Chapter 4

Human MHC-linked olfactory receptor genes

4.1. Introduction

The analysis of the human extended MHC class I region revealed 2 clusters of MHC-linked olfactory receptor genes (Figure 3.2). These 2 clusters were designated the major and the minor cluster according to the relative sizes of clusters (Younger *et al.*, 2001). Olfactory receptor genes encode 7 transmembrane proteins with similarity to the G-protein coupled receptor superfamily that are thought to be selectively expressed in olfactory sensory neurons where they are involved in the binding of odorants, triggering action potentials in olfactory sensory neurons (Buck and Axel, 1991, Buck, 1992). OR genes are commonly arranged in clusters on most chromosomes throughout the human genome (Ben-Arie *et al.*, 1994, Glusman *et al.*, 1996, Vanderhaeghen *et al.*, 1997, Carver *et al.*, 1998, Trask *et al.*, 1998). The existence of a potential MHC-linked cluster was first reported by Fan *et al.* (1995).

The availability of the genomic sequence of both these clusters allowed a further investigation of MHC-linked OR genes. Each cluster was analysed with regard to the number of 1 Kb-long exons that code for genes and pseudogenes, the relationship of these genes to other OR genes within the extended MHC class I region, and the conservation of amino acids within olfactory receptor proteins in the cluster. OR genes have been classified into families and subfamilies based on their shared sequence identity (Ben-Arie *et al.*, 1994). Mechanisms creating the OR pseudogenes within each cluster were also considered, as well as any local duplications that could have been responsible for producing new OR loci. Repetitive genomic elements have been suggested to have a critical role in mediating duplication events creating new ORs (Glusman *et al.*, 1996). In

addition, the genomic environment of the MHC-linked ORs was considered. The genomic environment of OR genes may provide clues as to how the highly controlled expression of olfactory receptor genes in olfactory sensory neurons is created and maintained (Chess *et al.*, 1994, Malnic *et al.*, 1999).

4.2. Identification of olfactory receptor genes and subfamilies

Olfactory receptor genes were identified using their sequence similarity to other OR genes within a database of OR genes ('ROLF', Chaper 2). Within the human MHC extended class I region, 34 olfactory receptor genes were found. Of these 34 genes, 8 (-10, -29P, -30P, -31P, -32, -33P, -34P, and –35) are located within the minor cluster with the remaining 26 found in the major cluster which is located just telomeric of the classical MHC class I region. The major and minor cluster distinction was made primarily by taking into account the size of the 2 regions (approximately 800 Kb and 200 Kb respectively), but a distinction could also be made by taking into account the synteny breakpoint between mouse and human. The minor cluster is located between HFE and RFP, whilst the major cluster is located between HLA-F and RFP. This means that the minor cluster is not linked to the MHC in mouse, as the mouse MHC is located on mouse chromosome 17 but mouse Rfp and mouse Hfe are located on chromosome 13 (Szpirer *et al.*, 1997, Yoshino *et al.*, 1997). The major OR gene cluster is therefore the only olfactory receptor cluster that is MHC-linked in human and mouse (and rat).

The 34 MHC-linked ORs can also be divided into genes and pseudogenes. From the genomic sequence, 15 of the OR genes appear to have complete open reading frames (defined according to criteria outlined in Chapter 2), whilst the other 19 are disrupted in some way (Appendix 6). The ratio of genes to pseudogenes across the 2 human MHC-linked OR clusters is therefore, 0.8 which is significantly higher than the genome average of 0.3 (Rouquier *et al.*, 1998). Discounting

the minor cluster, which has a gene to pseudogene ratio of 0.6, the major MHC-linked OR cluster has an even higher gene to pseudogene ratio of 0.9.

These 34 olfactory receptor genes can be allocated to subfamilies according to their protein sequence similarities. Sequence similarities within the cluster generally range from 40% upwards: in this analysis, a similarity greater than 70% was considered as a cut-off value for OR genes to belong to a subfamily. This cut-off was assigned because the majority of the OR genes in the cluster had a shared protein identity of 50-60% with the other ORs, so the 70% value allowed only the closest relationships to be considered. According to the 70% cut-off the majority of OR genes within the major and minor clusters are isolated genes lacking closely related subfamily members but 5 subfamilies containing 14 of the 34 OR genes in the human MHC extended class I region were defined. (Table 4.1).

Subfamily	Subfamily member genes			
1	hs6M1-1	hs6M1-10	hs6M1-32	
2	hs6M1-3	hs6M1-4P	hs6M1-5P	hs6M1-6
3	hs6M1-12	hs6M1-13P	hs6M1-16	
4	hs6M1-19P	hs6M1-20		
5	hs6M1-23P	hs6M1-24P		

Table 4.1: Subfamily designations of human MHC-linked OR genes.

Subfamily designations are also supported by the phylogenetic tree showing proposed evolutionary relationships between the OR genes in the MHC extended class I region (Figure 4.1). In this phylogenetic analysis, only branches with a bootstrap value greater than 70 were considered to be significant. The 5 subfamily groupings can all be seen to have significant relationships with other genes in their subfamily. In contrast to this, in cases such as hs6M1-30P and hs6M1-34P where the 2 genes end up clustered closely together in the tree, the lack of a

significant bootstrap value meant, in this analysis, the implied close relationship between the 2 genes was not considered to be evolutionarily important. Protein sequence similarities between these genes (64.2%) support this lack of a subfamily designation.

In addition to the significant branches associated with the subfamilies, significant branch points are also indicated in 2 other places on the tree (figure 4.1, marked by 'A' and 'B'). Branch point A suggest that 4 genes, hs6M1-35, hs6M1-27, hs6M1-20 and hs6M1-19P have a distinctive evolutionary history from the other 30 OR genes in the extended MHC. Branch point B, meanwhile, suggests that of these 30 genes, hs6M1-3, hs6M1-4P, hs6M1-5P and hs6M1-6 are significantly distant from the other 26 genes in the cluster. From the phylogenetic tree, therefore, the ancient pattern of evolution (with the exception of the creation of the subfamilies which presumably occurred later in evolutionary time) that can be predicted is shown in figure 4.2. Alternatively, the lack of shared history between the 3 precursors of hs6M1-27, hs6M1-35 and hs6M1-20/19 with the other 30 OR genes may suggest these genes are recent insertions into the cluster from other regions of the genome. Evidence from the mouse extended MHC (Chapter 5) and from other olfactory receptor genes (Chapter 8), however, suggest this is not the case: these genes appear to have been part of the extended MHC for a significant period of evolutionary time.

