Chapter 1

Introduction

1.1 Genome evolution

Genomes have evolved and increased in complexity owing to a number of evolutionary processes acting upon them, such as insertions, deletions and inversions. Gene duplications are also believed to have played a major role in the evolution and development of vertebrate genomes. Susumu Ohno (1970) first suggested that the increase in organismal complexity during vertebrate evolution could only have occurred if there was a considerable increase in gene number and proposed that this happened by the duplication of entire genomes in a process termed polyploidisation.

When Ohno first proposed the theory of polyploidisation it generated a lot of excitement and outrage in the field of genetics, but, by the late 1980s, many had lost interest owing to the lack of evidence. With the expansion of genomic information generated during the 1990s, duplicated genes and chromosomal regions were identified in the human, and other genomes, and the theory became popular again although it remains controversial. Duplicated genes and regions are believed by some to represent remnants of whole-genome duplication events, whilst others have argued that they are the result of the duplication of chromosomal regions or of individual genes brought together by selective forces (reviewed by Wolfe, 2001; Lundin *et al*, 2003; Hughes and Friedman, 2003).

1.2 Homologues, paralogues and orthologues

Three definitions are commonly used to describe the relationship between genes: homologues, paralogues and orthologues (reviewed by Sharman, 1999). Homologous genes are members of the same family or superfamily and share a common ancestor at some point back in evolutionary time. Homologues can be further subdivided into two groups; orthologues, genes that have been separated by speciation, and paralogues, genes that have resulted from a duplication event. Orthologues can be traced by descent to the common ancestor of two organisms and will both encode equivalent evolutionarily conserved proteins. Paralogues, however, are genes within the same species that have originated through duplication of an ancestral gene; whether as part of a whole genome, chromosomal segment or a single gene duplication event. The evolutionary fate of paralogues and orthologues are very different. Orthologues often take over the function of the precursor gene in the species of origin and thus tend to be conserved. In contrast, young paralogues have redundant functions, which are an evolutionary unstable situation, thus, in the long run – with a few exceptions – paralogues either diverge functionally, or all but one of the versions are lost.

1.3 Paralogous genes and the evolution of the human genome

Paralogous genes have been identified throughout the human genome. Ohno (1973) identified duplicated chromosomal segments within the human genome containing two pairs of duplicate genes on chromosomes 11 and 12, which he proposed as being evidence of polyploidisation. In the 1990s, molecular mapping data was used to identify a number of chromosomal regions containing clusters of paralogues in the human and mouse genomes that were believed to be remnants of genome duplication

events (termed paralogous regions; Lundin, 1993).

Intriguingly, the number of paralogous regions and paralogous genes investigated at the time was generally four (this phenomenon was termed tetralogy), or less, suggesting that at least two rounds of large-scale block or genome duplications have occurred during the course of mammalian evolution. For example, Spring and co-workers (1994) found that vertebrates have four copies of a gene for a cell-surface protein called syndecan, whereas the fruit fly *Drosophila* has only one. More than fifty examples of this so-called 1-to-4 gene rule have now been identified (Spring, 1997). Independently, Sidow (1996) observed the 1-to-4 gene rule during phylogenetic and sequence surveys of developmental regulator families, in which he concluded that two large-scale gene duplication events, most likely of entire genomes, occurred in an ancestor of vertebrates (Sidow, 1996).

Ohno (1970) originally suggested that there were large-scale gene duplication events, possibly involving the whole genome, in early chordates; specifically on the lineage leading to both cephalochordates (including amphioxus) and vertebrates (including hagfish, lampreys and jawed vertebrates). He also suggested a second, and maybe a third, large-scale duplication event at the time of fish or amphibian divergence. The number of duplications and the mechanisms involved have been heavily debated, and many modifications of Ohno's model have been proposed. For example, Holland and co-workers (1994) proposed that there were two phases of duplication on the vertebrate lineage, but suggested that the first duplication occurred on the vertebrate lineage after divergence of the amphioxus lineage, and the second on the jawed vertebrate lineage after the divergence of jawless fish. Kasahara and colleagues (1996) proposed that two polyploidisation events occurred later in the vertebrate

lineage, after the divergence of lampreys. The most popular version of events has been termed the 2R hypothesis as it involved two rounds of polyploidisations; one prior to the divergence of agnatha (jawless fish, exemplified by lampreys and hagfish) and gnathostomata (jawed vertebrates), while the second occurred after the divergence of agnatha but before the divergence of chondryichthyes (cartilaginous fish) (Sidow, 1996). This is simplified in figure 1.1.

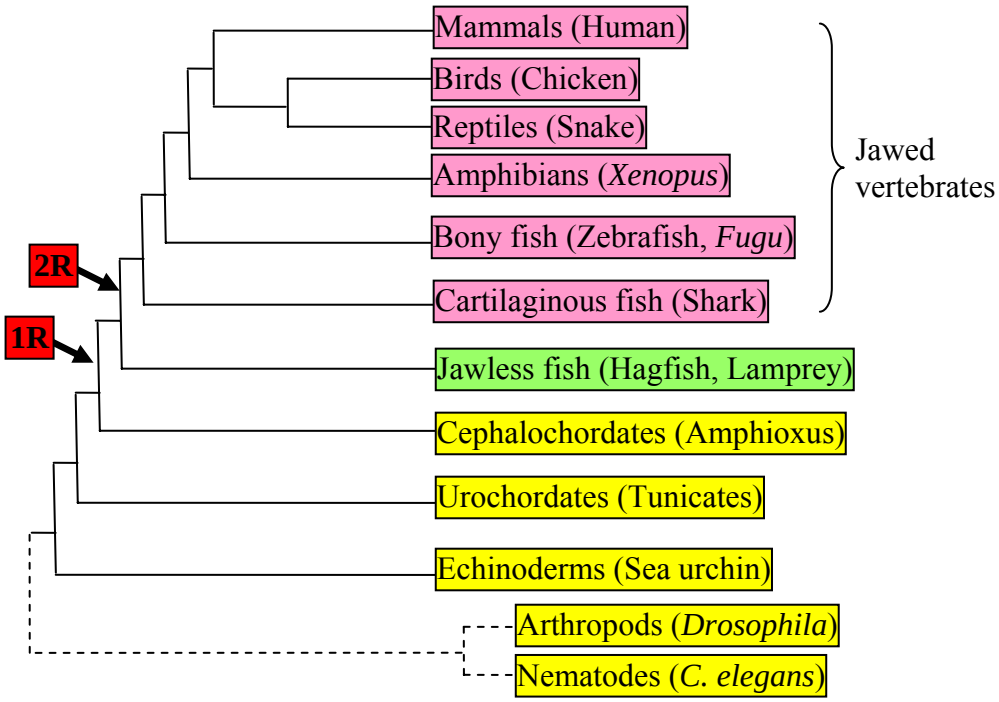

Figure 1.1 The 2R hypothesis. The two rounds of duplication are indicated by arrows. 1R corresponds to the first round of whole-genome duplication, after the emergence of amphioxus, and 2R corresponds to the second round of whole-genome duplication prior to the emergence of jawed vertebrates, more specifically cartilaginous fish.

