Testing the hypothesis of genome duplications in vertebrates

Lucy Skrabanek

Ph.D. thesis

Trinity College, Dublin

October, 1999

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 previously as an exercise for a degree at this or any other university.

Apart from the advice and assistance mentioned in the acknowledgments and in the text, this thesis is entirely my own work. I agree that the library may lend or copy this thesis freely on request.

Lucy Skrabanek

Lucy Skrabanek

October, 1999

ACKNOWLEDGMENTS

I owe a particular debt of gratitude to my supervisor, Ken Wolfe, for his patience, criticism, advice and encouragement throughout the duration of this project. Thanks are also due to the other members of the lab, Amanda Lohan, Sarah Goulding, Colin Semple, Aoife McLysaght, Karsten Hokamp and Mike McLean, but most especially to Cathal Seoighe for his comments and our ever-useful discussions. Thanks are also due to Denis Shields and Andrew Lloyd for computer troubleshooting, Jinghui Zhang at NCBI for help with downloading sequences from the 1997 NCBI Genomes Division, Yasuo Ina and Wojtiech Makałowski for advice about synonymous and non-synonymous substitution rates, Nick Gray for help with the equations in Chapter 5 and Feargal Murphy for the use of his G3 computer. Also thanks to my family for continual support and for proofreading the final draft of this thesis.

I owe all of these people a great debt, although the responsibility for all errors, omissions and opinions is entirely my own.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

FIGURES:

TABLES:

SUMMARY

The hypothesis that the human genome has undergone at least two rounds of genome duplication has become widely accepted, but has never been rigorously tested. In this study, several map-based inter- and intra-species comparisons and simulations were used to critially examine this hypothesis. Interspecific map comparisons between human and mouse were thought to be the most likely to yield useful information. Discovery of conserved duplicated regions between human and mouse would allow the estimation of the date of divergence of the paralogous genes, and therefore also the date of the last genome duplication. But disappointingly, we observed very few such regions. Human-human intraspecies comparisons also yielded very little information. We demonstrated that most paralogous regions within the human genome, defined by genes separated by more than about 2 - 10 cM, are likely to be artefacts. It has also been shown that a simple model of evolution by two genome duplications (and no other duplication events) is likely to be misleading. A system of filtering output from TBLASTX to find more biologically significant matches by comparing the rates of synonymous versus non-synonymous rates of nucleotide substitution is presented.

The work presented here shows that the amount of both sequencing and map data currently available is not sufficient to be able to draw any conclusions about how many genome duplications occurred in the evolution of the human genome, nor when they may have occurred. It is also indicated that the value of the discovery of paralogous regions within the human genome is limited in its ability to aid in the resolution of this hypothesis. It is suggested that a combination of interspecies comparisons and simulations may be the best way of determining the history of the evolution of the human genome.

ABBREVIATIONS

BLAST	basic local alignment search tool
cM	centimorgan
DCR	duplicated conserved region
DNA	deoxyribonucleic acid
EST	expressed sequence tag
HSA	*Homo sapiens*
HSP	high scoring pair
kb	kilobases
Mb	megabases
MHC	major histocompatibility complex
MMU	*Mus musculus*
Mya	million years ago
Myr	million years
NCBI	National Center for BioInformatics
ORF	open reading frame
PG	paralogy group
RNA	ribonucleic acid

CHAPTER 1 —

GENERAL INTRODUCTION

1.1 EVOLUTIONARY IMPORTANCE OF GENE DUPLICATION

Major evolutionary change is made possible by the acquisition of genes with new functions. Any mutation, resulting in a new function or specificity in a gene which occurs only once in the genome and which codes for a protein with a function essential to the organism and whose loss would be deleterious, cannot be fixed unless the original function can also be maintained (Ohno, 1970). Gene duplication, followed by functional divergence of the two daughter genes, is one class of mutation that allows the evolution of novel biological functions and permits major evolutionary change. The ubiquity of multigene families is evidence of the frequent occurrence of gene duplications.

1.1.1 CO-EVOLUTION OF FUNCTIONALLY RELATED FAMILIES OF GENES

Gene duplicates in functionally related gene families often show similarities in divergence dates, functional specificities and phylogenetic tree topologies, implying that such gene families co-evolved (Fryxell, 1996). Examples of such a relationship include the insulin-NGF family, and the interleukin-8 family. Two of the four insulin-like genes in vertebrates are insulin and IGF-1 (insulin-like growth factor 1), both of which are related to the nerve growth factor, NGF. Insulin binds to the InsR receptor (IGF-1R), which is related to the IGF-1 receptor. These associations form two different pathways: the insulin pathway, which regulates energy metabolism, and the IGF-1 pathway, which regulates embryonic growth. NGF is a member of the neurotrophic polypeptide hormone family, which includes NT3, NT4 and BDNF. All these genes have receptors that are also related to the insulin and IGF receptors (references in Fryxell, 1996). Such correlations, together with the high rate of spontaneous gene duplication (e.g., 10^{-3} to 10^{-5} / locus / generation in bacteria (Anderson and Roth, 1977) or 10^{-9} / base pair / year in vertebrates (Ohno, 1985)), suggest that the co-evolution of functionally related gene families has been a rather general phenomenon, and that the rate-limiting step in the evolutionary process of gene duplication and divergence is not the duplication rate itself (Fryxell, 1996).

1.1.2 GENE NETWORKS

Gene duplication is particularly attractive from the theoretical standpoint of duplicating entire genetic networks (for example, signal transduction cascades) (Wagner, 1994). Genes involved in development, which become integrated into developmental pathways that are hierarchical and highly interdependent, cannot readily mutate or take on new functions without major disruption to the developmental processes. Gene duplication provides a way around this by retention of the pre-existing developmental inter-relations and the incorporation of newly developed genes into new pathways and relationships. Insertion of new genes into an interactive network of genes (sets of genes encoding regulators (e.g., transcription factors) that mutually regulate each other's activity) can be expected to introduce new degrees of freedom and thereby multiple possible avenues of evolutionary divergence (Ruddle *et al.*, 1994b).

It has been shown that in the evolution of gene networks, the probability that a gene duplication event alters the equilibrium expression pattern of network genes is a unimodal function of the fraction of network genes that are duplicated in a single duplication event. The least likely scenario to be tolerated is when around 40% of the genes in a network are duplicated (Wagner, 1994). Preferentially, either single genes are duplicated, or all genes involved in a network are duplicated. It follows then, that a duplication event which involves genes incorporated in a network that are organized either in a tight linkage of all genes on one chromosome, as in the case of any of the *Hox* gene clusters, or as individual genes dispersed throughout the genome, is more likely to result in a viable organism.

1.2 TYPES OF HOMOLOGY

1.2.1 ORTHOLOGUES

Orthologues are sequences descended from a common ancestor through speciation (Lundin, 1979). Orthology implies that the original gene was retained in both species due to natural selection and in most cases has also kept its original function. Examples of orthologues include the human α-globin gene, and the mouse α-globin gene.

1.2.2 PARALOGUES

Paralogues are sequences arising from gene duplication, and existing within a single species (Lundin, 1979). A paralogous region is a segment of chromosome that contains a set of distinct genes that have paralogues on another segment of chromosome elsewhere in the same genome.

Paralogues can also exist between species. If a gene duplication is conserved between two species, then the two pairs of similar genes between species are orthologues, but the other two pairs that can be formed are paralogues, e.g., the human α-globin gene and the mouse ß-globin gene are paralogues.

1.2.3 TETRALOGUES

Sidow (1996) noted that many developmental regulatory genes are coded for by a single gene locus in protosomes (such as *Drosophila*), and primitive chordates (such as amphioxus), whereas in vertebrates these regulators are coded for by gene families. The *Hox* gene clusters are a much studied example of this phenomenon. The homeotic complex *HOM-C* in *Drosophila* corresponds to four *Hox* clusters found on four different chromosomes in vertebrates (Holland and Garcia-Fernàndez, 1996). Other such examples include the vertebrate *Cdx* (*Drosophila* homologue is caudal (*cad*): homeobox transcription factors) (Gamer and Wright, 1994), *MyoD* (*nau*: bHLH transcription factors) (Atchley *et al.*, 1994), *Notch* (*N*: cell-cell interaction receptors) (Uyttendaele *et al.*, 1996). Not all *Drosophila* genes have four copies in vertebrates; some only have two, some three, homologues.

Jürg Spring (1997) compiled a set of 53 such families where one *Drosophila* gene corresponds to two, three, or four or more genes in higher vertebrates, and introduced the term 'tetralogue' to distinguish putative paralogues resulting from tetraploidy from other sorts of paralogues, such as paralogues in tandem (e.g., *HoxA4*, *HoxA5*). Tetralogues, which correspond to a single invertebrate gene, form groups of four paralogous vertebrate genes, at four different chromosomal locations. There were only

12 instances, out of the 53 examples given, where one *Drosophila* gene was related to four vertebrate genes. There were also four examples where more than four genes were given as being homologous to one or two *Drosophila* genes. Spring suggested that in these instances, the tetralogue concept still held. He proposed that these situations were caused by genes that had not yet been identified either in vertebrates or in *Drosophila*, and that these larger sets could break up into two tetralogue sets. He cited an example where a newly discovered *Drosophila* gene (*Src41A*), homologous to *Src29A*, had helped to break up the vertebrate *src*-family, with eight members, into two sets of tetralogues.

1.3 TYPES OF GENE DUPLICATION

The DNA content of a cell can be altered by a variety of duplicative processes including tandem duplication, transposition, aneuploidy and polyploidy. These differ only in how much DNA is duplicated and where the duplicate copy is located in the genome.

1.3.1 TANDEM DUPLICATION

Gene duplication arises from an error in the mitotic or meiotic processes. Frequent exchange between sister chromatids, in the synthetic phase of the meiotic cycle, was first noticed in 1957 (Taylor *et al.*, 1957). Occasionally this exchange is unequal between the two chromatids, resulting in one daughter cell becoming heterozygous for a duplicated gene, the other hemizygous. Unequal crossing over during meiosis results in a similar situation. During the first meiotic prophase, two homologous chromosomes pair up and form chiasmata, and exchange genetic material by a process of genetic recombination, placing two former alleles of the same gene locus onto the same chromosome. Unequal crossing over can lead to a considerable expansion and contraction of gene family size. Subsequent chromosomal rearrangements can either increase the distance between duplicates or distribute duplicates to other chromosomes, without regard to the chromosomal location of other members of the gene family. Such rearrangements are rare, however, due to the original proximity of the duplicates, and tandem duplications tend to remain physically linked. Examples include the *Hox* gene clusters and globin clusters.

1.3.2 RETROTRANSPOSITION

Retrotransposition produces a single gene duplication that lacks introns and that moves to new chromosomal locations via an RNA intermediate (Vanin, 1985). The chromosomal location of the processed genes is apparently independent of other members of the gene family from which they originated. In general, retrotransposition results in processed pseudogenes possessing no promoter, no introns, and a remnant of a 3' poly-A sequence, often flanked by short direct repeats (Rogers, 1983). A few examples of retroposons that are intact, expressed and that lack introns, are known. One such example is that of the human phosphoglycerate kinase 2 gene which arose from the *pgk-1* gene by retrotransposition before the human-mouse split (Boer *et al.*, 1987; McCarrey and Thomas, 1987). The rat preproinsulin I gene is an example of a retrotransposed gene which still contains introns (Soares *et al.*, 1985).

1.3.3 POLYPLOIDY

The simultaneous duplication of all the genes in the genome is termed whole genome duplication or polyploidy. Polyploidy is often used to refer to recent events whereas genome duplication usually refers to postulated ancient events. Polyploidy is further categorized into auto- and allo-polyploidy depending on whether the polyploid had one or two parent species. For older genome duplications it is generally impossible to distinguish between these two possibilities, and it may be difficult or impossible to distinguish duplication of the complete genome from other large-scale processes such as aneuploidy (duplication of one or more chromosomes). Large-scale chromosomal duplication or genome duplication is identifiable as clusters of adjacent dissimilar genes found in duplicated regions or blocks, usually on different chromosomes. Genes duplicated by polyploidy tend to show greater sequence divergence than tandemly duplicated genes, which may be homogenized by gene conversion (Ohno, 1993). Using the method of Crow and Kimura (1970), Bailey *et al.* (1978) calculated that for a single locus, there is a 50% probability of loss of function at that locus after 4.9 x 10^5 generations. Unlinked duplicates are likely to survive for 100 times that long.

1.3.2.1 AUTOPOLYPLOIDY AND ALLOPOLYPLOIDY

Autopolyploids are polyploids that have descended from a single ancestral diploid species that have undergone a doubling of the genome within that species. Allopolyploidy involves the chromosomal hybridization between different species and the assembly of pairs of corresponding genes which have probably already given rise to distinct alleles during the separate history of the parental species. Allotetraploidy may be important in the restoration of fertility to plant hybrids (Hilu, 1993).

Allopolyploidy between two different species would imply that the faster evolving genes were already quite different, and would therefore reduce the likelihood that these genes would be functionally redundant in the hybrid. Conversely, highly conserved genes would form a redundant gene pool, and would be more likely to be reduced rapidly to a single copy. Spring (1997) suggested that allopolyploidy by interspecific hybridization could be responsible for the number of tetralogues observed in the human genome.

In some allopolyploid species, referred to as segmental allopolyploids, a proportion of chromosomes form quadrivalents during meiosis whilst others assume bivalent formation. Duplicated genes resulting from segmental allopolyploidy tend to have different times of divergence: some indicate the time of divergence of the two parent species, and others indicate the time of formation of the segmental allopolyploid. The ancestor of maize appears to have been a segmental allotetraploid (Gaut and Doebley, 1997).

1.3.2.2 Resolution of tetraploidy

A newly arisen tetraploid has four identical chromosomes for each linkage group, which form quadrivalents at meiosis (so that a gamete can contain any two of the four chromosomes). Diploidization involves the re-establishment of disomic inheritance by functional diversification of the four original homologues of any linkage group, so that the original linkage group is split into two separate linkage groups (with gametes containing one chromosome from each pair). The preferential formation of two separate bivalents instead of one quadrivalent is the prerequisite for diploidization, a process aided by gene deletions, reciprocal translocations, transpositions, inversions, and Robertsonian fusions as well as by nucleotide substitution. This process is not necessary in allopolyploids because some sequence divergence will already exist between homologous chromosomes, allowing them to form bivalents. Functional divergence cannot begin until disomy is re-established.

The rate of chromosomal rearrangements, in particular reciprocal translocations, seems to be greatly increased following polyploidization, because they return the genome to a more stable diploid state (Leipold and Schmidtke, 1982; Ahn and Tanksley, 1993). In synthetically generated *Brassica* polyploids, rearrangements following polyploidization began in the F2 generation (Song *et al.*, 1995). Chromosomes showing many distinct paralogous regions, as would be expected after polyploidization (Lundin, 1993), have been observed in *Zea mays* (Helentjaris *et al.*, 1988; Ahn and Tanksley, 1993). Only reciprocal translocations involving non-homologous chromosomes are easily detected as disruptions in duplicated segments.

1.4 FATE OF DUPLICATED GENES

There are three possible fates for a duplicated gene pair: both daughter genes are retained; one is retained and one is lost; both are lost. It is presumed in general that if a gene is present in a genome, its product plays some essential role, and therefore an organism that loses both duplicates will be non-viable (e.g., Nadeau and Sankoff, 1997). Furthermore, it is assumed that recovery ("resurrection") of a pseudogene becomes increasingly improbable due to the accumulation of secondary deleterious mutations (Dollo's Law; Marshall *et al.*, 1994). For a population of infinite size, unlinked duplicates must both remain functional indefinitely, because the mutational rate would be balanced through the selective elimination of defective double homozygotes (Fisher, 1936). But for real populations, mutation and random drift lead to the silencing of duplicates in proportion to the effective population size, where duplicates in a larger population are less likely to be silenced.

Under the classical interpretation, after gene duplication, one of the two daughter genes becomes redundant, and is free to accumulate previously forbidden mutations, changing the character of the protein for which it codes (Ohno, 1970). There are two likely evolutionary reasons for retaining both copies of a duplicated gene: selection for increased levels of expression, or divergence of gene function. Different genes may have different effective neutral mutation rates because of their size, structure and function (Allendorf, 1978). This was confirmed by Bulmer *et al.* (1991) who showed that silent substitution rates differed between genes. Iwabe *et al.* (1996) have also shown that different functional classes of genes have remained in duplicate to different degrees during vertebrate evolution. Proteins that are more variable than others are predicted to be less likely to remain in duplicate (Allendorf, 1978). For example, creatine kinase has two duplicated loci in salmonids, CK-1 and CK-2, which exhibit low heterozygosity, and are kept in duplicate in eight out of 10 salmonid species (Utter *et al.*, 1979).

Because most mutations are disadvantageous, and when a mutant is advantageous it is so only under restricted conditions, it is expected that the probability of gene loss will be greater by an order of magnitude that the rate of functional divergence (Kimura and Ohta, 1974), although there might be a slight advantage in fixing a new duplicate because it can potentially mask the accumulation of null alleles (Clark, 1994). Therefore a gene duplicate is likely to be quickly lost if there is no positive selection acting on it (Anderson and Roth, 1977). It has also been shown mathematically that the probability that a gene is retained is dependent on the ratio of advantageous to null allele mutation rates, the selective advantage of an advantageous allele, and the effective population size. At an effective population size of 5,000, with a selective advantage of 0.01, and a ratio of 5 x 10^{-5}, there is a 1% probability that a duplicate will be retained. It is only when N_e increases to 500,000, that the ratio increases to 50% (Walsh, 1995). When N_e is 5,000,000 and 50,000,000, that ratio increases to 90% and 99%, respectively.

However, in vertebrates that are recent tetraploids, the fraction of the genome retained in duplicate is surprisingly high. Some examples include 30% of genes kept in duplicate 50 Myr after tetraploidization in loaches (Ferris and Whitt, 1977), 50% after 50 - 100 Myr in salmonids and after 50 Myr in catastomids (although the effective population size of salmonids may extend into the millions) (Bailey *et al.*, 1978) and 50% after 30 Myr in *Xenopus* (Hughes and Hughes, 1993). Yeast, on the other hand, has retained only 8% of its genes in duplicate after 100 Myr despite having a very large effective population size (Seoighe and Wolfe, 1998). Differences in the rate of gene loss among species may be a result of differences in generation time and effective population size.

Nadeau and Sankoff (1997) have demonstrated, using a dataset of 661 human genes and 419 mouse genes, that duplicated genes in the human and mouse genomes are almost as likely to acquire new functions as they are to be lost through the accumulation of disadvantageous mutations. There are a number of reasons why the classical model may give a gross underestimate of the percentage of duplicates retained. Hughes (1994) pointed out that, in *Xenopus* at least, purifying selection acted on both duplicates, not just one. He proposed a theory of bifunctionality, whereby usually no new functions were created but instead the daughter genes each specialized for one of the functions carried out by an ancestral bifunctional gene. In cases where duplicated genes evolved new functions, positive Darwinian selection aided the adaptation of each gene (Hughes, 1994). One recent example of positive Darwinian selection is shown in the case of the genes for eosinophil-derived neurotoxin and eosinophil cationic protein (Zhang *et al.*, 1998). This argument was further propounded by Fryxell (1996) who stated that pre-existing heterogeneity could facilitate the retention and divergence of duplicated copies of functionally interacting genes. Another variation on this is the theory that multidomain proteins are more likely to be retained in duplicate. Assuming that point mutations occur more often than naturally occurring deletions, because multidomain proteins have multiple interactions and are often found within multiprotein complexes, there is a stronger dominant negative deleterious phenotype exhibited by loss-of-function point mutations than by null mutations (Gibson and Spring, 1998). However, genes vary widely in the extent to which they might display this effect, and it may be caused by only a small number of point mutations (Cooke, 1998).

Another factor not taken into account by the classical model is the concept of developmental gene networks. Genes involved in development often become integrated into hierarchical and interdependent pathways (Ruddle *et al.*, 1994b). Whereas housekeeping genes tend to be reduced to a single copy, duplicated developmental genes can be recruited to new specificities (Cooke *et al.*, 1997). Many developmental genes with just one copy in invertebrates such as *Drosophila* have two to four counterparts in higher organisms (Sidow, 1996). It has been shown that redundancy should be more common in developmental genes expressed in specific spatio-temporal patterns, and that very evolutionarily stable configurations consist of complex networks in which each function is performed by several genes and each gene performs several functions (Nowak *et al.*, 1997). Redundancy may also be maintained by appropriately balanced mutation rates (Nowak *et al.*, 1997).

These considerations have focused entirely on the coding regions of the duplicated genes. But it is not just the coding regions but also the regulatory regions that are duplicated during genome duplication. Spring (1997) mentioned that the regulatory regions of genes can evolve faster than the coding regions, and could lead to tissue specificity of expression of functionally redundant genes. Lim and Bailey (1977) also pointed out that the duplicated LDH A_4 and B_4 loci in salmonids may be drifting towards non-functionality because of the accumulation of mutations in the regulatory DNA, not in the structural DNA. Recently, it has been suggested that genes with a number of subfunctions regulated by more than one *cis*-regulatory region are more likely to be preserved when the total subfunction mutation rate relative to the total null mutation rate is between 0.3 and 0.7 (Force *et al.*, 1999), an idea also mentioned by Tautz (1998) in relation to sea urchin enhancers. A mutation in different *cis*-regulatory regions on both duplicates implies that both have to be retained so that both of those subfunctions remain active. The higher the number of subfunctions, the more likely it is that both genes are retained. A corollary of this is that after a second round of genome duplication, because each duplicate now has fewer subfunctions than it did before the first genome duplication, the likelihood that both of the new duplicates are retained decreases.

1.5 DUPLICATION EVENTS IN THE LINEAGE LEADING TO VERTEBRATES

Free-living organisms can be divided into three categories based on their gene number: prokaryotes, which have only a few thousand genes; eukaryotes (excluding vertebrates) which contain from 6,000 genes (*S. cerevisiae*; Dujon, 1996) to 12,000 ± 5,000 genes (*Drosophila*; Miklos and Rubin, 1996; Ashburner *et al.*, 1999), 15,000 ± 3,700 genes (*C. intestinalis*; Simmen *et al.*, 1998), 19,000 genes (*C. elegans*; Waterston and Sulston, 1995; C. elegans Sequencing Consortium, 1998) and 25,000 genes (*S. purpuratus* (sea urchin); Poustka *et al.*, 1999) and vertebrates, which are thought to have between 50,000 and 100,000 genes (Brenner *et al.*, 1993; Fields *et al.*, 1994). Gene number in vertebrates is fairly constant, and is several-fold greater than in invertebrates (Miklos and Rubin, 1996). *Msx*, *Cdx* and insulin, for example, are all represented once in invertebrates such as *Drosophila* or amphioxus, but several times in vertebrates (Chan *et al.*, 1990; Holland *et al.*, 1994; Sidow, 1996; Spring, 1997). Because mammals have about 5.8 times as many genes in their genomes as do *Drosophila* and *C. elegans* (Miklos and Rubin, 1996), the basic vertebrate gene number may be similar to that of invertebrates. This supposition is made plausible by the discovery that out of the ~80,000 genes in mouse, only 5,000-26,000 of them are essential for viability, a number which brackets the invertebrate gene number (Miklos and Rubin, 1996). Bird (1995) suggested that major transitions during evolution (from prokaryotes to eukaryotes, and from invertebrates to vertebrates) were accompanied by large increases in gene number, where only functionally different genes are counted (counting functional genes is a more meaningful parameter of complexity than genome size, since genome size is not an indicator of gene number). It has also been hypothesized that gene number is limited by the efficiency with which spurious gene expression can be repressed (Bird, 1995; Hurst, 1995). The more cell types an organism has, the more tissue-specific genes it requires, and the greater the number of genes that need to be repressed in any one cell-type. The development of the nuclear envelope, histones and perhaps most importantly, DNA methylation, was important in such a reduction of transcriptional background noise (Bird, 1995).

The scarcity of polyploids among vertebrates and invertebrates is due to the establishment of the chromosomal sex determination mechanism. While the sex chromosomes of fish and amphibians appear to have remained morphologically indistinct, it has been widely held that the development of the amniote egg prevented the possibility of polyploidy in mammals, birds and reptiles, because duplication of the sex chromosomes would result in infertility (Ohno, 1970). The single known exception to this is the case of the red viscacha rat, *Tympanoctomys barrerae*, which has been shown to be tetraploid (4N = 102), with a single XY sex-chromosome system (Gallardo *et al.*, 1999). If genome duplication events occurred in vertebrates (as a means of increasing gene number), the most recent one can only have taken place around or before the divergence of reptiles from the mammalian lineage. Subsequent to this, once polyploidization became unfeasible, gene duplication could only be accomplished by aneuploidy, unequal crossing over, unequal exchange, transpositions and other means of tandem duplication (Ohno, 1970).

1.5.1 OHNO'S HYPOTHESIS

In 1970, Susumu Ohno suggested that whole genome duplication, via a state of tetraploidy, was not only possible, but was also an important evolutionary mechanism (Ohno, 1970; Ohno, 1993). In 1967, Atkin and Ohno reported the haploid genome content of the amphioxus *Branchiostoma lanceolatum* as being approximately 0.6 pg (Atkin and Ohno, 1967), which is about three times larger than that estimated for tunicates (e.g., *Ciona intestinalis*, 0.2 pg, or 160 Mb) but similar to the smallest vertebrate genomes (e.g., *Fugu rubripes*, at 0.5 pg). That of placental mammals is approximately 3.5 pg. Based on the comparison of basic levels of DNA in the nuclei of the different vertebrate classes, Ohno proposed that there may have been as many as three consecutive tetraploidizations, or genome-wide duplications, in the descent from primitive chordates to mammals (see Figure 1.1).

Lundin (1993) argued that it was possible that three rounds of duplication had occurred, but that evidence for the earliest round was likely to be obscured.
Sidow (1996) considered only two rounds of duplication to have occurred, based on his analysis of the phylogenetic relationships among extant deuterosome classes and phyla. Spring (1997) argued for two consecutive genome-wide duplications because of the presence of numerous tetralogues, or the almost ubiquitous 1:4 concordance between many *Drosophila* genes and their mammalian homologues. Based on the available comparative data in 1994, Holland *et al.* (1994) postulated two phases of genome expansion, the first being close to the origin of vertebrates, just after the divergence of the lineage leading to amphioxus (550 - 450 Mya) (contradicting Ohno's argument that amphioxus had a similar genome size to vertebrates), and the second duplication event occurring just before the appearance of jawed vertebrates (450 - 400 Mya), using the dates of Doolittle *et al.* (1996). Conversely, Sharman and Holland (1996) proposed that the first phase of genome expansion involved mostly tandem duplications, rather than tetraploidy, because some genes have only two copies in the vertebrate genome, e.g., HMG1/2 did not undergo duplication near the origin of vertebrates, but was duplicated only once after the divergence of jawed and jawless fish.

The most common model, which is influenced largely by studies on the *Hox* gene clusters, involves two duplications: one in the common ancestor of vertebrates (after the divergence from amphioxus, about 500 Mya; Holland *et al.*, 1992), and one in the common ancestor of cartilaginous fish and tetrapods (after the divergence from lamprey/hagfish, about 430 Mya). Figure 1.1 shows a summary of the various proposals of occurrences of duplication in the vertebrate lineage. Other proposals not

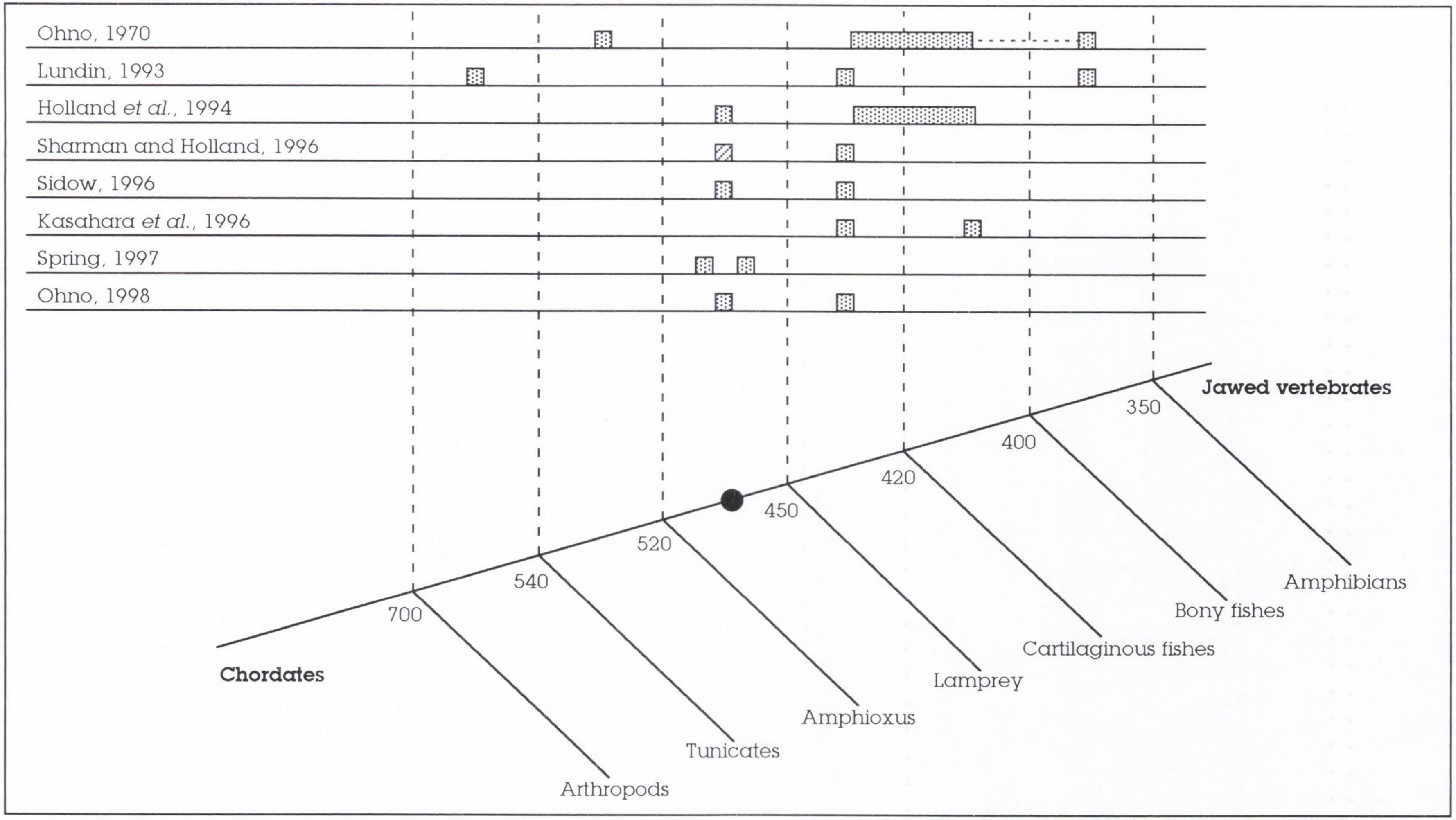


Figure 1.1. Summary of the various proposals of the occurrences of tetraploidy in the vertebrate lineage. Divergence of various lineages from the vertebrate lineage are drawn schematically, not to scale. Shaded boxes indicate each proposed genome duplication and are drawn at the centre of possible time ranges. The hatched box indicates a proposed wave of tandem duplications, as opposed to a tetraploidization event (Sharman and Holland, 1996). Ohno (1970) postulated a tetraploidization event at the divergence of fish and/or amphibians, shown by the two boxes connected by a dotted line. The circle at ~ 500 Myr denotes the origin of vertebrates.

shown in Figure 1.1 include Koop and Nadeau's suggestion (1996) that the most recent duplication in the mammalian lineage may have occurred after its divergence from bony fish, and the suggestion that an additional (chromosomal or whole genome) duplication may have occurred in the zebrafish lineage after its divergence from the lineage leading to mammals (Postlethwait *et al.*, 1998).

One reason for Ohno's exclusion of a genome-wide duplication event at the origin of vertebrates (see Figure 1.1) could stem from the fact that he was working only with genome sizes, before the discovery of "junk" (non-coding) DNA. But genome sizes are only a very approximate indication of gene number. For example, the human genome is thought to contain 97% "junk" DNA (Ohno, 1972; Miklos and Rubin, 1996). Repetitive DNA can comprise from between 20% to 50% of metazoan genomes, and this fraction can change dramatically during the course of evolution, probably without concomitant changes in gene number (Lewin, 1990). Secondly, Ohno's assessment of genome expansion relied heavily on the assumption that *Fugu*, which has the smallest vertebrate genome, was representative of early vertebrates. Genome sizes vary widely within the fishes, and it is likely that the genome of *Fugu* is unusually small and compact, and cannot be taken as representative (Holland *et al.*, 1994).

The idea that genome duplications occurred around the time of vertebrate origins is tempting because it is consistent with the hypothesis that gene duplications facilitate developmental evolution. Holland *et al.* (1992) speculated that multiple gene duplications may have occurred at that time; new genes could then have been co-opted to new roles, facilitating the evolution of new developmental pathways and the morphological diversity of the Cambrian explosion (Ohno, 1998). A comprehensive analysis of duplicate genes in vertebrates by Iwabe *et al.* (1996) demonstrated that there had been an apparent burst of gene duplications producing tissue-specific isoforms of proteins in the vertebrate lineage, after it had diverged from arthropods but before it diverged from bony fish.

1.6 POLYPLOIDY IN NON-MAMMALIAN SPECIES

The strongest evidence that genome duplications do occur comes not from vertebrates, but from plants and fungi. Polyploidy is common in plants, fungi and in some vertebrates such as fish and amphibians, but is unknown in those vertebrates (mammals and birds) whose sex determination mechanisms depend on having distinct sex chromosomes (Muller, 1925; Ohno, 1970), with a single exception of the red viscacha rat (Gallardo *et al.*, 1999).

1.6.1 *XENOPUS*

Xenopus, a genus of African clawed frogs, is a bisexual species which has undergone several independent polyploidization events. The only diploid member of the *Xenopus* family is *X. tropicalis*, with a chromosome number of 20. All other members such as the tetraploid *X. epitropicalis* (40 chromosomes; forms a group with *X. tropicalis*) and *X. laevis* (36 chromosomes), the octaploid *X. vestitus* and *X. wittei* (both with 72 chromosomes) and the dodecaploid *X. ruwenzoriensis* (108 chromosomes) have undergone one or more polyploidization events. The tetraploid species consist of pairs of morphologically distinct chromosomes; the higher polyploids have quartets and sextets of chromosomes. In all species, chromosomes generally assemble as bivalents during meiosis. All *Xenopus* species are thus expected to display the inheritance patterns predicted by disomy rather than polysomy (Kobel and Du Pasquier, 1986).

There are numerous examples of unlinked duplicated genes within the *X. laevis* genome (Jeffreys *et al.*, 1980; Hosbach *et al.*, 1983; Graf, 1989). It has been estimated that a first polyploidization event within the *Xenopus* genus took place around 30 Mya (Bisbee *et al.*, 1977; Thiébaud and Fischberg, 1977; Hosbach *et al.*, 1983). It has been noted that during the evolution of duplicated genes in *X. laevis*, both copies of a duplicate gene are subject to purifying selection. The sequences of 17 duplicated *X. laevis* genes were analysed by comparing rates of synonymous and non-synonymous nucleotide substitution between the two copies with the corresponding rates between the orthologous genes of human and rodent (Hughes and Hughes, 1993). Neither of the duplicated copies was found to be free to accumulate mutations without constraint. However, if both duplicated genes are still present and functioning in the genome, then they are by definition both still identifiable. Thus this dataset is biased because there must also be many examples of duplicated genes where one or other of the duplicates has been lost, and therefore it cannot be shown that there is some restraint working on them.

1.6.2 JAWED FISHES

Several groups of fishes, including all salmonids and catastomids, as well as carp and some loaches, have arisen via tetraploid ancestors (Bailey *et al.*, 1978). A large percentage of loci have remained as functioning duplicates over long periods of time in these fishes. Both salmonid and catostomic fish still retain 50% - 75% of the gene duplicates produced by tetraploidy over 50 - 100 Mya.

The Salmonidae appear to be an autotetraploid family that probably originated from a diploid ancestor relatively recently, around 50 - 100 Mya (Lim *et al.*, 1975; Allendorf and Thorgaard, 1984), and appear to be progressing towards diploidization to various degrees (Kavsan *et al.*, 1993). The Pacific chum salmon (*Oncorhynchus keta*) is thought to have almost completed this process, containing two non-allelic insulin genes that have structurally diversified.

Catostomids (catfish) have been proposed to have undergone a tetraploidization event around 50 Mya (Uyeno and Smith, 1972). Catostomids have 100 chromosomes, as opposed to 50 for most cypriniformes, and also contain a DNA content twice that of typical diploid fishes. The catostomids have returned to a functionally diploid state at over half of their enzyme loci, a process that may not be entirely random. In many cases, where two catostomid species have the same percentage of duplicates expressed, different combinations of loci have undergone the diploidization process. There would appear to be no obvious correlation with the metabolic function of an enzyme and the extent to which the duplicate loci encoding it have undergone diploidization. Because some of the duplicate loci have diverged in function and regulation, and observed levels of duplicate gene expression are much higher than predicted by gene loss, it is likely that these duplicate genes are being either maintained or silenced by selection (Ferris and Whitt, 1977), or perhaps in the manner described by Allendorf (1978) (see Section 1.4).

The cyprinids (carp) are also a tetraploid species, thought to have diverged from the catostomids 50 Mya (Uyeno and Smith, 1972). They appear to have undergone a separate allotetraploidization event from the catostomids around 16 Mya (Larhammar and Risinger, 1993). The two parent diploid species are thought to have been quite different from each other since no chromosomes appear to have been lost in the tetraploidization event and there are no quadrivalents evident at meiosis (Ohno *et al.*, 1967).

There is a general pattern of extensive, expressed enzyme gene duplications in teleost fishes (Morizot, 1990). There are examples of teleost fishes having twice the number of enzyme gene duplicates as mammals (e.g., there are two cytosolic creatine kinase loci in mammals, but four in fishes; Ferris and Whitt, 1977). Perhaps the most studied jawed fishes to date are *Fugu* and zebrafish. *Fugu* was first proposed as a model organism to aid in the discovery of human genes (Brenner *et al.*, 1993). Although its genome is 7.5 times smaller than that of human, it seems to have a similar gene repertoire (Brenner *et al.*, 1993), and its absolute gene number is comparable to that of mammals (Koop and Nadeau, 1996). Most of the work done on *Fugu* has involved the investigation of synteny between the human and *Fugu* genomes (Baxendale *et al.*, 1995; Schofeld *et al.*, 1997; Miles *et al.*, 1998; Brunner *et al.*, 1999; Gellner and Brenner, 1999; Gilley and Fried, 1999; Lim and Brenner, 1999). However, it has been suggested that an ancestor of *Fugu* did double its genome, most of which was then lost (Amores *et al.*, 1998; Postlethwait *et al.*, 1998; Vogel, 1998).

Postlethwait *et al.* (1998) developed a linkage map for zebrafish (*Danio rerio*) which included 144 genes. They showed that the large conserved groups of duplicated genes that are linked to the four *Hox* clusters in mammals are also present in zebrafish, as well as some parts of the HSA 1/6/9/19 paralogy group. This implies that the duplications producing these chromosomal regions occurred before the bony fish/tetrapod divergence. Although zebrafish and mammals show conservation of synteny, gene order is often rearranged in these examples (Aparicio, 1998). Postlethwait *et al.* (1998) also report some examples where a pair of linked genes in a mammal seems to correspond to two linked pairs in zebrafish. Similarly, Gates *et al.* (1999) identified 18 cases where two zebrafish genes seemed to be orthologous to a single mammalian gene (but also conversely, three cases where two mammalian genes were orthologous to a single zebrafish gene) and propose that additional duplications (either of chromosomal fragments or of the whole genome) may have occurred in this species (Postlethwait *et al.*, 1998). It is probable that a genome duplication occurred in a common ancestor of *Fugu* and zebrafish, after this lineage diverged from other vertebrates.

1.6.3 JAWLESS FISHES

Lamprey and hagfish are the only living jawless vertebrates, and evidence from 18S rRNA studies shows that they form a natural group (Stock and Whitt, 1992). The lamprey (*Lampetra fluviatilis*), a diploid, is one of the earliest vertebrate lineages (Figure 1.1). Studies of the HMG locus in lampreys yielded a single gene, LfHMG1, which is homologous to the functional duplicate pair HMG1/2 in mammals (Sharman *et al.*, 1997). This implies that the duplication event in the HMG1/2 family occurred after the divergence of the jawed and jawless fishes around 450 Mya. This finding supports and refines the hypothesis that there was a period of extensive gene duplication early in vertebrate evolution. Lamprey, however, also appears to have an intermediate number of three *Hox* gene clusters (Sharman and Holland, 1998), which implies that if there was a genome doubling event in the vertebrate lineage, it occurred before jawless fishes split off from the chordate lineage. Spring (1997) has also suggested that hagfish may be an allohexaploid, or perhaps one of the allotetraploid ancestors that were the parents of an allooctaploid which evolved into eukaryotes such as human and mouse.

1.6.4 PLANTS

Both autopolyploidy and allopolyploidy are common phenomena in domestic plant evolution. A large number of major food crops are polyploid, such as potato, sweet potato, orchard grass and alfalfa (autopolyploids), or cotton, tall fescue, wheat and tobacco (allopolyploids). A number of selective advantages conferred by polyploidy (such as longer life span, greater defense against pathogens, larger seeds) appear relevant to domestication (Hilu, 1993). Both monocots (such as cereals and grasses) and dicots (such as potato, tomato, beans, *Arabidopsis* and *Brassica*) exhibit polyploidy. Monocots and dicots diverged from each other around 130 - 300 Mya (Wolfe *et al.*, 1989), and Paterson *et al.* (1996) have predicted that 43 - 58% of chromosomal segments ≤ 3 cM should remain collinear between the two taxa after such a period of time. Larger conserved blocks between monocots and dicots may reflect unusual structural features and genomic processes that confer fitness advantages (Paterson *et al.*, 1996). Despite these predictions, comparative plant analysis has largely been made within monocots and dicots, not between them.

1.6.4.1 MONOCOTS

Among the most intensively studied cereal species are maize, wheat, sorghum and rice, all of which are polyploid with the exception of rice. Although their genome sizes differ 40-fold (16,000 Mb in wheat, 400 Mb in rice), synteny and gene order between them is found to be remarkably well conserved (Ahn *et al.*, 1993).

Maize is by far the best characterized degenerate polyploid plant. The duplicated structure of the maize genome was first recognized by Helentjaris *et al.* (1988). Ahn and Tanksley (1993) found that 72% of cDNA clones that mapped to a single genomic locus in rice were duplicated in maize, and that there is still extensive conservation of gene order between these species. They tentatively identified two polyploidy-derived pairs of maize chromosomes. In maize, the most recent map (Gale and Devos, 1998) shows approximately 14 translocations or inversions and 10 chromosome arm fusions or fissions after genome duplication. It has been proposed that the duplicated regions of the maize genome arose from segmental allotetraploidy and that one of its diploid parents may have been related to sorghum (Gaut and Doebley, 1997). Phylogenetic analysis using outgroup sequences for 14 duplicated genes yielded a range of divergence dates (8 - 23 Mya), interpreted as forming two non-overlapping groups (centered on 11.4 and 20.5 Mya), corresponding to the speciation time between the two progenitors (20.5 Mya), and to the time of establishment of disomy (11.4 Mya). Interestingly, if the ancestral maize genome constructed by Wilson *et al.* (1999) is correct, then the diploidization process in maize must also have involved a reduction in chromosome number from 16 chromosomes to 10.

Wheat is a recently derived allohexaploid (Feldmann, 1976) constructed from three different diploid species. There is some evidence that even though wheat is a relatively recent polyploid, intergenomic rearrangements have taken place within this species (Ahn *et al.*, 1993). Gene order appears to be well conserved between rice and wheat (Moore *et al.*, 1995b), with most rice genes showing three wheat orthologues.

Because there is a high level of synteny and gene order conservation between some chromosomes of the cereal genomes (possibly implying that certain gene orders are advantageous and preserved by evolution (Ahn *et al.*, 1993)) a "lego" model has been proposed where the rice genome, assumed to be similar to the ancestral cereal genome, is divided into linkage blocks which are then compared with the genomes of other cereals (Moore, 1995; Moore *et al.*, 1995a; Moore *et al.*, 1995b). For example, the 10 chromosomes in the maize genome can be divided into two sets of five chromosomes. Each set contains a set of linkage blocks that reflect the rice genome, but they are rearranged in a different order, which supports the idea that maize is an allotetraploid, and also that one of its diploid parents was related to sorghum. However, the circular representation of the aligned genomes from the several species examined (wheat, maize, foxtail millet, sugar cane and sorghum) implies that either there have been no chromosomal fusions or translocations during grass evolution (in which case the ancestor of the grasses must have had just a single, giant, chromosome) or else that chromosomal fusions occur but that each chromosome is only permitted to fuse with a particular designated partner.

While this model is unlikely to be an accurate representation of the evolutionary history of the cereal family, it can be useful for locating genes in other genomes, especially in wheat, because the rice genome can be used as a tool to jump across regions of repetitive DNA (Moore *et al.*, 1995b).

1.6.4.2 DICOTS

Soy bean (*Glycine soja*) is another ancient tetraploid whose genome has undergone diploidization. Significant chromosome pairing during meiosis is indicative of extensive sequence similarity (Crane *et al.*, 1982). Segments ranging from 1.5 - 106.4 cM (averaging 43.5 cM) are present in up to six copies (an average of 2.55 copies) (Shoemaker *et al.*, 1996). The occurrence of nested duplications and the presence of triplicated and quadruplicated markers implies that one of the original genomes may have undergone an additional round of tetraploidization. Thus, tetraploidization, along with large segmental duplications, has been the mode of evolution in soy bean. The fact that soy bean multigene families contain two distinct subgroups of more closely related genes supports the theory of tetraploidization (Lee and Verma, 1984; Hightower and Meagher, 1985).

The soy bean genome has undergone many chromosomal rearrangements, a process which has been proposed to account for the diversity and success of many ancient polyploid plant lineages (Song *et al.*, 1995). Mungbean and common bean share relatively large conserved linkage blocks, but the soy bean genome is greatly rearranged with respect to either of these species: one linkage group in mungbean is split into 16 groups in soy bean, one in common bean is split into nine in soy bean (Boutin *et al.*, 1995). The high level of apparent rearrangement could be due to incomplete mapping of triplicated and quadruplicated markers.

1.6.5 *SACCHAROMYCES CEREVISIAE*

Saccharomyces cerevisiae was the first eukaryotic genome to be fully sequenced (Goffeau *et al.*, 1996). It is 12 Mb long and has a haploid number of 16 chromosomes (Dujon, 1996; Goffeau *et al.*, 1996). The genome contains approximately 6,000 protein coding genes, 40 snRNAs and 275 tRNAs (Goffeau *et al.*, 1996; Hani and Feldmann, 1998). 72% of the genome is occupied by open reading frames, with the average length of an ORF about 1450 bp (Dujon, 1996).

S. cerevisiae has been shown to be a degenerate tetraploid (first suggested by Smith, 1987) which underwent genome duplication around 100 Mya (Wolfe and Shields, 1997), and has retained only 8% of the original set of genes in duplicate (Seoighe and Wolfe, 1998). There have been 55 non-overlapping duplicated chromosomal regions found in the yeast genome (Coissac *et al.*, 1997; Mewes *et al.*, 1997; Wolfe and Shields, 1997), covering approximately 50% of the genome. The paired chromosomal regions contain duplicated genes with conserved gene order and transcriptional orientation, but the duplicated genes are outnumbered by unique (non-duplicated) genes lying between them. These unique genes must have been duplicated as part of the original genome duplication, but then returned to a single-copy state by deleting one of the copies, presumably because there was no selective advantage to keeping both. Evidence in favour of the degenerate tetraploid hypothesis is threefold: a) the transcriptional orientation of the entire block with respect to the centromere is conserved between duplicated blocks, which is to be expected if reciprocal translocations were involved in genome rearrangement after tetraploidization; b) no triplicated regions were found, contrary to expectations if there had been a series of regional or chromosomal duplication events; c) the duplicated regions are found scattered throughout the genome. By comparison with gene order in *Kluyveromyces lactis*, a model of yeast gene order evolution through tetraploidization, gene deletion and reciprocal translocation has been proposed (Keogh *et al.*, 1998).

1.7 QUADRUPLICATED REGIONS FOUND IN THE HUMAN GENOME

There have been three potentially quadruplicated regions found in the human genome: the *Hox* gene clusters, the MHC regions, and FGFR regions.

1.7.1 THE *HOX* CLUSTERS ON HSA 2/12/7/17

Homeobox genes encode transcription factors of the helix-turn-helix type, regulating various aspects of cell regulation and differentiation. They are typified by the homeobox domain, a highly conserved sequence of 183 nucleotides encoding a domain which is capable of binding a DNA motif and regulating gene transcription (Gehring *et al.*, 1994). The Antennapedia-class of homeobox (*Hox*) genes are involved in the determination of pattern formation along the anterior-posterior axis of the animal embryo (McGinnis and Krumlauf, 1992; Lawrence and Morata, 1994) and are found clustered together. In the developmental process, the 3' end gene of the Antp-class *Hox* gene complex is transcribed first, and the other genes are transcribed successively from the 3' end to the 5' end of the complex. There are four *Hox* gene clusters in mammals, altogether composed of 39 genes (Acampora *et al.*, 1989; Zeltser *et al.*, 1996). They are located on HSA 7p, 17q, 12q and 2q, and on MMU 6, 11, 15 and 2, and each cluster extends over 100 kb. All genes are expressed, there are no identifiable pseudogenes and the genes are all transcribed in the same direction with respect to each other (see Figure 1.2). There are 13 paralogy groups (PG), numbered 1-13 and clusters are labelled A-D, e.g., *HoxA3* is the third *Hox* gene in the *HoxA* cluster (found on HSA 7p and MMU 6) (Scott, 1992).