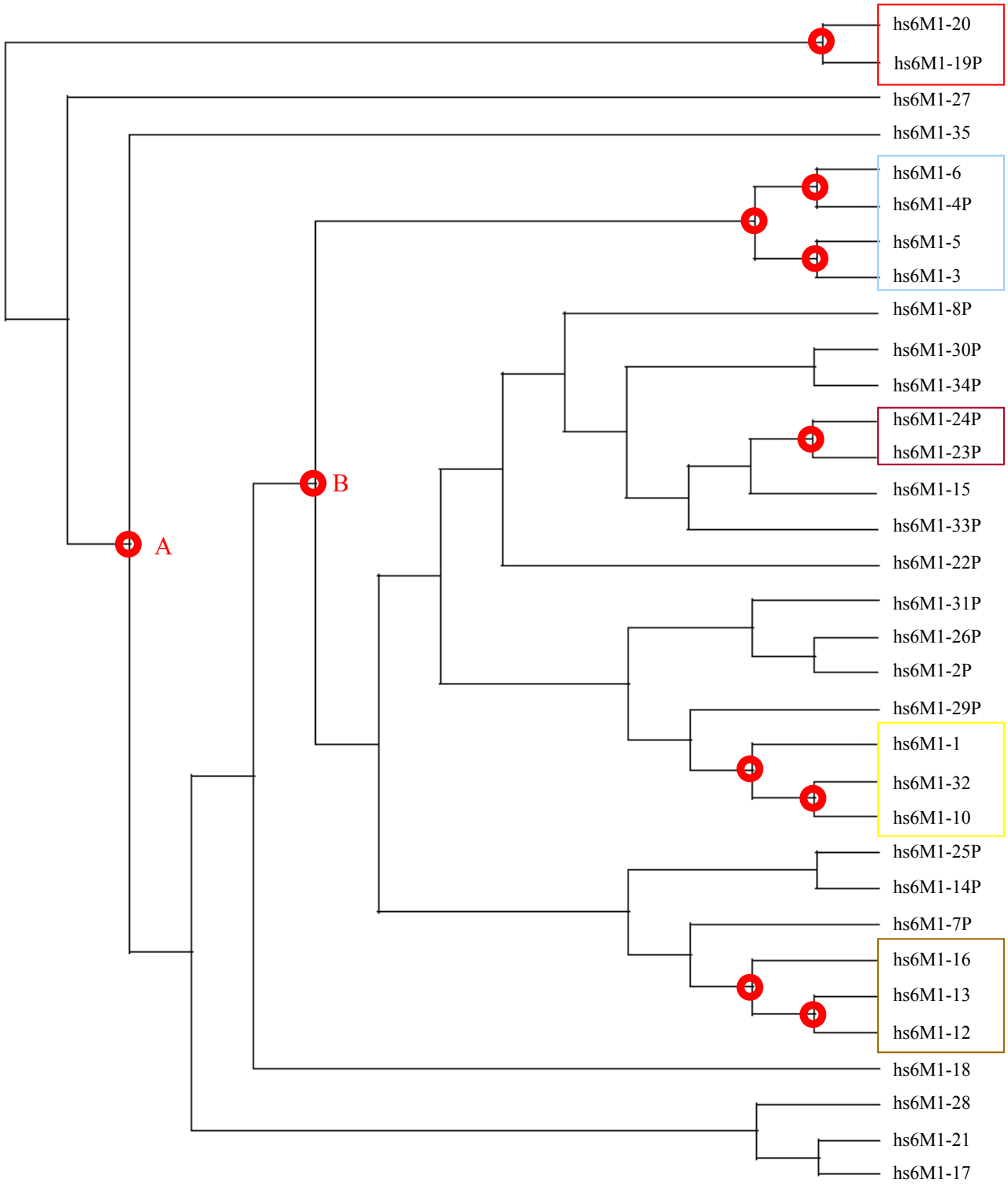


Figure 4.1: Phylogenetic tree (parsimony method) of human MHC-linked OR genes. 175 sites were used and 250 bootstrap replicates were performed. Subfamilies are boxed in different colours, whilst the red rings at branch points indicate where bootstrap values are over 70%.

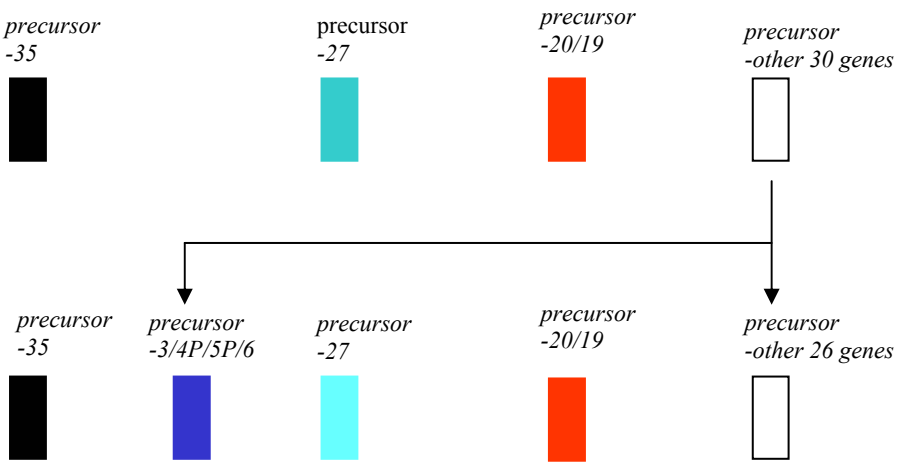


Figure 4.2: A proposed model of evolution for the MHC-linked OR genes, based on significant branch points from the phylogenetic tree (figure 4.1).

4.3. Conservation of amino acids in olfactory receptor proteins

A protein alignment of the 34 MHC-linked OR genes (Appendix 7) was also used to analyse the conservation of amino acids across the cluster. This alignment was used to create a consensus sequence based on shared amino acids identities across the 34 proteins. Within the consensus, the starting methionine was defined as the position where 10 of the 34 (29.4%) proteins have their starting methionine. With the exception of hs6M1-26P (a fragmented OR), all the other ORs in the cluster actually have starting methionines that are further upstream: generally the genes start 1-4 amino acids further upstream. The starting methionines of hs6M1-28 and hs6M1-25 are further upstream of this starting methionine (20-21 extra amino acids). This extended amino terminus may have functional implications for these two proteins. It has been suggested for some G-protein coupled receptors, such as the V2Rs (Matsunami and Buck, 1997) and metabotropic glutamate receptors (mGluRs) (O'Hara *et al.*, 1993, Takahashi *et al.*, 1993), that a large amino terminus plays a role in ligand binding. A functional role for the extended amino termini in these 2 ORs is therefore something that should be considered: however, both of these OR exons also have other methionines located closer to the first motif ('FILLG') that could also represent the

start of the gene. Translation for both of these genes could start at these methionines creating a protein the same length as the majority of human MHC-linked OR genes.

Alternatively, it may be that both putative starting methionines are used by these OR genes. Alternative translational start sites have been found within the subtelomeric olfactory receptor gene, 'OR-A', although these rely on splicing and the starting methionines are located within alternative 5' exons rather than within the 1 Kb major coding exon (Linardopoulou *et al.*, 2001). Nevertheless, the results from OR-A support the idea that olfactory receptor genes can have alternative forms with different length amino termini.

The carboxy terminal of the consensus sequence was shortened to the last position where the consensus protein shared an amino acid with over 25% identity to the alignment. The majority of the OR proteins (21 out of 34) end within 13 amino acids of this amino acids; hs6M1-10 is exceptional in having a very long carboxy terminal that extends 50 amino acids past this point. The pseudogene hs6M1-26P and hs6M1-35 also have longer carboxy termini that extend by 29 amino acids. As with the extended amino termini, these larger carboxy termini are predicted to have structural and functional implications for these 3 genes.

