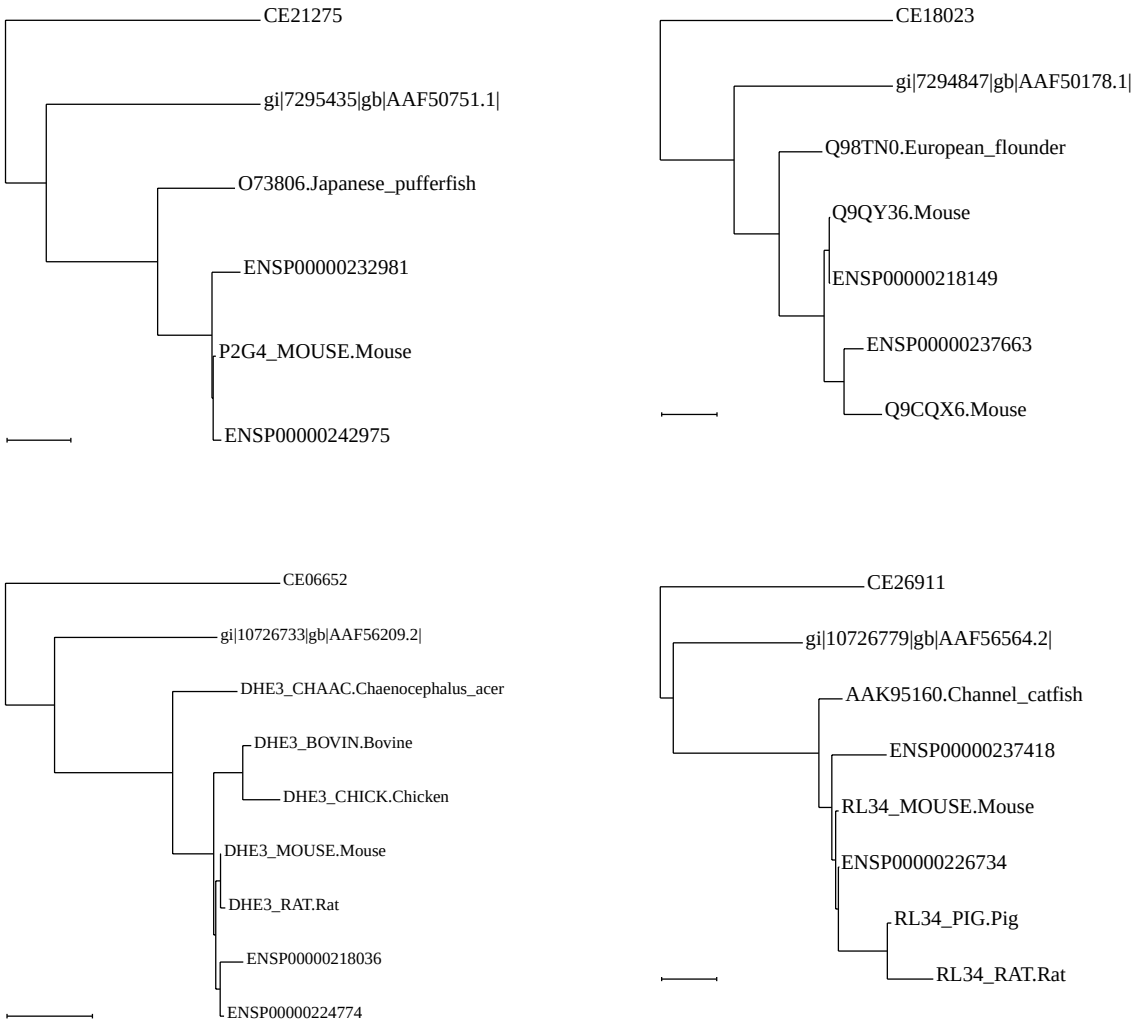


Figure 4.15: Phylogenetic tree topologies indicating duplication of the human genes after divergence of the bony fish lineage. Human genes are listed by their Ensembl accession number (beginning ENSP), for other species GenBank, Swissprot, or Ensembl identifiers are given beside species names. The scale bar for each tree indicates a distance of 0.1 substitutions per site.

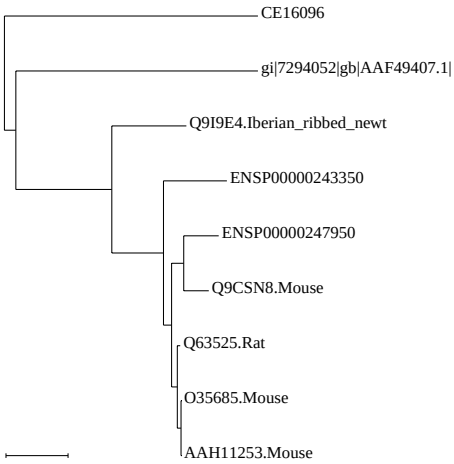


Figure 4.16: Phylogenetic tree topology indicating duplication of the human genes after divergence of the amphibian lineage. Human genes are listed by their Ensembl accession number (beginning ENSP), for other species GenBank, Swissprot, or Ensembl identifiers are given beside species names. The scale bar indicates a distance of 0.1 substitutions per site.

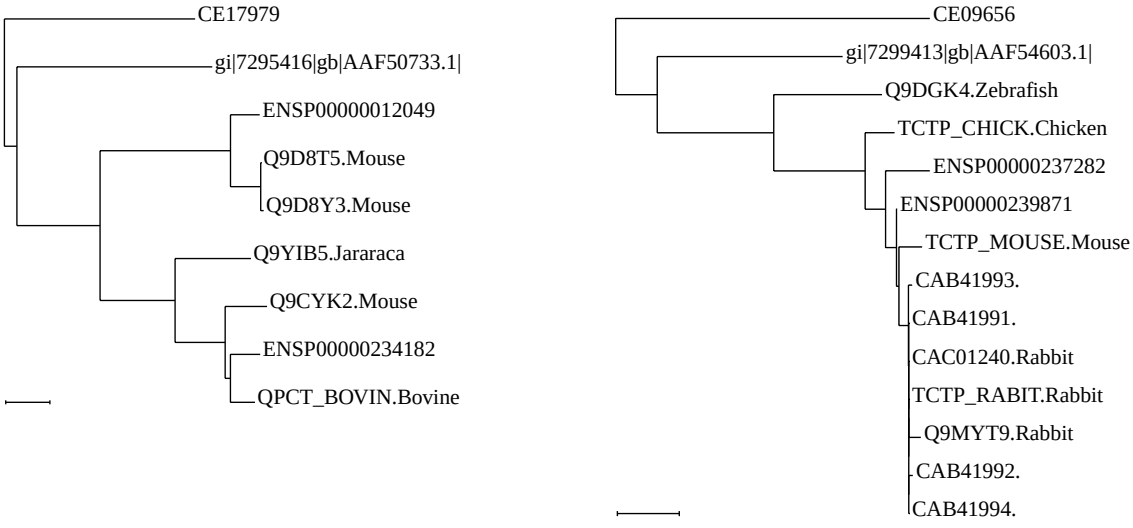


Figure 4.17: Phylogenetic tree topologies indicating duplication of the human genes after divergence of the lineage leading to birds and reptiles. Human genes are listed by their Ensembl accession number (beginning ENSP), for other species GenBank, Swissprot, or Ensembl identifiers are given beside species names. The scale bar for each tree indicates a distance of 0.1 substitutions per site.

