Chapter 3

The human MHC extended class I region

3.1. Introduction

The MHC extended class I region has been defined as the sequence on chromosome 6 between the HLA-F locus (the end of the classical MHC) and the hereditary haemochromatosis locus (HFE, originally known as HLA-H) (Stephens *et al.*, 1999). This definition was supported by two pieces of evidence obtained through a transcript map of the hereditary haemochromatosis locus (Ruddy *et al.*, 1997). Firstly, the transcript map revealed several members of the butyrophilin family and the RoRet gene which share an exon of common evolutionary origin called B30-2. This B30-2 exon was originally isolated from the HLA class I region, leading to the suggestion that it may have "shuffled" into several genes telomeric to the MHC. The "shuffling" of this exon, together with the fact that the hereditary haemochromatosis locus (HFE) has a certain level of amino acid homology to MHC class I molecules, was taken as evidence that the area around the hereditary haemochromatosis locus was related to the MHC. In addition to these observations, some studies have suggested that there is a strong linkage disequilibrium between the HLA-A locus and the HFE locus, leading to the initial proposal that HFE was located in the classical MHC class I region (subsequently displaced by the extended MHC class I hypothesis) (Simon *et al.*, 1987, Malfroy *et al.*, 1997). The extension of synteny beyond the HLA-F gene between human and mouse also provided support for the idea of an extended MHC (Yoshino *et al.*, 1997).

In the light of data provided by the Human Genome Project and other sources, this definition of the MHC extended class I region could be considered to be outdated. The B30.2 domain, for example, is known to exist across the genome in a variety of locations (Henry *et al.*, 1998).

Similarly, there is evidence suggesting that recombination between HFE and the MHC does occur (Roetto *et al.*, 1997). In addition, the mouse synteny does not extend as far as HFE: mouse Hfe is located on chromosome 13 rather than on chromosome 17, next to the MHC (Szpirer *et al.*, 1997). Although the similarity between HFE and the two HLA genes at the telomeric end of the classical MHC has withstood the influx of additional data (a 'BLAST' search of the entire human genome using the HFE sequence reveals that these two genes are the closest relative the HFE locus has (over 60% shared nucleotide identity)), the definition of a MHC extended class I region encompassing HFE is questionable. In spite of these inconsistencies associated with defining the MHC extended class I region, however, for the lack of a better description of this region, the term MHC extended class I region is used throughout this thesis.

One of the aims of this project was to examine the relationship between the MHC and the MHC-linked olfactory receptor genes. Identification of all human MHC-linked olfactory receptor genes, therefore, involved the complete analysis of the region between HLA-F and HFE.

3.2. Sequence assembly and gene content

Mapping of the human MHC extended class I region was carried out by a number of groups; by those specifically interested in the MHC region (Gruen *et al.*, 1992, Volz and Ziegler, 1996, Ruddy *et al.*, 1997), and as part of a large scale approach to map the Human Genome (McPherson *et al.*, 2001). Sequencing was carried out at the Sanger Centre, as part of the Sanger Centre's contribution to the Human Genome Project. From these clones, a consensus sequence was put together using information from the FPC fingerprinting databases and information in EMBL files, in conjunction with programs used to assess the overlaps between clones such as 'cross_match' and 'dotter' (Chapter 2). Clones containing HLA-F and HFE were taken as the end points in the

sequence (although extra genes located next to these genes within these clones are included in the analysis).

The extended MHC class I region was found to consist of 3913358 bases. Details of the assembly of the 43 clones that contribute to this sequence are found in Appendix 5: this includes AL031983, dJ271M21, which was sequenced and assembled by me as part of this project.

The human MHC extended class I region was analysed using a variety of tools, ranging from the 'NIX' program (*HGMP*) to an analysis of the olfactory receptor content performed using 'dotter'. (Chapter 2). In some cases, where analysis had been performed by the human annotation group at the Sanger Centre, the program 'seeclone' (Chapter 2) was used to obtain the gene content of a clone.

The gene content of the 3913358 bases is shown in Figure 3.1. Genes of the same type are highlighted in the same colour, whilst genes with no clear relatives in the sequence are black. In total, 178 loci have been identified: of these 34 are olfactory receptor genes, whilst 5 are pheromone receptor loci (VNO type 1-like genes). In addition to these genes, the extended MHC class I region also contains 51 histone gene loci, 7 butyrophilin-like genes (Rhodes *et al.*, 2001), and 20 zinc finger-like genes.

The olfactory receptor and pheromone receptor genes will be discussed in more detail in later chapters (Chapter 4, MHC-linked olfactory receptor genes, Chapter 9, MHC-linked pheromone receptor genes). However, in order to gain an idea of the overall gene content of the MHC extended class I region and to consider how OR genes fit into the region, a brief description of the

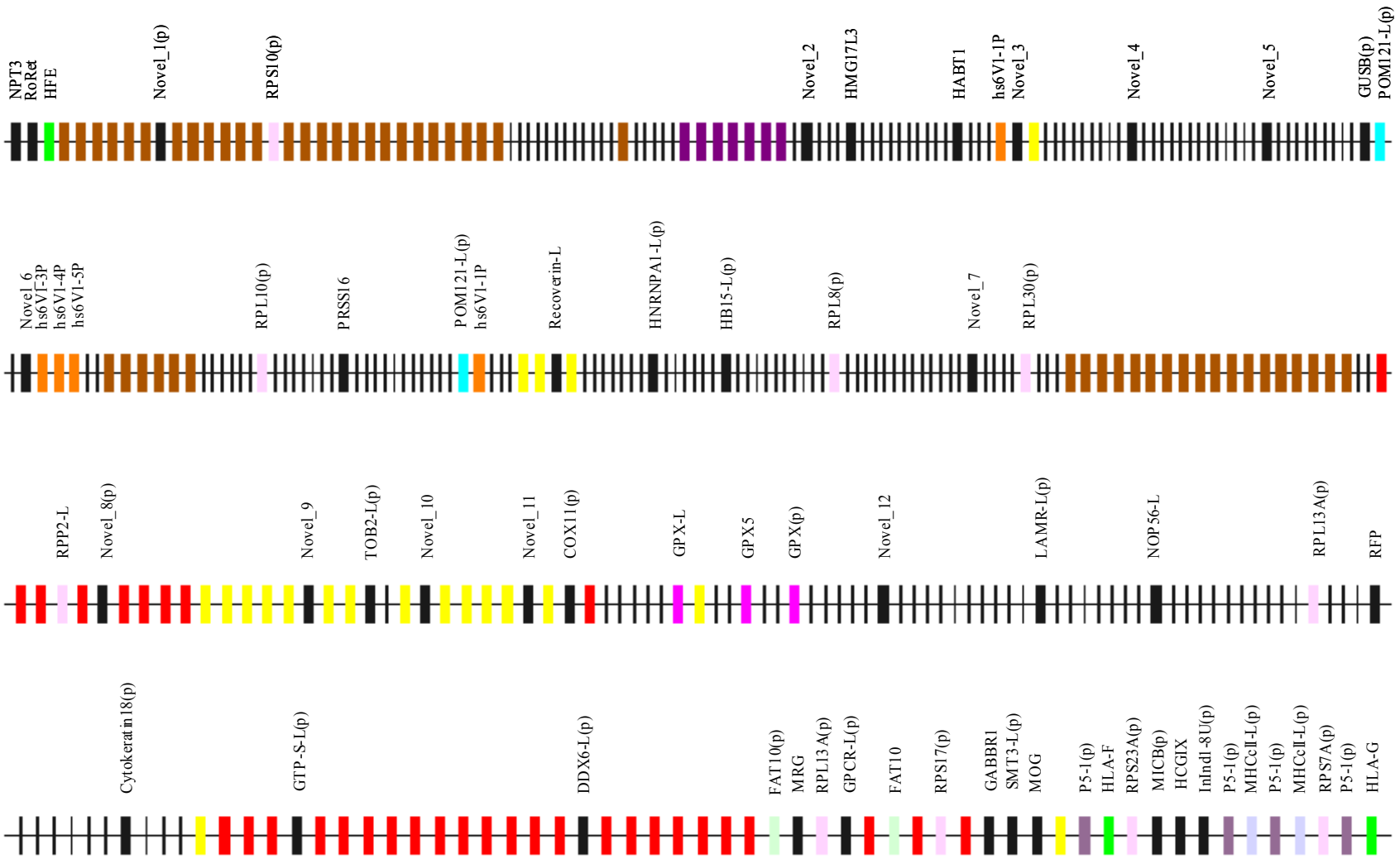


Figure 3.1: Gene content of human MHC extended class I region. 1 Mb of sequence is displayed per track, and genes are coloured according to the family to which they belong: HLA genes (green), histones (brown), ribosomal proteins (light pink), butyrophilins (purple), pheromone receptors (orange), POM121-like (blue), zinc finger proteins (yellow), olfactory receptors (red), GPX-like (pink), FAT10-like (light green), P5-1-like (light purple), and MHC class I-like (light blue). The tRNA genes are represented by the thin black lines. Genes not belonging to a family are coloured black and are labelled above the gene track. A number of landmark genes are also labelled above the gene tracks.

other genes within the extended class I is featured in this chapter. It is interesting to note, given that duplication and diversification have been proposed as hallmarks of MHC evolution (Shiina *et al.*, 1999), that the extended class I region consists of a 'cluster of clusters,' with histone genes, zinc finger protein genes and olfactory receptor genes all prevalent within the sequence of this region.