Figure 4.3 shows the consensus protein with amino acids in different colours according to the level of conservation at that position within the alignment. The positions of transmembrane domains were predicted using the program 'Tmpred' (Chapter 2). According to the predicted structure of this protein, high amounts of conservation exist throughout the protein, notably in the amino terminal (indicated by the 'A', figure 4.3), transmembrane domain (TM) 2, transmembrane domain 6, transmembrane domain 7 and in the second half of transmembrane domain 5. Buck and Axel (1991) suggested transmembrane domains 3, 4 and 5 were 'hypervariable' and represented the ligand binding domains of the protein: in the MHC-linked ORs there is variability in TM4,

but the variability in the second half of transmembrane domain 5 and transmembrane domain 3 is less pronounced and transmembrane domain 1 shows a greater amount of variability than would be expected from this model.

The conservation profile of the MHC-linked olfactory receptor proteins can also be used to predict amino acids that may be structurally important. As membrane proteins, olfactory receptor proteins fall into a category of proteins that have little structural information attached to them. This lack of structural knowledge is due to the difficulties involved in applying conventional methods of structure determination such as solution

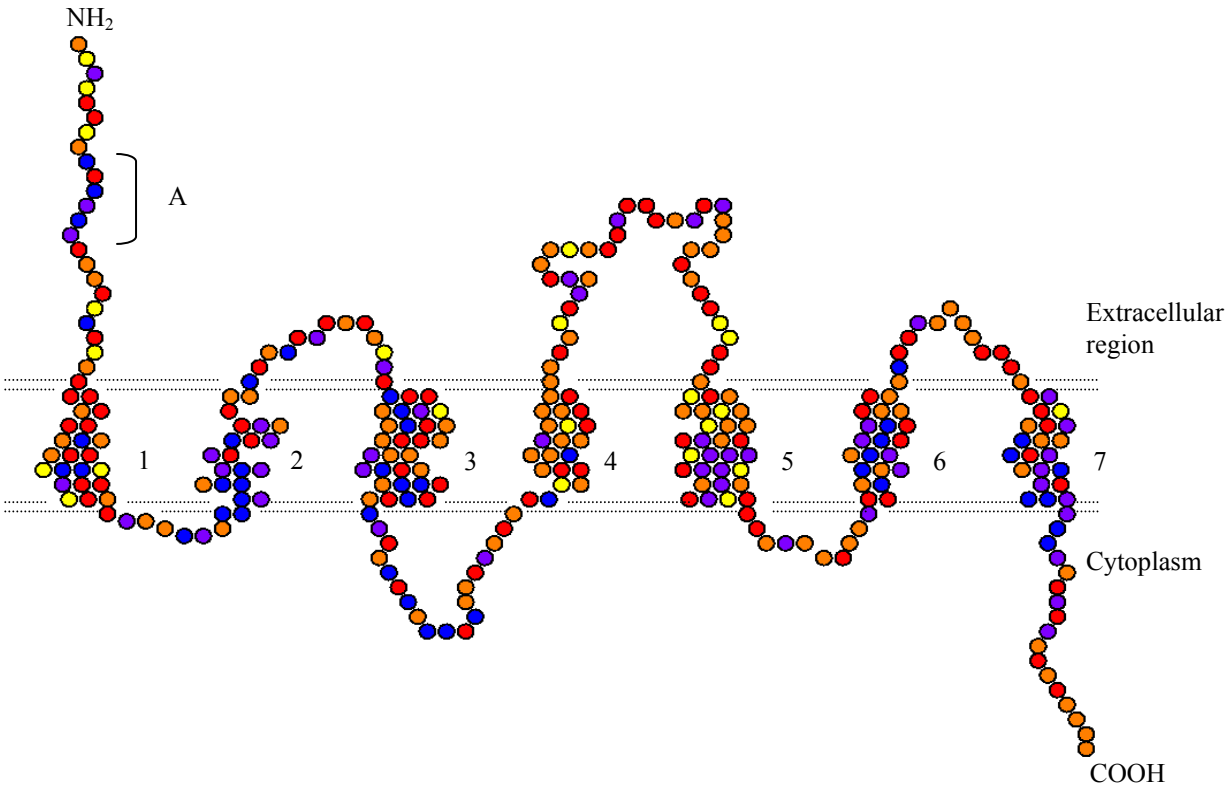


Figure 4.3: Amino acid conservation in human MHC-linked ORs. Schematic diagram showing the conservation of amino acids at predicted positions within a consensus human MHC-linked olfactory receptor protein. The degree of conservation ranges from 90%+ (blue), 75-90% (purple), 50-75% (red), 25-50% (orange) and less than 25% (yellow).

nuclear magnetic resonance (NMR) and x-ray crystallography to transmembrane proteins. These difficulties stem from the hydrophobic nature of transmembrane proteins: they do not easily dissolve, and any structure obtained in solution is likely to be different from the structure the protein adopts in the membrane. Limitations in the amount of structural data available about G-protein coupled receptors, therefore, mean it is not possible to construct an accurate structural representation of olfactory receptor proteins from the sequence data.

One structure OR proteins can be compared against is the rhodopsin structure (Meng and Bourne, 2001, Sakmar, 2002). This was the first GPCR to have its 3D structure elucidated at a high resolution (2.8 angstrom) (Palczewski *et al.*, 2000). The consensus OR protein shares 1 pair of cysteines with the rhodopsin structure (Figure 4.4). In rhodopsin, the disulphide bridge formed by these 2 cysteine residues acts to stabilize the second extracellular loop which appears to form a 'cap' to the pocket formed by the transmembrane domains in the inactive state. It has also been suggested that the main role of this cap might be to regulate the stability of the active state of the receptor; if this were the case this loop might also be involved in ligand interactions. In the consensus OR protein, there are a number of highly variable residues within this loop (Figure 4.3), suggesting this loop may play a role in ligand interactions.