The four Hox gene clusters in the human genome exemplify the 2R hypothesis (figure 1.2). The homoeotic complex (HOM-C) occurs as a single cluster in invertebrates such as *Drosophila, Caenorhabditis elegans* and amphioxus, but is found as four paralogous Hox gene clusters in vertebrates like mice and humans (Schughart *et al*,

1988). Interestingly, not only was the order of genes in the mammalian Hox clusters found to be conserved between human and mouse, but it was also conserved among the four mammalian clusters. The quadruplication of the Hox genes and the discovery of other paralogous genes linked to the Hox clusters provide evidence to support the involvement of large-scale chromosomal or whole-genome duplications in the evolution of vertebrate genomes (Larhammar *et al*, 2002).

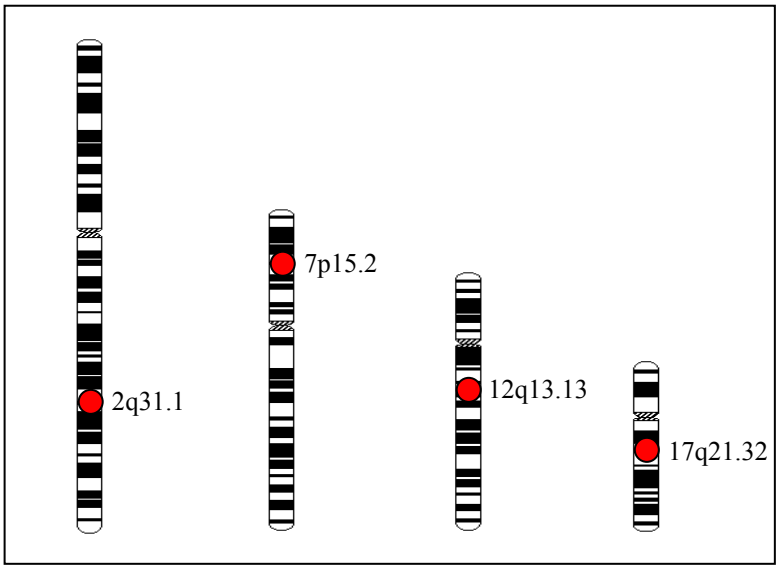

Figure 1.2 Distribution of Hox gene clusters in the human genome (represented by the red circles).

The 2R hypothesis is controversial and continues to be heavily discussed in the literature (reviewed by Wolfe, 2001). It was widely believed that the debating over the evolution of the human genome would be resolved once the entire human genome sequence was available. However, the initial analysis of the draft human genome sequence did not reveal overwhelming evidence for tetralogy and the 2R hypothesis remains controversial (International Human Genome Sequencing Consortium (IHGSC), 2001; Venter *et al*, 2001).

1.4 Genome sequencing projects

Between 1977 and 1982 the genomes of the bacterial virus ΦX174 (Sanger *et al*, 1977a, 1978), bacteriophage lambda (Sanger *et al*, 1982), animal virus SV40 (Fiers *et al*, 1978) and the human mitochondrion (Anderson *et al*, 1981) were successfully sequenced and assembled. During the early 1990s, the genomes of the yeast *Saccharomyces cerevisiae* (Oliver *et al*, 1992) and the nematode worm *Caenorhabditis elegans* (Wilson *et al*, 1994) were sequenced, thus demonstrating the feasibility of large-scale genome sequencing. By September 2003, the sequencing of 160 genomes had been completed, with 393 prokaryotic and 242 eukaryotic genome-sequencing projects still ongoing (http://igweb.integratedgenomics.com/GOLD/). The time-line of a number of genome sequencing projects is shown in figure 1.3.

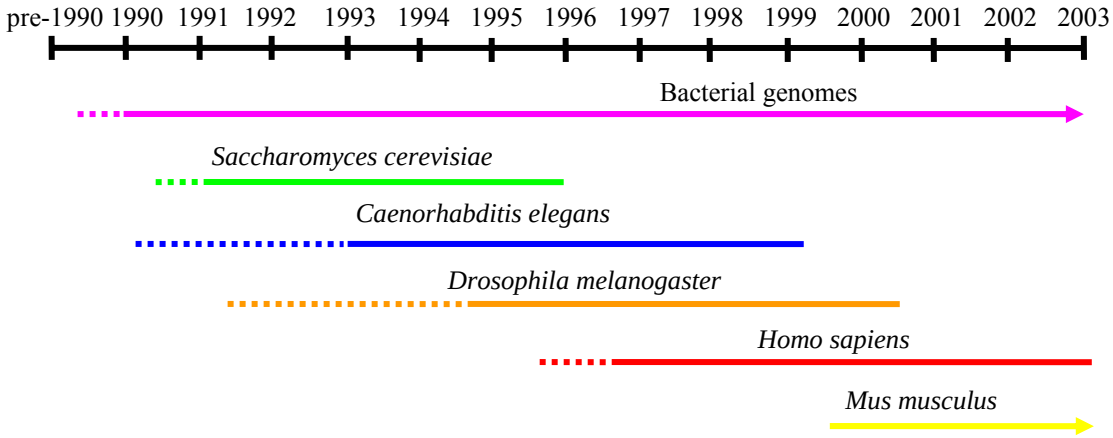

Figure 1.3 Time-line of a range of genome sequencing projects. The arrows signify ongoing sequencing projects.

1.5 The Human Genome Project

The Human Genome Project (HGP) was established in 1990 with the aim of

sequencing the entire human genome by 2005. In 1999, the year I started this project, the HGP effort moved into full-scale production, and the overall sequencing output increased significantly (figure 1.4). By 2000, the 'draft' human sequence was completed consisting of mainly 'unfinished' sequence covering approximately 90% of the human genome. Two 'draft' sequences were published by separate organisations (IHGSC, 2001; Venter *et al*, 2001) offering the chance to compare the genomic data produced. The data generated by the International Human Genome Sequencing Consortium (IHGSC) was a collaborative effort involving 20 groups from around the world. Venter and colleagues were part of the biotechnology company Celera Genomics which was formed in 1998. The completion of the HGP was announced by the IHGSC in 2003, two years ahead of schedule.