The clustered organization of *Hox* genes seems to be related to function, since the genes are generally expressed in a spatial and temporal order that is collinear with their

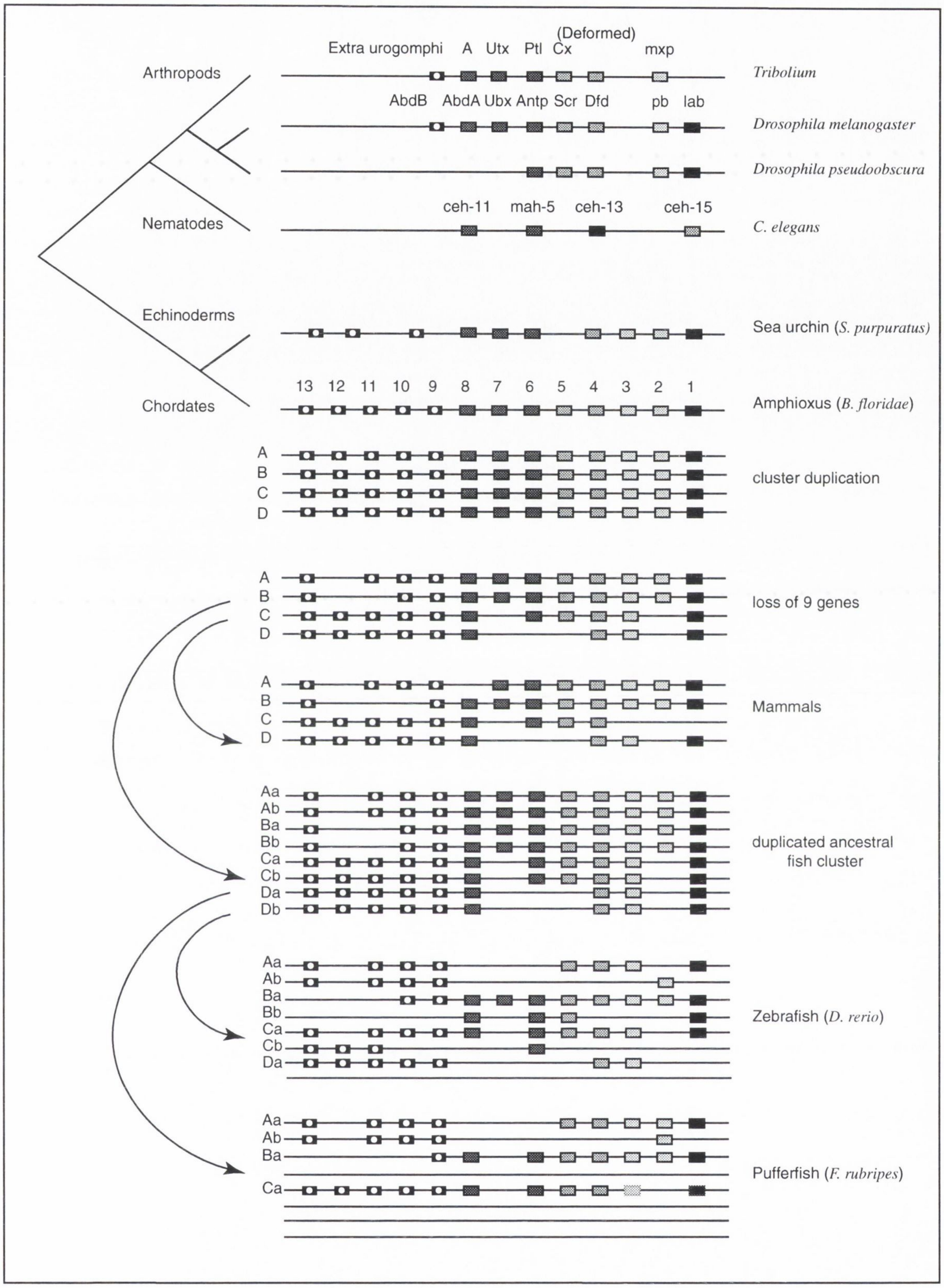


Figure 1.2. Evolution of the *Hox* gene clusters. An hypothetical ancestral cluster which contained approximately five or six *Hox* genes underwent tandem duplications of some of the *Hox* genes to give the pattern seen in examples such as *Drosophila* or amphioxus. The cluster in *C. elegans* underwent secondary gene losses (and a transposition of *ceh-13* and *ceh-15*), and now has only four *Hox* genes. In chordates, after the divergence of the lineage leading to amphioxus, the *Hox* clusters underwent two duplication events, yielding four clusters. A maximum of nine genes can be lost from the complete complement of the four clusters to giv an ancestral configuration which can give rise to the mammalian configuration and also the pre-duplicated fish configuration. A further duplication of all four clusters, probably as part of a genome-wide duplication event, occurred in the ray-finned fish lineage. Differential loss of genes and clusters led to the differences in the zebrafish and pufferfish *Hox* clusters. The phylogenetic tree on the left shows the evolutionary relationships of the species shown. Genes Ca1 and Ca3 in *Fugu* are pseudogenes.

physical order (Duboule and Dollé, 1989; Graham *et al.*, 1989). Genes of the greatest similarity occupy identical serial positions along the clusters. Sharkey *et al.* (1997) have identified characteristic residues that define the different paralogy groups. Most of these are orientated away from the DNA which they bind, in positions where they might engage in protein-protein interactions.

The Antp-class *Hox* genes have a near universal distribution, indicating the presence of a highly conserved developmental process shared by metazoans. It must have originated very early in metazoan evolution: even sponge, one of the most primitive metazoans, has Antp-class *Hox* genes but flagellates, amoeboids, ciliate protozoans, fungi, algae and plants do not (Degnan *et al.*, 1995). It has been suggested that *Hox* genes may be a key component of the evolution of diverse metazoan body forms (Amores *et al.*, 1998). The homeobox gene system represents the best example where duplications of entire gene clusters and the subsequent conservation of the organization of genes over such a long period of evolutionary time can be observed.

There is only one *Hox* cluster in *C. elegans* (Bürglin and Ruvkun, 1993), echinoderms such as the sea urchin and starfish (Martinez *et al.*, 1999), hemichordates such as the acorn worm (Pendleton *et al.*, 1993), *Drosophila* (Ruddle *et al.*, 1994a) and amphioxus (Holland *et al.*, 1992; Garcia-Fernàndez and Holland, 1994; Holland and Garcia-Fernàndez, 1996). There appear to be four or more clusters in birds (Stein *et al.*, 1996), amphibians (Belleville *et al.*, 1992; Stein *et al.*, 1996), teleost fish (Misof and Wagner, 1996; Aparicio *et al.*, 1997; Amores *et al.*, 1998; Postlethwait *et al.*, 1998; Prince *et al.*, 1998) and mammals (Ruddle *et al.*, 1994a). Lamprey, a jawless vertebrate, contains 21 genes, as yet unmapped, and appears to have an intermediate number of three clusters (Sharman and Holland, 1998). Paralogy groups PG11-13 have so far been only reported in jawed vertebrates, with the possible exception of lamprey, which appears to possess one PG13 gene, possibly indicating an early expansion of the posterior group genes (Holland and Garcia-Fernàndez, 1996).

Within all the species studied, the teleost fish seem to have the most dramatically different complements of *Hox* genes, which has been suggested to play some role in their evolutionary diversity (Kurosawa *et al.*, 1999). 31 *Hox* genes have been identified in *Fugu* (see Figure 1.2), which form four clusters on four different chromosomes (Aparicio *et al.*, 1997). However, instead of having one cluster of each type found in mammals, *Fugu* appears to have one *HoxA* , one *HoxB* and one *HoxC* cluster, and another cluster which is unlike either of the *HoxD* clusters found in mammals or zebrafish and which has been suggested to be a *HoxA*-like cluster (Aparicio *et al.*, 1997; Amores *et al.*, 1998). This presented the possibility that some fish had and may still have more than four *Hox* clusters, and that *Fugu* had lost the *HoxD* cluster completely, an idea which was made plausible when it was found that zebrafish has at least seven *Hox* clusters — two *HoxA* clusters, two *HoxB* clusters, two *HoxC* clusters, but only one *HoxD* cluster (Figure 1.2; Amores *et al.*, 1998; Prince *et al.*, 1998). John Postlethwait has suggested that there may have been a genome doubling in the ray-finned fish lineage, around the time of the fish radiation around 300 Mya (Postlethwait *et al.*, 1998; Vogel, 1998). *Fugu* has a highly compressed genome and an unusual morphology (e.g., no pelvis), and may have lost four of its duplicated *Hox* clusters, to give the configuration seen today.

The reconstruction of the history of the *Hox* clusters and their duplications, and its relation to vertebrate evolution has been problematic, due to the lack of alignable sequence information. Within a cluster, collinearity between gene expression and gene arrangement seems to have been generated by successive tandem gene duplications and that gene arrangement may be maintained by some sort of selection (Zhang and Nei, 1996). There appear to have been at least five genes in the last common ancestor of nematodes, *Drosophila* and chordates – PG1, PG2, PG3, PG4-8 and PG9-13, which may already have formed PG9-10 and PG11-13 (Kappen and Ruddle, 1993). The *Hox* cluster in the red flour beetle, *Tribolium*, containing just six *Hox* genes, is thought to be similar to the ancestral cluster (Figure 1.2; Beeman, 1987; Kappen and Ruddle, 1993). *C. elegans* then lost the PG2 and PG3 genes, and duplicated the PG4-8 gene, to give its present complement. Zhang and Nei (1996) proposed that there had been a tandem duplication of one *Hox* gene 1000 Mya to give two genes, PG1-8 and PG9-13, the first of which underwent tandem duplication to give rise to PG1-2 and PG3-8 around 730 Mya, and then PG3-8 duplicated to give PG3 and PG4-8 around 600 Mya. PG9-13 gave rise to PG9-10 and PG11-13 around 800 Mya. The five subgroups underwent different amounts of tandem duplication to produce the situation seen in *Drosophila*, which contains eight *Hox* genes, which share orthologous collinearity with the mammalian clusters for the first nine paralogy groups, with the exception of PG3 which is absent in *Drosophila* (Graham *et al.*, 1989). Before amphioxus diverged, the *Hox* cluster contained ≥ 10 genes (Garcia-Fernàndez and Holland, 1994). The four cluster state must have arisen in the lineage leading from primitive chordates to higher vertebrates because all invertebrates examined have a single cluster only (Bailey *et al.*, 1997), and it has been suggested that the generation of these clusters contributed to the evolution of more complex organisms (Kappen and Ruddle, 1993).

A two-step model has been described for the evolution of the *Hox* genes (Kappen *et al.*, 1989; Schughart *et al.*, 1989). The ancestral gene cluster was expanded by tandem duplication of individual homeobox genes, as described above, and then that entire expanded cluster was duplicated several times, to give the four clusters seen in vertebrates, suggested by the fact that homeobox sequences from each cluster seem to have diverged from the corresponding genes in other clusters to approximately the same extents. Although the *Hox* gene complement differs between vertebrates, there is no evidence for tandem duplications after cluster duplication (Schughart *et al.*, 1989). It was initially assumed that the *Hox* clusters duplicated twice, at least once close to the origin of vertebrates, since amphioxus only has one *Hox* cluster, to give a "one to two to four" cluster history (Holland and Williams, 1990; Kappen and Ruddle, 1993; Holland *et al.*, 1994). However, by assuming that the collagen genes which are found in close linkage to *Hox* clusters share the same duplication history as the *Hox* genes

(Ruddle *et al.*, 1994b), Bailey *et al.* (1997) showed that the *Hox* cluster duplication history may be more complex, possibly involving three duplication events, with subsequent cluster loss. They also rejected the hypothesis that the *Hox* cluster duplications had occurred within a relatively short period of time (as suggested by Kappen and Ruddle (1993)), and showed that the *HoxD* cluster is the oldest cluster, and the *HoxB* and *HoxC* clusters are the most recently duplicated (Bailey *et al.*, 1997). If this hypothesis is correct, then if lamprey has only three clusters, it should be more likely to contain a *HoxD*-like cluster rather than a *HoxB* or a *HoxC*-like cluster (Meyer, 1998). The suggestion that the *HoxD* cluster is the oldest is also consistent with the observation that the *HoxD* cluster is the most deteriorated in mammals (Aparicio *et al.*, 1997).

The *Hox* clusters extend beyond the *Hox* genes themselves. There are at least 28 gene families that are associated with the *Hox* clusters on at least one of the *Hox* chromosomes, although only the collagens are associated with all four of the *Hox* chromosomes in human. These clusterings have been shown to have a statistically significant tendency to remain linked to the *Hox* genes (Ruddle *et al.*, 1994b), although little is known about the precise gene order or the distances between genes within their associated clusters. Although phylogenetic analysis of the non-*Hox* genes seems to imply that these genes were not duplicated along with the *Hox* genes (Hughes and Friedman, submitted), the fact that these genes are also seen to cluster with the *Hox* genes in *C. elegans*, *Drosophila* (Ruddle *et al.,* 1994a), *Fugu* (McLysaght *et al.*, submitted) and zebrafish (Amores *et al.*, 1998) indicates that the clustering of the non-*Hox* genes with the *Hox* genes themselves is part of an ancestral linkage pattern. It has been hypothesized that the retention of these ancestral linkages may have some biological role (Ruddle *et al.*, 1994b). It may be significant that many of these genes have developmental roles (such as *Evx* and *Dlx*), and that some have functional interactions with the *Hox* genes (such as the retinoic acid receptors and keratins). This may imply that there is a functional advantage afforded by the conserved linkage of these clusters, perhaps due to the possibility of transcriptional read-through or of multigene regulation by enhancers (Bentley *et al.*, 1995), although duplicates of linkage groups associated with the *Hox* genes have also been found on non-*Hox* clusters (Koh and Moore, 1999).

Examination of the genes associated with each cluster may make it easier to determine the history of the *Hox* clusters, and the relationship of clusters between species. For example, in zebrafish, there are seven *Hox* gene clusters. There are two representatives of all *Hox* clusters, except for *HoxD* which is only represented once. There are, however, two linkage groups for the genes associated with the remaining *HoxD* cluster, which implies secondary loss of the *HoxD* duplicate (Amores *et al.*, 1998). *Dlx* genes are another gene family associated with the *Hox* clusters, which are present in vertebrates as tandem duplicates. However, the tandem *Dlx* pair which should be associated with the *HoxC* cluster seems to have been lost from the human genome (Stock *et al.*, 1996). One interpretation of this could be to suggest that one of the rounds of duplication events leading to the formation of four clusters in the vertebrate genome was a chromosomal duplication as part of a genome duplication event, while the other was a chromosomal segment duplication.

Expansion of the *Hox* gene family appears to have been a key event in vertebrate evolution, perhaps permitting the evolution of greater complexity in vertebrate embryonic development (Pendleton *et al.*, 1993; Garcia-Fernàndez and Holland, 1994). The *Hox* clusters, along with the non-*Hox* genes which may be preferentially associated with them (Ruddle *et al.*, 1994b), are probably the most cited example of genes whose organization supports the hypothesis of two rounds of genome duplication in vertebrates (Holland *et al.*, 1994). A possible reason that these clusters are so evident may be because of their highly important developmental role and the fact that not only does their clustered organization seem to be important to their function (Duboule and Dollé, 1989), but also that the associated non-*Hox* genes may have an adaptive biological function. Such a tendency to retain linkage relationships may not extend to other regions of the genome, and other clusters may be more difficult to find.

A further interesting development is the discovery of a ParaHox cluster in amphioxus, which points to an even older duplication around 520 Mya (Dickman, 1997; Brooke *et al.*, 1998). The three *Hox* genes that make up the ParaHox cluster were identified by cloning *Gsx*, *Xlox* and *Cdx* genes, homeobox genes that in mammals do not form part of the Antp-class *Hox* clusters. Brooke *et al.* (1998) proposed that the second cluster originated before an ancestor of amphioxus, and was instrumental in the creation of the endoderm. Interestingly, Pendleton (1993) suggested that there were two *Hox* clusters in amphioxus based on the fact that five of the paralogy groups contained two members. There are five "orphan" *Hox* genes in the mammalian genome, the three examples here, and *Evx* and *Mox*. It would be interesting to see if the ParaHox cluster can be extended to include these genes also, (which were not cloned by Brooke *et al.*).

1.7.2 The MHC regions on HSA 1/6/9/19

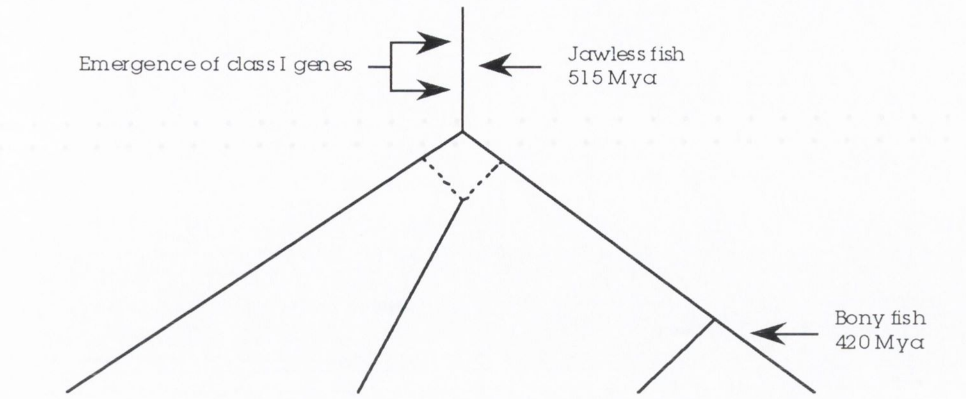
Gene duplication has played a major role in the evolution of the vertebrate immune system. The immunoglobulins, T-cell receptors and MHC gene families are all evolutionarily related (Hughes and Yeager, 1997; Nei *et al.*, 1997) and all derive from a single respective common ancestor via repeated rounds of gene duplication (Hood *et al.*, 1985; Klein and O'hUigin, 1993).

The major histocompatibility complex (MHC), which encodes highly polymorphic leukocyte antigens (HLA) responsible for antigen presentation to T cells (Mizuki *et al.*, 1997), is a region spanning over 4 Mb on 6p21.3, contains over 100 genes and is important in self-nonself discrimination. The MHC is divided up into three regions. The HLA class I and class II antigens (encoded in the MHC class I and class II regions, respectively) are involved in the genetic control of the immune response. The MHC class I genes are located nearest to the telomere, the MHC class II genes nearest to the centromere. The MHC class III genes are positioned between these, along with other non-MHC genes, some of which have functions related to the immune system, such as complement components, heat shock proteins and genes involved in antigen processing (Campbell and Trowsdale, 1993). The class I and class II regions of the human MHC region have been mapped to different linkage groups in zebrafish, and the class III region in *Fugu* contains genes that are not present in the human class III region. Because of this, it has been proposed that the class I and class II regions in the human MHC region were brought together by chromosomal translocations, during which process the class III region was formed (Bingulac-Popovic *et al.*, 1997; Lim and Brenner, 1997).

It has also been postulated that the MHC class II was the first class of MHC genes to evolve (Hughes and Nei, 1993). The ancestral class II was composed of an immunoglobulin-like domain, a membrane-anchoring domain, and a peptide binding domain (Klein and O'hUigin, 1993). This was then tandemly duplicated, first to give the ancestral class IIa and class IIb regions, then again resulting in four ancestral clusters side by side. One pair of IIa and IIb regions lost an immunoglobulin-like domain and a membrane-anchoring domain, and gave rise to the class Ia region, while the other pair became the class II region. The class Ib gene, a ß2-microglobulin gene, is not linked to the main MHC. Class I and class II molecules exist in all vertebrate phyla except agnathan fishes (Klein and O'hUigin, 1993), implying that the divergence of the different regions occurred around 400 Mya. Class IIa and IIb genes are estimated to have diverged approximately 500 Mya (Hughes and Yeager, 1997). The HLA-B and HLA-C genes within the class I region are products of an extended segmental duplication between 44 and 81 Mya (Kulski *et al.*, 1997). Many HLA pseudogenes have been found within the class I and class II regions (Beck *et al.*, 1996), consistent with the birth-and-death model of evolution (caused by ongoing gene duplication and dysfunctional mutation) proposed for these highly variable gene families (Nei *et al.*, 1997). The enormous diversity within the immunoglobulin heavy- and light-chain gene families (Hunkapiller and Hood, 1989) has also been shown to be generated by divergent evolution due to diversifying selection and evolution by the birth-and-death process (Ota and Nei, 1994; Sitnikova and Nei, 1998).

Some genes present within the MHC class III region have been shown to have counterparts elsewhere in the genome. Members from gene families present in the MHC region on HSA 6p21.3 (MMU 17) have been found on HSA 9q33-q34 (MMU 2) and HSA 1q21-q25 (MMU 1 and MMU 3) (Katsanis *et al.*, 1996; Kasahara *et al.*, 1997). An additional paralogous segment on HSA 19p13.1-13.3 (broken up in mouse on MMU 7, 10 and 17; DeBry and Seldin, 1996) has also been suggested (Kasahara *et al.*, 1997). Based on the existence or non-existence of a eukaryotic-like immune system in various classes of vertebrates, it has been proposed that this group of genes was duplicated twice as a block early in vertebrate history, perhaps as part of a polyploidization event, once after the divergence of jawless fishes and once at the emergence of cartilaginous fish (Figure 1.3; Kasahara *et al.*, 1996; Kasahara, 1997; Kasahara *et al.*, 1997).

If the homologous clusters did evolve by chromosomal or genomic duplication, then phylogenetic analysis of the gene families involved should highlight this fact. This was tested by Hughes (1998) and Endo *et al.* (1997). The genes involved in these regions duplicated at widely different times, spread out over at least 1.6 billion years (see Table 1.1). Some are estimated to have duplicated before the divergence of eukaryotes and eubacteria while others appear to have duplicated early in vertebrate history (Endo *et al.*, 1997; Hughes, 1998). Out of the 11 gene pairs that have been proposed on HSA 6 / HSA 9, six may have had a simultaneous origin before the divergence of vertebrates and may have been duplicated together. For three of the six gene pairs on HSA 6 and 9, there is a third copy on HSA 1 which is in each case more closely



6p23 (HLA)	19p13.1-p13.3	1q21-q25	9q33-q34
		ABL2	ABL1
		SPTA	SPTAN1
	LMNB2	LMNA	
	VAV1		VAV2
MHC class I		CD1	
HSP70-1			GRP78
HSP70-2			
HSP70-HOM			
C4	C3		C5
CYP21	CYP2		
TNX		TNR	TNC
PBX2		PBX1	PBX3
NOTCH4 (INT3)	NOTCH3	NOTCH2	NOTCH1
TAP1			ABC2
TAP2			
PSMB9			PSMB7
PSMB8			
LMP7			
LMP2			
RING3			RING3L
COL11A2			COL5A1
RXRB		RXRG	RXRA

Figure 1.3. Genes involved in the paralogous regions on HSA 1/6/9/19. Gene order is not conserved between them, but are listed in groups as above to allow easier visualization of the orthologous relationships. The phylogenetic tree indicates the presumed relationship of the four chromosomes involved: HSA 1 and 9 seem to be more similar to each other than either is to HSA 6 or 19. The relationship of HSA 19 to the other three chromosomes remains unknown. The regions are thought to have been formed by two genome duplications, the first one occurring in a common ancestor of jawed fishes after its separation from the lineage leading to the jawless fish, the second before bony fish diverged. Adapted from Kasahara *et al.*, 1997, Endo *et al.*, 1997, and Smith *et al.*, 1999.

gene family	Estimated age (Mya)		
	HSA 6 vs HSA 9		HSA 1 vs HSA 9
	Hughes	Endo *et al.*	Hughes
RXR	648	268 - 372	596
COL	561	258 - 344	
RING		331 - 450	
TAP/ABC	**2140**		
NOTCH	**1896**	364 - 909	610
PBX	612	580	599
C3/4/5	579	161 - 279	
HSPA	**1604**		
TEN	696		
PSMB	**2268**		

Table 1. 1. Dates of divergence calculated for members of gene families involved in the paralogous regions on HSA 6, 9 and 1. All single dates calculated by Hughes (1998) refer to genes on HSA 6 and 9. All second dates refer to genes on HSA 1 and 9. All dates calculated by Endo *et al.* (1997) are for genes on HSA 6 and 9. Gene familes with very ancient duplication dates are highlighted in bold. There appears to be no common ancestor for LMP2/7.

related to the HSA 9 paralogue than it is to the HSA 6 paralogue (Katsanis *et al.*, 1996; Hughes, 1998).

Smith *et al.* (1999) proposed that an ancestral cluster duplicated first to give the precursors for the paralogous clusters on HSA 6 / HSA 19 and on HSA 1 / HSA 9 (although the relationship of the region on HSA 19 to the other three regions is ambiguous; Kasahara *et al.*, 1997) and then again to give the four clusters present in the genome now. Each duplication was followed by extensive gene loss. To explain the widely differing divergence dates, they postulated various tandem duplications prior to the origin of vertebrates dating from as far back as the eukaryote-prokaryote split, where differential silencing of the duplicates leads to misallocation of paralogy (see Section 6.1, p. 99).

Hughes (1998) decisively rejected the hypothesis that all 11 of these gene families were involved in a block duplication and instead suggested that those gene families with ancient duplication dates (ABC, PSMB, NOTCH and HSP70, highlighted in bold in Table 1.1) had all undergone separate duplication events, and had all independently translocated to HSA 6 and 9, possibly because there was some selective functional advantage to the clustering of these genes.

Endo *et al.* (1997) put forward a complicated model (Figure 1.4) to explain the paralogous regions on HSA 6 and 9. They suggested that the four most anciently duplicated gene families were originally clustered on a primordial chromosome, and duplicated together. An insertion of members of the other seven gene families occurred on one of the duplicated chromosomes, between the HSPA and NOTCH homologues, followed by a regional duplication, whereupon the inserted region on one of the chromosomes was likewise inserted into the similar region on the other duplicated chromosomes, presumably by a process of recombination. This was followed by various rearrangements and single gene duplication events. However, if the initial

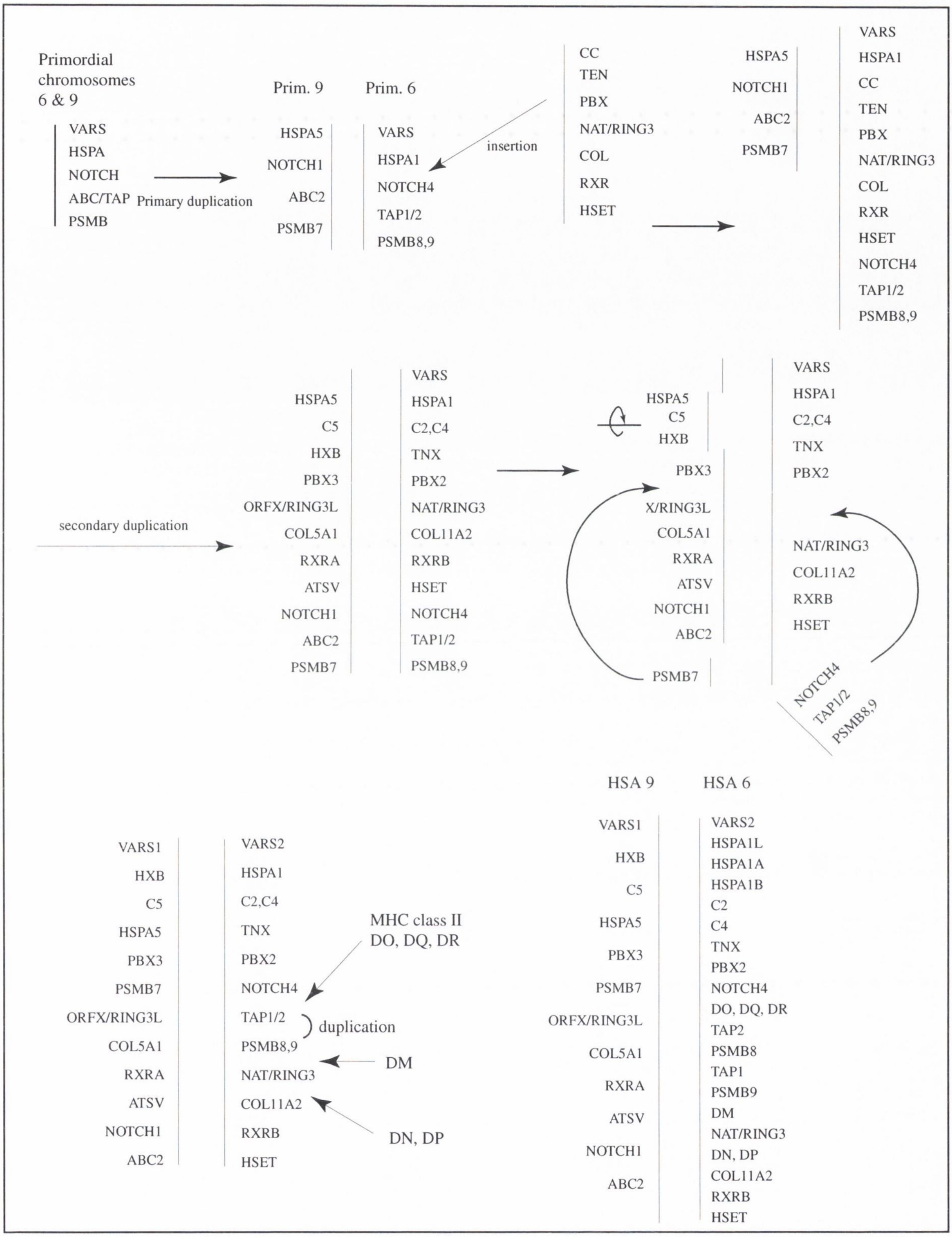


Figure 1.4. The model proposed by Endo *et al.* (1997) for the evolutionary history of the MHC regions of HSA 6 and 9. The common ancestral band region of 6p21.3 and 9q33-q34 contained the primordial *HSPA, NOTCH, ABC/TAP* and *PSMB* genes. The region was duplicated and a segment containing *CC, TNX, PBX, NAT/RING3, COL, RXR* and *HSET* was inserted into the region between *HSPA* and *NOTCH* on one of the duplicated chromosomal segments. A second duplication event occurred, followed by numerous rearrangements and tandem duplications. Adapted from Endo *et al.* (1997).

duplication event occurred about 1 Gya and the second regional duplication event occurred around 600 Mya, then it would be highly improbable that the "sister" chromosomes were still similar enough to be able to undergo such a recombination event.

It is more difficult to determine the history of the paralogous regions on HSA 1/6/9/19 than it is for the *Hox* gene clusters, because there is little or no MHC class III data available from any other species. These regions appear to have undergone much rearrangement in the mouse genome (DeBry and Seldin, 1996).

1.7.3 FGFR REGIONS ON HSA 4/5/8/10

A third possible example of a quadruplicated region was described recently by Pébusque *et al.* (1998). There is detailed information for a chromosomal region on HSA 8p12-21 (Adélaïde *et al.*, 1998). By searching the genome for paralogues of the genes in this region, they identified seven gene families (plasminogen activators, ankyrins, fibroblast growth factor receptors, adrenergic alpha and beta receptors, intermediate-early transcription factors, vesicular monoamine transporters and lipoprotein lipases). The quadruplicated region described is centred on the four members of the FGFR family which belong to the superfamily of tyrosine kinase receptors. Each of these genes is located close to adrenergic receptor genes on human chromosomes 4p16, 5q33-35, 8p12-21 and 10q24-26. A member of each of the seven gene families identified in the 8p12-21 region also has members on 10q24-36. However, there are only two more genes that can be added to the region on HSA 5, and the fourth member of the PLAT, PLAU, FXII group (HGFA) has recently been mapped to 4p16 by Miyazawa *et al.* (1998), which is a perfect fit to the prediction by Pébusque *et al.* Synteny, but not gene order, is conserved within these regions (Figure 1.5).

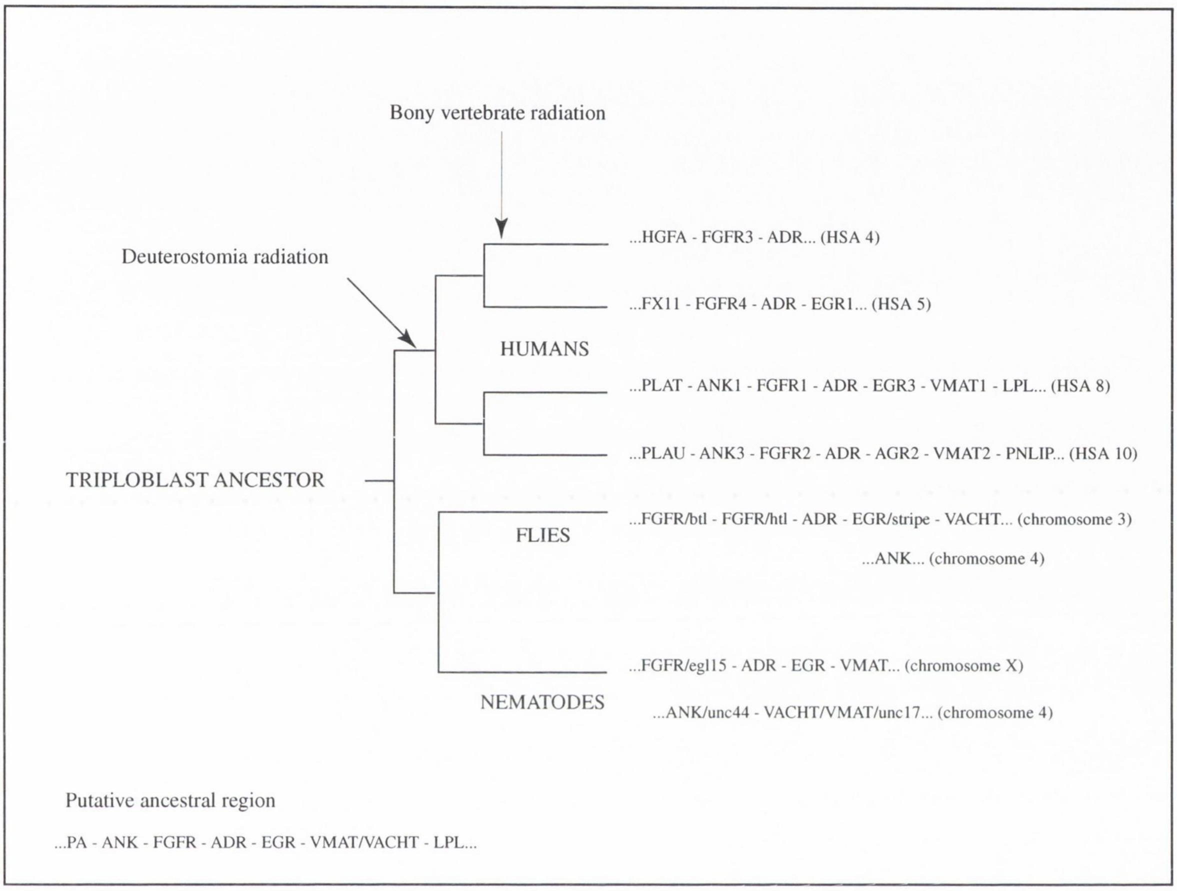

Bony vertebrate radiation
Deuterostomia radiation
...HGFA - FGFR3 - ADR... (HSA 4)
...FX11 - FGFR4 - ADR - EGR1... (HSA 5)
HUMANS
...PLAT - ANK1 - FGFR1 - ADR - EGR3 - VMAT1 - LPL... (HSA 8)
...PLAU - ANK3 - FGFR2 - ADR - AGR2 - VMAT2 - PNLIP... (HSA 10)
TRIPLOBLAST ANCESTOR
FLIES
...FGFR/btl - FGFR/htl - ADR - EGR/stripe - VACHT... (chromosome 3)
...ANK... (chromosome 4)
...FGFR/egl15 - ADR - EGR - VMAT... (chromosome X)
NEMATODES
...ANK/unc44 - VACHT/VMAT/unc17... (chromosome 4)
Putative ancestral region
...PA - ANK - FGFR - ADR - EGR - VMAT/VACHT - LPL...

Figure 1.5. Reconstruction of events leading to the FGFR regions on HSA 4/5/8/10. Gene order is not known precisely except for HSA 8p12-21, so linkage groups are drawn so as to facilitate orthologue comparison. Chromosomal localization is given in brackets for each region. Taken from Pébusque *et al.* (1998).

Where data are available, phylogenetic analysis indicates a 1:4 relationship between invertebrate and mammalian sequences for these genes. Phylogenetic trees do not, however, allow determination of whether the two duplications occurred at different times, or within a short space of time. Pébusque *et al.* argue that these duplicated genes arose before the bony fish divergence but after the echinoderm / chordate split, but they did not use molecular clocks to estimate dates for each gene. Syntenous regions are found in *Drosophila* (on chromosome 3, ANK moved to chromosome 4), and *C. elegans* (on chromosomes X and 4), neither of which include members of the plasminogen activator or lipoprotein lipase gene families. Because the ankyrin gene is not located on the same chromosome as the FGFR-ADR linkage group in either *Drosophila* or *C. elegans*, perhaps ANK was moved to the FGFR-ADR linkage group after the chordate lineage diverged, and was then involved in a regional duplication of that region. It is possible that some adaptive function has caused these ancestral linkages to remain linked: plasminogen activators, fibroblast growth factor receptors and intermediate-early transcription factors are all involved in some stage in the FGF stimulatory pathway.

CHAPTER 2 —

CONSERVED DUPLICATED SEGMENTS BETWEEN THE HUMAN AND MOUSE GENOMES

2.1 AIM

Lineages deriving from the same ancestor should contain conserved chromosomal segments, delineated by (random) breakpoints. If genome duplication took place in such a common ancestor, then we should be able to see not only orthologous segments between them, but also paralogous segments within each species which are mirrored in the other species. Using data from the human and mouse genomes, we tried to find such conserved duplicated segments, draw trees using the genes involved, and from this to put a possible date on when genome duplication may have occurred in the common ancestor of the human and mouse lineages. The rationale behind this was that genes making up conserved orthologous segments between the human and mouse genomes must have been in place before these species diverged, and consequently should be more reliable markers of a genome duplication (assumed to pre-date the human / mouse divergence) than either the human or mouse map alone.

2.2 INTRODUCTION

2.2.1 USES OF COMPARATIVE GENE MAPPING

The extent of genome conservation is of great practical and evolutionary interest (O'Brien and Graves, 1991; O'Brien *et al.*, 1999). The uses of comparative gene mapping in mammals are at least three-fold:

1. It helps identify new genes or identify homologues of disease traits mapped in other species. For example, a segment on MMU 1 containing the Pax3 gene is homologous to a segment on HSA 2, which was found to contain the equivalent human gene which leads to the expression of the Waardenburg syndrome (Tassabehji *et al.*, 1992). The identification of chromosomal regions with both conserved gene order and synteny is important for such gene hunting purposes. At present, the level of mapping detail between human and mouse is still insufficient to provide a complete description of all chromosomal blocks with conserved synteny and gene order.

2. It gives insight into the forces guiding genome organization and evolution (Clark, 1999). Are synteny and linkage conservation between species due to chance ("frozen accidents"; Ohno, 1973), or do the functions of the genes within any given segment play some role in whether some genes must remain in close linkage to one another? For example, retinoic acid receptors are functionally related to the *Hox* clusters and the keratin clusters, and are also found in close proximity to them (Boncinelli *et al.*, 1991; Nadeau *et al.*, 1992). Are the breakpoints at which chromosomal rearrangements occur random, or are some genetic regions more prone to rearrangement than others (Purandare and Patel, 1997)? Regions of DNA which contain repeated sequences may facilitate illegitimate inter- and intra-chromosomal recombination events (Dutly and Schnizel, 1996; Carver and Stubbs, 1997; Kehrer-Sawatzki *et al.*, 1997). Are some stretches of DNA more likely to be duplicated or translocated? Evidence suggests that this may be the case, depending on the nature of the DNA sequences involved (Eichler, 1998; Pennisi, 1998). Do population factors such as generation time and effective

population size play a role in determining the rate of disruption? The mouse genome is three or fourfold more rearranged with respect to the human genome than are either the feline or bovine genomes (O'Brien, 1991), and mice also have a shorter generation time and potentially a larger population size than the larger mammals.

3. It allows an estimation of the rate of chromosomal rearrangement within and between species. Despite strong selection against reciprocal translocations, inter-chromosomal rearrangements occurred approximately fourfold more often than inversions and other intrachromosomal rearrangements in lineages leading to humans and mice (Ehrlich *et al.*, 1997; Nadeau and Sankoff, 1998). However, this estimation excludes the possibility of many small inversions, which may be an important part of genome evolution (Eppig, 1996). This disrupted linkage but not synteny is seen as rearranged segments embedded within larger conserved segments, as though a small inversion has occurred within a long conserved segment (Nadeau, 1989). In their comparative map between the human and mouse genomes, DeBry and Seldin tried to minimize their inclusion of such rearrangements (DeBry and Seldin, 1996). But the point of DeBry and Seldin's work was to help identify unknown genes (see point 1 above). If a segment in one species includes a gene of interest, then we can sequence through the whole segment so that even if slight rearrangements have taken place, the gene that we are looking for, if it has not been deleted or involved in some other rearrangement process, should be easily identifiable. Recent evidence has indicated that these 'waltzing genes' are common in yeast (Seoighe *et al.*, in press), but examples have been found in higher eukaryotes also: although synteny is strongly conserved in the Huntington disease region of HSA 4p16.3 in human, and on MMU 5 in mice, several genes on a 1.5 cM segment (Idna, Dagk4, Pdeb) have been transposed within that segment; there are two small rearrangements between HSA 9 and MMU 2, marked by Spna2-Abl and Dbh-Rxra (Nadeau and Sankoff, 1998). There are four small rearrangements within a 10 cM segment of HSA 5 and MMU 11, where the two segments showing synteny but not linkage conservation have lengths of 1.5 cM and 1.7 cM (Watkins-Chow *et al.*, 1997) There have been three transpositions of three

gene-rich segments and a local inversion within a 23 cM conserved syntenic region on HSA 19q and MMU 7, involving 42 markers, although within these segments, gene content, order and spacing are remarkably well conserved (Stubbs *et al.*, 1996).

At high resolution, many syntenically homologous regions are shown to have undergone significant rearrangements (Carver and Stubbs, 1997). For example, genes from a 6.5 Mb region of 22q11 (the DiGeorge syndrome / velocardiofacial syndrome region) are seen to have homologues on MMU 6, 10 and most impressively on MMU 16. The 19 genes and nine EST groups from 22q11 which have homologues on MMU 16 form at least four regions where gene order appears to be conserved, but the relative order and orientation of these regions is different in human and mouse (Botta *et al.*, 1997; Puech *et al.*, 1997; Sutherland *et al.*, 1998). It has been proposed that the high number of duplicated regions and low-copy repeat families in this region may cause instability in this region (Puech *et al.*, 1997).

Ohno's law states that chromosomal rearrangements involving the X-chromosome and autosomes are strongly selected against (Ohno, 1967), with one possible exception of CLCN4 having been transposed to MMU 7 in lab mice (Palmer *et al.*, 1995). There have, however, been numerous rearrangements within the human and mouse X chromosomes (Blair *et al.*, 1994; Dinulos *et al.*, 1996).

2.2.2 RATES OF REARRANGEMENT IN MAMMALIAN LINEAGES

The recent rapid accumulation of mapping and sequencing data for mammalian genomes has allowed the investigation of orthologous gene order and paralogous duplicated regions and other issues concerning the genomic events during the early evolution of vertebrates. The most detailed mammalian map data come from the human and mouse genomes. The number of chromosomal segments conserved during the divergence of two species can be used to measure their genomic distance. Nadeau and Taylor (1984) estimated interchromosomal exchange rates based on the rearrangement of chromosomal segments in the human versus mouse genomes. They found 13 conserved linkage groups (on comparison of 83 homologous loci) between the two genomes, and estimated that there had been 178 ± 39 chromosomal rearrangements since the divergence of the lineages leading to humans and mice, the average length of a segment being 8.1 ± 1.6 cM (Nadeau and Taylor, 1984). As the human-mouse comparative map became more dense, although the number of conserved segments discovered between human and mouse increased, these initial estimates of the number of rearrangements and the average length of a segment did not change significantly (see Table 2.1).

The rate of change in the gene maps has not been uniform among the mammalian orders. Rates of synteny disruption vary approximately 25-fold among mammalian lineages (Ehrlich *et al.*, 1997). However, comparison of orthologous chromosomal regions reveals linkage conservation for extensive parts of mammalian chromosomes (Eppig, 1996). To facilitate the comparison of genomes from different mammalian lineages, O'Brien *et al.* proposed a list of reference anchor loci (comparative anchor tagged sequences (CATS)) spaced approximately 5 - 10 cM apart, offering relatively even coverage over all chromosomes (O'Brien, 1991; O'Brien *et al.*, 1993; Lyons *et al.*, 1997). Evidence seems to indicate that the linkage map has been broken up to a greater extent in the rodent map than in that of the primates, compared to an ancestral mammalian genome, with the human genome having undergone the least number of linkage group changes (Lundin, 1993; Ohno, 1993; Graves, 1996; Ehrlich *et al.*, 1997).

While large chromosomal segments of mouse and man have remained relatively intact (Copeland *et al.*, 1993) (e.g., a region spanning 227 kb on MMU 6 and 223 kb on HSA 12p13 has been shown to display conserved gene number, order and orientation;

YEAR	AUTHOR	# genes used	# linkage groups	# rearrange-ments	average length (cM)
1984	Nadeau and Taylor	83	13	178 ± 39	8.1 ± 1.6
1989	Nadeau	241	26	138 ± 32	10.1 ± 2.2
1991	Nadeau	425	>100	-	7.4 ± 1.7
1993	Copeland *et al.*	917	101	144	8.8
1996	DeBry and Seldin	1416	181	-	-
1997	Ehrlich *et al.*	1152	91	122*	-
1997	Sankoff *et al.*	1423	130	181	-

Table 2.1. Estimates of the number of rearrangements and average length of segments between the human and mouse genomes. These estimates have not changed significantly as more genes are put onto the human-mouse comparative map.

* Intrachromosomal effects not included in this calculation

Ansari-Lari *et al.*, 1998), for most chromosomal regions, the gene order has been rearranged by multiple translocations and/or inversions. For example, on analysis of the proximal MMU 9 linkage map, Seldin (1991) concluded that segments of MMU 2, 7, 9 and 19 and HSA 11, 15 and 19 derived from a single common ancestral chromosome, where four rearrangements (two translocations and two transpositions) had occurred in the human map to give the present day arrangement and six rearrangements (three translocations, three transpositions) had occurred in mouse. In a more extreme example, HSA 6p has loci on MMU 4, 9, 10 and 13, and HSA 6q has loci on MMU 9, 10 and 17 (Davisson *et al.*, 1991).

Comparison of the mouse and rat genomes indicates that rearrangements have also occurred within the rodent lineage. There are at least 49 conserved autosomal segments between rat and mouse, of which 41 have two or more markers. The estimated mean size of a conserved segment is 39 ± 6 cM, and it is thought that most, if not all, of the conserved segments have been identified (Watanabe *et al.*, 1999). Some of the rat chromosomes span whole mouse chromosomes (e.g., RNO 5 (RNO: R*attus* NO*rvegicus*) is equivalent to MMU 4, as is RNO 8 to MMU 9. Both of these regions also seem to have highly conserved gene order).

There are 109 conserved segments observed between the human and rat genomes (Watanabe *et al.*, 1999). It is clear that while large regions of conserved synteny do exist between the two genomes (Remmers *et al.*, 1992), many rearrangements have also taken place between them. For example, 71 genes on RNO 1 form six syntenic segments in mouse (on MMU 7, 10, 13, 17, 19), but 13 in human, over eight chromosomes.

That the rate of rearrangement within the rodent genomes is exceptionally rapid is highlighted by comparison of the human genome with other mammalian genomes. The genome organization of the cat can be reorganized to the human status by as few as 13 translocation steps (O'Brien *et al.*, 1999), implying a single translocation every 10 - 12 Myr (O'Brien *et al.*, 1997). For example, HSA 11 is syntenic to a single chromosome B4 in cats, but is composed of five different segments on four mouse chromosomes (MMU 2, 7, 9, 19), a number that increases to 20 distinct ordered linkage segments at a higher map resolution.

There is also a greater conservation of synteny between the cattle and human genomes than between the human and mouse genomes (Womack and Moll, 1986), although it is not as great as originally proposed (Womack and Kata, 1995). The bovine maps show fewer and larger blocks of synteny with human than the mouse maps (Georges and Andersson, 1996; Andersson *et al.*, 1997). For example, only one rearrangement has been observed between HSA 13q and BOV 12. In contrast to this, there are two translocations separating this conserved segment into three linkage groups in mouse (Sun *et al.*, 1997). A minimum of 40 rearrangements have taken place between the bovine and human genomes, yielding 70 homologous segments (Womack and Kata, 1995). However, within the boundaries of conserved synteny between cattle and human, it is likely that extensive disruptions of conserved linkage remain to be discovered (Johansson *et al.*, 1995). Seven loci on 117 cM on HSA 2 and 86.9 cM in BOV 2 show conserved synteny but analysis has shown that at least three translocations have taken place between the two segments (Sonstegaard *et al.*, 1998).

2.3 DATA

In September 1995, the most comprehensive human protein database was that coordinated by The Institute of Genomic Research (TIGR) (http://www.tigr.org/tdb/) (Adams *et al.*, 1995). The TIGR human cDNA database contained tentative human consensus (THC) sequences, derived by combining 174,472 new partial cDNA sequences and full-length cDNAs from GenBank with 118,406 ESTs from the dbEST database, and assembled as for a shotgun sequence assembly project, yielding 29,599 THCs and 58,384 additional non-overlapping ESTs. Among these, there were 10,214 previously identified genes, or sequences with similarity to known genes. In theory, this dataset should be a non-redundant set of human protein sequences. Each gene has associated with it a unique human transcript (HT) number, a non-unique human gene (HG) number (since genes can have alternative transcripts) and DNA and protein sequences. The TIGR THC database entries also have cross-references to the GDB database, where GDB is the Genome Database at John Hopkins University which contains human map information (e.g., serum amyloid A1 (SAA1) is designated HT732, with HG732, and a GDB number of 120364).

Each HT number was used to query the TIGR database on the World Wide Web, using http://www.tigr.org/docs/tigr-scripts/egad_scripts/ht_report.spl?htnum=<ht#> (where <ht#> is the unique HT number for every human transcript), to obtain the relevant human transcript. From these we identified 4554 HTs which had associated GDB numbers. Mapping positions of the human genes and mouse genes were both taken from the Mouse Genome Database (http://www.informatics.jax.org/mgd.htm, or its mirror site http://mgd.hgmp.mrc.ac.uk/mgd.html), run by the Jackson Laboratory.

Human mapping data vary widely in their precision, with many locations being indicated by a range, rather than a specific position. Mapping positions of human genes were given in the MGD in cytogenetic units of chromosome bands and subbands (e.g., SAA1: 11p15.1 – p14).

Mouse genetic mapping data, measured in centimorgan units, tend to be more precise due to the ability to do crosses, although may still include error ranges of several centimorgans. Most of the mouse mapping data were originally compiled by the Mouse Chromosome Committee Report (Anonymous, 1996). The protein sequences of mouse genes which were annotated by MGD as homologous to sequenced human genes were taken from GenBank (release 93) (e.g., Saa1: MMU 7@23.5, MGD-MRK-14278, with a GenBank accession number of M17798). MGD gives a number of links to GenBank for each mouse gene. In general, only one of these links has a full-length cDNA. Where mouse sequences were not available, the corresponding rat sequences were used instead, if available.

2.4 METHODS

2.4.1 NEIGHBOUR-JOINING METHOD OF CONSTRUCTING PHYLOGENETIC TREES

Phylogenetic trees were constructed for each set of paralogues and their orthologues (i.e., sets of two human and two mouse sequences; Figure 2.1). The sequences were first aligned using CLUSTALW, then examined by eye, the tree calculated using the neighbour-joining method (Saitou and Nei, 1987), and the time of divergence between paralogues, or the time since duplication of the original gene, was estimated using equation 2.1.

$$\text{Relative time of divergence} = \frac{c + \tfrac{1}{2}(a+b+d+e)}{\tfrac{1}{2}(a+b+d+e)} \qquad \textbf{(Eq. 2.1)}$$

where a, b and d, e are opposing pairs of branches, and where c is the central branch (see Figure 2.1). This gives the time of gene duplication relative to the speciation time between human and mouse, which for convenience we took to be 100 Myr.