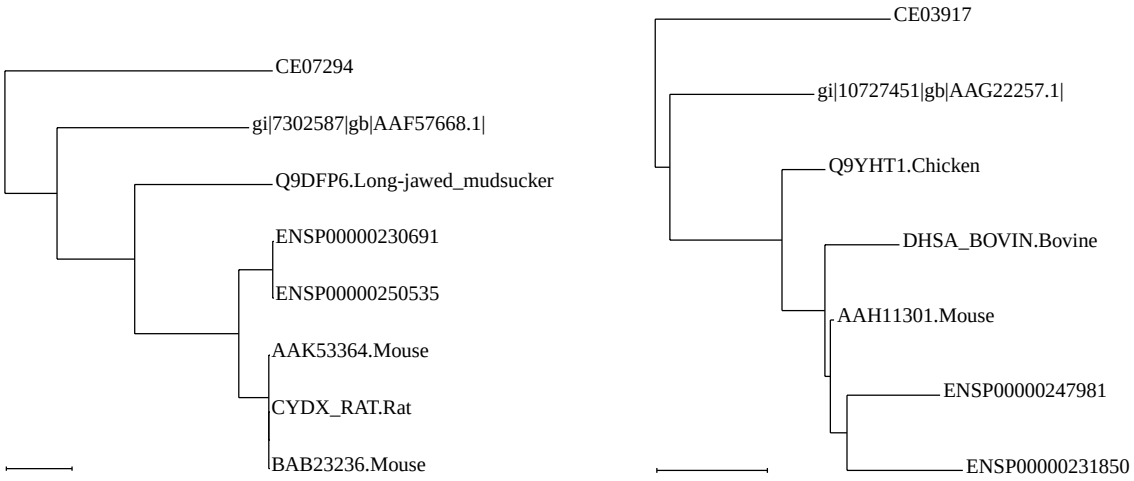


Figure 4.18: Phylogenetic tree topologies indicating duplication of the human genes within the mammalian lineage. Human genes are listed by their Ensembl accession number (beginning ENSP), for other species GenBank, Swissprot, or Ensembl identifiers are given beside species names. The scale bar for each tree indicates a distance of 0.1 substitutions per site.

Chapter 5

Conclusions

5.1 Rapid genome rearrangement following polyploidy?

The extent of genome rearrangement between the human and *Fugu* genomes reported in Chapter 3 (4,000-16,000 chromosomal rearrangement events) seems high in the context of reported values for other genome comparisons. Estimates for the number of rearrangement events since the divergence of the mouse and human genomes are consistently around 200 events (*e.g.*, Nadeau and Taylor, 1984; DeBry and Seldin, 1996). The estimates for human-mouse rearrangements differ from those for human-*Fugu* rearrangements by a factor of 20-80, but the time since the divergence of the bony fish lineage (about 450 Mya; Kumar and Hedges, 1998) is only 4-5 times longer than the time since the divergence of the rodent lineage (about 96 Mya; Nei *et al.*, 2001). These estimates imply a faster rate of genome rearrangement than seen in the mammalian lineage.

A tetraploid stage specific to the bony fish lineage (including *Fugu* and zebrafish) has been proposed on the basis of paralogous regions within the zebrafish genome that are present in single copy in mammalian genomes (Amores *et al.*, 1998; Gates *et al.*, 1999). Models of diploidisation (*i.e.*,

the change from chromosomal tetravalency to chromosomal bivalency) often invoke structural rearrangement of chromosomes to explain the changes in chromosomal affinities (*e.g.*, Allendorf and Thorgaard, 1984, and discussion in Chapter 1). If this model is accurate it may explain the apparent increase in the rate of genome rearrangement in the fish lineage as an artefact of diploidisation of the ancient tetraploid genome. Furthermore, there is evidence from polyploid plant genomes that recombination between homœologous (similar) chromosomes may be common, and that there may be increased transposable element activity in polyploid genomes (Song *et al.*, 1995; Wendel, 2000), which would contribute to the rearrangement of the genome.

This hypothesis predicts that the increased rate in genome rearrangement is fish lineage specific. This could be tested by examination of genome rearrangement between genomes that shared the polyploidy event. Genome rearrangement events may lead to speciation (White, 1978) so it reasonable to assume that different fish lineages that radiated after the genome duplication event (which probably took place 300-450 Mya; Taylor *et al.*, 2001a) will have fixed different genome rearrangements. This hypothesis predicts that comparisons among teleost fish such as *Fugu* and zebrafish should reveal a high rate of genome rearrangement, or, a burst of rearrangement associated with polyploidy, followed by a 'normalised' rate afterwards.

Similarly, the paralogous regions that we identified in the human genome (Chapter 4, *e.g.*, Figures 4.2 on page 86 and 4.3 on page 87) require a high rate of rearrangement if they are to be compatible with a hypothesis of block or whole genome duplication. The rearranged nature of paralogous regions found in this type of map-based analysis of the vertebrate genome has led to criticism of the conclusion that block duplication is a parsimonious explanation for the origin of these regions (Hughes, 1998; Hughes *et al.*, 2001).

5.2 A question of parsimony

One way in which the genome duplication hypothesis is more parsimonious than alternative hypotheses that explain the distribution of paralogues in the genome is in the number of words it takes to describe it, a fact which may be related to its popularity as a hypothesis. Austin Hughes has challenged the assumption that block duplication is the most parsimonious way to generate paralogous regions within a genome using his own parsimony statistic. He considers the relative parsimony of the hypothesis that paralogous regions were made by a block duplication event (perhaps as part of a whole genome duplication event) or the alternative hypothesis that they are the result of tandem duplication of genes followed by translocation (abbreviated here to the 'TDT model'; Hughes, 1998; Hughes *et al.*, 2001). Hughes found for both the *Hox* cluster regions and the 1/6/9/19 region that the TDT model was more parsimonious than the block duplication model as an explanation for the observed gene orders and phylogenetic trees. However, his reasoning may have been flawed as shown below.

Here we consider the simple case of a single genome duplication. The genome duplication hypothesis has an inbuilt disadvantage in the parsimony count method of Hughes (Hughes, 1998; Hughes *et al.*, 2001). Each gene that is no longer present in duplicate is counted as an individual deletion event (or equally, as a single translocation event removing it from the scope of detection as part of a paralogous region). By contrast, the method is very generous to the TDT model, assuming that a single translocation event brings each gene to its current position within a paralogous region. In fact, it can be shown that Hughes' TDT model will always require fewer events than a genome duplication model, so long as fewer than $\frac{1}{3}$ of genes are retained in duplicate.

Let G be the number of genes in the pre-duplication genome. Let x be the proportion of the pre-duplication genome retained in duplicate in the

modern genome, and y the proportion in single copy.

$$\text{Then } x + y = 1 \tag{5.1}$$

$$\text{and } Gx + Gy = G \tag{5.2}$$

Gx is the number of genes retained in duplicate.

Hughes' TDT model requires Gx tandem duplication events, and a further Gx translocation events, totalling $2Gx$ events. The whole genome duplication hypothesis requires one genome duplication event followed by Gy gene deletion events (the number of genes seen in single copy). For these two hypotheses to have an equal number of events (*i.e.*, to be equally parsimonious) then:

$$2Gx = 1 + Gy \tag{5.3}$$

Replace y with $1 - x$ from Equation 5.1:

$$\Rightarrow 2Gx = 1 + G(1 - x) \tag{5.4}$$

$$\Rightarrow x = \frac{1}{3G} + \frac{1}{3} \tag{5.5}$$

For genomes with a large number of genes (*e.g.*, $G = 6000$ for yeast):

$$\Rightarrow x \simeq \frac{1}{3} \tag{5.6}$$

Hughes' TDT model will be more parsimonious than the genome duplication model if $x < \frac{1}{3}$, *i.e.*, whenever the retention of genes in duplicate is less than $\frac{1}{3}$ of the pre-duplication genome.

According to the work of Walsh (1995) the retention of duplicate genes will be less than 33% for any effective population size (N_e) smaller than 250,000 assuming a selective advantage of advantageous alleles of 0.01, and

Table 5.1: The effect of population size on duplicate gene retention. Calculations are based on Equation 1.1 on page 21 (Walsh, 1995) with $\rho = 5 \times 10^{-5}$ and $s = 0.01$. N_e is the effective population size and $P(r)$ is the probability of retention of a duplicated gene.

$\mathbf{N_e}$	$\mathbf{P(r)}$
5,000	0.01
10,000	0.02
20,000	0.04
30,000	0.06
100,000	0.16
250,000	0.33
500,000	0.50

a ratio of advantageous to null mutation rate of 5×10^{-5}. This increases to 50% retention for N_e of 500,000 (Table 5.1). The difference between these values is within the margin of error of N_e estimation for vertebrates which may be as large as one order of magnitude (Hartl and Clark, 1997). Therefore Walsh's formula has little predictive value for estimating the amount of gene loss following any hypothesised genome duplication in an early vertebrate.

