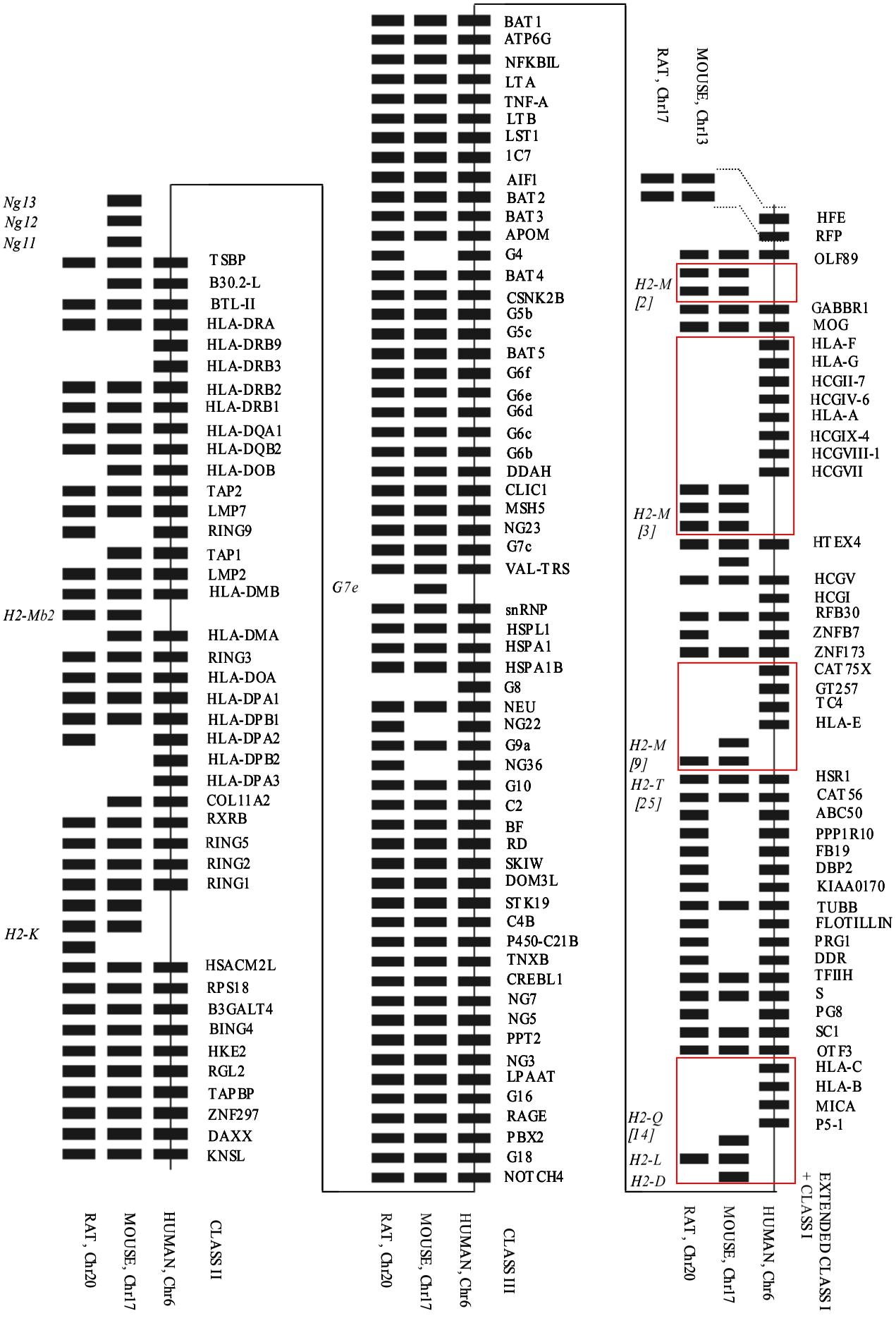
Chapter 5

The mouse MHC-linked contig and comparative analysis

5.1. Introduction.

The conservation of a cluster of olfactory receptor genes next to the MHC classical class I region in mouse, human and rat was discovered as part of an investigation into the synteny breakpoints between these various species (Szpirer *et al.*, 1997, Yoshino *et al.*, 1997). The three species all show a strong conservation of gene order in class II and class III of the MHC, and although conservation is much less marked in class I (see figure 5.1), the conservation of some genes, led to the proposal of the 'framework hypothesis', which suggests that some genes within the class I region are highly conserved between species, whilst there are permissive regions, where duplications and deletions resulted in areas of the MHC class I region sharing different evolutionary histories in different species (Amadou, 1999). In the extended MHC class I region, therefore, interest was focussed on whether the conservation of olfactory receptor genes followed the pattern observed in the class II and class III regions (strong conservation) or whether the pattern of conservation would follow the 'framework hypothesis' of the class I region.

Figure 5.1 (next page): Comparative gene map of the human, mouse and rat MHC. This is a schematic diagram (not to scale) showing human MHC genes known to be conserved in the rat and mouse species. All MHC genes are located on chromosome 17 in mouse and chromosome 20 in rat, with the exception of Hfe and Rfp which are located on chromosome 13 in mouse and chromosome 17 in rat. All human genes are labelled above the human track; additional rat/mouse loci are labelled in italics below their position. 'Permissive' areas of the class I region are indicated by the red boxes. This diagram was created using data from The MHC Sequencing Consortium (1999), Gunther and Walter, (2001), Amadou (1999) and Allcock *et al.*, (2000).


Ng13
Ng12
Ng11
TSBP
B30.2-L
BTL-II
HLA-DRA
HLA-DRB9
HLA-DRB3
HLA-DRB2
HLA-DRB1
HLA-DQA1
HLA-DQB2
HLA-DOB
TAP2
LMP7
RING9
TAP1
LMP2
HLA-DMB
H2-Mb2
HLA-DMA
RING3
HLA-DOA
HLA-DPA1
HLA-DPB1
HLA-DPA2
HLA-DPB2
HLA-DPA3
COL11A2
RXRB
RING5
RING2
RING1
H2-K
HSACM2L
RPS18
B3GALT4
BING4
HKE2
RGL2
TAPBP
ZNF297
DAXX
KNSL
RAT, Chr20
MOUSE, Chr17
HUMAN, Chr6
CLASS II
BAT1
ATP6G
NFKBIL
LTA
TNF-A
LTB
LST1
1C7
AIF1
BAT2
BAT3
APOM
G4
BAT4
CSNK2B
G5b
G5c
BAT5
G6f
G6e
G6d
G6c
G6b
DDAH
CLIC1
MSH5
NG23
G7c
VAL-TRS
G7e
snRNP
HSPL1
HSPA1
HSPA1B
G8
NEU
NG22
G9a
NG36
G10
C2
BF
RD
SKIW
DOM3L
STK19
C4B
P450-C21B
TNXB
CREBL1
NG7
NG5
PPT2
NG3
LPAAT
G16
RAGE
PBX2
G18
NOTCH4
RAT, Chr20
MOUSE, Chr17
HUMAN, Chr6
CLASS III
RAT, Chr17
MOUSE, Chr13
HFE
RFP
OLF89
H2-M [2]
GABBR1
MOG
HLA-F
HLA-G
HCGII-7
HCGIV-6
HLA-A
HCGIX-4
HCGVIII-1
HCGVII
H2-M [3]
HTEX4
HCGV
HCGI
RFB30
ZNFB7
ZNF173
CAT75X
GT257
TC4
HLA-E
H2-M [9]
H2-T [25]
HSR1
CAT56
ABC50
PPP1R10
FB19
DBP2
KIAA0170
TUBB
FLOTILLIN
PRG1
DDR
TFIIH
S
PG8
SC1
OTF3
HLA-C
HLA-B
MICA
P5-1
H2-Q [14]
H2-L
H2-D
RAT, Chr20
MOUSE, Chr17
HUMAN, Chr6
EXTENDED CLASS I + CLASS I

Previous work on the mouse MHC class I had been done by a number of groups, including the group of Kirsten Fischer Lindahl (HHMI, Dallas, TX) who had been involved in the mapping and sequencing of the H2-M region of the mouse MHC for a number of years (Jones *et al.*, 1995, Yoshino *et al.*, 1997, Yoshino *et al.*, 1998a, Yoshino *et al.*, 1998b, Amadou *et al.*, 1999, Jones *et al.*, 1999). In order to analyse the MHC-linked OR genes in mouse and to compare them to the human OR genes, a collaboration was agreed with Claire Amadou and Kirsten Fischer-Lindahl of the Dallas group to map and sequence the mouse region from the Gabbr1 receptor to the synteny breakpoint on mouse chromosome 17. This region was expected to contain a number of olfactory receptor genes (Amadou 1996) and at least 2 MHC class I like genes (Wang *et al.*, 1991).

This chapter describes my part of the collaboration to map, sequence and assemble a mouse contig that represents the mouse extended class I region. It also describes the gene content of the region, and how this gene content relates to that found in the syntenic human region.

5.2. Constructing the mouse MHC-linked OR contig

Mapped and unmapped BAC clones (from the Research Genetics 129 mouse BAC library (CITB-CJ7-B)) and PAC clones (from the Children's Hospital Oakland BACPAC resources RPCI-21 library (129S6/SvEvTac)) together with marker data were provided by Claire Amadou and Kirsten Fischer Lindahl (Amadou *et al.*, 1999). Using these resources, a clone contig was constructed using restriction digest fluorescent fingerprinting (see Chapter 2). This method used fluorescently tagged dideoxy ATPs to label the *Hind*III sites created in a double digest of the clone, allowing the restriction patterns from the various clones to be compared. From these restriction patterns, the degree of overlap between clones (calculated according to whether clones share the same restriction sites) was used to assemble the mouse

contig (Gregory *et al.*, 1997). In total, 387 mouse clones were fingerprinted and these were assembled into 12 contigs, using the program, 'FPC' (Soderlund *et al.*, 1997). These contigs varied greatly in size, ranging from 2 or 3 clones up to as many as 73 clones. The contig selected for sequencing was the second largest contig that contained 50 clones which were predicted to produce 1 Mb of sequence (Figure 5.3). From the FPC database, the following clones were selected for sequencing: bM573K1, bM87K14, bM332P19, bM104O10, dM538M10, dM639N14, and bM350K7. These clones were considered to represent the minimal tiling path across the contig, ensuring the smallest amount of sequencing possible was done. Prior to sequencing, it was also considered important to confirm the contig was located on mouse chromosome 17, in the region syntenic to the human extended MHC. In order to do this, a clone in the middle of the contig, bM332P19 was mapped using fluorescent *in situ* hybridisation (FISH). The result of this analysis confirmed that bM332P19 was located on chromosome 17 in the region C-D. (Figure 5.2).

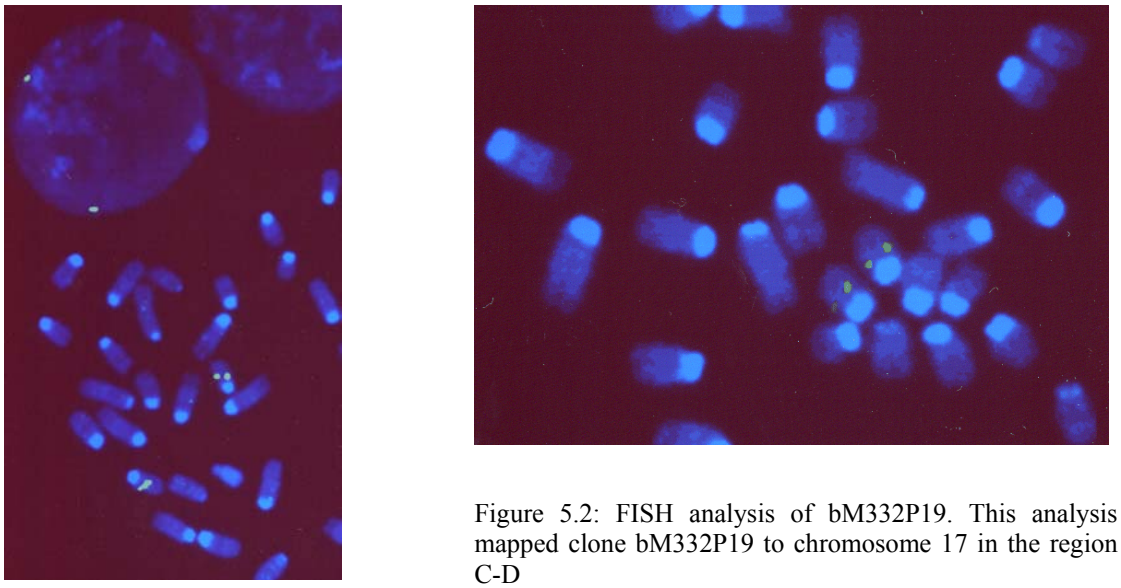
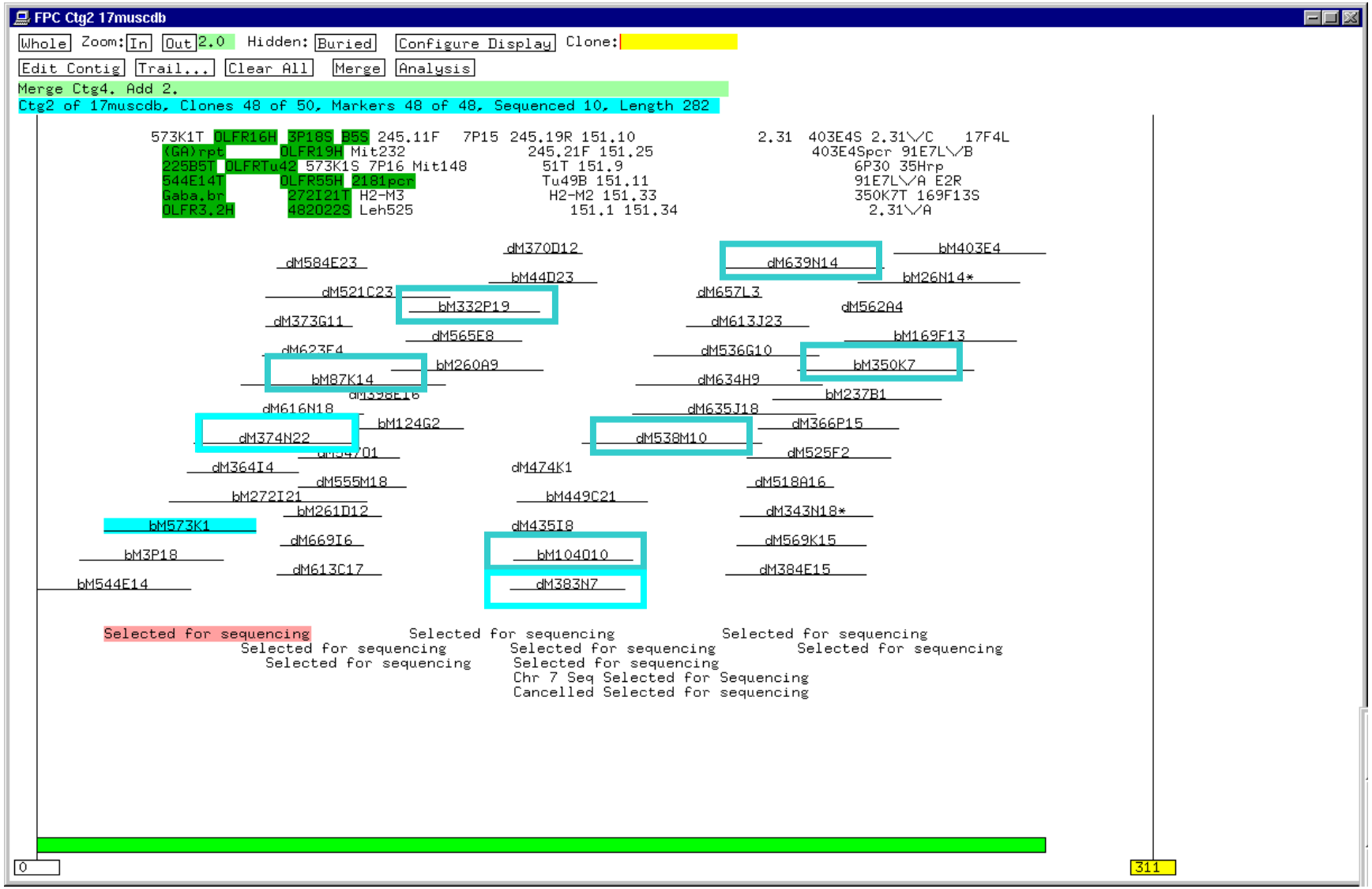