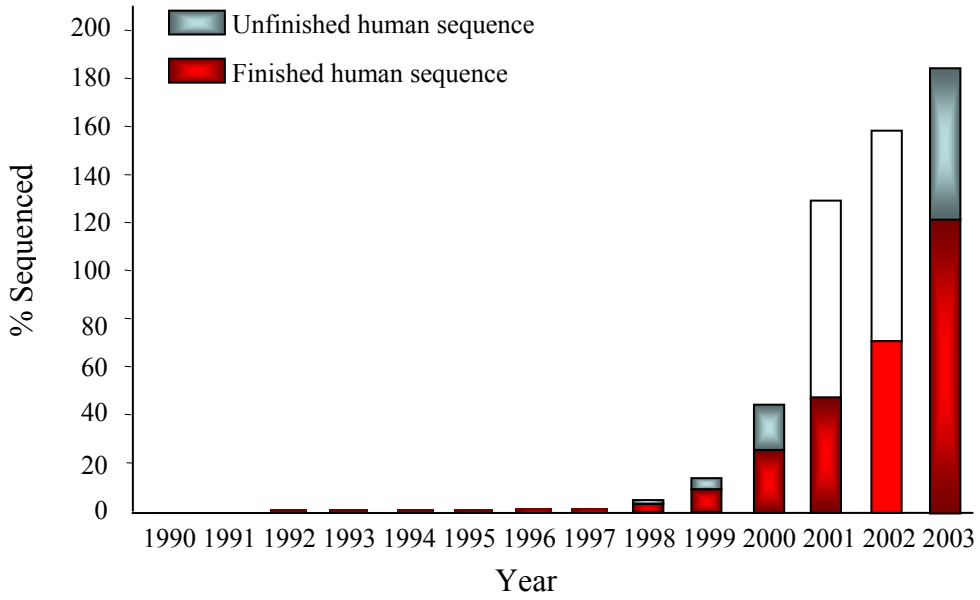

Figure 1.4 Progress of the Human Genome Project from the launch in 1990 to its completion in 2003. The % of finished (red) and unfinished (grey) sequence was calculated for January of each year using the Genome Monitoring Table (http://www2.ebi.ac.uk/genomes/mot/). Finished sequence is the final stage of the sequencing project when the sequence is contiguous with reads covering depths of greater than 8 times redundant sequence with 99.99% accuracy. Unfinished sequence is a working draft covering depths of 2-4 times redundant sequence and contains gaps.

1.6 Analysis of the human genome sequence

The sequence of the entire human genome has enabled a number of key aspects of the genome to be investigated in order to test the theory of polyploidisation and the 2R hypothesis. These are discussed below.

1.6.1 Gene numbers

One of the most important pieces of information revealed by sequencing projects is the number of genes. Ohno (1970) first observed that the gene number and genome sizes increased when looking at more complex organisms. This observation has been confirmed by various sequencing projects (table 1.1).

Table 1.1 Gene number and genome size for a range of organisms.

Organism	***Genome size (Mb)***	***Gene Number***	***Reference***
Homo sapiens	3000	~30,000	IHGSC, 2001 and Venter *et al*, 2001
Mus musculus	3000	30,000	Marshall, 2001
Fugu rubripes	365	31,059	Aparacio *et al*, 2002
Drosophila melanogaster	135.6	13,061	Adams *et al*, 2000
Caenorhabditis elegans	97	19,099	*C.elegans* Sequencing Consortium, 1998
Saccharomyces cerevisiae	12.1	6,034	Mewes *et al*, 1997
Escherichia coli	4.67	3,237	Blattner *et al*, 1997

The human genome was originally estimated to have over 80,000 genes while invertebrates have less than 20,000. The fourfold increase between human and invertebrate gene numbers was previously used as evidence in support of the 2R hypothesis (Makalowski, 2001). One of the most interesting discoveries in the human

sequencing projects has been the identification of only approximately 30,000 protein-coding genes in the human genome (IHGSC, 2001; Venter *et al*, 2001). On average, this would give only two paralogues in humans for every invertebrate gene and would support only one round of genome duplication. However, it could be argued that extensive gene loss may follow genome duplication i.e. if two rounds of duplication occurred then a significant proportion of duplicate genes were lost after each duplication event leaving no obvious trace of two genome duplication events.

1.6.2 1-to-4 gene rule

Initial analysis of the protein coding genes in the draft human genome does not support a strict 1-to-4 gene rule (IHGSC, 2001; Venter *et al*, 2001). The International Human Genome Sequencing Consortium employed an all-against-all sequence comparison to identify orthologous groups in human, *C. elegans* and *Drosophila* genomes. A total of 1308 groups were identified with a mean of 2.4 genes per human orthologue group and 1.1 genes per group in *C. elegans* or *Drosophila*. On closer analysis, almost half of the identified orthologue groups had just a single gene in the human genome, and the remainder had two, three, four or more genes. When the ratio of the number of orthologue groups with a single gene in *C. elegans* and *Drosophila* and the number of genes in human were plotted for each analysis, the peak of this distribution was found over the 1:1 ratio and not the 1:4 ratio needed to support a strict 1-to-4 gene rule. In both cases, there are a significant number of gene families (greater than 50%) with two or more members implying that gene families have expanded via duplication (IHGSC, 2001; Venter *et al*, 2001). These gene family expansions could have been generated through whole-genome duplication events.

1.6.3 Paralogy and the human genome

With the advent of the 'draft' human genome sequence, a number of analyses have now been performed to identify all the paralogous regions. Prior to the release of the draft sequence several lists of paralogous regions had been published and were believed to represent only a small percentage of the total (Lundin, 1993; Lundin and Larhammar, 1998; Skrabanek and Wolfe, 1998; Pollard and Holland, 2000).

The International Human Genome Sequencing Consortium (2001) concluded that approximately 5% of the human genome consists of paralogous regions. The duplicated regions tend to be large, greater than 10 kb, and highly homologous. Evidence of ancient duplications, characterised by high sequence similarity between coding regions, were identified along with evidence of more recent segmental duplications. The latter duplicated regions share high sequence identity between both exons and introns, with many showing less than 6% nucleotide divergence between paralogous regions. Such duplications seem to have emerged very recently in evolution as they are absent from closely related species.

Analysis of the draft human genome sequence by Venter and co-workers (2001) using a multiple alignment algorithm, identified 1077 blocks of paralogy spread throughout the genome. Out of the 1077 blocks, 159 contained only three genes, 137 contained four genes and 781 contained five or more genes thus illustrating the extent of duplications in the human genome. McLysaght and colleagues (2002) conducted one of the most thorough investigations into duplicate genes in the human genome. Of the 24,046 genes used in the analysis, 6,120 (almost a quarter) were identified located in 1642 paralogous regions containing two or more linked duplicated genes. The Hox gene clusters were amongst the largest paralogous regions identified; they found 28

paralogous genes on chromosomes 7p and 17q, and 26 genes on chromosomes 2q and 12q. Owing to the number and sizes of the paralogous regions identified in all three analyses the most likely explanation is that they arose by whole-genome or large-scale block duplication events rather than through duplication of individual genes.

1.6.4 Evolutionary analysis of paralogous gene families

A number of phylogenetic studies have been conducted in order to understand the evolutionary histories of the paralogous gene families. The 2R hypothesis proposes that one round of duplication occurred after the divergence of cephalochordates (exemplified by amphioxus) and the second after the divergence of jawless fish (including hagfish and lamprey). Therefore, the phylogenetic trees of the gene families should show similar histories.

The phylogenetic analyses of gene families supporting the 1-to-4 (or less) gene rule revealed that the evolution of the human genome is complicated. Wang and Gu (2000) analysed 49 vertebrate gene families, each consisting of three or four gene members, generated in the early stages of vertebrates, and/or shortly before the origin of vertebrates, including the early growth response protein, EGR, and the glycine receptor, GLR. Of the 49 gene families studied, they determined that 26 families with three members were consistent with the 2R hypothesis but the evolution of the remaining 23 gene families with four members was more complicated. Of these 23, only five were consistent with the 2R hypothesis, with 11 families supporting a third round of genome duplication and the remaining seven families suggesting at least one round of duplication prior to the divergence between *Drosophila* and vertebrates.