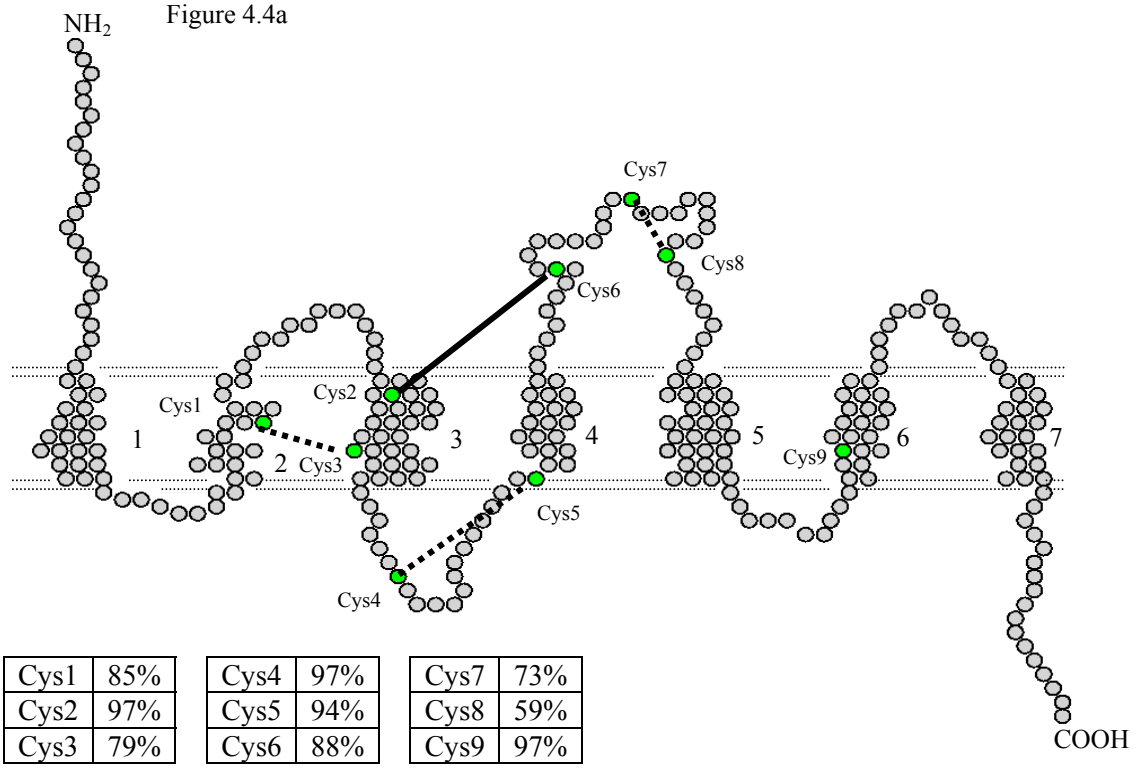
Other cysteines within the consensus OR protein are highly conserved across the MHC-linked ORs, suggesting they may play a structural role. On average, each olfactory receptor protein contains 11 cysteine residues, although the number ranges from 8 to 15 in other proteins. In the consensus sequence, 9 cysteine residues are conserved in more than 50% of the proteins. The positions of these residues and the percentage of MHC-linked olfactory receptor proteins containing a cysteine in this position are shown in Figure 4.4. As with the rhodopsin protein, disulphide bridges may act to stabilize the pocket that is created for ligand binding. Disulphide bridges in other GPCRS have been found to be critical for ligand recognition and membrane

trafficking (Le Gouill *et al.*, 1997, Blanpain *et al.*, 1999, Zeng and Wess, 1999). These disulphide bridges have been predicted given the relative conservation of cysteines in this position across the MHC-linked OR genes, however, another study based on a multiple sequence alignment of 197 ORs from human, rat, mouse, dog and fish predicts disulphide bridges between 2 cysteines (Cys 7 and Cys8 in figure 4.4) in extracellular loop 2 and between 1 cysteine in intracellular loop 2 (Cys4 in Figure 4.4) and intracellular loop 3 (there is no corresponding conserved cysteine in the MHC-linked ORs) (Sharon *et al.*, 1998). Discrepancies between predicted disulphide bridges in these 2 models suggest multiple sequence alignments can produce hypotheses, but conclusions require experimental work on the structure of this family of genes.

Other predictions about the structure of OR genes can be made from the conservation of a short stretch of amino acids into the amino terminal of the consensus gene. This short stretch of amino acids is associated with the predicted formation of a β-strand structure (predicted using by the programs DSC and Simpa96) but the role of this stretch of amino acids is unknown. However, all the MHC-linked OR genes do contain a predicted N-linked glycosylation site (NXT/S). This has been observed to exist within all OR human proteins (Zozulya *et al.*, 2001). Within the carboxyl terminal it has been suggested that 80% of all functional human ORs have a consensus sequence for phosphorylation, consisting of 2 serine or threonine residues located in the vicinity of positively charged amino acids. This does not appear to be the case for the human MHC-linked ORs where 15 (of which 9 are pseudogenes) of the 34 genes have less than 2 serine or threonine residues in their carboxyl terminal. The MHC-linked OR genes, however, do conform to the rule that only about 25% of human ORs have a cysteine in their carboxyl terminal (10 out of 34: 29%). Cysteines in the carboxyl terminal have been implicated in palmitoylation in other GPCRs (rhodopsin and the β_2-adrenergic receptor) (Zozulya *et al.*, 2001). It may be significant that the

genes with the longer carboxyl termini (-10, -26P, and –35) possess both putative palmitoylation and phosphorylation sites.

Figure 4.4a



Cys1	85%	Cys4	97%	Cys7	73%
Cys2	97%	Cys5	94%	Cys8	59%
Cys3	79%	Cys6	88%	Cys9	97%

Figure 4.4b

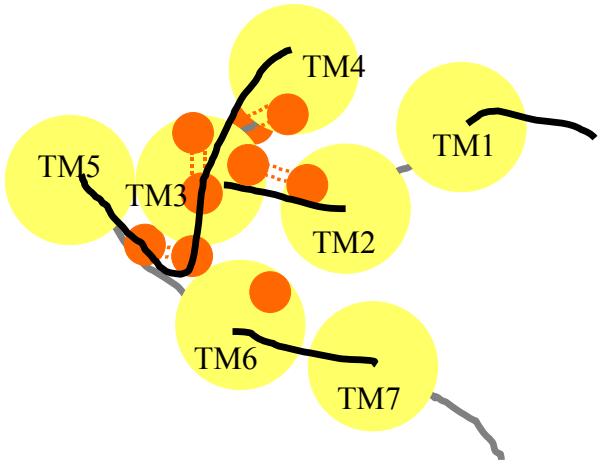


Figure 4.4a. Cysteine conservation in the human MHC-linked ORs. Schematic diagram showing where cysteines are conserved within a consensus human MHC-linked olfactory receptor protein. The percentage of OR proteins containing a cysteine in this position is shown in the table underneath. The disulphide bridge shared with the rhodopsin protein is indicated by a bold line, whilst a dotted line indicates hypothetical disulphide bridges.

Figure 4.4b. Proposed OR structure showing potential disulphide bridges. The large yellow circles represent transmembrane domains. Black lines show amino acid stretches linking TM domains on the extracellular side of the plasma membrane whilst grey lines show amino acid stretches linking TM domains in the cytoplasm. The small orange circles represent cysteine residues. Potential disulphide bridges between these residues are shown by the dotted orange lines.

4.4. Human MHC-linked olfactory receptor pseudogenes

19 of the human MHC-linked ORs are pseudogenes (Appendix 6). Of these, hs6M1-9 and hs6M1-26 are fragments of OR pseudogenes with numerous stops and frameshifts but at the other extreme, 10 OR pseudogenes only have 1 mutation that prevents them from having open reading frames (through frameshifts or stop codons). There is a possibility, therefore, that these pseudogenes exist as functional genes in other individuals (see Chapter 7, Ehlers *et al.*, 2000). Of the 19 pseudogenes, therefore, there is a chance that over half could be functional in other individuals.

The distribution of mutations within the pseudogenes suggests that there are mutational hotspots within OR genes. In total (excluding the fragments hs6M1-9P and hs6M1-26P), 19 mutations are observed across the 318 amino acid OR consensus protein. 9 of these mutations are located within 2 regions of the protein (positions 24-39: 4 mutations, and positions 180-200: 5 mutations). 47% of the mutations, therefore, are located within 11% of the protein suggesting mutations within OR genes are not random events occurring equally across the gene. This is also supported by the fact that 2 pseudogenes share a frameshift (hs6M1-30P and hs6M1-31P at position 108) within the consensus OR protein sequence. This frameshift appears to have evolved independently in the 2 OR genes. (Figure 4.5 shows a deletion has produced the hs6M1-30P frameshift whilst with hs6M1-31P, the insertion of a guanine has produced the frameshift). Mutational hotspots within OR genes are considered further in Chapter 8.