Figure 5.2: FISH analysis of bM332P19. This analysis mapped clone bM332P19 to chromosome 17 in the region C-D

Figure 5.3 (next page): Mouse MHC-linked OR contig. A screen dump from the FPC database created in order to sequence the mouse contig is displayed. Marker data is shown at the top of the screen, and markers present in the highlighted blue clone are highlighted green. bM573K1 (highlighted in blue) was one of the clones sequenced; other clones involved in the tiling path across the contig are indicated by the unfilled light blue boxes. 48 clones are displayed on screen; 2 further clones are hidden behind clones indicated by the asterix.


FPC Ctg2 17muscdb
Whole Zoom: In Out 2.0 Hidden: Buried Configure Display Clone:
Edit Contig Trail... Clear All Merge Analysis
Merge Ctg4. Add 2.
Ctg2 of 17muscdb, Clones 48 of 50, Markers 48 of 48, Sequenced 10, Length 282
573K1T OLFR16H 3P18S B5S 245.11F 7P15 245.19R 151.10
(GA)rpt OLFR19H Mit232 245.21F 151.25
225B5T OLFRTu42 573K1S 7P16 Mit148 51T 151.9
544E14T OLFR55H 2181pcr Tu49B 151.11
Gaba.br 272I21T H2-M3 H2-M2 151.33
OLFR3.2H 482O22S Leh525 151.1 151.34
2.31 403E4S 2.31\/C 17F4L
403E4Spcr 91E7L\/B
6P30 35Hrp
91E7L\/A E2R
350K7T 169F13S
2.31\/A
dM370D12
dM584E23
bM44D23
dM521C23
bM332P19
dM373G11
dM565E8
dM623F4
bM260A9
bM87K14
dM616N18
bM124G2
dM374N22
dM364I4
dM555M18
bM272I21
bM261D12
bM573K1
dM669I6
bM3P18
dM613C17
bM544E14
dM639N14
bM403E4
bM26N14*
dM657L3
dM562A4
dM613J23
bM169F13
dM536G10
bM350K7
dM634H9
bM237B1
dM635J18
dM366P15
dM538M10
dM525F2
dM474K1
dM518A16
bM449C21
dM343N18*
dM435I8
dM569K15
bM104O10
dM384E15
dM383N7
Selected for sequencing
Selected for sequencing
Selected for sequencing
Selected for sequencing
Selected for sequencing
Selected for sequencing
Selected for sequencing
Selected for sequencing
Chr 7 Seq Selected for Sequencing
Cancelled Selected for sequencing
0
311

With confirmation that the clone was located in the correct area of mouse chromosome 17, the 7 mouse clones were sequenced. During this process, it became obvious that in spite of the fingerprint analysis suggesting an overlap between bM573K14 and bM87K14, this could not be confirmed at the sequence level. Another clone, dM374N22, was selected for sequencing to fill the gap in the contig. An additional clone, dM383N7, was also selected for sequencing at a later stage as, again contrary to the fingerprint analysis, bM104O10 failed to bridge the gap between the 2 clones, bM332P19 and dM538M10.

The final tile path, then, consisted of 8 clones that represented a region of the mouse 'extended MHC class I' from Gabbr1 to a number of MHC olfactory receptor genes. However, this contig did not extend to the synteny breakpoint, as a marker that was used to help define the breakpoint (Olf89) was located in a smaller contig that was not sequenced. Attempts were made to anchor this small contig using the marker and fingerprint data but it became obvious that these approaches were not going to yield quick results. With plans for a Mouse Genome Project on the horizon (Smaglik and Abbott, 2000, Rogers and Bradley, 2001), therefore, it was decided to stop trying to map the region and wait for draft sequence that could be used to help assemble the region.

5.3. Sequence assembly of the mouse OR contig

The 8 tile path clones were assembled at the sequence level using 'Gap4' software by the Core Sequencing Department (Team 44) at the Sanger Centre and myself (Bonfield *et al.*, 1995, Staden *et al.*, 2000). The problems experienced in mapping, owing to the high number of repeat units within the region, were also reflected in the difficulties that were encountered in assembling the sequence. As with the fingerprinting software ('FPC') which 'stacks-up' contigs if they have very similar restriction site positions, the software for assembling sequences also has a tendency to

align sequences with a very similar nucleotide content. For example, in the extreme case of bM332P19, a segment of sequence about 6 Kb in size had duplicated and translocated right next to the original segment of sequence. The only difference between the two duplicated areas was 1 base pair, which meant the sequence required very careful assembly, using additional parameters to basic similarity, such as read pair information. Other clones within the region showed less extreme duplications, but the repeat units meant assembly was still difficult.

After individual assembly of each clone, the consensus sequence of all tile path clones was assembled using the programs, 'dotter' (Sonnhammer and Durbin, 1995) and 'cross_match'. As can be seen from table 5.1, in spite of the selection of a minimal tiling path from the FPC database, there was some redundancy (33.2%) across the region with only 897213 bases of the 1343061 bases sequenced being used. This amount of redundancy in the sequencing is expected using this fingerprinting approach.

Clone name	Accession number	Size, bp	Sequence used, bp
bM350K7	AL359352	162935	162935
dM639N14	AL365336	208443	54434
dM538M10	AL136158	190737	190737
dM383N7	AL450393	119416	36803
bM332P19	AL133159	170749	170745
bM87K14	AL359381	232383	102817
dM374N22	AL590433	103784	24128
bM573K1	AL078360	154614	154614
	Total:	1343061	897213

Table 5.1: Clones contributing to mouse contig. The size of clones is compared to how much of the sequence of the clone contributed unique sequence in the final consensus sequence.

5.4. Identification of mouse MHC-linked olfactory receptor genes

The 897213 bp contig was analysed using a variety of programs to search for genes, with specific emphasis being placed on the identification of olfactory receptor genes (Chapter 2). Within the

sequence, 46 olfactory receptor genes were identified, suggesting an average of one OR gene locus per 19.5 Kb. This density of OR genes is higher than the equivalent density of the human (major) MHC-linked cluster. (580000 / 25 = 23.2 Kb/OR gene), but the real difference lies in the ratio of pseudogenes. Of the 46 mouse MHC-linked OR genes, 36 have complete open reading frames, whilst the other 10 appear to be pseudogenes (Appendix 8). The ratio of genes to pseudogenes is therefore 3.6 which is significantly higher than the 0.8 ratio within the human MHC-linked OR cluster.

These results are consistent with observations that have been made about the mouse olfactory receptor gene repertoire, namely, that the number of olfactory receptor genes is higher in mouse. 1300-1500 mouse OR genes have been found in the mouse genome (Young *et al.*, 2002, Zhang and Firestein, 2002) compared to about 900 human OR genes (Glusman *et al.*, 2001). (These figures replaced older figures suggesting there were 1000 ORs in the mouse compared to 500-750 ORs in humans (Buck, 1996, Mombaerts, 1999)). Results from the analysis of the mouse MHC-linked ORs are also consistent with the gene to pseudogene ratio of the 2 studies of the entire mouse OR genomic repertoire (Young *et al.*, 2002, Zhang and Firestein, 2002) (In both papers, the pseudogene total is approximately 20%, producing a gene to pseudogene ratio of 4.0, similar to the 3.6 for the MHC-linked cluster.)

5.5. Identification of mouse MHC-linked OR subfamilies

The 46 olfactory receptor genes can be divided into several subfamilies according whether they share 70% or over protein similarity with other OR genes in the cluster (Table 5.2). In contrast to the human MHC-linked OR cluster, where few genes shared this degree of protein similarity with other MHC-linked OR genes, this subfamily designation can be applied to the majority of OR genes found within this region. This subfamily allocation leaves out 5 genes (mm17M1-29,

mm17M1-39, mm17M1-44P, mm17M1-6, mm17M1-40P) which appear to be only distantly related to other ORs within the cluster.

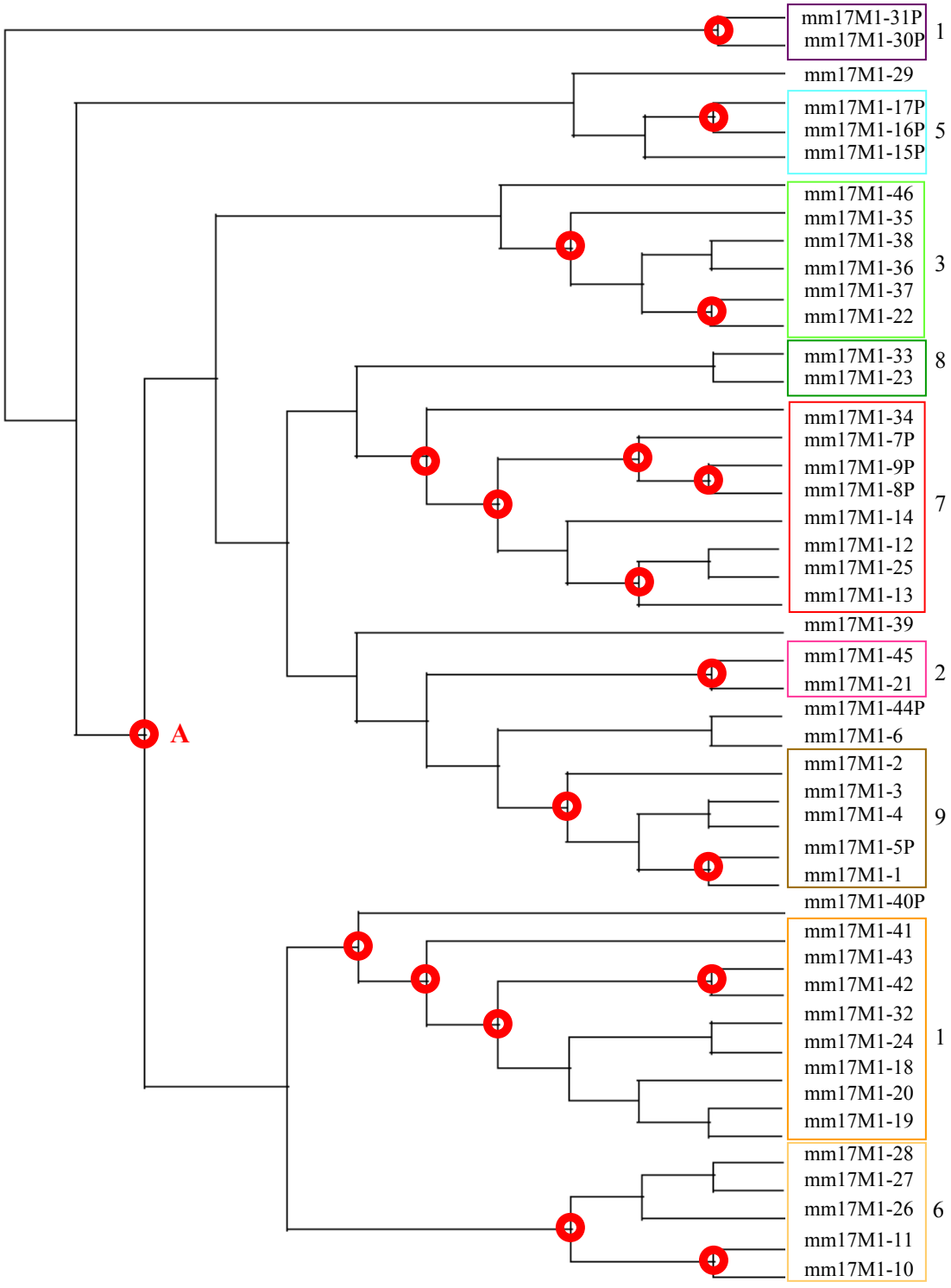
Subfamily	Genes in subfamily					
1	mm17M1-43 mm17M1-19	mm17M1-42 mm17M1-20	mm17M1 –41	mm17M1-32	mm17M1-24	mm17M1-18
2	mm17M1-45	mm17M1-21				
3	mm17M1-38	mm17M1-37	mm17M1-36	mm17M1-35	mm17M1-22	mm17M1-46
4	mm17M1-23	mm17M1-33				
5	mm17M1-17P	mm17M1-16P	mm17M1-15P			
6	mm17M1-10	mm17M1-11	mm17M1-28	mm17M1-27	mm17M1-26	
7	mm17M1-9P mm17M1-12,	mm17M1-7P mm17M1-34	mm17M1-8P mm17M1-25	mm17M1-14	mm17M1-13	
8	mm17M1-31P	mm17M1-30P				
9	mm17M1-5P	mm17M1-2	mm17M1-1	mm17M1-4	mm17M1-3	

Table 5.2: Subfamily designations of mouse MHC-linked OR genes. These were made based on a shared protein identity of 70% and over.