In contrast, Friedman and Hughes (2001) found that of a total of 134 families with four members 70% were not consistent with the 2R hypothesis. Similar results were also reported for a smaller number of gene families by the International Human Genome Sequencing Consortium (2001). However, it is considered by some that organisms, such as amphioxus, hagfish and lamprey, are more appropriate to study vertebrate evolution than *Drosophila* as they are actually on the vertebrate lineage (Holland, 2003). Escriva and colleagues (2002) investigated 33 gene families, where the sequence was available for both lamprey and hagfish. According to their phylogenetic analyses, all 33 families were found to support the 2R hypothesis.

1.7 Polyploidy

Humans and other species are generally diploid and have two copies of each gene; one from each parent. As stated earlier, it has been suggested that the vertebrate genome evolved via whole-genome duplication events, in which the chromosome complement doubled at some point in time. Therefore, the vertebrate genome underwent a stage when it was polyploid, then, through processes such as gene silencing and mutation, reverted to a diploid-like state. Several polyploid species have been identified in both the animal and plant kingdoms. One example is the amphibian, *Xenopus laevis,* which is tetraploid and has double the number of chromosomes than its cousin, *Xenopus tropicalis*. In 1999, the first polyploid mammal, the red viscacha rat (*Tympanoctomys barrerae*) was discovered (Gallardo *et al*, 1999). The rodent is unaffected by having double the number of chromosomes showing that the vertebrate genome can duplicate and that organisms can survive with multiple copies of a genome.

In addition to the two rounds of genome duplication in the vertebrate lineage Ohno (1970) proposed a round of genome duplication in fish; after the divergence of lobed-fin fish that led to land-based organisms. Evidence in support of the third duplication was detected in zebrafish (Postlethwait *et al,* 1998) and *Medaka* (Wittbrodt *et al,* 1998) based on the observation that they generally have larger multigene families than mammals. In particular, Amores and co-workers (1998) observed that zebrafish had seven Hox gene clusters in comparison to the four present in mammals and one in amphioxus (Garcia-Fernandez and Holland, 1994). Further mapping and sequence data has shown that for any four paralogous (or tetralogous) genes or regions in mammals there are probably an additional three or four in teleost fish.

1.8 Mechanisms of gen(om)e duplication

Gene duplication has played a major role in the evolution of the human genome. Duplication may involve part of a gene, a single gene, part of a chromosome, an entire chromosome, or the whole genome. The duplication of part of, or a whole, gene is also referred to as a tandem duplication event. Chromosomal regions are duplicated as part of either a block or segmental duplication event. Segmental duplications are defined as involving the transfer of genomic sequence to one or more locations in the human genome and, because of the strong sequence identity between both exons and introns, are relatively recent events (IHGSC, 2001). More ancient duplication events are characterised by similarities only in the coding regions and, in this thesis, are referred to as block duplication events.

The duplication of an entire chromosome is also known as aneuploidy or polysomy. There are several examples of trisomy of human chromosomes that are linked to a

number of conditions. A well known example of this is the trisomy of chromosome 21, which causes Down's syndrome. As discussed previously in this thesis the duplication of an entire genome is referred to as whole-genome duplication or polyploidisation.

There are several mechanisms by which duplication can occur; they are unequal crossing-over, unequal sister chromatid exchange, duplicative transposition, replication slippage and polyploidisation. Unequal crossing-over is a recombination event initiated by similar nucleotide sequences that are not at identical places in a pair of chromosomes. Unequal sister chromatid exchange is essentially the same as unequal crossing-over except that it involves chromatids from a pair of homologous chromosomes. The result can be duplication of a segment of DNA in one of the recombination products. This mechanism can create both small families, such as the five related genes of the β-globin cluster on chromosome 11, and large ones, such as the olfactory receptor gene clusters, which together contain nearly 1,000 genes and pseudogenes.

Transposition is defined as the movement of genetic material from one chromosomal location to another. During the process termed duplicative transposition, the transposable element is copied, therefore if this element contains a gene, the original copy is retained at the original site while a new copy is inserted elsewhere in the genome. The process, replication slippage is more commonly associated with the duplication of very short sequences, such as repeat units in microsatellites, but can also result in gene duplication if the genes are relatively short. In either case, the recombination occurs between two different copies of a short repeat sequence leading to duplication of the sequence between the repeats.

Whole-genome duplication occurs as a consequence of the lack of separation between all daughter chromosomes following DNA replication. Since it immediately doubles the size of a genome it is considered as the most effective mechanism for increasing genome size. Whole-genome duplication can occur via two mechanisms; allotetraploidy and autotetraploidy (figure 1.5). Autotetraploidy occurs within a single species and allotetraploidy occurs between genomes from different individuals (Wendel, 2000).

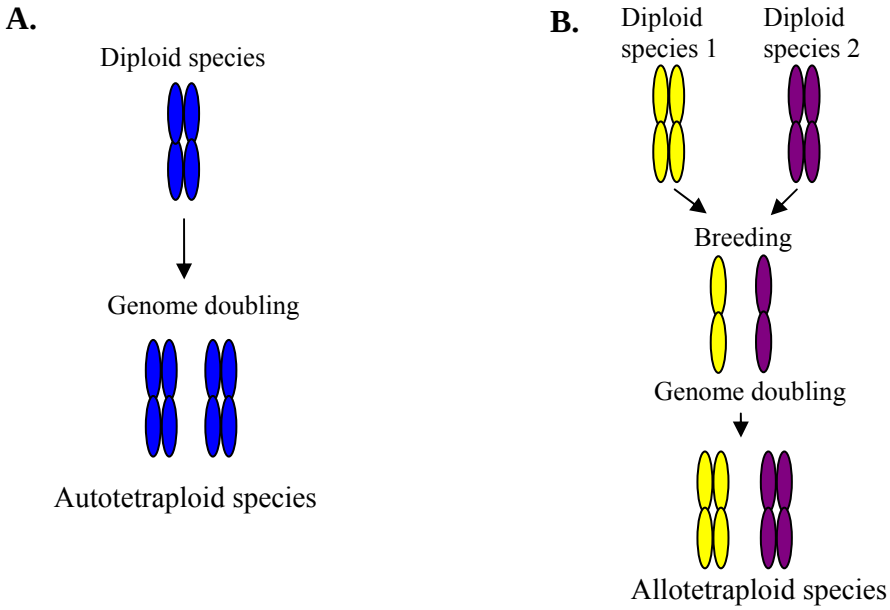

Figure 1.5 Models of genome duplication by (A) autotetraploidisation and (B) allotetraploidisation.

In plants, polyploidy is widespread and numerous studies have been conducted to understand the prevalence and consequence of polyploidy. Artificially produced autopolyploids are generally inferior to their diploid progenitors, and have lower fertility and, often, lowered ability to compete with diploid species owing to physiological effects such as, genetic imbalances and irregularities in chromosomal segregation (reviewed by Li, 1997).

Polyploidy is extremely rare in bisexually reproducing animals. Muller (1925) proposed that this is because in bisexual animals the two sexes are differentiated by means of a process involving the diploid mechanism of segregation and combination, and polyploidy invariably disturbs this process. In amphibians and fish, where there is evidence of successful polyploidy, the chromosomal determiners of the opposite sexes are still in a rather initial state of differentiation, and the X and the Y or Z and the W chromosomes can substitute for each other (Ohno, 1970). In these animals, genome duplication would not result in sexual imbalance, and many tetraploid species have been found (Ohno, 1970; Bogart, 1980; Schultz, 1980). It is interesting to see that, the only example of a tetraploid mammal identified to date is tetraploid for all autosomal chromosomes but is diploid for the sex chromosomes (figure 1.6; Gallardo *et al*, 1999).