3.3. Gene clusters in the human MHC extended class I region

3.3.1. The histone cluster

The histone cluster contains 47 genes encoding a variety of different proteins that can be divided into five subfamilies of basic nuclear protein. These proteins are responsible for the nucleosome structure of the chromosomal fibre in the eukaryotic genome. The core structure of the nucleosome is formed by two of the core histones (subtypes H2A, H2B, H3, and H4), whilst the linker histone of subtype H1 anchors two rounds of nucleosome DNA on the surface of the nucleosome core (Maxson *et al.*, 1983).

The cluster of histones on chromosome 6 is the largest cluster of histones in the human genome, although there is a small group on 1q21 (Albig and Doenecke, 1997) and there appear to be isolated single copy genes located in other regions of the genome. The arrangement of the genes within the cluster is interesting: the majority of H2A and H2B genes are located in pairs as has previously been observed (Trappe *et al.*, 1999) but the partners are always located on opposite strands. It is also interesting to note that the histone cluster is subdivided into four separate subclusters by a number of other genes. All of these subclusters contain H1 genes, with the exception of subcluster 3.

3.3.2. The zinc finger protein cluster

In contrast to the histone cluster, which is the largest histone cluster in the human genome, the zinc finger protein gene cluster can be regarded as relatively small, containing only 20 zinc finger proteins (ZNF).

Zinc finger proteins are involved in binding nucleic acids which can have a number of implications, the most notable function being the regulation of transcription (Laity *et al.*, 2001). There are a number of types of ZNF that are characterized according to certain properties, for example the C2H2-ZNF family is characterized by repeated zinc finger motifs of approximately 28 amino acids that have been shown to trap zinc ions using 2 cysteine residues and 2 histidine residues. Of these C2H2-ZNFs, a large number are classified as Krüppel-like. Krüppel-like ZNFs are defined according to the possession of conserved 6 amino acid histidine-cysteine links in the regions connecting successive finger repeats (Bellefroid *et al.*, 1991). A further level of classification that is applied to zinc finger proteins is KRAB-like. This refers to Krüppel-like ZNFs that contain a conserved, approximately 75-amino acid motif, called the Krüppel-associated box (KRAB), in their N-terminal nonfinger region. The KRAB is composed of 2 modules, the A box and the B box (Bellefroid *et al.*, 1993).

3.3.3. The ribosomal protein cluster.

In addition to zinc finger proteins, which have been implicated in the regulation of transcription, and histone genes, which are implicated in allowing transcriptional factors to reach certain areas of chromosome, the MHC extended class I region also contains a number of proteins implicated in the control of translation, the ribosomal proteins (Warner and Nierras, 1998). These ribosomal

proteins combine with other ribosomal proteins and 4 species of RNA in order to produce functional ribosomes.

These ribosomes are composed of 1 large 60S subunit and 1 small 40S subunit. It is predicted that 80 different ribosomal proteins are available to become involved in these subunits. Within the human genome, however, the number of ribosomal protein loci exceeds 80. Ribosomal protein genes are members of multigene families, most of which are composed of 1 single functional intron-containing gene plus multiple processed pseudogenes (Davies *et al.*, 1989).

Within the extended MHC there are 10 ribosomal protein-like genes but 9 of these appear to be pseudogenes. As detailed in table 3.1, most of these pseudogenes have been mapped as functional genes in other regions of the genome (Kenmochi *et al.*, 1998), so it can be assumed that these loci represent processed pseudogenes. The two exceptions to this are RPS10 which appears to be a pseudogene in this genomic sequence, even though it is expected that a functional form should exist in this region of the genome, and the RPP2-L gene which exists in a functional form in the extended MHC despite possessing a functional form located on chromosome 11.

One of the ribosomal proteins within the extended MHC class I region is the pseudogene version of the BBC1 (Breast Basic Conserved) gene. The functional version of the cDNA was identified as a representation of an mRNA showing significantly higher levels of expression in benign breast lesions than in carcinomas. The cDNA hybridized to multiple sequences within both human and other mammalian genomes and although only one major transcript was identified in human cells, the existence of several pseudogenes was suspected (Adams *et al.*, 1992).

Gene	Chromosome localization	Orientation, consensus	Start position, consensus	End position, consensus	Size
RPS10 (p)	6	>	306514	307011	497
RPL10 (p)	X	>	1284053	1284993	640
RPL8 (p)	8	<	1725509	1725734	225
RPL30 (p)	8	<	1853492	1853920	428
RPP2 -L	11	>	2037983	2038264	281
RPL13 (p)	16	<	2934239	2934474	235
RPL13A (p)	19	>	3654213	3655744	1531
RPS17 (p)	15	>	2934239	2934474	235
RPL23A (p)	17	<	3799447	3799917	470
RPL7A (p)	9	<	3876031	3876829	798

Table 3.1: The distribution of ribosomal protein genes in the extended MHC class I region. Columns reveal the orientation of the gene (telomere to centromere) and the position and orientation of the gene within the consensus sequence. The size is calculated according to the predicted start and stop positions within the consensus. The chromosome localization shows where the functional version of the gene is located according to Kenmochi *et al.* (1998).

3.3.4. The butyrophilin cluster.

Located centromeric of the first histone subcluster, the butyrophilin cluster is composed of 7 genes, belonging to three subfamilies (BTN1, BTN2 and BTN3) of the B7/butyrophilin-like group. Butyrophilin (BTN) is a member of the immunoglobulin superfamily, that in many species, is specifically expressed on the surface of mammary gland epithelial cells during lactation. As milk is produced, the butyrophilin protein becomes incorporated into the fat globule membrane of the milk (Jack and Mather, 1990). The human ortholog of this protein and the other BTN genes have been shown to be expressed on the cell surface in transfected cells (HeLa and CHO), but the function of these proteins in humans remains unknown (Rhodes *et al.*, 2001).

The 3 subfamilies of butyrophilin were defined according to their sequence similarity to bovine butyrophilin. BTN1A1 was defined as the ortholog of this gene: it was found to be located about 25 Kb centromeric of the other 6 genes within the cluster. The other 6 genes, from subfamilies, BTN2 and BTN3, are arranged in 3 sets of pairs; an arrangement that is likely to have arisen through the duplication of an original block of two genes, one from each subfamily (Rhodes *et al.*, 2001).

3.3.5. The tRNA cluster.

Another cluster of genes located within the extended MHC class I region is the tRNA cluster. 194 of these small single exon genes (50-100 bp in length) are located within the human extended MHC class I region. The draft genome sequence suggested there was a total of 497 human tRNA genes within the genome (IHGSC, 2001): the 194 found within the extended MHC class I region therefore represents 42.4% of the total human tRNA repertoire. The tRNAs produced by these genes correspond to 19 out of the 20 commonly used amino acids (see Table 3.2). The different types of tRNA are distributed across the region suggesting local duplications were not the major force behind the evolution of this cluster. The one missing tRNA within this cluster is the cysteine tRNA: according to the analysis of the draft genome, the majority of cysteine tRNAs (18 out of 30) are found in cluster spanning a 0.5 Mb stretch of chromosome 7.

tRNA type	No. in extended class I
tRNA-Ala	37
tRNA-Arg	12
tRNA-Asn	1
tRNA-Asp	3
tRNA-Gln	12
tRNA-Glu	1
tRNA-Gly	3
tRNA-His	5
tRNA-Ile	17
tRNA-Leu	11

tRNA type	No. in extended class I
tRNA-Lys	8
tRNA-Met	22
tRNA-Phe	9
tRNA-Pro	2
tRNA-Ser	18
tRNA-Thr	10
tRNA-Trp	2
tRNA-Tyr	4
tRNA-Val	17

Table 3.2: tRNA types and the number of genes per type found within the extended MHC class I region.