Relationships between the mouse MHC-linked OR genes were considered further by phylogenetic analysis of an alignment of these ORs (Appendix 9). As with the analysis of human MHC-linked OR genes, only branches supported by a bootstrap value of over 70% were considered to represent valid tree subdivisions. Considering valid branches, it is clear that ORs in the same subfamily cluster together with high bootstrap values reflecting a consistent relationship between the two branches. These relationships are also maintained when a larger number of sites are taken into account (251 as opposed to 118, results not shown).

In addition to confirming the assignment of mouse ORs to subfamilies, the phylogenetic tree also suggests an old relationship between mm17M1-40P and subfamily 1 (the association has a significant bootstrap value), implying that this pseudogene may have originally been a member of

Figure 5.4: Phylogenetic tree of mouse MHC-linked OR genes. This is a maximum parsimony tree showing the relationships between the OR genes in the mouse extended MHC. 118 sites were used and 250 bootstrap replicates were performed. Subfamilies are boxed in different colours and the number of the subfamily is indicated to the left of the box. Red rings at branch points indicate where bootstrap values are over 70%.


mm17M1-31P
mm17M1-30P
1
mm17M1-29
mm17M1-17P
mm17M1-16P
mm17M1-15P
5
mm17M1-46
mm17M1-35
mm17M1-38
mm17M1-36
mm17M1-37
mm17M1-22
3
mm17M1-33
mm17M1-23
8
mm17M1-34
mm17M1-7P
mm17M1-9P
mm17M1-8P
mm17M1-14
mm17M1-12
mm17M1-25
mm17M1-13
7
mm17M1-39
mm17M1-45
mm17M1-21
2
mm17M1-44P
mm17M1-6
A
mm17M1-2
mm17M1-3
mm17M1-4
mm17M1-5P
mm17M1-1
9
mm17M1-40P
mm17M1-41
mm17M1-43
mm17M1-42
mm17M1-32
mm17M1-24
mm17M1-18
mm17M1-20
mm17M1-19
1
mm17M1-28
mm17M1-27
mm17M1-26
mm17M1-11
mm17M1-10
6

this subfamily in spite of its low shared identity (30.4%). The phylogenetic tree also suggests an ancient association between subfamily 1 and subfamily 6. This is supported by the branch point A (Figure 5.4) which has a bootstrap value of 100%, suggesting that at some point in evolution a common ancestor duplicated to produce subfamilies 1 and 6.

5.5. Conservation of amino acids in mouse MHC-linked OR proteins

The protein alignment of all 46 mouse OR loci (Appendix 9) reveals positions where amino acids have been highly conserved, suggesting these amino acids may be functionally important sites across all the genes in this cluster. The consensus protein sequence produced from this alignment was also analysed to see if any of these hypothetically important sites could be linked to a putative function. The starting methionine was taken as the position where 14 of the 46 OR proteins (30.4%) share a methionine start codon. This is located 9 amino acids away from the first conserved motif (F I/L L L G F S). Within the cluster, however, there are some OR proteins that have a start codon that lies further away from the first conserved motif, for example, an extreme case is mm17M1-6 which has a start codon that is 84 amino acids upstream from the first conserved motif. This extended amino terminus clearly has structural implications for the protein: it may be, as with the V2R pheromone receptors (Matsunami and Buck, 1997) and metabotropic glutamate receptors (mGluRs) (O'Hara *et al.*, 1993, Takahashi *et al.*, 1993), that this long terminus is implicated in ligand binding.

The carboxy termini of the mouse MHC-linked OR genes are far less variable: the longest termini belong to mm17M1-41 and mm17M1-43 both of which only extend 14 amino acids beyond the last conserved residue. In contrast, the human MHC-linked ORs are much more variable; hs6M1-10 and hs6M1-35 both have much longer carboxy termini. These 2 OR genes, however, are both from the minor cluster OR in the human MHC extended class. If the orthologs of these 2 genes

were considered, it may be that this variability does exist within the syntenic mouse OR cluster (located on mouse chromosome 13 between the mouse loci, Rfp and Hfe, identified later in this chapter).

Figure 5.5 compares the conservation of amino acid residues across the hypothetical consensus protein. From this it is obvious that transmembrane regions 4 and 5 can be considered to be highly variable, whilst the other transmembrane regions appear to be more conserved. The third predicted hypervariable region, transmembrane domain 3 (Buck and Axel, 1991), shows a number of highly conserved residues. In both the human and mouse MHC-linked ORs, therefore, the 3 proposed hypervariable regions appear to be less hypervariable than might be expected: both species show a high number of conserved residues in transmembrane domain III and in the human MHC-linked ORs, amino acid residues in the second half of transmembrane domain V are highly conserved.

The comparison of the human MHC-linked ORs with the rhodopsin structure (Chapter 4) revealed that a pair of cysteines that stabilize the ligand binding pocket in the rhodopsin structure are conserved in a consensus sequence of human MHC-linked OR genes. An analysis of the cysteine content of the mouse MHC-linked OR genes (not shown) reveals the same pair of cysteines are conserved in the mouse MHC-linked ORs. By comparison with the rhodopsin structure it is likely that these 2 cysteine amino acids form a disulphide bridge involved in forming the ligand binding pocket. Other putative disulphide bridges are also predicted to exist in the same position as in the human MHC-linked ORs (Figure 4.4) because cysteines are conserved in the same position in both the mouse and the human MHC-linked OR consensus sequence.

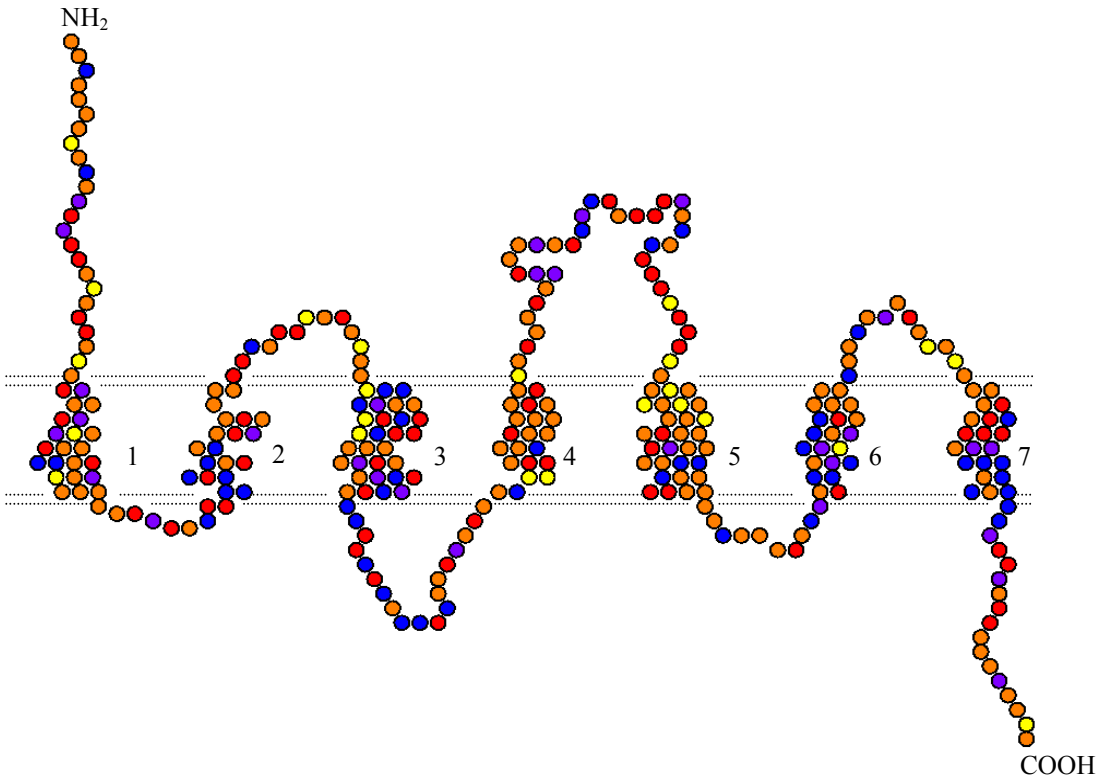


Figure 5.5: Conserved amino acids in mouse MHC-linked OR proteins. Schematic diagram showing the conservation of amino acids at predicted positions within a consensus mouse MHC-linked olfactory receptor protein. The degree of conservation ranges from 90%+ (blue), 75-90% (purple), 50-75% (red), 25-50% (orange) and less than 25% (yellow).

5.6. Mouse MHC-linked OR pseudogenes

The percentage of pseudogenes within the mouse MHC-linked OR cluster is significantly lower than that found within the human MHC-linked OR cluster. This percentage of pseudogenes within the mouse cluster, however, may be lower than the recorded 28% as the classification of 3 pseudogenes (mm17M1-7P, mm17M1-8P and mm17M1-9P) is based on the lack of a starting methionine at the predicted position within the open reading frame, and a splicing mechanism may add a methionine in front of the valine that replaces the methionine in what is considered to be the start position. There is evidence of the upstream splicing of ORs in mouse (Lane *et al.*, 2001), rat (Walensky *et al.*, 1998) and human (Linardopoulou *et al.*, 2001). These 5' exons have generally contained untranslated sequence, although Linardopoulou *et al*. (2001) do suggest that 5' exons are involved in producing coding sequence.

An alternative possibility for these genes is that they may utilise a different start codon downstream of the 'FILLG' (or equivalent) motif. In the case of the human OR gene, hs6M1-16, the methionine at amino acid position 79 is likely to be used as an alternative start codon in some transcripts (Chapter 6, Younger *et al.*, 2001) so mm17M1-7P, mm17M1-8P and mm17M1-9P may use a similar mechanism.

Mm17M1-5P is another pseudogene that may be functional. One substitution that creates a stop codon disrupting the open reading frame exists in the genomic sequence, but again this locus is well conserved, and it may be that the stop codon (TAG) exists as a glutamine (CAG) or some other functional codon in other haplotypes. This type of change was observed in different human haplotypes (Chapter 7, Ehlers *et al.*, 2000).

With the exception of these 4 loci (mm17M1-5P, mm17M1-7P, mm17M1-8P and mm17M1-9P), all the remaining pseudogenes within the mouse cluster have open reading frames that are significantly disrupted (containing several stops, several frameshifts or insertions) compared to other OR genes with open reading frames. The positions of these mutations are, in some cases, conserved between pseudogenes suggesting either that these pseudogenes were duplicated or that there are positions within the sequence that mutated more rapidly, or are more likely to have pieces of DNA inserted into them. The 2 shared mutations and the LTR insertion that mm17M1-30P and mm17M1-31P share suggests that these two pseudogenes were formed by a duplication event. A less obvious relationship is that between the frameshift at position 796 in mm17M1-30 and the frameshift at position 433 in mm17M1-16P (and mm17M1-17P). These frameshifts are located in a very similar position with regard to the protein sequence alignment, suggesting either an old duplication event or a gene conversion event from mm17M1-30P creating mm17M1-16/17P, or it may be that region within the sequence mutates at a faster rate, or is under less

selectional pressure than other sites (Figure 5.6). Analysis of the nucleotide sequence appears to suggest the latter as opposed to gene duplication or conversion events, as the sequence is significantly different in the two pseudogenes.

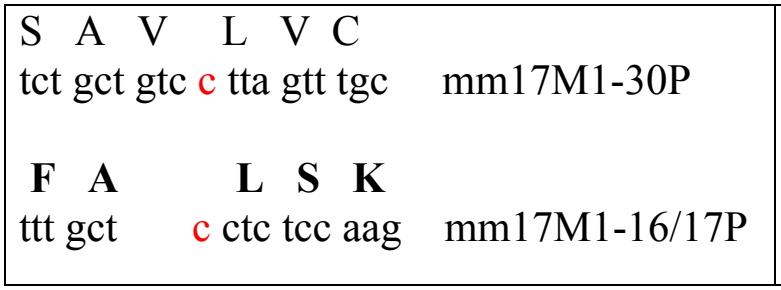
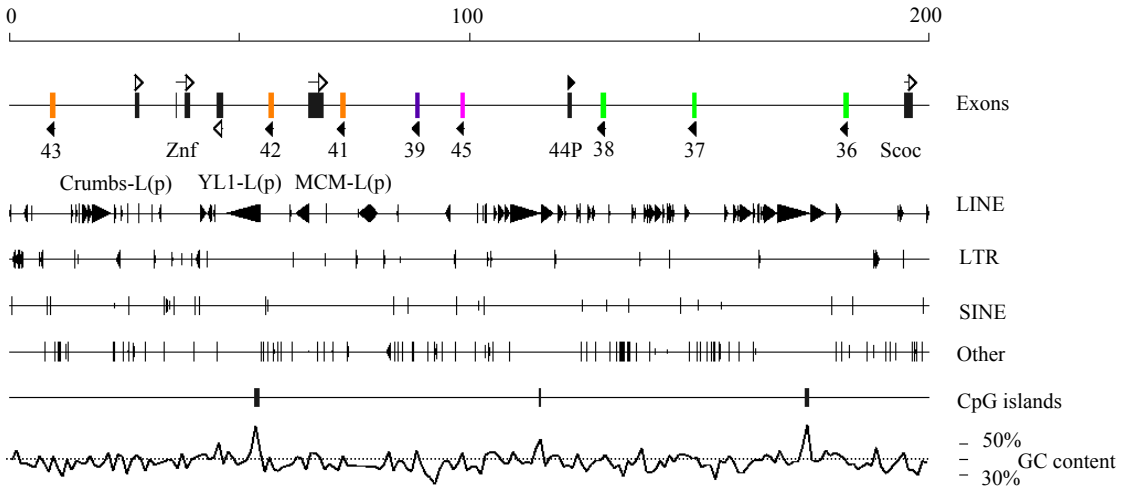


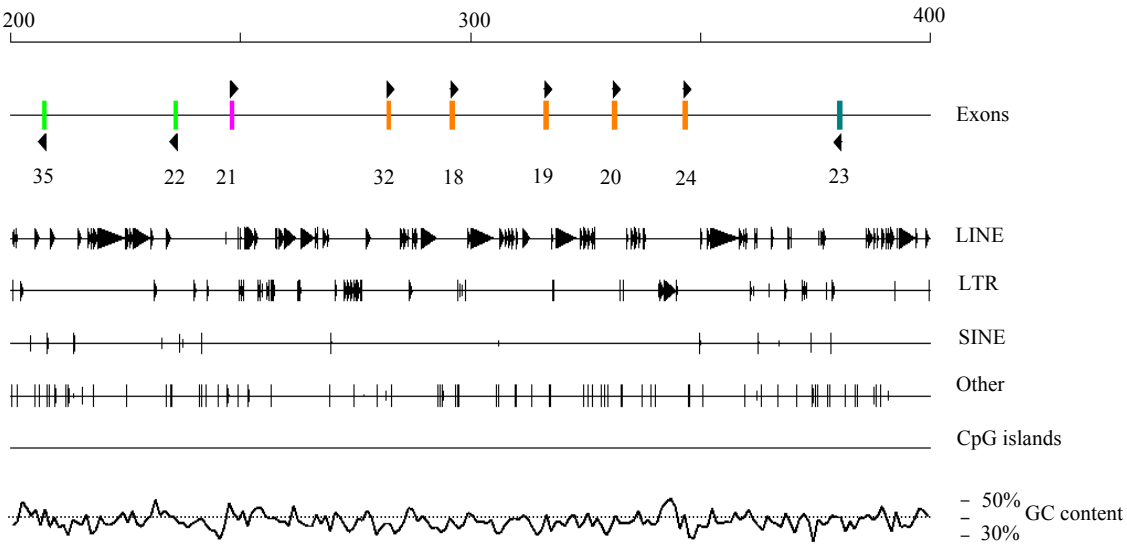
Figure 5.6: A base insertion in mouse pseudogenes mm17M1-30P and mm17M1-16/17P. An insertion of a cytosine has occurred in a similar position in both the mm17M1-30P and mm17M1-16/17P pseudogene, causing a frameshift.