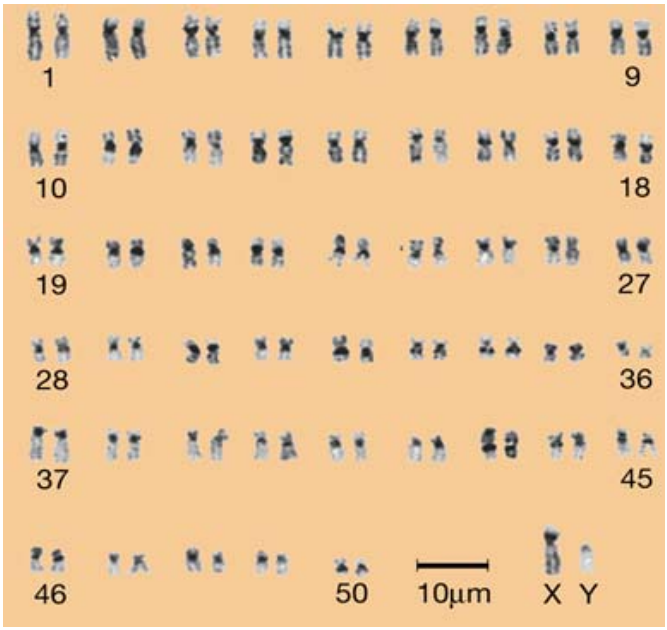

Figure 1.6 Karyotype of a male tetraploid *Tympanoctomys barrerae* from Mendoza, Argentina taken from Gallardo *et al* (1999). The karyotype contains 36 pairs of metacentric to submetacentric chromosomes and 14 pairs of subtelomeric autosomes. The X chromosome is the largest element (present in two copies in females) and the Y chromosome is the only acrocentric element of the karyotype.

Ohno (1970) argued that genome duplication has been more important than tandem duplication because the latter may duplicate only parts of the genetic system of structural genes and regulatory genes. This may disrupt the function of the duplicate genes, whereas polyploidisation duplicates the entire genetic system. However, evidence from the human genome has shown that most genes do not exist as a single copy in the genome but rather as clusters. Thus, showing that tandem duplication has played an important role in moulding the present-day structure of the human genome. It has also been seen that tandem duplication is both important for increasing the number of genes with the same function, exemplified by the HLA class I genes, and for generating genes with new functions, such as the human β-globin genes.

1.9 What happens after gen(om)e duplication?

Gene duplication is an important mechanism for the creation of new gene function (Ohno, 1970; Lynch and Conery, 2000; Wagner, 2001). After gene duplication the two resulting paralogues can evolve in different ways. The classical model of functional diversification after duplication indicates that one copy of a gene maintains the original function of the ancestral gene whereas the other gene is redundant and will either diverge functionally or be lost from the genome altogether by the accumulation of random mutations (Ohno, 1970). More recently an alternative model has been proposed, in which the two gene copies acquire complementary loss-of-function mutations and develop independent sub-functions, such that both genes are required to produce the full complement of functions of the ancestral gene (Force *et al*, 1999). The process is known as both the sub-functionalisation model and the duplication-degeneration-complementation (DDC) model.

1.10 The extended Major Histocompatibility Complex

The human Major Histocompatibility Complex (MHC) is located on the short arm of chromosome 6 (6p22.2-p21.3). The region was identified in humans over 50 years ago because of its role in tissue transplant rejection (Dausset, 1958) and is now one of the best characterised and studied regions in the human genome. It contains a high density of immune-related genes responsible for recognising foreign antigens and eliciting an adaptive immune response. The region has been linked with more diseases than any other region in the human genome (Price *et al*, 1999). It is of particular biological importance due to its association with a number of autoimmune diseases, including insulin dependent diabetes mellitus, multiple sclerosis, systemic lupus erythmatosus and rheumatoid arthritis (Thomson, 1995). In addition it has been linked with a range of aetiologies from cancer to sleeping disorders (The MHC Sequencing Consortium, 1999).

The MHC has traditionally been divided into three regions: the class I (most telomeric), class III and class II (most centromeric). The complete 3.6 Mb contiguous sequence of the three MHC regions was published in 1999 by the MHC Sequencing Consortium prior to the release of the 'draft' human genome sequence. It was estimated that 40% of the 224 genetic loci (of which 128 are expressed) have an immune function, although many still have an unknown function. Work on these three regions revealed that sequence conservation and possibly linkage disequilibrium extended further; the immediate flanking regions were termed the extended class I region and extended class II regions of the MHC (Stephens *et al*, 1999). The region of the human genome encompassing all five regions is now termed the 'extended Major Histocompatibility Complex' and spans almost 8 Mb and contains over 390 genetic

loci. The extended MHC represents a well characterised region of the human genome and is one of the best examples for the involvement of gene duplication events during its evolution.

1.10.1 The extended class I region

The extended MHC class I region (figure 1.7) has been defined as the region between the hereditary haemochromatosis locus (HFE) and the MOG locus, spanning almost 4 Mb of genomic sequence (Stephens *et al*, 1999). The region is characterised by a number of gene clusters suggesting that this region has evolved via numerous local duplication events, or through the recruitment of similar genes into the region.

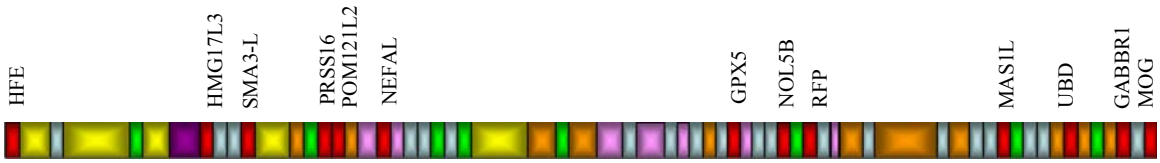

Figure1.7 Schematic representation of the extended MHC class I region. Gene clusters and individual genes are coloured according to family: histones (yellow), ribosomal proteins (green), butyrophilin receptors (purple), zinc finger proteins (pink) and olfactory receptor genes (orange). The expressed genes (red) that do not belong to a gene family cluster are labelled accordingly. Pseudogenes that do not belong to a gene family cluster are coloured grey.

There are 55 histone genes within this region, which is the largest cluster of histone genes in the human genome (Marzluff *et al*, 2002). There are also over 160 small single exon (50-100 bp in length) tRNA genes which produce 18 out of the 20 commonly used amino acids and represents approximately 25% of the human tRNA repertoire (not shown on figure 1.7). Other clusters located within the extended MHC class I region include; 20 zinc-finger proteins, 10 ribosomal proteins, two clusters of olfactory receptor genes and seven butyrophilin genes. There is further evidence of

local duplication events involving the GPX5 gene and POM121L2 gene, which both have a pseudogene in close proximity. In addition, there are also a number of expressed single copy genes, as well as pseudogenes in the extended class I region.