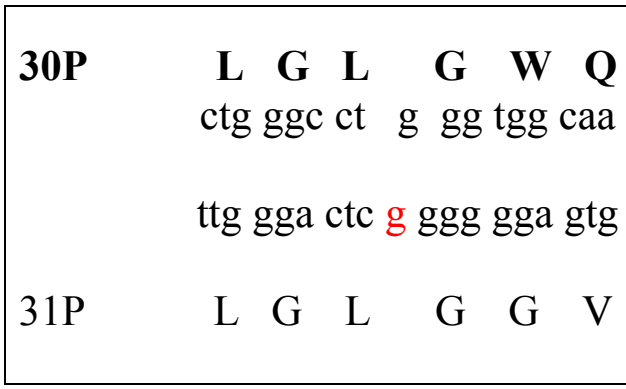


Figure 4.5: A comparison of the mutations causing frameshifts in hs6M1-30P and hs6M1-31P. 2 different mutations are responsible for changes at the same amino acid position. A deletion of guanine is likely to have disrupted hs6M1-30P whilst a guanine insertion has disrupted hs6M1-31P.

4.5. The genomic environment of the human MHC-linked olfactory receptor genes

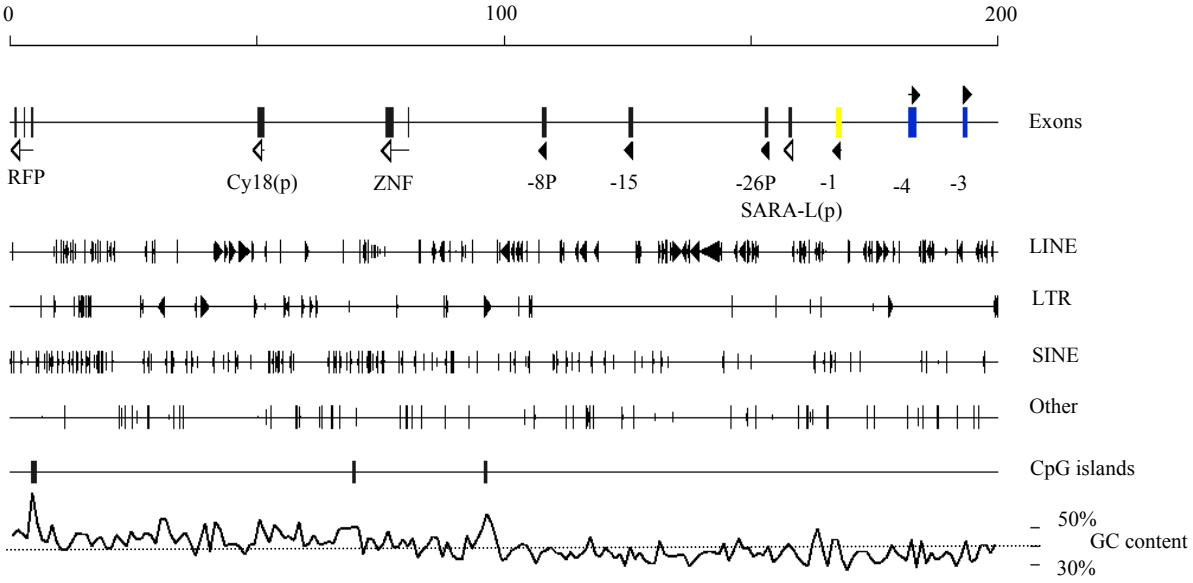
Figure 4.6 shows the genomic environment of the human MHC-linked olfactory receptor genes in the major cluster and the minor cluster. In contrast to some OR clusters, where it was been reported that OR sections of the genome form an exclusive environment, containing no other non-OR genes (Glusman *et al.*, 1996), the MHC-linked OR clusters contain a number of other genes including FAT10 (Liu *et al.*, 1999), the human counterpart of Zfp57 (Okazaki *et al.*, 1994), a novel Mas-like G-protein coupled receptor (Mas-GPCR-L), a novel zinc finger protein (ZNF311) and a number of pseudogenes (Younger *et al.*, 2001).

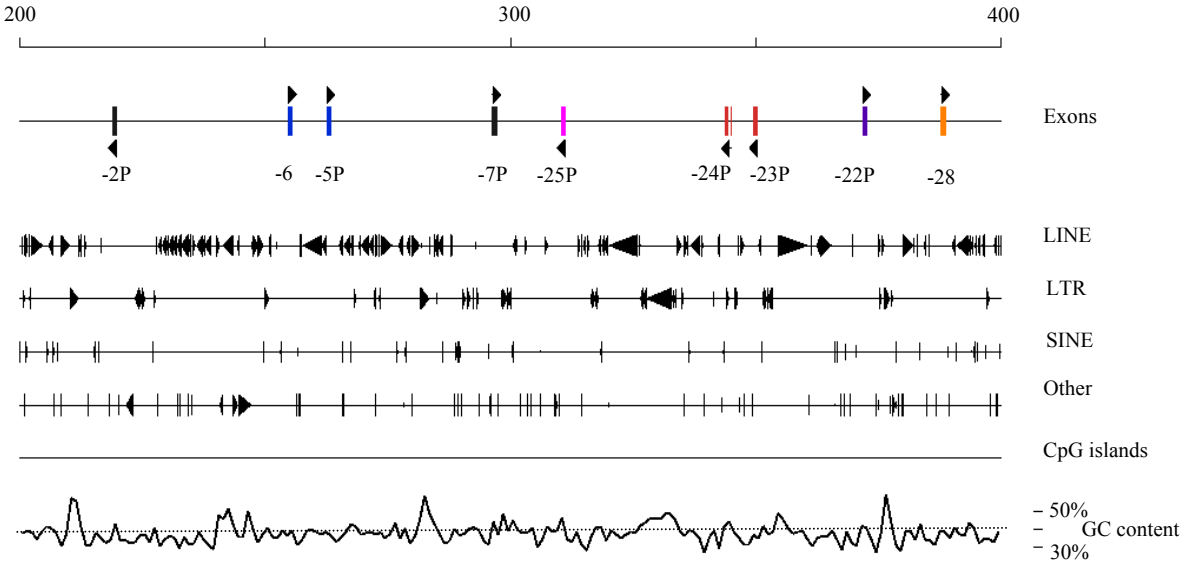
As has been previously discussed (Chapter 3), the human MHC-linked OR genes are located in distinct isochores associated with a low GC content (generally the GC content does not rise above 40%) and a high number of long interspersed nuclear elements (LINEs). This trend is particularly pronounced within the olfactory specific region of the major olfactory cluster sequence (100 Kb-500 Kb in figure 4.6); GC content picks up in the area around the FAT10 pseudogene, whilst LINEs decrease in frequency with SINEs increasing around the FAT10 gene.