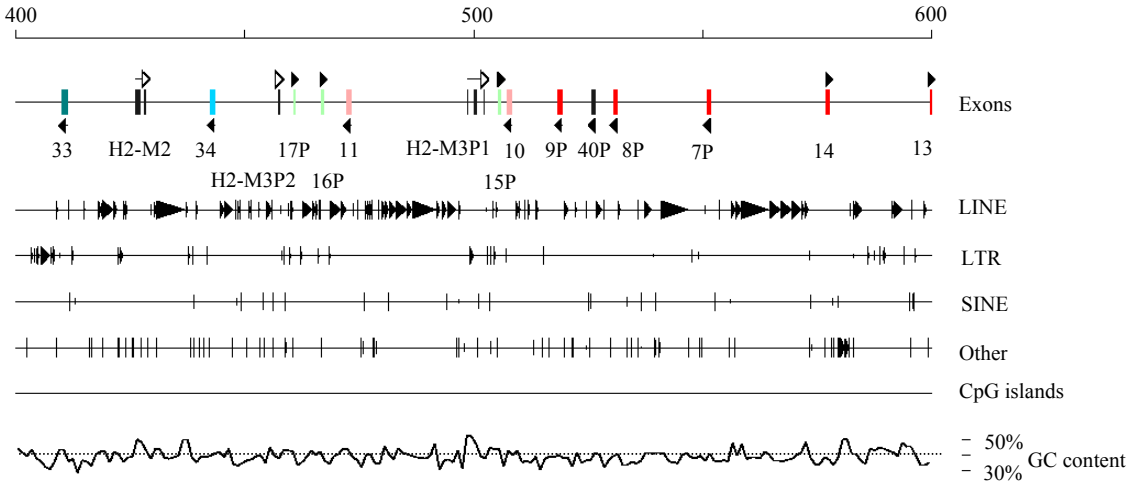
5.8. The genomic environment of the mouse MHC-linked olfactory receptor genes

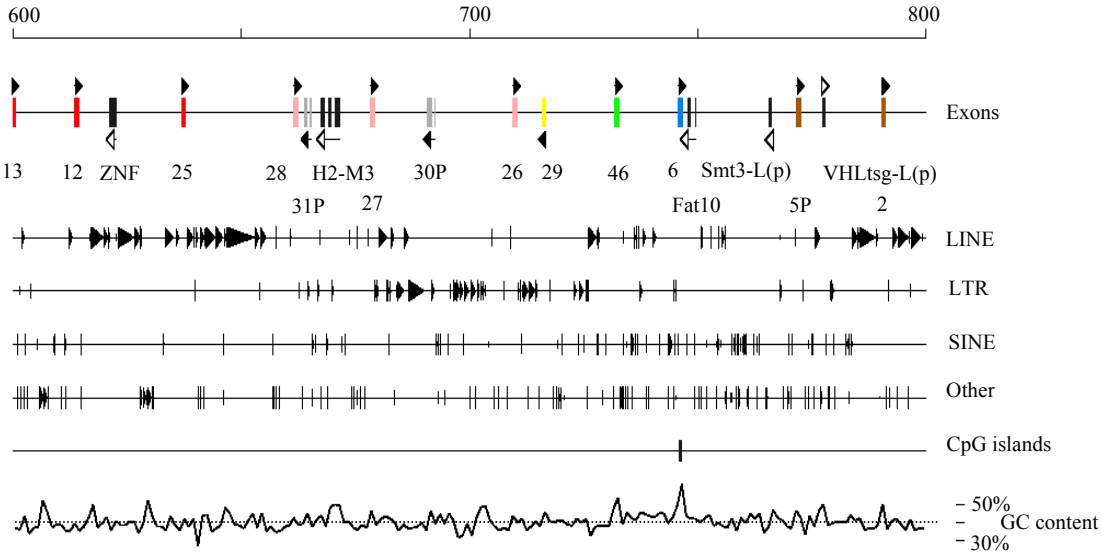
Figure 5.7 shows the genomic environment surrounding the MHC-linked olfactory receptor genes. From this diagram it is possible to see that the region is similar to that occupied by the human MHC-linked ORs. As in the human MHC-linked OR cluster, in terms of repeats the region is generally dominated by long interspersed nuclear elements (LINEs). This is particularly apparent in the first 700 Kb of the cluster, where LINEs comprise 74% of the total repeat content of the region. Long terminal repeat (LTR) elements, also known as retroviral-like elements, are also prevalent within this region: they contribute 15% of the repeat content within the first 700 Kb of the region. The SINE content is generally low but after this first 700 Kb, the genomic environment of the region changes, and SINEs become increasingly common. Within the last 197213 bases, for example, SINEs account for 23% of the repeat content, whilst LINEs and LTRs account for 38% and 25% respectively.

6 CpG islands are identified within the region. 2 of these are located within the olfactory receptor cluster, whilst 2 are associated with the Gabbr1 receptor, 1 appears to be associated with the Scoc gene and 1 appears to be associated with the YL1-like pseudogene. The Gabbr1 associated CpG


0
100
200
Exons
43
Znf
42
41
39
45
44P
38
37
36
Scoc
Crumbs-L(p)
YL1-L(p)
MCM-L(p)
LINE
LTR
SINE
Other
CpG islands
50%
30%
GC content


200
300
400
Exons
35
22
21
32
18
19
20
24
23
LINE
LTR
SINE
Other
CpG islands
50%
30%
GC content


400
500
600
Exons
33
H2-M2
34
17P
11
H2-M3P1
10
9P
40P
8P
7P
14
13
H2-M3P2
16P
15P
LINE
LTR
SINE
Other
CpG islands
50%
30%
GC content


600
700
800
Exons
13
12
ZNF
25
28
H2-M3
30P
26
29
46
6
Smt3-L(p)
VHLtsg-L(p)
31P
27
Fat10
5P
2
LINE
LTR
SINE
Other
CpG islands
50%
GC content
30%

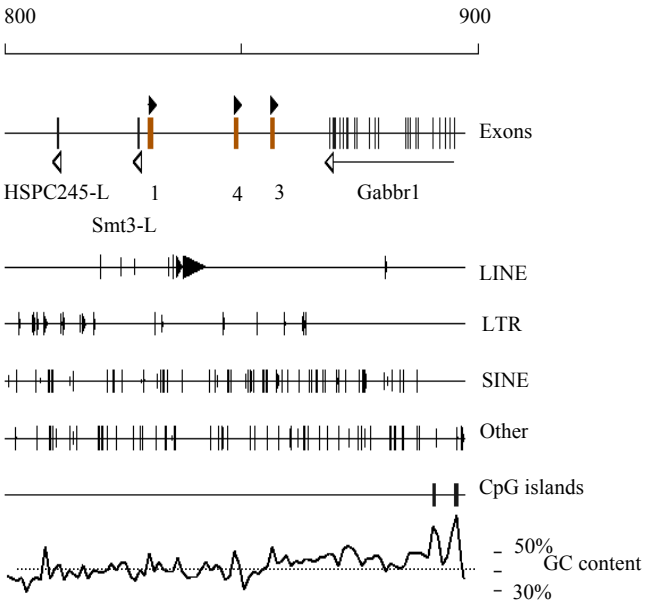


Figure 5.7: The genomic organisation of the mouse MHC-linked OR cluster. Predicted exons are shown as boxes on the first track. The orientation of the gene is indicated by an arrow either above or below the gene, with the line indicating exons that belong to the same gene. OR genes are indicated by filled arrows and non OR genes are indicated by unfilled arrows. The subfamily an OR gene belongs to is indicated by the colour of its exons. 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 human genome average figure of 40%

islands are likely to be important in the regulation of this gene, as they are located in positions upstream of the two alternative transcriptional start sites and appear to be conserved in both the human and mouse genomes. Similarly, the YL1-like associated CpG island may once have been important in the regulation of this gene before it lost its open reading frame. The 3 CpG islands associated with the OR cluster cannot be completely disregarded as having a regulatory role, however, the lack of CpG islands in the human OR cluster and the number of CpG islands (2) compared to the number of OR loci (46) does suggest these islands do not play a major role in the regulation of OR gene transcription.

5.9. MHC Class I and Class I-like genes within the MHC-linked OR contig

As in the human MHC-linked OR cluster, the mouse OR cluster is not an exclusive environment: a number of other genes are located between olfactory receptor genes. A major difference between the 2 clusters, however, is that the mouse MHC-linked OR cluster contains 4 loci that are related to MHC class I genes. MHC class I genes code for molecules that present antigens to CD8+ T cells in protective immunity against intracellular infection. In mouse, these genes are located in clusters, subdivided into 5 regions (H2-K, H2-D, H2-Q, H2-T and H2-M, Figure 5.1). The most telomeric subregion is the H2-M subregion, so-called because the presenter of the Mta (maternally transmitted antigen)(Loveland *et al.*, 1990) was mapped to that subregion (Richards *et al.*, 1989). 21 class I genes were identified within the H2-M region (Jones *et al.*, 1995), including H2-M3 and H2-M2 which represent the most telomeric genes in the H2-M region.

On mouse chromosome 17, therefore, the olfactory receptor gene cluster is clearly MHC-linked, since some of the olfactory receptor genes (mm17M1-34 to mm17M1-3) can be regarded as being located within the H2-M region of the mouse MHC. This suggests there could be an ancient association between the olfactory receptor genes and the class I genes on mouse chromosome 17.

Two class I genes, H2-M2 and H2-M3 had previously been characterised, and these were both identified within the mouse MHC-linked OR cluster. H2-M3 is a MHC class I molecule that presents the maternally transmitted antigen of mice (Mta) to cytotoxic T lymphocytes (Wang *et al.*, 1991). It also presents N-formylated peptides from the amino terminal of bacterial and mitochondrial proteins (Lindahl *et al.*, 1997). H2-M3 is closely related to the rat RT-M3 locus (83% shared protein identity), and the ability of the peptide to present the antigen varies according to allelic variations (Wang *et al.*, 1991). The allele present in this version of the

sequence has a shared protein identity of greater than 99% with the allele reported in the original paper (within the coding stretch of the protein 3 out of 336 amino acids differ).

The second previously-characterised gene is H2-M2 (initially known as 'Thy19.4'). This was identified as the most telomeric mouse class I gene on chromosome 17, but its function remains unknown (Yoshino *et al.*, 1998a, Yoshino *et al.*, 1998b). In the genomic sequence, it appears to be a pseudogene as the reading frame of the first exon is disrupted by the insertion of a additional guanine in a 'ggg' stretch of sequence. This suggests H2-M2, which has previously been observed as having an open reading could be pseudogenic in some haplotypes.

In addition to H2-M2 and H2-M3, 2 class I-like loci were found within the MHC-linked OR cluster. Both of these loci are pseudogenes: the more complete pseudogene (H2-M3P1) is missing a start codon, contains at least 4 stop codons and goes through 3 frameshifts. It appears to be related to H2-M3 and it also has some sequence identity to the rat MHC class I gene, RT1-M3 (Q62708). The second pseudogene (H2-M3P2) is a gene fragment that is similar to a number of class I genes. The fragment is too small to be able to deduce any significant homology or orthology relationships: duplication events suggest it is descended from H2-M3P1.

5.10. Other genes located within the MHC-linked OR cluster

Within the mouse OR cluster, several other genes and pseudogenes are also found. These include the Crumbs-like pseudogene, 2 zinc finger protein genes, a YL1-like pseudogene, a MCM-like pseudogene and the Scoc gene. In addition, towards the centromeric end of the cluster, the Gabbr1 gene, 2 Smt3-like loci, the HSPC245-like gene, the VHLtsg-like pseudogene and the Fat10 gene were all identified.

Of these identified genes, three genes found in the human MHC extended class I region (discussed in Chapter 3) are located in this region of mouse chromosome 17. The Gabbr1 gene and the Fat10 gene are conserved in the same position in both species, but the Smt3-like loci (of which one is a pseudogene and the other is predicted to be functional in mouse) are found telomeric of the Gabbr1 gene on mouse chromosome 17. In contrast to this, on human chromosome 6, the Smt3-like pseudogene is located centromeric of the Gabbr1 gene.

In mouse, the functional version of the Crumbs-like protein precursor appears to be involved in the production and migration of neurons in adulthood (den Hollander *et al.*, 2002). This includes the olfactory bulb where olfactory neurons are continually regenerated through an organism's lifetime, but the fact that this gene is pseudogenic, and is not found in the human region suggests there is no significant linkage between this gene and the cluster of OR genes. This Crumbs-like pseudogene is also distantly related to the Notch gene, a MHC class III gene that has three known paralogs within the human genome. These loci are all located within regions considered to be putative MHC paralogous regions and all these regions also have clusters of OR genes associated with them. The similarity to Notch could suggest some ancient relationship between the two loci, although the two proteins share EGF (epidermal growth factor)-like domains so similarity may be due to shared functional properties shaping evolution in a similar fashion.