3.4. Related genes within the human MHC extended class I region

In addition to these clusters there are a number of genes which have closely related family members within the same cluster. Among the known genes this includes the POM121-like gene, the glutathione peroxidase precursor (GPX5), the Mas-related G-protein coupled receptor, and the FAT10 gene. In addition, moving into the classical MHC, the P5-1 locus has many copies that have duplicated alongside a number of MHC class I-like fragments. Other close relationships include the 2 HLA loci, HLA-H and HLA-F to the HFE locus, and the RoRet and Ret Finger Protein (RFP) also share a level of sequence similarity. Two novel genes, designated Novel_4 and Novel_5 also appear to be related to each other.

3.4.1. The glutathione peroxidase loci.

The glutathione peroxidase precursor gene, GPX5, is found in the human MHC extended class I region, together with a gene sharing 75% identity (GPXL) and a pseudogene fragment. GPX5 has been identified as an enzyme involved in protecting mammalian sperm membranes from the effects of lipid peroxidation (Vernet *et al.*, 1997, Hall *et al.*, 1998). However, glutathione peroxidase-like genes have also been isolated from the olfactory mucosa (Dear *et al.*, 1991), and it has been suggested that GPX-like genes have a function in olfactory-related biotransformations. These processes may include a function such as clearing odorants from the neuroepithelium, preventing the initiation of new olfactory signals from residual odorants, or they may act as detoxification enzymes, metabolising potentially harmful chemicals into less harmful forms. The location of GPX5 and a GPX-like gene in an area of the genome with a number of olfactory receptor genes is interesting, and it could be hypothesized that these loci are not found together by chance.

3.4.2. The FAT 10 gene and pseudogene.

FAT10 (HLA-F associated transcript 10) is a gene that encodes a protein with two domains found in the ubiquitin gene (FAT10 is also known as diubiquitin). It was first isolated as a 1.1 Kb cDNA located near the HLA-F gene (HLA-F associated transcript 10 -FAT10) in B-cell lines transformed by Epstein-Barr virus (Fan *et al.*, 1996). The full-length cDNA was isolated from dendritic cell libraries: it was named diubiquitin owing to the prediction of 2 ubiquitin-like domains in the protein structure. Ubiquitin domains are generally involved in protein degradation that is instrumental to various cellular processes, such as cell-cycle progression, transcription and antigen processing. Expression of diubiquitin was detected in dendritic cells, B cells and a kidney carcinoma cell line (Bates *et al.*, 1997), whilst immunoprecipitation studies revealed that the FAT10 protein was associated with MAD2, a protein involved in checking for spindle assembly during anaphase in the cell cycle. FAT10 may, therefore, be implicated in controlling cell growth during B cell or dendritic cell development and activation (Liu *et al.*, 1999). FAT10 has also been proposed to be involved in inducing apoptosis mediated by the tumor necrosis factor alpha cytokine (Raasi *et al.*, 2001). The relationship between FAT10 and the pseudogene found approximately 100 Kb away appears to be fairly close, with about 47% similarity detected at the protein sequence level.

3.4.3. The GPCR loci.

In addition to the olfactory receptors and the pheromone receptors which are members of the G-protein coupled receptor superfamily, the human extended class I region contains 3 other genes that are G-protein coupled receptors. Beyond this shared classification, the GTP-SARA-related gene shows little similarity to the other 2 receptors, and the 2 GPCR genes located within 5 KB of

sequence (MRG and GPCR-L) also share little similarity on the protein level. This suggests that all 3 loci have different evolutionary histories.

The mas-related GPCR was originally classified according to its similarity (35%) to the mas-related oncogene, that is involved in the physiological response to angiotensin in model systems (Monnot *et al.*, 1991). Other mas-related GPCRs were classified by Dong *et al.* (2001), but MRG was not among these. The GTP-SARA-L (GTP-S-L) gene was identified by sequence similarity to a cDNA that was retrieved from a pituitary tumour cDNA library. This cDNA was named according to its homology with the *Saccharomyces cerevisiae* SAR1 gene which codes for an essential protein required for transport of secretory proteins from the endoplasmic reticulum to the Golgi apparatus. The GPCR-L locus, meanwhile, has a generalised similarity to the G-protein coupled receptor superfamily, but no further information is provided by homology searches of this protein against the databases.

3.4.4. The POM121-like loci.

The POM121 gene was described as a gene that coded for a protein that is located specifically in the pore membrane domain of the nuclear matrix (Hallberg *et al.*, 1993). The majority of the protein was predicted to be exposed on the pore side of the pore membrane, with a nucleoporin-like domain likely to anchor components of the nuclear pore complex to the pore membrane. There are 2 loci within the human extended MHC class I region that are related to the POM121 gene, but, as with the G-protein coupled receptors within the region, at the protein level the 2 loci are not significantly similar to each other, suggesting recent local duplications have not been responsible for the 2 loci within the human extended MHC class I region, or, as these loci are both pseudogenes they may have duplicated from each other but the lack of selective pressures means that these sequences may have diverged from each other at a fast rate.

3.4.5. The RoRet gene and the Ret Finger Protein gene

The RoRet gene was initially described in the paper that detailed the mapping of the HFE locus (Ruddy *et al.*, 1997). In this paper, RoRet took its name based on the strong similarity to both the Ro/SSA lupus and Sjogren's syndrome autoantigen and the Ret finger protein (RFP). These two genes have both been characterized: Ro/SSA genes code for nucleocytoplasmic ribonucleoprotein (RNP) particles implicated in autoantigenic responses (Ben-Chetrit *et al.*, 1989), whilst the RFP gene codes for a DNA-binding protein associated with the nuclear protein involved in the activation of the ret proto-oncogene (Isomura *et al.*, 1992). Although the RoRet gene is similar to the RFP gene, however, the RoRet gene also shares a similar amount of identity with the butyrophilins so the two genes, RFP and RoRet cannot be regarded as a subfamily within the extended MHC.

3.5. Other known genes within the human MHC extended class I region

Single copy genes within the human extended MHC include NPT3, a sodium phosphate transporter (Ruddy *et al.*, 1997), HMG17L3, a member of the high mobility nonhistone chromosomal protein group, and HABT1, a basal transcriptional activator. Interestingly, HMG17L3 and HABT1 are both considered to play a role in transcription and they are located within 50 Kb of each other. HMG17-like genes are thought to be able to confer specific conformations to transcriptionally active regions of chromatin (Landsman *et al.*, 1986). HABT1 was identified in mouse (mABT1) as a nuclear protein that associates with the TATA-binding protein (TBP) and enhances basal transcription of class II promoters. The close proximity of HMG17L3 and HABT1 in the extended MHC class I region may be important in triggering transcription across the MHC as a whole.

Moving further towards the classical MHC, GUSB is a pseudogene. The functional version of this gene, which codes for the beta glucuronidase enzyme, has been mapped to chromosome 7 (Knowles *et al.*, 1977). Deficiency of this enzyme in fibroblasts has been associated with an autosomal mucopolysaccharidosis (MPS VII), (Sly *et al.*, 1973) and attempts were made to localise the gene by considering the relation of chromosomal deletions to the MPS VII phenotype (mental retardation, short stature, 'coarse' facial appearance, mild skeletal involvement and recurrent lower respiratory tract infection). A more precise localisation mapped the locus to 7q11.21-q11.22 in a position proximal to the elastin gene (Speleman *et al.*, 1996). This localization, made using fluorescence *in situ* hybridisaton (FISH) did draw attention to the fact that there appear to be a number of pseudogene fragments of the GUSB gene, including 2 on 5p13 and 5q13, but the pseudogene version on chromosome 6 was not detected by this study. The functional version of the GUSB gene has also been associated as a cause of hydrops fetalis (Kagie *et al.*, 1992).