1.10.2 The class I region

The MHC class I region (figure 1.8) contains the three functional, classical class I genes, HLA-A, HLA-B and HLA-C, which are highly polymorphic and are expressed by most nucleated cells. In addition, there are several other functional class I loci, including the non-classical class I genes, HLA-E, HLA-F and HLA-G, which are less poylmorphic and have restricted expression. These genes are termed HLA class I genes in this thesis. The HLA-H, HLA-J and HLA-K gene fragments are thought to be pseudogenes.

The HLA class I genes encode the heavy (α) chain of the cell-surface class I molecule which, along with the β chain encoded by the β2-microglobulin locus on chromosome 15, is responsible for presenting antigens (short, specific processed peptides) to T-cells. The peptides loaded onto the class I molecules are generally derived from an endogenous (intracellular) source by the proteasome, of which PSMB8 and PSMB9 (both found within the MHC class II region; Driscoll *et al*, 1993) are subunits. The peptides are then transported to the endoplasmic reticulum by the TAP1/TAP2 molecules (also encoded by genes within the MHC class II region; Ortmann *et al*, 1994), where they are loaded onto MHC class I molecules and proceed as a complex to the cell surface via the Golgi apparatus. At the cell surface, the MHC class I molecule-peptide complex is accessible to $CD8^+$ cytotoxic T lymphocytes that elicit an immune response, which results in the lysis of the cell presenting the antigen (for a

review of class I antigen presentation see Monaco, 1992).

Of the 50 or more non-HLA related genes within the class I region, there are genes that are distantly related to the conventional class I sequence, namely the MIC genes. There are also a large number of pseudogenes (almost half of the genes) and multigene families, such as the P5 and HCG families; suggesting that duplication events contributed to the evolution of the class I region (Shiina *et al*, 1999).

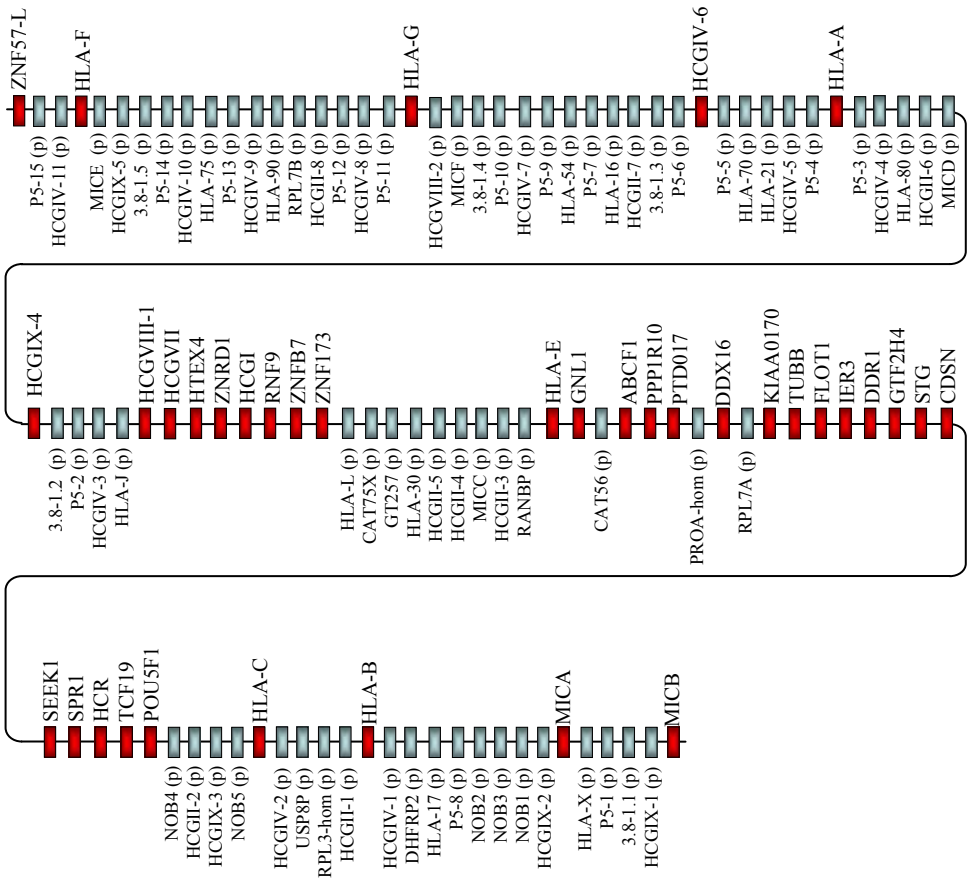

Figure 1.8 The MHC class I region. The expressed genes (red) are labelled above the gene track and the pseudogenes (grey) are labelled with a 'p' below the gene track.

1.10.3 The class III region

The MHC class III region (figure 1.9) spans approximately 0.7 Mb and is extremely

gene dense with 58 genes (this corresponds to 1 gene every 12 kb of DNA). The extent of gene density is demonstrated by the overlapping genes AGPAT1 and C6orf8, which are transcribed in different directions but overlap by 87 bp at their 3-prime ends. Furthermore, the TNXB and CYP21A2 genes overlap in the 3-prime untranslated regions. The high gene content of the class III region is complemented by a corresponding high GC content (53%). This produces a distinct boundary between the class III region and the rest of the MHC (reviewed by Beck and Trowsdale, 2000). The genes encoded in the class III region have a variety of functions and are associated with diseases, such as congenital adrenal hyperplasia and C2 deficiency (reviewed by Gruen and Weissman, 2001). There are a number of genes with an immune-related function, including members of the complement cascade (C2, C4 and BF) and the tumour necrosis factor family (TNF, LTA and LTB). Genes that are expressed in specialised cells of the immune system, such as LST1 and 1C7, are located next to each other (Holzinger *et al*, 1995).

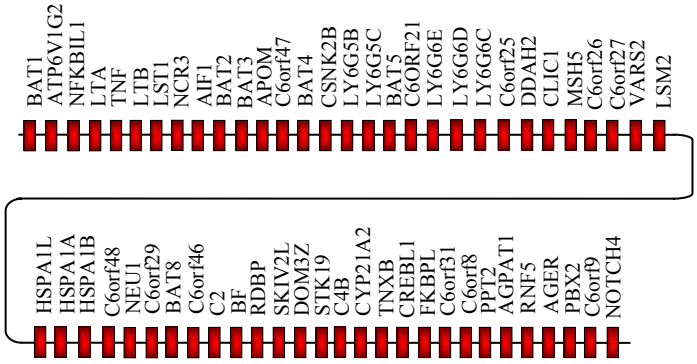

Figure 1.9 The MHC class III region. Expressed genes labelled in red are plotted from the most telomeric (BAT1) to centromeric (NOTCH4) end of the chromosome.

Gene duplication has occurred to a lesser extent in the class III region as compared with other regions of the MHC. Three genes, C4/CYP21A/TNX, have undergone

tandem duplications yielding a complex comprising overlapping genes and genes within genes (Bristow *et al*, 1993). There are also small clusters of gene families, including the three heat shock proteins located next to each other and members of the Ly6 superfamily. The class III region is unique compared with the rest of the MHC region as it does not contain any pseudogenes (with the exception of the C4 duplicate in certain haplotypes), has few duplicated genes and the genes have diverse functions, suggesting a distinct origin of the class III region.