The YL1-like pseudogene is located centromeric of the Crumbs-like pseudogene. In its functional form, this gene would be expected to code for a nuclear protein with DNA-binding ability, like its relation located on 1q21 in the human genome. The gene on 1q21 is predicted to be a transcriptional regulator, based on observations that various transformed phenotypes of Kirsten sarcoma virus-transformed NIH 3T3 cells were suppressed by introduction of a normal human chromosome 1. Cells that re-acquired the transformed phenotype were found to have lost the human 1q21 and 1q23-q24 regions, suggesting a transformation suppressor gene(s) was located

on the proximal portion of 1q. YL1 was considered to be one of these transformation suppressor genes. (Horikawa *et al.*, 1995)

The MCM4-like (mini chromosome maintenance deficient 4-like, also known as Cdc21) pseudogene belongs to a family of genes that encode proteins that appear to be 'replication licensing factors.' These factors are part of the cellular mechanism that ensures the replication of DNA occurs only once per cell cycle in eukaryotic cells (Blow and Laskey, 1988, Chong *et al.*, 1996). Cell fractionation studies indicate that differentially-phosphorylated forms of MCM4 are associated with the nucleus; the less phosphorylated form appeared to be more tightly bound to a nuclear structure. MCM4 also appears to forms a stable complex with 2 other MCM proteins and to be loosely associated with MCM2 (Musahl *et al.*, 1995). In the human genome the MCM4 gene has been mapped to 8q12-q13 by fluorescence *in situ* hybridisation (Ladenburger *et al.*, 1997, Satoh *et al.*, 1997).

The Scoc gene was identified through its similarity to the mRNA for the short coiled coil protein SCOCO (Scoc, AF115778). In a yeast two-hybrid assay, this protein was found to interact with metaxin 1, which is a component of the protein import apparatus of the mitochondrial outer membrane. However, this interaction could not be confirmed in mammalian cells or tissues so the exact function of Scoc remains unknown (Armstrong *et al.*, 1999).

The VHLtsg-like pseudogene, meanwhile, is similar to the tumour suppressor gene implicated in causing Von Hippel-Lindau syndrome (VHL), a dominantly inherited familial cancer which produces a number of benign and malignant neoplasms. In humans the functional version of the gene is located on chromosome 3p25 (Latif *et al.*, 1993). Finally, within the mouse OR cluster, the HSPC245-like gene was identified during an EST screen of a collection of CD34+

haemopoietic stem/progenitor cells (Zhang *et al.*, 2000). It appears to be a gene with low levels of expression in haemopoietic cells and in other tissues.

5.11. Local duplications of MHC-linked OR genes

A dot-matrix plot of the 897213 bp of the mouse MHC-linked OR contig against itself reveals that a large number of genomic duplications have occurred over evolutionary time (Figure 5.8). A detailed analysis of these large duplications reveals that in a number of cases, mouse olfactory receptor genes have duplicated through these events. The duplication in figure 5.8, box A, for example, accounts for the 2 OR genes, mm17M1-43 and mm17M1-42, which appear to have been generated by the duplication of a 5 Kb block. A mouse SINE, B1_MM, may have been duplicated as part of this block, suggesting a relatively recent time for the event. This is supported by the high nucleotide identity (94.7%) between the two ORs.

Figure 5.8, box B reveals the relationship between 5 ORs of subfamily 3, mm17M1-38, mm17M1-37, mm17M1-36, mm17M1-35, and mm17M1-22. Within this region, there appear to be 4 distinct duplication events. A 15 Kb block appears to have duplicated twice to produce mm17M1-38, mm17M1-37, and mm17M1-35. From mm17M1-37, meanwhile, a smaller 11-13 Kb block has duplicated twice to produce mm17M1-36 and mm17M1-22. The two duplication blocks are characterised by different repeat breakpoints. The larger block is delineated by a Lx repeat at both ends, whilst the smaller block is flanked by a B2_Mm2 SINE repeat and a tract of $(CT)_n$ repeats. As with box A, these blocks are associated with OR genes with a high degree of nucleotide similarity (92.4-95.1%).

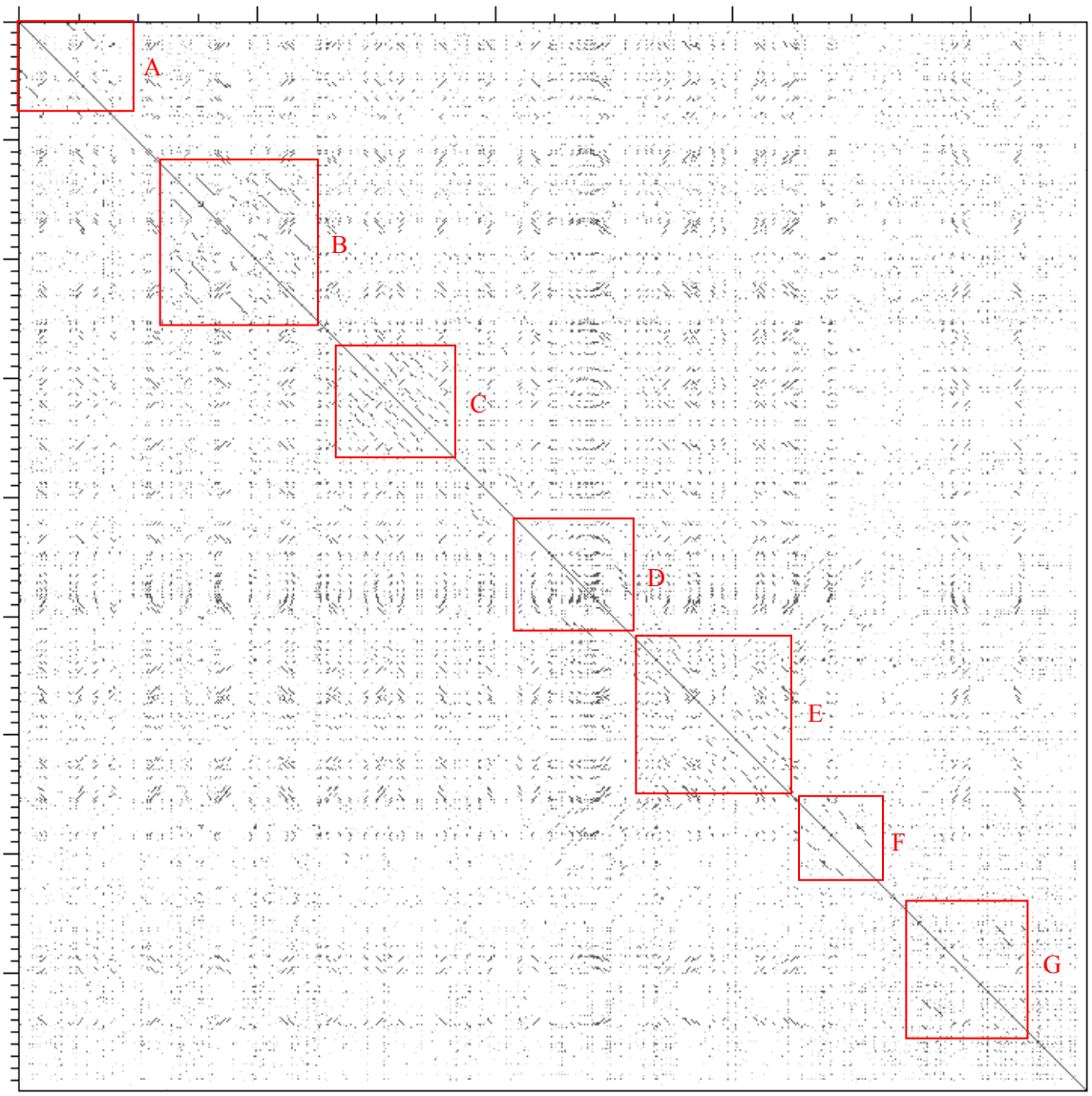


Figure 5.8: Dot-matrix plot of the mouse OR contig. The 897213 bp region was plotted against itself, revealing several duplications indicated by the red boxes which are indicated by a designated letter.

Further large scale duplications (Figure 5.8, box C) are responsible for producing 5 closely related (90.5%-90.6% nucleotide identity) ORs of subfamily 4. In this case, a 8-9 Kb block of sequence flanked by L1_MM and Lx5 repeats has duplicated 4 times to produce the 5 OR genes. The repeat content suggests the history of the duplication was as follows: mm17M1-32 -> mm17M1-20, mm17M1-20 ->mm17M1-24, mm17M1-20 -> mm17M1-18 and mm17M1-19. These duplications appear to predate the duplications of

box A and box B, since no SINE appears to have been carried within the block at any point.

Box D consists a number of small duplications. Within this sequence, however, the largest duplications are associated with a 5 Kb block delineated by LINE repeats, which has produced mm17M1-15P, mm17M1-16P and mm17M1-17P. The blocks associated with mm17M1-16P and mm17M1-17P are virtually identical in terms of base composition, suggesting this is a very recent event, whilst the mm17M1-15P/mm17M1-16P split obviously occurred earlier in time. The difference in timing of these events correlates with the nucleotide similarity of the 3 ORs: mm17M1-15P and mm17M1-16P are 77.8% identical, whilst mm17M1-16P and mm17M1-17P are 100% identical.

In figure 5.8, box E, there are a number of duplications associated with OR genes. These appear to have been fairly small local duplications consisting of around 6 Kb of sequence and generally flanked by SINEs. The mm17M1-12/mm17M1-13 duplication, for instance involves a B4A and PB1D10 repeat. The mm17M1-12/mm17M1-25 duplication involves a RMER1A repeat at one end, whilst the other end of the block is difficult to discern. Similarly, the mm17M1-7P/mm17M1-9P duplication is flanked by a RSINE1 at one end, whilst it is difficult to find a shared repeat that could resemble the end of the block. The mm17M1-7P/mm17M1-8P repeat unit duplication is flanked by the same SINE and a L1_MM repeat. As with the other OR genes in this region, the ability to detect local duplications is associated with a high nucleotide similarity between OR genes in these blocks (above 90% similarity in all these cases).

Duplications producing mm17M1-31P, mm17M1-30P and mm17M1-29 are shown in box F. A 14-18 Kb duplication with SINEs, B1_MM and B2_Mm2 appears to have produced mm17M1-30P and mm17M1-29. The mm17M1-30P/mm17M1-31P duplication event is also delineated by an SINE at one end, with a L1_MM repeat at the other. This event is interesting because a LTR element, RMER4, is present to disrupt both OR genes. The most parsimonious explanation for this is that a pseudogenic OR gene duplicated to produce two pseudogenes, although it may be that the two duplicated functional genes contained a favourable insertion site for this retroviral element and this retroviral insertion event occurred twice in two different regions. This region presents other difficulties in trying to predict an evolutionary history. Another retroviral element, MYSERV, is present in the sequence in both mm17M1-29 and mm17M1-30P blocks but it is absent in the mm17M1-31P block. Nucleotide sequence identities (87.2% mm17M1-30P and mm17M1-31P, compared to <80% mm17M1-29 against either of the other two genes) suggest mm17M1-31P is a copy of mm17M1-30P, rather than descending from mm17M1-29, but this means hypothesizing that the retroviral element, MYSERV inserted independently in the two blocks containing mm17M1-29 and mm17M1-30P. In contrast to the conservation of repeats within these regions, the ORs are less well conserved: nucleotide identities range from 80.1-87.2%.

One large local duplication in figure 5.8, box G accounts for the duplication of an olfactory receptor gene (mm17M1-1 and mm17M1-5P), and it also resulted in the duplication of another gene within the region. This event involved a 10-11 Kb piece of sequence, flanked by MIRs and a B1 SINE. The inclusion of SINEs within the segment suggests it was a fairly recent event, followed by a mutational process that turned

mm17M1-5P into a pseudogene. The other ORs within this region that belong to the same subfamily as mm17M1-1 and mm17M1-5P, however, are not associated with block duplications. A comparison of the nucleotide identities of these 5 OR genes suggests that the reason for failing to detect block duplications can be attributed to the different nucleotide similarities (Table 5.3). It is only mm17M1-1 and mm17M1-5 that share over 90% nucleotide identity, which suggests the other OR genes must either have duplicated less recently or that mutational forces have been acting on this area at a faster rate. If there has been a faster rate of mutation at these loci, the rate must also have affected the repeats around these genes. A final consideration is that some of these genes duplicated through a different mechanism to many of the other ORs in the cluster.

	mm17M1-1	*mm17M1-2*	*mm17M1-3*	*mm17M1-4*
mm17M1-2	86.8			
mm17M1-3	86.8	85.5		
mm17M1-4	89.9	87.6	87.1	
mm17M1-5P	96.4	88.9	86.7	86.4

Table 5.3: Nucleotide percentage identities within subfamily 9. The highest percentage identity is between mm17M1-1 and mm17M1-5P. The subfamily members mm17M1-1 and mm17M1-5P are the only 2 OR genes within this family that appear to have arisen in a recent block duplication event.

Mm17M1-23 and mm17M1-33 belong to the subfamily 4. There is some shared sequence similarity in the regions where the two genes located, notably an imperfect $(GA)_n$ repeat, but there are no repeats that are shared between the two regions. This lack of detectable block duplication is something that could provide evidence that some genes have other methods of duplication, since the 2 OR genes have a shared nucleotide identity of 98.1%. This close identity suggests a different method of duplication or, alternatively, a degree of selectional pressure to conserve these genes independent of their repeats seen nowhere else in this cluster.

Genes mm17M1-10 and mm17M1-11 (nucleotide identity of 90.9%) are associated with the duplication of a 4-5 Kb block flanked by two LINE repeats and carrying a MTE repeat. Mm17M1-27 and mm17M1-26 (nucleotide identity of 95.1%) are associated with a 5-6 Kb block also carrying a MTE repeat but flanked by B1_MM and BGLII repeats. Mm17M1-28 is very similar to both of these genes (about 93.1%) but the three regions only have a MTE repeat in common.