The neighbour-joining method of constructing a phylogenetic tree involves the calculation of the distances (pij, percent divergence between sequences i and j) between all pairs of sequence from a multiple alignment. All sites with gaps (in any sequence) were discarded, which excludes the most ambiguous parts of the alignments.

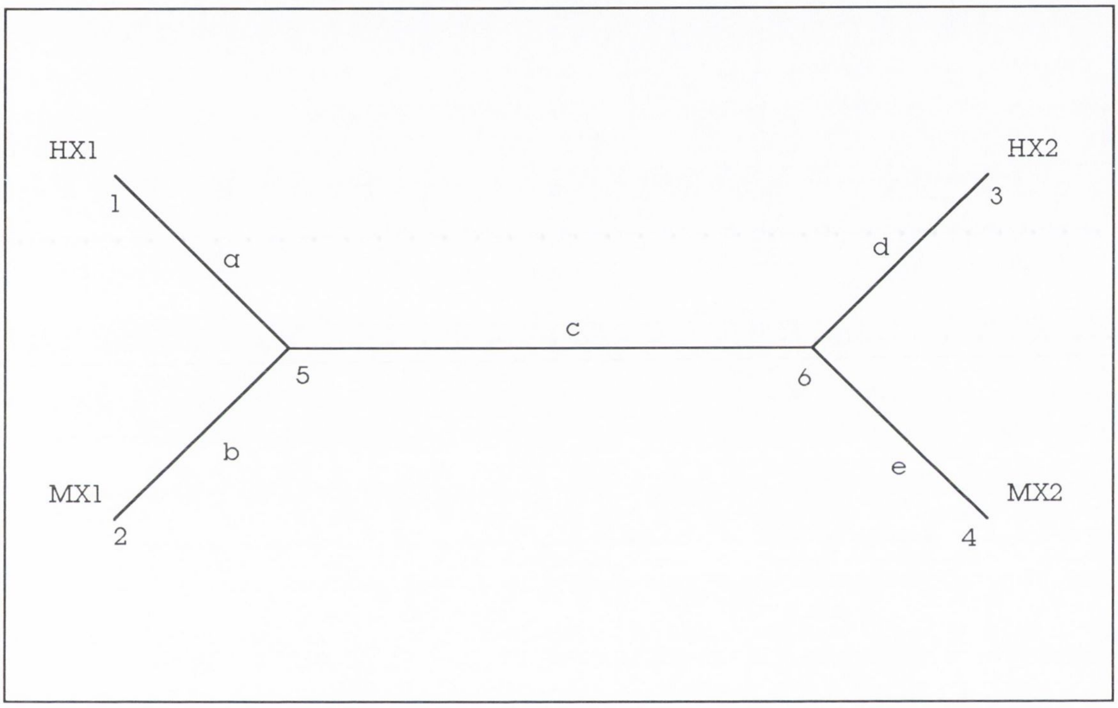


Figure 2.1. Tree relating human and mouse sequences. HX1 and MX1 are orthologues, as are HX2 and MX2. Branch lengths are indicated by the letters a, b, c, d and e. Nodes are labelled 1 - 6.

We used f = 19/20 as a correction factor. This is the Jukes-Cantor formula for proteins. Kimura's corrections for multiple substitutions were not used. From these pairwise distances, the corrected mean of amino acid replacements per site between sequences i and j (dij) can be calculated. Corrections for multiple substitutions were incorporated into the equation for the mean number of amino acid replacements per site (Eq. 2.2).

$$dij = -f(\log(1 - {}^{pij}\!/_{f})) \tag*{\textbf{(Eq. 2.2)}}$$

Using these values, the branch lengths of the tree can be calculated:

$$a = \tfrac{1}{2}d12 + \tfrac{1}{4}(d13 - d23 + d14 - d24) \tag*{\textbf{(Eq. 2.3)}}$$

$$b = d12 - a \tag*{\textbf{(Eq. 2.4)}}$$

$$c = \tfrac{1}{4}(d13 + d23 + d14 + d24) - \tfrac{1}{2}(d12 + d34) \tag*{\textbf{(Eq. 2.5)}}$$

$$d = \tfrac{1}{2}d34 + \tfrac{1}{4}(d13 + d23 - d14 - d24) \tag*{\textbf{(Eq. 2.6)}}$$

$$e = d34 - d \tag*{\textbf{(Eq. 2.7)}}$$

where sequences 1, 2, 3, 4 refer to HX1, MX1, HX2, MX2, respectively, in Figure 2.1.

The variances of these branch lengths can be obtained using equations as follows. These were derived using Li (1989, 1990):

$$\begin{aligned} Va = {} & (\tfrac{1}{4}Vd12) + \tfrac{1}{2}(Vd15 - Vd25) + \tfrac{1}{16}(Vd13 + Vd14 + Vd23 + Vd24) + \\ & \tfrac{1}{8}(Vd16 + Vd26 - Vd35 - Vd45 - 2Vd56) \end{aligned} \tag*{\textbf{(Eq. 2.8)}}$$

$$\begin{aligned} Vb = {} & \tfrac{1}{4}Vd12 + \tfrac{1}{2}(Vd25 - Vd15) + \tfrac{1}{16}(Vd13 + Vd14 + Vd23 + Vd24) + \\ & \tfrac{1}{8}(Vd16 + Vd26 - Vd35 - Vd45 - 2Vd56) \end{aligned} \tag*{\textbf{(Eq. 2.9)}}$$

$$Vc = (\tfrac{1}{16}(Vd13 + Vd23 + Vd14 + Vd24 + 2Vd16 + 2Vd26 + 4Vd56 + 2Vd35 + 2Vd45)) - \tfrac{1}{2}(Vd15 + Vd25 + Vd36 + Vd46) + \tfrac{1}{2}Vd12 + \tfrac{1}{4}Vd34 \qquad \textbf{(Eq. 2.10)}$$

$$Vd = \tfrac{1}{4}Vd34 + \tfrac{1}{2}(Vd36 - Vd46) + \tfrac{1}{16}(Vd13 + Vd23 + Vd14 + Vd24) + \tfrac{1}{8}(Vd35 + Vd45 - Vd16 - Vd26 - 2Vd56) \qquad \textbf{(Eq. 2.11)}$$

$$Ve = \tfrac{1}{4}Vd34 + \tfrac{1}{2}(Vd46 - Vd36) + \tfrac{1}{16}(Vd13 + Vd23 + Vd14 + Vd24) + \tfrac{1}{8}(Vd35 + Vd45 - Vd16 - Vd26 - 2Vd56) \qquad \textbf{(Eq. 2.12)}$$

where:

$$Vdij = \frac{(pij(1 - pij))}{L(1 - {}^{pij}/_{f})^2} \qquad \textbf{(Eq. 2.13)}$$

and $Vdij$ is the variance of the number of amino acid replacements between sequences i and j, and L is the number of sites used in the comparison. Errors were calculated using variations of the equations described by Li (1989; 1990).

2.4.2 Finding conserved duplicated regions between human and mouse

We attempted to find duplicated chromosomal regions (DCRs) that are conserved in both human and mouse genomes, which could define ancestral conserved chromosomal linkage regions. To find these, sets of genes needed to be identified that are related as shown in Figure 2.2. To be able to define the timing of the divergence (and therefore of the duplication) of the gene families involved, we also had to be able to construct phylogenetic trees, based on the protein sequences of the relevant genes.

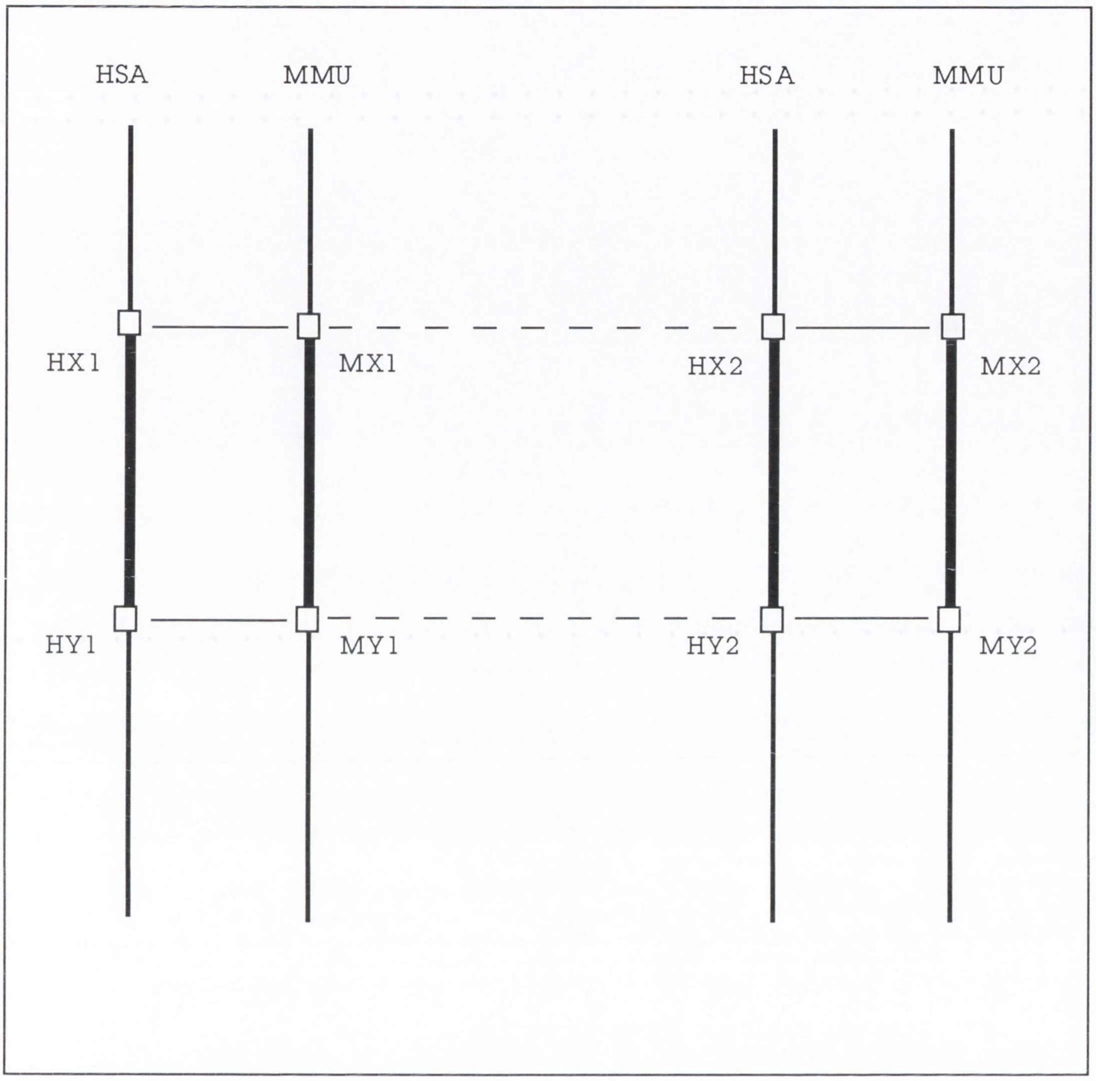


Figure 2.2. Orthology and paralogy relationships minimally needed to define a duplicated conserved region (DCR) conserved between human (HSA) and mouse (MMU). Each vertical line represents a chromosome or non-overlapping segment of chromosome. X and Y denote two unrelated gene families. X-type genes are related in sequence and usually in function, as are Y-type genes. The DCR is the region between X-type and Y-type genes (thickened lines) and is present in two copies in both species. Orthologous relationships between human and mouse are shown by solid horizontal lines. Dashed lines represent paralogous gene pairs, created by gene duplication prior to the human / mouse speciation event. In addition to the paralogous MX1 / HX2 and MY1 / HY2 pairs indicated, all other X1 / X2 and Y1 / Y2 pairs are paralogous.

Analysis of the data was carried out as shown in Figure 2.3. We searched for mouse homologues of each human gene in the TIGR THC database and then looked for their genetic map locations in MGD.

A set of human paralogues was obtained by performing BLASTP searches of all the human protein sequences in the TIGR database against each other. Cut-offs of a BLASTP score of over 100, a percentage identity of over 35%, and an alignment length threshold of over 35% of the shorter protein being aligned were imposed. This low BLASTP cut-off point was chosen because, although more spurious sequence similarities would be found, we were also more likely to find most of the true paralogues. Using only human paralogues which have orthologues in mouse, we initially (1.) put together a set of paralogues which were adjacent on a chromosome, as in Figure 2.2. To complement this, the conditions for finding a group of paralogous genes were relaxed (2.) to allow non-adjacency of at most one pair of paralogues in either or both species (Figure 2.4). The cut-off score was also changed to 300 to offset the relaxed criteria.

Genes were taken as belonging to a paralogous region if

a) genes on the same chromosome were less than 30 cM apart (for mouse mapping data), or less than two bands apart (for human mapping data). The total size of the mouse genome is 3.45 x 10^6 Mb or 1593.6 cM (Database of Genome Sizes: http://www.cbs.dtu.dk/databases/DOGS/index.html). 30 cM therefore corresponds to an interval of about 65 Mb. The human genome is 3.4 x 10^6 Mb and contains 542 cytogenetic bands at the level of resolution provided in MGD. Two bands therefore corresponds to (very roughly) 12.5 Mb. However, some bands span very large regions, and others are very small. In the cases where the human mapping positions were given as ranges (which generally span two or three bands), genes with overlapping ranges, or with ranges within two bands of each other were kept;

b) genes had scores over 100 or 300 as described above. The BLASTP alignments were also examined by eye;

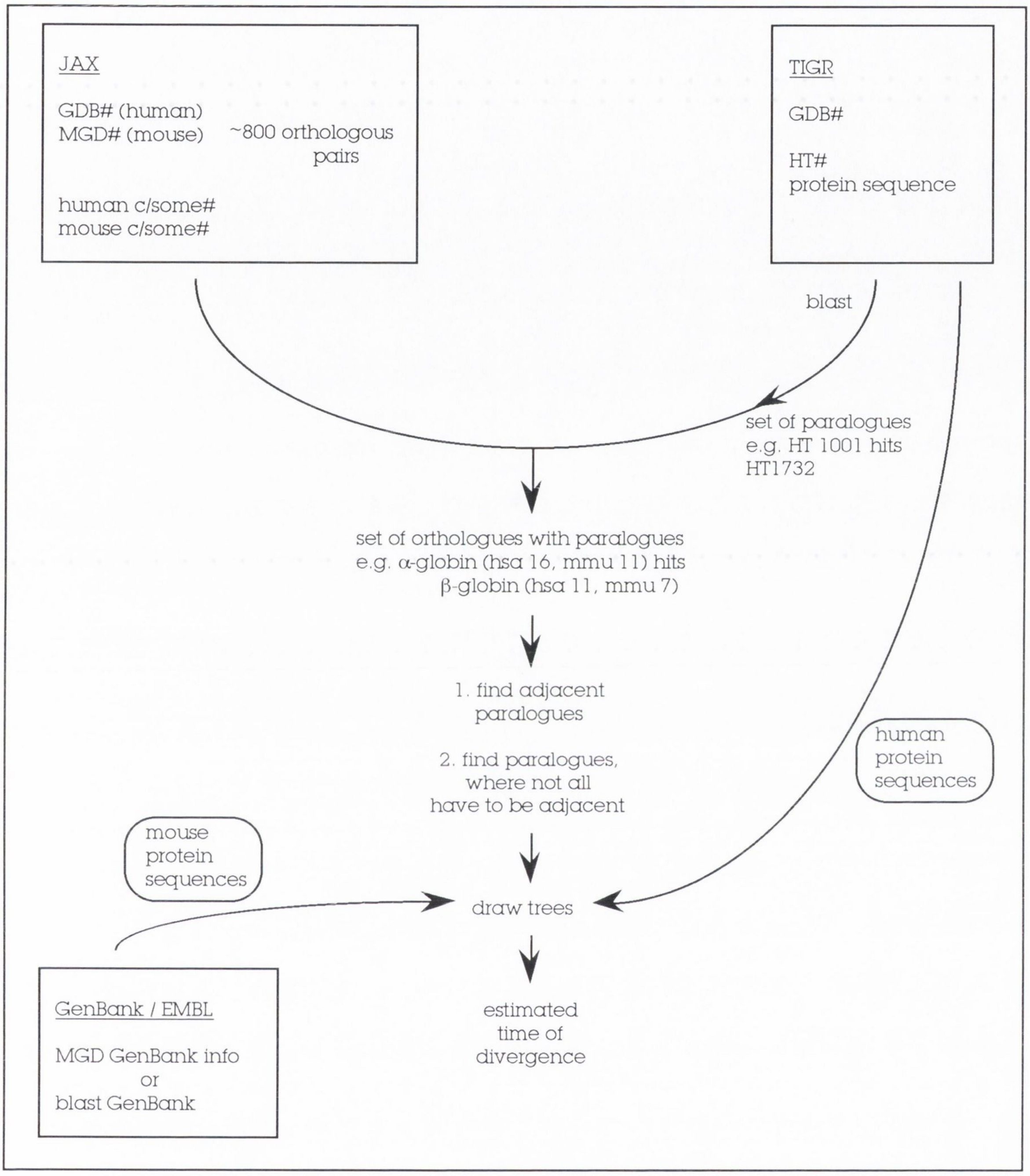


Figure 2.3. Flowchart of data collection process using the TIGR THC dataset. Human and mouse mapping information was obtained from the Jackson Laboratory database. The Institute of Genome Research provided the human protein sequences, and mouse sequence data, where available, was taken from the GenBank / EMBL databases.

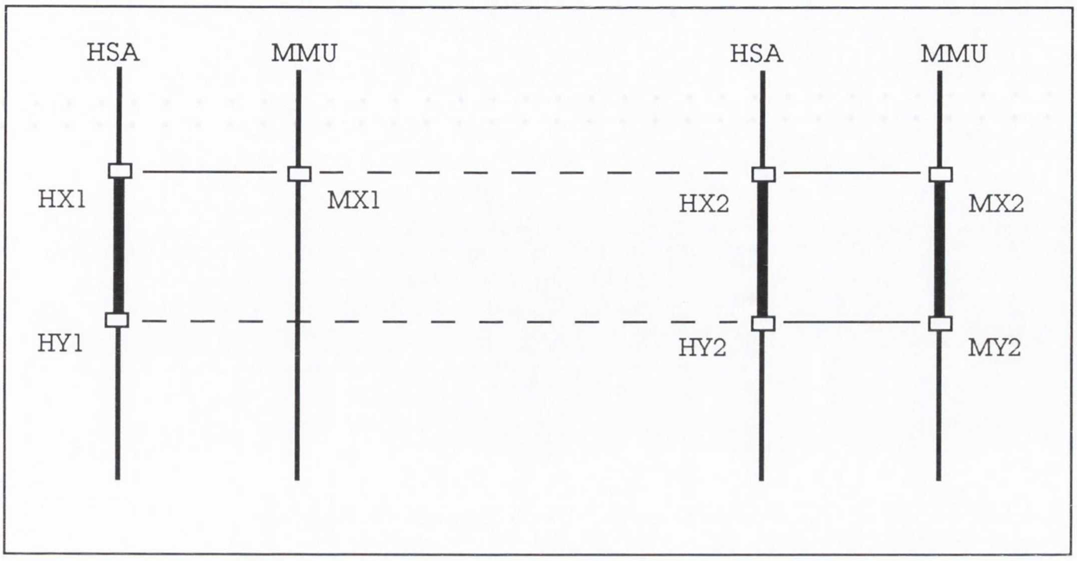


Figure 2.4A

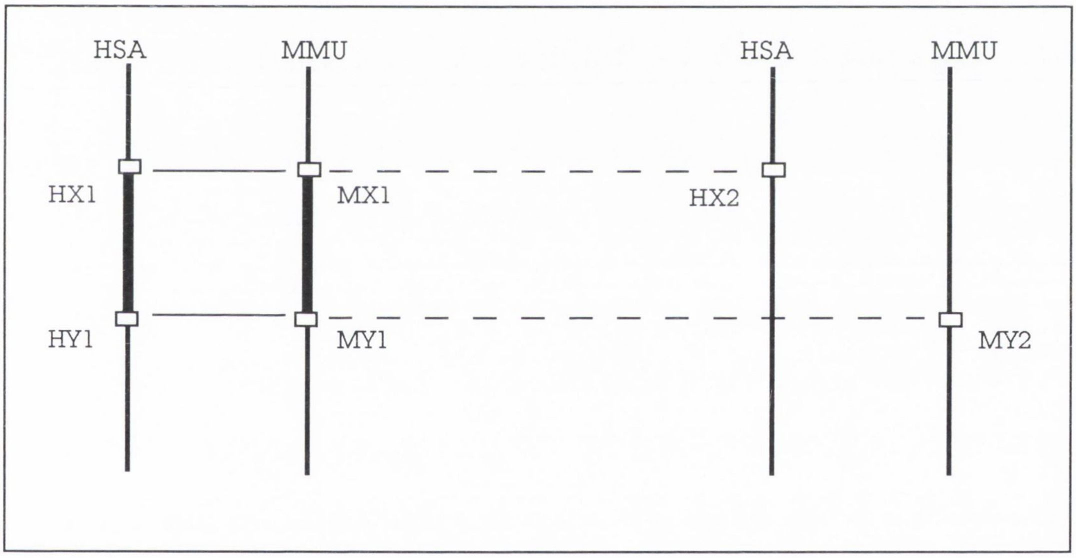


Figure 2.4B

Figure 2.4. Duplicated conserved regions after chromosomal rearrangements. Figure 2.4A reflects the situation where the segment carrying one of the genes defining a DCR (here, the original MY1 gene) has undergone a chromosomal rearrangement event, either moving the MY1 gene to another mouse chromosome, or resulting in its deletion, or where MY1 has simply not yet been sequenced. Similarly, in Figure 2.4B, two chromosomal rearrangements have taken place, once in the human genome, and once in the mouse genome (or again it may be due to lack of sequence). A conserved region still exists, but it no longer appears as being duplicated.

c) the times of divergence calculated for the two delimiting paralogous pairs were similar within an approximate error margin of ± 1 SD (standard deviation).

Table 2.2 summarizes the number of genes in the dataset at each stage of the verification process.

2.5 RESULTS FROM THE ANALYSIS OF TIGR DATA

Table 2.2 shows the initial number of human/mouse gene sets found (a maximum of eight genes per set - two paralogues and two orthologues each, for two unlinked loci), and the number of sets used in the final analysis, detailing how many sets were discarded, and on what basis. We found only four sets of human / mouse genes satisfying the scenario in Figure 2.2, where a) HX1 and HY2 (and all other X- and Y-type gene pairs within the same organism and on the same chromosome; X and Y denote two gene types with unrelated sequences) were less than 30 cM apart (or approximately less than two chromosomal bands for the human data); b) where HX1 and HX2 were true homologues, as judged by looking at sequence alignments; and c) where the times of divergence calculated for the original X and Y paralogue duplications were similar. Five sets were discarded on the basis of disparate, or very old (see below), times of divergence between sets of paralogues.

Figure 2.5 gives an indication of the number of groups that were discarded overall on the basis of inconsistent times of divergence between pairs. The over-estimation of the timing of divergence events tends to occur when the proteins are essential to the organism and their orthologues are therefore highly conserved, but they may be quite different to their paralogues, yielding a tree with a very long internal branch, and very short external branches. With the *Hox* genes, for example, the *HoxB5* genes in human

number of genes in set	initial number of sets	> 30 cM apart / unmapped	*non-homo-logous	*not sequenced	disparate, or too old, time of divergence	final number of sets used
4	62	36	10	7	5	4
3	199	142		17	42	9
2	344	314		21	20	10

Table 2.2. Summary of number of sets in dataset at each stage of processing the TIGR data. Those categories marked with an asterisk arose because some of the human-mouse "homologies" listed by MGD have been made on the basis of mutant phenotypes, not gene sequences. Some sets were discarded on the basis of failure of more than one criterion.

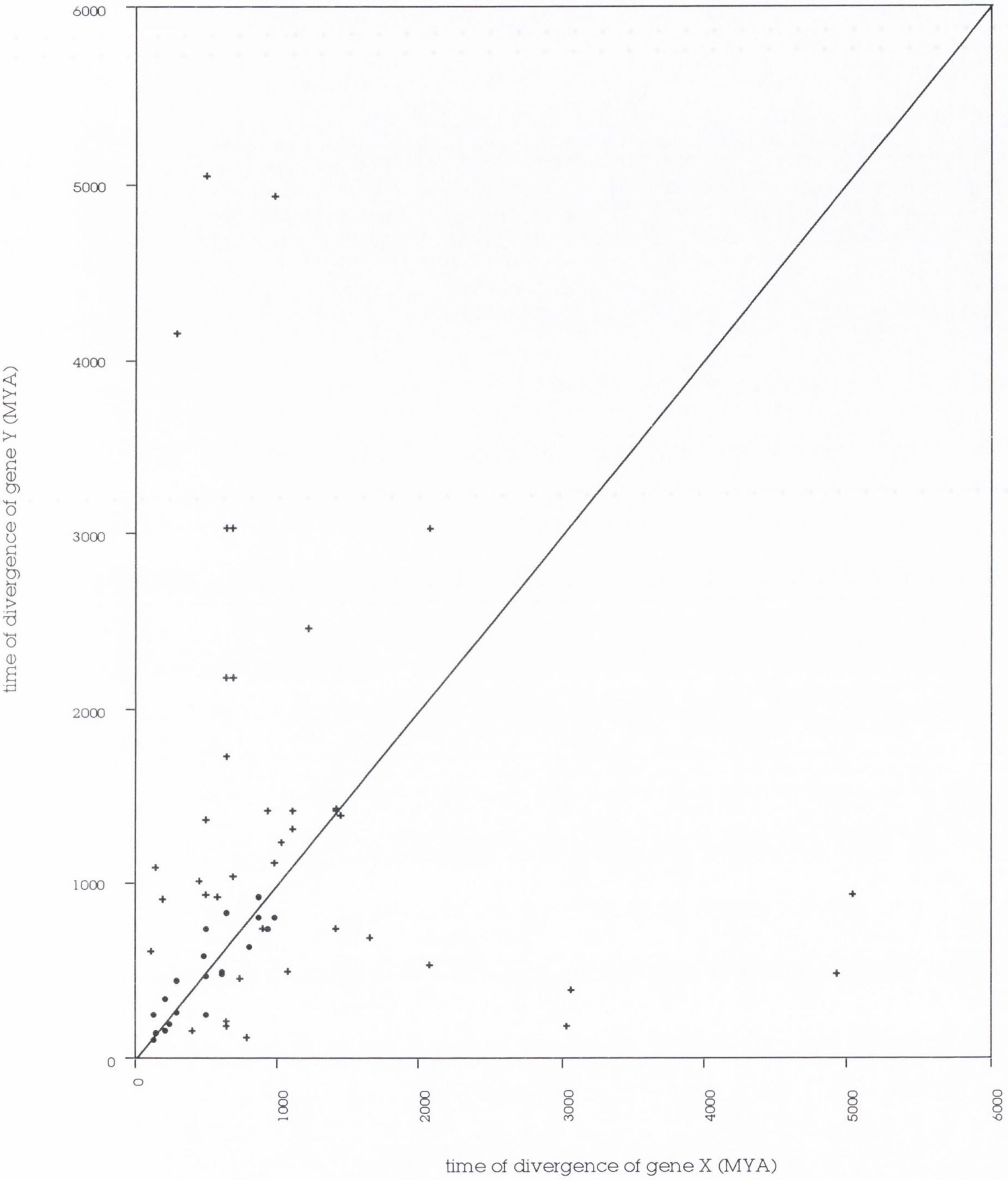


Figure 2.5. Scatterplot of the times of divergence between pairs of genesets defining a possible conserved duplicated region between human and mouse. Points coloured black are those that were retained in this study. All points marked with a cross were deemed to show times of divergence too different from each other, or too large to be taken as being correct.

and mouse are very similar, but their homology to *HoxC5*, although still evident in the homeobox domain, is weak in the flanking sequence. Ruddle *et al.* (1994b) calculated a divergence time of 350 Mya for the *Hox* clusters, by assuming that silent sequence distances are linear to divergence time, which is much more reasonable than our estimate of over 3,000 Mya, using the substitution rate over the whole protein, not just the highly conserved homeobox domain. Under the less stringent conditions imposed for the situations outlined in Figure 2.4 (columns 3 and 4 in Table 2.2), where non-adjacency of at most one pair of paralogues in either or both species was allowed, a further 19 groups of genes were retained as possibly representing duplicated regions conserved between human and mouse.

Figure 2.6 is a histogram detailing the 23 groups of genes that were kept. All human chromosomes except for HSA 8, 9, 13, 15, 16, 18, 20, 21 and 22 are represented. Four candidate paralogous regions (indicated on the left hand side of Figure 2.6) are discussed below.

1. HSA 4 has three sets of orthologues linking it to MMU 5: FGFR3, TEC, PDEB (all in the Huntington disease region). Of these, all have paralogues on HSA 5: (PDGFRB, CSF1R), (FER, ITK), PDEA, where names in brackets indicate two paralogues on HSA 5 for the respective single gene on HSA 4. Two of these paralogous groups have orthologues on MMU 18 (Pdgfrb, Pdea). The other orthologue is to be found on MMU 11 (Itk). This suggests an ancestral conserved duplicated region on HSA 4 and HSA 5 (first proposed by Comings (1972)), which correspond to MMU 5 and MMU 18, respectively. See Figure 2.7. The fact that the segment is broken in mouse, into sections on both MMU 11 and MMU 18 is consistent with other observations that the rodent map has undergone more rearrangements than that of mammals (Lundin, 1993). If we assume 100 Mya to have elapsed since the divergence of the lineages leading to human and mouse, we get an

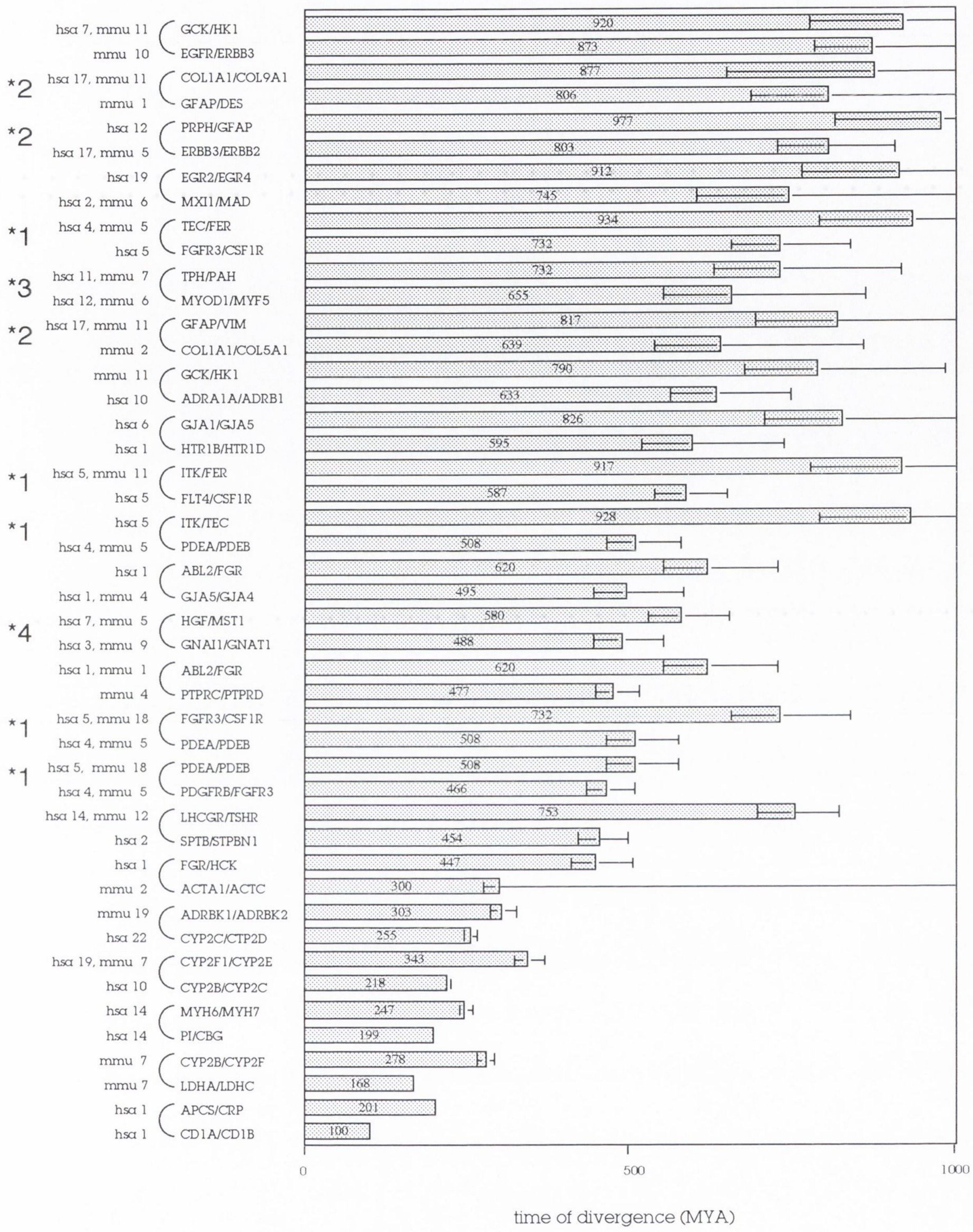


Figure 2.6. Histogram of paired times of divergence between genesets defining possible conserved duplicated regions between human and mouse. For example, in the first set, GCK is near to EGFR, and HK1 is near ERBB3. GCK is a paralogue of HK1, and EGFR is a paralogue of ERBB3.The estimated time of divergence is written in the centre of each bar. The chromosomes indicated are the chromosomes involved in the duplication. Sets marked with an *1, *2, *3 or an *4 indicate those sets defining the possible conserved duplicated segments between the human and mouse genomes described in the text.

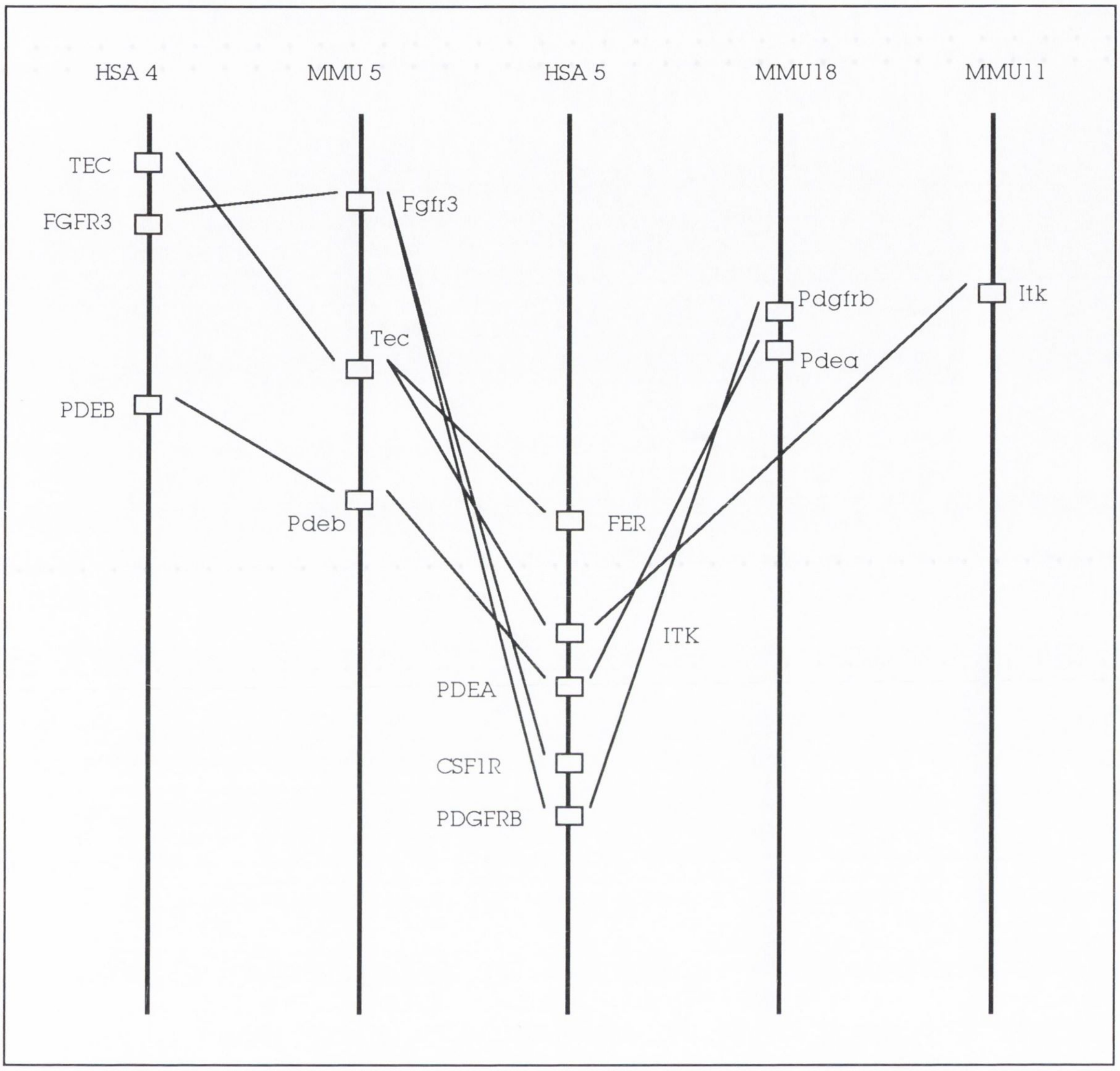


Figure 2.7. Possible conserved duplicated segment between HSA 4 and HSA 5, with MMU 5 and MMU 18. Lines indicate sequence relationships. Scale is approximate only, and lengths of chromosomes are not indicated. Mapping positions of human genes involved (as defined by chromosomal bands): TEC: 4p12, FGFR3: 4p16.3, PDEB: 4p16.3; FER: 5q21, ITK: 5q31-32, PDEA: 5q31.2-q34, CSF1R: 5q33.3-q34, PDGFRB: 5q33-q35. Mapping positions of mouse genes involved (in cM): Fgfr3: 5@20; Tec: 5@41; Pdeb: 5@57, Pdea: 18@31, Pdgfrb: 18@30, Itk: 11@27.

average divergence time of 730 Mya for these sets of paralogues, and therefore for a possible genome duplication event.

2. The genes found on HSA 17 all have homologues on MMU 11 (see Figure 2.8). Unlike the case with the segment on HSA 4, 5 / MMU 5, 18, there is no clear-cut corresponding duplicated region. Three sets of genes on HSA 17 (TCF2, GFAP, ERBB2) have paralogues on HSA 12 (TCF1, PRPH, ERBB3), of which TCF1 has a mouse orthologue on MMU 5, PRPH has a homologue on MMU 15, and ERBB3 has a homologue on MMU 10. GFAP has three mouse orthologues: Gfap (MMU 11), Vim (MMU 2) and Des (MMU 1). Des also has an orthologue on HSA 12. COL1A1, on HSA 17, has a mouse orthologue on MMU 2. There is a possible duplicated segment on HSA 17 and HSA 12, with a segment on MMU 11 corresponding to HSA 17. The rodent map is too broken to be able to define an orthologous region for HSA 12. The average time of divergence for this set of paralogues is approximately 850 Mya. A segment has already been described between MMU 11 and HSA 17. It is approximately 47 cM long, and consists of 62 loci (Eppig and Nadeau, 1995; Eppig, 1996). On comparison of this map with our segment, we added all genes which also had homologues on HSA 12 and MMU 15 such as the *Hox* clusters, the keratin clusters, the retinoic acid receptors and CD4 and CD7 (see white boxes in Figure 2.8). These genes were overlooked in our analysis either because they were not in the TIGR/MGD dataset or because they did not achieve the thresholds for BLAST or permitted intergenic distances.

There are two other gene sets with the full complement of eight genes.

3. On HSA 11, there is a segment comprising TPH and MYOD1, with a homologous segment on HSA 12 (PAH, MYF5) with orthologous segments on MMU 6 and MMU 7, respectively, and an average divergence time between paralogues of 700 Mya.

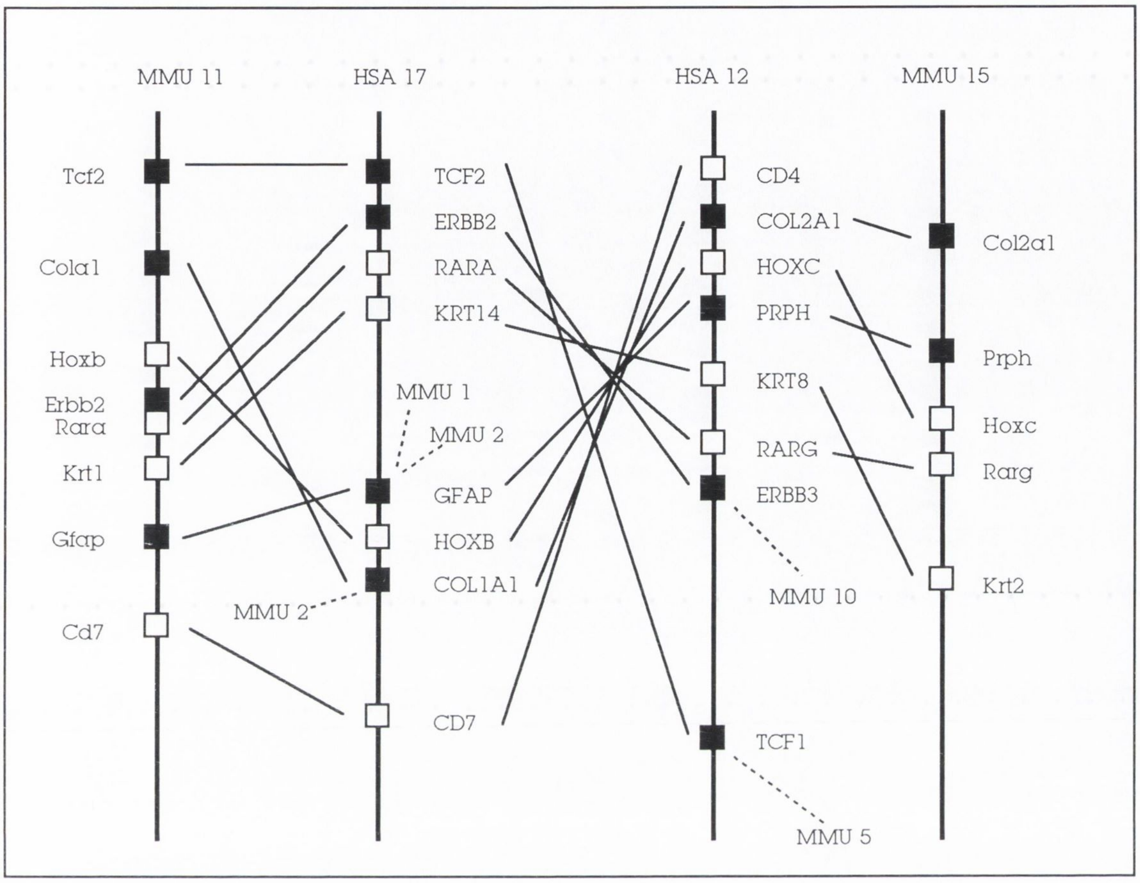


Figure 2.8. Possible conserved duplicated segment between HSA 17 and HSA 12, with paralogous segments on MMU 11 and MMU 15. Black boxes indicate positions of those genes involved in this study. White boxes are those genes which have been added by reference to Eppig (1995, 1996). Solid lines indicate homology relationships. Dashed lines show the locations of homologous genes which have been transposed in the mouse genome. Mapping positions of human genes involved (as defined by chromosomal bands) in order of position: TCF2: 17q11.2-q12, ERBB2: 17q11.2-q12, RARA: 17q12, KRT14: 17q12-q21, GFAP: 17q21, HOXB: 17q21-q22, COL1A1: 17q21.3-q22, CD7: 17q25.2-25.3; CD4: 12pter-p12, COL2A1: 12q12-q13.2, HOXC: 12q12-q13, PRPH: 12q13, KRT8: 12p13.2-q24.1, ERBB3: 12q13, TCF1: 12q24.3. Mapping positions of mouse genes involved (in cM) in order of position: Tcf2: 11@44, Cola1: 11@50, Hoxb: 11@56, Erbb2: 11@57, Rara: 11@57, Krt1: 11@58, Gfap: 11@62, Cd7: 11@74; Col2a1: 15@56.8, Prph: 15 (syntenic), Hoxc: 15@57.1, Rarg: 15@57.1, Krt2: 15@58.7

4. On HSA 3, HGF and GNAI1 form a segment homologous to MST1 and GNAT1 on HSA 7, and orthologous to MMU 5 and MMU 9, with an average time of divergence of 530 Mya.

It is of interest to note that the last three groups in Figure 2.6, showing the lowest divergence times, only contain paralogous regions within the same chromosome. Although this could be due to a tetraploidization event, a more parsimonious explanation is to suppose that they are the result of regional duplications.

2.6 DISCUSSION

Ancestral paralogous segments should also be seen as orthologous segments between human and mouse (Nadeau, 1991). We have found at least four candidate paralogous segments, of which two are only marked by two pairs of paralogous genes. Other sets found define incomplete segments without the full complement of eight genes. This could imply that these segments are just coincidental artefacts. Apparent conserved syntenies marked by only a single gene may also be due to a possible error in the mapping process, as is illustrated by the following example from Sankoff *et al.* (1997a): “In April 1996, the MGD contained 28 genes which each constituted the sole evidence of a homologous segment in some human chromosome and some mouse chromosome, out of ~110 conserved syntenies in all. By August 1996, 5 of these genes had been removed from either the human or mouse data, 4 had been reassigned in one or both genomes, and only 2 segments were confirmed by the mapping of additional genes on both the human and mouse chromosomes. An additional 6 single-gene segments also appeared in the database at this date.”

A second problem with finding genuine conserved segments is that of identifying paralogues. While any pair of sequences with a significant BLAST score are homologues, it can often be difficult to distinguish orthologues from paralogues, and to distinguish between multiple paralogues. Genome duplication, coupled with tandem duplication and successive rearrangements, can complicate the identification of paralogy among families of unlinked genes, so that members of large superfamilies have a more limited predicative value in this context than closely related members of small gene and protein families (Lundin, 1993). For example, in the segment between HSA 4 and HSA 5 described above, we have identified PDGFRB (and CSF1R) as homologues of FGFR3. In fact, a more likely homologue is FGFR4 (Lundin, 1993) which was not included in our study, not having been mapped in mouse, but it is in the "correct" position at 5q33-5qter. Furthermore, PDGFRB and CSF1R, which lie close together on HSA 5 not only have orthologues close together on MMU 18, but also form a paralogous segment with PDGFRA and KIT on HSA 4 (4q11-q12; 4q12, on the other arm from the segment we have described) which also shares an orthologous segment on MMU 5 (Pdgfra: 5@42; Kit: 5@42) which is in the middle of our segment.

Thirdly, differential silencing of duplicated genes and subsequent rearrangements may also confound the evidence. On duplication of a genome, each chromosome now forms a paralogous segment with its duplicate. As the genome progresses towards a re-establishment of disomy, some of the gene duplicates will be lost, either by deletion or mutation. Which of the two duplicates is lost would appear to be, for the most part, a random process, unless there is some functional reason for any pair of genes to remain linked. If no rearrangements occur, then each chromosome still forms a paralogous segment with its duplicate, but now there would appear to be genes on one segment that do not have a paralogue on the other. If approximately 50% of gene duplicates are silenced and rearrangements take place, the number of paralogous segments observed increases, the number of genes forming those segments decreases, and it becomes less likely that we will see the evidence of such a genome duplication event. Also, small inversions will change the gene order of a paralogous segment, while conserving synteny. In both of our segments that included more than two pairs of paralogues, neither of them conserved gene order across the whole segment.

Another problem which is highlighted here is the determination of duplication and divergence dates. Genes from paralogous regions have all been duplicated together, and their duplication dates should indicate this fact by being consistent with one another. In our analysis, for example for the four candidate paralogous regions discussed above, the divergence dates (i.e., dates for the genome duplication event) range from 530-800 Mya, all of which are older than what we would expect (~450-500 Mya). However, these dates are dependent on the 100 Mya which we took as the time of divergence between the human and mouse lineages. If a date of 80 Mya is more accurate, then the dates of divergence which we have calculated would also become proportionally younger, and would therefore become consistent with a duplication event around the origin of vertebrates. It would probably be more feasible to use orthologues from a number of other species to determine the timing of gene duplication events.

The dates shown in Figure 2.6 are also inconsistent with one another, spanning a range of divergence dates from 100 Mya to almost 1,000 Mya. There are several possible reasons why this might occur. Firstly, little is known about how disomic inheritance becomes re-established. In particular, it is possible that individual chromosomes could become disomic at different times, resulting in multiple different divergence times for gene pairs (Gaut and Doebley, 1997). Secondly, gene conversion events between genes on opposing branches (e.g., branches a and d in Figure 2.1) may give dates which appear younger than the duplication event. Gene conversion events can also give trees which instead of having an (AB)(CD) topology have an (A)(BCD) topology, further confusing the timing of gene duplication events. Conversely, undetected paralogy may produce date disparities between duplicated genes. If genes have been tandemly duplicated before the genome duplication event, and then are differentially silenced, the divergence dates for the two remaining genes will yield more ancient dates than that of the genome duplication event (see Section 6.1, p. 99). Lack of sequence data will also yield a similar result to this. Spring (1997) has suggested an amphioxus-like animal underwent allotetraploidy, creating primitive vertebrates around 530 Mya. In this case, the divergence times between the two sequences at a duplicated locus could correspond either to the speciation time between the two progenitors, or the time of establishment of disomy, the outcome being random for any particular locus.