1.10.4 The class II region

The MHC class II region (figure 1.10) takes its name from the classical and non-classical HLA class II genes, termed HLA class II genes in this thesis. The classical HLA class II genes (HLA-DP, HLA-DQ, HLA-DR) either encode proteins with α chains (HLA-DPA, HLA-DQA, HLA-DRA) or proteins with β chains (HLA-DPB, HLA-DQB, HLA-DRB). The α and β chains combine to form class II MHC molecules. The class II molecules are polymorphic and are expressed on specialised antigen-presenting cells (e.g. dendritic cells, B lymphocytes, macrophages) and present peptides mainly derived from extracellular proteins to $CD4^+$ T cells.

MHC class II molecules differ from MHC class I molecules in that the groove of the peptide-binding region (PBR) is open-ended, thus allowing longer peptides to be bound. Prior to a peptide binding, the class II molecules are assembled in the endoplasmic reticulum (ER) with a membrane-bound chaperone protein (known as the MHC class II associated invariant chain or γ chain) acting to stabilise the complex. This γ chain is degraded by proteases in the trans-Golgi reticulum with the exception of a small fragment that is buried in the PBR. The removal of this small fragment

(prior to peptide binding) is catalysed by gene products of the HLA-DM gene – a non-classical class II gene. After binding, the MHC class II molecule-peptide complex is transported to the cell surface where it is recognised by $CD4^+$ helper T lymphocytes (for a review of class II antigen presentation see Neefjes and Ploegh, 1992 and Pieters, 1997).

Within the class II region there are also a number of other genes that have an immune related function. The PSMB8, PSMB9, TAP1 and TAP2 genes are involved with antigen processing of MHC class I molecules as described in section 1.10.2. There are also a number of genes with quite diverse functions, such as the butyrophilin-like gene BTNL2, the testis-specific basic protein TSBP and the bromodomain-containing protein BRD2. Furthermore, a number of pseudogenes are located within this region, including a ribosomal protein pseudogene and the pseudogene of the extended class II gene COL11A2.

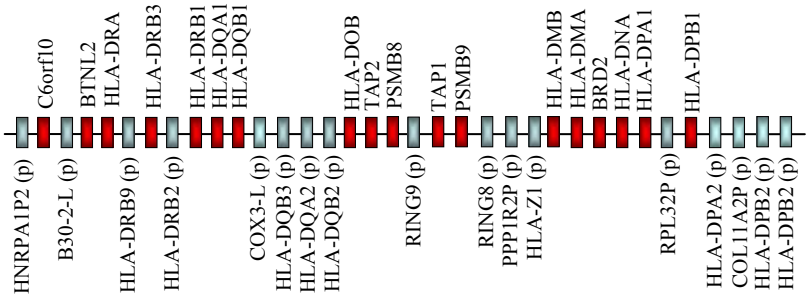

Figure 1.10 The MHC class II region. The expressed genes are shown in red and are labelled above the gene track and the pseudogenes (labelled with a 'p') are in grey and are labelled below.

1.10.5 The extended class II region

The identification of the tapasin gene, required for antigen presentation by MHC class

I molecules, in the region flanking the MHC class II region suggested that the MHC extended further than previously thought (Herberg *et al*, 1998a; 1998b). Detailed analysis of the region centromeric to the MHC class II region, now termed the extended class II region (figure 1.11), revealed several other genes, including collagen gene type 11A2 (COL11A2), a ribosomal protein RPS18 and, the most centromeric gene in the extended MHC, KNSL2.

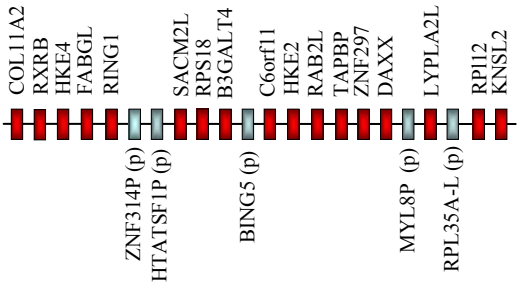

Figure 1.11 The extended MHC class II region. The expressed genes are shown in red and are labelled above the gene track and the pseudogenes (labelled with a 'p') are in grey and labelled below.

1.11 Origin of the extended MHC

There is conservation of some of the genes within the extended MHC regions between species suggesting that there is an evolutionary advantage in conserving the MHC as a unit. This MHC 'unit' can be observed in species evolving after the divergence of the jawless vertebrates. In particular, the MHC class I and class II region genes have been identified in all jawed vertebrates studied to date, but have not been identified in the jawless vertebrates, hagfish and amphioxus (Kasahara *et al*, 1996b; Flajnik *et al*, 1999). Jawless vertebrates also lack other molecules of the adaptive immune system, such as RAG1 and RAG2, as well as the lymphoid organs thymus and spleen. Thus, the adaptive immune system has arisen in a very short period of geological time since

the emergence of jawed vertebrates (Bernstein *et al*, 1996). Several MHC genes (including NOTCH4, RXRB and PBX2) are syntenic in invertebrate genomes, such as *Drosophila* and *C. elegans* indicating that the origin of the MHC locus predates the emergence of the adaptive immune system

The three classical regions of the human MHC (class I, class III and class II) appear to have been subject to different evolutionary mechanisms: whilst MHC class II and class III genes often appear to have direct orthologues, the MHC class I genes appear to have expanded and contracted in different species. The class III region is considered to be the oldest region of the MHC (reviewed by Beck and Trowsdale, 2000). It is evident that both the class I and class II regions have evolved via a series of duplications, but it is not known which region came first. One hypothesis claims the class II region evolved first (Hughes and Nei, 1993), whereas another hypothesis holds that the class I region originated first as a result of a recombination between an immunoglobulin-like C-domain and the peptide-binding domain of an HSP70 heat shock protein (Flajnik *et al*, 1991). Phylogenetic analysis supported the prior hypothesis, albeit with low statistical support (reviewed by Hughes and Yeager, 1997; Klein and Sato, 1998).

1.12 MHC Paralogy

MHC paralogous genes were observed during the study of MHC class III genes (Sugaya *et al*, 1994; Katsanis *et al*, 1996) and the class II proteasome genes (Kasahara *et al*, 1996a). It was concluded that the region 9q33-q34 was paralogous to the MHC. Furthermore, Katsanis and colleagues (1996) also noted two additional regions in the human genome, 1q21-q25/1p11-p32 and 19p13.1-p13.3, which contained MHC

paralogues (Figure 1.12). Initially, only a few genes were reported to have paralogues on chromosomes 1, 9 and/or 19 but the number has increased to 40, approximately one third of the expressed MHC genes (reviewed by Kasahara, 1999b).