There is, therefore, evidence that local duplication processes have produced most (34) of the OR genes within the cluster. A summary of these local duplication is shown in Figure 5.9. There are, however, a number of OR genes that could not be found to have duplicated through local duplication. These exceptions are mm17M1-41, mm17M1-39, mm17M1-45, mm17M1-44P, mm17M1-21, mm17M1-34, mm17M1-46, mm17M1-6, mm17M1-2, mm17M1-4, and mm17M1-3. Of these exceptions, mm17M1-44P, mm17M1-34, mm17M1-29, and mm17M1-6 are unique within the region, in having no subfamily members so local duplications are not expected, but the fact that the other exceptions do have subfamily members present in the region suggests either that these local duplications happened a relatively long time ago and all similar repeats have been displaced or changed, or it suggests that there is a different mode and mechanism of duplication for these genes.

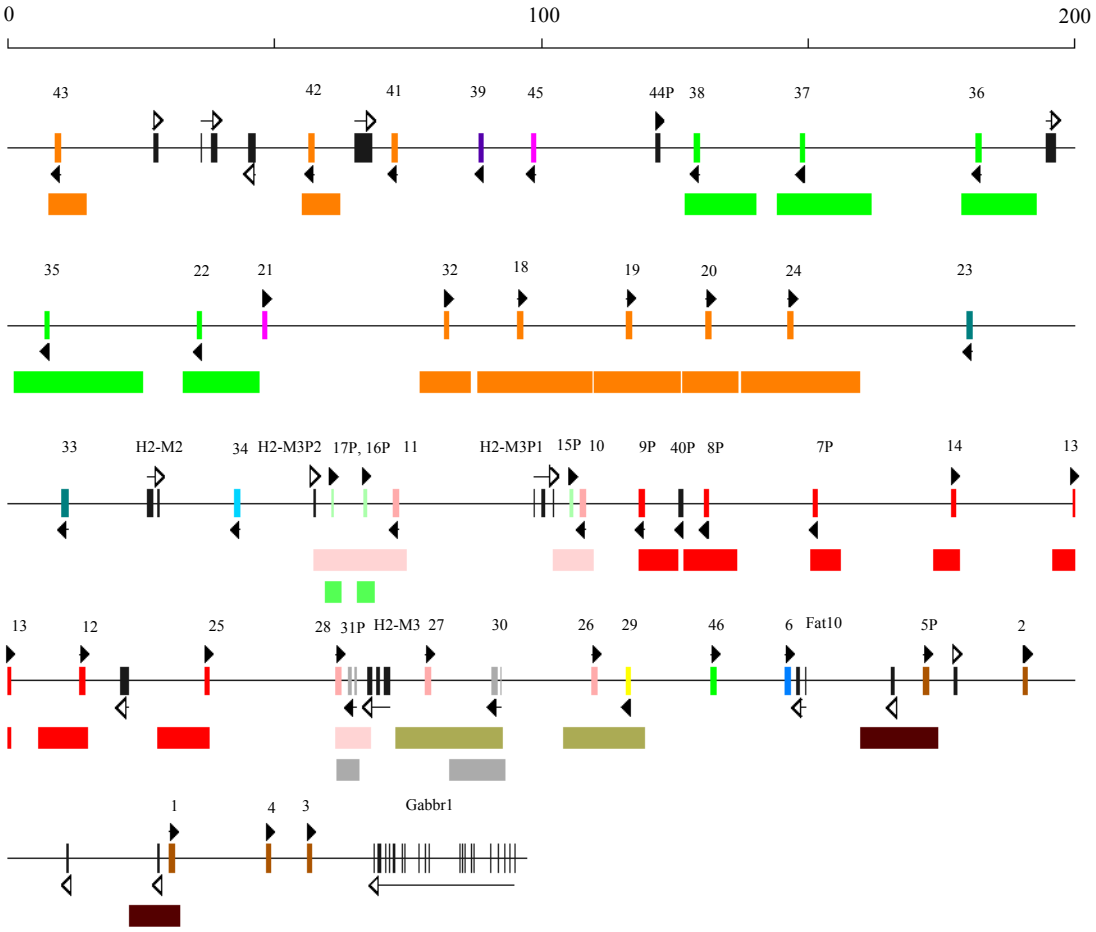


Figure 5.9: Block duplications within the mouse MHC-linked OR cluster. OR genes are named, and the colours represent the subfamily to which the OR genes belong (referred to previously in Table 5.3). Putative blocks are coloured according to the OR genes they are involved in duplicating, with the exception of the block involved in the duplication of mm17M1-27 and mm17M1-30 and the duplication of the block involved in the duplication of mm17M1-26 and mm17M1-29. The blocks involved in the duplication of mm17M1-11, mm17M1-10 and mm17M1-28 appear to be implicated in the duplication of the MHC class I-like genes (with the exception of H2-M2) in addition to their role in duplicating the three OR genes. Repeats in these regions are less highly conserved than in blocks associated with just OR gene duplications, suggesting this was a much earlier event in the history of the mouse MHC. In places where there are two blocks, two separate duplication events have occurred. For example, after the duplication of the mm17M1-11 and MHC class I-like fragment block, a later duplication produced mm17M1-17P and mm17M1-16P. Similarly, the block mm17M1-30P duplicated to form the mm17M1-31P block after a much earlier duplication involving mm17M1-28 and H2-M3.

5.12. Local duplications of MHC class I and class I-like genes.

Local duplications can also be associated with the duplication of other genes within the MHC-linked OR contig. These local duplications are shown in Figure 5.10: 3 blocks with similar nucleotide content have been delineated. Block 1 consists of the region around H2-M3, including 2 OR genes, mm17M1-31P and mm17M1-28. This has duplicated to produce block 2 which contains a MHC class I-like pseudogene (H2-M3P1) located next to the OR genes, mm17M1-15P and mm17M1-10. A block of 13 Kb can be implicated in this duplication event, which has a L1_MM repeat at one end. Since this duplication event, a large number of mutations have reshaped these two blocks of sequence. In block 2, for example, a 3 Kb piece of sequence containing 4 H2-M3 exons has been excised whilst between mm17M1-15P and H2-M3P1, a number of repeats have been inserted removing the first part of the mm17M1-15P pseudogene.

The origins of the second H2-M3 pseudogene, H2-M3P2 located in block 3, also appear to be associated with duplications involving OR genes. This pseudogene appears to have originated from H2-M3P1 and it involved a block duplication of 8 Kb (from block 2). This duplicon is flanked by a L1_MM repeat at one end, and alongside a fragment of the H2-M3P1 pseudogene, it carried mm17M1-15P and mm17M1-10 which were mutated, becoming mm17M1-17P and mm17M1-11. Repeats were inserted into block 3 alongside mm17M1-17P, and a block of repeats containing mm17M1-17P duplicated to produce another block of repeats and mm17M1-16P.

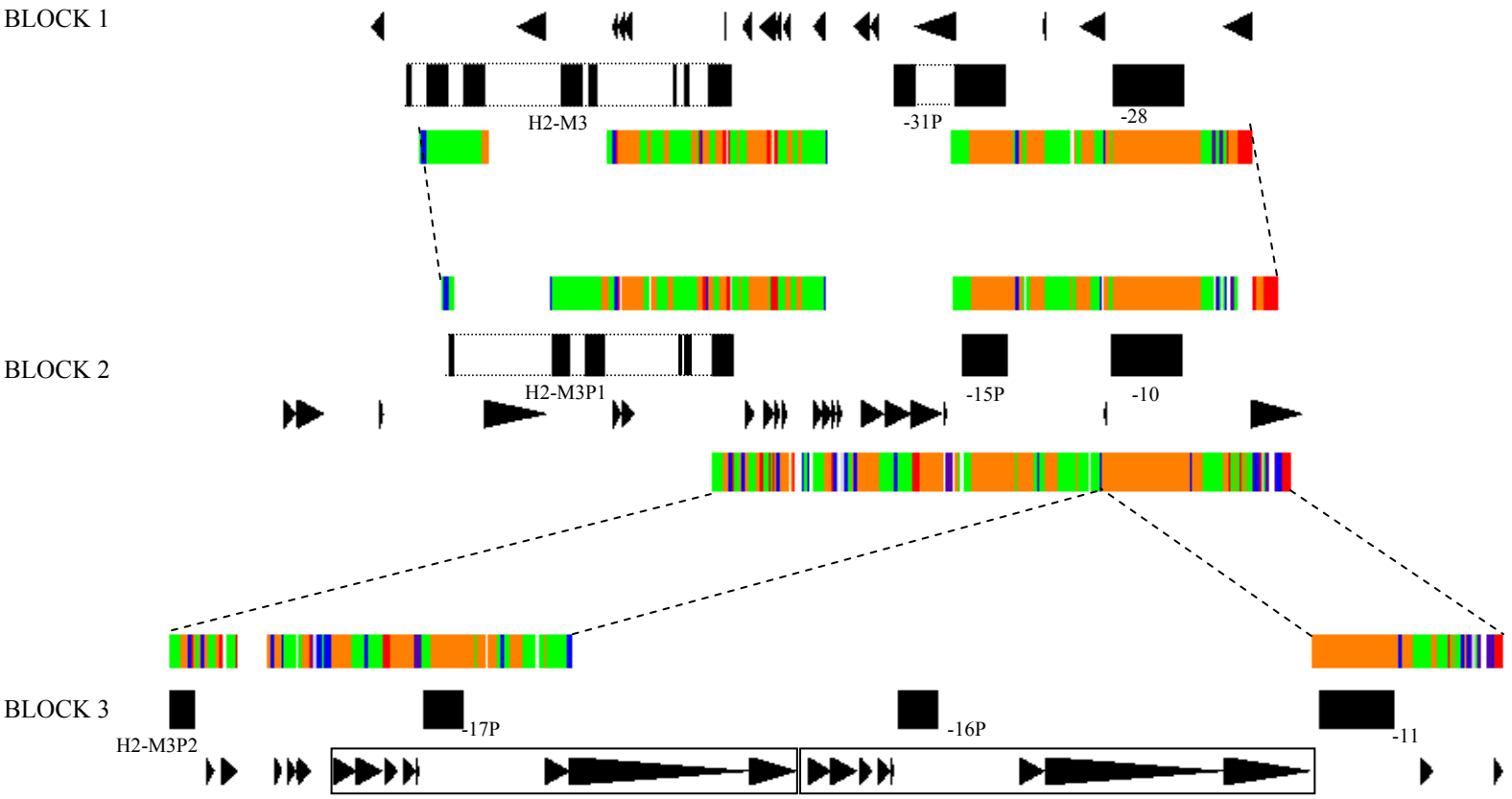


Figure 5.10: Duplications associated with the H2-M3 loci in the mouse MHC-linked OR cluster. The exon content and the repeat content of each block is indicated by the square blocks (exons) and triangles (repeats). Conserved blocks of sequence are indicated by coloured blocks which represent shared nucleotide identity of > 90% (red), > 80% (orange), > 70% (green), > 60% (blue), > 50% (purple) and > 40% (grey). Block 1 initially duplicated to form block 2. Within block 2 repeat insertions and deletions rendered H2-M3P1 a pseudogene, and the first part of mm17M1-15P was also lost. Part of block 2 then duplicated to produced block 3. Within block 3 repeats were inserted and one block of repeats containing mm17M1-17P (indicated by the boxed repeats) duplicated to produce another block of repeats containing mm17M1-16P (also indicated by the boxed repeats).

5.13. Identification of MHC-linked OR orthologs in mouse and human

Separate analyses of both the human and the mouse MHC-linked OR clusters, therefore, reveal that within the mouse lineage a number of relatively recent duplications (classed as relatively recent owing to high degree of shared nucleotide identity in both repeat content and the coding region of OR gene) have occurred to shape the cluster. In contrast to this, only one major duplication could be detected within the human MHC-linked OR cluster. Comparing the two regions, therefore, a number of mouse OR genes would be expected to have duplicated from an ancestral gene that may not have duplicated at all in the human lineage.

A simple comparison of the protein sequences of all the human and mouse OR genes within the two assembled sequences reveals that there are 10 groups of what can be considered to be orthologous genes (Orthologous genes were defined as sharing over 70% protein identity with a gene in the other species). The relationship that is suggested by this analysis is shown in figure 5.11.

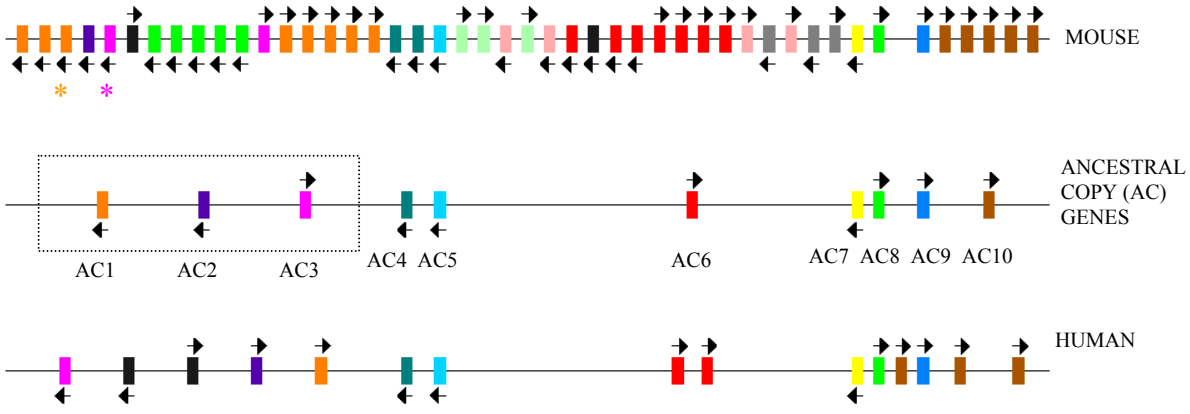


Figure 5.11: Orthologous olfactory receptor gene within the MHC-linked OR clusters in human and mouse (not to scale, other genes within the region not shown). A putative ancestral arrangement of OR genes is represented on the middle line: the boxed OR genes could be in either order as it appears an inversion of these genes has occurred in either the mouse or the human lineage.

10 putative ancestral framework genes (ancestral copy (AC) genes) appear to have existed and these have followed separate evolutionary histories in the two species. For example, AC10 has duplicated to produce 2 copies in the human extended MHC, whilst on the mouse chromosome 17 it has duplicated, producing 5 copies. Similarly, AC6 has 2 copies in the human region and 7 copies in the mouse region. In contrast to AC6 and AC10, 4 other ancestral genes (AC5, AC7, AC8 and AC9) appear to have duplicated in neither species.