CHAPTER 3 —

ASSESSMENT OF DISTRIBUTION OF DUPLICATED GENES IN A GENOME

3.1 AIM AND INTRODUCTION

The human genome is thought to contain between 50,000 to 100,000 genes (estimates average about 80,000; Fields *et al.*, 1994), of which about 50% had been sampled in the form of ESTs by 1996 (Schuler *et al.*, 1996). The Schuler database (1996) contains 2,317 genes which have been both sequenced and placed on physical maps of chromosomes. Because mammalian map data is very sparse, it becomes difficult to determine the extent of duplication within the human genome using standard methods. To estimate the extent of evidence for one or more tetraploidization events in the human genome, we developed an algorithm to calculate the average pairwise physical or genetic (not sequence) distance between all possible pairs of paralogues within a genome. For example, if there are two unrelated pairs of paralogues, A/*a* and B/*b*, we examine the distance from A to B and from *a* to *b*. If we calculate the average pairwise distances, we have an estimate of the length of a possible paralogous segment. Evidence of genome duplication should be evident by an excess of short candidate DCRs. For example, for a strand of DNA with genes AB duplicated to give *ab* (as in Figure 3.1), one of the calculated distances will be 1/2((AB)+(*ab*)), where AB is the

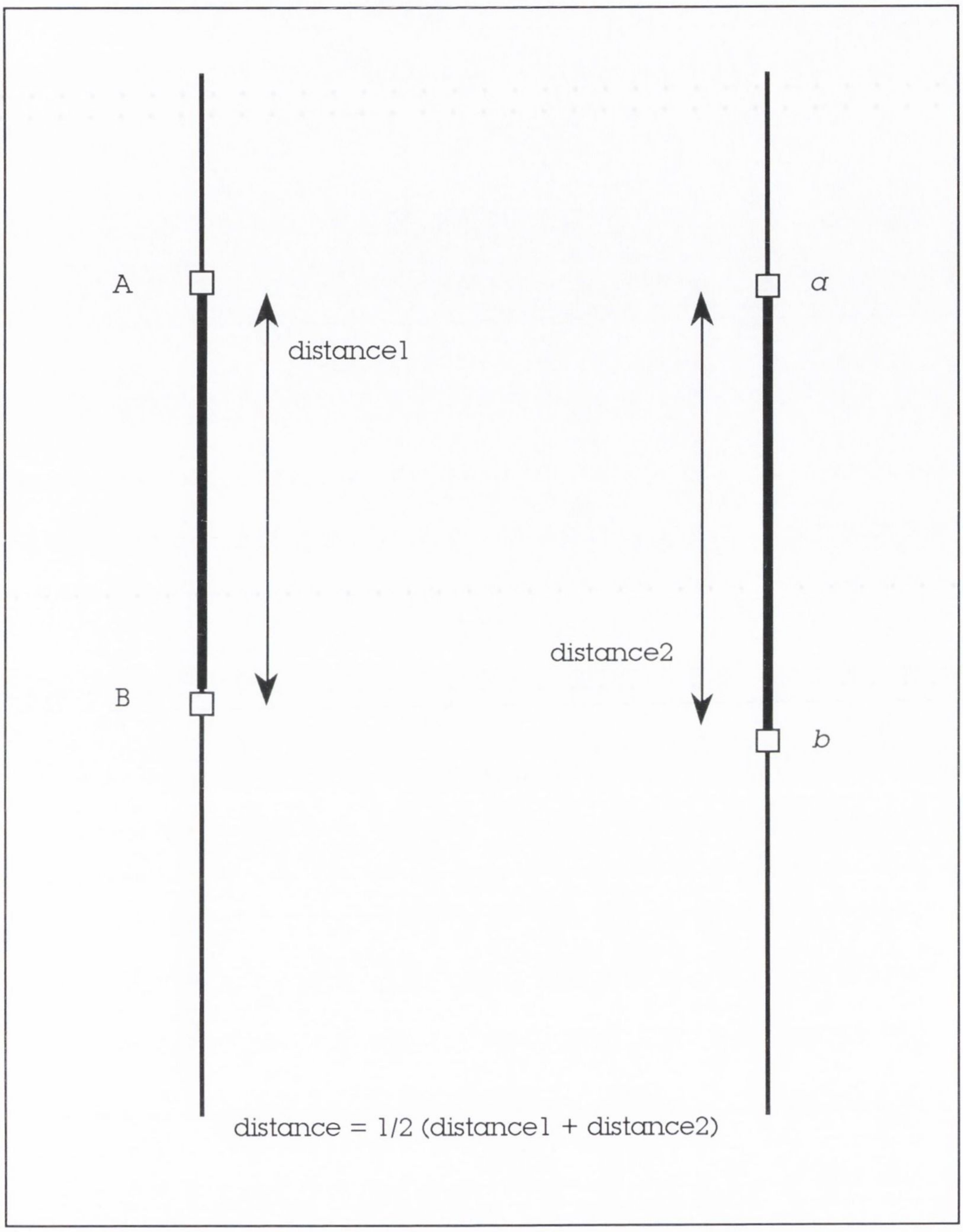


Figure 3.1. Representation of the non-overlapping duplicated region required to calculate distances between genes delimiting a possible paralogous region. A is a paralogue of B, and *a* is a paralogue of *b*. A- and B-type genes may be related. The distance calculated is the average distance between the two adjacent unrelated genes.

shortest distance between A and B, and *ab* is the shortest distance between *a* and *b*. On each segment, A and B are adjacent as are *a* and *b* so the calculated distance will be short. If A and B have been duplicated by other means (or even if they have been duplicated as part of the strand and then been rearranged), then the distance between them will be much greater.

To be able to understand and evaluate our results, we used the genome of yeast, an ancient tetraploid (Wolfe and Shields, 1997) as a positive control. Because it is unlikely that bacteria have undergone a similar genome duplication event, we used the genomes of the five bacterial genomes which had been completely sequenced in 1997 as negative controls.

3.2 DATA

3.2.1 HUMAN SCHULER DATA

The publication by Schuler *et al.* (1996) of a human gene map, integrating much of the previously uncollated sequence and map data, presented a wealth of data. A gene may be represented in the current databases by multiple short cDNA fragments also known as expressed sequence tags (ESTs), which correspond to different parts of a transcript, or alternatively spliced transcripts. In the Schuler *et al.* (1996) database, each unique gene was represented by a single representative sequence, by focusing on the 3' untranslated region (UTR) of the mRNAs, whose sequences can be converted to gene-specific sequence tagged site markers (STSs) for mapping. These STSs were put into the Unigene database. An international consortium, IMAGE, yielding mapping data for 20,104 STSs in Unigene, corresponding to 16,354 distinct loci (possibly as much as 20% of the protein coding regions in the human genome).

For each STS mapped by Schuler *et al.*, their database lists corresponding entries in at least one of the Unigene, EMBL or SWISSPROT databases. The World Wide Web site (http://www.ncbi.nlm.nih.gov/SCIENCE96) that provides the data supporting the Schuler *et al.* paper includes lists of cross-references to these other databases for each STS. Our aim was to place as many complete protein sequences from the Schuler *et al.* map for our work as possible. We assigned SWISSPROT and EMBL entries (in that order of precedence) where they existed, to the map. The data at this stage of processing comprised 3,745 mapped SWISSPROT entries, and 766 mapped EMBL entries. The other 12,000 or so loci on the Schuler *et al.* map are ESTs for which the full-length cDNA or protein sequence is not known.

On examination of the data, we found multiple instances where the same SWISSPROT entry had been assigned by Schuler *et al.* to more than one locus, often on different chromosomes. The map was estimated by Schuler *et al.* to have an error rate of 1% of loci placed on different chromosomes by two different laboratories (Schuler *et al.*, 1996). We examined 101 SWISSPROT proteins that had been assigned to more than one chromosome by comparing the EST sequences from the conflicting map positions to SWISSPROT, using BLASTP. In 37/101 cases, the ESTs of both map positions were found to hit the relevant protein in SWISSPROT, which suggests that two loci exist coding for two similar proteins, only one of which appears in SWISSPROT. In another 37/101 instances, the ESTs from only one of the given locations gave the correct result, indicating mistakes in the database annotations of Schuler *et al.* Of the 101 proteins tested, 27 were mapped differently by different labs, and one or both of the two reported chromosomal assignment was flagged as uncertain by Schuler *et al.* This is equivalent to a 27% error rate. All singly mapped genes tested were found to be correct.

To counteract the problem of accuracy in the mapping data, we excluded all SWISSPROT proteins with multiple localizations. This increased the accuracy of the map, at the cost of the loss of much data. Our final dataset contained 1,610 SWISSPROT proteins, and

697 mRNAs from EMBL. Of 69 mRNAs which had no CDS in EMBL, another 10 were added by performing BLASTP searches with them against the EMBL database, bringing the number of mRNA-based proteins to 707, giving a total of 2,317 mapped proteins used from Schuler *et al.* (1996).

3.2.2 *SACCHAROMYCES CEREVISIAE*

The yeast genome has 5,908 identified genes, of which the 55 duplicated blocks, containing 2,949 genes, span 50% of the genome. The protein sequences and location information were obtained from Ken Wolfe at: http://acer.gen.tcd.ie/~khwolfe/yeast.

3.2.3 BACTERIA

3.2.3.1 *HAEMOPHILUS INFLUENZAE* RD

The *Haemophilus influenzae* genome was the first bacterial genome to be sequenced by whole-genome random sequencing and assembly (Fleischmann *et al.*, 1995), with an estimated error rate of one base every 5,000-10,000 bases. The *Haemophilus* genome is 1,830,137 bp long, with 1,743 identified ORFs, of which the functions of 58% are reasonably well defined. Locations and protein sequences were obtained from the TIGR ftp site: ftp://ftp.tigr.org/pub/data/h_influenzae.

3.2.3.2 *METHANOCOCCUS JANNASCHII*

The genome of the autotrophic archaeon contains 1,682 identified ORFs in the large circular chromosome of 1,664,976 bp, 44 in large extrachromosomal element (ECE) of

58,407, and 12 in 16,550 bp small circular ECE (Bult *et al.*, 1996). Of these 1,738 ORFs, only 38% have been assigned a putative cellular role. Locations and protein sequences of the large circular chromosome only were obtained from the TIGR ftp site: ftp://ftp.tigr.org/pub/data/m_jannaschii.

3.2.3.3 *MYCOPLASMA GENITALIUM*

At 580,070 bp, this is the smallest known genome of any free-living organism (Fraser *et al.*, 1995). Of 468 predicted ORFs, 374 were assigned putative biological roles. Data were obtained from: ftp://ftp.tigr.org/pub/data/m_genitalium

3.2.3.4 *SYNECHOCYSTIS SP.* STRAIN PCC6803

The 3,573,470 bp cyanobacterium genome sequence contains 3,168 potential protein genes, 45% of which had no similarity to any known genes (Kaneko *et al.*, 1996). 145 (4.6%) were identical to reported genes and 1,257 (39.6%) were similar to reported genes, while 340 (10.8%) are similar to hypothetical proteins. Sequence data and mapping locations are available from: ftp://ftp.kazusa.or.jp/pub/cyano.

3.2.3.5 *ESCHERICHIA COLI* K12

E. coli, 4,638,858 bp, was sequenced at the E. coli Genome Centre at the University of Wisconsin (Burland *et al.*, 1993). The *E. coli* genome contains 4,285 ORFs, of which over 70% have a known function. Peptide sequences and location data were obtained from: ftp://ftp.genetics.wisc.edu/pub/sequence/ecoli.seq.

3.3 METHODS

Both the human and yeast genomes were systematically searched for paralogous regions. Instead of finding single pairs of genes that had been duplicated and trying to fit them into larger regions, the whole genome was utilized (using all the mapped genes with protein sequences) and distances calculated between every possible pair of paralogues. Genes in the human dataset were taken as being paralogous if their BLASTP scores were greater than 150, or for shorter proteins, one third of the lower BLASTP score of the two self-hits. For yeast and all bacterial data, the SEG filter was not used, and a BLASTP score of 200 was taken as the cut-off point. This high threshold was imposed to exclude alignments of questionable biological significance.

3.3.1 MULTIPLE LINEAR CHROMOSOMES

Every possible pair of blast hits were evaluated and if the pair of hits delimited two non-overlapping regions (as in Figure 3.1), the average distance between them was calculated. To determine what proportion of these duplicated segments was due to chance, the gene locations were shuffled, and the average pairwise distances between paralogues were again calculated as above. The results from 20 sets of shuffled data were compared to the real data. By subtracting the number of DCRs found in the shuffled data from the number found by examining the real data, we can get an estimate of the number of DCRs not occurring by chance. Because the shuffling process for multiple linear chromosomes can change the chromosome on which the genes are located, the number of non-overlapping paralogous regions that exist varies not only between the real and shuffled datasets, but also between shuffled datasets.

3.3.2 CIRCULAR CHROMOSOMES

Changes were made to the algorithm for circular bacterial chromosome calculations. For any pair of genes, there are two possible distances between them, going either in a clockwise or anticlockwise direction around the circular genome, of which the shorter of the two was chosen. If the distance between any pair of genes is greater than half the length of the genome, then the shortest distance between them is the genome length minus that distance. The average distance between any two pairs of paralogues cannot, therefore, be greater than half the genome length. Potentially paralogous regions which overlapped were not included. Gene location data were also shuffled and, as before, the results of 20 sets of shuffled data calculations was compared to the real data. The number of distances calculated in the real and shuffled sets of data does not differ when dealing with a circular chromosome.

3.4 RESULTS

3.4.1 ANALYSIS OF SIZE DISTRIBUTION OF POSSIBLE DCRS

We put the numbers of DCRs found for both the real and shuffled data into bins of 0.1% of the length of an average chromosome for the human and yeast genomes (0.05% of the genome length for bacterial genomes) (Figures 3.2 and 3.3) and into bins of 1% of the length of an average chromosome (Figures 3.4 and 3.5). Figures 3.2 and 3.4 are basically histograms, detailing how many DCRs were found for each bin. Because it is difficult to see the differences between the numbers of DCRs found for the real and shuffled datasets from these, we also made histograms where each bin now contained numbers of all DCRs found up to that bin (i.e., that fraction of the length of an average chromosome (or genome, for bacteria)), not just the number found at that length (Figures 3.3 and 3.5). These help us to distinguish more clearly between the real

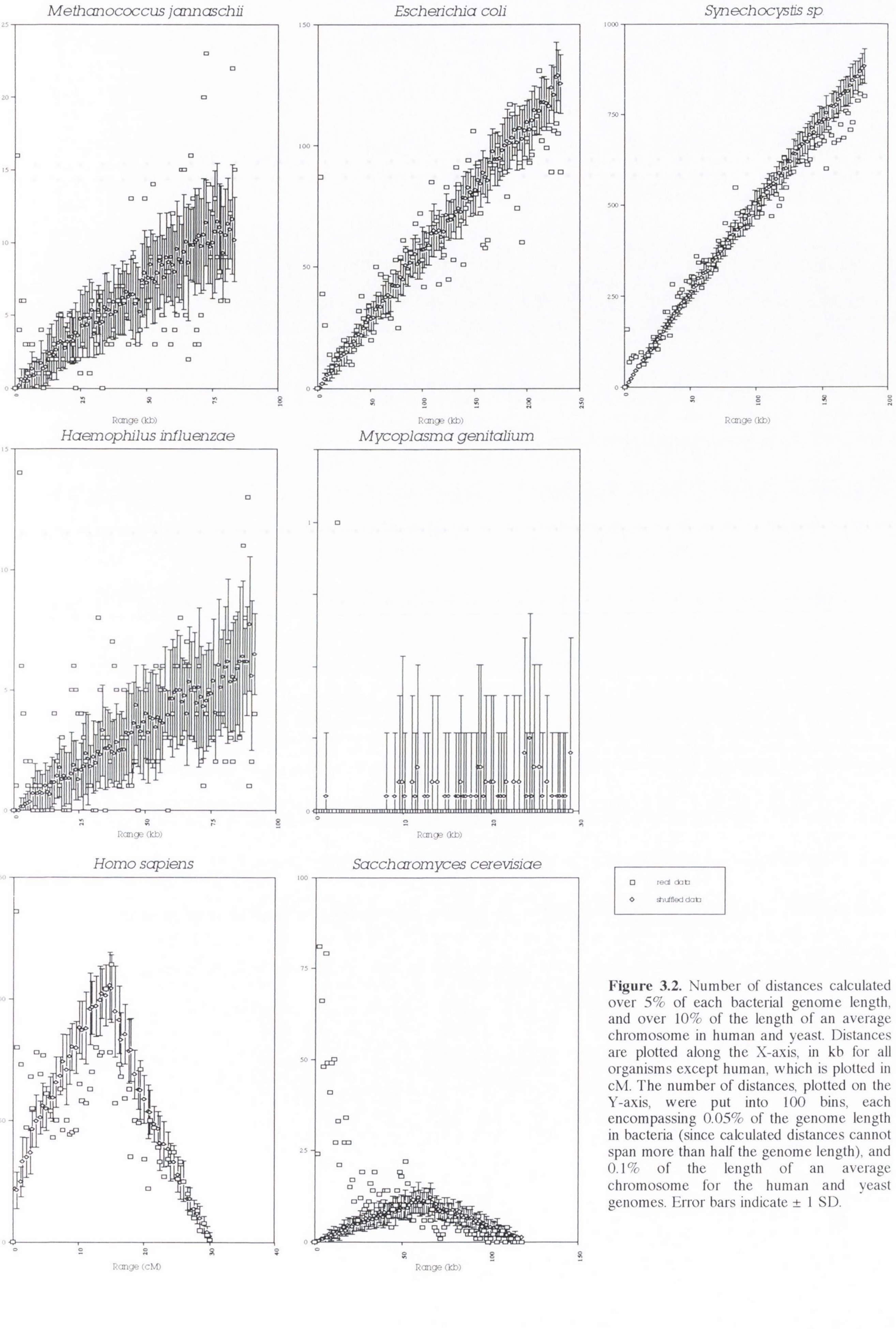


Figure 3.2. Number of distances calculated over 5% of each bacterial genome length, and over 10% of the length of an average chromosome in human and yeast. Distances are plotted along the X-axis, in kb for all organisms except human, which is plotted in cM. The number of distances, plotted on the Y-axis, were put into 100 bins, each encompassing 0.05% of the genome length in bacteria (since calculated distances cannot span more than half the genome length), and 0.1% of the length of an average chromosome for the human and yeast genomes. Error bars indicate ± 1 SD.

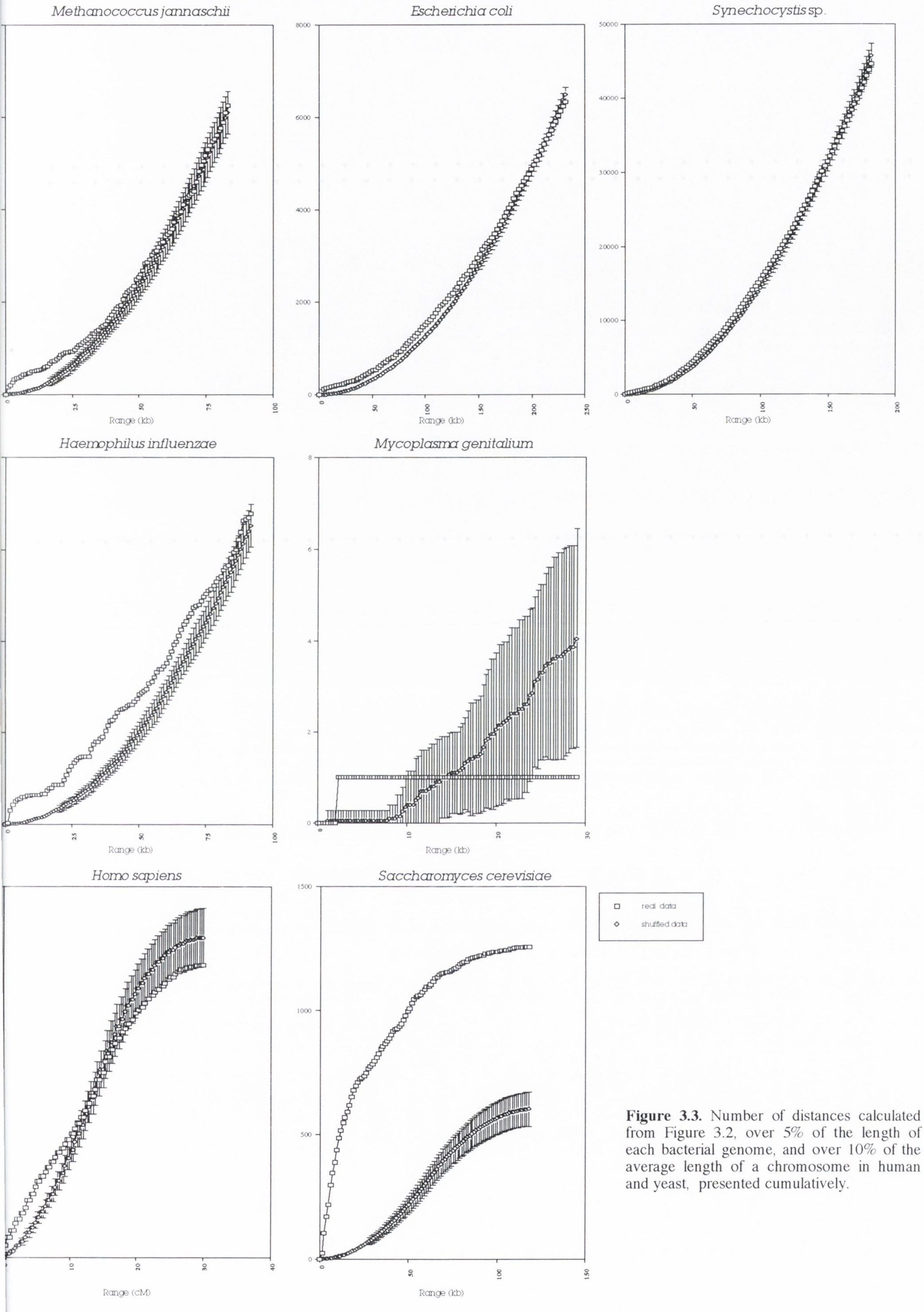


Figure 3.3. Number of distances calculated from Figure 3.2, over 5% of the length of each bacterial genome, and over 10% of the average length of a chromosome in human and yeast, presented cumulatively.

and shuffled datasets. The qualitative difference in the shape of the simulated plots in Figure 3.2 (Normal curves for yeast and human, linear slopes for bacteria) is a consequence of circular versus linear chromosomes.

3.4.1.1 BACTERIA

The five bacterial chromosomes can be divided into two categories: those that coincide with the simulated data, which would be expected for a genome that has undergone no genome-wide duplication event, and those that veer from it. Those that are more similar to the simulated data are *Methanococcus jannaschii*, *Escherichia coli*, and *Synechocystis* sp., all of which are also non-pathogenic. This implies a random organization of the genome, where any duplicated genes are not part of an ancestral duplicated segment, as would be expected. The other two bacterial genomes, *Haemophilus influenzae* and *Mycoplasma genitalium*, both pathogens, show a higher proportion of short DCRs, up to about 4% of the genome length, when compared to the shuffled data (see Figures 3.3 and 3.5), although the difference in *Mycoplasma* is minimal because there is only one DCR found at distances up to 0.05% of the genome length. It is known that bacterial genomes contain duplicated genes and multigene families, with estimates of between 38-50% for *E. coli* (Koonin *et al.*, 1995; Labedan and Riley, 1995; Koonin *et al.*, 1996), and lower estimates for the two pathogens: 30% of *Haemophilus* genes exist as duplicates (Brenner *et al.*, 1995; Koonin *et al.*, 1996); only 25% of genes are duplicated in *Mycoplasma* (Koonin *et al.*, 1996). The pathogens may include duplicates of some genes important to their lifestyle. For example, many of the short DCRs in *Haemophilus* are composed of transporter proteins and permeases. Perhaps the nature of a pathogen makes it necessary for its genomic structure and organization to be laid out in a specific non-random way.

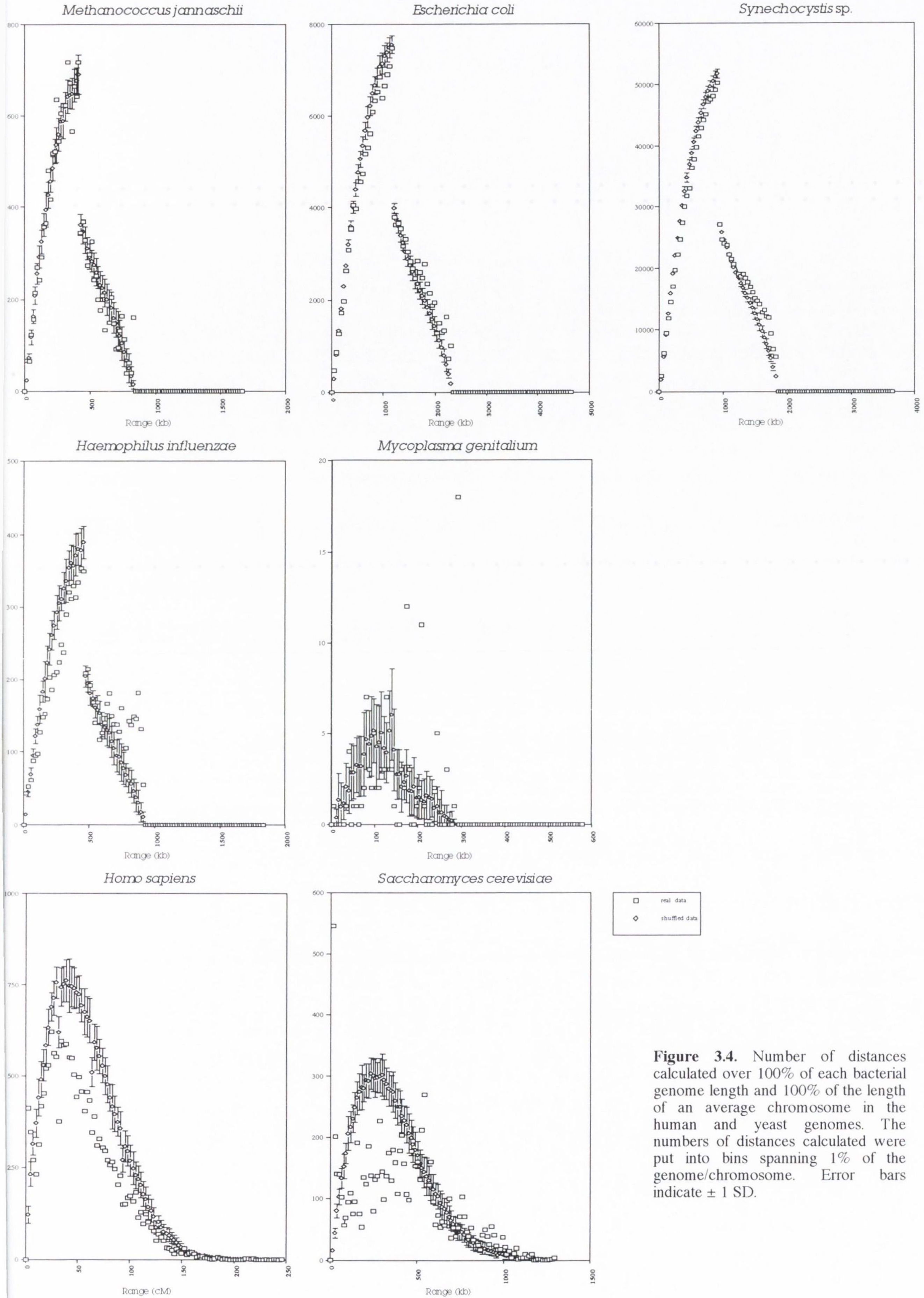


Figure 3.4. Number of distances calculated over 100% of each bacterial genome length and 100% of the length of an average chromosome in the human and yeast genomes. The numbers of distances calculated were put into bins spanning 1% of the genome/chromosome. Error bars indicate ± 1 SD.

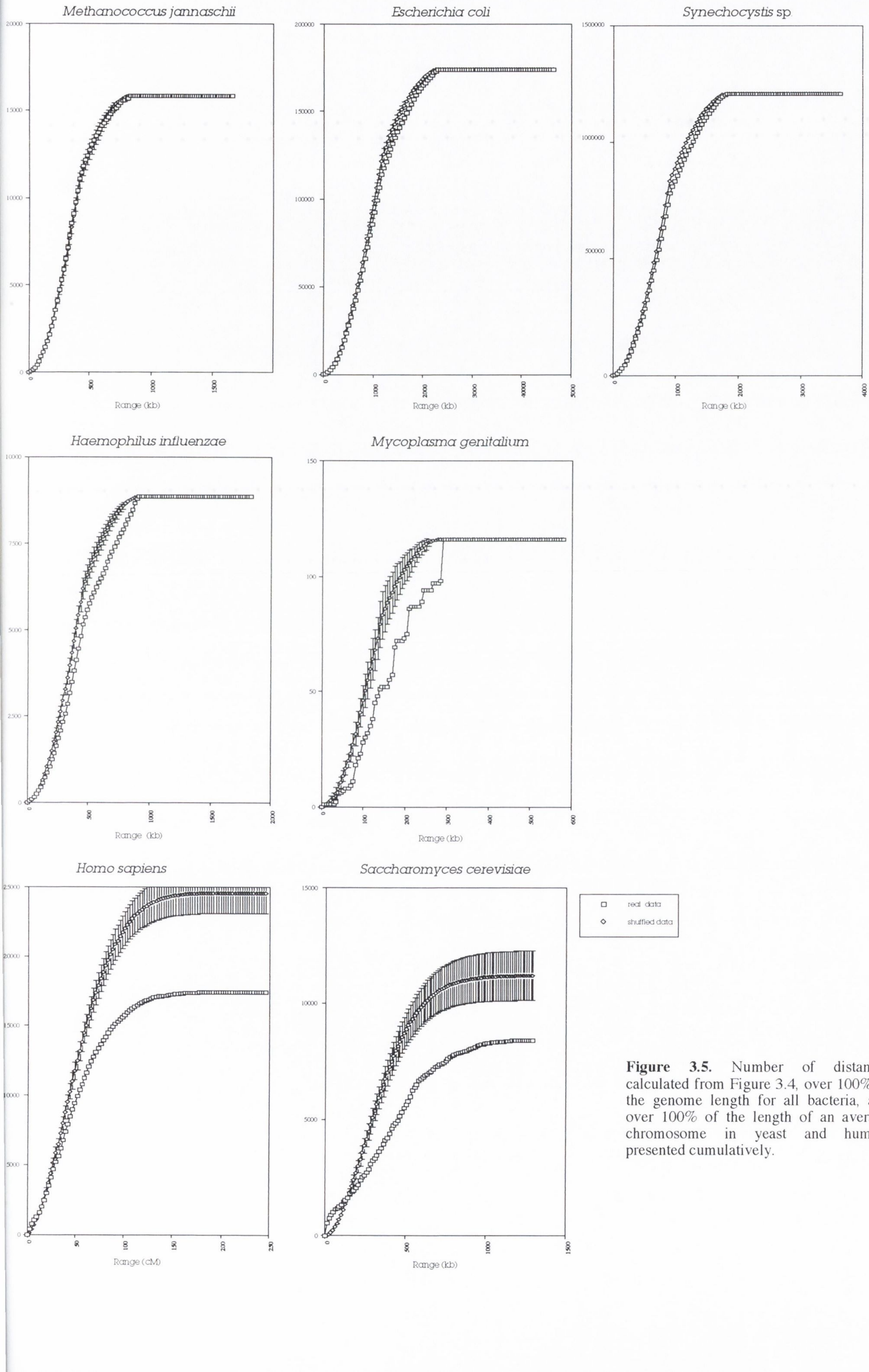


Figure 3.5. Number of distances calculated from Figure 3.4, over 100% of the genome length for all bacteria, and over 100% of the length of an average chromosome in yeast and human, presented cumulatively.

Synechocystis stands out as containing a very large number of DCRs (note the scale of the Y-axis for *Synechocystis* in Figure 3.2). This may be attributable to the presence of large gene families present in the genome, such as the transposase family. Table 3.1 indicates family sizes of as large as 38 highly similar genes. It also shows that *Synechocystis* is the only bacterial genome that has an average (per gene) of more than one BLAST hit with a score of greater than 200, indicating that the sizes of gene families in *Synechocystis* are large. This is reflected in the huge number of DCRs calculated.

3.4.1.2 *SACCHAROMYCES CEREVISIAE*

Saccharomyces cerevisiae is a degenerate tetraploid, and should exhibit an excess of short candidate DCRs, as evidence of a genome duplication. It can be seen from the number of distances calculated and put into bins at 0.1% of the average length of a yeast chromosome, that a higher number of short DCRs are found up to a 50 kb range, as compared to those distances which would be calculated if the genes present were randomly distributed (Figure 3.2). This is consistent with the lengths of duplicated segments found by Wolfe and Shields (1997), which average 56 kb.

Figure 3.3 shows that the number of DCRs found in the yeast genome up to distances of 10% of the length of an average yeast chromosome are significantly higher than the number that would be expected by chance (shuffled data). Furthermore, comparison with the other species examined here also indicates that the extent of duplicated segments at short distances in the yeast genome is vastly different from all other genomes.

When numbers of DCRs which cover the entire length of an average yeast chromosome are examined, it becomes clear that while the shuffled data appear to form a slightly skewed Normal curve, in the real data, the number of DCRs decreases after 50 kb.

	genome size (kb)	# genes	# BLAST hits > 200	mean # BLAST hits > 200 per gene	maximum # hits	standard deviation	standard error
Mycoplasma genitalium	580	468	16	0.034	2	0.19	0.009
Methanococcus jannaschii	1,665	1,682	188	0.112	12	0.59	0.015
Haemophilus influenzae	1,830	1,743	138	0.080	10	0.45	0.011
Escherichia coli K12	4,639	4,285	1,597	0.761	31	1.98	0.031
Synechocystis sp.	3,573	3,168	1,637	1.059	38	3.66	0.066

Table 3.1. Comparison of the extent of duplication within selected bacterial genomes. Genes were taken to be similar if they had a BLASTP score of greater than 200. It can be seen that in all genomes there is, on average, one gene per 1 kb. The larger the genome, the greater the number of similar genes, and the greater the family sizes.

Therefore, DCRs above this length are less likely to survive the rearrangements and gene loss that follow genome duplication.

3.4.2 HUMAN

The human genome also shows a higher number of DCRs formed up to about 10 cM that would be expected by chance (Figures 3.2 and 3.3). After this distance, the numbers of DCRs found for both the real and shuffled datasets become similar. This implies that paralogous regions which have been described in the literature as being composed of two or three genes and spanning large fractions of chromosomes are likely to be artefacts. It seems that it is unlikely that paralogous regions where genes are separated by more than 10 cM represent ancient paralogous regions (unless they have some adaptive function). However, the difference between the numbers of DCRs found at distances of up to 10% of the length of an average human chromosome is not as markedly different from the shuffled data as it is for yeast, despite the fact that it has been hypothesized that the human genome has undergone more than one round of genome duplication, along with multiple other chromosomal and tandem duplications. However, because the human genome contains at least 10 times as many genes as the yeast genome, there is a greater absolute number of DCRs found in human compared to yeast.

The observation that it is only DCRs of up to 10 cM that represent ancient paralogous regions is strengthened when we look at the numbers of distances calculated over the length of an average human chromosome (Figures 3.4 and 3.5). While many DCRs are still found, the difference between the numbers found in the real and shuffled datasets is slight (as it is for yeast). This does, however, indicate the significance of the difference between the numbers found at shorter distances.

Figure 3.5 shows that in human there is a difference between the randomly shuffled data and the real data, implying that the paralogues found in the human genome are not placed at random, but that rather, there is a regularity in the distribution of conserved segments within the genome. Far fewer map distances were calculated for the real data than there were for the shuffled data. This may underline a specific genome organization in mammals (O'Brien, 1991), where the conserved regions of duplication comprise only a small percentage of the chromosome, the rest having been destroyed by gene deletions, chromosomal rearrangements, inversions, or the differential silencing of genes to become pseudogenes (Lundin, 1993).

3.5 EXTENT OF DUPLICATION

We were concerned that some of the apparent short DCRs in human could have been caused by the presence of tandem repeat genes at multiple sites within the genome (i.e., if genes A and B in our model (Figure 3.1) were not distinct genes but members of a superfamily). There were 1,262 pairs of paralogues in the human genome which had average intergenic distances of under 10 cM, using the data of Schuler *et al.* (1996). To determine the actual difference between the real and the shuffled data, and to assess the extent to which tandem duplications contributed to the increased frequency of short DCRs as compared to the shuffled data, we were interested in the number of distances calculated within bins of 1 cM, the number of those distances accounted for by tandem repeats, and the difference between the real and the shuffled data. Table 3.2 shows such a comparison between human and yeast data. If there have been two genome-wide duplications events during the evolution of vertebrates, it would be expected that the human genome would show a higher percentage of short DCRs than would be expected by chance, and would also show a higher number of short DCRs than yeast, which is a degenerate tetraploid (Wolfe and Shields, 1997). Very few of the short putative DCRs in yeast are attributable to tandem repeats. Unfortunately, this is

Length of DCR	total # distances calculated		# tandem		chromosomes involved in tandem duplications[a]		proportion of DCRs that involve tandem repeats		# distances calculated for shuffled data		difference = (real - tandem) - shuffled[b]	
0 - 1 (0 - 5)	216	(233)	174	(3)	11(35) 12(2) 14(18) 17(4) 19(10) 2(4) 3(3) 6(94) 8(4)	24/3 8/2 7/1 10/2 13/3 9/2 6/1 19/2 9/2	0.81	(0.01)	30	(2)	12	(228)
1 - 2 (5 - 10)	150	(222)	89	(1)	11(1) 12(2) 17(2) 19(17) 3(2) 4(1) 6(64)	4/1 5/1 6/1 10/1 5/1 4/1 21/1	0.59	(0.005)	57	(6)	4	(215)
2 - 3 (10 - 15)	115	(131)	44	(1)	1(4) 10(3) 11(19) 17(6) 19(12)	9/2 6/1 15/2 6/1 8/1	0.38	(0.008)	80	(10)	-9	(120)
3 - 4 (15 - 20)	133	(109)	40	(0)	17(1) 19(18) 2(10) 8(18) 4(1) 3(1)	4/1 13/1 4/1 12/1 4/1 4/1	0.30	(0)	100	(15)	-7	(94)
4 - 5 (20 - 25)	146	(44)	24	(0)	1(3) 10(12) 19(4) 3(4) 8(1)	9/2 8/1 5/1 7/1 4/1	0.16	(0)	106	(23)	16	(21)
5 - 6 (25 - 30)	115	(51)	13	(0)	11(8) 19(2) 2(3)	9/1 5/1 5/1	0.11	(0)	127	(23)	-25	(28)
6 - 7 (30 - 35)	93	(61)	5	(0)	1(2) 2(3)	5/1 5/1	0.05	(0)	135	(26)	-47	(35)
7 - 8 (35 - 40)	109	(51)	2	(1)	4(2)	5/1	0.02	(0.02)	126	(29)	-19	(21)
8 - 9 (40 - 45)	94	(37)	14	(0)	1(5) 19(2) 2(1) 21(6)	15/3 6/1 4/1 6/1	0.15	(0)	161	(42)	-81	(-5)
9 - 10 (45 - 50)	91	(60)	1	(0)	19(1)	4/1	0.01	(0)	170	(36)	-80	(24)

Table 3.2. Number of distances calculated attributable to tandem repeats over 3% of the length of an average chromosome for yeast and human. All numbers in brackets indicate yeast data. The size of yeast bins versus human bins are not comparative by distance but rather by percentage average chromosome length. Lengths of DCRs are given in cM for human data and kb for yeast data.

[a] The first half of the column refers to the chromosomes involved, and the number of tandem repeats calculated for each chromosome. The second half of the column refers to the number of genes involved, and the number of paralogous regions they define.

[b] This column is presented in graph format in Figure 3.6.

not the case in the human genome where at shorter distances, up to 81% of DCRs are due to the presence of tandem repeats (first half of column 7 in Table 3.2). The huge number of tandem repeats at 0 - 2 cM can be largely explained by the haemoglobin gene cluster on HSA 6p, which forms DCRs with genes 1A02, 1A03, HLAE (histocompatibility antigens) and MOG (myelin/oligodendrocyte glycoprotein), all of which are found on HSA 6.

It is clear that the human genome has undergone more internal tandem duplications than yeast has, and this fact may hide the evidence for genome-wide duplications. If a gene is duplicated, it may be from a single random tandem duplication (or retrotransposition, etc.) or it may be from a tetraploidy event. The results from our method of calculating physical distances between pairs of duplicates as a means of estimating the extent of duplication in a genome may be harder to understand in the case of human because there will be more tandem repeats which will therefore increase the number of short DCRs we see.

Since it is an excess of short DCRs that point to a tetraploidization event, to be able to more truly assess the evidence for polyploidization in human, therefore, it was necessary to determine the number of DCRs found at short distances that were attributable neither to chance nor to tandem repeats (column 7 in Table 3.2; Figure 3.6). In yeast, it can be seen that very few of the short DCRs are attributable either to chance or to tandem repeats, as would be expected, since there are very few tandem repeats in the yeast genome. 78% of the DCRs defined by pairs of genes spaced at distances of up to 3% (50 kb) of an average chromosome are likely to be representative of ancient paralogous regions. In human between 60-80% of the DCRs found are attributable to tandem repeats at the shorter distances. At least 14-38% of the DCRs found which span 2 cM may be occurring by chance.

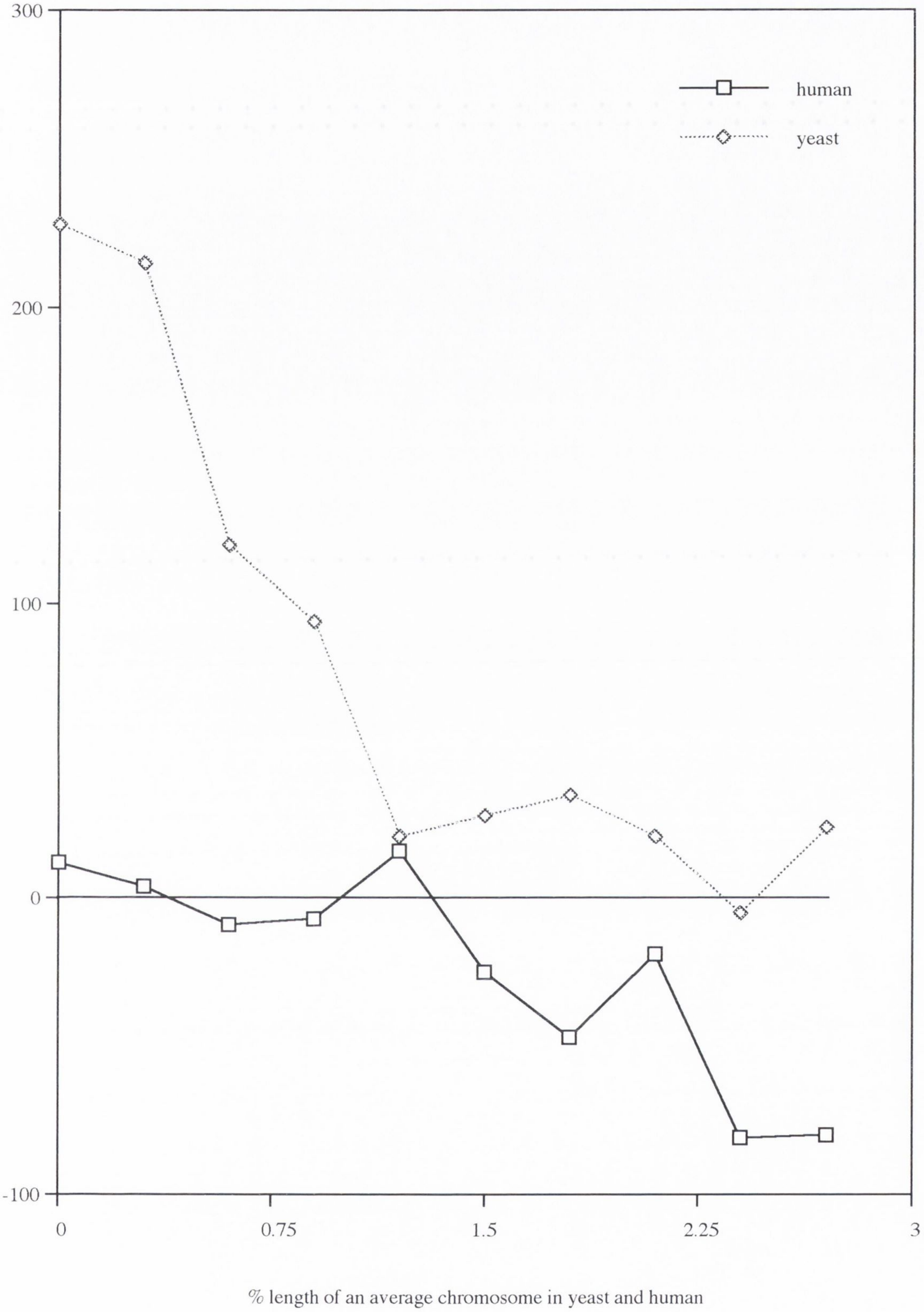


Figure 3.6. Number of DCRs composed of two pairs of paralogues attributable neither to tandem repeats, nor to chance, at 3% of the length of an average chromosome in yeast and human (data from Table 3.2).

CHAPTER 4 —

PARALOGOUS REGIONS IN THE HUMAN GENOME

4.1 AIM

Because we found very few duplicated conserved regions between mouse and man, we instead attempted to find duplicated paralogous regions within the human genome alone, because there was a greater amount of sequence data now available. By looking at human mapping data, we investigated whether we could find any evidence for one or more periods of polyploidy during the evolution of mammals. We used two different sets of sequence data — one protein and one nucleotide — in this analysis. We performed all-against-all BLAST searches with these data, and plotted the results on dot-matrix plots.

4.2 INTRODUCTION

The hypothesis that the eukaryotic genome has undergone numerous duplications has gained credence from the presence of unlinked duplicated genes (Ohno *et al.*, 1968; Ohno, 1970). Evidence of duplications should be seen by the presence of closely linked pairs of duplicated genes within the genome (Nadeau and Kosowsky, 1991).

Approximately 60% of unlinked duplicated genes appear to be part of duplicated chromosomal segments (Nadeau, 1991). The human genome appears to be composed of a number of short duplicated segments which are not interrupted by unrelated genes and long segments that are always interrupted (Nadeau, 1991).

There are at least four factors which could obscure the evidence for duplications within (and of) the genome. Both members of a duplicated gene pair must be retained if there is to be evidence that duplication took place. However, some duplicated genes will be lost, either by deletion or mutation, and will no longer be present in the genome (see Section 1.4). Secondly, of those that are retained, if a long enough period has elapsed since the duplication event, their sequences may have diverged sufficiently so as to make identification of paralogy difficult. Differential (random) silencing of genes within paralogous segments also causes changes in linkage relationships (Lundin, 1993). When a syntenic group is duplicated, genes on one group may remain active and linked, while the majority of the genes on the duplicated segment may be lost, obscuring the evidence that this segment had been duplicated. Similarly, a number of genes on either segment could be lost, with perhaps only one or two of the duplicated genes being retained. A long paralogous segment may thus only have a few genes left to show that a duplication had occurred. Likewise, differential silencing could cause a single paralogous segment to appear as two or more smaller segments. Thirdly, chromosomal rearrangements, large or small, will disrupt ancestral linkages, and reduce the size and increase the number of paralogous segments within a genome (see Section 2.2.2). However, because paralogous segments are still visible in the human genome, there must be a certain stability of syntenic groups (Lundin, 1993). This could be due to a conservation of certain linkage groups for functional reasons, or because chromosomal rearrangements are selected against. Finally, if many duplication events have occurred, then it becomes more difficult to attribute paralogous segments to any particular duplication event. The presence of overlapping segments, coupled with the fact that average number of gene family members is greater than two, is evidence for multiple duplication events.

Because of these factors, it becomes more difficult to find paralogous segments within a single genome than it is to identify orthologous segments between two genomes. Orthologues are readily identifiable between the human and mouse genomes, where orthologous genes have an amino acid sequence identity of approximately 85% (Makałowski *et al.*, 1996; Makałowski and Boguski, 1998). Speciation involves a single duplication event (i.e., each lineage carries a copy of the same genome), and all conserved linkages and chromosomal rearrangements can be traced to that one duplication event. Within a single genome in which more than one duplication event has occurred (either by regional, chromosomal or genomic duplication), the resulting overlapping segments confound the evidence for any single duplication event, and also make it more difficult to characterize the rate of rearrangement within that genome (Nadeau, 1991).

Despite these complications, many authors have attempted to find paralogous regions within the human genome, and to put these forward as evidence that one or more genome duplications have taken place as proposed by Ohno (1970). One of the earliest examples involved the comparison of cytogenetic bands of the different chromosomes. On this basis, Comings (1972) proposed a series of chromosomal pairs which he thought were ancestral homologues (1 and 2; 4 and 5; 7 and 8; 11 and 12; 14 and 15; 16 and 17; 19 and 20; 21 and 22). In his opinion, the proposal that HSA 11 and 12 were ancestral homologous chromosomes was strengthened by the presence of a single gene pair (LDHA and LDHB), which had members on both of these chromosomes. The pairing of HSA 11 and 12 has, however, been supported by the presence of at least 15 paralogous pairs of genes on these chromosomes (Lundin, 1993) and has been explained by chromosomal duplication due to tetraploidization (Lundin, 1985). Similarly, HSA 4 and 5 have many paralogous genes linking them (Section 2.5;

Lundin, 1993; Imanishi *et al.*, 1997). Other proposals of paralogous regions include HSA 8, 10 and 16 (Lundin, 1993) and HSA 6 and 9 (Lundin, 1989). Nadeau (1991) found a further 19 different paralogous regions, based of a dataset of 126 duplicated genes, which formed 52 different protein families. Examples of possible quadruplicated regions include the *Hox* cluster regions on HSA 2/7/12/17, the MHC regions on HSA 1/6/9/19, and the FGFR regions on HSA 4/5/8/10 (see Section 1.7) and 10q24, 17q21, 19q13, 22q12 (Imanishi *et al.*, 1997).

Recently, Imanishi *et al.* (1997) found 128 pairs of chromosomal regions which involved three or more pairs of homologues, and concluded that extensive chromosomal regions had been frequently duplicated in the past. They also demonstrated that the MHC region of 6p21.3 has more homologous regions than just those proposed by Kasahara *et al.* (Kasahara *et al.*, 1996; Kasahara, 1997; Kasahara *et al.*, 1997) and Katsanis *et al.* (1996) (see Section 1.7.2). In fact, 6p21.3 seems to have 11 homologous regions (1q23, 3p21, 5q21, 7q22, 9q33-q34, 11q21-q23, 17q21, 19p13, 20q13, 21q22, 22q11).

There is a huge diversity in the chromosome pairs that have been proposed. Of the 231 possible pairs that can be made from 22 autosomes, at least 64 (28%) have been proposed in the literature to contain ancient paralogous segments. Many of these proposals have been made on the strength of two or three homologous genes being located on the two chromosomes. This is clearly insufficient, given that an average human chromosome may contain about 4,000 genes, many of which are members of large multigene families. Analysis of simulated maps of a random gene distribution show that half the map could result from coincidental combinations of duplicated genes. From this, Nadeau (1991) concluded that short duplicated segments represent ancestral linkages, while long segments are probably coincidental.

4.3 DATA

4.3.1 SCHULER DATA

Schuler data were obtained as described in Section 3.2.1.

4.3.2 NCBI GENOMES DIVISION DATA

There were two main problems with the Schuler dataset: a) many genes that had already been mapped were not included in their analysis and b) many genes were given multiple locations and had to be discarded. As an alternative, we used the NCBI human dataset from the Genomes Division of the NCBI Entrez Database, which contained 13,401 mapped sequences when we downloaded it in December 1997. Each chromosome map details the sequences found on that chromosome and gives their map locations in base pair units. The base pair units were estimated by NCBI using the positions of STS markers on physical maps of chromosomes. These sequences are not all complete protein sequences or cDNAs: many of them are ESTs or genomic DNAs. Some previously mapped genes that were not included in the Schuler dataset were included in the NCBI data. However, because the map positions are given in precise base pair notation on the NCBI maps, many genes were also excluded because, while the general region in which they are located is known, their precise location in terms of base pair position is not (e.g., the *Hox* genes).