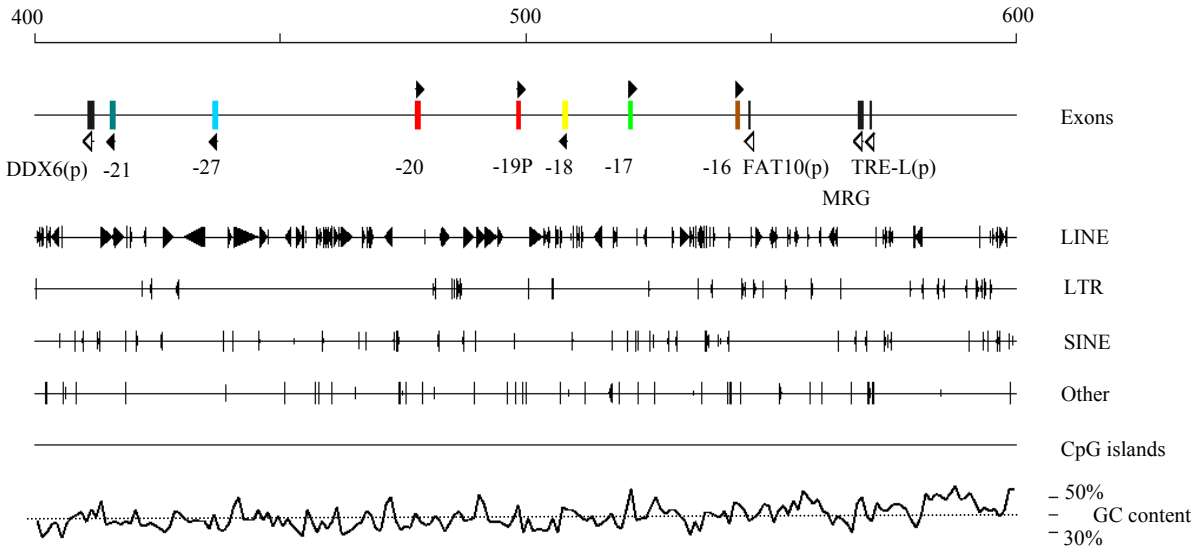
8 CpG islands are identified within the MHC extended class I region analysed here. Of these, the first 3 appear to be associated with RFP, the cytokeratin 18 pseudogene and the zinc finger protein gene that all exist telomeric of the major OR cluster. 2 other CpG islands can be attributed to the GABBR1 gene. This gene has 2 experimentally-proven alternative splice forms (Schwarz *et al.*, 2000), the start sites of which correspond the positions of these 2 islands. Centromeric to the GABBR1 gene, 2 CpG islands are likely to be associated with a zinc finger protein gene and a P5-1 pseudogene. This leaves 1 CpG island within the major OR cluster, but the position of this gene suggests a possible involvement with the Tre/Mas1 oncopseudogene or with the mas-related GPCR. CpG islands, therefore, do not seem to play a role in transcriptional control of the major OR cluster.

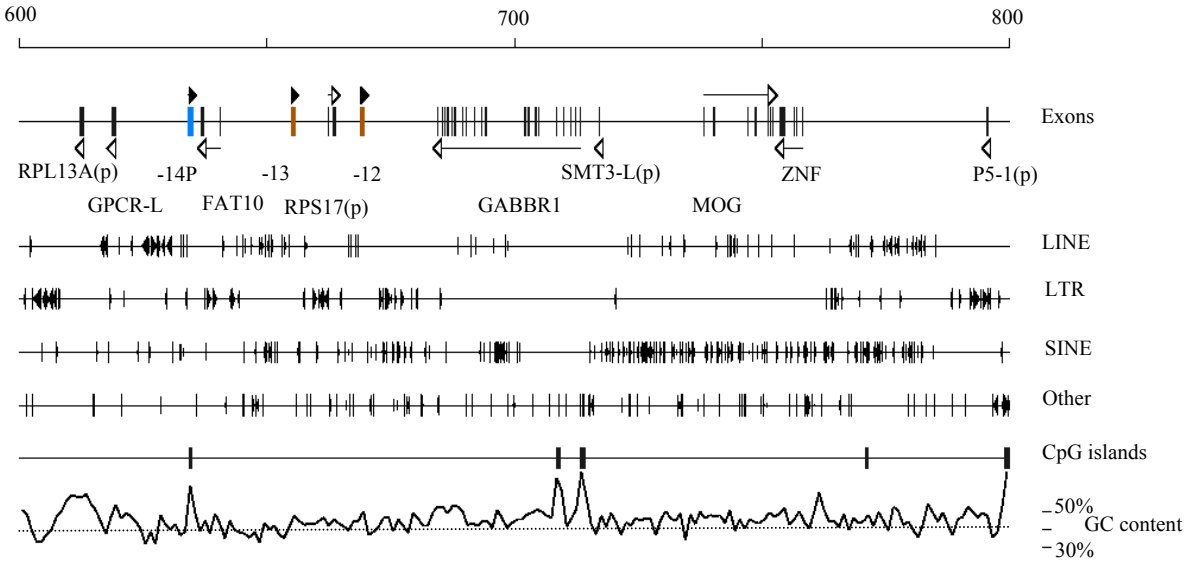
As with the major cluster, the genomic environment of the minor cluster ORs is characterised by a lower than average GC content (typically it is less than 40%). A high proportion of LINE repeats and a small proportion of SINE elements is also associated with the minor cluster. Another shared feature is the lack of CpG islands: the only CpG island within this 200 Kb region is associated with the zinc finger telomeric of the OR genes.

This association of OR genes with a low GC environment is consistent with observations made on the chromosome 17 cluster. The chromosome 17 cluster, which contains 17 OR coding regions was also reported to be located in a low GC region. This region also contained CpG islands, but as with the MHC-linked cluster, none of these islands appeared to be coupled to OR genes (Glusman *et al.*, 2000).


0
100
200
Exons
RFP
Cy18(p)
ZNF
-8P
-15
-26P
-1
-4
-3
SARA-L(p)
LINE
LTR
SINE
Other
CpG islands
50%
GC content
30%


200
300
400
Exons
-2P
-6
-5P
-7P
-25P
-24P
-23P
-22P
-28
LINE
LTR
SINE
Other
CpG islands
50%
GC content
30%


400
500
600
Exons
DDX6(p)
-21
-27
-20
-19P
-18
-17
-16
FAT10(p)
TRE-L(p)
MRG
LINE
LTR
SINE
Other
CpG islands
50%
GC content
30%


600
700
800
Exons
RPL13A(p)
-14P
-13
-12
SMT3-L(p)
ZNF
P5-1(p)
GPCR-L
FAT10
RPS17(p)
GABBR1
MOG
LINE
LTR
SINE
Other
CpG islands
50%
GC content
30%