The localization of a GUSB pseudogene to chromosome 5q13 links in with another feature of this region. During mapping and sequencing, this was the most difficult part of the extended MHC class I region to map, and several clones originally designated as chromosome 6 clones were reassigned to chromosome 5. These clones were all largely associated with the region of chromosome 5 (5q11.2-13.3) considered to be involved in chronic childhood-onset spinal muscular atrophy (SMA) (Brzustowicz *et al.*, 1990). Gene sequences isolated from this region showed sequence homologies to exons of beta-glucuronidase, and in addition, putative gene sequences showed a complex, repetitive arrangement. This arrangement appeared to be polymorphic between individuals, suggesting that this SMA may be caused by novel genomic rearrangements arising from aberrant recombination events (Theodosiou *et al.*, 1994). The relationship of this region on chromosome 5q13, the region on chromosome 5p13 (which also shows the same complex repetitive arrangement of putative gene sequences), and the region in

the extended class I region is clearly something that requires further investigation as recombination within and between these regions may be implicated in a number of diseases.

Other single copy genes in the human MHC extended class I are the PRSS16 gene, a recoverin-like gene, a HNRNPA1 pseudogene, a HB15L pseudogene, a TOB2-L pseudogene, a COX11 pseudogene, a LAMR-L pseudogene, and a NOP56L gene. Of the 3 functional genes, the PRSS16 gene codes for a serine protease enzyme that, in mice, is expressed specifically within the thymus (Carrier *et al.*, 1999), whilst the NOP56L gene is a nucleolar protein, similar to the NOP56 gene isolated from Drosophila melanogaster (Adams *et al.*, 2000).

The third functional gene, the recoverin-like gene is related to the recoverin gene that is specifically expressed in the retina and encodes a protein with 3 calcium binding sites (Dizhoor *et al.*, 1991). Recoverin was shown to activate guanylate cyclase when the amount of free calcium in the cell was lowered. This ability to respond to calcium concentrations is a property it shares with several related other proteins; for example, visinin which may be involved in the calcium dependent regulation of rhodopsin phosphorylation (Yamagata *et al.*, 1990) and neurocalcin which is expressed in the rat olfactory bulb, together with other calcium binding proteins (Brinon *et al.*, 1999). Neurocalcin is also expressed in the rat accessory olfactory bulb (Porteros *et al.*, 1996). The similarity between this recoverin-like gene found in the extended MHC and neurocalcin which appears to have a role in the olfactory bulb suggests that this recoverin-like gene could have a role in the olfactory system, possibly reinforcing action potentials generated within the system through its response to calcium concentrations.

The pseudogenes found in the extended class I region were compared against functional versions to consider the potential ‘old’ functions of these loci. The HNRNPA1 gene, for example, would code for a heterogeneous nuclear ribonucleoprotein (hnRNPs) (Biamonti *et al.*, 1994). These

proteins are a large family of nucleic acid binding proteins that are often found in, but not restricted to, the 40S-ribonucleoprotein particle. HNRNP1A is a polypeptide that appears to be involved in binding nascent hnRNA in the nucleus to form the so called hnRNP complexes which are involved in pre-mRNA processing and in the export of mRNA from the nucleus to the cytoplasm. After exportation in the complex, the hnRNPs are immediately re-imported back into the nucleus (Weighardt *et al.*, 1995). The existence of this pseudogene within the extended MHC is interesting given the number of ribosomal protein pseudogenes within this sequence; it could be suggested that the functional version of this gene formed part of a transcription-translation complex of genes that existed in this region of the genome at one stage of evolution.

The HB15L pseudogene on a protein level is similar to the CD83 antigen. The gene was isolated as a cDNA coding for a 205-amino acid protein containing a pair of cysteine residues in positions to permit the disulfide bonding that creates an Ig-like domain (Zhou *et al.*, 1992). CD83 expression was observed in lymph nodes, spleen, tonsils, scattered interfollicular cells, and in a subpopulation of dendritic cells in the epidermis. Further work showed that CD83 binds to a 72-kD ligand containing sialic acid, which led to the classification of the molecule as an adhesion receptor belonging to the SIGLEC family (Scholler *et al.*, 2001). Human CD83 was mapped to 6p23 (Olavesen *et al.*, 1997), a location which is supported by the sequence data (it is located on clone AL133259, in the 6p23 region). The mouse version of the gene has been mapped to chromosome 13A5 suggesting the synteny of mouse chromosome 13 extends past the Hfe locus (Twist *et al.*, 1998).

The TOB2-L pseudogene is classified according to its similarity to the gene located on chromosome 22 that codes for the TOB2 protein. TOB (Transducer of ErbB-2) proteins interact with the c-erbB-2 gene product p185erbB2 and they are considered to have a role in negatively regulating cell proliferation (Matsuda *et al.*, 1996). The interactions with p185 could negatively

regulate the TOB-mediated antiproliferative pathway, resulting possibly in growth stimulation by p185. TOB has been found to be expressed in primary peripheral blood T lymphocytes and it has to be downregulated for T-cell activation (Tzachanis *et al.*, 2001). TOB2 is also involved in the negative regulation of cell proliferation, inhibiting cell cycle progression from the G0/G1 to S phase of the cell cycle. A high level of expression of TOB2 in oocytes suggested a role for TOB2 in oogenesis (Ikematsu *et al.*, 1999).

COX11 is a gene that codes for a cytochrome-c oxidase protein, a constituent of the inner mitochondrial membrane. It is thought that, like the COX10 protein, the COX11 protein may be involved in the biosynthesis of heme, a prosthetic group of the cytochrome oxidase complex. The COX11 gene was mapped to 17q22 (confirmed by sequence as 17q23.1), with the pseudogene corresponding to this locus mapped to 6p23-p22 (Petruzzella *et al.*, 1998).

The LAMR-L pseudogene is located in clone AL390196. This sequence is related to the adhesive basement membrane protein laminin which is classified as a member of the integrin family of cell adhesion molecules. Incorporation of the receptor into lysosomal membranes allowed lysosomes to attach to surfaces coated with laminin (Gehlsen *et al.*, 1988). A number of laminin receptor pseudogenes are found within the human genome: pseudogenes on chromosomes 3, 12, 14 and X were identified by Bignon *et al.*(1991) who suggested the laminin receptor belongs to a retroposon family in mammals. This explanation would explain the high number of pseudogene copies of this gene: in addition to the above locations for pseudogenic copies, laminin pseudogenes are annotated on chromosome 6 (2, plus the 1 in the human extended MHC), chromosome 1, and chromosome 20. The laminin-binding protein can also be classified as a 40S ribosomal subunit since sequences are 99% identical (Tohgo *et al.*, 1994).

Within the olfactory receptor cluster, there are a number of other single-copy pseudogenes, including a cytokeratin18-like pseudogene, a DDX6-like pseudogene, a pseudogene copy of the

TRE-like oncogene and the SMT3H2 pseudogene. The functional form of the cytokeratin 18 gene has been located on chromosome 12 (Yoon *et al.*, 1994); this encodes for a protein with a distinctive alpha-helical 'rod' domain. This rod domain is shared by all intermediate filaments (IF), a large group of proteins that are all involved in maintaining the cytoskeleton. The mutation of human cytokeratin 18 gene has been associated with cryptogenic cirrhosis (Ku *et al.*, 1997) and as a bronchial epithelial autoantigen, it has also been implicated in nonallergic asthma (Nahm *et al.*, 2002). Pseudogenic forms of the cytokeratin 18 gene appear in at least 4 other locations within the human genome.

DDX6 is a putative RNA helicase, distinguished by a characteristic Asp-Glu-Ala-Asp (DEAD) box which is 1 of 8 highly conserved sequence motifs (Akao and Matsuda, 1996). The location of the functional form of this gene in the human genome is currently unknown. The TRE-like oncogene is similar in sequence to genetic elements found to contribute to the TRE gene that was first isolated from cells transfected with human Ewing sarcoma DNA (Nakamura *et al.*, 1988). In the original identification of this gene, genetic elements from chromosomes 5, 17 and 18 recombined to form the transcripts which could be detected in a wide variety of cancer cells but not in human cells from normal tissue (Huebner *et al.*, 1988). The SMT3-like pseudogene, meanwhile, is related to the SMT3 genes found in yeast (Lapenta *et al.*, 1997): in their functional form they code for ubiquitin-like proteins which can be conjugated to other proteins, such as RanGAP1, a Ran GTPase-activating protein critically involved in nuclear transport (Kamitani *et al.*, 1998).