The data in the NCBI set are divided into genes from Unigene and single ESTs from GenBank. The EST sequences were obtained directly from GenBank. The Unigene database was created at NCBI by taking sequences from GenBank and dbEST databanks, and clustering them into similar sets. The release of Unigene available in December 1997 was composed of sequences from GenBank (release 101), and ESTs from dbEST (up till June 30, 1997), making 522,643 sequences subjected to clustering analysis, of which 11,751 were full length mRNAs, 213,885 were 3' EST reads, 270,012 were 5' EST reads, and 26,995 were other EST sequences. From these, there were 45,918 sets resulting from clustering. 6,719 sets contained at least one known gene, 44,525 sets contained at least one EST, and 5,326 sets contained both. Sets consisting of a single 3' end which did not have a poly-A signal were eliminated.

Unigene does not generate consensus sequences from EST clusters but merely lists the sequences making up the cluster. A list of all the set identification numbers with their representative clones was available from http://www.ncbi.nlm.nih.gov/repository/ Unigene/Hs.seq.all. All the sequences used in the cluster analysis were contained in http://www.ncbi.nlm.nih.gov/repository/Unigene/Hs.data. We retrieved these sequences from the consensus displays on the NCBI genomes web site. Table 4.1 details the numbers and lengths of the sequences used from the NCBI database.

Finally, because TBLASTX cannot cope with very large sequences, all sequences greater in length than 25,000 bp were split into chunks of 20,160 bp. Furthermore, longer genomic sequences are likely to contain more than one gene, which may make output from such long sequences awkward to interpret.

4.4 METHODS

4.4.1 DOT-MATRIX PLOTS

Dot matrix plots are a diagrammatic means of representing the relative positions of homologues within a genome (Gibbs and McIntyre, 1970). Each point represents a homologous relationship between the gene at that position on the X-axis and the gene at that position on the Y-axis. Plots were arranged so that the locus on the lower-

c/some	number of sequences	total number of base pairs sequenced	number of sequences > 1 kb	number of sequences > 10 kb	number of sequences > 100 kb	number of sequences > 1 Mb
1	1,248	1,588,841	307	13	1	0
2	1,056	1,245,535	205	16	0	0
3	934	1,188,272	194	4	1	0
4	767	2,703,302	128	18	1	1
5	738	767,949	121	3	1	0
6	753	1,609,037	139	16	3	0
7	1040	3,829,984	190	22	9	0
8	583	722,399	99	3	2	0
9	540	658,413	94	2	1	0
10	631	702,298	100	7	1	0
11	762	1,291,200	170	15	1	0
12	620	1,212,176	151	12	2	0
13	323	1,464,737	38	4	2	1
14	462	531,895	85	5	0	0
15	437	517,412	82	5	0	0
16	420	842,733	74	6	2	0
17	440	893,663	115	11	2	0
18	303	282,889	53	2	0	0
19	324	494,677	68	6	0	0
20	325	341,357	67	4	0	0
21	165	239,727	26	4	0	0
22	229	1,713,182	74	34	4	0
X	303	5,694,733	93	35	20	1
TOTAL	13,401	30,535,411	2,673	247	53	3

Table 4.1. Numbers and lengths of sequences used in the NCBI dataset.

numbered chromosome was always plotted along the X-axis so that all points on the plot would be seen together in one half of the diagram. Possible duplicated segments where both linkage and gene order are conserved will be seen as diagonals of points, where a diagonal is minimally composed of three points and the outermost points delimit the boundaries of the putative duplicated conserved region.

4.4.1.1 SCHULER (SWISSPROT AND EMBL) DATA

To investigate duplicated conserved regions in the human genome, we systematically searched for diagonals in two cases: 1) where the dot-matrix plot included all possible similarities (with a BLASTP score of ≥ 150 (BLOSUM62 matrix)), so that any one gene can form part of many points; and 2) where only the highest-scoring match (i.e. the most similar) for any gene was included, as long as it was over a BLASTP score of 150. The Schuler database places genes into 'bins' (intervals along a chromosome) so consequently many genes have identical positions. The three genes forming a diagonal are not necessarily precisely adjacent, i.e. there may be other genes interrupting the diagonal (see Figure 4.1). These interrupting genes can have homologues either elsewhere on the same chromosome or elsewhere in the genome. Their homologues may also form part of the original paralogous region but may have become silenced to become pseudogenes.

For a diagonal to be considered as representing a potential duplicated conserved region, only diagonals corresponding to regions in which at least approximately 6% of the genes were duplicated were considered. This figure was based on an estimate of a limit of 10 genes in the bin into which a duplicated gene fell, and an average of 10 interrupting genes between duplicated genes (3/50 = 0.06). In addition, interrupting regions with ≥ 15 (sequenced and mapped) unduplicated genes between duplicate pairs,

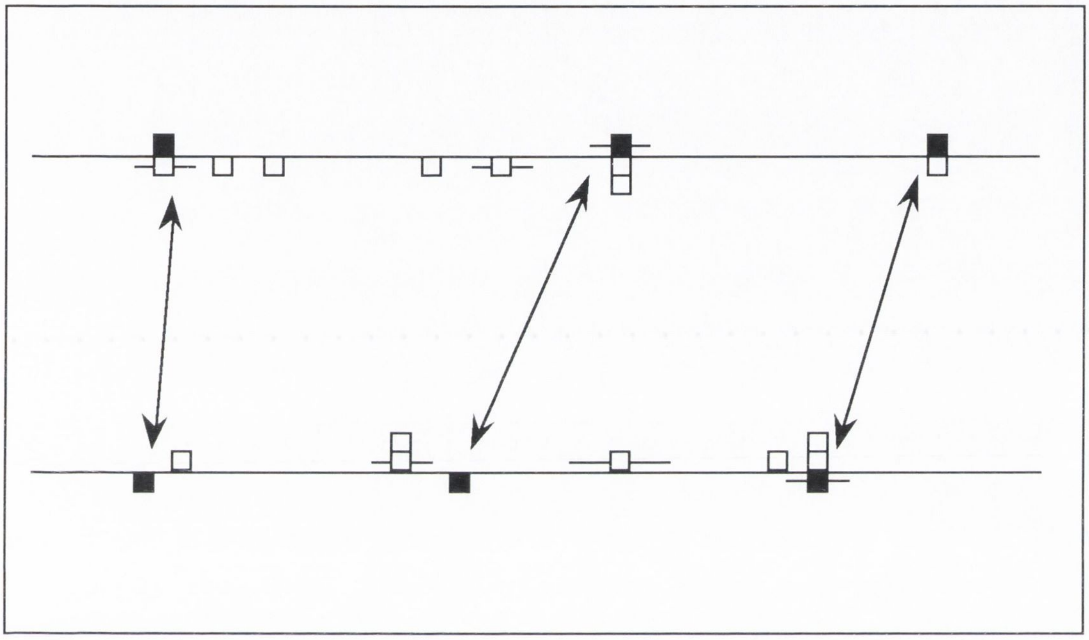


Figure 4.1. Genes and ranges in a potential diagonal. Boxes shaded black are those genes contributing to the diagonal, and paralogues are linked by arrowed lines. All other genes are non-contributing, and interrupting. Boxes stacked on top of each other indicate genes mapped to the same position. Lines through boxes represent the mapping data as a range. Such genes are taken to be positioned at the midpoint of their range.

on either chromosome, were also discarded, as being less likely to reflect a true paralogous region. Distances (in cM) were not taken into account.

4.4.1.2 NCBI Genomes Division data

The dot-matrix plots for the NCBI data were analysed both according to map position on the chromosome (in base pairs) and by the physical order of the mapped genes. All possible similarities above a TBLASTX score of 200 (BLOSUM62 matrix) were included on the plots. Regions were considered to be potential duplicated conserved segments if at least three genes were within 200 genes of each other (or approximately 30 Mb). Gene order was not required to be conserved.

4.4.2 Filtering

The Schuler dataset was composed of protein sequences from both the SWISSPROT and EMBL databases. The NCBI dataset on the other hand is composed of nucleotide sequences, many of which are ESTs. Therefore there are many potential non-coding sequences included in this dataset, such as introns, intergenic sequences and repetitive elements. Because of this, we needed a system of filtering the data to ensure that any similarities we found were likely to be genes and not repetitive DNA elements. BLAST tends to output many false positives which are either biologically insignificant or irrelevant, and these are then difficult to distinguish from the matches that are truly relevant. Here, a distinction must be drawn between merely statistically significant and biologically significant, where biologically significant is meant to imply a genuine common ancestry (which may be unexpected, and might clarify the function of a protein). The most likely causes of highly redundant and/or biologically irrelevant matches are repetitive elements and low complexity regions. To circumvent these problems, PowerBLAST (Zhang and Madden, 1997) was developed at NCBI. PowerBLAST masks repeats and low complexity regions, searches the resulting sequence against the database of choice (non-redundant, SWISSPROT, etc.) using BLAST, and then uses SIM (Huang *et al.*, 1990) to make the best gapped alignment. The output can then be viewed with a graphical browser, Chromoscope. Unfortunately, this client-server is not available in local versions, so searches cannot be performed against one's own database and there is no information on the repeat regions or low complexity regions that are filtered out, so we developed our own version of PowerBLAST.

4.4.2.1 HUMAN REPEAT DATABASE

We developed our own system of filtering. DUST (ftp://ncbi.nlm.nih.gov/pub/tatusov/dust) was used to mask low complexity regions in DNA. Repetitive elements were removed by using TBLASTX with XBLAST (Claverie and States, 1993; Claverie, 1994) against a human repeat database, and regions of the sequence which resulted in matches with a score of ≥ 150 were masked. The human repeat database we used was based on REPBASE (ftp://ncbi.nlm.nih.gov/repository/repbase/REF/humrep.ref). This was modified when we realized that some repeat sequences were not being masked by this method. One of the modifications involved the inclusion of a wider range of Alu sequences, which are known to be frequent in higher eukaryotic genomes and to yield highly significant matches (Gish and States, 1993). We tested three sets of Alu sequences, from NCBI (http://ncbi.nlm.nih.gov/repository/repbase/ALU-ALN), from GenBank, and from GIRI (http://www.girinst.org/~server/publ; Klonowski, 1997). The NCBI set comprised over 18,000 Alu sequences which were all very similar to the one already in the REPBASE database. Both the set of GenBank Alu sequences and those from GIRI eliminated all spurious hits against a few test sequences but using BLAST against the GenBank set was very time-consuming, so we added the GIRI Alu sequences to our human repeat database. We also found that retroviral sequences were another source of non-biologically significant hits. We extracted the regions involved in matches from those sequences which resulted in a number of high-scoring (≥ 150) hits and searched them using XBLAST against a non-redundant database. From this process, we added another 14 retroviral sequences to our human repeat database.

To prevent other spurious hits, we also tagged any base on the query sequence which was hit more than five times in a single BLAST search. This procedure was intended to remove any undiscovered repetitive elements (i.e. those not in REPBASE). Only matches which included a minimum sequence overlap of 200 residues and which were not tagged in this way were considered relevant.

4.4.2.2 SYNONYMOUS VS. NON-SYNONYMOUS SUBSTITUTION RATES

Because of the high density of potentially non-coding sequences in the NCBI database, we added a further filter so that only matches involving coding sequences would be included. It is known that the rate of synonymous substitution is generally much greater than that of non-synonymous substitutions in the amino acid coding regions of genes, where the non-synonymous rate should reflect protein sequence conservation, and the synonymous rate should reflect the local mutation rate and any possible codon selection (Li *et al.*, 1985; Wolfe and Sharp, 1993).

There have been many papers describing how best to estimate the rates of synonymous and non-synonymous substitutions, from the simpler methods (e.g., Miyata and Yasunaga, 1980; Perler *et al.*, 1980; Li *et al.*, 1985) to the more complex and arguably more precise (e.g., Muse and Gaut, 1994; Muse, 1996). All these methods use the codon, as opposed to the nucleotide, as the unit of evolution.

We developed a method which can be used in conjunction with the TBLASTX program which gives the ratio of synonymous to non-synonymous substitutions for each HSP defined by the BLAST program. This allows more certainty as to whether an HSP spans a coding region or not. The use of this method helps eliminate biologically insignificant matches to which high statistical significance might be ascribed because local biases in amino acid composition are not accounted for by the random sequence model assumed in Karlin-Altschul statistics. It may however hide pseudogenes, which are true homologues, but will also have a higher-than-usual non-synonymous substitution rate.

To be able to estimate the rates of synonymous and non-synonymous substitution, we have referred to the older Li-Wu-Luo substitution matrix (Li *et al.*, 1985). Each nucleotide site in a codon is classified as being either degenerate, twofold degenerate, or fourfold degenerate. A site is fourfold degenerate if any change at that site is synonymous, i.e., results in the same amino acid. The third positions of 32 of the 61 sense codons fall into this category. The term twofold degenerate is used if one of the three possible changes is synonymous, usually transitions being synonymous, transversions being non-synonymous. All changes at nondegenerate sites lead to an amino acid replacement. Li *et al.* (1985) define K_S (the number of synonymous substitutions per synonymous site) as:

$$K_S=(L_2A_2 + L_4K_4)/(L_{2/3} + L_4) = 3(L_2A_2 + L_4K_4)/(L_2 + 3L_4)$$

and K_A (the number of non-synonymous substitutions per non-synonymous site) as:

$$K_A=(L_2B_2 + L_0K_0)/(2L_{2/3} + L_0) = 3(L_2B_2 + L_0K_0)/(2L_2 + 3L_0)$$

where L_i = total number of i-fold degenerate sites, i = 0, 2, 4

A_i = transitional substitutions per i-fold site

B_i = transversional substitutions per i-fold site

and K_i = total number substitutions per i-fold site, using Kimura's 2-parameter method to correct for multiple hits separately at zero-, two- and fourfold degenerate sites.

The ratio K_A/K_S now gives us a parameter by which to measure the relative certainty that an HSP actually describes a coding region. We used a K_A/K_S ratio of 0.55 (an approximate excess of synonymous substitutions) to determine whether an HSP is in the coding region or not. In compilations of genes compared between mouse and rat (Wolfe and Sharp, 1993), 95.8% of comparisons were below this threshold.

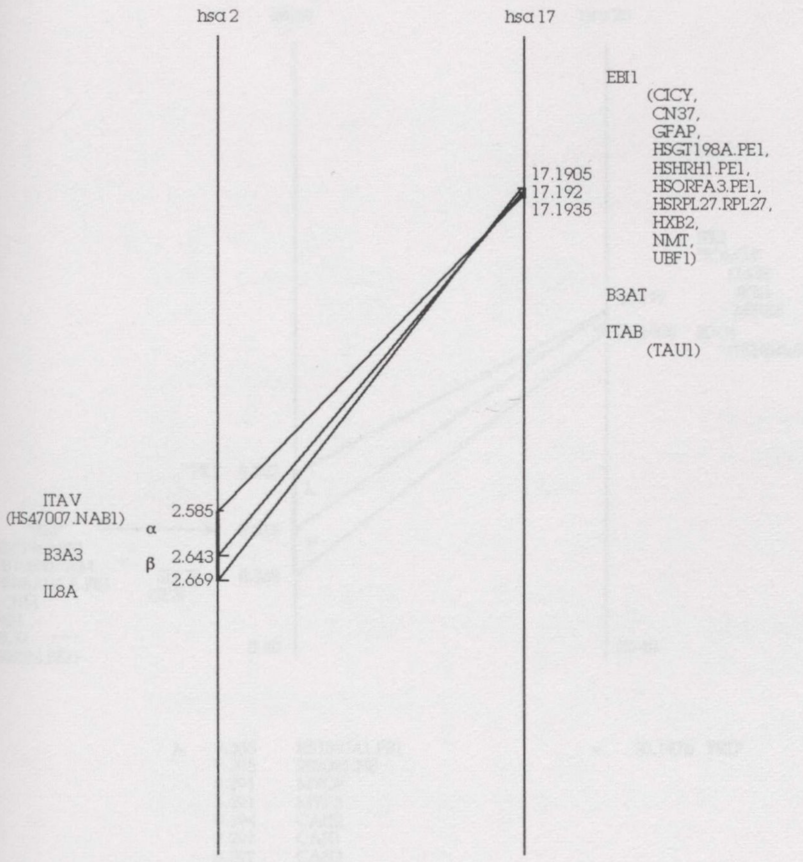
4.5 RESULTS

4.5.1 ANALYSIS OF SCHULER DATA

4.5.1.1 ONE-HIT DOT-MATRIX PLOT

For the one-hit scenario in analysis of the Schuler *et al.* data, where each gene was only permitted to have one paralogue, 15 different diagonals were found with a frequency of duplicated genes of ≥ 0.06 (see section 4.4.1.1). Of these, only eight had less than 16 interrupting genes between any of the genes involved in the diagonal (see Figure 4.2). The eight diagonals found for the one-hit case are listed in Table 4.2. The distances spanned by the FLT3-VGR1-UBL3 and the ANX8-CYPM-CPT7 groups are very long for a region consisting of just three genes (41.5 cM and 61 cM, respectively), and more evidence will be needed to determine whether these regions really represent ancient paralogous regions with KKIT-VGR2-UBL1 and ANX2-CYPB-CP12, respectively, although the sequence similarity between the genes involved is high. No quadruplicated regions were found using the one-hit dot-matrix plot.

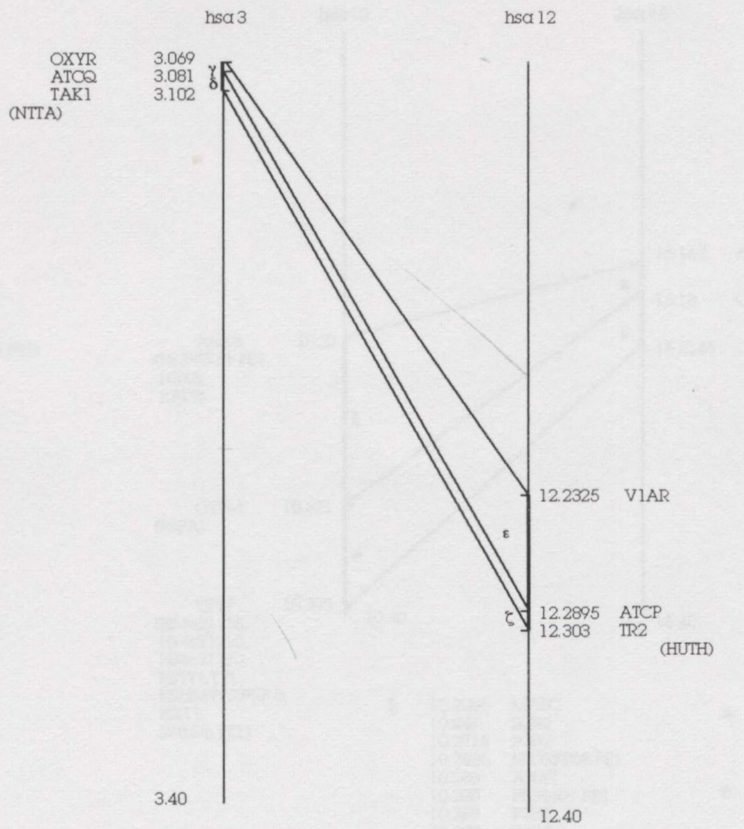
The paralogous region between HSA 8 and HSA 20 includes members of the syndecan (cell surface heparan sulfate proteoglycans) gene family. Spring (1997), in his list of tetralogous groups found in the human genome, included four members of this family (SDC1: 2p24-p23, SDC2: 8q22-q23, SDC3: 1p36-p32, SDC4: 20q12-



α:	2.588	HSLSTAT4R.STAT4
	2.588	SMD2
	2.5955	FRIH
	2.6	INPP
	2.6075	CA25
	2.609	ADO
	2.609	CD28
	2.6135	HS234351.PE1
	2.6135	HS310891.ABLBP3
	2.6135	HS402681.HSORC2
	2.615	PTR2
	2.627	NUAM
	2.6375	CREB
	2.6375	MAP2
	2.6375	MLE1

β:	2.6585	FINC
	2.6585	HSBIFUN.PE1
	2.6585	R37A
	2.6585	STP1
	2.663	KU86
	2.6645	VIL1
	2.666	SG2

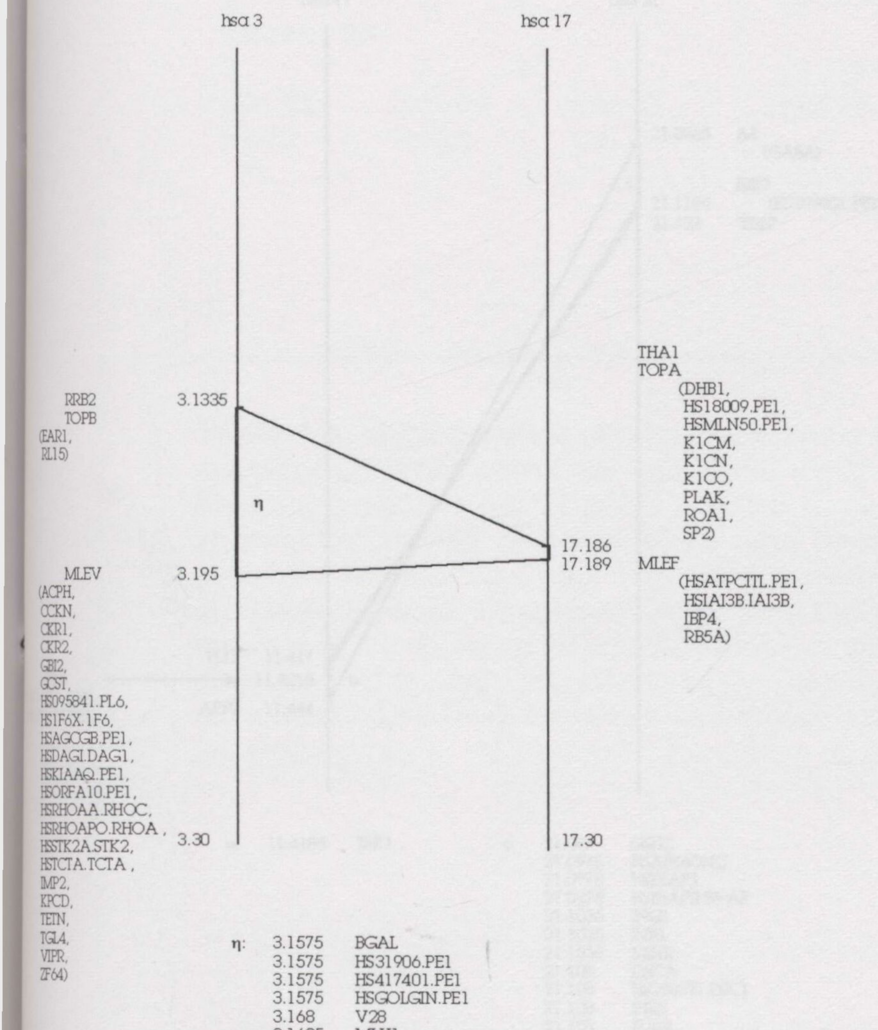
Group 1. Possible conserved duplicated region between hsa 2 and hsa 17.



γ:	3.079500	HH1R
δ:	3.090000	NTG1
	3.090000	NTG3
	3.094500	KRAF
	3.099000	HSAGGCRB.PE1

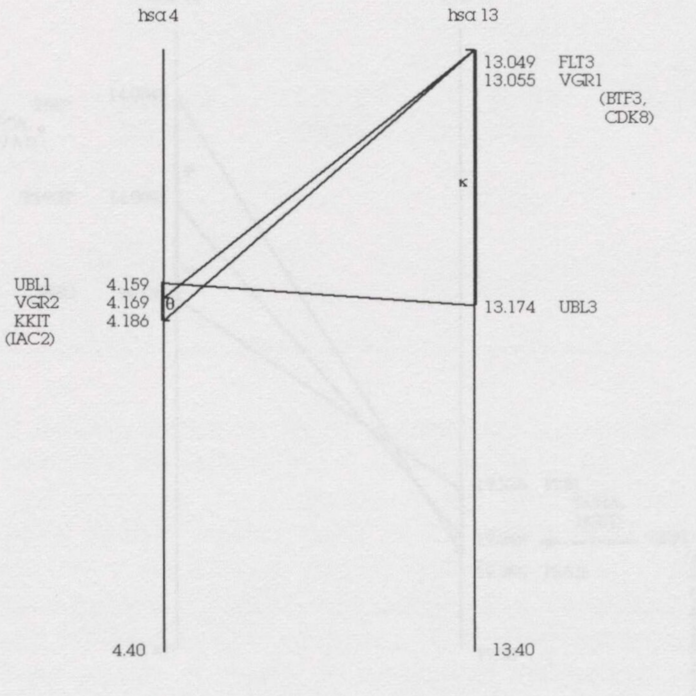
ε:	12.238500	LYC
	12.249000	CO02
	12.259500	GLIP
	12.259500	HS557661.PE1
	12.268500	MYF5
	12.268500	MYF6
	12.268500	SYT1
	12.288000	SCF
ζ:	12.297000	MSSP

Group 2. Possible conserved duplicated region between hsa 3 and hsa 12.



η:	3.1575	BGAL
	3.1575	HS31906.PE1
	3.1575	HS417401.PE1
	3.1575	HSGOLGIN.PE1
	3.168	V28
	3.1695	MLH1
	3.1695	THIK
	3.1785	RSP4
	3.1785	ZN38
	3.192	RL1X

Group 3. Possible conserved duplicated region between hsa 3 and hsa 17.

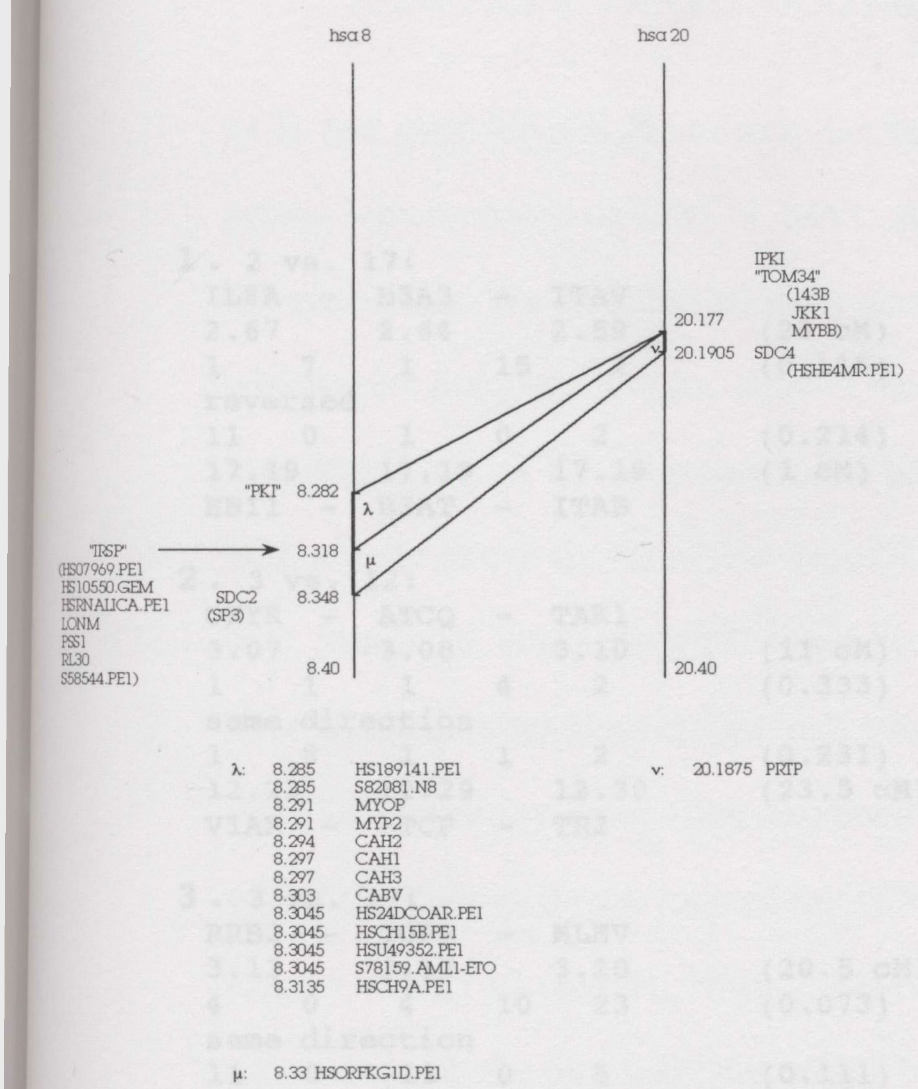


θ:	4.171	NEUU
	4.174	HSU62325.HFE65L
	4.1815	HS13877.PE1
	4.183	CGCC
	4.183	TEC
	4.183	TXK
	4.184	S62907.PE1

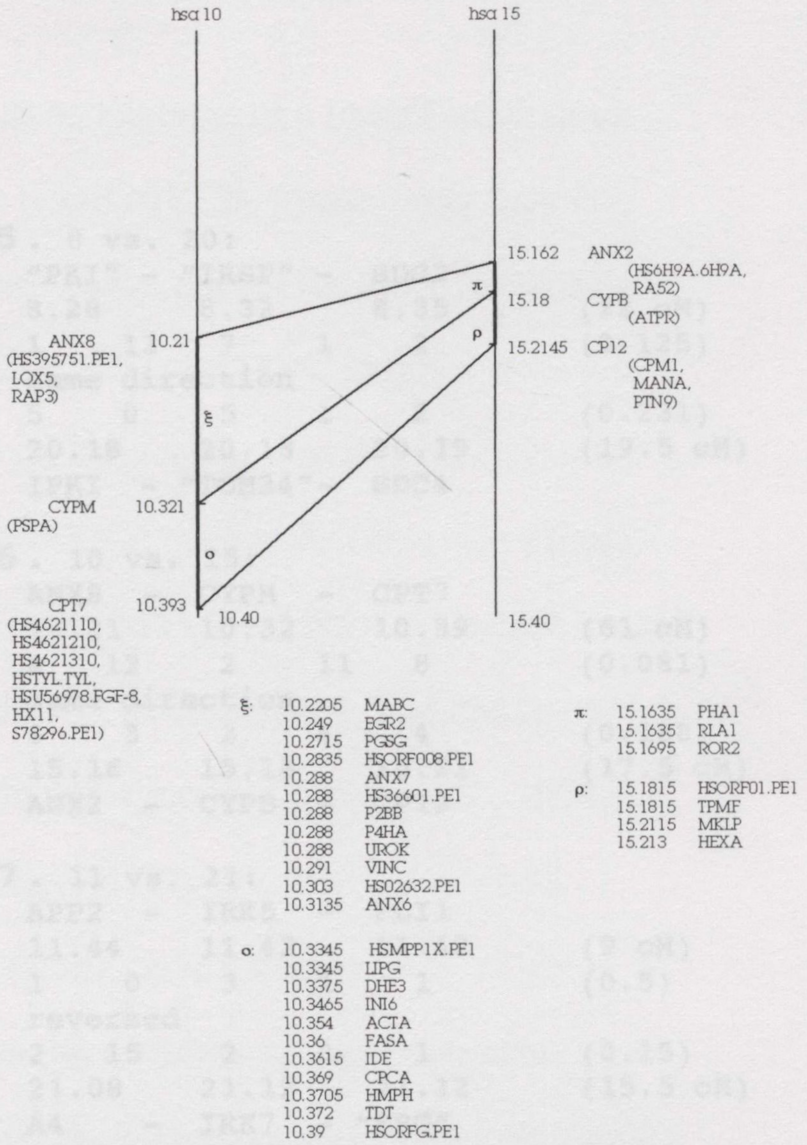
κ:	13.069	HSD13S106.PE1
	13.096	AC13
	13.096	ERR1
	13.096	HSOSF2OS.OSF-2
	13.096	HSOSF2P1.OSF-2
	13.117	PAX3
	13.126	T2FB
	13.129	5H2A
	13.129	HSU35048.PE1
	13.135	RB
	13.138	CBPC
	13.138	PLSL
	13.1395	GCRT

Group 4. Possible conserved duplicated region between hsa 4 and hsa 13.

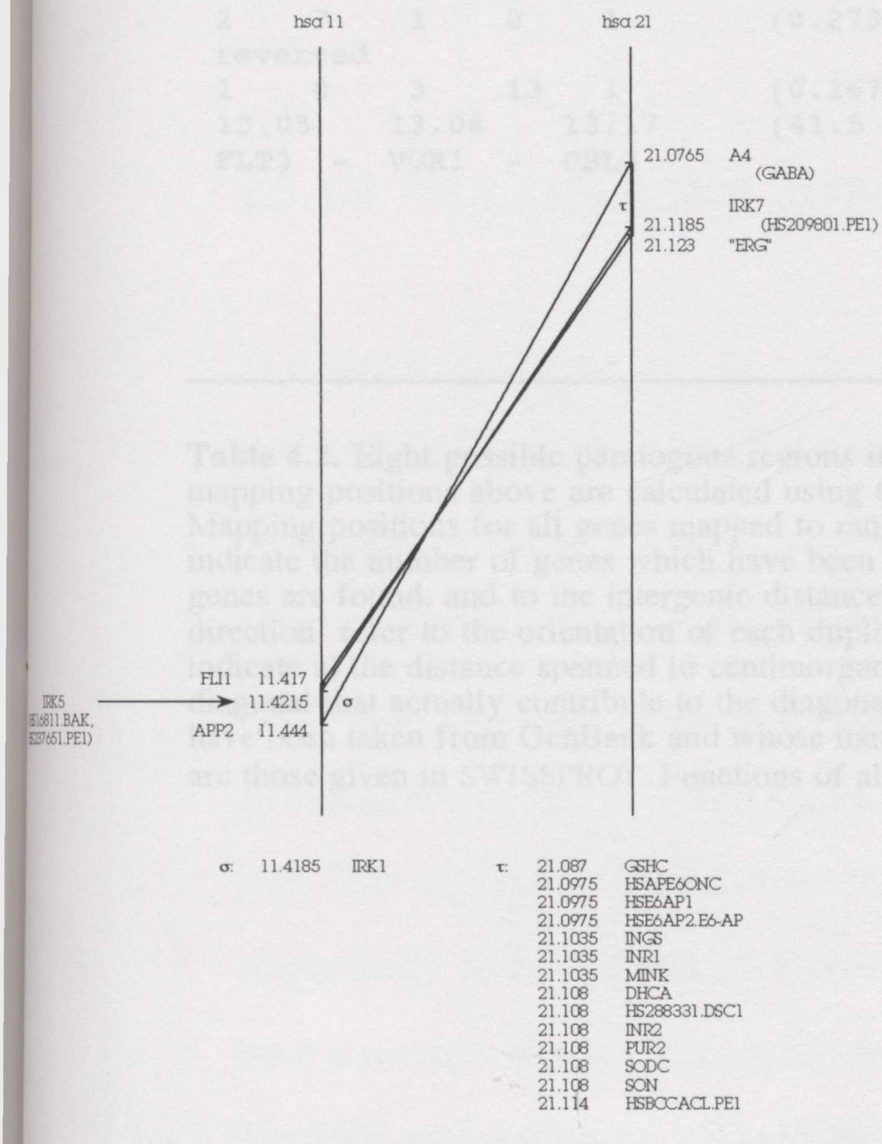
Figure 4.2a. Paralogous regions defined by diagonals of three genes on the one-hit dot-matrix plot. Vertical lines represent the chromosome. Bold lines indicate the region of paralogy between chromosomes. Lines connecting vertical lines indicate homology relationships. Gene names in brackets are genes which have been mapped to the same position as genes contributing to the diagonal. All map locations have been calculated according to the equation: chromosome + (3.location in cM)/1000. Greek letters indicate that genes have been mapped between contributing genes, and those genes are given, marked with the appropriate Greek letter, below each figure.



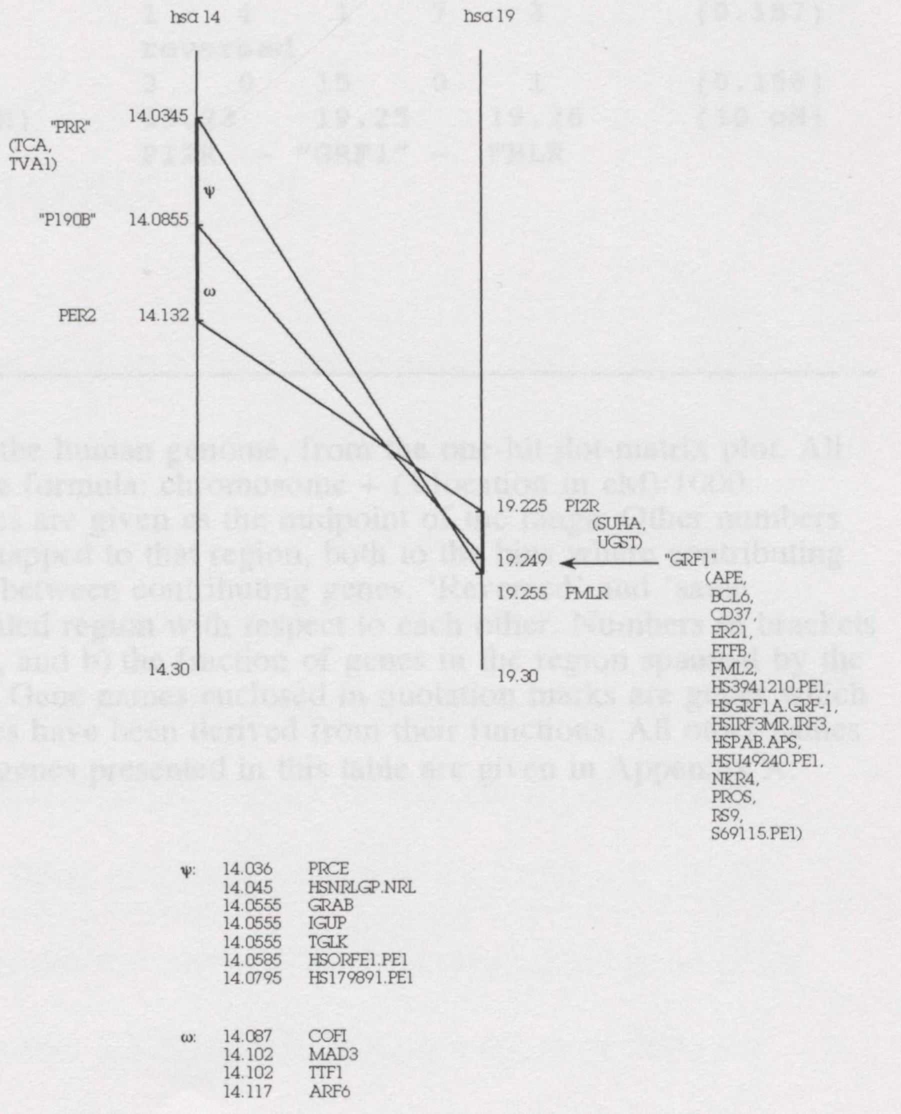
Group 5. Possible conserved duplicated region between hsa 8 and hsa 20 .



Group 6. Possible conserved duplicated region between 10 and hsa 15.



Group 7. Possible conserved duplicated region between hsa 11 and hsa 21.



Group 8. Possible conserved duplicated region between hsa 14 and hsa 19.

Figure 4.2b.

1. 2 vs. 17:

IL8A	-	B3A3	-	ITAV	
2.67		2.64		2.59	(28 cM)
1	7	1	15	2	(0.115)
reversed					
11	0	1	0	2	(0.214)
17.19		17.19		17.19	(1 cM)
EBI1	-	B3AT	-	ITAB	

2. 3 vs. 12:

OXYR	-	ATCQ	-	TAK1	
3.07		3.08		3.10	(11 cM)
1	1	1	4	2	(0.333)
same direction					
1	8	1	1	2	(0.231)
12.23		12.29		12.30	(23.5 cM)
V1AR	-	ATCP	-	TR2	

3. 3 vs. 17:

RRB2	-	TOPB	-	MLEV	
3.13		3.13		3.20	(20.5 cM)
4	0	4	10	23	(0.073)
same direction					
11	0	11	0	5	(0.111)
17.19		17.19		17.19	(1 cM)
THA1	-	TOPA	-	MLEF	

4. 4 vs. 13:

KKIT	-	VGR2	-	UBL1	
4.19		4.17		4.16	(9 cM)
2	7	1	0	1	(0.273)
reversed					
1	0	3	13	1	(0.167)
13.05		13.06		13.17	(41.5 cM)
FLT3	-	VGR1	-	UBL3	

5. 8 vs. 20:

"PKI"	-	"IRSP"	-	SDC2	
8.28		8.32		8.35	(22 cM)
1	13	7	1	2	(0.125)
same direction					
5	0	5	1	2	(0.231)
20.18		20.18		20.19	(19.5 cM)
IPKI	-	"TOM34"	-	SDC4	

6. 10 vs. 15:

ANX8	-	CYPM	-	CPT7	
10.21		10.32		10.39	(61 cM)
4	12	2	11	8	(0.081)
same direction					
3	3	2	4	4	(0.188)
15.16		15.18		15.21	(17.5 cM)
ANX2	-	CYPB	-	CP12	

7. 11 vs. 21:

APP2	-	IRK5	-	FLI1	
11.44		11.42		11.42	(9 cM)
1	0	3	1	1	(0.5)
reversed					
2	15	2	0	1	(0.15)
21.08		21.12		21.12	(15.5 cM)
A4	-	IRK7	-	"ERG"	

8. 14 vs. 19:

PER2	-	"P190B"	-	"PRR"	
14.13		14.09		14.03	(32.5 cM)
1	4	1	7	3	(0.187)
reversed					
3	0	15	0	1	(0.158)
19.23		19.25		19.26	(10 cM)
PI2R	-	"GRF1"	-	FMLR	

Table 4.2. Eight possible paralogous regions in the human genome, from the one-hit dot-matrix plot. All mapping positions above are calculated using the formula: chromosome + (3.location in cM)/1000. Mapping positions for all genes mapped to ranges are given as the midpoint of the range. Other numbers indicate the number of genes which have been mapped to that region, both to the bins where contributing genes are found, and to the intergenic distances between contributing genes. 'Reversed' and 'same direction' refer to the orientation of each duplicated region with respect to each other. Numbers in brackets indicate a) the distance spanned in centimorgans, and b) the fraction of genes in the region spanned by the diagonal that actually contribute to the diagonal. Gene names enclosed in quotation marks are genes which have been taken from GenBank and whose names have been derived from their functions. All other names are those given in SWISSPROT. Functions of all genes presented in this table are given in Appendix A.

q13), and gave the bHLH transcription factors as an example of a linked tetralogous group, which consist of MYCN, MYC, MYCL1, and MYCB, respectively linked to the syndecans. We included all these genes in our study, except for MYCB and SDC3, which were not present in the Schuler dataset. The SDC and MYC families were not involved in any of the diagonals found because the only syndecans that hit each other with a score above the threshold were SDC2 and SDC4 (Group 5; Table 4.2; Figure 4.2). Similarly, while all MYC genes hit each other, MYCB on HSA 20, which would have added another gene to Group 5, was not present in our dataset. The linkage of these two tetralogy groups has also been described for mouse (Spring *et al.*, 1994), where the gene pairs are located on: MMU 2 (Synd4, Bmyc), MMU 4 (Synd3, Lmyc), MMU 12 (Synd1, Nmyc), MMU 15 (Synd2, myc). Spring *et al.* (1994) concluded that the physical relationship between the members of these two gene families appears to be ancient and conserved after the two genome duplications thought to have occurred during vertebrate evolution.

4.5.1.2 MULTIPLE-HIT DOT-MATRIX PLOT

72 putative paralogous regions, indicated by diagonals of three genes with a frequency of ≥ 0.06, were found on the all-hit plot, which included all hits found above a threshold of 150 (see Figure 4.3). Of these, 19 had ≤ 15 unique genes between any genes involved in the diagonal. These are detailed in Table 4.3. Five of them recapitulate diagonals found on the one-hit dot-matrix plot. One diagonal (group 19) is an elongation of the HSA 8/20 region identified previously on the one-hit dot-matrix plot. This region now extends over five pairs of genes, spanning 43.5 cM and 14 cM, respectively on HSA 8 and 20. However, all members of both newly added gene pairs have a stronger hit elsewhere in the genome: LYN hits HCK with a score of 985, but hits SRC with a score of 1041; HCK hits LCK with a score of 1768. Likewise, both MYBA and MYBB hit MYB more strongly (scores of 875 and 715, respectively) than

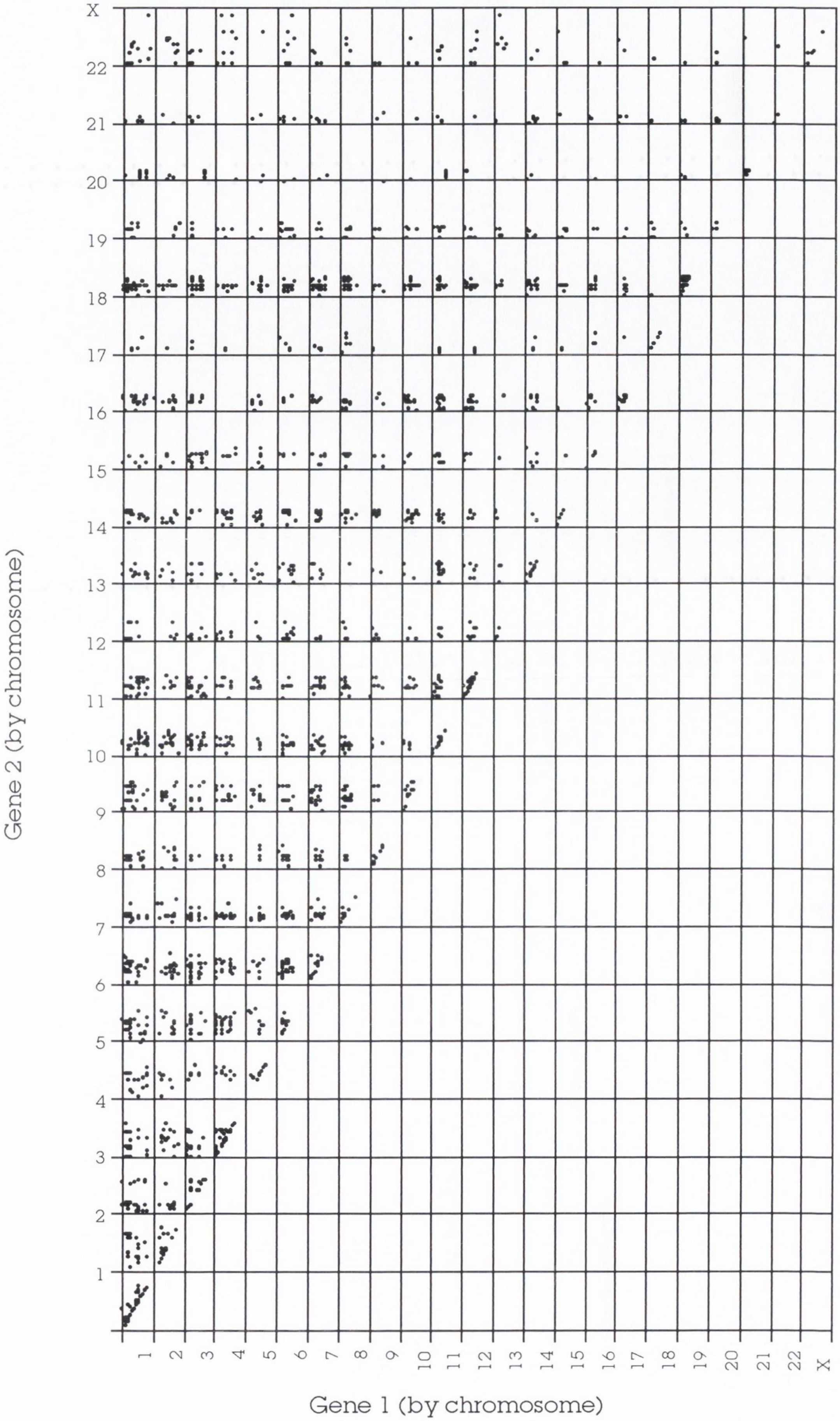


Figure 4.3. Dot-matrix plot of all homologous relationships within the human genome using the Schuler *et al*. (1996) data. Homologues were plotted by the lower number chromosome on the X-axis. Each chromosome was given a maximum value of 330 cM, by using the equation: chromosome + (3.location)/1000. There are 4,072 homologous relationships plotted. The dataset was composed of 2,317 genes from the Schluer dataset.

```
1. 2 vs. 7:
 GRAN  -  TFP1  -  PTR2
 2.52     2.58     2.62        (32.5 cM)
 1    14    2   12    1        (0.100)
 same direction
 2    3     3    0     3       (0.273)
 7.29     7.32     7.32        (10.5 cM)
 SORC  -  TFP2  -  CALR

2. 2 vs. 8:
 NTC1  -  GLVR1 -  PLMN
 2.42     2.37     2.37        (16.5 cM)
 2     11   3    0    3        (0.158)
 reversed
 1    0     3    0    3        (0.428)
 8.19     8.19     8.19        (0.5 cM)
 ANK1  -  GLVR2 -  UROT

3. 2 vs. 17:
 IL8A  -  B3A3  -  ITAV
 2.67     2.64     2.59        (28 cM)
 1    7     1    15   2        (0.11)
 reversed
 11   0     1    0    2        (0.214)
 17.19    17.19    17.19       (1 cM)
 EBI1  -  B3AT  -  ITAB

4. 3 vs. 12:
 "C338" - GTR2  -  CRAR
 3.52     3.55     3.63        (3 cM)
 1    3     4    15   9        (0.094)
 same direction
 9    1     9    0    9        (0.107)
 12.05    12.05    12.05       (1 cM)
 C5AR  -  GTR3  -  C1S

5.
 OXYR  -  ATCQ -  TAK1
 3.07     3.08     3.10        (11 cM)
 1    1     1    4    2        (0.333)
 same direction
 1    8     1    1    2        (0.231)
 12.23    12.29    12.30       (23.5 cM)
 V1AR  -  ATCP  -  TR2

6. 3 vs. 17:
 RRB2  -  TOPB  -  ZN38
 3.13     3.13     3.18        (15 cM)
 4    0     4    7    2        (0.176)
 same direction
 11   0     11   0    11       (0.091)
 17.19    17.19    17.19       (0 cM)
 THA1  -  TOPA  -  SP2

7. 5 vs. 8:
 DADR  -  KFMS  -  EGR1
 5.54     5.45     5.43        (38.5 cM)
 2    13    2    4    4        (0.120)
 reversed
 8    7     2    3    1        (0.143)
 8.19     8.22     8.25        (19 cM)
 B3AR  -  LYN   -  ZN07
```

```
8. 5 vs. 10:
 PDGR  -  EGR1  -  CTNA
 5.45     5.43     5.43        (7.5 cM)
 2    4     4    0    4        (0.214)
 reversed
 3    5     1    7    1        (0.176)
 10.21    10.25    10.29       (28 cM)
 RET   -  EGR2  -  VINC

9. 6 vs. 11:
 5H1E  - "PTPK" -  OPRM
 6.28     6.38     6.47        (62 cM)
 4   13     2   12    1        (0.094)
 same direction
 7    0     7    0    2        (0.333)
 11.18    11.18    11.19       (0.5 cM)
 ACM4  -  PTPB  -  APJ

10. 8 vs. 17:
 NFL   -  EGR3  -  NFM
 8.13     8.14     8.15        (7 cM)
 2    0     3    2    5        (0.250)
 same direction
 11   0     11   0    11       (0.091)
 17.19    17.19    17.19       (0 cM)
 K1CO  -  SP2   -  K1CM

11. 8 vs. 18:
 KG1P  -  1D12A - "NFAT4"
 8.15     8.17     8.18        (8.5 cM)
 5    2     2    1    4        (0.214)
 reversed
 7    7     1    4    1        (0.150)
 18.19    18.28    18.36       (57.5 cM)
 LIV1  - "CGRP" - "NFATC"

12. 9 vs. 13:
 TRKA  -  PAX5  -  TYR1
 9.25     9.18     9.07        (62 cM)
 1    7     5    15   1        (0.103)
 reversed
 3    5     1    9    1        (0.158)
 13.06    13.12    13.23       (57 cM)
 VGR1  -  PAX3  -  TYR2

13. 10 vs. 15:
 ANX6  -  CYPM  -  CPCA
 10.31    10.32    10.37       (18.5 cM)
 1    0     2    7    1        (0.273)
 same direction
 3    3     2    4    4        (0.188)
 15.16    15.18    15.21       (17.5 cM)
 ANX2  -  CYPB  -  CP12

14. 10 vs. 17:
 VIM   -  FBRNP -  EGR2
 10.12    10.21    10.25       (5 cM)
 3   10     3    5    1        (0.136)
 same direction
 11   0     11   0    11       (0.273)
 17.19    17.19    17.19       (0 cM)
 K1CO  -  ROA1  -  SP2
```

15. 12 vs. 16:

```
NTBE   -   VWF    -   C1R
12.01      12.05      12.06         (17.5 cM)
1    8     9   10     1             (0.103)
same direction
1    7     1    8     16            (0.091)
16.16      16.23      16.26         (32 cM)
NTNO   -   MT1A   -   HPT2
```

16. 12 vs. 18:

```
CPSS   -   MLRV   -   PTNB
12.33      12.35      12.37         (12.5 cM)
4    7     3    1     12            (0.111)
same direction
2    0     2    0     1             (1.000)
18.02      18.02      18.07         (17.5 cM)
"C338" -   MLRM   -   PTPM
```

17. 12 vs. X:

```
ATCE   -   PAXI   -   UBIQ
12.37      12.42      12.48         (30.5 cM)
12    4     2    2     3            (0.130)
same direction
5    11     1    9     10           (0.083)
X.29       X.46       X.58          (97.5 cM)
AT7A   -  "FHL1"  -   UBIL
```

18. 13 vs. 19:

```
CDK8   -   VGR1   -   GCRT
13.06      13.06      13.14         (28 cM)
3    0     3   12     1             (0.188)
reversed
5    0     5   11     12            (0.107)
19.19      19.22      19.25         (4 cM)
ERK3   -   UFO    -   GPR4
```

19. 8 vs. 20:

```
LYN    -   MYBA   -  "PKI"   -  "IRSP"  -   SDC2
8.22       8.26       8.28       8.32       8.34        (43.5 cM)
2    5     1    0     1    13    7    1     2           (0.281)
same direction
2   10     5    0     5    0     5    1     2           (0.150)
20.15      20.18      20.18      20.18      20.19       (14 cM)
HCK    -   MYBB   -   IPKI   -  "TOM34"-    SDC4
```

Table 4.3. 19 possible paralogous regions in the human genome, from the multiple-hit dot-matrix plot. All mapping positions above are calculated using the formula: chromosome + (3.location in cM)/1000. Mapping positions for all genes mapped to ranges are given as the midpoint of the range. Other numbers indicate the number of genes which have been mapped to that region, both to the points where contributing genes are found, and to the intergenic distances between contributing genes. 'Reversed' and 'same direction' refer to the orientation of each duplicated region with respect to each other. Numbers in brackets indicate a) the distance spanned in centimorgans, and b) the frequency of genes contributing to the diagonal over the genes present in the region spanned by the diagonal. Gene names enclosed in quotation marks are genes which have been taken from GenBank and whose names have been derived from their functions (listed below). All other names are those given in SWISSPROT. Functions of all genes presented in this table are given in Appendix A.

they hit each other (score of 700). Four groups (9, 11, 12 and 17) span very long regions (over 55 cM) and are probably not indicative of real paralogous regions, being more likely to have arisen by chance.