Further away from the classical MHC, moving out towards the telomere, 3 other ancestral OR genes have been involved in an inversion in either the mouse or the human species. One of the genes involved in this inversion, AC1, has duplicated numerous times in mouse, but it remains as a single copy gene in the human region. The original mouse descendant of AC1 is indicated by the asterix in figure 5.11: this gene (mm17M1-41) shares 80% protein identity with the human descendant (hs6M1-28). AC2 has only one descendant in each species, whilst AC3 has duplicated once in the mouse genome. In the case of the original mouse descendant of AC3, protein identities of the two OR genes compared to the human counterpart are very similar (76% and 71%); the protein with the higher protein identity was considered to be the original descendant of AC3 (indicated by the asterix).

Other orthologous genes existing within the extended MHC region are the Gabbr1 gene and the Fat10 gene, which both share high protein identity and a similar position in both species. Other genes which are found in the mouse and human MHC-linked OR clusters are not orthologous: they must have been inserted or deleted since human-mouse divergence.

5.14. Conservation of sequence outside OR coding regions in mouse and human

The existence of a large number of orthologs across the region, and the formation of orthologous gene groups suggested that across the region there may be a high degree of conservation. A dot-matrix plot of the MHC-linked major olfactory receptor cluster against the contiguous 897 Kb of mouse sequence (Figure 5.12), however, revealed very little general conservation across the region, with the exception of a well conserved region around the GABBR1 locus.

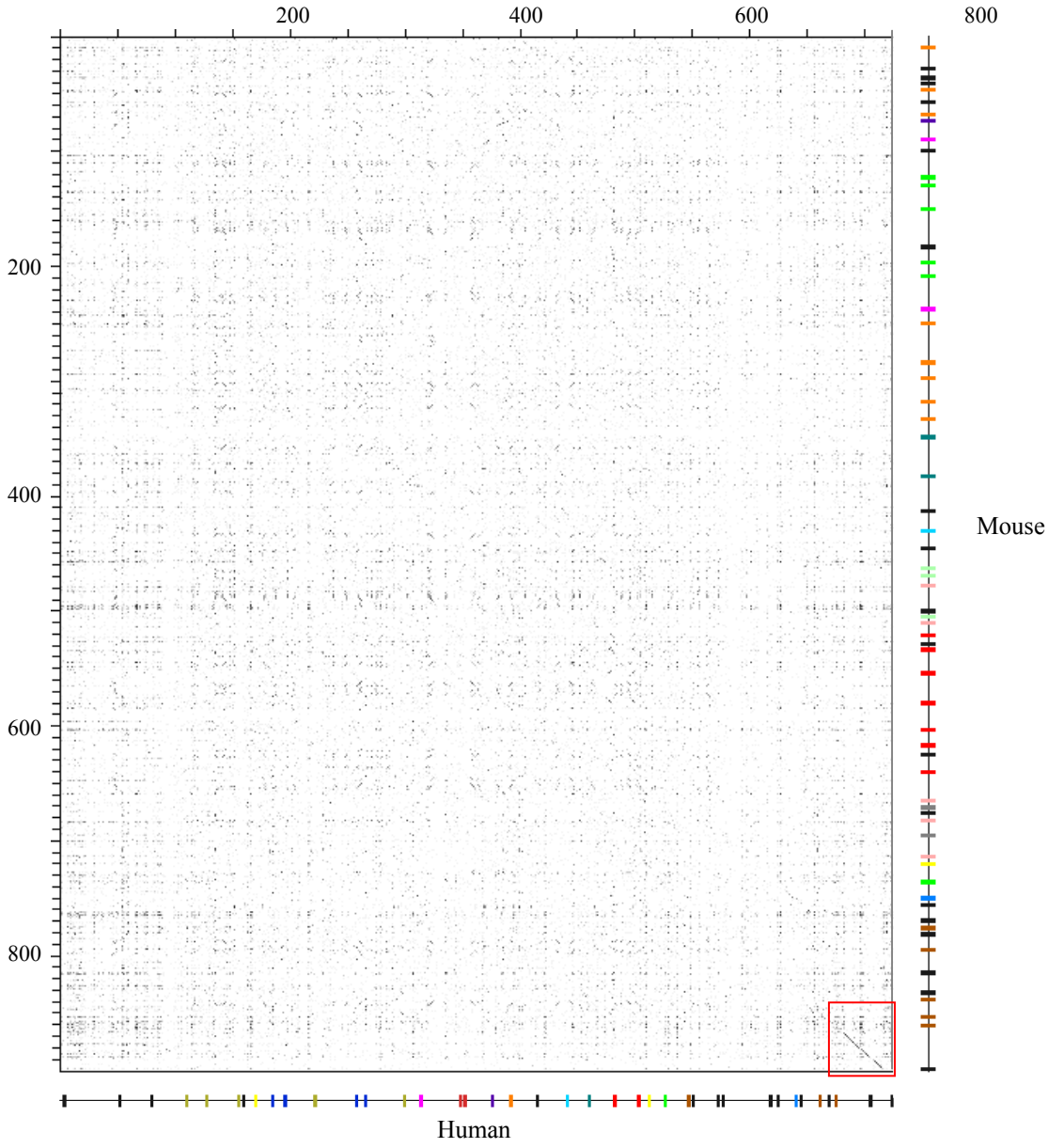


Figure 5.12 (previous page): Dot-matrix plot of human major MHC-linked OR cluster against the mouse MHC-linked cluster. The mouse sequence is plotted on the vertical axis whilst the human sequence is plotted on the horizontal axis. Bars under/to the right of the plot show the gene content of the region: ORs are coloured according to their subfamily (or pale green where they have no subfamily). For gene names refer to Figure 4.6 (human) and Figure 5.7 (mouse). The red square shows the only region that is highly conserved at this resolution which corresponds to GABBR1 locus.

This lack of conservation is something that can also be seen in percentage identity plots (PIPs) that were generated using the mouse and human sequence (data not shown). Disregarding the region around the GABBR1 locus, the largest amount of conservation that is detectable is located around the olfactory receptor genes. The open reading frames of all the olfactory receptors show a considerable amount of conservation (over 60%). In addition to this conservation, in some cases, additional conservation is found in sequences located next to the OR loci. This conservation is generally located within 3 Kb of the 5' end of the OR gene, although in some cases, it extends further. Conservation of 3' untranslated regions can also be observed although this is less common and it does not extend as far as conservation at the 5' end of the OR gene.

Upon closer analysis, it is clear that the upstream conservation (or lack thereof) can be used to classify orthologous relationships. An example of this is provided by mm17M1-45. This OR is closely related to hs6M1-25P, and although there is another mouse gene with similarity to hs6M1-25P (mm17M1-21), it is clear from the conservation in the upstream regions (mm17M1-45 has extensive upstream conservation, whilst mm17M1-21 has no upstream conservation) that mm17M1-45 is the real ancestor (ortholog) of hs6M1-25P. Figure 5.13 shows this upstream conservation in mm17M1-45 and hs6M1-25P. Upstream conservation also confirms that mm17M1-41 is the ortholog of hs6M1-28, mm17M1-39 is the ortholog of hs6M1-22P, mm17M1-24 is the ortholog of hs6M1-27, and mm17M1-6 is the ortholog of hs6M1-14P.

Figure 5.13a

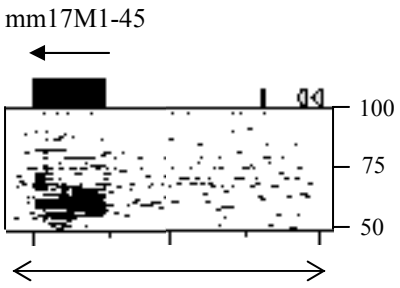


Figure 5.13b

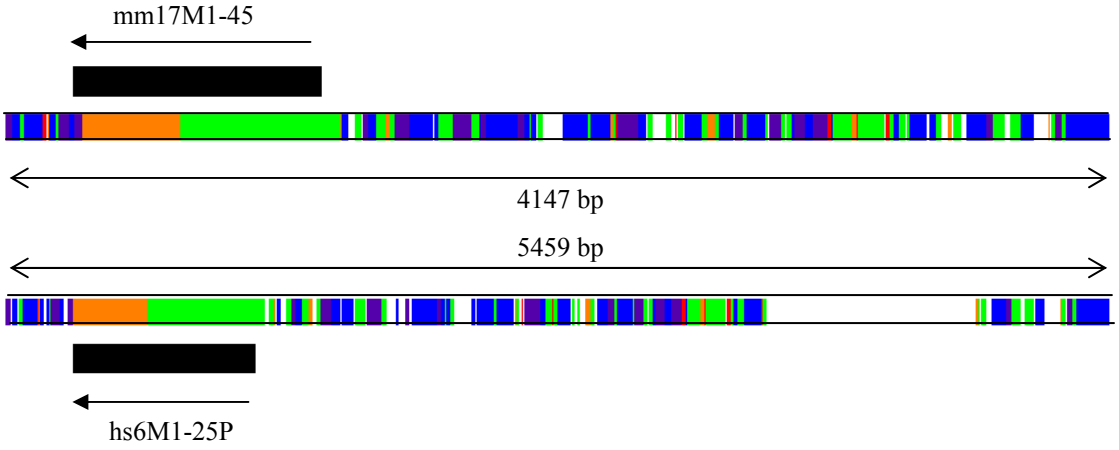


Figure 5.13a: mm17M1-45 PIP identity plot. Coding exons (shaded box) and repeats (triangles) are plotted on top of the box which contains lines showing segments of sequence conserved in the human MHC-linked OR cluster. The vertical position of these lines shows the similarity of these segments which can range from 50% to 100%. This plot reveals that a number of sequences in the region of mm17M1-45 are conserved in the human extended MHC. The large number of sequences conserved within the coding region of mm17M1-45 reflects the large number of olfactory receptor genes found in the human extended MHC. Analysis of the upstream region, however, reveals that the sequences conserved upstream of mm17M1-45 are all located upstream of hs6M1-25P.
Figure 5.13b: mm17M1-45 plotted against hs6M1-25P. Conserved blocks of sequence are plotted in their position relative to the olfactory receptor genes. The colours represent different levels of conservation ranging from >90% (red), >80% (orange), >70% green, >60% (blue), > 50% (purple) and > 40% (grey). Uncoloured blocks (white) indicate sequence which shares less than 40% nucleotide identity either owing to insertions or deletions, or to faster mutation rates.

Upstream conservation can also be used to consider the relationships of genes defined as belonging to orthologous groups. The subfamilies consisting of hs6M1-12 and hs6M1-13P, and mm17M1-1, -2,-3,-4 and –5P, for example, have been defined as belonging to an orthologous group of genes according to the protein sequence similarity they share in their coding regions. Analysis of conservation of untranslated regions around these genes, however, suggests that mm17M1-3 has an ancestral relationship with hs6M1-12, whilst the other 4 mouse genes (mm17M1-1, -2, -4 and –5P) have been derived from an ancestor of hs6M1-13P. (Figure 5.14

shows the difference in the degree of conservation between mm17M1-1 and hs6M1-12 and hs6M1-13P.) Similarly, with the mouse OR subfamily 7 (containing mm17M1-7P, -8P, -9P, -12, -13, -14, -25 and –34), 4 of these genes (mm17M1-7P, -8P, -9P and –13) appear to have been derived from hs6M1-20 rather than hs6M1-19P. The other 4 genes are either equally closely related to hs6M1-20 and hs6M1-19P (mm17M1-12, -14, and –25) or an ortholog of hs6M1-28 (mm17M1-34).

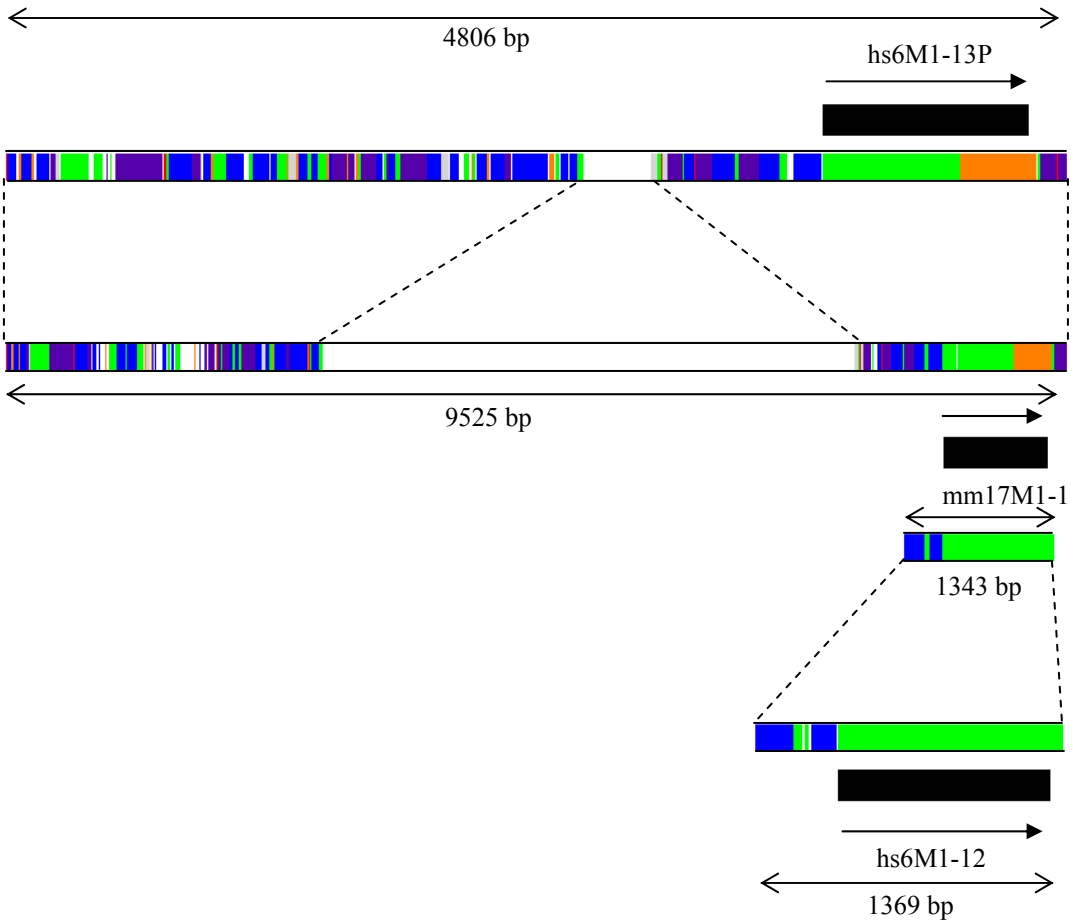


Figure 5.14: mm17M1-1 plotted against hs6M1-12 and hs6M1-13P. Conserved blocks of sequence are plotted in their position relative to the olfactory receptor genes. The colours represent different levels of conservation ranging from >90% (red), >80% (orange), >70% green, >60% (blue), > 50% (purple) and > 40% (grey). Uncoloured blocks (white) indicate sequence which shares less than 40% nucleotide identity either owing to insertions or deletions, or to faster mutation rates. The sequence upstream of hs6M1-13P shows a much higher amount of conservation than that the sequence upstream of hs6M1-12.