Finally, at the telomeric end of the OR cluster, 2 functional single-copy genes have been identified, the GABBR1 gene and the MOG gene. The GABBR1 gene encodes a protein from the family involved in the gamma-aminobutyric acid (GABAergic) neurotransmissions of the mammalian central nervous system (CNS). As in other neurotransmitter families, GABA contains

both ionotropic receptors that directly cause change in activity of cell through influencing ion flow by opening ion channels, and it also contains metabotropic receptors that influence the activity of cell indirectly by initiating a metabolic change in the cell. The GABBR1 gene produces the $GABA_B$ receptor which is a metabotropic type of receptor that inhibits cAMP formation and inositol phosphate turnover (Kerr and Ong, 1995). The $GABA_B$ receptor is also associated with the inhibition of adenylyl cyclase activity, and it interacts with serotonin receptors within the CNS (Kasture *et al.*, 1996). As inhibitory neurotransmitters, the $GABA_B$ receptor forms have clinical relevance to a number of diseases. $GABA_B$ receptor agonists are used to treat spasticity following spinal injuries, and they can induce catatonia in rats. Specific $GABA_B$ antagonists, meanwhile, have been shown to improve cognitive processes in several animals (Mondadori *et al.*, 1993). $GABA_B$ functions also appear to produce variable effects in animal models of anxiety (Dalvi and Rodgers, 1996). On the molecular level, the $GABA_B$ receptor was first cloned in rats (Kasture *et al.*, 1996) and it was identified as a chromosome 6p21.3 gene by 2 separate groups (Goei *et al.*, 1998, Grifa *et al.*, 1998).

The MOG (myelin-oligodendrocyte glycoprotein) gene produces a protein that is a membrane molecule with one or two transmembrane domains and a N-terminal, extracellular region with the characteristics of an immunoglobulin variable domain (Pham-Dinh *et al.*, 1993). Six alternatively splicing forms of the eight exons have been identified; these isoforms differ from one another in their cytoplasmic domains and their carboxyl terminal (Pham-Dinh *et al.*, 1995, Roth *et al.*, 1995). In mammals, MOG is a minor component of the central nervous system myelin, a multilamellar membrane that ensheathes segments of axons and facilitates conduction of electrical impulses. It is expressed in the later stages of myelination by oligodendrocytes on the outermost surface of mature myelin, suggesting that its contributes to myelin maturation and maintenance.

3.6. 'Novel' genes identified within the human MHC extended class I region

In addition to the genes that have been fairly well characterised in humans or in other species, a number of other genes have been predicted to exist within the sequence. Table 2.3 shows the evidence for these genes, which ranges from EST data to the isolation of a cDNA for the gene, such as Novel_1, Novel_8 and Novel_11 which were all identified in large scale cDNA projects. (Nomura *et al.*, 1994, Seki *et al.*, 1997, Nagase *et al.*, 1998).

Gene name	Accession number	Evidence for prediction	Comment
Novel_1	AL353759	2 cDNAs: AL133034, AB018274	KIAA0731-like gene
Novel_2	AL121936	6 cDNAs: AK021459, AL050030, AK001535, J03802, AK024111, AK007344	
Novel_3	AL513548	2 ESTs: N51465, N45115	Possibly part of centromeric ZNF
Novel_4	AL591044	3 ESTs: BF089861, BG951300, AW951856	Splicing suggests the existence of 2 isoforms.
Novel_5	AL591044	1 EST: BF808163	
Novel_6	AL590062	5 ESTs: AI640588, BE673560, AI208304, AW182240, AW013982	Splicing suggests the existence of 2 isoforms.
Novel_7	AL009179	6 predicted proteins: Q9XIP7, Q9U1S3, Q17912, Q9T087, Q9V792, Q9P0S4	Similar proteins in *Caenorhabditis elegans, Drosophila, Arabadoptosis*
Novel_8	AL121944	1 cDNA: D25278	KIAA0036-like pseudogene form.
Novel_9	AL358993	1 cDNA: AB056432	Possibly part of centromeric ZNF
Novel_10	AL390721	7 cDNAs: AK011492, AK012826, AF085870, BC000940, AK014812, AK026279, AK015912	Similar proteins from *C.elegans*, hypothetical β-adrenergic fragment
Novel_11	AL358785	1 cDNA: AB007886	KIAA0462-like gene
Novel_12	AL121932	3 ESTs: AA913908, AI688709, AI187831	

Table 3.3: Novel genes in the extended MHC class I region. The table shows the name of gene, the accession number the gene is located in, the supporting evidence for suggesting this is a gene locus, and any addition comments.

3.7. The genomic environment of the human MHC extended class I region

The MHC extended class I region can be divided into 12 subsections, indicated on figure 3.2. These subsections were defined according to the type of gene they contain as entire regions contain distinct types of genes, for example, histones or olfactory receptors.

Subsections were compared according to GC content (see figure 3.2), gene density per Kb (table 3.4) and repeat content (figure 3.3). From this it can be seen that subsections defined by gene content have distinctive genomic characteristics. The histone clusters, for example, are associated with very high gene density ranging from 1 gene per 6026.1 bp to 1 gene per 11851.0 bp. The histone clusters are also associated with a GC content ranging from 40% to 50%, and with a high Alu-low LINE content. The MHC class I region is also associated with GC rich sequence (> 50% in some places in the subsection) and high Alu-low LINE content. As with the histone cluster, the gene density is fairly high with 1 gene per 13579.2 bp on average.

This analysis also highlights the distinctive genomic environment of the MHC-linked olfactory receptor genes. As small, single exon genes they have a high gene density: 1 gene per 18216.0 bp (major cluster) and 1 gene per 18286.4 bp (minor cluster). The OR clusters are also characterised by a low GC content that rarely rises above the genome average of 40%: it tends to be nearer 35%. The repeat content of the major and minor olfactory clusters is also distinctive: the percentage of LINE repeats is higher than anywhere else in the MHC extended class I region (Figure 3.3). This is particularly true in the major cluster, where over 60% of the total repeat content of the area consists of LINE repeats. Interestingly, a similar genomic environment to that associated with the major and minor clusters is also associated with subsection 4 of the MHC extended class