Possible quadruplicated region between HSA 8/10/17/12

Examination of the 19 possible paralogous regions found on the multiple-hit dot-matrix plot indicates that there is one trio of genes on HSA 17 which includes genes which are involved in diagonals with two other chromosomes, HSA 8 and HSA 10. These are K1CO and SP2 on HSA 17, which are paired both with NFL and EGR3 on HSA 8, and with VIM and EGR2 on HSA 10 (groups 10 and 14 in Table 4.3). A diagonal between HSA 8 and HSA 10 involving an adjacent region (although not the same genes) is also found (P2AB-UROT-ANXD on HSA 8 with P2BB-UROK-ANX7 on HSA 10) but was not included in the 72 diagonals because it had a frequency of lower than 0.06 (0.058 and 0.2). HSA 8 and HSA 10 also form diagonals around that region with HSA 12 (HSA 8 versus HSA 12: LYN-PTNA-"TTNB" with ERB3-PAC1-CPSS; HSA 10 versus HSA 12: "NF66"-EGR2-RET with K2CD-SP1-ERB3), where ERB3 is in the centre of both regions, possibly indicating a fourfold duplication of that region (see Figure 4.4). By inspecting the BLAST results for all pairwise combinations of the four chromosomes involved, we added other gene pairs within the boundaries of the putative paralogous segments. It should be noted that, unlike the paralogous segments found in yeast by Wolfe *et al.* (Wolfe and Shields, 1997; Seoighe and Wolfe, 1999), the genes here are not in positional order in the segments relative to each other. All genes contributing to diagonals involved in this region are listed in Table 4.4.

Table 4.4 shows that there are 28 main groups of genes in this region. There are seven groups which contain gene family members on three or more chromosomes. Of these,

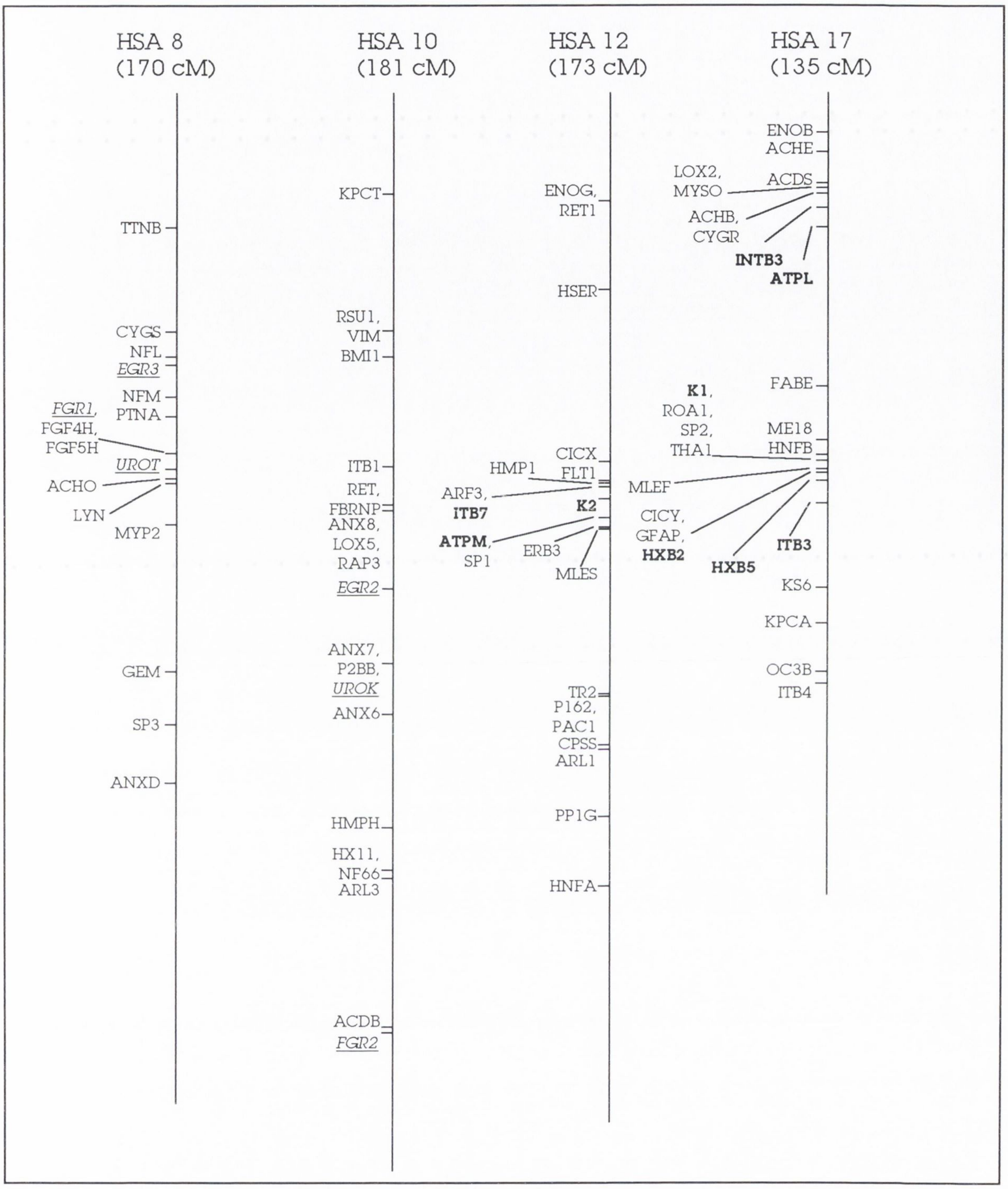


Figure 4.4. Possible quadruplicated region on HSA 8/10/12/17. Vertical lines indicate the chromosomes which are drawn to scale (length in cM given in brackets beneath the chromosome name). Genes which have been mapped to a range rather than a specific point have been mapped to the midpoint of their given ranges. Genes highlighted in italics and which are underlined indicate overlap of this region with the Pébusque *et al.* (1998) FGFR region. Genes highlighted in bold mark those genes which are associated with the *Hox* clusters.

Group number	functions	hsa 8	hsa 10	hsa 12	hsa 17
1	transcription factors	SP3 (116) EGR3 (45.3)	EGR2 (83)	SP1 (71)	SP2 (62)
2	intermediate filament proteins	NFL (44) NFM (51)	VIM (39.5) "NF66" (131)	K2C1 (64.5) K2C4 (64.5) K2C5 (64.5) K2CD (64.5) K22E (64.5) K22O (64.5)	K1CM (62) K1CN (62) K1CO (62) K1CT (62) GFAP (63.5)
3	plasminogen activators	UROT (64.5)	UROK (96)		
4	protein kinase (PK) receptors	LYN (72.5) FGR1 (63) FGF4H (63) FGF5H (63)	RET (69) FGR2 (158)	ERB3 (73)	
5	protein phosphatases	P2AB (60.5)	P2BB (96)	PP1G (122)	
6	PK phosphatases	PTNA (54.5)		PAC1 (101.5)	
7	annexins	ANXD (132)	ANX6 (104.5) ANX7 (96) ANX8 (70)		
8	C-proteins	"TTNB" (4)		CPSS (109.5)	
9	ribonucleoproteins		"FBRNP" (69)		ROA1 (62)
10	hormone receptors			TR2 (101)	THA1 (62)
11		GEM (106)	RAP3 (70)		
12	myelin, retinol- and fatty acid-binding	MYP2 (97)		RET1 (18)	FABE (49)
13	guanylate cyclase, enterotoxin receptor	CYG5 (22.5)		HSER (33)	CYGR (17)
14	acetylcholinesterases	ACHO (65.5)			ACHE (10) ACHB (17)
15	Ras suppressor proteins		RSU1 (39.5)	"FLT1" (65)	
16	fibronection receptors (integrin ß)		ITB1 (62.5)	ITB7 (66)	ITB3 (69) ITB4 (99)
17	ADP ribosylation factors		ARL3 (132.5)	ARF3 (66) ARL1 (110.5)	
18	acyl CoA dehydrogenases		ACDB (157)		ACDS (15)
19	arachidonate 5-lipoxygenases		LOX5 (70)		LOX2 (16)
20	DNA-binding proteins		BMI1 (44.5)		ME18 (57.5)
21	homeobox proteins		HMPH (123.5) HX11 (131)		HXB2 (63.5) HXB5 (65)
22	protein kinase C, synaptotagmin		KPCT (17)		KS6 (83) KPCA (89)
23	enolases			ENOG (18)	ENOB (7)
24	ATP synthases			ATPM (71)	ATPL (22.5)
25	hepatocyte nuclear factors			HNFA (134)	HNFB (61.5)
26	myosin light chains			MLES (73.5)	MLEF (63)
27	dihydropyridine-sensitive L-type calcium channel subnunits			CICX (62)	CICY (63.5)
28	transcription factors			HMP1 (65.5)	OC3B (97)

Table 4.4. Genes mapped to the regions covered by the potential four-fold duplication on HSA 8/10/12/17. The order of contributing genes was not taken into account, since gene order varies between different chromosomes. The *Hox* genes, if two members of any paralogy group had been included in this study, would contribute to the regions on HSA 12 and 17. We do have two *Hox* genes from the *HoxB* cluster included here (HXB2 and HXB5). All spaces left blank indicate that there is no orthologous gene known for that family on that chromosome. Numbers in brackets indicate the map location of the gene in centimorgans.

only one has members on all four chromosomes (intermediate filament proteins). Five are composed of unique or almost unique sequences (i.e. from small families) (serine/threonine protein phosphatases, guanylate cyclases and enterotoxin receptor, myelin and retinol- and fatty acid-binding receptors, fibronectin receptors (integrin ß) and the SP transcription factors). Another group is composed of protein kinases, which form a large and varied multigene family. Protein kinase genes are found on most chromosomes, so any data from them are unlikely to be of much significance. Table 4.5 shows the topology of the trees formed by the most informative groups. The results from groups 1 and 5, pairing HSA 8 and HSA 12, are contradicted by the results from groups 2 and 13, which pair HSA 12 with HSA 17 instead. However, as pointed out by Toby Gibson (pers. comm.), if two genome duplications occurred within a short time-frame of each other around 500 Mya, then it would be difficult to get reliable and informative trees.

Table 4.6 gives a statistical breakdown of the genes involved in the possible quadruplicated region. HSA 12 and 17 appear to be the most similar, having the highest number of gene pairs in common. HSA 8 and 17 seem to be the least similar. The number of similarities between these four chromosomes, given the limited amount of data available to us, suggests that they may indeed indicate a paralogous segment, but whether that segment is really a quadruplicated one or just two distinct segments, possibly on HSA 8/HSA 10, and HSA 12/HSA 17, is open to conjecture. The area discussed here spans some very large regions (35-77% of the chromosome lengths; Table 4.6), and it is unlikely that such regions would have remained intact during the 250-500 Myr which have elapsed since the last hypothesized tetraploidization event. If this is a quadruplicated region, then as the human gene map is extended, more gene families should appear in the gaps to help link these four chromosomes.

If we look at the other chromosomes which genes involved in this region also hit, genes on HSA 8 and 12 hit HSA 11 more than once, genes on HSA 8 and HSA 10

Group number (from Table 4.4)	Topology (chromosome numbers)	Gene family
1	(8, 12), 10	SP transcription factors
2	(8, 10), (12, 17)	intermediate filament proteins
5	(8, 12), 17	PK phosphatases
12	(8, 17), 12	myelin and retinol- and fatty acid-binding receptors
16	(10, 12), 17	fibronectin receptors (integrin ß)
13	(12, 17), 8	guanylate cyclases and enterotoxin receptor

Table 4.5. Topology of phylogenetic trees formed by members of gene families found on three or more chromosomes involved in the potential quadruplicated region. Trees were constructed using CLUSTALW. Chromosomes bracketed together indicate that genes on those chromosomes have a higher degree of similarity to each other than they do to any other member of that gene family involved in this region.

c/somes involved	# sets	# sets involving this pair of the four c/somes only	# sets from column 3 which hit no other c/some	length spanned in cM / length of c/some1 in cM	length spanned in cM / length of c/some 2 in cM
8 vs 10	7	3	1	88/170 (0.52)	118.5/181 (0.65)
8 vs 12	8	2	1	112/170 (0.66)	104/173 (0.60)
8 vs 17	5	1	0	93.5/170 (0.55)	53.5/135 (0.40)
10 vs 12	7	2	1	118.5/181 (0.65)	60/173 (0.35)
10 vs 17	9	5	3	140/181 (0.77)	84/135 (0.62)
12 vs 17	13	8	4	116/173 (0.67)	92/135 (0.68)

Table 4.6. Pairwise comparisons of all chromosomes involved in the potential quadruplicated region. Numbers in brackets indicate the fraction of the length of that chromosome covered by the duplicated segment.

hit genes on HSA 2, 4 and 5, and genes on HSA 12 and 17 hit genes on HSA 1, 2 and 7. This just illustrates that while it is tempting to put forward chromosomal regions which contain a number of similar genes as paralogous or even as quadruplicated regions, the real story is much more complex, with many rearrangements or independent gene duplications having occurred since the last postulated genome duplication event.

If there is a genuine quadruplicated segment on HSA 8, 10, 12 and 17, we would expect the topology of phylogenetic trees drawn from the sequences to indicate this. Instead, the tree topologies are confused. If these are real ancient paralogous segments then, although gene order need not be conserved (due to the "waltzing" genes phenomenon; see Section 2.2.1), the genes involved should be spaced quite close together (allowing for differential silencing of intermediate genes) over smaller fractions of the chromosomes involved. It is, however, notable that genes on HSA 8 and 10 hit genes on HSA 4 and 5 more often than they hit genes on any other chromosome, and that genes on HSA 12 and 17 hit genes on HSA 2 and 7. Pébusque *et al.* (1998) have described a quadruplicated region over HSA 4, 5, 8 and 10 (see Section 1.7.3), and our potential quadruplicated region includes some of the genes used in their analysis; see Figure 4.4). The *Hox* clusters, one of the most widely cited examples of a quadruplicated region put forward as evidence for a double duplication event in the evolution of vertebrates, are found on HSA 2, 7, 12 and 17 (see Section 1.7.1). Of the non-*Hox* genes that are associated with the *Hox* clusters, there are five gene families that have homologues on both HSA 12 and 17 that we could expect to see. Of these, we have included the integrins and ATP synthase genes (marked on Figure 4.4). The three remaining gene families (collagens, epidermal growth factor receptors and synaptobrevin) only have one member included in the Schuler dataset. This could either indicate that what has already been described by other authors really are quadruplicated regions, but equally, without more genes to fit into these regions, it could imply that each "quadruplicated" region is in fact just two separate paralogous regions, although the clustering of genes on HSA 2/7/12/17 has been shown to be highly significant (Ruddle *et al.*, 1994b).

4.5.2 ANALYSIS OF NCBI GENOMES DIVISION DATA

In Figure 4.5, we show the results of all-against-all TBLASTX searches using the DNA sequences that appear on the NCBI physical map of human chromosomes (dataset described in Section 4.3.2). From 13,401 sequences, we found only 348 matches. These define 102 protein families (detailed in Appendix B), ranging in size from 71 families with two members only, to a family of 20 members (the zinc finger proteins) and a 24-member family (the receptor tyrosine kinases). Only 14 families have four or more members. There are 254 possible pairs among the 23 human chromosomes (excluding Y). We were able to identify at least three pairs of similar sequences in 48 of these. However, in none of these pairwise comparisons were three genes spaced closer together than 100 (mapped) genes apart on both chromosomes. The only segment where there were three genes spaced closer than 200 genes apart on both chromosomes was on HSA 1/3. This segment spans 107 mapped genes on HSA 1 and 63 on HSA 3, and includes two HEK ephrin receptor genes, glucose transporters SGLT1 and GLUT2, and two diacylglycerol kinases. Gene order within this segment is not conserved. There were seven other chromosome pairs, including two which involved the *Hox* chromosomes, which could be considered as potentially duplicated regions, although all the genes involved tend to be quite widely spaced and are more likely to be a product of coincidence. These are detailed in Table 4.7.

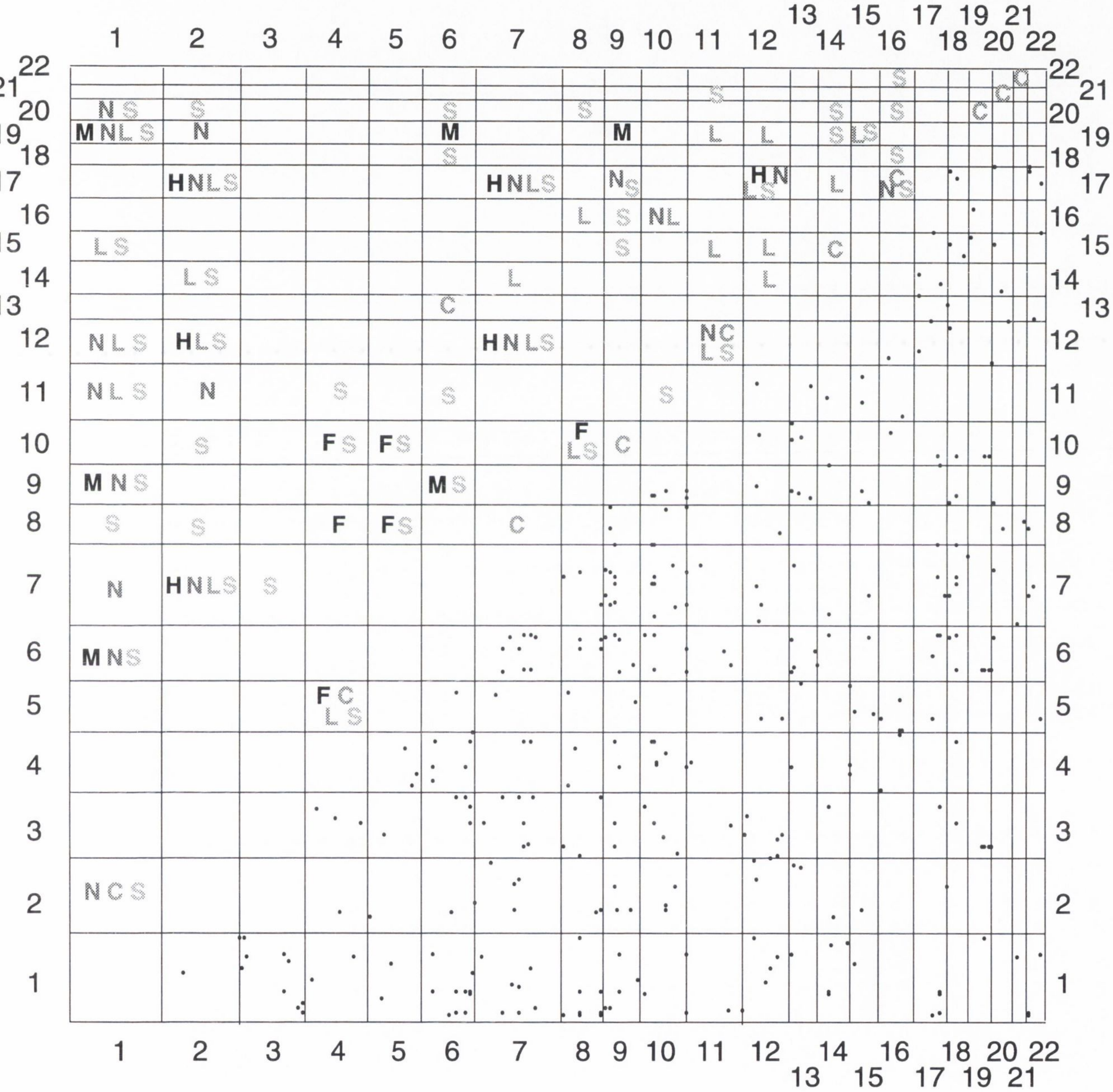


Figure 4.5. Dot-matrix plot of all homologous relationships wihin the human genome using the NCBI Genomes Division data, and comparison to previous proposals. Above diagonal (bottom left to top right): human chromosome pairs that have been proposed to contain duplicated regions. H: *Hox* regions, M: MHC regions, F: FGFR regions, N: Nadeau (1991), C: Comings (1972), L: Lundin (1993), S: Spring (1997). Below diagonal: dot-matrix summary of TBLASTX search results. Human sequence pairs with significant similarity are plotted at the rank-order positions of the sequences on a physical map of chromosomes.

c/somes involved	# gene pairs*	length spanned on the two chromosomes (in mapped gene units)	gene order conserved?	Additional notes
1 vs. 3	3	107, 63	no	
2 vs. 7	3	445, 86	yes	2 pairs spaced 78 and 75 genes apart
3 vs. 10	4	673, 480	yes	
3 vs. 12	5	361, 388	yes	~ group 5 in Table 3.3
4 vs. 10	4	179, 425	no	
5 vs. 15	3	406, 337	yes	
7 vs. 12	3	486, 64	no	
9 vs. 13	3	130, 285	yes	

Table 4.7. Potentially duplicated regions found on the NCBI Genomes Division dot-matrix plot, identified as described in the text.

* tandem repeats were considered to be part of a single hit in this calculation

4.5.3 Comparison of NCBI and Schuler dataset results

Taking into consideration that from a dataset of 2,317 genes, we found 4,072 matches, it was disappointing to find only 348 matches from a much larger dataset of 13,401 sequences. This can in part be explained by the fact that of the 13,401 sequences, only 2,673 were ≥ 1 kb in length. Also, many of the longer sequences were composed primarily of intergenic sequences and repeat sequences. Moreover, many sequences were ESTs, which may not encode a protein sequence. Out of the 13,401 sequences in the NCBI Genomes Division database, only 46.6% (or around 7,800) of them seem to be protein-coding sequences.

The basic make-up of the two datasets was also different. The Schuler dataset was composed of previously sequenced full-length cDNAs, where in many cases, genes in the same family would all have been dealt with in the same study, or in the same lab, therefore the database was biased towards finding duplicates. The NCBI database, on the other hand, was composed of STSs derived from ESTs or cDNAs, precisely mapped, which should have given a much better estimate of the overall make-up of the human genome, and should be completely unbiased with respect to duplicates. In fact, the bias tends in the opposite direction, because while the general chromosomal location of many genes is known, their precise chromosomal position is not, so that many of the previously sequenced genes (i.e., the duplicates included in the Schuler dataset) were not included in the NCBI dataset, even though it is more recent.

CHAPTER 5 —

COMPUTER SIMULATION OF THE 2R HYPOTHESIS

5.1 AIM

The hypothesis that there have been two rounds of polyploidization in the evolution of the chordate lineage (the 2R hypothesis) has become widely accepted, and is often cited in evolutionary literature. In reality, the current scant amount of mapped sequence information for vertebrate genomes has made this hypothesis difficult to prove or disprove conclusively. In the previous chapter, we compared 13,401 mapped and sequenced human DNA sequences to each other in TBLASTX searches. We were surprised to find no duplicated regions, using a rule where a maximum distance of 50 other genes was allowed between paralogues. This prompted us to carry out computer simulations to test whether, given the current state of knowledge about the human genome, we could detect Ohno-style duplications assuming that they have occurred. We investigated whether it is possible, without having the whole human genome sequence, to estimate the probability of gene deletion as opposed to retention and the number of subsequent genome rearrangements (transpositions, etc.). Genome duplication should result in duplicated chromosomal regions being found throughout the genome. We used computer simulations to model a double genome duplication event, followed by deletions and transpositions.

For simplicity, we assumed that genes were only duplicated via polyploidization. An algorithm to find duplicated chromosomal regions was applied to the simulated genomes. The parameters of the simulation (probability of a gene being deleted, and number of transpositions) were adjusted to establish parameter ranges that could correspond to the observed situation. The simulations were modified to model situations where complete sequence information is not available. Genes were discarded after the simulation so that the number remaining mirrored the real number of mapped and sequenced human genes. This provided an estimate of the minimum fraction of the human genome that must be mapped before any analysis can be informative.

5.2 INTRODUCTION

The incompleteness of genome sequence information for vertebrates has made Ohno's hypothesis difficult to prove or disprove (Skrabanek and Wolfe, 1998). There are three major unknown quantities concerning the hypothesis: the number of rounds of duplication (assumed here to be two; see Section 1.5.1), the fraction of genes retained in duplicate (as opposed to subsequently being deleted) (see Section 1.4), and the number of chromosomal rearrangements following duplication (see Section 2.2.2). The ease with which duplicated chromosomal segments resulting from genome duplication can be identified is expected to depend strongly on these unknown parameters.

Our simulations follow the models proposed by Nadeau and Sankoff (1997), summarized in Figure 5.1. They examined the numbers of human and mouse genes in families arising from genome duplication, and estimated the relative rates of gene loss versus functional divergence from the relative numbers of four-, three-, and two-membered families. On the presupposition that two rounds of genome duplication occurred in the evolution of the chordate lineage, they looked at three possible models of this process. Model 1 assumes that the two rounds occurred almost simultaneously,

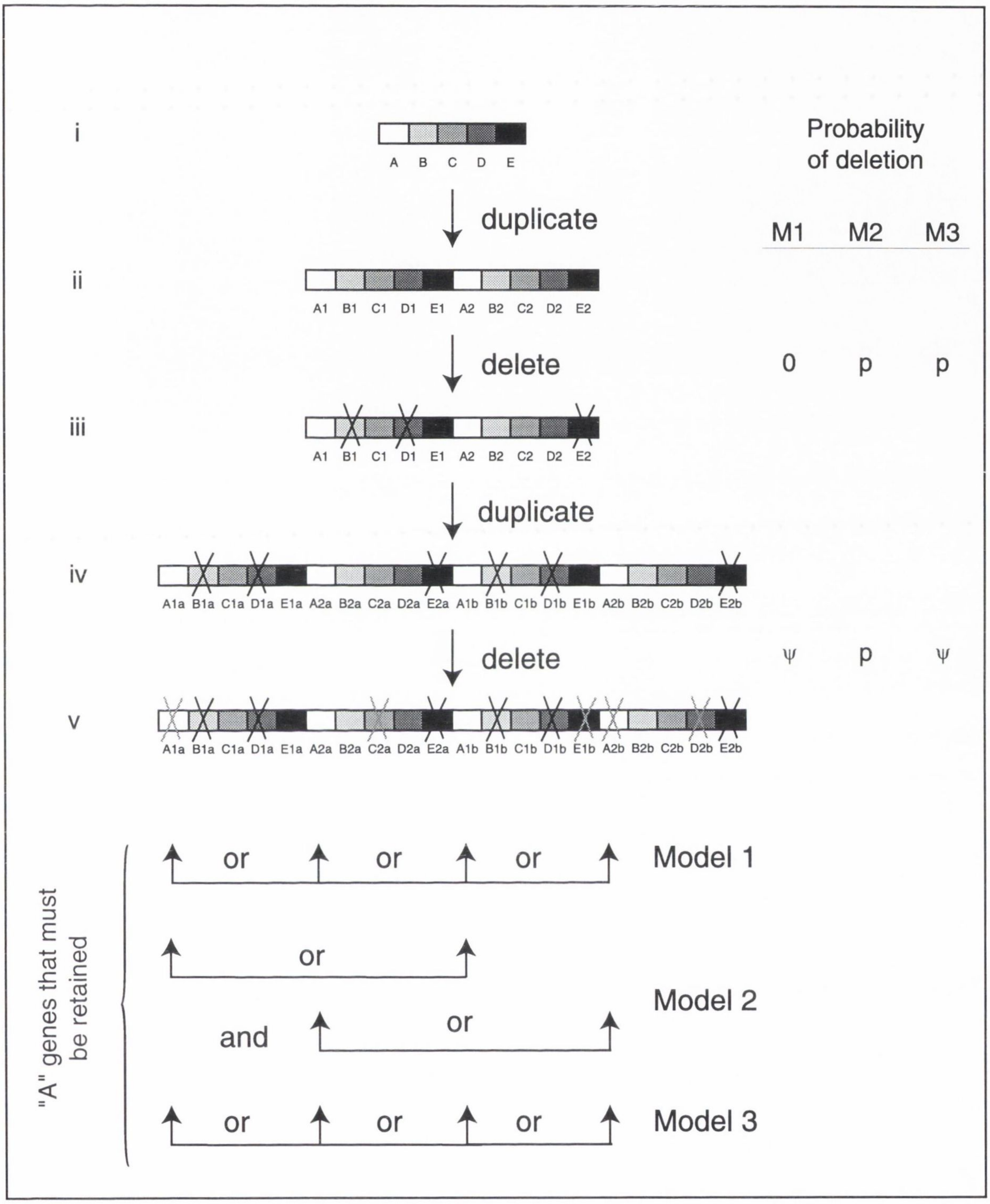

i
A B C D E
Probability of deletion
duplicate
M1 M2 M3
ii
A1 B1 C1 D1 E1 A2 B2 C2 D2 E2
delete
0 p p
iii
A1 B1 C1 D1 E1 A2 B2 C2 D2 E2
duplicate
iv
A1a B1a C1a D1a E1a A2a B2a C2a D2a E2a A1b B1b C1b D1b E1b A2b B2b C2b D2b E2b
delete
ψ p ψ
v
A1a B1a C1a D1a E1a A2a B2a C2a D2a E2a A1b B1b C1b D1b E1b A2b B2b C2b D2b E2b
"A" genes that must be retained
or or or
Model 1
or
and
or
Model 2
or or or
Model 3

Figure 5.1. Schematic representation of the duplication and deletion processes simulated, following the three models of Nadeau and Sankoff (1997). Each box represents a gene, and similarly shaded boxes indicate paralogues. Deleted genes are shown scored through with a cross. Black crosses highlight genes deleted during the first round of gene deletions; grey crosses indicate those deleted during the second round. Each round of genome duplication consists of one duplication and one deletion process. In Model 1, there are no deletions after the first round. The probability of deleting a gene at each round under the three models (M1, M2, M3) is shown on the right.

and not enough time elapsed between them for deletion or functional divergence to take place. In both Models 2 and 3, there is a longer interval between rounds so that the fate of the duplicated genes is partially or wholly established before the second round. Model 2 assumes that, in cases where both genes are retained after the first round, complete divergence takes place before the second duplication event, so that they now both have essential functions. Thus one, but not both, of the genes in each of the new pairs can lose function. In the third model, functional divergence between the pairs is incomplete so that any gene of the possible final four duplicates can still compensate for the function of any other gene in that family. A basic assumption common to all these models is that any gene function present before duplication is essential, so that all the genes in a duplicated family cannot be deleted. Nadeau and Sankoff (1997) discarded Model 1 as improbable, and suggested that, in mammals, some genes underwent complete divergence but the fate of the remainder was not resolved between duplications.

5.3 METHODS

There are between 50,000 and 100,000 genes in the human genome (Fields *et al.*, 1994). In our simulations we aimed to produce a post-duplicative genome containing approximately 80,000 genes, the product of two rounds of genome duplication and subsequent gene deletion. Because the extent of gene deletion was a variable parameter, the number of genes in the starting genome also had to be variable and ranged from 20,000 to 75,000 genes. We started with the same number of genes for all three models at any given deletion probability (*p*, see below). Because Model 1 consistently had the lowest number of genes deleted (because there is effectively only one round of deletion), we occasionally had as few as 60,000 genes left for the final arrays of the other two models, but approximately 80,000 genes left for Model 1. This only makes a difference regarding the number of duplicated regions seen at higher deletion probabilities. All other parameters, including the percentage of the genome in duplicated regions, remain the same regardless of the size of the genome.

For simplicity, the genome was represented as a single chromosome. The following assumptions were made in all the models used: genes are deleted singly and randomly, genome rearrangement is via transposition only, and the breakpoints created by these rearrangements are at random intergenic points. It was also assumed that there are no functional constraints on gene order. After gene deletion and transpositions, the genome was scanned for duplicated chromosomal regions (blocks), hallmarks of genome duplication. We use the term 'block' to mean detectable duplicated chromosomal regions. As the genome becomes rearranged, the smaller duplicated segments (as demarcated by the breakpoints made by transpositions) will become unidentifiable and any block that is identified will be part of a larger segment (Seoighe and Wolfe, 1998). Genes contributing to a block (paralogues) could be no more than 50 genes distant from the next paralogue in the same block. A minimum of three pairs of paralogues was required to form a block. To simulate the incompleteness of the human genome sequence, once the double duplication-deletion-transposition process was complete, genes were randomly discarded so that only a fraction of the genome remained (analogous to the human genes currently mapped and sequenced).

5.3.1 CHROMOSOMAL REARRANGEMENTS

The number of chromosomal rearrangements in vertebrates since a putative genome duplication is also unknown. Rearrangements of a duplicated genome break it up into duplicated conserved segments, in each of which the number and order of paralogous genes is the same. The number of conserved segments increases as the genome is disrupted by new events, and the duplicated segments become shorter over time (Nadeau and Taylor, 1984; Sankoff *et al.*, 1997a). Chromosomal rearrangements between human and mouse have occurred at a rate of approximately one per Myr of evolution (Nadeau, 1991; Copeland *et al.*, 1993; DeBry and Seldin, 1996; Ehrlich *et al.*, 1997; Sankoff *et al.*, 1997b).

Assuming that the last genome duplication occurred about 250 Mya (Lundin, 1993), and that the time between successive genome duplications was of the order of 50-100 Myr (enough time for the fate of some, but not all, gene duplicates to be totally resolved (Nadeau and Sankoff, 1997)), we can expect that there might have been a total of approximately 350 rearrangements since the first of the two postulated genome duplication events. Because the last duplication could in fact have occurred as long ago as 500 Mya (Holland *et al.*, 1994; Sharman and Holland, 1996), in our simulations, we modelled 150 - 525 transpositions after each duplication event (except in Model 1, where all deletions and transpositions were performed after a double duplication event). We used a simple method of genome shuffling, where fragments to be transposed were chosen at random and transposed to random sites in the genome. Transposed fragments were limited to being up to 900 genes in length, which corresponds to approximately half a chromosome.

5.3.2 GENE DELETION

In our simulations, genes were deleted by assigning them a random number between zero and one. We then went through the genome systematically. If the random number allocated to a gene was below a "deletion threshold", p, and other members of that gene family were still present, then it was deleted. To describe the possible fates of a duplicated gene-pair, Nadeau and Sankoff (1997) used a variable, ψ, which is the probability that a gene is lost rather than retained, without taking into account the constraint that not all members of a family can be deleted. Because this constraint is not inherent in ψ, the variable on its own does not have any real biological meaning; only the probabilities of the retention of one or both genes (expressed in terms of ψ) are significant. To be able to compare our simulations directly with Nadeau and Sankoff's predictions, we had to relate our probability of deletion (*p*) to their probability (ψ). To do this, we equated the total numbers of genes being retained. In our model, because we search the array systematically, it follows that after the first round of duplication, from the first half of the resulting genome (A1 through E1 in line iii of Figure 5.1), we will delete a fraction *p* of genes, keeping (1 - *p*). In the second half of the genome, the *p* genes whose paralogues were deleted must now be retained, and of the (1 - *p*) genes that can be deleted, a fraction (1 - *p*) will be kept. Nadeau and Sankoff (1997, bottom of p. 1261) give the probability of retaining two genes. We multiplied this by two (to get the number of duplicate genes retained) and added the number of single genes retained. Thus,

$$(1-p)+p+(1-p)^2 = 2\frac{1-\psi}{1+\psi}+\frac{2\psi}{1+\psi} \qquad \textbf{(Eq. 5.1)}$$

$$\Rightarrow (1-p)^2 = \frac{2}{1+\psi}-1$$

The above relation between ψ and *p* is only valid for a single duplication process producing just two daughter genes. For quartets of genes (a four-membered family), their relation becomes an equality between a quartic and a tertiary equation, where ψ is approximately equal to *p*. In our simulations, we used ψ as an input parameter. Deletions were made after each round of duplication (see Figure 5.1). In twofold redundant gene pairs (both stages of Model 2 and the first stage of Model 3), we used *p* (calculated from ψ using Eqn. 1) as the probability of deletion. For fourfold families (Model 1 and the second stage of Model 3), we assumed that *p* was equal to ψ. That *p* and ψ are equivalent can be seen from Table 5.1, which compares the relative frequencies of four-, three-, and two-membered families predicted by Nadeau and Sankoff (1997) to those observed using our simulations.

	4:3		3:2	
	Predicted	Observed	Predicted	Observed
ψ				
Model 1				
0.1	2.25	2.25	6.00	5.85
0.2	1.00	1.00	2.67	2.67
0.3	0.58	0.58	1.56	1.55
0.4	0.38	0.38	1.00	1.00
0.5	0.25	0.25	0.67	0.67
0.6	0.17	0.17	0.44	0.44
0.7	0.11	0.10	0.29	0.29
0.8	0.06	0.06	0.17	0.17
0.9	0.03	0.02	0.07	0.08
Model 2				
0.1	2.25	2.25	1.38	1.38
0.2	1.00	1.00	1.00	1.00
0.3	0.58	0.58	0.73	0.73
0.4	0.38	0.38	0.55	0.55
0.5	0.25	0.25	0.40	0.40
0.6	0.17	0.17	0.29	0.28
0.7	0.11	0.11	0.19	0.19
0.8	0.06	0.06	0.12	0.12
0.9	0.03	0.03	0.05	0.05
Model 3				
0.1	2.25	2.24	1.27	1.27
0.2	1.00	1.00	0.84	0.87
0.3	0.58	0.58	0.56	0.6
0.4	0.38	0.38	0.38	0.42
0.5	0.25	0.25	0.25	0.29
0.6	0.17	0.17	0.15	0.18
0.7	0.11	0.11	0.08	0.11
0.8	0.06	0.07	0.04	0.05
0.9	0.03	0.00	0.01	0.02

Table 5.1. Predicted relative frequencies (from Nadeau and Sankoff, 1997, p. 1262) of four-membered:three-membered families, and three-membered:two-membered families, at varying values of the deletion probability,ψ, compared with simulated frequencies (mean results from 1000 simulations).

5.4 RESULTS

The two parameters we varied were the number of transpositions and the deletion probability, ψ. As the number of transpositions increases, the underlying number of segments increases but their mean length decreases (Nadeau and Taylor, 1984). From a practical point of view, it is more interesting to calculate the total fraction of the genome that is assigned to duplicated regions, because this gives an indication of whether we might expect to be able to discover chromosomal regions supporting the 2R hypothesis. The percentage of the genome covered by duplicated blocks is calculated by adding all block lengths together (where any overlap is counted only once), and dividing by the total genome length for that deletion probability. Included in both of these numbers are the deleted genes, which although they can no longer be seen, are still likely to contribute to the length of the genome (as pseudogenes or "junk" DNA). The number of transpositions does not make a difference to the percentage of the genome in blocks for a fully sequenced genome because each transposition only breaks the existing blocks up into smaller blocks, and very few blocks will be lost, unless they are very small.

Figure 5.2A shows the percentage of the genome covered by duplicated blocks found at varying numbers of translocations, and varying ψ deletion probabilities, for each of the three models described in Nadeau and Sankoff (1997), for a completely sequenced genome containing 80,000 genes. Models 1 and 2 show very similar percentages of the genome covered by blocks. In Model 3, families can be deleted down to one member only, which cannot then form part of a block. The higher the deletion probability, the more likely it is that this will happen. This model, which allows the most genes to be deleted, accordingly has the fewest blocks at the higher deletion ranges, and shows a sharp decrease in the percentage of the genome in blocks compared to the other two models. It can be noted that the number of transpositions performed does not dramatically alter the percentage of the genome in blocks for any given model. The

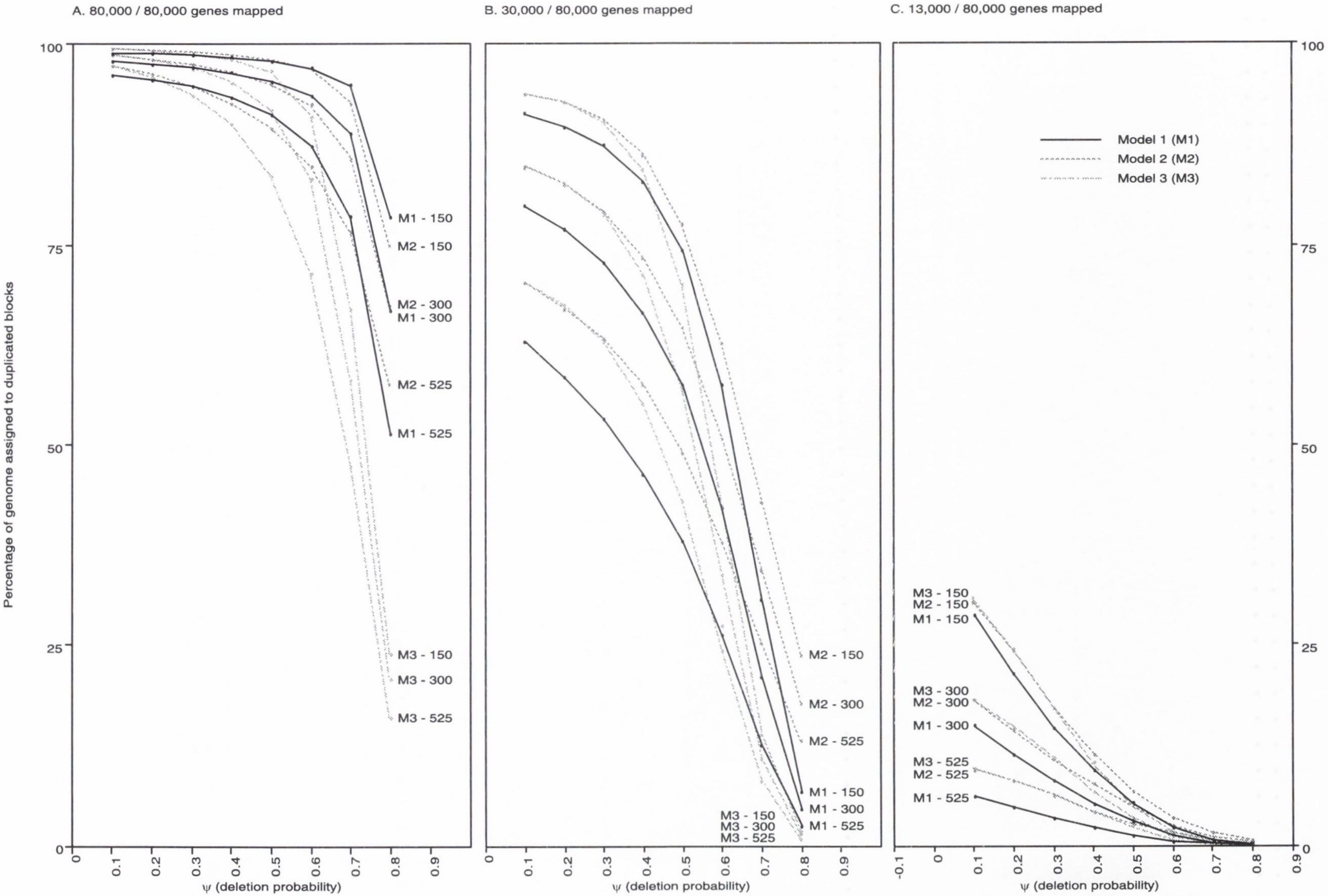


Figure 5.2. Percentage of the genome assigned to duplicated blocks at varying numbers of transpositions and deletion probabilities (ψ) for (A) a fully mapped and sequenced genome of 80,000 genes, (B) a genome where only 30,000 genes out of a possible 80,000 have been identified, and (C) a genome where 13,000 genes out of 80,000 have been mapped and sequenced. For each of the three models (M1, M2, M3), only the results for 150, 300 and 525 chromosomal transpositions after each genome duplication are shown.

number of blocks found, however, does depend on the number of transpositions, increasing as the number of transpositions performed increases (data not shown).

As ψ, the probability of gene deletion, increases, we expect the density of paralogues within the genome to decrease quite sharply. The sparser the genome becomes, the less likely genes are to form part of a block. Because a block requires at least three paralogues to be less than 50 genes apart, as more genes are deleted, the number of blocks decreases. Contrary to intuition, however, the mean length of the blocks increases (data not shown). This is because as more duplicates are deleted, if the allowed distance between paralogues is quite large, the smaller blocks are almost completely eliminated, while the larger blocks will remain, with some shearing at the edges. As the mean length of the blocks increases, we also find that there are fewer duplicated genes within them. Because there are fewer blocks found, the percentage genome in blocks also decreases. It should, however, be noted that even at high deletion probability ranges, 25-50% of the genome is still found in blocks (Figure 5.2A), which suggests that with a fully sequenced genome, it should be very easy to identify a large number of duplicated regions within the genome if the 2R hypothesis is correct.

Over 30,000 genes, corresponding to just under 40% of the estimated number of genes present in the human genome, have now been mapped and at least partially sequenced (Deloukas *et al.*, 1998). Figure 5.2B shows the percentage genome covered by duplicated blocks for 30,000 genes out of an estimated 80,000. Here again, at low probabilities of gene deletion, the percentage genome in blocks is still very high, about 95%. The blocks that were evident in the fully sequenced genome are shortened, and the smaller blocks are eliminated entirely, because there are now gaps in the genome where genes have not yet been mapped and sequenced. As both the deletion probability and the number of transpositions increase, blocks are broken up and eliminated. In this case, unlike the case for the fully sequenced genome, the three models are more similar to each other at any given number of transpositions than any model is to itself at a different number of transpositions, except at very high deletion probabilities. Here again, Model 3 loses the most genes. Model 2 tends to keep more genes because after the second duplication event, paralogues from the previous duplication event can no longer compensate for each other. At high gene deletion probabilities, the percentage of the genome in blocks is quite low (5 - 25%). The ease with which duplicated blocks are seen with this fraction of mapped and sequenced genes depends greatly on the probability of gene deletion.

When we analysed the 13,000 randomly mapped genes from the NCBI database, we used strict criteria to define paralogous relationships, and we observed no blocks (Skrabanek and Wolfe, 1998). Although some duplicated regions have been found in the data currently available, such as the *Hox* clusters and their neighbours, these genes have not been sampled at random. To see whether our results were consistent with expectation, we repeated the simulation, this time randomly deleting all but 13,000 elements after the double duplication-deletion-rearrangement process. Figure 5.2C shows that with this number of genes, we can expect to see anything from 0-30% of the genome in blocks if the genes identified are randomly distributed within the genome. Also, interestingly, we can expect to find 0 - 300 duplicated blocks (Table 5.2). A finding of no blocks, therefore, corresponds to a deletion probability of 0.8 or higher (i.e., ≤ 11% of genes retained in duplicate). At low deletion probabilities, there are as many as 300 duplicated blocks, a number that decreases with the number of transpositions made. Because blocks can minimally consist of at least three genes spread over at most 100 genes, any increase in transpositions will break up these minimal blocks, and they will be lost. Likewise, as the deletion probability increases, the number of duplicated genes with which blocks can be formed decreases, and so too does the number of duplicated blocks.