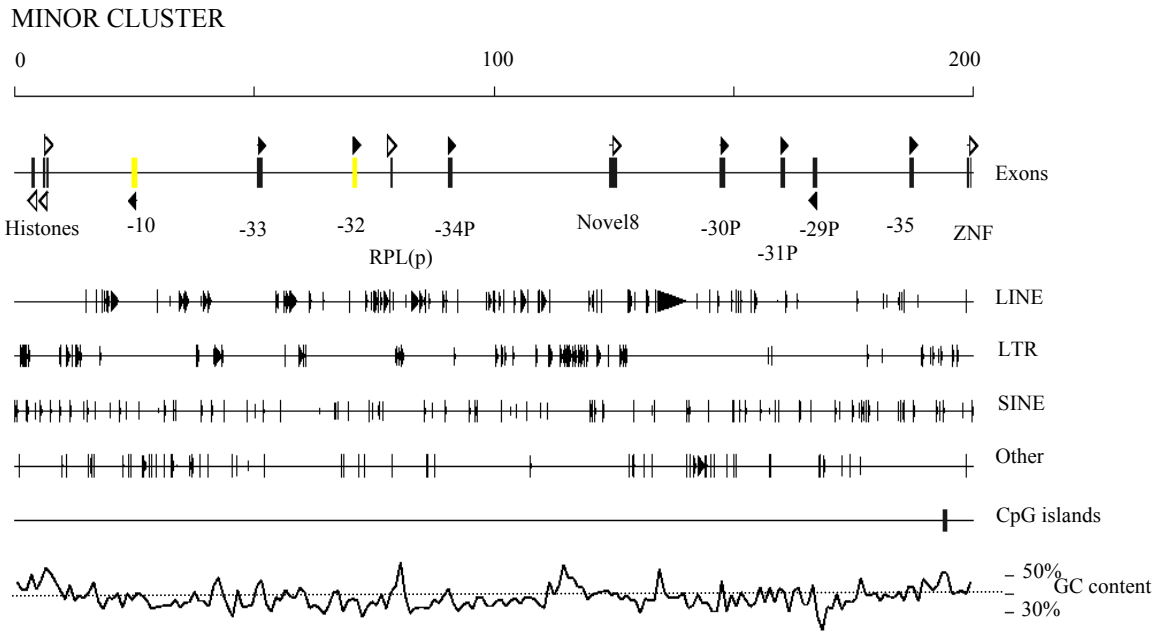


Figure 4.6: The genomic organisation of the human MHC-linked major and minor OR clusters. Predicted exons are shown as boxes on the first track. The orientation of genes is indicated by an arrow either above or below the gene, with a line indicating exons that belong to the same gene. OR genes are indicated by filled arrows and non OR genes are indicated by unfilled arrows.Where ORs belong to a subfamily, the subfamily designation is indicated by the colour of the OR exon. Below the gene track, arrows indicate where repeats are found. The second track shows LINE repeats, the third track shows LTR and retroviral elements, and the fourth track shows SINE repeats. Repeats that could not be classified according to these criteria (for example, low complexity repeats) are shown on the fifth track ('Other'). The sixth track shows boxes where the CpG islands within the sequence are found. Beneath this track, the GC content of the sequence is plotted per 1 Kb: the dotted line indicates the genome average figure of 40%

4.6. Local duplications within the human MHC-linked OR cluster

Figure 4.7 is a dot-matrix plot of the 718800 bp sequence of the major OR cluster (from the RFP locus to the SMT3H2 pseudogene locus).

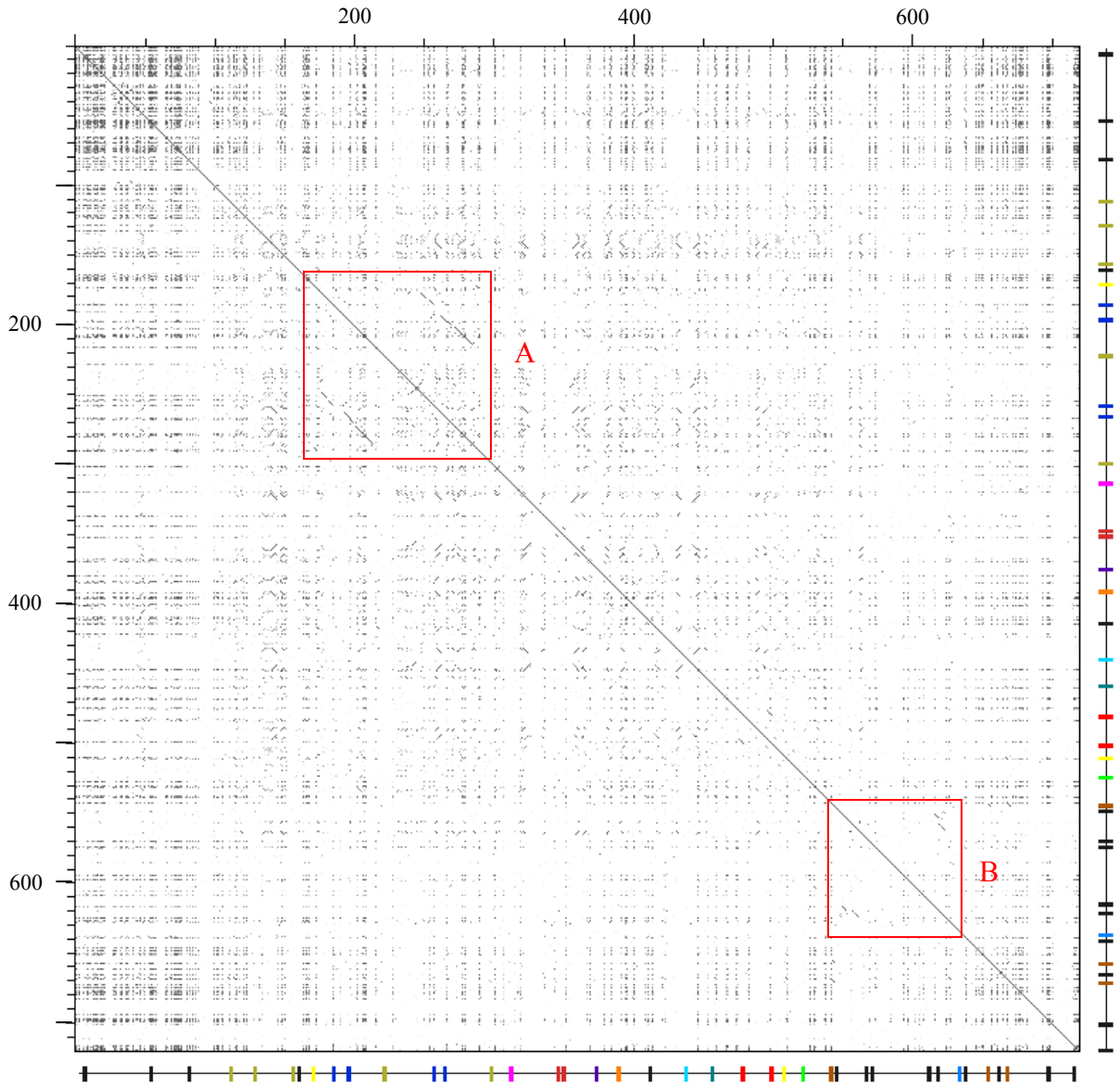


Figure 4.7: A dot-matrix plot showing the major OR cluster plotted against itself. The gene line shows the positions of exons within the sequence. OR subfamilies are shown in colours corresponding to those in figure 4.6. Other ORs are indicated by the light green colour, with non-ORs black.