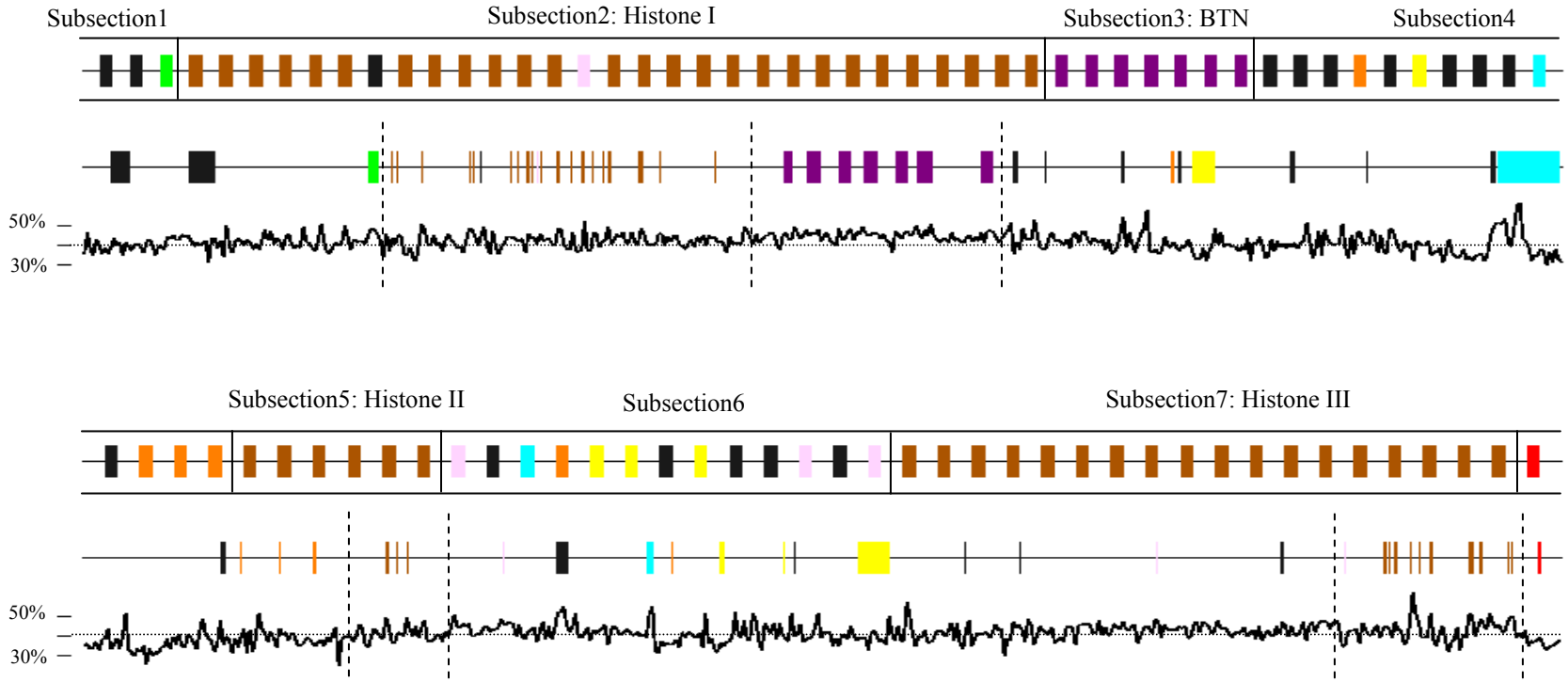
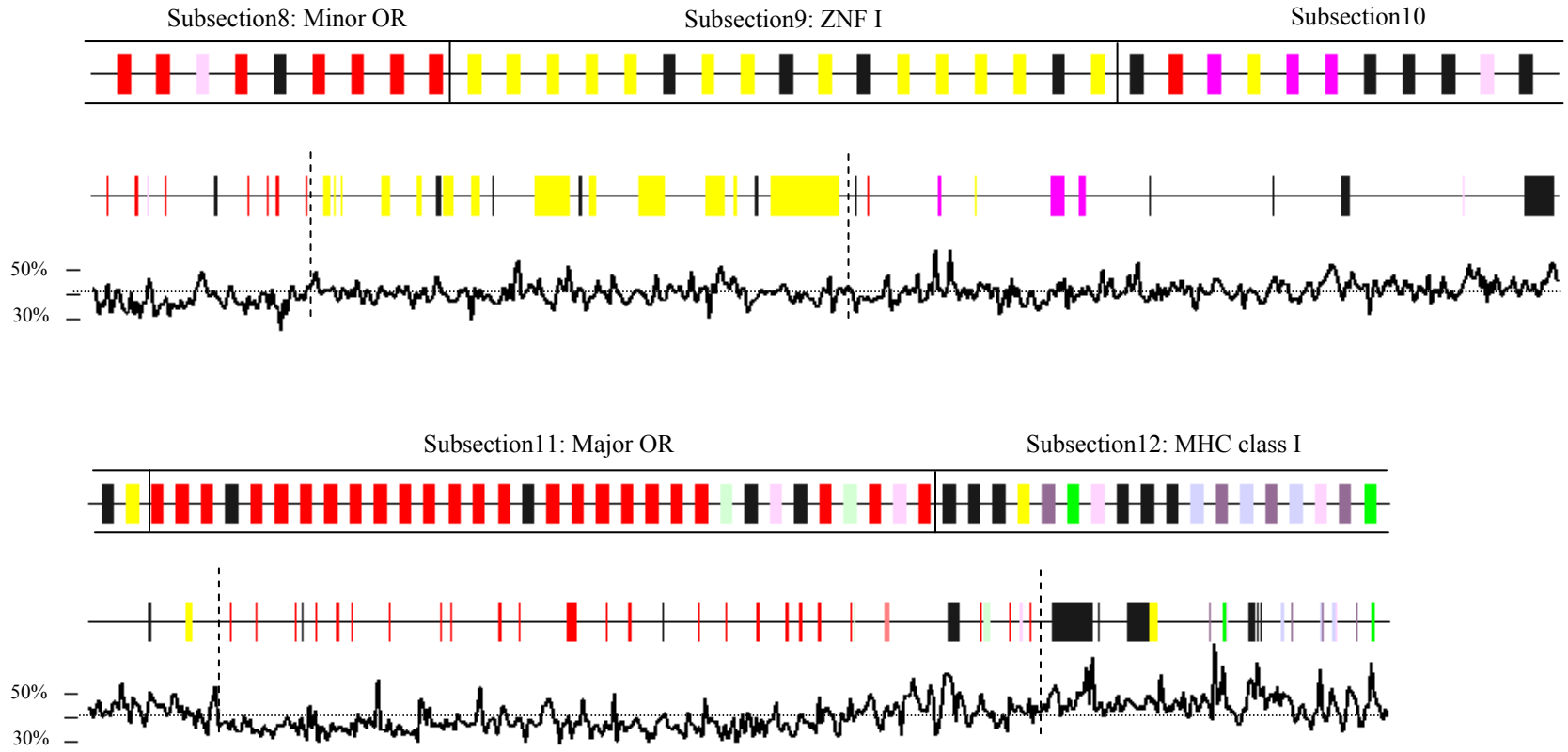


Figure 3.2 (cont. overleaf): Schematic diagram showing subsections of the extended MHC class I. The 3913358 bp sequence has been divided into 12 subsections, defined according to majority gene content. The first track shows a schematic represenation of the genes, for names refer to Figure 3.1. The second track shows the genes plotted to scale according to the length they span and the distance between this genes and their neighbouring loci. Dashed lines indicate the subsection divisions shown in the first track. The third track shows the GC content of the sequence, calculated per 2 Kb, and ranging from 30% to 50%. The dotted line indicates the genomic average GC content of 40%.


Subsection8: Minor OR
Subsection9: ZNF I
Subsection10
50%
30%
Subsection11: Major OR
Subsection12: MHC class I
50%
30%