	Number of transpositions					
	150	225	300	375	450	525
ψ	*Model 1*					
0.1	314 ± 10	259 ± 13	209 ± 14	167 ± 13	139 ± 12	114 ± 8
0.2	248 ± 14	202 ± 14	162 ± 11	128 ± 10	108 ± 8	87 ± 8
0.3	182 ± 14	146 ± 12	121 ± 9	99 ± 10	75 ± 8	66 ± 7
0.4	127 ± 10	101 ± 7	81 ± 8	66 ± 9	56 ± 7	45 ± 9
0.5	77 ± 8	64 ± 8	50 ± 6	42 ± 6	33 ± 5	28 ± 6
0.6	37 ± 7	31 ± 5	24 ± 5	20 ± 5	17 ± 4	12 ± 4
0.7	12 ± 5	11 ± 3	8 ± 3	7 ± 2	6 ± 2	5 ± 2
0.8	2 ± 1	1 ± 1	2 ± 1	1 ± 1	1 ± 1	1 ± 1
	Model 2					
0.1	309 ± 11	260 ± 11	215 ± 16	183 ± 11	157 ± 10	134 ± 8
0.2	257 ± 16	209 ± 13	172 ± 11	151 ± 13	129 ± 7	113 ± 10
0.3	191 ± 12	167 ± 10	134 ± 9	116 ± 10	103 ± 7	91 ± 8
0.4	140 ± 10	120 ± 8	103 ± 8	87 ± 7	78 ± 6	66 ± 6
0.5	92 ± 8	75 ± 9	69 ± 9	57 ± 7	52 ± 7	45 ± 7
0.6	54 ± 7	48 ± 6	40 ± 4	36 ± 8	30 ± 4	28 ± 5
0.7	27 ± 3	24 ± 5	21 ± 5	19 ± 4	17 ± 5	14 ± 3
0.8	13 ± 4	11 ± 3	8 ± 3	8 ± 3	8 ± 3	6 ± 3
	Model 3					
0.1	317 ± 12	258 ± 13	215 ± 11	184 ± 10	158 ± 10	132 ± 9
0.2	258 ± 14	212 ± 12	179 ± 11	152 ± 11	132 ± 9	114 ± 10
0.3	188 ± 13	161 ± 11	136 ± 8	115 ± 10	102 ± 9	88 ± 8
0.4	124 ± 9	103 ± 12	88 ± 9	79 ± 8	67 ± 9	61 ± 6
0.5	63 ± 8	54 ± 7	47 ± 5	43 ± 6	39 ± 6	33 ± 5
0.6	23 ± 4	20 ± 4	18 ± 4	15 ± 4	14 ± 4	11 ± 3
0.7	5 ± 2	4 ± 2	4 ± 2	3 ± 2	3 ± 2	2 ± 1
0.8	0 ± 1	0 ± 1	1 ± 1	0 ± 1	0 ± 1	0 ± 1

Table 5.2. Numbers of duplicated blocks in simulated genomes for 13,000 mapped genes. The average number of duplicated blocks (± 1 SD) found after 100 simulation runs is shown for each of the three models described by Nadeau and Sankoff (1997). The genome contained approximately 80,000 genes, of which 13,000 were mapped and sequenced. The parameters varied are the number of transpositions performed after each duplication event, and the probability of gene deletion (ψ).

5.5 DISCUSSION

As our earlier analysis of a 13,000-gene human data set yielded no blocks, we can draw one of two conclusions: either the probability of duplicated gene deletion is very high ($\psi \geq 0.8$; which corresponds to $\leq$ 11% of genes being retained in duplicate) or the 2R hypothesis as modelled here is incorrect (cf. Hughes, 1999). The first conclusion is inconsistent with recent literature which suggests that ψ should be around 0.3 (50% of genes retained in duplicate, see Section 1.4; Walsh, 1995; Nadeau and Sankoff, 1997; Force *et al.*, 1999). Our simulations, however, are based on several assumptions including that only two genome duplications occurred in the evolution of the human genome and that paralogues are easy to identify and result from genome duplication only. We also ignored the possible role of tandem gene duplications. If the redundancy generated by tandem duplication means that one or more of the genes duplicated by polyploidy are more likely to be lost, then there will be fewer duplicated blocks found in the human genome than we predict. Furthermore, we assumed the probability of gene deletion to be the same after both genome duplication events. If the recent DDC model proposed by Force *et al.* (1999) is realistic, and the likelihood of a (subfunctionalized) gene being retained decreases with each successive duplication event, then the probability of deletion should increase after each duplication event. This also would imply that there would be fewer genes representative of any one gene family present in the genome, and thus also a smaller number of blocks. It is also worth noting that from the method described here, it is almost impossible to predict the number of transpositions that have occurred since the first of the two proposed polyploidization events.

TBLASTX searches with the 13,000-gene set took 17 days processor time on our DEC Alpha station 500 333 Mhz laboratory computer and it was currently impractical for us to do an equivalent all-against-all search with the 30,000-gene set of Deloukas *et al.* (1998). From Figure 5.2B however, we can estimate that less than 25% of the genome will be found to lie in duplicated blocks if 2R hypothesis is correct and if $\psi \geq 0.8$ (as appears from Table 5.2). If the 2R hypothesis is incorrect, it is still unclear how much of the genome would be placed into blocks artefactually: this depends on the distribution of family sizes in the human proteome, which is currently unknown and was not simulated here.

It may be plausible to suggest replacing one of the postulated rounds of polyploidy with a series of single gene or regional duplications, as first proposed by Sharman and Holland (1996). If polyploidization occurred first, followed by a large number of smaller duplications with subsequent genome rearrangements, we could explain not only the fact that we find fewer duplicated blocks than expected, but also the phylogenetic tree topologies that Hughes (1999) has put forward as an argument against the 2R hypothesis.

Our results indicate that if the hypothesis is correct, then contrary to recent proposals, the frequency at which duplicated genes were subsequently retained in the genome was at most only approximately 10%. We also suggest that, if the probability of retention is so low, the amount of sequence information currently available (30,000 genes; Deloukas *et al.*, 1998) will still not be enough to draw any conclusions about the veracity of the 2R hypothesis.

Many of the genes present in the NCBI dataset that we used are ESTs, which may not encode a protein sequence. The true count of mapped protein-coding sequences may actually be as low as 7,800. If this is actually the case, then to get a more accurate estimate of what the deletion probability is, which will probably be higher and more in keeping with recent estimates, the simulations should be redone randomly deleting all but 7,800 elements after the double duplication-deletion-rearrangement process.

CHAPTER 6 —

GENERAL DISCUSSION

In 1972, Comings said of Ohno's hypothesis: "Implications of the theory are so wide-ranging that it is worth stating the hypothesis to encourage its being tested." In fact, the hypothesis has never been rigorously tested. Paralogous regions in the genome can be derived in at least four ways (Smith *et al.*, 1999). They can be a product of genome duplication, duplication of a subset of the genome (such as chromosomal duplication), they can be formed from pre-duplicated genes which are clustered together for functional reasons or they can result from a coincidental clustering. Nadeau (1991) demonstrated that up to 50% of duplicated regions which are marked by two or three genes may in fact be coincidental. Construction of molecular phylogenies may give some indication as to which of the above mechanisms may have been involved in the evolution of any particular paralogous segment.

6.1 PHYLOGENY-BASED APPROACHES TO TESTING THE GENOME DUPLICATION HYPOTHESIS

The difficulty in interpreting phylogenetic data for complex gene families is illustrated by two recent studies on a single data set. Baker (1997) analysed the steroid hormone receptor family, which is a subset of the nuclear receptor (NR) superfamily analysed independently by Escriva *et al.* (1997). Both groups concluded that they had identified two rounds of gene duplication in their trees. The two rounds identified by Baker both occurred approximately around the origin of vertebrates and gave rise to a 2+2 topology for the androgen, progesterone, glucocorticoid and mineralocorticoid receptors. In contrast, Escriva *et al.* regarded these events as a single "wave" of duplications and they proposed that there was also a much earlier "wave," prior to the Hydra divergence over 700 Mya, which created the six NR subfamilies (of which the steroid hormone receptors are one). Despite these differences both groups interpreted their results as supporting Ohno's hypothesis, which illustrates the ease with which the hypothesis can be interpreted to accommodate almost any tree for a multigene family.

Two other examples show how easy it is to make the data fit any preconceived opinion regarding the 2R hypothesis. Patton *et al.* (1998), having assumed that paralogous regions on HSA 11 and 12 arose by genome duplication early in vertebrate evolution, constructed phylogenetic trees for two families of closely linked genes (the aromatic amino acid hydroxylase genes (AAAH) and insulin-type genes (IGF)) on these chromosomes (TPH and TH are located on HSA 11, PAH on HSA 12; INS and IGF2 are found on HSA 11, IGF1 on HSA 12). They found that the gene duplication yielding the AAAH genes occurred at a vastly different time than the duplication giving rise to the IGF genes (the AAAH genes duplicated early in metazoan evolution, predating the divergence of nematodes, arthropods and chordates, whereas the IGF genes duplicated around the origin of vertebrates). They concluded that, far from disproving the 2R hypothesis, these results indicated that "paralogous regions can have a complex history." They proposed a tandem repeat and differential silencing model to explain the pattern of genes, and the discrepancy in their dates of divergence, seen in the genome today. This proposal is similar to that of Smith *et al.* (1999). A corollary of this work is that caution must be exercised in extrapolating gene duplication dates from one gene family to its neighbouring gene family, as was involved in the Bailey *et al.* (1997) study.

Hughes (1999), on the other hand, was skeptical about the veracity of the 2R hypothesis, and cautions that there is no clear-cut evidence at the moment that supports it. Phylogenetic analysis of nine protein families important in development yielded many different dates of divergence, ranging from prior to the divergence of deuterosomes and proteosomes, to after vertebrate origins, and also a number of topologies of an (A)(BCD) type. He concluded that these results provided strong evidence against the 2R hypothesis, but conceded that all the phylogenies were consistent with a single tetraploidization event occurring around the origin of vertebrates. However, the genes under investigation were all isolated genes (i.e., they are not physically linked). If the tandem repeat and differential silencing model proposed by Patton *et al.* (1998) is important throughout the genome, then possibly some or all of the discrepancies illustrated by Hughes' analysis could be explained. Evidence that this mode of evolution may be widespread throughout the genome comes from a recent study by Loftus *et al.* (1999) who showed that 12 Mb of DNA sequence from the centromeric region of HSA 16 contains a large number of highly homologous, recently duplicated tracts of sequence.

Although using a phylogenetic approach would appear to be a good way to address the examination of the 2R hypothesis, several issues make analysis of the data difficult. For example, gene conversion between duplicates makes them appear to have been duplicated more recently than they actually were. The tandem repeat and differential silencing model means that duplicates appear to have duplicated much earlier than they actually did (the time of tandem duplication versus the time of segmental duplication). These factors imply that it may not be valid to use the convergence of gene duplication dates within paralogous segments as a test of the tetraploidy hypothesis.

6.2 MAP-BASED APPROACHES

In our attempt to examine the 2R hypothesis, we used a map-based method rather than phylogenetic or tree-based methods. We performed systematic searches, both comparative between the human and mouse genomes, and within the human genome itself, on a number of different datasets, looking for duplicated regions. While we did find some putative duplicated segments, our overall conclusion was that there was not enough data available to draw any real conclusions about the duplication history of the human genome. However, the interspecies comparative approach would seem to be more effective in elucidating such a history.

Comparative analysis has been shown to be useful in the determination of the evolutionary history of genomes. For example, *Saccharomyces cerevisiae* is thought to have undergone a genome-wide duplication event round 100 Mya (Wolfe and Shields, 1997). Because only 8% of duplicated genes have been retained and there have been many rearrangements (mainly reciprocal translocations) after this event (Seoighe and Wolfe, 1998), many of the linkage groups in *S. cerevisiae* have been broken up, and paralogous regions contain many genes which are no longer duplicated. From comparisons of the paralogous regions in *S. cerevisiae* with *Kluyveromyces lactis* (an unduplicated yeast), it can be seen that some regions of the *K. lactis* genome have gene orders that correspond to an amalgamation of genes from both copies of the duplicated regions in *S. cerevisiae* (Wolfe and Shields, 1997; Keogh *et al.*, 1998). This not only lends credence to the hypothesis that genome duplication did occur in the *S. cerevisiae* genome, but also allows the reconstruction of the ancestral linkage group.

Several species have been proposed as good comparative models for the human genome, including mouse, *Fugu*, *Drosophila* and amphioxus. There are problems associated with most of these species. Brenner *et al.* (1993) were the first to propose that *Fugu* might be an excellent model vertebrate for use in the dissection of the human genome because it appears to have the same total number of genes as the human genome. Several regions of conserved synteny (but not necessarily gene order) have been described between the *Fugu* and human genomes (Aparicio *et al.*, 1997; Armes *et al.*, 1997; Brunner *et al.*, 1999; Gellner and Brenner, 1999; Gilley and Fried, 1999; Kehrer-Sawatzki *et al.*, 1999; Reboul *et al.*, 1999; McLysaght *et al.*, submitted), although it is not known how large the syntenic regions are or how well gene order is conserved within them. Conserved synteny between the two genomes appears to be of the order of 40% (McLysaght *et al.*, submitted). Despite this, *Fugu* may not be the best species to use in comparative analysis with the human genome, for at least three reasons. Firstly, *Fugu* genes must show sufficient similarity to their human orthologues to be identifiable. The human and mouse lineages diverged from each other about 80 Mya, and show an 85% amino acid sequence identity between orthologues (Makałowski *et al.*, 1996). The human and *Fugu* lineages diverged from each around 400 Mya, so their sequences may have diverged quite significantly, which will complicate the issue of distinguishing paralogues from orthologues. Secondly, the genome of *Fugu* may have undergone a genome duplication event after it diverged from the human lineage. The reason that the *Fugu* genome seems to contain approximately the same number of genes as the human genome has been explained by a widespread loss of the genes duplicated after this second duplication event (Vogel, 1998), again making paralogues harder to distinguish from orthologues. Thirdly, if the process of diploidization entails a high rate of rearrangement immediately subsequent to the duplication event, then the *Fugu* genome may be too rearranged in comparison to the human genome to be informative about ancestral gene orders. McLysaght *et al.* have estimated the number of rearrangements between the human and *Fugu* genomes to be as high as 8000 - 32000.

The *Drosophila* genome has also been used in initial comparative analysis with the human genome, because the *Drosophila* genome, which contains an estimated 8,600 - 17,000 genes (Ashburner *et al.*, 1999) (an average of 12,000, which is approximately 5.8-fold less than is found in vertebrate genomes (Miklos and Rubin, 1996)), may be indicative of the basic chordate genome. However, *Drosophila* has a very small genome (170 Mb; Rasch *et al.*, 1971; Rudkin, 1972; Kavenoff and Zimm, 1973), and it is possible that *Drosophila* has an unusually small number of genes, which would make it a poor comparison for reconstructing possible events of genome duplications early in vertebrate history (Miklos and Rubin, 1996).

Amphioxus, on the other hand, may be a good species to use in comparisons with the human genome, because it is the sister group to the vertebrates and is thought to have branched off from the chordate lineage just before the putative tetraploidization event (Patton *et al.*, 1998).

6.3 SIMULATIONS

The evolution of mammalian genomes is thought to include at least two whole genome duplications of an ancestral genome (Holland *et al.*, 1994), as well as duplication of subchromosomal segments together with extensive gene duplication that has given rise to many large multigene families (Lundin, 1993). Gene loss may occur frequently (as seen in yeast; Seoighe and Wolfe, 1998) and multiple small inversions will disrupt paralogous regions (Seoighe *et al.*, in press). Because of these factors, it becomes difficult to actually prove Ohno's hypothesis by finding paralogous regions in the genome, because it is hard to devise ways to discriminate between this hypothesis and other alternatives. Even regions that may have been duplicated around the time of the origin of vertebrates and/or eukaryotes do not prove that genome duplications occurred at those times. Computer simulations of the evolution of the human genome may be a superior way of proving the hypothesis or at least gaining some insight into how the human genome evolved. By predicting how many paralogous regions should be identifiable under any given model, and comparing that number with the number actually found, we can determine how accurate that model is. In our simulations, which were based on a simplistic model of a double genome duplication event only, with no additional chromosomal or tandem duplications, we concluded that either the model we had used was incorrect, or the probability that a duplicated gene is lost is very high, contrary to recent findings. Of these, the more likely explanation is that a simple model of two rounds of genome duplication is wrong. Simulations can be performed which include other duplication events. The most likely outcome of this type of work would be that there are many different ways to get to the situation we see today.

6.4 FUTURE PROSPECTS

While the proposition that genome duplication has contributed greatly to the evolution of chordates / vertebrates is a simple and wide reaching one, it would appear to be only too easy to make the data fit the hypothesis. Because of the paucity of available detailed mapping and sequence data for vertebrates, it is difficult to reach any definite conclusion about the veracity of Ohno's original 1970 hypothesis, and any attempts to do so at present without rigorous statistical analysis should be treated with caution.

The human genome project aims to have the complete human genome sequenced by 2001 (Marshall, 1998; Wadman, 1998), at which time there should be enough data to reach a definitive conclusion as to whether there have been two genome duplications during the evolution of chordates. We can investigate this either by looking at the human genome itself (a study which should be helped by the availability of additional information, such as gene orientation), or by comparing the human genome to the genomes of other chordates. The complete sequencing of chordates such as amphioxus, *Fugu* and mouse should make the task of elucidating the history of the human genome dramatically easier. From studies of *Fugu* (or preferably from the zebrafish genome), we may be able to learn what a genome duplication in a vertebrate looks like. General comparisons between the human and mouse genomes could be informative about the evolution of gene order (i.e., inversions and translocations) that is necessary to model the 2R hypothesis. However, because of the large numbers of rearrangements that have occurred in *Fugu*, and due to the fact that the rodent lineage is more rearranged than that of human (although many small inversions are likely to have occurred in both lineages), it is more feasible to use the amphioxus genome to determine the history of the human genome, and then to use the ancestral configuration of the human genome to elucidate the genomic evolution of the *Fugu* and mouse genomes.

The work presented here shows that the amount of both sequencing and map data currently available is not sufficient to be able to draw any conclusions about how many genome duplications occurred in the evolution of the human genome (if any), nor when they may have occurred. It is also indicated that the discovery of paralogous regions within the human genome is unlikely to aid in the resolution of this hypothesis. It is suggested that a combination of interspecies comparisons and simulations may be the best way of determining the history of the evolution of the human genome.

REFERENCES

Acampora, D., D'Esposito, M., Faiella, A., Pannese, M., Migliaccio, E., Morelli, F., Stornaiuolo, A., Nigro, V., Simeone, A. and Boncinelli, E. (1989) The human HOX gene family. *Nucl. Acids Res.* **17:** 10385-10402.

Adams, M. D., Kerlavage, A. R., Fleischmann, R. D., Fuldner, R. A., Bult, C. J., Lee, N. H., Kirkness, E. F., Weinstock, K. G., Gocayne, J. D., White, O. *et al.* (1995) Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* **377:** 3-174.

Adélaïde, J., Chaffanet, M., Imbert, A., Allione, F., Geneix, J., Popovich, C., van Alewijk, D., Trapman, J., Zeillinger, R., Borresen-Dale, A. L., Liderman, R., Birnbaum, D. and Pébusque, M. J. (1998) Chromosome region 8p11-p21: refined mapping and molecular alterations in breast cancer. *Genes Chromosom. Cancer* **22:** 186-199.

Ahn, S., Anderson, J. A., Sorrells, M. E. and Tanksley, S. D. (1993) Homeologous relationships of rice, wheat and maize chromsomes. *Mol. Gen. Genet.* **241:** 483-490.

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

Allendorf, F. W. (1978) Protein polymorphism and the rate of loss of duplicate gene expression. *Nature* **272:** 76-78.

Allendorf, F. W. and Thorgaard, G. H. (1984) Tetraploidy and the evolution of salmonid fishes, pp. 1-53 in *Evolutionary genetics of fishes*, edited by Turner, B. J. Plenum, New York.

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

Anderson, R. P. and Roth, J. R. (1977) Tandem genetic duplications in phage and bacteria. *Ann. Rev. Microbiol.* **31:** 473-505.

Andersson, L., Archibald, A., Ashburner, M., Audun, S., Barendse, W., Bitgood, J., Bottema, C., Broad, T., Brown, S. and Burt, D. (1997) The first international workshop on comparative genome organisation. *Mamm. Genome* **7:** 717-734.

Anonymous (1996) Mouse chromosome committee reports. *Mamm. Genome* **6**.

Ansari-Lari, M. A., Oeltjen, J. C., Schwartz, S., Zhang, Z., Muzny, D. M., Lu, J., Gorrell, J. H., Chinault, A. C., Belmont, J. W., Miller, W. and Gibbs, R. A. (1998) Comparative sequence analysis of a gene-rich cluster at human chromosome 12p13 and its syntenic region in mouse chromosome 6. *Genome Res.* **8:** 29-40.

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

Aparicio, S., Kelvin, H., 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:** 79-83.

Armes, N., Gilley, J. and Fried, M. (1997) The comparative genomic structure and sequence of the surfeit gene homologs in the pufferfish Fugu rubripes and their association with CpG-rich islands. *Genome Res.* **7:** 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., Harbey, 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., Kimmel, B., Celniker, S. and Rubin, G. M. (1999) An exploration of the sequence of a 2.9-Mb region of the genome of *Drosophila melanogaster*: the *Adh* region. *Genetics* **153:** 179-219.

Atchley, W. R., Fitch, W. M. and Bronner-Fraser, M. (1994) Molecular evolution of the MyoD family of transcription factors. *Proc. Natl. Acad. Sci. USA* **91**.

Atkin, N. B. and Ohno, S. (1967) DNA values of four primitive chordates. *Chromosoma* **23:** 10-13.

Bailey, G. S., Poulter, R. T. M. and Stockwell, P. A. (1978) Gene duplication in tetraploid fish: Model for gene silencing at unlinked duplicated loci. *Proc. Natl. Acad. Sci. USA* **75:** 5575-5579.

Bailey, W. J., Kim, J., Wagner, G. P. and Ruddle, F. H. (1997) Phylogenetic reconstruction of vertebrate Hox cluster duplications. *Mol. Biol. Evol.* **14:** 843-853.

Baker, M. E. (1997) Steroid receptor phylogeny and vertebrate origins. *Mol. Cell. Endocrin.* **135:** 101-107.

Baxendale, S., Abdulla, S., Elgar, G., Buck, D., Berks, M., Micklem, G., Durbin, R., Bates, G., Brenner, S., Beck, S. and Lehrach, H. (1995) Comparative sequence analysis of the human and pufferfish Huntington's disease genes. *Nat. Genet.* **10:** 67-75.

Beck, S., Abdulla, S., Alderton, R. P., Glynne, R. J., Gut, I. G., Hosking, L. K., Jackson, A., Kelly, A., Newell, W. R., Sanseau, P., Radley, E., Thorpe, K. L. and Trowsdale, J. (1996) Evolutionary dynamics of non-coding sequences within the class II region of the human MHC. *J. Mol. Biol.* **255:** 1-13.

Beeman, R. W. (1987) A homeotic gene cluster in the red flour beetle. *Nature* **327:** 247-249.

Belleville, S., Beauchemin, M., Tremblay, M., Noiseux, N. and Savard, P. (1992) Homeobox-containing genes in the newt are organized in clusters similar to other vertebrates. *Gene* **114:** 179-186.

Bentley, K. L., Bradshaw, M. S. and Ruddle, F. H. (1995) Human HOXB cluster and the nerve growth factor receptor gene: comparison with an orthologous domain in mouse. *Genomics* **30:** 18-24.

Bingulac-Popovic, J., Figueroa, F., Sato, A., Talbot, W. S., Johnson, S. L., Gates, M., Postlethwait, J. H. and Klein, J. (1997) Mapping of the *Mhc* class I and class II regions to different linkage groups in the zebrafish, *Danio rerio*. *Immunogenetics* **46:** 129-134.

Bird, A. P. (1995) Gene number, noise reduction and biological complexity. *Trends Genet.* **11:** 94-100.

Bisbee, C. A., Baker, M. A., Wilson, A. C., Hadji-Azami, I. and Fischberg, M. (1977) Albumin phylogeny for clawed frogs (*Xenopus*). *Science* **195:** 785-787.

Blair, H. J., Reed, V., Laval, S. H. and Boyd, Y. (1994) New insights into the man-mouse comparative map of the X chromosome. *Genomics* **19:** 215-220.

Boer, P. H., Adra, C. N., Lau, Y.-F. and McBurney, M. W. (1987) The testis-specific phosphoglycerate kinase gene pgk-2 is a recruited retroposon. *Mol. Cell Biol.* **7:** 3107-3112.

Boncinelli, E., Simeone, A., Acampora, D. and Mavilio, F. (1991) *HOX* gene activation by retinoic acid. *Trends Genet.* **7:** 329-334.

Botta, A., Lindsay, E. A., Jurecic, V. and Baldini, A. (1997) Comparative mapping of the DiGeorge syndrome region in mouse shows inconsistent gene order and differential degree of gene conservation. *Mamm. Genome* **8:** 890-895.

Boutin, S., Young, N., Olson, T., Yu, Z.-H. and Shoemaker, R. (1995) Genome conservation among three legume genera detected with DNA markers. *Genome* **38:** 928-937.

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

Brenner, S. E., Hubbard, T., Murzin, A. and Chothia, C. (1995) Gene duplications in *H. influenzae*. *Nature* **378:** 140.

Brooke, N. M., Garcia-Fernàndez, J. and Holland, P. W. H. (1998) The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster. *Nature* **392:** 920-922.

Brunner, B., Todt, T., Lenzner, S., Stout, K., Schulz, U., Ropers, H.-H. and Kalscheuer, V. M. (1999) Genomic structure and comparative analysis of nine *Fugu* genes: conservation of synteny with human chromosome Xp22.2-p22.1. *Genome Res.* **9:** 437-338.

Bulmer, M., Wolfe, K. H. and Sharp, P. M. (1991) Synonymous nucleotide substitution rates in mammalian genes: implications for the molecular clock and the relationship of mammalian orders. *Proc. Natl. Acad. Sci. USA* **75:** 5575-5579.

Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton, R. A., Gocayne, J. D. *et al.* (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273:** 1058-1073.

Bürglin, T. R. and Ruvkun, G. (1993) The *Caenorhabditis elegans* homeobox gene cluster. *Curr. Opin. Genet. Dev.* **3:** 615-620.

Burland, V., Daniels, D. L., Plunkett, G. D. and Blattner, F. R. (1993) Genome sequencing on both strands: the Janus strategy. *Nucl. Acids Res.* **21:** 3385-3390.

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

Campbell, R. D. and Trowsdale, J. (1993) Map of the human MHC. *Immun. Today* **14:** 349-352.

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

Chan, S. J., Cao, Q.-P. and Steiner, D. F. (1990) Evolution of the insulin superfamily: cloning of a hybrid insulin / insulin-like growth factor cDNA from amphioxus. *Proc. Natl. Acad. Sci. USA* **87:** 9319-9323.

Clark, A. G. (1994) Invasion and maintenance of a gene duplication. *Proc. Natl. Acad. Sci. USA* **91:** 2950-2954.

Clark, M. S. (1999) Comparative genomics: the key to understanding the Human Genome Project. *BioEssays* **21:** 121-130.

Claverie, J.-M. (1994) Large-scale sequence analysis, pp. in *Automated DNA sequencing and analysis*, edited by Adams, M. D., Fields, C. and Venter, J. C. Academic Press Ltd., New York.

Claverie, J. M. and States, D. J. (1993) Information enhancement methods for large-scale sequence analysis. *Computers Chem.* **17:** 191-201.

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

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

Cooke, J. (1998) Reply to Tony Gibson and Jürg Spring. *Trends Genet.* **14:** 49-50.

Cooke, J., Nowak, M. A., Boerlijst, M. C. and Maynard-Smith, J. (1997) Evolutionary origins and maintenance of redundant gene expression during metazoan development. *Trends Genet.* **13:** 360-364.

Copeland, N. G., Jenkins, N. A., Gilbert, D. J., Eppig, J. T., Maltais, L. J., Miller, J. C., Dietrich, W. F., Weaver, A., Lincoln, S. E., Steen, R. G., Stein, L. D., Nadeau, J. H. and Lander, E. S. (1993) A genetic linkage map of the mouse: current applications and future prospects. *Science* **262:** 57-66.

Crane, C. F., Beversdorf, W. B. and Bingham, E. T. (1982) Chromosome pairing and associations at meiosis in haploid soybean. *Can. J. Genet. Cytol.* **24:** 293-300.

Crow, J. F. and Kimura, M. (1970) *An Introduction to Population Genetics Theory.* Harper and Row, New York.

Davisson, M. T., Lalley, P. A., Peters, J., Doolittle, D. P., Hillyard, A. L. and Searle, A. G. (1991) Report of the comparative committee for human, mouse and other rodents. *Cytogenet. Cell Genet.* **58:** 1152-1189.

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

Degnan, B. M., Degnan, S. M., Giusti, A. and Morse, D. E. (1995) A *hox/hom* homeobox gene in sponges. *Gene* **155:** 175-177.

Deloukas, P., Schuler, G. D., Gyapay, G., Beasley, E. M., Soderlund, C., Rodriguez-Tomé, P., Hui, L., Matise, T. C., McKusick, K. B., Beckmann, J. S. *et al.* (1998) A physical map of 30,000 human genes. *Science* **282:** 744-746.

Dickman, S. (1997) Possible new roles for *HOX* genes. *Science* **278:** 1882-1883.

Dinulos, M. B., Bassi, M. T., Rugarli, E. I., Chapman, V., Ballabio, A. and Disteche, C. M. (1996) A new region of conservation is defined between human and mouse X chromosomes. *Genomics* **35:** 244-247.

Doolittle, R. F., Da-Fei, F., Tsang, S., Cho, G. and Little, E. (1996) Determining divergence times of the major kingdoms of living organisms with a protein clock. *Science* **271:** 470-477.

Duboule, D. and Dollé, P. (1989) The structural and functional organization of the murine *Hox* gene family resembles that of *Drosophila*. *EMBO J.* **8:** 1497-1505.

Dujon, B. (1996) The yeast genome project: what did we learn? *Trends Genet.* **12:** 263-270.

Dutly, F. and Schnizel, A. (1996) Unequal interchromosomal rearrangements may result in elastin gene deletions causing the Williams-Beuren syndrome. *Hum. Mol. Genet.* **5:** 1893-1898.

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

Eichler, E. E. (1998) Masquerading repeats: paralogous pitfalls of the human genome. *Genome Res.* **8:** 758-762.

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

Eppig, J. T. (1996) Comparative maps: adding pieces to the mammalian jigsaw. *Curr. Opin. Genet. Dev.* **6:** 723-730.

Eppig, J. T. and Nadeau, J. H. (1995) Comparative maps: the mammalian jigsaw puzzle. *Curr. Opin. Genet. Dev.* **5:** 709-716.

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

Feldmann, M. (1976) Wheats, pp. 120-128 in *Evolution of crop plants*, edited by Simmonds, N. W. Longman, New York.

Ferris, S. D. and Whitt, G. S. (1977) Loss of duplicate gene expression after polyploidisation. *Nature* **265:** 258-260.

Fields, C., Adams, M. D., White, O. and Venter, J. C. (1994) How many genes in the human genome? *Nature Genet.* **7:** 345-346.

Fisher, R. A. (1936) *Am. Nat.* **69:** 446-455.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J.-F., Dougherty, B. A., Merrick, J. M. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269:** 496-512.

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:** 1531-1545.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., Fritchman, J. L., Weidman, J. F., Small, K. V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T. R., Saudek, D. M., Phillips, C. A., Merrick, J. M., Tomb, J.-F., Dougherty, B. A., Bott, K. F., Hu, P.-C., Lucier, T. S., Peterson, S. N., Smith, H. O., Hutchinson III, C. A. and Venter, J. C. (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science* **270:** 397-403.

Fryxell, K. J. (1996) The coevolution of gene family trees. *Trends Genet.* **12:** 364-369.

Gale, M. D. and Devos, K. M. (1998) Comparative genetics in the grasses. *Proc. Natl. Acad. Sci. USA* **95:** 1971-1974.

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

Gamer, L. W. and Wright, C. V. E. (1994) Murine *Cdx-4* bears striking resemblance to the *Drosophila* caudal gene in its homeodomain sequence and early expression pattern. *Mech. Dev.* **43:** 71-81.

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

Gates, M. A., Kim, L., Egan, E. S., Cardozo, T., Sirotkin, 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 Res.* **9:** 334-347.

Gaut, B. S. and Doebley, J. F. (1997) DNA sequence evidence for the segmental allotetraploid origin of maize. *Proc. Natl. Acad. Sci. USA* **94:** 6809-6814.

Gehring, W. J., Qian, Y. Q., Billeter, M., Furukubo-Tokunaga, K. and Schier, A. F. (1994) Homeodomain-DNA recognition. *Cell* **78:** 211-223.

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

Georges, M. and Andersson, L. (1996) Livestock genomics comes of age. *Gen. Res.* **6:** 907-921.

Gibbs, A. J. and McIntyre, G. A. (1970) The diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequences. *Eur. J. Biochem.* **16:** 1-11.

Gibson, T. J. and Spring, J. (1998) Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins. *Trends Genet.* **14:** 46-49.

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. *Hum. Mol. Genet.* **8:** 1313-1320.

Gish, W. and States, D. J. (1993) Identification of protein coding regions by database similarity search. *Nature Genet.* **3:** 266-272.

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:** 546-567.

Graf, J.-D. (1989) Genetic mapping in *Xenopus laevis*: eight linkage groups established. *Genetics* **123:** 389-398.

Graham, A., Papalopulu, N. and Krumlauf, R. (1989) The murine and Drosophila homeobox gene complexes have common features of organization and expression. *Cell* **57:** 367-378.

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

Hani, J. and Feldmann, H. (1998) tRNA genes and retroelements in the yeast genome. *Nucl. Acids Res.* **26:** 689-696.

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

Hightower, R. C. and Meagher, R. B. (1985) Divergence and differential expression of soybean actin genes. *EMBO J.* **4:** 1-8.

Hilu, K. W. (1993) Polyploidy and the evolution of domesticated plants. *Am. J. Bot.* **80:** 1494-1499.

Holland, P. W. H. and Garcia-Fernàndez, J. (1996) *Hox* genes and chordate evolution. *Developmental Biology* **173:** 382-395.

Holland, P. W. H., Garcia-Fernàndez, J., Williams, N. A. and Sidow, A. (1994) Gene duplications and the origins of vertebrate development. *Development* **Suppl.:** 125-133.

Holland, P. W. H., Holland, L. Z., Williams, N. A. and Holland, N. D. (1992) An amphioxus homeobox gene: sequence conservation, spatial expression during development and insights into vertebrate evolution. *Development* **116:** 653-661.

Holland, P. W. H. and Williams, N. A. (1990) Conservation of *engrailed*-like homeobox sequences during vertebrate evolution. *FEBS Letts* **277:** 250-252.

Hood, L., Kronenberg, M. and Hunkapiller, T. (1985) T cell antigen receptors and the immunoglobulin supergene family. *Cell* **40:** 225-229.

Hosbach, H. A., Wyler, T. and Weber, R. (1983) The *Xenopus laevis* globin gene family: chromosomal arrangement and gene structure. *Cell* **32:** 45-53.

Huang, X., Hardison, R. C. and Miller, W. (1990) A space-efficient algorithm for local similarities. *CABIOS* **6:** 373-381.

Hughes, A. L. (1994) The evolution of functionally novel proteins after gene duplication. *Proc. R. Soc. Lond. B* **256:** 119-124.

Hughes, A. L. (1998) Phylogenetic tests of the hypothesis of block duplication of homologous genes on human chromsomes 6, 9, and 1. *Mol. Biol. Evol.* **15:** 854-870.

Hughes, A. L. (1999) Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history. *J. Mol. Evol.* **48:** 565-576.

Hughes, A. L. and Friedman, R. (submitted) Ancient genome duplications did not structure the vertebrate genome: evidence from the human Hox chromosomes. .

Hughes, A. L. and Nei, M. (1993) Evolutionary relationships of the classes of major histocompatibility complex genes. *Immunogenetics* **37:** 337-346.

Hughes, A. L. and Yeager, M. (1997) Molecular evolution of the vertebrate immune system. *Bioessays* **19:** 777-786.

Hughes, M. K. and Hughes, A. L. (1993) Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*. *Mol. Biol. Evol.* **10:** 1360-1369.

Hunkapiller, T. and Hood, L. (1989) Diversity of the immunoglobulin gene superfamily. *Adv. Immunol.* **44:** 1-63.

Hurst, L. D. (1995) The silence of the genes. *Curr. Biol.* **5:** 459-461.

Imanishi, T., Endo, T. and Gojobori, T. (1997) An exhaustive search for extensive chromosomal regions duplicated within the human genome. *HGM '97 Poster*, Toronto, Canada.

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

Jeffreys, A. J., Wilson, V., Wood, D., Simons, J. P., Kay, R. M. and Williams, J. G. (1980) Linkage of adult alpha- and beta-globin genes in X. laevis and gene duplication by tetraploidization. *Cell* **21:** 555-564.

Johansson, M., Ellegren, H. and Andersson, L. (1995) Comparative mapping reveals extensive linkage conservation — but with gene order rearrangements — between the pig and human genomes. *Genomics* **25:** 682-690.

Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., Kimura, T., Hosouchi, T., Matsuno, A., Muraki, A., Nakazaki, N., Naruo, K., Okumura, S., Shimpo, S., Takeuchi, C., Wada, T., Watanabe, A., Yamada, M., Yasuda, M. and Tabata, S. (1996) Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* **3:** 109-136.

Kappen, C. and Ruddle, F. H. (1993) Evolution of a regulatory gene family: *HOM/HOX* genes. *Curr. Biol.* **3:** 931-938.

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

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

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. *Proc. Natl. Acad. Sci. USA* **93:** 9096-9101.

Kasahara, M., Nakaya, J., Satta, Y. and Takahata, N. (1997) Chromosomal duplication and the emergence of the adaptive immune system. *Trends Genet.* **13:** 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:** 101-108.

Kavenoff, R. and Zimm, B. H. (1973) Chromosome-sized DNA molecules from *Drosophila. Chromosoma* **41:** 1-27.

Kavsan, V., Koval, A., Petrenko, O., Roberts Jr., C. T. and LeRoith, D. (1993) Two insulin genes are present in the salmon genome. *Biochem. Biophys. Res. Comm.* **191:** 1373-1378.

Kehrer-Sawatzki, H., Haussler, J., Krone, W., Bode, H., Jenne, D. E., Mehnert, K. U., Tummers, U. and Assum, G. (1997) The second case of a t(17;22) in a family with neurofibromatosis type 1: sequence analysis of the breakpoint regions. *Hum. Genet.* **99:** 237-247.

Kehrer-Sawatzki, H., Maier, C., Moschgath, E., Elgar, G. and Krone, W. (1999) Characterization of three genes, AKAP84, BAW and WSB1. *Gene* **235:** 1-11.

Keogh, R. S., Seoighe, C. and Wolfe, K. H. (1998) Evolution of gene order and chromosome number in *Saccharomyces*, *Kluyveromyces* and related fungi. *Yeast* **in press**.

Kimura, M. and Ohta, T. (1974) On some principles governing molecular evolution. *Proc. Natl. Acad. Sci. USA* **71:** 2848-2852.

Klein, J. and O'hUigin, C. (1993) Composite origin of major histocompatibilty complex genes. *Curr. Opin. Genet. Dev.* **3:** 923-930.

Kobel, H. R. and Du Pasquier, L. (1986) Genetics of polyploid *Xenopus*. *Trends Genet.* **2:** 310-315.

Koh, Y.-S. and Moore, D. D. (1999) Linkage of the nuclear hormone receptor genes NR1D2, THRB, and RARB: evidence for an ancient, large-scale duplication. *Genomics* **57:** 289-292.

Koonin, E. V., Mushegian, A. R. and Rudd, K. E. (1996) Sequencing and analysis of bacterial genomes. *Curr. Biol.* **6:** 404-416.

Koonin, E. V., Tatusov, R. L. and Rudd, K. E. (1995) Sequence similarity analysis of *Escherichia coli* proteins: Functional and evolutionary implications. *Proc. Natl. Acad. Sci. USA* **92:** 11921-11925.

Koop, B. F. and Nadeau, J. H. (1996) Pufferfish and a new paradigm for comparative genome analysis. *Proc. Natl. Acad. Sci. USA* **93:** 1363-1365.

Kulski, J. K., Gaudieri, S., Bellgard, M., Balmer, L., Giles, K., Inoko, H. and Dawkins, R. L. (1997) The evolution of MHC diversity by segmental duplication and transposition of retroelements. *J. Mol. Evol.* **45:** 599-609.

Kurosawa, G., Yamada, K., Ishiguro, H. and Hori, H. (1999) *Hox* gene complexity in medaka fish may be similar to that in pufferfish rather than zebrafish. *Biochem. Biophys. Res. Comm.* **260:** 66-70.

Labedan, B. and Riley, M. (1995) Widespread protein sequence similarities: Origins of *Escherichia coli* genes. *J. Bact.* **177:** 1585-1588.

Larhammar, D. and Risinger, C. (1993) Molecular genetic aspects of tetraploidy in the common carp *Cyprinus carpio*. *Mol. Phylogenet. Evol.* **3:** 59-68.

Lawrence, P. and Morata, G. (1994) Homeobox genes: their function in Drosophila segmentation and pattern formation. *Cell* **78:** 181-189.

Lee, J. S. and Verma, D. P. S. (1984) Structure and chromosomal arrangement of leghemoglobin genes in kidney bean suggest divergence in soybean leghemoglobin gene loci following tetraploidization. *EMBO J* **3:** 2745-2752.

Leipold, M. and Schmidtke, J. (1982) Gene expression in phylogenetically polyploid organisms, pp. 219-236 in *Genome evolution*, edited by Dover, G. and Flavell, R. Academic Press, New York.

Lewin, B. (1990) *Genes IV*. Oxford University Press, Oxford.

Li, W.-H. (1989) A statistical test of phylogenies estimated from sequence data. *Mol. Biol. Evol.* **6:** 424-435.

Li, W.-H. (1990) Statistical tests of molecular phylogenies. *Methods in Enzymology* **183:** 645-659.

Li, W.-H., Wu, C.-I. and Luo, C.-C. (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* **2:** 150-174.

Lim, E. H. and Brenner, S. (1997) Short-range linkage relationships of the valyl-tRNA synthetase gene in *Fugu rubripes*. *Immunogenetics* **46:** 332-336.

Lim, E. H. and Brenner, S. (1999) Short-range linkage relationships, genomic organisation and sequence comparisons of a cluster of five *HSP70* genes in *Fugu rubripes*. *Cell. Mol. Life. Sci.* **55:** 668-678.

Lim, S. T. and Bailey, G. S. (1977) Gene duplication in salmonid fishes: evidence for duplicated but catalytically equivalent A_4 lactate dehydrogenases. *Biochem. Genet.* **15:** 707-721.

Lim, S. T., Kay, R. M. and Bailey, G. S. (1975) Lactate dehydrogenase isozymes of salmonid fish. Evidence for unique and rapid functional divergence of duplicated H_4 lactate dehydrogenases. *J. Biol. Chem.* **250:** 1790-1800.

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:** 295-308.

Lundin, L. G. (1979) Evolutionary conservation of large chromosomal segments reflected in mammalian gene maps. *Clin. Genet.* **16:** 72-81.

Lundin, L. G. (1985) Possible paralogous chromosomal regions in mouse and man. *Genet. Res.* **45:** 214.

Lundin, L. G. (1989) Gene homologies with emphasis on paralogous genes and chromosomal regions. *Life Sci. Adv. (Genet.)* **8:** 89-104.

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

Lyons, L. A., Laughlin, T. F., Copeland, N. G., Jenkins, N. A., Womack, J. E. and O'Brien, S. J. (1997) Comparative anchor tagged sequences (CATS) for integrative mapping of mammalian genomes. *Nature Genet.* **15:** 47-56.

Makałowski, W. and Boguski, M. S. (1998) Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl. Acad. Sci. USA* **95:** 9407-9412.

Makałowski, W., Zhang, J. and Boguski, M. S. (1996) Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences. *Genome Res.* **6:** 846-857.

Marshall, C. R., Raff, E. C. and Raff, R. A. (1994) Dollo's law and the death and resurrection of genes. *Proc. Natl. Acad. Sci. USA* **91:** 12283-12287.

Marshall, E. (1998) NIH to produce a 'working draft' of the genome by 2001. *Science* **281:** 1774-1775.

Martinez, P., Rast, J. P., Arenas-Mena, C. and Davidson, E. H. (1999) Organization of an echinoderm Hox gene cluster. *Proc. Natl. Acad. Sci. USA* **96:** 1469-1474.

McCarrey, J. R. and Thomas, K. (1987) Human testis-specific PGK gene lacks introns and possesses charcteristics of a processed gene. *Nature* **326:** 501-505.

McGinnis, W. and Krumlauf, R. (1992) Homeobox genes and axial patterning. *Cell* **66:** 283-302.

McLysaght, A., Enright, A., Skrabanek, L. and Wolfe, K. H. (submitted) Estimation of synteny conservation and genome compaction between pufferfish (Fugu) and human. .

Mewes, H. W., Albermann, K., Bähr, M., Frishmann, D., Gleissner, A., Hani, J., Heumann, K., Kleine, K., Maierl, A., Oliver, S. G., Pfeiffer, F. and Zollner, A. (1997) Overview of the yeast genome. *Nature* **387:** 7-65.

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

Miklos, G. L. G. and Rubin, G. M. (1996) The role of the genome project in determining gene function: insights from model organisms. *Cell* **86:** 521-529.

Miles, C., Elgar, G., Coles, E., Kleinjan, D.-J., 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. *Proc. Natl. Acad. Sci. USA* **95:** 13068-13072.

Misof, B. Y. and Wagner, G. P. (1996) Evidence for four Hox clusters in the killifish Fundulus heteroclitus (Teleostei). *Mol. Phylogenet. Evol.* **5:** 309-322.

Miyata, T. and Yasunaga, T. (1980) Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *J. Mol. Evol.* **16:** 23-36.

Miyazawa, K., Wang, Y., Minoshima, S., Shimizu, N. and Kitamura, N. (1998) Structural organization and chromosomal localization of the human hepatocyte growth factor activator gene — phylogenetic and functional relationship with blood coagulation factor XII, urokinase and tissue-type plasminogen activator. *Eur. J. Biochem.* **258:** 355-361.

Mizuki, N., Ando, H., Kimura, M., Ohno, S., Miyata, S., Yamazaki, M., Tashiro, H., Watanabe, K., Ono, A., Taguchi, S., Sugawara, C., Fukuzumi, Y., Okumura, K., Goto, K., Ishihara, M., Nakamura, S., Yonemoto, J., Kikuti, Y. Y., Shiina, T., Chen, L., Ando, A., Ikemura, T. and Inoko, H. (1997) Nucleotide sequence analysis of the HLA class I region spanning the 237-kb segment around the HLA-B and -C genes. *Genomics* **42:** 55-66.

Moore, G. (1995) Cereal genome evolution: pastoral pursuits with 'Lego' genomes. *Curr. Opin. Genet. Dev.* **5:** 717-724.

Moore, G., Devos, K. D., Wang, Z. and Gale, M. D. (1995a) Grasses, line up and form a circle. *Curr. Biol.* **5:** 737-739.

Moore, G., Foote, T., Helentjaris, T., Devos, K., Kurata, N. and Gale, M. (1995b) Was there a single ancestral cereal genome? *Trends Genet.* **11:** 80-81.

Morizot, D. C. (1990) Use of fish gene maps to predict ancestral vertebrate genome organization, pp. 207-234 in *Isozymes: structure, function, and use in biology and medicine*, edited by Ogita, Z.-I. and Markert, C. L. Wiley-Liss, Inc., New York.

Muller, H. (1925) *Am. Nat.* **59:** 346-353.

Muse, S. V. (1996) Estimating synonymous and nonsynonymous substitution rates. *Mol. Biol. Evol.* **13:** 105-114.

Muse, S. V. and Gaut, B. S. (1994) A likelihood approach for comparing synonymous and nonsynonmous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.* **11:** 715-724.

Nadeau, J. H. (1989) Maps of linkage and synteny homologies between mouse and man. *Trends Genet.* **5:** 82-86.

Nadeau, J. H. (1991) Genome duplication and comparative gene mapping, pp. 269-296 in *Advanced Techniques in Chromosome Research*, edited by Adolph, K. W. Marcell Dekker, New York.

Nadeau, J. H., Compton, J. G., Giguère, V., Rossant, J. and Varmuza, S. (1992) Close linkage of retinoic acid receptor genes with homeobox- and keratin-encoding genes on paralogous segments of mouse chromosomes 11 and 15. *Mamm. Genome* **3:** 202-208.

Nadeau, J. H. and Kosowsky, M. (1991) Mouse map of paralogous genes. *Mamm. Genome* **1:** S433-S460.

Nadeau, J. H. and Sankoff, D. (1997) Comparable rate of gene loss and functional divergence after genome duplications early in vertebrate evolution. *Genetics* **147:** 1259-1266.

Nadeau, J. H. and Sankoff, D. (1998) The lengths of undiscovered conserved segments in comparative maps. *Mamm. Genome* **9:** 491-495.

Nadeau, J. H. and Taylor, B. A. (1984) Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc. Natl. Acad. Sci. USA* **81:** 814-818.

Nei, M., Gu, X. and Sitnikova, T. (1997) Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proc. Natl. Acad. Sci. USA* **94:** 7799-7806.

Nowak, M. A., Boerlijst, M. C., Cooke, J. and Maynard Smith, J. (1997) Evolution of functional redundancy. *Nature* **388:** 167-171.

O'Brien, S. J. (1991) Mammalian genome mapping: lessons and prospects. *Curr. Opin. Genet. Dev.* **1:** 105-111.

O'Brien, S. J. and Graves, J. A. M. (1991) Report of the committee on comparative gene mapping. *Cytogenet. Cell Genet.* **58:** 1124-1151.

O'Brien, S. J., Menotti-Raymond, M., Murphy, W. J., Nash, W. G., Wienberg, J., Stanyon, R., Copeland, N. G., Jenkins, N. A., Womack, J. E. and Graves, J. A. M. (1999) The promise of comparative genomics in mammals. *Science* **286:** 458-480.

O'Brien, S. J., Wienberg, J. and Lyons, L. A. (1997) Comparative genomics: lessons from cats. *Trends Genet.* **13:** 393-399.

O'Brien, S. J., Womack, J. E., Lyons, L. A., Moore, K. J., Jenkins, N. A. and Copeland, N. G. (1993) Anchored reference loci for comparative genome mapping in mammals. *Nature Genet.* **3:** 103-112.

Ohno, S. (1967) Sex chromosomes and sex linked genes. In *Monographs on Endocrinology*, Vol. 1, edited by Labhart, A., Mann, T. and Samuels, L. T. Springer-Verlag, Heidelberg.