The above calculations were based on Hughes' assumption that genes are deleted individually, *i.e.*, that only one gene is deleted per deletion event. It may be more biologically realistic to allow for several neighbouring genes to be deleted in a single event. If d is the average number of genes deleted in a gene deletion event, then Equation 5.4 can be rephrased as:

$$2x = \frac{1}{G} + \frac{(1-x)}{d} \tag{5.7}$$

$$\Rightarrow 2x \simeq \frac{(1-x)}{d} \tag{5.8}$$

$$\Rightarrow d = \frac{1}{2x} - \frac{1}{2} \tag{5.9}$$

Solving Equation 5.9 for different values of x shows the average size of a deletion event that is required for the two hypotheses to have equal

Table 5.2: Crossover values for x (proportion of genes retained in duplicate) and d (average number of genes deleted in a single event) at which the whole genome duplication and TDT models are equally parsimonious[a]

x	d
0.33	1.0
0.20	2.0
0.10	4.5
0.08	5.8
0.05	9.5
0.01	49.5

[a] Calculated from Equation 5.9

probability for different frequencies of duplicate gene retention (Table 5.2).

One of the observations of the well-documented case of paleopolyploidy in yeast was that only 8% of the pre-duplication genome was retained in duplicate (Seoighe and Wolfe, 1998). For $x = 0.08$ the average size of a deletion event (d) needs to be 6 genes or larger (Table 5.2) to favour the genome duplication hypothesis by the simple parsimony statistic. Intuitively this seems like a biologically feasible size. Indeed it seems more acceptable than another assumption built-in to the alternative tandem-duplication and translocation model, *i.e.*, that selection can create genomic regions with similar gene contents by favouring particular translocations (Hughes, 1999a).

One of the blocks identified in the analysis of the yeast genome is shown in Figure 5.1 (Pohlmann and Philippsen, 1996; Wolfe and Shields, 1997). Within this block there are six duplicated genes and 20 unduplicated genes (eight on chromosome XIV and 12 on chromosome IX). Under Hughes' TDT model the formation of this block would require 12 steps (six tandem duplications, and six translocations). Under a whole genome duplication model, with each gene deleted individually, the formation of this block would

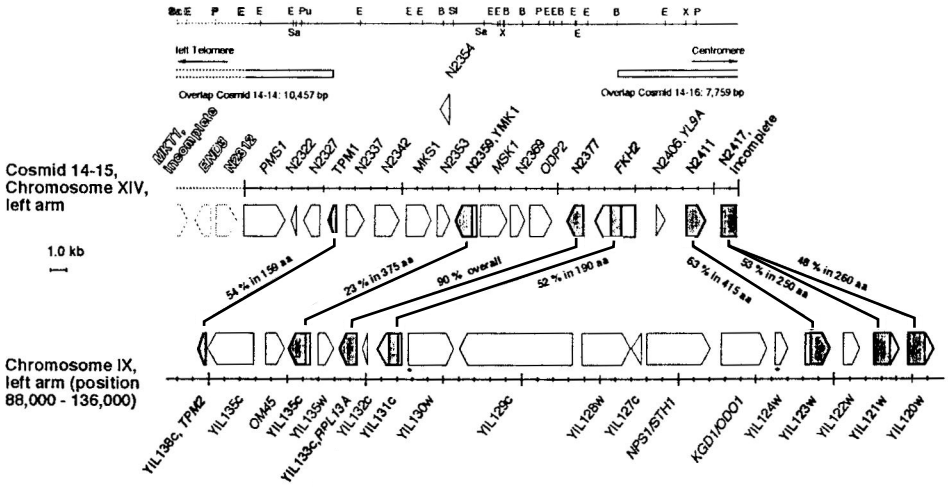


Figure 5.1: Example of a paralogous region between chromosomes XIV and IX identified in the yeast genome. This figure is taken from Pohlmann and Philippsen (1996). This block was labelled block 39 by Wolfe and Shields (1997).

require 21 events (one whole genome duplication, and 20 gene deletions) and would thus be less parsimonious by this statistic. However, if each deletion event included on average three genes then only seven deletion events would be required to explain the current state of this paralogous region, and the block duplication model would be more parsimonious.

Thus it appears that, even in the well-documented case of yeast which Hughes and colleagues agree is a likely polyploid (Friedman and Hughes, 2001), this simple parsimony statistic is not appropriate to determine the relative probability of paralogous region formation by block duplication or by tandem duplication and translocation.

Bibliography

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., George, R. A., Lewis, S. E., Richards, S., Ashburner, M., Henderson, S. N., Sutton, G. G., Wortman, J. R., Yandell, M. D., Zhang, Q., Chen, L. X., Brandon, R. C., Rogers, Y. H., Blazej, R. G., Champe, M., Pfeiffer, B. D., Wan, K. H., Doyle, C., Baxter, E. G., Helt, G., Nelson, C. R., Gabor Miklos, G. L., Abril, J. F., Agbayani, A., An, H. J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R. M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E. M., Beeson, K. Y., Benos, P. V., Berman, B. P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M. R., Bouck, J., Brokstein, P., Brottier, P., Burtis, K. C., Busam, D. A., Butler, H., Cadieu, E., Center, A., Chandra, I., Cherry, J. M., Cawley, S., Dahlke, C., Davenport, L. B., Davies, P., de Pablos, B., Delcher, A., Deng, Z., Mays, A. D., Dew, I., Dietz, S. M., Dodson, K., Doup, L. E., Downes, M., Dugan-Rocha, S., Dunkov, B. C., Dunn, P., Durbin, K. J., Evangelista, C. C., Ferraz, C., Ferriera, S., Fleischmann, W., Fosler, C., Gabrielian, A. E., Garg, N. S., Gelbart, W. M., Glasser, K., Glodek, A., Gong, F., Gorrell, J. H., Gu, Z., Guan, P., Harris, M., Harris, N. L., Harvey, D., Heiman, T. J., Hernandez, J. R., Houck, J., Hostin, D., Houston, K. A., Howland, T. J., Wei, M. H., Ibegwam, C., *et al.* (2000). The genome sequence of *Drosophila melanogaster*. *Science*, **287**(5461), 2185–2196.

Ahn, S. and Tanksley, S. D. (1993). Comparative linkage maps of the rice and maize genomes. *Proceedings of the National Academy of Sciences U.S.A.*, **90**, 7980–7984.

Allendorf, F. W. and Thorgaard, G. (1984). Tetraploidy and the evolution of salmonid fishes. In B. Turner, editor, *Evolutionary genetics of fishes*, pages 1–46. Plenum, New York.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**(17), 3389–3402.

Amores, A., Force, A., Yan, Y., Joly, L., Amemiya, C., Fritz, A., Ho, R., Langeland, J., Prince, V., Wang, Y., Westerfield, M., Ekker, M., and Postlethwait, J. (1998). Zebrafish Hox clusters and vertebrate genome evolution. *Science*, **282**(5394), 1711–1714.

Antequera, F. and Bird, A. (1993). Number of CpG islands and genes in human and mouse. *Proceedings of the National Academy of Sciences U.S.A.*, **90**(24), 11995–11999.

Aparicio, S., Hawker, K., Cottage, A., Mikawa, Y., Zuo, L., Venkatesh, B., Chen, E., Krumlauf, R., and Brenner, S. (1997). Organization of the *Fugu rubripes* Hox clusters: evidence for continuing evolution of vertebrate Hox complexes. *Nature Genetics*, **16**(1), 79–83.

Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**(6814), 796–815.

Armes, N., Gilley, J., and Fried, M. (1997). The comparative genomic structure and sequence of the surfeit gene homologs in the puffer fish *Fugu rubripes* and their association with CpG-rich islands. *Genome Research*, **7**(12), 1138–1152.

Ashburner, M., Misra, S., Roote, J., Lewis, S. E., Blazej, R., Davis, T., Doyle, C., Galle, R., George, R., Harris, N., Hartzell, G., Harvey, D., Hong, L., Houston, K., Hoskins, R., Johnson, G., Martin, C., Moshrefi, A., Palazzolo, M., Reese, M. G., Spradling, A., Tsang, G., Wan, K., Whitelaw, K., and Celniker, S. e.

(1999). An exploration of the sequence of a 2.9-Mb region of the genome of *Drosophila melanogaster*: the Adh region. *Genetics*, **153**(1), 179–219.