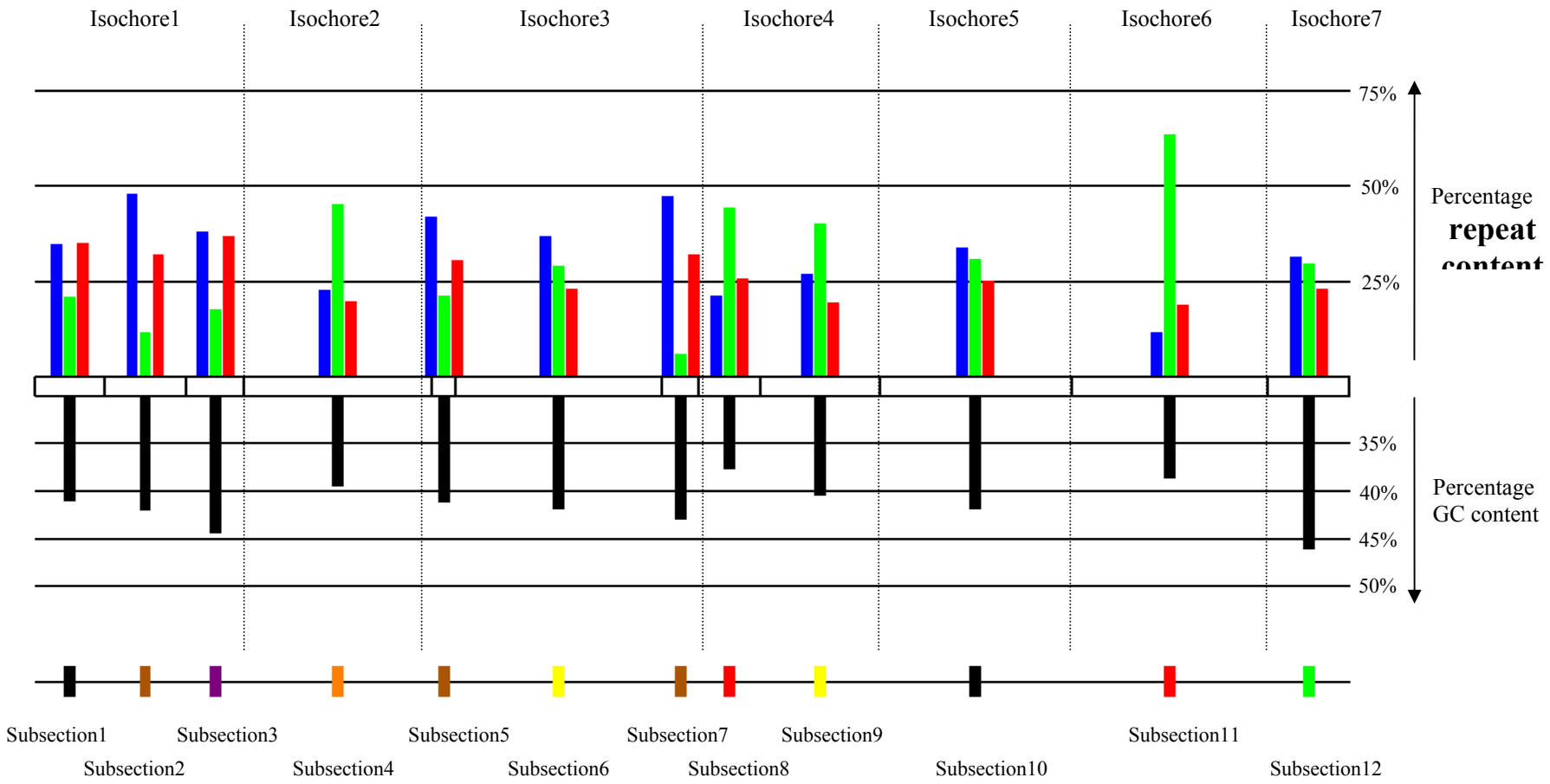


Figure 3.3: Isochores of the MHC extended class I region. Schematic diagram showing sequence containing the genes from HFE to HLA-G divided into 12 subsections, defined according to gene content. Each subsection is plotted to scale, showing how much of the sequence is composed of the various clusters. (See figure 3.1/3.2 for a detailed diagram of the gene content of the region). The blocks above the sequence line show the repeat content of each subsection, plotted according to the percentage each type of repeat contributes to the repeat content of the subsection. Blue blocks are SINE repeats, green blocks are LINE repeats and red blocks are LTR/Retroviral repeats. Below the sequence line, black blocks indicate the percentage GC content for each subsection. Dotted lines represent the divisions between the 7 potential isochores. The gene line shows gene clusters found within the subsections; colours refer to Figure 3.1/Figure 3.2.

Subsection	Size, bp	Number of genes	Gene density/bp
1	204311	3	68103.3
2	246184	29	8489.1
3	171075	7	24439.3
4	559513	15	37300.9
5	71106	6	11851.0
6	614758	13	47289.1
7	108479	18	6026.1
8	182864	10	18286.4
9	356308	17	20959.3
10	571421	13	43955.5
11	582913	32	18216.0
12	244426	18	13579.2
Total	3913358	181	21620.8

Table 3.4: Size and gene density of subsections of human MHC extended class I region. The subsections were defined in figure 3.2, whilst the gene density per subsection is calculated (number of genes/size): it takes no account of the size of individual genes.

I region. This subsection, which contains 4 VNO-type olfactory receptors, shows a GC content that generally falls below 40% and it has a high LINE-low Alu repeat content.

Across the sequence as a whole, an analysis of the repeat content (figure 3.3) and GC content (figure 3.2) suggests the sequence can be divided into 7 domains that may represent isochores (figure 3.3, indicated by dotted lines). Isochores were defined in 1976 (Macaya *et al.*, 1976) (although they were named in 1981 (Cuny *et al.*, 1981)) as long regions of DNA (longer than 300 Kb) that are fairly homogeneous in terms of base composition compared to the heterogeneity

present in other (non-satellite) DNA in the human genome. The idea was formalised by Bernardi *et al.* (1985) who suggested that warm-blooded vertebrates had a 'mosaic' genome consisting of 5 distinct types of isochore (L1, L2, H1, H2 and H3 with GC contents of <38%, 38-42%, 42-47%, 47-52% respectively.) In contrast to this view of the vertebrate genome, the draft sequence paper published by the International Human Genome Sequencing Consortium (IHGSC, 2001) suggested there was no evidence of the existence of compositional homogeneous isochores, and suggested that owing to the heterogeneity of the genome, "isochores do not appear to deserve the name 'iso.'" In response to this Bernardi *et al.* (2001) argued that the genome does contain large regions of distinctive GC content that can be used to partition chromosomes, since the denial of a compositionally discontinuous sequence organization results in the denial of "a fundamental level of genome organization" (Eyre-Walker and Hurst, 2001).

In spite of criticism of the term 'isochore', in the absence of other terms to define genomic partitions, therefore, isochore will continue to be used here to describe regions of the genome showing different GC content and a different repeat content. LINEs and SINEs have been established as repeats that are located preferentially in GC-poor and GC-rich areas respectively (Soriano *et al.*, 1981, Meunier-Rotival *et al.*, 1982, Soriano *et al.*, 1983, Zerial *et al.*, 1986, Jurka *et al.*, 1996, Smit, 1996, Jabbari and Bernardi, 1998, Smit, 1999) and so in this analysis they have been use as an additional predictor of isochores. Isochores predicted by this analysis seem to be of the predicted size of greater than 300 Kb (table 3.5: isochore 7 is not complete, it probably includes the rest of the class I region isochore described in the MHC sequencing consortium's complete sequence and gene map of a human major histocompatibility complex (The MHC Sequencing Consortium, 1999)).

Isochore	Size, bp	No. of gene loci	Percentage of pseudogenes	Size of repeat content, bp.	Percentage of SINEs/repeat content	Percentage of LINEs/repeat content
1	621570	39	15.4	277396	40.5	16.6
2	559513	14	42.9	273581	22.7	45.2
3	749343	37	27.0	418327	38.5	25.9
4	539172	27	55.6	245071	25.0	41.6
5	571421	13	30.8	303222	33.7	30.8
6	582913	32	43.8	312386	11.7	63.5
7	244426	18	66.7	121004	31.6	29.7

Table 3.5: Isochores of the human MHC extended class I region. These were defined according to GC content and repeat content (figure 3.2 and figure 3.3). For each isochore, the number of gene loci (including pseudogenes), and the percentage of loci that are pseudogenes was calculated. The (base pair) size of the repeat content, and the percentage of this repeat content that are SINEs and LINEs is also recorded.

The association of specific classes of repeats with sequences that differ according to GC content and gene content could be due to a number of mechanisms. These include selective targeting and selectional pressures leading to either retention or loss of repeat elements. Theoretically, any of these mechanisms or any combination of these mechanisms could be operating, and all of these mechanisms could have implications for the genes within a specific isochore. For example, the selective targeting of LINEs for GC-poor, AT-rich DNA may disrupt genes within GC-poor regions, leading to a higher number of pseudogenes within this type of isochore. There is some support for this idea: isochores associated with a higher percentage of pseudogenes appear richer in LINE repeats (table 3.5).

Alternatively, some type of excision process may operate to remove certain types of repeat from certain isochores. For example, isochores may be low in SINEs because SINEs in this isochore are excised by an enzymatic mechanism. This mechanism may also act to disrupt functional

genes by excising valuable sequence around the unwanted repeat. (This is something that is not supported by the higher percentage of pseudogenes in low-SINE, high-LINE areas of the MHC extended class I region). In spite of this potential disruption of functional units, however, negative selectional pressures are favoured as the mechanism by a number of authors who suggest that Alu sequences have not been fixed within human populations for a long enough period of time for positive selection to act upon these sequences (Brookfield, 2001).