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

Ohno, S. (1972) So much junk in our genome, pp. 366-370 in *Evolution of genetic systems (Brookhaven Symposia in Biology No. 23)*, edited by Smith, H. H. Gordon and Breach, New York.

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

Ohno, S. (1985) Dispensable genes. *Trends Genet.* **1:** 160-164.

Ohno, S. (1993) Patterns in genome evolution. *Curr. Opin. Genet. Dev.* **3:** 911-914.

Ohno, S. (1998) The notion of the Cambrian pananimalia genome and a genomic difference that separated vertebrates from the invertebrates, pp. in *Molecular Evolution: evidence for monophyly of metazoa (Progress in molecular and subcellular biology)*, edited by Müller, W. E., Kuchino, Y., Jeantur, P. and Paine, P. L. Springer-Verlag, New York.

Ohno, S., Muramoto, J., Christian, L. and Atkin, N. B. (1967) Diploid-tetraploid relationship among old-world members of the fish family Cyprinidae. *Chromosoma* **23:** 1-19.

Ohno, S., Wolf, U. and Atkin, B. (1968) Evolution from fish to mammals by gene duplication. *Hereditas* **59:** 169-187.

Ota, T. and Nei, M. (1994) Divergent evolution and evolution by the birth-and-death process in the immunoglobulin V_H gene family. *Mol. Biol. Evol.* **11:** 469-482.

Palmer, S., Perry, J. and Ashworth, A. (1995) A contravention of Ohno's law in mice. *Nature Genet.* **10:** 472-476.

Paterson, A. H., Lan, T.-H., Reischmann, K. P., Chang, C., Lin, Y.-R., Liu, S.-C., Burow, M. D., Kowalski, S. P., Katsar, C. S., DelMonte, T. A., Feldmann, K. A., Schertz, K. F. and Wendel, J. F. (1996) Toward a unified genetic map of higher plants, transcending the monocot-dicot divergence. *Nature Genetics* **14:** 380-382.

Patton, S. J., Luke, G. N. and Holland, P. W. H. (1998) Complex history of a chromosomal paralogy region: insights from amphioxus aromatic amino acid hydroxylase genes and insulin-related genes. *Mol. Biol. Evol.* **15:** 1373-1380.

Pébusque, 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. *Mol. Biol. Evol.* **15:** 1145-1159.

Pendleton, J. W., Nagia, B. K., Murtha, M. T. and Ruddle, F. H. (1993) Expansion of the *Hox* gene family and the evolution of chordates. *Proc. Natl. Acad. Sci. USA* **90:** 6300-6304.

Pennisi, E. (1998) How the genome readies itself for evolution. *Science* **281:** 1131-1134.

Perler, R., Efstratiadis, A., Lomedico, P., Gilbert, W., Klodner, R. and Dodgson, J. (1980) The evolution of genes: the chicken preproinsulin gene. *Cell* **20:** 555-566.

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., Westerfield, M., Zon, L. I., Johnson, S. L. and Talbot, W. S. (1998) Vertebrate genome evolution and the zebrafish gene map. *Nature Genet* **18:** 345-349.

Poustka, A. J., Herwig, R., Krause, A., Hennig, S., Meier-Ewert, S. and Lehrach, H. (1999) Toward the gene catalogue of sea urchin development: the construction and analysis of an unfertilized egg cDNA library highly normalized by oligonucleotide fingerprinting. *Genomics* **59:** 122-133.

Prince, V. E., Joly, L., Ekker, M. and Ho, R. K. (1998) Zebrafish *hox* genes: genomic organization and modified colinear expression patterns in the trunk. *Development* **125:** 407-420.

Puech, A., Saint-Jore, B., Funke, B., Gilbert, D. J., Sirotkin, H., Copeland, N. G., Jenkins, N. A., Kucherlapati, R., Morrow, B. and Skoultchi, A. I. (1997) Comparative mapping of the human 22q11 chromosomal region and the orthologous region in mice reveals complex changes in gene organization. *Proc. Natl. Acad. Sci. USA* **94:** 14608-14613.

Purandare, S. M. and Patel, P. I. (1997) Recombination hot spots. *Genome Res.* **7:** 773-786.

Rasch, E. M., Barr, H. J. and Rasch, R. W. (1971) The DNA content of sperm of Drosophila melanogaster. *Chromosoma* **33:** 1-18.

Reboul, J., Gardiner, K., Monneron, D., Uze, G. and Lutfalla, G. (1999) Comparative genomic analysis of the interferon/interleukin-10 receptor gene cluster. *Genome Res.* **9:** 242-250.

Remmers, E. F., Goldmutz, E. A., Cash, J. M., Crofford, L. J., Misiewicz Poltorak, B., Zha, H. and Wilder, R. L. (1992) Genetic map of nine polymorphic loci comprising a single linkage group on rat chromosome 10: evidence for linkage conservation with human chromosome 17 and mouse chromosome 11. *Genomics* **14:** 618-623.

Rogers, J. H. (1983) Retroposons defined. *Nature* **301:** 460.

Ruddle, F. H., Bartels, J. L., Bentley, K. L., Kappen, C., Murtha, M. T. and Pendleton, J. W. (1994a) Evolution of *HOX* genes. *Ann. Rev. Genet.* **28:** 423-442.

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

Rudkin, G. T. (1972) Replication in polytene chromosomes, pp. 59-85 in *Developmental studies on giant chromosomes*, edited by Beermann, W. Springer-Verlag, Berlin.

Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4:** 406-425.

Sankoff, D., Ferretti, V. and Nadeau, J. H. (1997a) Conserved segment identification, pp. 252-256 in *RECOMB 97. Proceedings of the First Annual International Conference on Computational Molecular Biology*. ACM Press, New York.

Sankoff, D., Parent, M.-N., Marchand, I. and Ferretti, V. (1997b) On the Nadeau-Taylor theory of conserved chromosome segments, pp. 262-274 in *Combinatorial Pattern Matching. Eighth Annual Symposium*, edited by Apostolico, A. and Hein, J. Springer-Verlag.

Schofeld, J. P., Elgar, G., Greystone, 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:** 158-167.

Schughart, K., Kappen, C. and Ruddle, F. H. (1989) Duplication of large genomic regions during the evolution of vertebrate homeobox genes. *Proc. Natl. Acad. Sci. USA* **86:** 7067-7071.

Schuler, G. D., Boguski, M. S., Stewart, E. A., Stein, L. D., Gyapay, G., Rice, K., White, R. E., Rodriguez-Tomé, P., Aggarwal, A., Bajorek, E. *et al.* (1996) A gene map of the human genome. *Science* **274:** 540-546.

Scott, M. P. (1992) Vertebrate homeobox gene nomenclature. *Cell* **71:** 551-553.

Seldin, M. F., Saunders, A. M., Rochelle, J. M. and Howard, T. A. (1991) A proximal mouse chromosome 9 linkage map that further defines groups homologous with segments of human chromosomes 11, 15, and 19. *Genomics* **9:** 678-685.

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. (in press) Prevalence of small inversions in yeast gene order evolution. *Proc. Natl. Acad. Sci. USA* .

Seoighe, C. and Wolfe, K. H. (1998) Extent of genomic rearrangement after genome duplication in yeast. *Proc. Natl. Acad. Sci. USA* **95:** 4447-4452.

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

Sharkey, M., Graba, Y. and Scott, M. P. (1997) *Hox* genes in evolution: protein surfaces and paralog groups. *Trends Genet.* **13:** 145-151.

Sharman, A. C., Hay-Schmidt, A. and Holland, P. W. H. (1997) Cloning and analysis of an HMG gene from the lamprey *Lampetra fluviatilis*: gene duplication in vertebrate evolution. *Gene* **184:** 99-105.

Sharman, A. C. and Holland, P. W. H. (1996) Conservation, duplication and divergence of developmental genes during chordate evolution. *Neth. J. Zool.* **46:** 47-67.

Sharman, A. C. and Holland, P. W. H. (1998) Estimation of *Hox* gene cluster number in lampreys. *Int. J. Dev. Biol.* **42:** 617-620.

Shoemaker, R. C., Plozin, K., Labate, J., Specht, J., Brummer, E. C., Olson, T., Young, N., Concibido, V., Wilcox, J., Tamulonis, J. P., Kochert, G. and Boerma, H. R. (1996) Genome duplication in soybean (*Glycine* subgenus *soja*). *Genetics* **144:** 329-338.

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

Simmen, M. W., Leitgeb, S., Clark, V. H., Jones, S. J. and Bird, A. (1998) Gene number in an invertebrate chordate, *Ciona intestinalis*. *Proc. Natl. Acad. Sci. USA* **95:** 4437-4440.

Sitnikova, T. and Nei, M. (1998) Evolution of immunoglobulin kappa chain variable region genes in vertebrates. *Mol. Biol. Evol.* **15:** 50-60.

Skrabanek, L. and Wolfe, K. H. (1998) Eukaryote genome duplication - where's the evidence? *Curr. Opin. Genet. Dev.* **8:** 694-700.

Smith, M. M. (1987) Molecular evolution of the *Saccharomyces cerevisiae* histone gene loci. *J. Mol. Evol.* **24:** 252-259.

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

Soares, M. B., Schon, E., Henderson, A., Karathanasis, S. K., Cate, R., Zeitlin, S., Chirgwin, J. and Efstratiadis, A. (1985) RNA-mediated gene duplication: the rat preproinsulin I gene is a functional retroposon. *Mol. Cell Biol.* **5:** 2090-2103.

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. *Proc. Natl. Acad. Sci. USA* **92:** 7719-7723.

Sonstegaard, T. S., Kappes, S. M., Keele, J. W. and Smith, T. P. L. (1998) Refinement of bovine chromosome 2 linkage map near the *mh* locus reveals rearrangements between the bovine and human genomes. *Animal Genet.* **29:** 341-347.

Spring, J. (1997) Vertebrate evolution by interspecific hybridisation - are we polyploid? *FEBS Letts.* **400:** 2-8.

Spring, J., Goldberger, O. A., Jenkins, N. A., Gilbert, D. J., Copeland, N. G. and Bernfield, M. (1994) Mapping of the syndecan genes in the mouse: linkage with members of the myc gene family. *Genomics* **21:** 597-601.

Stein, S., Niss, K. and Kessel, M. (1996) Check-list — vertebrate homeobox genes. *Mech. Dev.* **55:** 519-533.

Stock, D. W., Ellies, D. L., Zhao, Z., Ekker, M., Ruddle, F. H. and Weiss, K. M. (1996) The evolution of the vertebrate Dlx gene family. *Proc. Natl. Acad. Sci. USA* **93:** 10858-10863.

Stock, D. W. and Whitt, G. S. (1992) Evidence from 18S ribosomal RNA sequences that lampreys and hagfish form a natural group. *Science* **257:** 787-789.

Stubbs, L., Carver, E. A., Shannon, M. E., Kim, J., Geisler, J., Generoso, E. E., Stanford, B. G., Dunn, W. C., Mohrenweiser, H., Zimmermann, W., Watt, S. M. and Ashworth, L. K. (1996) Detailed comparative map of human chromosome 19q and related regions of the mouse genome. *Genomics* **35:** 499-508.

Sun, H. S., Cai, L., Davis, S. K., Taylor, J. F., Doud, L. K., Bishop, M. D., Hayes, H., Barendse, W., Vaiman, D., McGraw, R. A., Hirano, T., Sugimoto, Y. and Kirkpatrick, B. W. (1997) Comparative linkage mapping of human chromosome 13 and bovine chromosome 12. *Genomics* **39:** 47-54.

Sutherland, H. F., Kim, U.-J. and Scambler, P. J. (1998) Cloning and comparative mapping of the DiGeorge syndrome critical region in the mouse. *Genomics* **52:** 37-43.

Tassabehji, M., Read, A. P., Newton, V. E., Harris, R., Balling, R., Gruss, P. and Strachan, T. (1992) Waardenburg's syndrome patients have mutations in the human homologue of the Pax-3 paired box gene. *Nature* **355:** 635-636.

Taylor, J. H., Woods, P. S. and Hughes, W. L. (1957) The organization and duplication of chromosomes as revealed by autoradiographic studies using tritium-labeled thymidine. *Proc. Natl. Acad. Sci. USA* **43:** 122-128.

Thiébaud, C. H. and Fischberg, M. (1977) DNA content in the genus *Xenopus*. *Chromosoma* **59:** 253-257.

Utter, F. M., Allendorf, F. W. and May, B. (1979) Genetic basis of creatine kinase isozymes in skeletal muscle of salmonid fishes. *Biochem. Genet.* **17:** 1079-1091.

Uyeno, T. and Smith, G. R. (1972) Tetraploid origin of the karyotype of catostomid fishes. *Science* **175:** 644-646.

Uyttendaele, H., Marazzi, G., Wu, G., Yan, Q., Sassoon, D. and Kitajewski, J. (1996) *Notch4/int3*, a mammary proto-oncogene, is an endothelial cell-specific mammalian *Notch* gene. *Development* **122:** 2251-2259.

Vanin, E. F. (1985) Processed pseudogenes: characteristics and evolution. *Ann. Rev. Genet.* **19:** 253-272.

Vogel, G. (1998) Doubled genes may explain fish diversity. *Science* **281:** 1119-1121.

Wadman, M. (1998) Human genome deadline cut by two years. *Nature* **395:** 207.

Wagner, A. (1994) Evolution of gene networks by gene duplications: a mathematical model and its implications on genome organization. *Proc. Natl. Acad. Sci. USA* **91:** 4387-4391.

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

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. *et al.* (1999) A radiation hybrid map of the rat genome containing 5,255 markers. *Nature Genet.* **22:** 27-36.

Waterston, R. and Sulston, J. (1995) The genome of *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. USA* **92:** 10836-10840.

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:** 114-122.

Wilson, W. A., Harrington, S. E., Woodman, W. L., Lee, M., Sorrells, M. E. and McCouch, S. R. (1999) Inferences on the genome structure of progenitor maize through comparative analysis of rice, maize and the domesticated panicoids. *Genetics* **153:** 453-473.

Wolfe, K. H., Gouy, M., Yang, Y.-W., Sharp, P. M. and Li, W. H. (1989) Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. *Proc. Natl. Acad. Sci. USA* **86:** 6201-6205.

Wolfe, K. H. and Sharp, P. M. (1993) Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. *J. Mol. Evol.* **37:** 441-456.

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

Womack, J. E. and Kata, S. R. (1995) Bovine genome mapping: evolutionary inference and the power of comparative genomics. *Curr. Opin. Genet. Dev.* **5:** 725-733.

Womack, J. E. and Moll, Y. D. (1986) A gene map of the cow: conservation of linkage with mouse and man. *J. Hered.* **77:** 2-7.

Zeltser, L., Desplan, C. and Heintz, N. (1996) Hoxb13 - a new Hox gene located in a distant region of the Hoxb cluster maintains colinearity. *Development* **122:** 2475-2484.

Zhang, J. and Madden, T. L. (1997) PowerBLAST: a new network BLAST application for interactive or automated sequence analysis and annotation. *Genome Res.* **7:** 649-656.

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

Zhang, J., Rosenberg, H. F. and Nei, M. (1998) Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc. Natl. Acad. Sci. USA* **95:** 3708-3713.

Appendix A: Functions of proteins coded for by genes involved in possible paralogous regions in the human genome defined by the Schuler dataset (1996) and detailed in Tables 4.2 and 4.3 (arranged alphabetically)

5H1E: 5-hydroxytryptamine 1E receptor
A4: Alzheimer's disease amyloid A4 protein (protease nexin-II)
ACM4: muscarinic acetylcholine receptor M4
ANK1: ankyrin 1
ANX2: annexin II, lipocortin II
ANX6: annexin VI, lipocortin VI
ANX7: annexin VII, synexin
ANX8: annexin VIII (vascular anticoagulant-β)
ANXD: annexin XIII
APJ: G protein-coupled receptor
APP2: amyloid-like protein 2
AT7A: copper-transporting atpase 1
ATCE: calcium-transporting atpase sarcoplasmic reticulum type, class 2 isoform
ATCP: calcium-transporting ATPase plasma membrane, isoform 1β
ATCQ: calcium-transporting ATPase plasma membrane, brain isoform 2
B3A3: anion exchange protein 3
B3AT: anion exchange protein 1
C1R: complement C1R component
C1S: complement C1S component
C5AR: C5A anaphylatoxin chemotactic receptor
CALR: calcitonin receptor
CDK8: cell division protein kinase 8
CP12: cytochrome P450, IA2
CPCA: cytochrome P450 2C10
CPSS: C-protein, skeletal muscle slow-isoform
CPT7: cytochrome P450, steroid 17α hydroxylase
CRAR: complement-activating component of ra-reactive factor
CTNA: alpha catenin
CYPB: peptidyl-prolyl *cis*-trans isomerase B
CYPM: peptidyl-prolyl *cis*-trans isomerase, mitochondrial
DADR: D(1A) dopamine receptor
EBI1: ebv-induced G protein-coupled receptor
EGR2: early growth response protein 2, zinc finger transcription factor
EGR3: early growth response protein 3, zinc finger transcription factor
ERB3: ERBB-3 receptor protein-tyrosine kinase
ERG (HSERG11.PE1): Erg transcription factor (ets-related)
ERK3: extracellular signal-regulated kinase 3
FBRNP (S63912.D10S102): heterogeneous ribonucleoprotein homologue
FGR1: basic fibroblast growth factor receptor 1 (BFGF-R)
FLI1: oncogene, ErgB transcription factor
FLT3: stem cell tyrosine protein kinase
FMLR: FMet-Leu-Phe receptor (N-formyl peptide receptor)
GFAP: glial fibrillary acidic protein, astrocyte
GLVR1: leukemia virus receptor 1
GLVR2: leukemia virus receptor 2

GPR4: probable g protein-coupled receptor GPR4
GRAN: grancalcin
GRF1 (HSGRF1A.GRF-1): glucocorticoid receptor
GTR2: glucose transporter 2
GTR3: glucose transporter 3
HCK: tyrosine protein kinase
HPT2: haptoglobin-2
IL8A: interleukin 8 receptor A
IPKI: cAMP dependent protein kinase inhibitor
IRK5: G-protein-activated inward rectifier potassium channel 4
IRK7: G-protein-activated inward rectifier potassium channel 7
IRSP (S58544.PE1): 75kd infertility-related sperm protein
ITAB: platelet membrane glycoprotein IIB (integrin α-IIB)
ITAV: vitronectin receptor α (integrin α-V)
K1CM: keratin, type I cytoskeletal 13 (cytokeratin 13)
K1CN: keratin, type I cytoskeletal 14 (cytokeratin 14)
K1CO: keratin, type I cytoskeletal 15 (cytokeratin 15)
K1CT: keratin, type I cytoskeletal 20 (cytokeratin 20)
K22E: keratin, type II cytoskeletal 2 epidermal (cytokeratin 2E)
K22O: keratin, type II cytoskeletal 2 oral (cytokeratin 2P)
K2C1: keratin, type II cytoskeletal 1 (cytokeratin 1)
K2C4: keratin, type II cytoskeletal 4 (cytokeratin 4)
K2C5: keratin, type II cytoskeletal 5 (cytokeratin 5)
K2CD: keratin, type II cytoskeletal 6D (cytokeratin 6D)
KFMS: macrophage colony stimulating factor 1 receptor
KKIT: mast/stem cell growth factor receptor
LYN: tyrosine protein kinase
MLEF: myosin light chain I, embryonic muscle/atrial isoform
MLEV: myosin light chain I, slow-twitch muscle B/ventricular isoform
MLRM: myosin regulatory light chain 2, nonsarcomeric
MLRV: myosin regulatory light chain 2, ventricular/cardiac muscle isoform
MT1A: metallothionine-IA
MYBA: MYB-related protein A
MYBB: MYB-related protein B
NF66 (S78296.PE1): neurofilament-66
NFL: neurofilament triplet L protein
NFM: neurofilament triplet M protein
NTBE: sodium- and calcium-dependent betaine transporter
NTC1: notch 1
NTNO: sodium-dependent noradrenaline transporter
OPRM: mu-type opioid receptor
OXYR: oxytocin receptor
P190B (HS170321.P190-B): P190-B, member of the Rho GAP family
P2AB: serine/threonine protein phosphatase PP2A-β, catalytic subunit
P2BB: serine/threonine protein phosphatase PP2B-β, catalytic subunit
PAC1: protein-tyrosine phosphatase
PAX3: paired box protein 3, transcription factor
PAX5: paired box protein 5, transcription factor
PAXI: paxillin
PDGR: platelet-derived growth factor receptor
PER2: prostaglandin E2 receptor
PI2R: prostacyclin receptor
PKI (S76965.PE1): protein kinase inhibitor
PLMN: plasminogen
PP1G: serine/threonine protein phosphatase PP1-γ catalytic subunit
PRR (HSU41070.PE1): purinergic receptor
PTNA: MAP kinase phosphatase-1
PTNB: protein-tyrosine phosphatase 2C

PTPB: protein tyrosine phosphatase beta
PTPM: protein-tyrosine phosphatase mu
PTR2: parathyroid hormone receptor
RET: proto-oncogene tyrosine-protein kinase receptor
ROA1: heterogeneous nuclear ribonucleoprotein A1
RRB2: retinoic acid receptor β2
SDC2: syndecan-2, fibroglycan
SDC4: syndecan-4, amphiglycan
SORC: sorcin
SP1: sperm protamine P1, transcription factor
SP2: transcription factor
SP3: transcription factor
TAK1: orphan receptor
TFP1: tissue factor pathway inhibitor 1
TFP2: tissue factor pathway inhibitor 2
THA1: thyroid hormone receptor α1
TOM34 (HSU58970.PE1): putative translocase
TOPA: DNA topoisomerase IIα
TOPB: DNA topoisomerase IIβ
TR2: steroid receptor
TRKA: high affinity nerve growth factor receptor
TTNB (HS165.PE1): titin-associated protein
TYR1: tyrosinase-related protein 1
TYR2: tyrosinase-related protein 1
UBIL: ubiquitin-like protein GDX
UBIQ: ubiquitin
UBL1: ubiquitin carboxyl-terminal hydrolase isozyme L1
UBL3: ubiquitin carboxyl-terminal hydrolase isozyme L3
UFO: tyrosine-protein kinase receptor UFO
UROK: urokinase-type plasminogen activator
UROT: tissue plasminogen activator
V1AR: vasopressin V1α receptor
VGR1: vascular endothelial growth factor receptor 1
VGR2: vascular endothelial growth factor receptor 2
VIM: vimentin
VINC: vinculin
VWF: von Willebrand factor
ZN07: zinc finger protein 7
ZN38: zinc finger protein 38

Appendix B: Protein families defined by the 348 matches found on the NCBI dot-matrix plot

1: CCT chaperonin γ, t-complex: 4 members					
hsa.1.1708_hsa.2.6457	1[a]	2[b]	731[c]	274[d]	**327**[e]
hsa.1.1708_hsa.6.90941	1	6	731	711	249
hsa.2.6457_hsa.5.90710	2	5	274	28	282
hsa.5.90710_hsa.6.90941	5	6	28	711	270
2: RAB: 3 members					
hsa.1.92042_hsa.3.91661	1	3	1219	76	345
hsa.1.92042_hsa.8.6744	1	8	1219	268	**427**
hsa.3.91661_hsa.8.6744	3	8	76	268	274
3: ryanodine / inositol receptors: 5 members					
hsa.1.91610_hsa.3.91674	1	3	1240	14	280
hsa.1.91610_hsa.12.6246	1	12	1240	156	274
hsa.1.91610_hsa.19.1953.1	1	19	1240	213	273
hsa.1.91610_hsa.19.1953.2	1	19	1240	214	287
hsa.3.91674_hsa.12.6246	3	12	14	156	**1015**
4: receptor tyrosine kinases: 24 members					
hsa.1.90760_hsa.3.90849	1	3	473	621	262
hsa.1.90760_hsa.6.73978	1	6	473	487	263
hsa.1.90760_hsa.6.94333	1	6	473	165	253
hsa.1.90760_M13880	1	6	473	625	315
hsa.1.90760_hsa.8.90603	1	8	473	576	280
hsa.1.90760_hsa.8.92961	1	8	473	254	214
hsa.1.90760_hsa.9.47860	1	9	473	215	367
hsa.1.91102_hsa.9.94451	1	9	240	105	278
hsa.1.92093_hsa.3.90849	1	3	983	621	208
hsa.1.92093_hsa.6.94333	1	6	983	165	204
hsa.1.92093_hsa.9.47860	1	9	983	215	307
hsa.1.92093_hsa.10.90622	1	10	983	616	205
hsa.1.92093_U02687	1	13	983	28	209
hsa.1.94453_hsa.3.2913	1	3	180	898	566
hsa.1.94453_hsa.6.73978	1	6	180	487	685
hsa.1.94453_M13880	1	6	180	625	218
hsa.1.94453_hsa.7.90579	1	7	180	599	446
hsa.1.94453_hsa.7.94304	1	7	180	346	209
hsa.1.94453_hsa.8.90603	1	8	180	576	269
hsa.1.94453_hsa.8.92961	1	8	180	254	250
hsa.1.94453_hsa.10.90622	1	10	180	616	204
hsa.1.94453_D00333	1	22	180	31	209
hsa.1.94706_hsa.8.90603	1	8	137	576	214
hsa.1.94706_hsa.8.92961	1	8	137	254	373

[a] chromosomal location of the first listed gene of the pair

[b] chromosomal location of the second listed gene of the pair

[c] gene order position of first listed gene of the pair

[d] gene order position of second listed gene of the pair

[e] TBLASTX score for that gene pair. The highest scoring match for each gene family is highlighted in bold.

hsa.1.94706_hsa.8.94657	1	8	137	44	259
hsa.1.94706_D00333	1	22	137	31	397
hsa.3.2913_hsa.6.73978	3	6	898	487	739
hsa.3.2913_M13880	3	6	898	625	243
hsa.3.2913_hsa.7.90579	3	7	898	599	530
hsa.3.2913_hsa.7.91937	3	7	898	787	233
hsa.3.2913_hsa.7.94304	3	7	898	346	250
hsa.3.2913_hsa.8.90603	3	8	898	576	292
hsa.3.90849_hsa.4.1751	3	4	621	403	206
hsa.4.1751_hsa.6.94333	4	6	403	165	341
hsa.4.1751_M13880	4	6	403	625	245
hsa.4.1751_hsa.9.47860	4	9	403	215	238
hsa.4.1751_hsa.10.90622	4	10	403	616	415
hsa.4.1751_hsa.13.92928	4	13	403	30	**919**
hsa.4.1751_U02687	4	13	403	28	536
hsa.6.73978_hsa.7.90579	6	7	487	599	325
hsa.6.73978_hsa.7.94304	6	7	487	346	262
hsa.6.73978_hsa.8.90603	6	8	487	576	285
hsa.6.73978_hsa.8.92961	6	8	487	254	269
hsa.6.73978_hsa.10.90622	6	10	487	616	258
hsa.6.94333_hsa.7.94304	6	7	165	346	224
hsa.6.94333_hsa.9.47860	6	9	165	215	232
hsa.6.94333_hsa.10.90622	6	10	165	616	393
hsa.6.94333_hsa.13.92928	6	13	165	30	337
hsa.6.94333_U02687	6	13	165	28	330
M13880_hsa.8.90603	6	8	625	576	271
M13880_hsa.8.92961	6	8	625	254	269
M13880_hsa.9.47860	6	9	625	215	265
M13880_hsa.13.92928	6	13	625	30	247
hsa.7.91937_hsa.8.92961	7	8	787	254	258
hsa.7.91937_hsa.9.94451	7	9	787	105	240
hsa.7.91937_hsa.10.90622	7	10	787	616	252
hsa.7.94304_hsa.8.90603	7	8	346	576	272
hsa.7.94304_hsa.9.94451	7	9	346	105	246
hsa.7.94304_hsa.10.90622	7	10	346	616	308
hsa.7.94304_hsa.12.96942	7	12	346	265	292
hsa.8.90603_hsa.9.94451	8	9	576	105	250
hsa.8.90603_hsa.10.90622	8	10	576	616	252
hsa.8.92961_hsa.9.94451	8	9	254	105	266
hsa.8.92961_hsa.20.94134	8	20	254	149	280
hsa.8.92961_D00333	8	22	254	31	537
hsa.9.47860_hsa.10.90622	9	10	215	616	267
hsa.9.47860_hsa.13.92928	9	13	215	30	248
hsa.9.47860_U02687	9	13	215	28	234
hsa.9.94451_hsa.10.90622	9	10	105	616	319
hsa.10.90622_hsa.13.92928	10	13	616	30	426
hsa.10.90622_U02687	10	13	616	28	397
5: glucose transporters: 2 members					
hsa.1.90693_hsa.3.92393	1	3	236	835	**281**
6: diacylglycerol kinases: 2 members					
hsa.1.91023_hsa.3.7140	1	3	287	890	**203**
7: calcium ATPases: 3 members					
hsa.1.995_hsa.3.90662	1	3	807	41	**978**
hsa.1.995_hsa.12.33170	1	12	807	393	665
hsa.3.90662_hsa.12.33170	3	12	41	393	767
8: E2F-related transcription factor, DP-2: 2 members					
hsa.1.1648_hsa.3.94714	1	3	889	675	**606**

9: RXRG, hap, steroid receptor TR2, MLR, GLR (mineralocorticoid, glucocorticoid receptor): 6 members					
hsa.1.26550_hsa.3.2712	1	3	974	92	267
hsa.1.26550_hsa.4.94291	1	4	974	664	202
hsa.1.26550_hsa.12.1084	1	12	974	500	289
hsa.3.2712_hsa.12.1084	3	12	92	500	291
hsa.3.520_hsa.12.1084	3	12	71	500	391
hsa.4.94291_hsa.5.94327	4	5	664	531	**638**
10: glutathione transferase, ABL?: 2 members					
hsa.1.90476.1_hsa.4.90005c.17	1	4	616	63	328
hsa.1.90476.1_hsa.9.1208.6	1	9	616	501	**468**
11: phosphodiestaerases: 2 members					
hsa.1.188_hsa.5.190	1	5	373	205	**1205**
12: thrombospondin: 3 members					
hsa.1.33454_hsa.5.92976	1	5	865	325	**499**
hsa.1.33454_hsa.15.94294	1	15	865	72	366
hsa.5.92976_hsa.15.94294	5	15	325	72	406
13: protein tyrosine phosphatases: 3 members					
hsa.1.92992_hsa.6.93152	1	6	141	386	560
hsa.1.92992_hsa.17.30043.1	1	17	141	255	**683**
hsa.6.93152_hsa.17.30043.1	6	17	386	255	448
14: protein kinases: 9 members					
hsa.1.1903_hsa.6.91920	1	6	441	688	282
hsa.1.1903_hsa.10.22071	1	10	441	30	215
hsa.1.1903_hsa.14.1880	1	14	441	188	231
hsa.1.1903_hsa.17.60762	1	17	441	364	253
hsa.1.1903_hsa.17.965	1	17	441	369	291
hsa.1.2079_hsa.17.965	1	17	148	369	454
hsa.1.69171_hsa.6.91920	1	6	457	688	257
hsa.1.69171_hsa.14.1880	1	14	457	188	263
hsa.1.69171_hsa.17.60762	1	17	457	364	270
hsa.1.69171_hsa.17.965	1	17	457	369	223
hsa.3.1904_hsa.6.91920	3	6	769	688	499
hsa.3.1904_hsa.10.22071	3	10	769	30	304
hsa.3.1904_hsa.14.1880	3	14	769	188	552
hsa.3.1904_hsa.17.60762	3	17	769	364	548
hsa.3.1904_hsa.17.965	3	17	769	369	422
hsa.6.91920_hsa.10.22071	6	10	688	30	425
hsa.6.91920_hsa.14.1880	6	14	688	188	555
hsa.6.91920_hsa.17.60762	6	17	688	364	649
hsa.6.91920_hsa.17.965	6	17	688	369	575
hsa.10.22071_hsa.14.1880	10	14	30	188	549
hsa.10.22071_hsa.17.60762	10	17	30	364	591
hsa.10.22071_hsa.17.965	10	17	30	369	288
hsa.14.1880_hsa.17.60762	14	17	188	364	**767**
hsa.14.1880_hsa.17.965	14	17	188	369	470
15: protein tyrosine phosphatases: 8 members					
hsa.1.2311_hsa.7.91890	1	7	245	832	241
hsa.1.2311_hsa.9.21593	1	9	245	39	802
hsa.6.91790_hsa.7.90530	6	7	659	460	262
hsa.6.91790_hsa.7.91890	6	7	659	832	245
hsa.6.91790_hsa.9.21593	6	9	659	39	416
hsa.6.91790_hsa.15.1924	6	15	659	285	234
hsa.6.91790_hsa.18.91089	6	18	659	34	**917**
hsa.6.91790_hsa.20.91769	6	20	659	8	247
hsa.7.90530_hsa.9.21593	7	9	460	39	258
hsa.7.90530_hsa.15.1924	7	15	460	285	266
hsa.7.90530_hsa.18.91089	7	18	460	34	276

hsa.7.91890_hsa.9.21593	7	9	832	39	363
hsa.7.91890_hsa.20.91769	7	20	832	8	301
hsa.9.21593_hsa.15.1924	9	15	39	285	270
hsa.9.21593_hsa.18.91089	9	18	39	34	403
hsa.9.21593_hsa.20.91769	9	20	39	8	579
hsa.15.1924_hsa.18.91089	15	18	285	34	251
hsa.15.1924_hsa.20.91769	15	20	285	8	215
hsa.18.91089_hsa.20.91769	18	20	34	8	268
16: collagens: 4 members					
hsa.1.92949_hsa.7.90671	1	7	533	601	213
hsa.2.90787_hsa.7.90671	2	7	813	601	**445**
hsa.2.90787_hsa.12.3231.2	2	12	813	201	297
hsa.7.90671_hsa.12.3231.2	7	12	601	201	277
17: GNA: 5 members					
hsa.1.73810_hsa.7.92105	1	7	562	481	333
hsa.7.92105_L22075	7	17	481	416	429
hsa.7.92105_hsa.18.92200	7	18	481	42	459
hsa.7.92105_hsa.22.90695.4	7	22	481	35	468
hsa.7.92105_hsa.22.90695.6	7	22	481	37	**470**
L22075_hsa.18.92200	17	18	416	42	379
L22075_hsa.22.90695.4	17	22	416	35	289
hsa.18.92200_hsa.22.90695.4	18	22	42	35	313
18: ?: 2 members					
hsa.1.93133_hsa.7.31404	1	7	788	765	**227**
19: SAP-1B, ETV1 (ets translocation unit): 3 members					
hsa.1.200_hsa.7.91281	1	7	969	76	248
hsa.1.200_hsa.21.78	1	21	969	38	**254**
hsa.7.91281_hsa.21.78	7	21	76	38	214
20: α-tubulins: 2 members					
hsa.1.64066_hsa.11.5700	1	11	197	579	**495**
21: ?: 2 members					
hsa.1.17075_hsa.12.90374	1	12	209	0	**501**
22: N-ras, ?: 2 members					
hsa.1.93085_hsa.12.56474	1	12	587	326	**424**
23: PP2A/ PP2B: 2 members					
hsa.1.91154_hsa.14.3106	1	14	1152	441	**1267**
24: actins: 2 members					
hsa.1.94446_hsa.14.90816	1	14	1124	230	**484**
25: ADP-ribosylation factor: 2 members					
hsa.1.44591_hsa.22.21019	1	22	1002	179	**386**
26: hormone receptors: 3 members					
hsa.2.1080_hsa.4.942	2	4	325	480	261
hsa.2.1080_hsa.8.36998	2	8	325	498	**520**
27: ?: 2 members					
hsa.2.3642_hsa.6.30167	2	6	327	441	**449**
28: ?: 2 members					
hsa.2.35397_hsa.6.2295	2	6	468	769	**467**
29: CDC10 homologues: 2 members					
hsa.2.91131_hsa.7.91088	2	7	1053	200	**267**
30: catenins: 2 members					
hsa.2.75616_hsa.7.91999	2	7	368	515	**711**
31: grancalcin, sorcin: 2 members					
hsa.2.91721_hsa.7.37884	2	7	735	526	**327**

32: annexin, lipocortin, synexin: 6 members					
hsa.2.90634_hsa.8.2776	2	8	370	544	246
hsa.2.90634_hsa.9.94867	2	9	370	202	**688**
hsa.2.90634_hsa.10.91642	2	10	370	322	477
hsa.2.90634_hsa.15.38368	2	15	370	183	459
hsa.2.94620_hsa.10.91642	2	10	433	322	568
hsa.8.2776_hsa.10.91642	8	10	544	322	320
hsa.9.94867_hsa.10.91642	9	10	202	322	469
hsa.9.94867_hsa.15.38368	9	15	202	183	397
33: stomatin: 2 members					
hsa.2.73922_hsa.9.185807	2	9	361	402	**560**
34: titin, slow MyBP-C: 3 members					
hsa.2.90877.1_hsa.9.90876	2	9	718	149	206
hsa.2.90877.2_hsa.18.2504	2	18	719	15	**225**
35: γ-actin: 2 members					
hsa.2.58189_hsa.10.665	2	10	709	475	**532**
36: ?: 2 members					
hsa.2.291209_hsa.13.90000.23	2	13	1005	61	**414**
37: serotonin receptors: 2 members					
hsa.2.90814_hsa.13.97047	2	13	959	155	**507**
38: ß-spectrins: 2 members					
hsa.2.90762_hsa.14.47431	2	14	260	240	**828**
39: 90 kDa heat shock proteins: 2 members					
hsa.3.91667_hsa.4.8533	3	4	727	122	**549**
40: zinc finger proteins: 20 members					
hsa.3.94682_hsa.4.71666	3	4	545	767	231
hsa.3.94682_hsa.6.73099	3	6	545	689	240
hsa.3.94682_hsa.7.90573	3	7	545	643	219
hsa.3.94682_hsa.9.822	3	9	545	147	221
hsa.3.94682_hsa.10.92024	3	10	545	168	213
hsa.3.94682_hsa.18.91147	3	18	545	124	207
L32164_hsa.7.90573	3	7	216	643	240
L32164_hsa.8.74769	3	8	216	40	236
L32164_hsa.9.822	3	9	216	147	216
L32164_hsa.19.2862	3	19	216	203	233
L32164_hsa.19.55503	3	19	216	275	207
L32164_hsa.19.92142	3	19	216	208	209
L32164_M29581	3	19	216	321	217
L32164_X78933	3	19	216	173	233
hsa.4.71666_hsa.6.1990	4	6	767	209	590
hsa.4.71666_hsa.6.73099	4	6	767	689	398
hsa.4.71666_hsa.7.90573	4	7	767	643	640
hsa.4.71666_hsa.7.90902	4	7	767	738	272
hsa.4.71666_hsa.9.822	4	9	767	147	930
hsa.4.71666_hsa.10.15971	4	10	767	152	209
hsa.4.71666_hsa.10.92024	4	10	767	168	611
hsa.4.71666_hsa.18.91147	4	18	767	124	396
hsa.6.1990_hsa.7.90573	6	7	209	643	482
hsa.6.1990_hsa.7.90902	6	7	209	738	257
hsa.6.1990_hsa.10.92024	6	10	209	168	476
hsa.6.1990_hsa.18.91147	6	18	209	124	349
hsa.6.1990_hsa.19.24148	6	19	209	325	218
hsa.6.1990_hsa.19.2862	6	19	209	203	556
hsa.6.1990_M29581	6	19	209	321	229
hsa.6.1990_U09412	6	19	209	287	521
hsa.6.1990_X78933	6	19	209	173	621
hsa.6.73099_hsa.7.90573	6	7	689	643	392
hsa.6.73099_hsa.7.90902	6	7	689	738	229

hsa.6.73099_hsa.9.822	6	9	689	147	365
hsa.6.73099_hsa.10.92024	6	10	689	168	359
hsa.6.73099_hsa.17.90571	6	17	689	319	353
hsa.6.73099_hsa.18.91147	6	18	689	124	332
hsa.7.90573_hsa.9.822	7	9	643	147	585
hsa.7.90573_hsa.10.15971	7	10	643	152	238
hsa.7.90573_hsa.10.33898	7	10	643	185	217
hsa.7.90573_hsa.10.92024	7	10	643	168	531
hsa.7.90573_hsa.18.91147	7	18	643	124	438
hsa.7.90902_hsa.8.74769	7	8	738	40	274
hsa.7.90902_hsa.9.822	7	9	738	147	258
hsa.7.90902_hsa.10.92024	7	10	738	168	249
hsa.7.90902_hsa.17.90571	7	17	738	319	261
hsa.7.90902_hsa.18.91147	7	18	738	124	274
U09847_hsa.9.822	7	9	354	147	699
hsa.8.74769_hsa.9.822	8	9	40	147	**1419**
hsa.8.74769_hsa.10.15971	8	10	40	152	233
hsa.8.74769_hsa.10.33898	8	10	40	185	222
hsa.8.74769_hsa.10.92024	8	10	40	168	601
hsa.8.74769_hsa.17.90571	8	17	40	319	858
hsa.8.74769_hsa.18.91147	8	18	40	124	405
hsa.9.822_hsa.10.15971	9	10	147	152	213
hsa.9.822_hsa.10.92024	9	10	147	168	573
hsa.9.822_hsa.18.91147	9	18	147	124	391
hsa.10.15971_hsa.17.90571	10	17	152	319	233
hsa.10.15971_hsa.18.91147	10	18	152	124	208
hsa.10.15971_hsa.19.2862	10	19	152	203	207
hsa.10.15971_hsa.19.55503	10	19	152	275	209
hsa.10.92024_hsa.18.91147	10	18	168	124	376
hsa.17.90571_hsa.18.91147	17	18	319	124	380
41: ARF/ARL: 3 members					
hsa.3.1520_hsa.5.792	3	5	375	227	438
hsa.3.1520_hsa.12.1971	3	12	375	544	**516**
hsa.5.792_hsa.12.1971	5	12	227	544	393
42: RII-A, PRXAR2A: 2 members					
hsa.3.2574_hsa.7.90757	3	7	238	722	**348**
43: integrins: 2 members					
hsa.3.92953_hsa.7.832	3	7	544	110	**254**
44: nuclear ribonucleoproteins: 2 members					
hsa.3.37495_hsa.10.96826	3	10	350	285	**350**
45: hormone receptors: 2 members					
hsa.3.91944_hsa.10.724	3	10	96	510	**331**
46: CBL: 2 members					
hsa.3.90894_hsa.11.92108	3	11	504	617	**1065**
47: calcium channel proteins: 2 members					
hsa.3.23838.1_hsa.12.88	3	12	374	18	**320**
48: angiotensin II type 1, anaphylatoxin C3A receptor: 3 members					
hsa.3.338_hsa.12.91182	3	12	650	69	220
hsa.12.91182_hsa.19.251	12	19	69	304	**376**
49: ?: 2 members					
hsa.3.90007.2_hsa.12.91696	3	12	313	501	**263**
50: dihydropyriminidase: 3 members					
hsa.4.91025_hsa.5.4778	4	5	113	621	890
hsa.4.91025_hsa.8.93002	4	8	113	117	1037
hsa.5.4778_hsa.8.93002	5	8	621	117	**1372**

51: GABRA/B: 5 members					
hsa.4.91984_hsa.5.45740	4	5	297	700	364
hsa.4.2677_hsa.15.91739	4	15	293	4	209
hsa.4.2677_hsa.15.94804	4	15	293	8	305
hsa.4.91984_hsa.15.91739	4	15	297	4	**860**
hsa.5.45740_hsa.15.91739	5	15	700	4	764
52: cyclin-dependent kinase: 2 members					
hsa.4.21029_hsa.6.94350	4	6	181	168	348
hsa.4.91984_hsa.5.45740	4	5	297	700	**364**
53: ankyrins: 2 members					
hsa.4.90532_hsa.8.94224	4	8	660	215	**478**
hsa.4.91984_hsa.5.45740	4	5	297	700	364
54: ?: 3 members					
hsa.4.35804_hsa.10.91975	4	10	426	201	**334**
hsa.4.35804_hsa.15.91676	4	15	426	22	292
55: JNK: 2 members					
hsa.4.90774_hsa.10.90620	4	10	473	191	**1324**
56: calcineurin, 2 members					
hsa.4.92_hsa.10.94566	4	10	582	324	**2087**
57: globins: 3 members					
hsa.4.39973.3_hsa.11.33958	4	11	460	55	322
hsa.4.39973.3_hsa.11.92304	4	11	460	54	**2114**
hsa.4.39973.4_hsa.11.92304	4	11	461	54	285
58: survival motor neuron, ?: 5 members					
hsa.4.90005c.15_hsa.16.90013.10	4	16	61	9	326
hsa.4.90005c.15_hsa.16.90013.12	4	16	61	11	304
hsa.5.2263.4_hsa.22.10741	5	22	251	177	279
hsa.5.2263.5_hsa.16.90013.12	5	16	252	11	338
hsa.5.2263.5_hsa.17.3783.4	5	17	252	242	280
hsa.16.90013.10_hsa.17.3783.4	16	17	9	242	313
hsa.16.90013.10_hsa.22.10741	16	22	9	177	314
hsa.16.90013.12_hsa.17.3783.4	16	17	11	242	303
hsa.16.90013.12_hsa.22.10741	16	22	11	177	**418**
hsa.17.3783.4_hsa.22.10741	17	22	242	177	313
59: GABR: 2 members					
M62400_M86868	5	6	609	504	**559**
60: UBCH5B, ?: 2 members					
hsa.5.32690_hsa.7.19196	5	7	555	264	**329**
61: UTH5P75, BiP: 2 members					
hsa.5.62218_hsa.9.5329	5	9	488	460	**263**
62: serine/threonine kinases: 2 members					
hsa.5.74035_hsa.12.94730	5	12	246	261	**432**
63: ?: 2 members					
hsa.5.34459_hsa.13.53258.6	5	13	727	174	**1088**
64: ras GTPase activating: 2 members					
hsa.5.93122_hsa.15.90737	5	15	294	341	**727**
65: cadherins: 5 members					
hsa.5.21552_hsa.16.91364	5	16	77	274	224
hsa.5.47075_hsa.16.90828	5	16	73	276	202
hsa.5.47075_hsa.16.91364	5	16	73	274	**668**
hsa.5.90593_hsa.16.91364	5	16	499	274	440
66: adenyl cyclase: 2 members					
hsa.5.2352_hsa.16.2454	5	16	19	250	**424**
67: tenascin, hexabrachion: 2 members					
hsa.6.90006.8_hsa.9.92049	6	9	250	416	**441**
68: ?: 2 members					
hsa.6.1537_hsa.11.3105	6	11	275	608	**392**

69: RING3 (female sterile homeotic (fsh)) homologs: 2 members					
hsa.6.91463_hsa.11.90865	6	11	466	517	**428**
70: thyroid receptor interactor, ?: 2 members					
hsa.6.10063_hsa.13.2212	6	13	467	374	**513**
71: ADP-ribosylation factor, M phase phosphoprotein 1, ?: 6 members					
hsa.6.90008.4_hsa.13.90000.27	6	13	231	65	**331**
hsa.7.90011.3_L16782	7	10	888	430	215
hsa.7.90011.3_hsa.13.90000.27	7	13	888	65	212
hsa.7.90011.3_hsa.13.90000.44	7	13	888	82	221
L16782_hsa.13.53258.3	10	13	430	171	258
72: proteasome subunits: 2 members					
hsa.6.38291.6_hsa.14.1169	6	14	270	29	**237**
73: SNRPB1, FBRNP: 2 members					
hsa.7.1036_hsa.10.62267	7	10	158	182	**298**
74: ?: 2 members					
hsa.7.91732_hsa.10.210144	7	10	312	454	**371**
75: ?: 2 members					
hsa.7.90011.18_hsa.11.90105.1	7	11	903	187	**292**
76: transcription factors SP1: 2 members					
X68561_hsa.12.2021	7	12	115	223	**623**
77: ?, neuroendocrine-specific protein A: 2 members					
hsa.7.6569_hsa.14.1834	7	14	198	184	**677**
78: T-cell receptors: 2 members					
hsa.7.46324.19_X57613	7	18	1033	281	**510**
79: ?: 3 members					
hsa.7.33043.2_Z68223.1	7	22	587	96	**389**
hsa.13.90000.6_Z68223.2	13	22	44	97	235
80: ?: 2 members					
hsa.8.57722_hsa.12.57750	8	12	212	533	**217**
81: AML1: 2 members					
hsa.8.91891_hsa.21.91851	8	21	373	117	**1164**
82: DAP-kinases: 2 members					
hsa.9.239476_hsa.12.90835	9	12	266	207	**378**
83: tyrosinases: 3 members					
hsa.9.2282_hsa.13.94721	9	13	85	317	**355**
hsa.11.2053_hsa.13.94721	11	13	522	317	220
84: transcription factors B5AP, PAX3: 2 members					
hsa.9.69959_hsa.13.92926	9	13	155	141	**313**
85: BAF60A, GPD1 (glycerol-3-phosphate dehydrogenase): 2 members					
hsa.10.91682_M36917	10	12	450	213	**289**
86: G1-WPI interferon inducible, ?: 2 members					
hsa.10.92076_hsa.13.90000.16	10	13	386	54	**300**
87: MRP (multidrug resistance proteins): 2 members					
hsa.10.91923_hsa.16.92075	10	16	486	134	**798**
88: ?, contactin: 2 members					
hsa.11.30935_hsa.12.21728	11	12	562	187	**231**
89: glycogen phosphorylases: 2 members					
hsa.11.46315_hsa.14.771	11	14	334	136	**236**
90: radixin, moesin B: 2 members					
hsa.11.1028_hsa.15.90442	11	15	640	178	**528**
91: casein kinase CK1, intergenic region: 2 members					
hsa.11.91020_hsa.15.676.2	11	15	279	180	**509**
92: casein kinase 2α subunits: 2 members					
hsa.11.92934_hsa.16.90980	11	16	88	300	**721**
93: N-methyl-D-aspartate receptor subunits: 2 members					
U11287_hsa.16.12368	12	16	119	98	**753**
94: heat stable enterotoxin receptor, retinal guanylyl cyclase: 2 members					
M73489_M92432	12	17	238	64	**450**

95: myosin light chains: 2 members					
hsa.12.74106_hsa.18.74102	12	18	580	16	**361**
96: L21 ribosomal protein, BRCA1: 3 members					
hsa.13.3309_hsa.17.3783.3	13	17	24	241	**447**
hsa.13.3309_hsa.20.9160	13	20	24	226	428
97: endothelin-B receptors: 2 members					
hsa.13.24_hsa.18.23	13	18	259	4	**284**
98: myosin heavy chains: 3 members					
hsa.14.929_hsa.17.92041	14	17	22	71	**267**
hsa.14.94231_hsa.17.92041	14	17	327	71	258
99: ?, actin depolymerizing factor: 2 members					
hsa.14.93051_hsa.20.64919	14	20	97	103	**277**
100: NEDD-4: 2 members					
hsa.15.1565_hsa.18.3550	15	18	129	241	**301**
101: IGF1 receptor, INSR (insulin receptor): 2 members					
hsa.15.94359_hsa.19.5929	15	19	394	35	**341**
102: haptoglobin, ?: 2 members					
hsa.16.1495.2_hsa.19.96962	16	19	345	46	**237**