Avner, P., Bruls, T., Poras, I., Eley, L., Gas, S., Ruiz, P., Wiles, M. V., Sousa-Nunes, R., Kettleborough, R., Rana, A., Morissette, J., Bentley, L., Goldsworthy, M., Haynes, A., Herbert, E., Southam, L., Lehrach, H., Weissenbach, J., Manenti, G., Rodriguez-Tome, P., Beddington, R., Dunwoodie, S., and Cox, R. D. (2001). A radiation hybrid transcript map of the mouse genome. *Nature Genetics*, **29**(2), 194–200.

Bailey, W. J., Kim, J., Wagner, G. P., and Ruddle, F. H. (1997). Phylogenetic reconstruction of vertebrate Hox cluster duplications. *Molecular Biology and Evolution*, **14**(8), 843–853.

Barbazuk, W. B., Korf, I., Kadavi, C., Heyen, J., Tate, S., Wun, E., Bedell, J. A., McPherson, J. D., and Johnson, S. L. (2000). The syntenic relationship of the zebrafish and human genomes. *Genome Research*, **10**(9), 1351–1358.

Baxendale, S., Abdulla, S., Elgar, G., Buck, D., Berks, M., Micklem, G., Durbin, R., Bates, G., Brenner, S., and Beck, S. (1995). Comparative sequence analysis of the human and pufferfish Huntington's disease genes. *Nature Genetics*, **10**(1), 67–76.

Bernardi, G. (1989). The isochore organization of the human genome. *Annu Rev Genet*, **23**, 637–661.

Blanc, G., Barakat, A., Guyot, R., Cooke, R., and Delseny, M. (2000). Extensive duplication and reshuffling in the *Arabidopsis* genome. *Plant Cell*, **12**(7), 1093–1101.

Blanchette, M., Kunisawa, T., and Sankoff, D. (1996). Parametric genome rearrangement. *Gene*, **172**(1), GC11–7.

Bork, P., Sander, C., and Valencia, A. (1992). An ATPase domain common to prokaryotic cell cycle proteins, sugar kinases, actin, and hsp70 heat shock

proteins. *Proceedings of the National Academy of Sciences U.S.A.*, **89**(16), 7290–7294.

Brenner, S., Elgar, G., Sandford, R., Macrae, A., Venkatesh, B., and Aparicio, S. (1993). Characterisation of the pufferfish (*Fugu*) genome as a compact model vertebrate genome. *Nature*, **366**(6452), 265–268.

Brunner, B., Todt, T., Lenzner, S., Stout, K., Schulz, U., Ropers, H., and Kalscheuer, V. (1999). Genomic structure and comparative analysis of nine *Fugu* genes: conservation of synteny with human chromosome Xp22.2-p22.1. *Genome Research*, **9**(5), 437–448.

Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic dna. *Journal of Molecular Biology*, **268**(1), 78–94.

Burt, D. W., Bruley, C., Dunn, I. C., Jones, C. T., Ramage, A., Law, A. S., Morrice, D. R., Paton, I. R., Smith, J., Windsor, D., Sazanov, A., Fries, R., and Waddington, D. (1999). The dynamics of chromosome evolution in birds and mammals. *Nature*, **402**(6760), 411–413.

Carvalho, A. B. and Clark, A. G. (1999). Intron size and natural selection. *Nature*, **401**(6751), 344.

Chowdhary, B. P. and Raudsepp, T. (2000). HSA4 and GGA4: remarkable conservation despite 300-Myr divergence. *Genomics*, **64**(1), 102–105.

Comings, D. E. (1972). Evidence for ancient tetraploidy and conservation of linkage groups in mammalian chromosomes. *Nature*, **238**(5365), 455–457.

Cottage, A., Clark, M., Hawker, K., Umrania, Y., Wheller, D., Bishop, M., and Elgar, G. (1999). Three receptor genes for plasminogen related growth factors in the genome of the puffer fish *Fugu rubripes*. *FEBS Letters*, **443**(3), 370–374.

Crooijmans, R. P., Dijkhof, R. J., Veenendaal, T., van Der Poel, J. J., Nicholls, R. D., Bovenhuis, H., and Groenen, M. A. (2001). The gene orders on human chromosome 15 and chicken chromosome 10 reveal multiple inter- and

intrachromosomal rearrangements. *Molecular Biology and Evolution*, **18**(11), 2102–2109.

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., and Lander, E. S. (2001). High-resolution haplotype structure in the human genome. *Nature Genetics*, **29**(2), 229–232.

DeBry, R. W. and Seldin, M. F. (1996). Human/mouse homology relationships. *Genomics*, **33**(3), 337–351.

Deloukas, P., Schuler, G. D., Gyapay, G., Beasley, E. M., Soderland, C., Rodriguez-Tome, P., Hui, L., Matise, T. C., McKusick, K. B., Beckmann, J. S., Bentolila, S., Bihoreau, M.-T., Birren, B. B., Browne, J., Butler, A., Castle, A. B., Chiannilkulchai, N., Clee, C., Day, P. J. R., Dehejia, A., Dibling, T., Drouot, N., Duprat, S., Fizames, C., Fox, S., Gelling, S., Green, L., Harrison, P., Hocking, R., Holloway, E., Hunt, S., Keil, S., Linjnzaad, P., Louis-Dit-Sully, C., Ma, J., Mendis, A., Miller, J., Morissette, D., Muselet, D., Nusbaum, H. C., Peck, A., Rozen, S., Simon, D., Slonim, D. K., Staples, R., Stein, L. D., Stewart, E. A., Suchard, M. A., Thangarajah, T., Vega-Czarny, N., Webber, C., Wu, X., Auffray, J. C., Nomura, N., Sikela, J. M., Polymeropoulos, M. H., James, M. R., Lander, E. S., Hudson, T. J., Myers, R. M., Cox, D. R., Weissenbach, J., Boguski, M. S., and Bentley, D. R. (1998). A physical map of 30,000 human genes. *Science*, **282**(5389), 744–746.

Dunham, I., Shimizu, N., Roe, B. A., Chissoe, S., Hunt, A. R., Collins, J. E., Bruskiewich, R., Beare, D. M., Clamp, M., Smink, L. J., Ainscough, R., Almeida, J. P., Babbage, A., Bagguley, C., Bailey, J., Barlow, K., Bates, K. N., Beasley, O., Bird, C. P., Blakey, S., Bridgeman, A. M., Buck, D., Burgess, J., Burrill, W. D., O'Brien, K. P., and et al. (1999). The DNA sequence of human chromosome 22. *Nature*, **402**(6761), 489–495.

Duret, L., Mouchiroud, D., and Gautier, C. (1995). Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. *Journal of Molecular Evolution*, **40**(3), 308–317.

Ehrlich, J., Sankoff, D., and Nadeau, J. H. (1997). Synteny conservation and chromosome rearrangements during mammalian evolution. *Genetics*, **147**(1), 289–296.

El-Mabrouk, N. (2000). Recovery of ancestral tetraploids. In D. Sankoff and J. H. Nadeau, editors, *Comparative Genomics*, pages 465–477. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Elgar, G. (1996). Quality not quantity: the pufferfish genome. *Human Molecular Genetics*, **5**(Spec No), 1437–1442.

Elgar, G., Sandford, R., Aparicio, S., Macrae, A., Venkatesh, B., and Brenner, S. (1996). Small is beautiful: comparative genomics with the pufferfish (*Fugu rubripes*). *Trends in Genetics*, **12**(4), 145–150.

Elgar, G., Clark, M. S., Meek, S., Smith, S., Warner, S., Edwards, Y. J., Bouchireb, N., Cottage, A., Yeo, G. S., Umrania, Y., Williams, G., and Brenner, S. (1999). Generation and analysis of 25 Mb of genomic DNA from the pufferfish *Fugu rubripes* by sequence scanning. *Genome Research*, **9**(10), 960–971.

C. elegans Sequencing Consortium (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. The *C. elegans* Sequencing Consortium. *Science*, **282**(5396), 2012–2018.

Endo, T., Imanishi, T., Gojobori, T., and Inoko, H. (1997). Evolutionary significance of intra-genome duplications on human chromosomes. *Gene*, **205**(1-2), 19–27.

Eyre-Walker, A. (1993). Recombination and mammalian genome evolution. *Proceedings of the Royal Society of London. Series B: Biological sciences*, **252**(1335), 237–243.

Eyre-Walker, A. and Keightley, P. D. (1999). High genomic deleterious mutation rates in hominids. *Nature*, **397**(6717), 344–347.