There are 2 major duplication events that are highlighted by this plot (indicated by boxes A and B in figure 4.7). The first duplication event (box A) is associated with subfamily 2 (hs6M1-3, hs6M1-4P, hs6M1-5P and hs6M1-6). This duplicated block is a region of sequence approximately 35 Kb in length. At one end of this block, a LINE repeat, L1MA7 is shared, whilst at the other end the block has a MER52A LTR retroviral element. At one site the block contains hs6M1-4P and hs6M1-3, whilst at the other site the OR genes in the block are hs6M1-6 and hs6M1-5P. These genes are located in the same position relative to repeats in both sequences. Throughout the block, 23 repeat elements, accounting for 13 Kb of sequence, are conserved in both positions (figure 4.8). Conserved repeat elements include 2 AluSx repeats and 1 AluSq repeat. The inclusion of Alu repeats within the duplicated block suggests this duplication event is a relatively recent event since Alus are primate-specific repeats descended from a processed 7SL RNA gene. AluSx elements are estimated to be approximately 37 million years (Myr) old (+ or – 19 Myr) whilst AluSq are estimated as 44 Myr (+ or – 19 Myr) (Kapitonov and Jurka, 1996). The duplication involving these 4 olfactory receptor genes can, therefore, be dated within the last 63 Myr. The boundary sequences of this duplication event are both LINE repeats, suggesting the duplication was mediated by LINEs. A similar LINE-mediated mechanism has been shown to be responsible for the duplication of the γ-globin locus (Fitch *et al.*, 1991). SINE-mediated duplications have been implicated in the duplication of an olfactory receptor gene located in the chromosome 17 cluster, OR17-24 (hs17M1-1P) and OR17-25 (hs17M1-2) (Glusman *et al.*, 1996).

The second duplication block (box B, figure 4.7) is a sequence of about 8 Kb. Unlike the other duplication observed from the dot-matrix plot, this event only appears to have involved repeats: namely, a large LIM4 repeat, a L2 repeat, a MER81 retroviral element and a LIMB5 repeat. The absence of Alu repeats suggests this duplication event occurred much earlier in evolutionary time than the duplication involving OR genes.

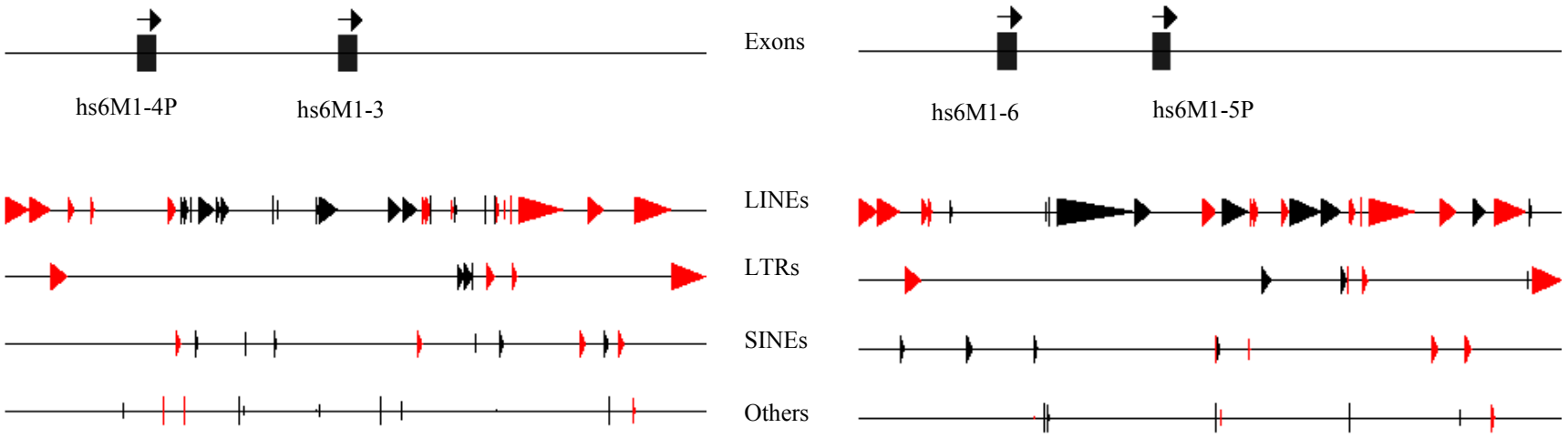


Figure 4.8: Local duplication within the major MHC-linked OR cluster creating 2 new OR genes. The block that has duplicated to create 4 OR genes from 2 original OR loci is shown above. Pairs of OR genes are 96% identical on the DNA level, suggesting this was a recent evolutionary event. Repeats conserved in both blocks of sequence are shown in red. The 2 blocks are separated by approximately 50 Kb of sequence containing another OR gene, hs6M1-2P (figure 4.6).

A dot-matrix analysis of the minor cluster against itself and against the major cluster was also performed in order to look for olfactory receptor gene duplications. In both cases, however, no large local duplications associated with OR genes were observed, although a broken 10 Kb block of sequence was shared around the region of hs6M1-10 and hs6M1-32 (both OR genes belong to subfamily 1). In contrast to the duplication associated with subfamily 2, however, few repeats are conserved across the sequence in both positions, and no Alu repeats are shared. This suggests the duplication creating the 2 genes, hs6M1-10 and hs6M1-32 occurred much earlier than the duplication creating 2 extra OR genes within subfamily 2.

4.7. Conclusions

In conclusion, 34 OR genes were identified in the 2 clusters of OR genes located within the human extended class I region. There is no evidence that the majority of these genes were created through recent duplication events, although five subfamilies can be defined, and an ancient duplication creating one of these subfamilies has been identified. Origins of the human MHC-linked OR cluster are discussed further in Chapter 8.

Across the human extended MHC olfactory receptor cluster, amino acids have been conserved within key regions of the protein structure, with hypervariability in transmembrane domains 3, 4 and parts of transmembrane domain 5. This hypervariability is associated with the idea that ligand binding takes place within a pocket formed by these 3 domains. In addition to this, comparison with the rhodopsin structure, and hypervariability within the 2nd extracellular region of the protein suggests this extracellular region may be involved in odorant ligand binding, possibly providing the ‘cap’ to the transmembrane pocket.

The 19 pseudogenes were analysed to examine whether there was a 'mutational bias' creating point mutation, insertion or deletions hotspots within certain regions of OR genes. This was suggested to account for the high number of OR pseudogenes with only 1 mutation disrupting the coding region. A shared frameshift between 2 distantly related genes suggested that this hypothesis might be valid: additional support was provided by the finding that 47% of pseudogene mutations occur within 11% of the gene. The low number of pseudogenes in this analysis, however, mean it is not possible to confirm or refute this idea conclusively.

The genomic environment of OR genes is distinct from other regions of the extended MHC. Trends associated with this region (low GC, high LINE content) are specifically related to the presence of olfactory receptor genes. OR genes are also associated with a lack of CpG islands, suggesting an as-yet-unknown mechanism is responsible for promoting the transcription of these genes. Finally, a local duplication within the major olfactory cluster is associated with the creation of 2 new OR loci. This local duplication, which also included Alu repeats has occurred within the last 63 Myr. A local duplication is also likely to have created 1 new OR loci within the minor cluster, however, the lack of shared repeats between the 2 blocks of sequence suggest this duplication occurred at a much later date, although the deletion of repeats from these duplicons cannot be excluded.