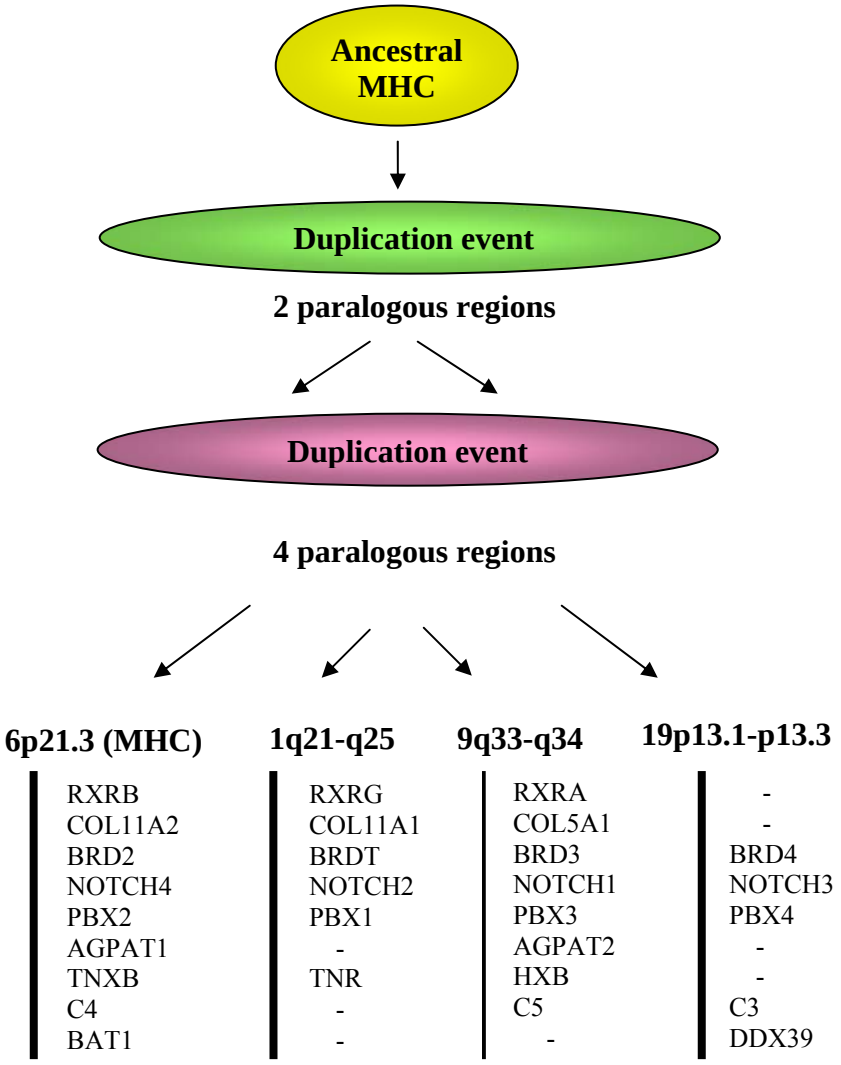

Figure 1.12 Summary of the MHC paralogous regions in the human genome.

The MHC genes with paralogues reside in both the classical and extended regions of the MHC region and constitute a diverse group of genes in terms of structure, function and gene size. Some families, such as NOTCH and PBX, have copies in all four

regions, but most only have two or three copies. MHC paralogues may not be identified in all regions, for each gene, as duplicated genes are likely to be silenced or lost from the genome altogether (Nadeau and Sankoff, 1997). Interestingly, there are also a number of other gene families that have copies in the 1, 9 and 19 paralogous regions but not in the MHC (reviewed by Kasahara *et al*, 2000).

1.12.1 Origin of the extended MHC paralogous regions

The origin of the MHC and the three paralogous regions is controversial. Currently there are two main hypotheses. The first is that they descended from a common ancestral region and emerged as a result of large-scale block duplications (Kasahara *et al*, 1996a; Kasahara, 1999a; Abi-Rached *et al*, 1999; Flajnik and Kasahara, 2001). The second is that they represent assemblies of independently duplicated genes and are grouped together by selective forces (Hughes, 1998). In general, the block duplication mechanism is preferred, as it can best explain why assortments of functionally and structurally varied genes are clustered on four specific regions of the human genome.

It is possible that the MHC and the paralogous regions on chromosomes 1, 9 and 19 arose from two rounds of whole-genome duplications (the 2R hypothesis) based on the estimated timings and numbers of duplications that appear to have occurred (reviewed by Flajnik and Kasahara, 2001). It is believed that the first round of duplication occurred close to the origin of jawed vertebrates. This is supported by the identification of a single orthologue of the MHC paralogues, exemplified by the complement genes C3, C4 and C5. All have been identified in cartilaginous fish, however, jawless fish lack C4 and C5 but do have a C3-like gene that shares features

with the common ancestor of C3 and C4 (Kasahara, 1999a).

In order to understand the evolution of the MHC and associated paralogous regions, numerous genes have been analysed in a number of organisms, including *Drosophila* and amphioxus which are thought to predate the whole genome duplication events proposed by the 2R hypothesis. The identification of 19 MHC paralogous genes in *Drosophila* (Danchin *et al*, 2003) and nine MHC paralogous genes in amphioxus (Abi-Rached *et al*, 2002) residing in close proximity to other genes found on human 1q21-q25/1p11-p32, 9q33-q34 and 19p13.1-p13.3 provides evidence for a block duplication event in vertebrates. Phylogenetic analysis has demonstrated that the duplications occurred prior to vertebrate emergence but after the divergence of amphioxus from the vertebrate lineage supporting the 2R hypothesis.

If the paralogous regions have a common origin there should be evidence of conserved synteny between them. Preliminary analysis of the gene order within the paralogous regions indicated that the order is poorly conserved (Endo *et al*, 1997). This is no surprise considering that more than 500 million years have passed since the last duplication event occurred and each region has undergone major structural rearrangements, including inversions and translocations. Rearrangements are particularly dramatic on chromosome 1, as the MHC paralogues are located on both arms of the chromosome. There is compelling evidence that chromosome 1 underwent a pericentromeric inversion after the divergence of the human and chimpanzee lineages and this is probably responsible for the occurrence of the paralogous genes on both arms (Maresco *et al*, 1996).

1.13 Thesis aims

With the discovery that approximately 10% of the human genome arose by duplication it is evident that this process has played a major role in the evolution of our genome. Whether the duplication events involved the entire genome (as proposed by the 2R hypothesis), chromosomal segments or individual genes is unclear. Therefore, the aim of this thesis was to investigate the mechanism(s) that gave rise to the present-day organisation of the human genome.

Using the MHC region as a model a number of aspects of paralogy and the genome were investigated. Firstly, in order to study the genomic regions containing MHC paralogous genes, the chromosomal region 9q32-q34.3 was mapped, sequenced and analysed. Comparison of 9q32-q34.3 with the MHC region on 6p22.2-p21.3 will reveal the level of synteny between these two regions and determine whether they have a common origin. Secondly, a survey of the entire human genome sequence was conducted to identify the MHC paralogues and determine their distribution. Thirdly, phylogenetic trees were used to reconstruct the evolutionary histories of the MHC paralogues. Analysis of the topology of the trees and the arrangement of the paralogues and orthologues will determine the mechanism(s) of evolution. Finally, the expression profiles were generated to understand how the MHC paralogues have evolved since their emergence. In summary, the data presented in this thesis aims to provide a unique insight into the evolution of the MHC paralogous genes and the human genome.