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Systematic Zoology*, **19**(2), 99–113.

Fitch, W. M. (2000). Homology a personal view on some of the problems. *Trends in Genetics*, **16**(5), 227–231.

Flajnik, M. F. and Kasahara, M. (2001). Comparative genomics of the MHC: glimpses into the evolution of the adaptive immune system. *Immunity*, **15**(3), 351–362.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, **151**(4), 1531–1545.

Friedman, R. and Hughes, A. L. (2001). Gene duplication and the structure of eukaryotic genomes. *Genome Research*, **11**(3), 373–381.

Gallardo, M. H., Bickham, J. W., Honeycutt, R. L., A., O. R., and Kohler, N. (1999). Discovery of tetraploidy in a mammal. *Nature*, **401**(6751), 341.

Garcia-Fernandez, J. and Holland, P. W. (1994). Archetypal organization of the amphioxus Hox gene cluster. *Nature*, **370**(6490), 563–566.

Gates, M. A., Kim, L., Egan, E. S., Cardozo, T., Sirtokin, H. I., Dougan, S. T., Lashkari, D., Abagyan, R., Schier, A. F., and Talbot, W. S. (1999). A genetic linkage map for zebrafish: Comparative analysis and localization of genes and expressed sequences. *Genome Research*, **9**, 334–347.

Gaut, B. S. and Doebley, J. F. (1997). DNA sequence evidence for the segmental allotetraploid origin of maize. *Proceedings of the National Academy of Sciences U.S.A.*, **94**, 6809–6814.

Gellner, K. and Brenner, S. (1999). Analysis of 148 kb of genomic DNA around the wnt1 locus of *Fugu rubripes*. *Genome Research*, **9**(3), 251–258.

Gibson, T. J. and Spring, J. (1998). A model for massive genetic redundancy in vertebrates: polyploidy followed by persistence of genes encoding multidomain proteins. *Trends in Genetics*, **16**.

Gibson, T. J. and Spring, J. (2000). Evidence in favour of ancient octaploidy in the vertebrate genome. *Biochemical Society Transactions*, **28**(2), 259–264.

Gilley, J. and Fried, M. (1999). Extensive gene order differences within regions of conserved synteny between the *Fugu* and human genomes: implications for chromosomal evolution and the cloning of disease genes. *Human Molecular Genetics*, **8**(7), 1313–1320.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H., and Oliver, S. G. (1996). Life with 6000 genes. *Science*, **274**(5287), 546, 563–546, 567.

Gottgens, B., Gilbert, J., Barton, L., Aparicio, S., Hawker, K., Mistry, S., Vaudin, M., King, A., Bentley, D., Elgar, G., and Green, A. (1998). The pufferfish SLP-1 gene, a new member of the SCL/TAL-1 family of transcription factors. *Genomics*, **48**(1), 52–62.

Graur, D. and Higgins, D. G. (1994). Molecular evidence for the inclusion of cetaceans within the order artiodactyla. *Molecular Biology and Evolution*, **11**(3), 357–364.

Graves, J. A. (1996). Mammals that break the rules: genetics of marsupials and monotremes. *Annu Rev Genet*, **30**, 233–260.

Gu, X. and Zhang, J. (1997). A simple method for estimating the parameter of substitution rate variation among sites. *Molecular Biology and Evolution*, **14**(11), 1106–1113.

Gyapay, G., Schmitt, K., Fizames, C., Jones, H., Vega-Czarny, N., Spillett, D., Muselet, D., Prud'Homme, J. F., Dib, C., Auffray, C., Morissette, J., Weissenbach, J., and Goodfellow, P. N. (1996). A radiation hybrid map of the human genome. *Human Molecular Genetics*, **5**(3), 339–346.

Hartl, D. L. and Clark, A. G. (1997). *Principles of Population Genetics*. Sinaur Associates, Inc., Massachusetts.

Hogenesch, J. B., Ching, K. A., Batalov, S., Su, A. I., Walker, J. R., Zhou, Y., Kay, S. A., Schultz, P. G., and Cooke, M. P. (2001). A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell*, **106**(4), 413–415.

Hokamp, K. (2001). *A bioinformatics approach to (intra-)genome comparisons*. Ph.D. thesis, University of Dublin.

Holland, P. W. (1999). Gene duplication: past, present and future. *Seminars in Cell and Developmental Biology*, **10**(5), 541–547.

Holland, P. W., Garcia-Fernandez, J., Williams, N. A., and Sidow, A. (1994). Gene duplications and the origins of vertebrate development. *Development Supplement*, pages 125–133.

How, G., Venkatesh, B., and Brenner, S. (1996). Conserved linkage between the puffer fish (*Fugu rubripes*) and human genes for platelet-derived growth factor receptor and macrophage colony-stimulating factor receptor. *Genome Research*, **6**(12), 1185–1191.

Hudson, T. J., Church, D. M., Greenaway, S., Nguyen, H., Cook, A., Steen, R. G., Etten, W. J. V., Castle, A. B., Strivens, M. A., Trickett, P., Heuston, C., Davison, C., Southwell, A., Hardisty, R., Varela-Carver, A., Haynes, A. R., Rodriguez-Tome, P., Doi, H., Ko, M. S., Pontius, J., Schriml, L., Wagner, L., Maglott, D., Brown, S. D., Lander, E. S., Schuler, G., and Denny, P. (2001). A radiation hybrid map of mouse genes. *Nature Genetics*, **29**(2), 201–205.

Hughes, A. L. (1994). The evolution of functionally novel proteins after gene duplication. *Proceedings of the Royal Society of London. Series B: Biological sciences*, **256**(1346), 119–124.

Hughes, A. L. (1998). Phylogenetic tests of the hypothesis of block duplication of homologous genes on human chromosomes 6, 9, and 1. *Molecular Biology and Evolution*, **15**(7), 854–870.

Hughes, A. L. (1999a). *Adaptive Evolution of Genes and Genomes*. Oxford University Press, New York.

Hughes, A. L. (1999b). Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history. *Journal of Molecular Evolution*, **48**(5), 565–576.

Hughes, A. L., da Silva, J., and Friedman, R. (2001). Ancient genome duplications did not structure the human Hox-bearing chromosomes. *Genome Research*, **11**(5), 771–780.

Hughes, M. K. and Hughes, A. L. (1993). Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*. *Molecular Biology and Evolution*, **10**(6), 1360–1369.

Hurst, L. D., Brunton, C. F., and Smith, N. G. (1999). Small introns tend to occur in GC-rich regions in some but not all vertebrates. *Trends in Genetics*, **15**(11), 437–439.

Ikemura, T. and Wada, K. (1991). Evident diversity of codon usage patterns of human genes with respect to chromosome banding patterns and chromosome numbers; relation between nucleotide sequence data and cytogenetic data. *Nucleic Acids Research*, **19**(16), 4333–4339.

Imanishi, T., Endo, T., and Gojobori, T. (1997). An exhaustive search for extensive chromosomal regions duplicated within the human genome. *HGM '97 Poster (March 1997, Toronto, Canada). http://www.cib.nig.ac.jp/dda/timanish/dup.html*.

Kappen, C., Schughart, K., and Ruddle, F. H. (1989). Two steps in the evolution of Antennapedia-class vertebrate homeobox genes. *Proceedings of the National Academy of Sciences U.S.A.*, **86**(14), 5459–5463.

Kasahara, M., Hayashi, M., Tanaka, K., Inoko, H., Sugaya, K., Ikemura, T., and Ishibashi, T. (1996). Chromosomal localization of the proteasome Z subunit gene reveals an ancient chromosomal duplication involving the major

histocompatibility complex. *Proceedings of the National Academy of Sciences U.S.A.*, **93**(17), 9096–9101.

Kasahara, M., Nakaya, J., Satta, Y., and Takahata, N. (1997). Chromosomal duplication and the emergence of the adaptive immune system. *Trends in Genetics*, **13**(3), 90–92.

Katsanis, N., Fitzgibbon, J., and 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**(1), 101–108.

Keller, B. and Feuillet, C. (2000). Colinearity and gene density in grass genomes. *Trends in Plant Science*, **5**(6), 246–251.

Kenmochi, N., Kawaguchi, T., Rozen, S., Davis, E., Goodman, N., Hudson, T. J., Tanaka, T., and Page, D. C. (1998). A map of 75 human ribosomal protein genes. *Genome Research*, **8**(5), 509–523.

Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge.

Kojima, S., Itoh, Y., Matsumoto, S., Masuho, Y., and Seiki, M. (2000). Membrane-type 6 matrix metalloproteinase (MT6-MMP, MMP-25) is the second glycosyl-phosphatidyl inositol (GPI)-anchored MMP. *FEBS Letters*, **480**(2-3), 142–146.

Koop, B. F. (1995). Human and rodent DNA sequence comparisons: a mosaic model of genomic evolution. *Trends in Genetics*, **11**(9), 367–371.

Ku, H. M., Vision, T., Liu, J., and Tanksley, S. D. (2000). Comparing sequenced segments of the tomato and arabidopsis genomes: large-scale duplication followed by selective gene loss creates a network of synteny. *Proceedings of the National Academy of Sciences U.S.A.*, **97**(16), 9121–9126.

Kumar, S. and Hedges, S. B. (1998). A molecular timescale for vertebrate evolution. *Nature*, **392**(6679), 917–920.

Kumar, S., Gadagkar, S. R., Filipski, A., and Gu, X. (2001). Determination of the number of conserved chromosomal segments between species. *Genetics*, **157**(3), 1387–1395.

Lahn, B. T. and Page, D. C. (1999). Four evolutionary strata on the human X chromosome. *Science*, **286**(5441), 964–967.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S.,

Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., and Szustakowki, J. (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**(6822), 860–921.

Lang, A. J., Mirski, S. E., Cummings, H. J., Yu, Q., Gerlach, J. H., and Cole, S. P. (1998). Structural organization of the human TOP2A and TOP2B genes. *Gene*, **221**(2), 255–266.

Levinson, B., Kenwrick, S., Lakich, D., r. Hammonds G, J., and Gitschier, J. (1990). A transcribed gene in an intron of the human factor VIII gene. *Genomics*, **7**(1), 1–11.

Li, W. H. (1980). Rate of gene silencing at duplicate loci: a theoretical study and interpretation of data from tetraploid fishes. *Genetics*, **95**(1), 237–258.

Li, W. H. (1993). So, what about the molecular clock hypothesis? *Current Opinion in Genetics and Development*, **3**(6), 896–901.

Li, W.-H. (1997). *Molecular Evolution*. Sinauer Associates, Inc., Sunderland.

Li, W. H., Ellsworth, D. L., Krushkal, J., Chang, B. H., and Hewett-Emmett, D. (1996). Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. *Molecular Phylogenetics and Evolution*, **5**(1), 182–187.

Lin, X., Kaul, S., Rounsley, S., Shea, T. P., Benito, M. I., Town, C. D., Fujii, C. Y., Mason, T., Bowman, C. L., Barnstead, M., Feldblyum, T. V., Buell, C. R., Ketchum, K. A., Lee, J., Ronning, C. M., Koo, H. L., Moffat, K. S., Cronin, L. A., Shen, M., Pai, G., Van Aken, S., Umayam, L., Tallon, L. J., Gill, J. E., Venter, J. C., and et al. (1999). Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature*, **402**(6763), 761–768.

Lindsay, E. A., Botta, A., Jurecic, V., Carattini-Rivera, S., Cheah, Y. C., Rosenblatt, H. M., Bradley, A., and Baldini, A. (1999). Congenital heart disease in mice deficient for the DiGeorge syndrome region. *Nature*, **401**(6751), 379–383.

Llorente, B., Malpertuy, A., Neuveglise, C., de Montigny, J., Aigle, M., Artiguenave, F., Blandin, G., Bolotin-Fukuhara, M., Bon, E., Brottier, P., Casaregola, S., Durrens, P., Gaillardin, C., Lepingle, A., Ozier-Kalogeropoulos, O., Potier, S., Saurin, W., Tekaia, F., Toffano-Nioche, C., Wesolowski-Louvel, M., Wincker, P., Weissenbach, J., Souciet, J., and Dujon, B. (2000). Genomic exploration of the hemiascomycetous yeasts: 18. comparative analysis of chromosome maps and synteny with *Saccharomyces cerevisiae*. *FEBS Letters*, **487**(1), 101–112.

Loftus, B. J., Kim, U. J., Sneddon, V. P., Kalush, F., Brandon, R., Fuhrmann, J., Mason, T., Crosby, M. L., Barnstead, M., Cronin, L., Deslattes Mays, A., Cao, Y., Xu, R. X., Kang, H. L., Mitchell, S., Eichler, E. E., Harris, P. C., Venter, J. C., and Adams, M. D. (1999). Genome duplications and other features in 12 Mb of DNA sequence from human chromosome 16p and 16q. *Genomics*, **60**(3), 295–308.

Long, M. and Thornton, K. (2001). Gene duplication and evolution. *Science*, **293**(5535), 1551.

Lucassen, A. M., Julier, C., Beressi, J. P., Boitard, C., Froguel, P., Lathrop, M., and Bell, J. I. (1993). Susceptibility to insulin dependent diabetes mellitus maps to a 4.1 kb segment of DNA spanning the insulin gene and associated VNTR. *Nature Genetics*, **4**(3), 305–310.

Lundin, L. G. (1993). Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. *Genomics*, **16**(1), 1–19.

Lynch, M. and Conery, J. C. (2001). Gene duplication and evolution. *Science*, **293**(5535), 1551.

Lynch, M. and Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science*, **290**(5494), 1151–1155.

Margoliash, E. (1963). Primary structure and evolution of cytochrome c. *Proceedings of the National Academy of Sciences U.S.A.*, **50**, 672–679.

Martin, A. (2001). Is tetralogy true? lack of support for the "one-to-four" rule. *Molecular Biology and Evolution*, **18**(1), 89–93.

Martin, A. P. (1999). Increasing genomic complexity by gene duplication and the origin of vertebrates. *American Naturalist*, **154**, 111–128.

Martin, G. R., Richman, M., Reinsch, S., Nadeau, J. H., and Joyner, A. (1990). Mapping of the two mouse engrailed-like genes: close linkage of En-1 to dominant hemimelia (Dh) on chromosome 1 and of En-2 to hemimelic extra-toes (Hx) on chromosome 5. *Genomics*, **6**(2), 302–308.

Martinez-Perez, E., Shaw, P., and Moore, G. (2001). The Ph1 locus is needed to ensure specific somatic and meiotic centromere association. *Nature*, **411**(6834), 204–207.

Mayer, K., Schuller, C., Wambutt, R., Murphy, G., Volckaert, G., Pohl, T., Dusterhoft, A., Stiekema, W., Entian, K. D., Terryn, N., Harris, B., Ansorge,

W., Brandt, P., Grivell, L., Rieger, M., Weichselgartner, M., de Simone, V., Obermaier, B., Mache, R., Muller, M., Kreis, M., Delseny, M., Puigdomenech, P., Watson, M., McCombie, W. R., and et al. (1999). Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. *Nature*, **402**(6763), 769–777.

McCarthy, L. C. (1996). Whole genome radiation hybrid mapping. *Trends in Genetics*, **12**(12), 491–493.

McLysaght, A., Enright, A. J., Skrabanek, L., and Wolfe, K. H. (2000a). Estimation of synteny conservation and genome compaction between pufferfish (*Fugu*) and human. *Yeast*, **17**(1), 22–36.

McLysaght, A., Seoighe, C., and Wolfe, K. (2000b). High frequency of inversions during eukaryote gene order evolution. In D. Sankoff and J. H. Nadeau, editors, *Comparative Genomics*, pages 47–58. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Meyer, A. and 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. *Current Opinion in Cell Biology*, **11**(6), 699–704.

Meyerowitz, E. M. (2001). Prehistory and history of *Arabidopsis* research. *Plant Physiology*, **125**(1), 15–19.

Miles, C., Elgar, G., Coles, E., Kleinjan, D., van Heyningen, V., and Hastie, N. (1998). Complete sequencing of the *Fugu* wagr region from WT1 to PAX6: Dramatic compaction and conservation of synteny with human chromosome 11p13. *Proceedings of the National Academy of Sciences U.S.A.*, **95**(22), 13068–13072.

Muller, H. J. (1925). Why polyploidy is rarer in animals than in plants. *American Naturalist*, **9**, 346–353.