It has, however, been suggested that SINEs are preferentially fixed in GC-rich DNA by positive selection (Smit, 1999, IHGSC, 2001). Schmid (1998) has suggested 3 putative functions for SINEs, 2 associated with variable SINE methylation and 1 associated with the control of protein translation. Some or all of these functions could lead to these repeats being maintained within GC-rich areas of the genome. Firstly, Alus are associated with a high proportion of CpG dinucleotides within the human genome which means they account for a substantial fraction of the genome's potential methylation sites (Britten *et al.*, 1988, Jurka and Smith, 1988). In sperm it appears, that despite the existence of an unmethylated subgroup of Alus (Hellmann-Blumberg *et al.*, 1993, Kochanek *et al.*, 1993, Rubin *et al.*, 1994), the majority of Alus are completely methylated. Complete methylation of most Alus in oocytes has also been observed, meaning embryos are likely to inherit different Alu methylation patterns from their father and their mother (Rubin *et al.*, 1994). This could mean Alus are involve in signal imprinting. Alternatively, differences in sperm methylation could be involved in directing the sequence-specific packing of two types of sperm chromatin. (85% of sperm DNA is organized by being bound to highly basic proteins called protamines and 15% is organized by being bound to histones as found in somatic cells (Gatewood *et al.*, 1987, Gardiner-Garden *et al.*, 1998)).

A third functional reason for the location of SINEs within certain regions of the genome concerns the effect Alus have on protein production. Overexpressed Alus have been found to increase protein synthesis, bind a particular protein kinase (PKR) and inhibit PKR activation (Chu *et al.*,

1998). These effects are heightened under conditions of cell stress and viral infection as normally very scarce SINE RNAs accumulate to very high levels (Panning and Smiley, 1993, Liu *et al.*, 1995, Schmid, 1998). These observations have led to suggestions that Alus act to repress the ability of PKR to inhibit protein translation, increasing the amount of protein available to the cell. This theory, therefore, suggests SINEs are retained in gene-rich GC-rich regions so, under conditions of cell stress they are able to (indirectly) promote the protein translation of these genes. Positive selection, then, may be acting on SINEs through their control of imprinting or chromatin packaging in sperm (controlled by differential methylation) or through their control of protein translation. According to this chromosome packaging-translational enhancing view of SINEs it would be expected that these repeats should be associated with transcriptionally active genes. This is supported by some genes within the MHC extended class I regions, for example, the histones would be expected to be located in a transcriptionally active area of the genome but the MHC class I region which would be expected to be transcriptionally active does not have a significant proportion of Alu repeats. The olfactory receptor genes clusters which are likely to be less transcriptionally active are found in regions containing a much lower proportion of Alu repeats.

A function for LINE elements, and therefore a selectional advantage for a region containing a large proportion of LINE repeats, has also been proposed. This function relates specifically to chromosome X inactivation, where a nearly 2-fold enrichment of LINE1 elements has been explained by the suggestion that LINE1 elements act as "boosters" to spread the X-inactivation signal (Lyon, 1998, Bailey *et al.*, 2000, Lyon, 2000). Additional support for LINEs functioning as inactivation elements is provided by the fact that genomic loci that escape X inactivation are significantly reduced in LINE1 content compared to other inactivated loci (Bailey *et al.*, 2000). Tandem reiterations of LINE1 repeats have also been shown to be able to form heterochromatin-like structures in other species, for example, in whales and dolphins, sequences with 63%

similarity to LINE1 elements make up the core of the α-heterochromatin satellite DNA (Kapitonov *et al.*, 1998) whilst in the short-tailed field vole (*Microtus agrestis*) and the Syrian hamster (*Mesocricetus auratus*) β-heterochromatin structures are enriched for LINE1 elements (Neitzel *et al.*, 1998).

This idea of a X inactivation signal being controlled by LINE content could, to some extent, be applied to the rest of the genome. It could be hypothesized that regions that are less transcriptionally active are associated with a larger proportion of LINE repeats which form complexes less accessible to transcription factors. Another hypothesis involves the association of the olfactory receptor cluster with a higher proportion of LINE repeats. The expression of one allele of one OR gene per olfactory neuron suggests a highly controlled method of regulation, including the silencing of 1 allele. By comparison with the X chromosome where one chromosome is silenced (inactivated) it may be that allelic silencing requires an inactivation of the OR cluster on 1 chromosome. LINE repeats in the OR cluster may be involved in this inactivation mechanism.

In conclusion, therefore, a number of isochores have been defined within the extended MHC. Certain classes of genes are associated with certain genomic environments, for example, the histone clusters are associated with GC-rich isochores containing a higher proportion of SINEs than LINEs. By contrast, the OR clusters are associated with GC-poor, LINE-rich isochores. The association of certain types of genes with certain genomic environments may be due to insertion or deletion mechanisms (supported by the higher percentage of pseudogenes in LINE-rich areas). Alternatively, selectional pressures may dictate that genes with a high rate of transcription (such as the histones) are associated with Alu-rich areas, whilst genes requiring allelic inactivation (such as the OR genes) are associated with LINE-rich areas. None of these mechanisms can be confirmed or refuted based on the data presented here: in any case, it may be that a combination

of all these mechanisms acts to preserve these genomic environments within the MHC extended class I region.

3.8. Duplications within the human MHC extended class I region

A large scale dot-matrix analysis of the 4 Mb region (not shown) revealed four major areas where large scale duplications appear to be involved in the formation of new genes. These areas are associated with 6 BTN genes (BTN3A2, BTN2A2, BTN3A1, BTN2A3, BTN3A3, and BTN2A1) (Rhodes *et al.*, 2001); the 2 novel genes in clone AL591044; olfactory receptor genes in the major cluster; and the P5-1 pseudogene (Kulski and Dawkins, 1999). Duplications associated with the olfactory receptor gene cluster are discussed in Chapter 4.

The lack of evidence for duplicated areas of sequence according to the dot-matrix analysis generally suggests that recent duplications have not been critical in forming the genomic environment within the MHC extended class I region. This is surprising, given the number of closely related genes found in clusters within the extended class I, and it suggests either that a different mechanism is responsible for forming these clusters, or it may be that these clusters were formed by ancient block duplications and the genomic footprints associated with these duplication events are no longer detectable.

3.9. Conclusions.

The analysis of the human MHC extended class I region revealed a number of themes giving insight into the function and evolution of this region. It is evident from the sequence that genes are organized into several clusters. The origin of these clusters could suggest that the extended

class I region is particularly permissive of local duplication events that act to create these clusters. Alternatively, clusters may have evolved through the recruitment of similar genes into the extended MHC. Evidence of recent local duplications creating these clusters has not been obtained in this study, but this does not totally exclude the idea of local duplications, since the syntenic regions of mouse chromosome 17 and 13 also appear to have gene clusters (Chapter 5, OR cluster, Chapter 9, VR cluster, histone cluster, (Albig *et al.*, 1998)). This suggested that local duplications implicated in forming these gene clusters may have occurred before the human-mouse lineages split.

Having originated through either local duplication or through the recruitment of paralogs, a second theme that appears is the clusters appear to have been conserved in the same locations over evolutionary time: indeed, recombination of these loci away from the classical MHC appears to be suppressed and to some extent, the MHC may represent an extended haplotype (Malfroy *et al.*, 1997). This could suggest a functional role for these loci that is linked to that of the MHC. Roles for a number of gene clusters within the MHC extended class I region could be hypothesized: for example, the histones, ribosomal protein, zinc finger proteins and the tRNAs may be important in the transcription and translation of the MHC. The OR gene cluster has been suggested to be involved in MHC-linked mate selection (through detection of favourable odours). Mediating against this hypothesis of an extended haplotype conserved for functional reason, synteny is not maintained within the mouse lineage and gene clusters telomeric of the OR gene cluster are likely to be found on mouse chromosome 13 rather than mouse chromosome 17. Any functional requirement to conserve the histone cluster, zinc finger cluster and tRNA cluster in the same chromosomal region as the MHC, therefore, does not exist in mouse. This lack of conservation across 2 species raises a clear counterpoint to this hypothesis that extended haplotypes exist owing to functional reasons: it may be that this assembly of gene clusters is a random occurrence and there is no selective advantage in maintaining this extended haplotype.

Another theme highlighted by this analysis of the extended class I region is the existence of isochores within the region. Distinct isochores, associated with specific groups of genes were identified. These isochores are characterised by distinct GC profiles and repeat content. It is interesting that, in general, the genes that are likely to be more commonly transcribed in cells are associated with LINE-poor, SINE-rich DNA. In contrast, the OR genes and other genes with more restricted patterns of expression are associated with LINE-rich, SINE-poor DNA. This association may be important with regards to transcriptional control, although differences in insertion and deletion of these elements may also be important in shaping these distinct isochores.