Nadeau, J. and Taylor, B. (1984). Lengths of chromosomal segments conserved since divergence of man and mouse. *Proceedings of the National Academy of Sciences U.S.A.*, **81**, 814–818.

Nadeau, J. H. (1989). Maps of linkage and synteny homologies between mouse and man. *Trends in Genetics*, **5**(3), 82–86.

Nadeau, J. H. and Kosowsky, M. (1991). Mouse map of paralogous genes. *Mammalian Genome*, **1 Spec No**, S433–S460.

Nadeau, J. H. and Sankoff, D. (1998). Counting on comparative maps. *Trends in Genetics*, **14**(12), 495–501.

Nanda, I., Shan, Z., Schartl, M., Burt, D. W., Koehler, M., Nothwang, H., Grutzner, F., Paton, I. R., Windsor, D., Dunn, I., Engel, W., Staeheli, P., Mizuno, S., Haaf, T., and Schmid, M. (1999). 300 million years of conserved synteny between chicken Z and human chromosome 9. *Nature Genetics*, **21**(3), 258–259.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, **48**(3), 443–453.

Nei, M. (1987). *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Nei, M., Xu, P., and Glazko, G. (2001). Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. *Proceedings of the National Academy of Sciences U.S.A.*, **98**(5), 2497–2502.

Newell, W., Beck, S., Lehrach, H., and Lyall, A. (1998). Estimation of distances and map construction using radiation hybrids. *Genome Research*, **8**(5), 493–508.

Nordborg, M., Charlesworth, B., and Charlesworth, D. (1996). The effect of recombination on background selection. *Genetical Research*, **67**(2), 159–174.

Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, **302**(1), 205–217.

Oeltjen, J. C., Malley, T. M., Muzny, D. M., Miller, W., Gibbs, R. A., and Belmont, J. W. (1997). Large-scale comparative sequence analysis of the human and murine Bruton's tyrosine kinase loci reveals conserved regulatory domains. *Genome Research*, **7**(4), 315–329.

Ohno, S. (1970). *Evolution by Gene Duplication*. George Allen and Unwin, London.

Ohno, S. (1985). Dispensable genes. *Trends in Genetics*, **1**, 160–164.

Ohno, S. (1999). The one-to-four rule and paralogues of sex-determining genes. *Cellular and Molecular Life Sciences*, **55**(6-7), 824–830.

Page, R. D. M. and Holmes, E. C. (1998). *Molecular Evolution - A Phylogenetic Approach*. Blackwell Science, Oxford.

Paterson, A. H., Bowers, J. E., Burow, M. D., Draye, X., Elsik, C. G., Jiang, C. X., Katsar, C. S., Lan, T. H., Lin, Y. R., Ming, R., and Wright, R. J. (2000). Comparative genomics of plant chromosomes. *Plant Cell*, **12**(9), 1523–1540.

Pebusque, M. J., Coulier, F., Birnbaum, D., and Pontarotti, P. (1998). Ancient large-scale genome duplications: phylogenetic and linkage analyses shed light on chordate genome evolution. *Molecular Biology and Evolution*, **15**(9), 1145–1159.

Pohlmann, R. and Philippsen, P. (1996). Sequencing a cosmid clone of *Saccharomyces cerevisiae* chromosome XIV reveals 12 new open reading frames (ORFs) and an ancient duplication of six orfs. *Yeast*, **12**(4), 391–402.

Postlethwait, J. H., Yan, Y. L., Gates, M. A., Horne, S., Amores, A., Brownlie, A., Donovan, A., Egan, E. S., Force, A., Gong, Z., Goutel, C., Fritz, A., Kelsh, R., Knapik, E., Liao, E., Paw, B., Ransom, D., Singer, A., Thomson, M., Abduljabbar, T. S., Yelick, P., Beier, D., Joly, J. S., Larhammar, D., Rosa, F., and et al. (1998). Vertebrate genome evolution and the zebrafish gene map. *Nature Genetics*, **18**(4), 345–349.

Postlethwait, J. H., Woods, I. G., Ngo-Hazelett, P., Yan, Y. L., Kelly, P. D., Chu, F., Huang, H., Hill-Force, A., and Talbot, W. S. (2000). Zebrafish comparative

genomics and the origins of vertebrate chromosomes. *Genome Research*, **10**(12), 1890–1902.

Poulter, R. and Butler, M. (1998). A retrotransposon family from the pufferfish (fugu) *Fugu rubripes*. *Gene*, **215**(2), 241–249.

Reboul, J., Gardiner, K., Monneron, D., Uze, G., and Lutfalla, G. (1999). Comparative genomic analysis of the Interferon/Interleukin-10 receptor gene cluster. *Genome Research*, **9**, 242–250.

Reeck, G. R., de Haen, C., Teller, D. C., Doolittle, R. F., Fitch, W. M., Dickerson, R. E., Chambon, P., D., M. A., Margolias, E., and Jukes, T. H. *et al.*. (1987). "Homology" in proteins and nucleic acids: a terminology muddle and a way out of it. *Cell*, **50**(5), 667.

Riley, R. and Kempanna, C. (1963). The homœologous nature of the non-homologous meiotic pairing in *Triticum æstivum* deficient for chromosome V (5B). *Heredity*, **18**, 287–306.

Rowen, L., Mahairas, G., and Hood, L. (1997). Sequencing the human genome. *Science*, **278**(5338), 605–607.

Ruddle, F. H., Bentley, K. L., Murtha, M. T., and Risch, N. (1994). Gene loss and gain in the evolution of the vertebrates. *Development Supplement*, pages 155–161.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**(4), 406–425.

Sandford, R., Sgotto, B., Burin, T., and Brenner, S. (1996). The tuberin (TSC2), autosomal dominant polycystic kidney disease (PKD1), and somatostatin type V receptor (SSTR5) genes form a synteny group in the *Fugu* genome. *Genomics*, **38**, 84–86.

Sandford, R., Sgotto, B., Aparicio, S., Brenner, S., Vaudin, M., Wilson, R., Chissoe, S., Pepin, K., Bateman, A., Chothia, C., Hughes, J., and Harris, P. (1997). Comparative analysis of the polycystic kidney disease 1 (PKD1) gene reveals an integral membrane glycoprotein with multiple evolutionary conserved domains. *Human Molecular Genetics*, **6**(9), 1483–1489.

Sankoff, D., Parent, M.-N., and Bryant, D. (2000). Accuracy and robustness of analyses based on numbers of genes in observed segments. In D. Sankoff and J. H. Nadeau, editors, *Comparative Genomics*, pages 299–306. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Sarich, V. M. and Wilson, A. C. (1967). Immunological time scale for hominid evolution. *Science*, **158**(805), 1200–1203.

Sarich, V. M. and Wilson, A. C. (1973). Generation time and genomic evolution in primates. *Science*, **179**(78), 1144–1147.

Schofield, J., Elgar, G., Greystrong, J., Lye, G., Deadman, R., Micklem, G., King, A., Brenner, S., and Vaudin, M. (1997). Regions of human chromosome 2 (2q32-q35) and mouse chromosome 1 show synteny with the pufferfish genome (*Fugu rubripes*). *Genomics*, **45**(1), 158–167.

Schuler, G. D. (1997). Sequence mapping by electronic PCR. *Genome Research*, **7**(5), 541–550.

Semple, C. and Wolfe, K. H. (1999). Gene duplication and gene conversion in the caenorhabditis elegans genome. *Journal of Molecular Evolution*, **48**(5), 555–564.

Seoighe, C. and Wolfe, K. H. (1998). Extent of genomic rearrangement after genome duplication in yeast. *Proceedings of the National Academy of Sciences U.S.A.*, **95**, 4447–4452.

Seoighe, C. and Wolfe, K. H. (1999a). Updated map of duplicated regions in the yeast genome. *Gene*, **238**(1), 253–261.

Seoighe, C. and Wolfe, K. H. (1999b). Yeast genome evolution in the post-genome era. *Current Opinion in Microbiology*, **2**(5), 548–554.

Seoighe, C., Federspiel, N., Jones, T., Hansen, N., Bivolarovic, V., Surzycki, R., Tamse, R., Komp, C., Huizar, L., Davis, R. W., Scherer, S., Tait, E., Shaw, D. J., Harris, D., Murphy, L., Oliver, K., Taylor, K., Rajandream, M. A., Barrell, B. G., and Wolfe, K. H. (2000). Prevalence of small inversions in yeast gene order evolution. *Proceedings of the National Academy of Sciences U.S.A.*, **97**(26), 14433–14437.

Sharman, A. C. (1999). Some new terms for duplicated genes. *Seminars in Cell and Developmental Biology*, **10**(5), 561–563.

Shimamura, M., Yasue, H., Ohshima, K., Abe, H., Kato, H., Kishiro, T., Goto, M., Munechika, I., and Okada, N. (1997). Molecular evidence from retroposons that whales form a clade within even-toed ungulates. *Nature*, **388**(6643), 666–670.

Sidow, A. (1996). Gen(om)e duplications in the evolution of early vertebrates. *Current Opinion in Genetics and Development*, **6**(6), 715–722.

Skrabanek, L. and Wolfe, K. H. (1998). Eukaryote genome duplication - where's the evidence? *Current Opinion in Genetics and Development*, **8**(6), 694–700.

Smith, N. G., Knight, R., and Hurst, L. D. (1999). Vertebrate genome evolution: a slow shuffle or a big bang? *Bioessays*, **21**(8), 697–703.

Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, **147**(1), 195–197.

Solovyev, V. V. and Salamov, A. A. (1999). Infogene: a database of known gene structures and predicted genes and proteins in sequences of genome sequencing projects. *Nucleic Acids Research*, **27**(1), 248–250.

Song, K., Lu, P., Tang, K., and Osborn, T. C. (1995). Rapid genome change in synthetic polyploids of brassica and its implications for polyploid evolution. *Proceedings of the National Academy of Sciences U.S.A.*, **92**(17), 7719–7723.

Spieth, J., Brooke, G., Kuersten, S., Lea, K., and Blumenthal, T. (1993). Operons in *C. elegans*: polycistronic mRNA precursors are processed by trans-splicing of SL2 to downstream coding regions. *Cell*, **73**(3), 521–532.

Spring, J. (1997). Vertebrate evolution by interspecific hybridisation–are we polyploid? *FEBS Letters*, **400**(1), 2–8.

Takezaki, N., Rzhetsky, A., and Nei, M. (1995). Phylogenetic test of the molecular clock and linearized trees. *Molecular Biology and Evolution*, **12**(5), 823–833.

Taylor, J. S., Van de Peer, Y., Braasch, I., and Meyer, A. (2001a). Comparative genomics provides evidence for an ancient genome duplication event in fish. *Philosophical Transactions of the Royal Society of London. Series B: Biological sciences*, **356**(1414), 1661–1679.

Taylor, J. S., Van de Peer, Y., and Meyer, A. (2001b). Genome duplication, divergent resolution and speciation. *Trends in Genetics*, **17**(6), 299–301.

Terryn, N., Heijnen, L., Keyser, A. D., Asseldonck, M. V., Clercq, R. D., Verbakel, H., Gielen, J., Zabeau, M., Villarroel, R., Jesse, T., Neyt, P., Hogers, R., Daele, H. V. D., Ardiles, W., Schueller, C., Mayer, K., Dehais, P., Rombauts, S., Montagu, M. V., Rouze, P., and Vos, P. (1999). Evidence for an ancient chromosomal duplication in *Arabidopsis thaliana* by sequencing and analyzing a 400-kb contig at the APETALA2 locus on chromosome 4. *FEBS Letters*, **445**(2-3), 237–245.

Tomsig, J. L. and Creutz, C. E. (2000). Biochemical characterization of copine: a ubiquitous Ca2+-dependent, phospholipid-binding protein. *Biochemistry*, **39**(51), 16163–16175.

Trower, M. K., Orton, S. M., Purvis, I. J., Sanseau, P., Riley, J., Christodoulou, C., Burt, D., See, C. G., Elgar, G., Sherrington, R., Rogaev, E. I., St. George-Hyslop, P., Brenner, S., and Dykes, C. W. (1996). Conservation of synteny between the genome of the pufferfish (*Fugu rubripes*) and the region on human chromosome 14 (14q24.3) associated with familial Alzheimer disease (AD3 locus). *Proceedings of the National Academy of Sciences U.S.A.*, **93**, 1366.

Uzzell, T. and Corbin, K. W. (1971). Fitting discrete probability distributions to evolutionary events. *Science*, **172**(988), 1089–1096.

Van de Peer, Y., Taylor, J. S., Braasch, I., and Meyer, A. (2001). The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. *Journal of Molecular Evolution*, **53**(4-5), 436–446.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Miklos, G. L. G., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Francesco, V. D., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C.,

Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigo, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001). The sequence of the human genome. *Science*, **291**(5507), 1304–1351.

Villard, L., Tassone, F., Crnogorac-Jurcevic, T., Clancy, K., and Gardiner, K. (1998). Analysis of pufferfish homologues of the AT-rich human APP gene. *Gene*, **210**, 17–24.

Vision, T. J., Brown, D. G., and Tanksley, S. D. (2000). The origins of genomic duplications in *Arabidopsis*. *Science*, **290**(5499), 2114–2117.

Walsh, J. B. (1995). How often do duplicated genes evolve new functions? *Genetics*, **139**, 421–428.

Wang, D. Y., Kumar, S., and Hedges, S. B. (1999). Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. *Proceedings of the Royal Society of London. Series B: Biological sciences*, **266**(1415), 163–171.

Wang, Y. and Gu, X. (2000). Evolutionary patterns of gene families generated in the early stage of vertebrates. *Journal of Molecular Evolution*, **51**(1), 88–96.

Watanabe, T. K., Bihoreau, M. T., McCarthy, L. C., Kiguwa, S. L., Hishigaki, H., Tsuji, A., Browne, J., Yamasaki, Y., Mizoguchi-Miyakita, A., Oga, K., Ono, T., Okuno, S., Kanemoto, N., Takahashi, E., Tomita, K., Hayashi, H., Adachi, M., Webber, C., Davis, M., Kiel, S., Knights, C., Smith, A., Critcher, R., Miller, J., and M. R. James, e. a. (1999). A radiation hybrid map of the rat genome containing 5,255 markers. *Nature Genetics*, **22**(1), 27–36.

Watkins-Chow, D. E., Buckwalter, M. S., Newhouse, M. M., Lossie, A. C., Brinkmeier, M. L., and Camper, S. A. (1997). Genetic mapping of 21 genes on mouse chromosome 11 reveals disruptions in linkage conservation with human chromosome 5. *Genomics*, **40**(1), 114–122.

Wendel, J. F. (2000). Genome evolution in polyploids. *Plant Molecular Biology*, **42**(1), 225–249.

White, M. J. D. (1978). *Modes of Speciation*. W. H. Freeman and Company, San Francisco.

Wolfe, K. (2000). Robustness–it's not where you think it is. *Nature Genetics*, **25**(1), 3–4.

Wolfe, K. H. (2001). Yesterday's polyploids and the mystery of diploidization. *Nature Reviews Genetics*, **2**(5), 333–341.

Wolfe, K. H. and Shields, D. C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, **387**, 708–713.

Wootton, J. and Federhen, S. (1996). Analysis of compositionally biased regions in sequence databases. *Methods in Enzymology*, **266**, 554–571.

Xu, W., Edmondson, D. G., Evrard, Y. A., Wakamiya, M., Behringer, R. R., and Roth, S. Y. (2000). Loss of Gcn5l2 leads to increased apoptosis and mesodermal defects during mouse development. *Nature Genetics*, **26**(2), 229–232.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences*, **13**(5), 555–556.

Zhang, J. and Nei, M. (1996). Evolution of Antennapedia-class homeobox genes. *Genetics*, **142**(1), 295–303.

Zhang, L., Gaut, B. S., and Vision, T. J. (2001). Gene duplication and evolution. *Science*, **293**(5535), 1551.

Zharkikh, A. and Li, W.-H. (1993). Inconsistency of the maximum-parsimony method: the case of five taxa with a molecular clock. *Systematic Biology*, **42**(2), 113–125.

Zuckerkandl, E. and Pauling, L. (1962). Molecular disease, evolution, and genetic heterogeneity. In M. Kasha and B. Pullman, editors, *Horizons in Biochemistry*, pages 97–166. Academic Press, New York.

Zuckerkandl, E. and Pauling, L. (1965). Molecules as documents of evolutionary history. *Journal of Theoretical Biology*, **8**(2), 357–366.