The two most centromeric orthologous groups related to hs6M1-12 and hs6M1-13P and hs6M1-19P and hs6M1-20, therefore, show a high degree of conservation in regions upstream of the

coding region of the olfactory receptor genes. In contrast to this, the two subfamilies (mm17M1-38, -37, -36, -35 and –22, and mm17M1-32, -19, -18, -20 and –24) that have duplicated near the telomeric end of the mouse area do not show upstream conservation compared to their human orthologs. This suggests that the duplications involved in the formation of these subfamilies occurred much later after human-mouse divergence than the duplications associated with the more centromeric orthologous groups.

Conservation of DNA between mouse and human has been suggested to imply that these sequences have a functional importance. Figure 5.15 shows regions that are conserved in hs6M1-21 compared to 2 related genes, mm17M1-23 and mm17M1-33. Upstream of the hs6M1-21 gene, the same sequence has been conserved in both genes supporting the idea of a functional role for this sequence, since otherwise chance would be expected to mutate different upstream sequences in both orthologous genes.

In conclusion, therefore, analysis of local conservation upstream of olfactory receptor genes reveals that some genes can be defined as orthologs according to conservation of sequences in the upstream region of these genes. This suggests the repertoire of MHC-linked olfactory receptor genes in the common ancestor shown in Figure 5.11 is an oversimplification: it is likely there were at least 2 copies of AC6, corresponding to hs6M1-20 and hs6M1-19P, and at least 2 copies of AC10 corresponding to hs6M1-12 and hs6M1-13P. Local conservation also supports the idea of conservation of sequence for a functional reason: conservation is generally only located upstream of olfactory receptor genes, and in genes descended from the same ancestral gene, it appears that the same upstream regions have been conserved.

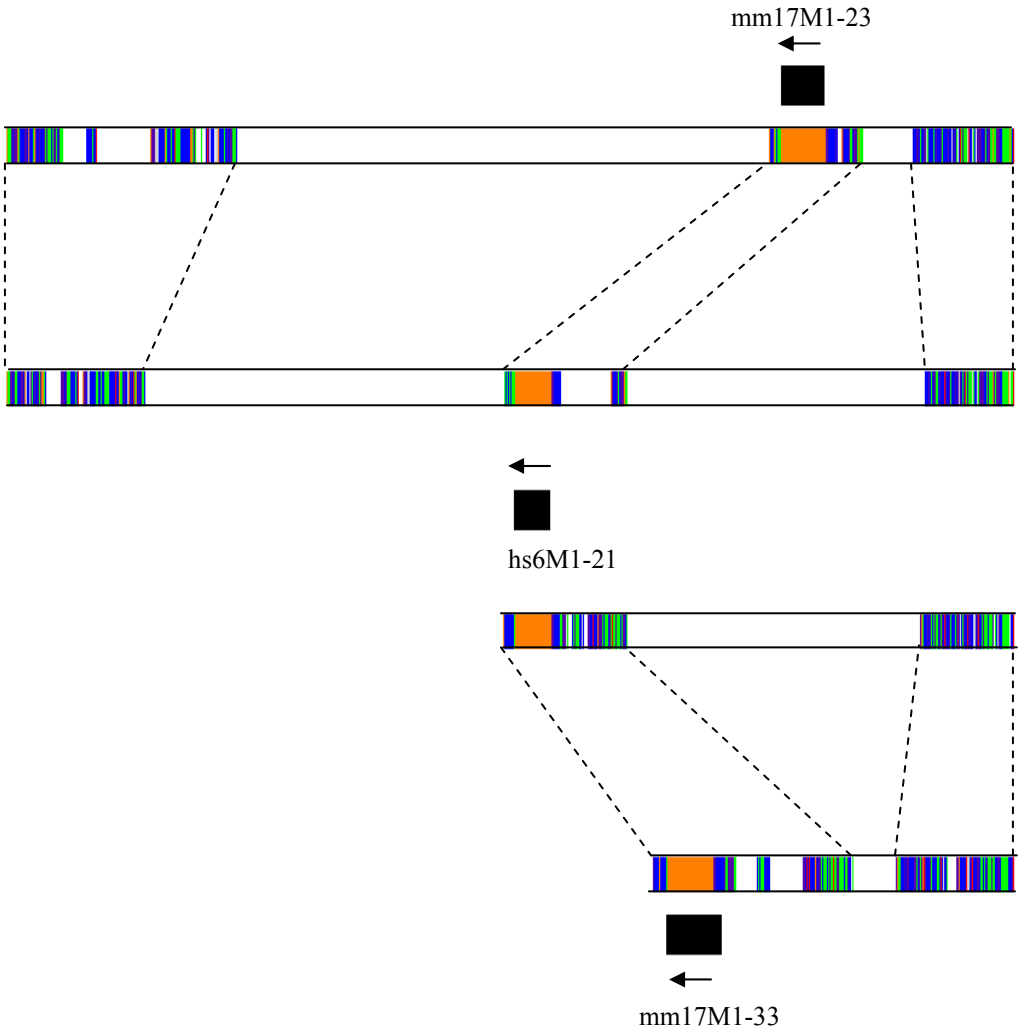


Figure 5.15: hs6M1-21 plotted against mm17M1-23 and mm17M1-33. Conserved blocks of sequence are plotted in their position relative to the olfactory receptor genes. The colours represent different levels of conservation ranging from >90% (red), >80% (orange), >70% green, >60% (blue), > 50% (purple) and > 40% (grey). Uncoloured blocks (white) indicate sequence which shares less than 40% nucleotide identity either owing to insertions or deletions, or to faster mutation rates. The same sequences upstream of hs6M1-21 are conserved in both mm17M1-23 and mm17M1-33, although mm17M1-23 shows a much greater amount of downstream conservation.

5.15. Non-orthologous OR genes

Figure 5.11 also shows a number of genes that do not have orthologs in the two regions analysed in detail in this project. Within the mouse region, 11 genes, 2 of which are highly pseudogenic and 9 of which come from 3 subfamilies have no human counterpart within the extended MHC.

Searching the human database of OR genes (Chapter 8), however, reveals no clear human ortholog so these genes are likely to represent OR genes that have been lost from the human genome. Similarly, 2 human OR genes, hs6M1-23P and hs6M1-24P, are only distantly related to mouse MHC-linked OR genes. This suggests that there has either been a loss of olfactory receptor genes since the two species diverged, or it suggests that OR genes have been recruited into this region since divergence. Hs6M1-16 in the human lineage is another gene that can be predicted to have duplicated from hs6M1-12 or hs6M1-13P after divergence.

A combination of duplication, deletion and insertion processes seems likely to have created the repertoire seen in both species today. Considering the data, however, it is possible to hypothesize that whilst gene loss in humans seems to be prevalent in the region located nearest to the MHC, telomeric of the OR gene mm17M1-23, duplications in mouse have occurred frequently since divergence. This is supported by upstream non-coding conservation of OR genes, and it is also supported by the analysis of block duplications in the mouse: duplications telomeric of mm17M1-23 in the mouse have been well-characterised, whilst duplications centromeric of mm17M1-23 are less well-characterised, suggesting they were earlier duplication events. The common ancestor of mouse and human therefore appears to have more genes than are present in human in the centromeric part of the cluster but fewer genes than are present in mouse in the telomeric part of the cluster.

5.16. Identification of orthologous ORs upstream of the original contig

The availability of mouse draft sequence (from mouse strain C57BL/6J) from the public sequencing effort, accessed using the UCSC genome browser (Mouse Feb. 2002 draft assembly) allowed the contents of mouse sequence telomeric of the partial MHC-linked OR contig to be

analysed. This resulted in the identification of 10 further mouse olfactory receptor genes on chromosome 17, and 13 olfactory receptor genes on mouse chromosome 13, located relatively near the murine version of hfe. As this sequence is unfinished, there may be more mouse olfactory receptor genes than those listed in Appendix 10, and the order may also change as more sequence becomes available. Nevertheless, adding this data to the data shown in Figure 5.11, (to produce Figure 5.16) allows a larger picture of the history of the MHC-linked olfactory receptor cluster to be built up.

From Figure 5.16, it is obvious that the majority of human and mouse genes possess orthologs or orthologous groups in the other species. Across both the major and minor MHC OR clusters, only 28% of the mouse ORs lack an orthologous relative, whilst in human only 17% show no obvious orthologous relationship within the cluster. It also appears that the order of the genes is broadly conserved in the two species, and the breakpoint in synteny can be defined as occurring around the olfactory receptor genes orthologous to hs6M1-2P. Interestingly enough, this is also the region at which at a local duplication has been observed in the human lineage: possibly the sequence around this locus has a higher rate of recombination that may play a role in local duplication processes or mechanisms involved in separating or bringing together clusters of genes.

Figure 5.16a

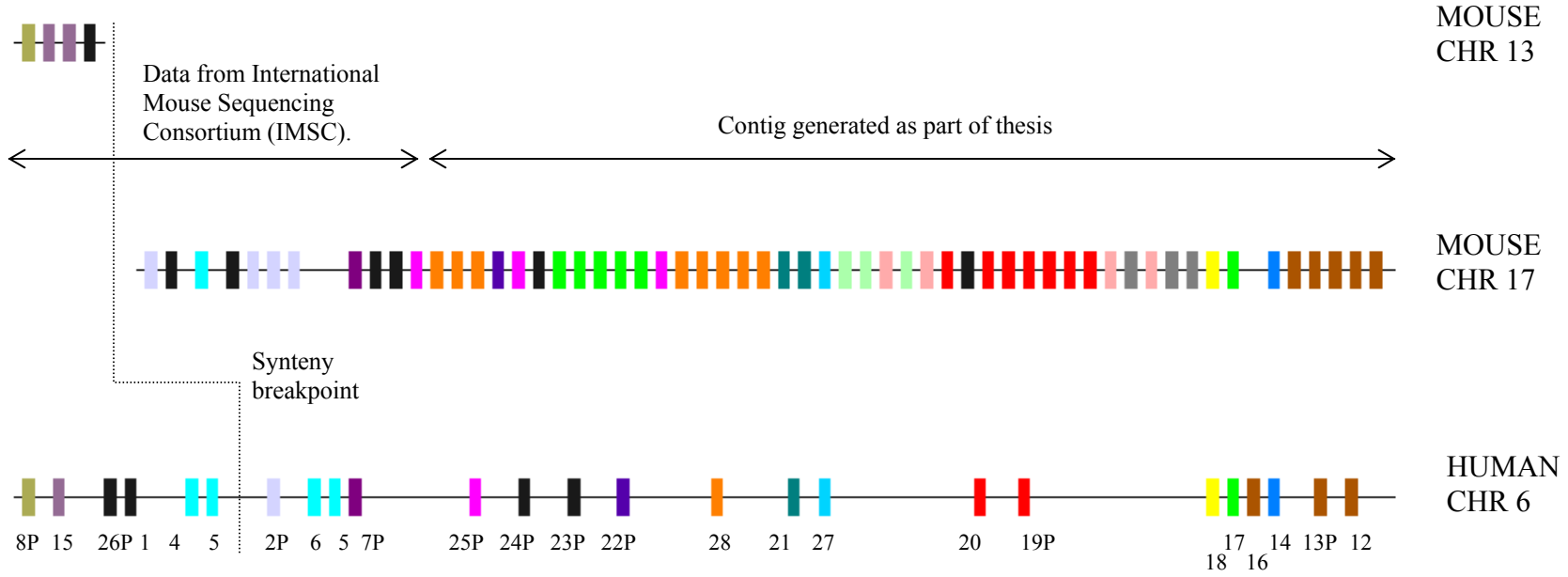


Figure 5.16b

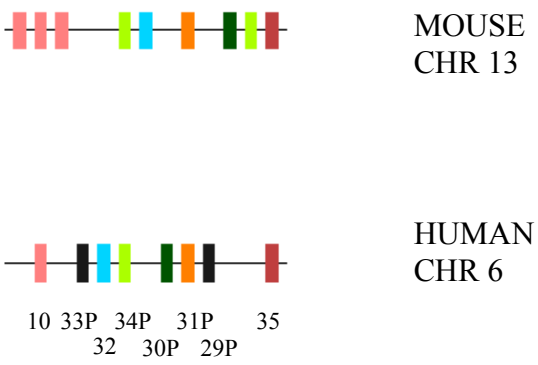


Figure 5.16: Orthologous ORs within the extended MHC major OR cluster (a) and within the extended MHC minor cluster (b). The genes are coloured according to their relationship to OR genes within the other species. The breakpoint in synteny appears to occur within the major OR cluster, around the hs6M1-2P gene. Human OR genes are labelled with their number: species and chromosome designations have been left out for reasons of clarity.

Another observation that can be made with regard to Figure 5.16 is that there appears to have been either more duplication of ORs within the mouse lineage or more loss of human ORs since speciation in the major MHC-linked cluster as opposed to the minor cluster. In the major cluster, there are a number of examples where 1 orthologous human gene has around 4 OR relatives within the mouse genome. In contrast to this, the largest orthologous cluster within the minor OR cluster possesses 3 OR genes. In the light of the fact that the extended MHC class I region, and indeed the MHC region in general can be seen to be a region of the genome where local duplication is a common phenomena, it could be hypothesized that the increased number of orthologs in the major MHC-linked is due to the proximity of this region to the MHC.

From the literature, it is possible to compare other human and mouse orthologous clusters to find out whether this theory of increased duplication owing to proximity to the MHC stands up. Four orthologous clusters of OR genes have been analysed, although all these analyses were on a smaller scale compared to the MHC-linked OR cluster. The chromosome 17 human OR cluster was compared against a mouse OR cluster located on mouse chromosome 11B3-11B5 (Lapidot *et al.*, 2001). The human chromosome 17 cluster contains 17 genes, compared to 13 amplified from mouse genomic clones. Considering results from this paper, and applying definitions of orthologs used in this project, it appears that 7 orthologous groups can be defined. One of these groups shows a significant increase in the number of genes in mouse (2 in human compared to 5 in mouse) and another shows a significant increase in the number of genes in human (4 in human compared to 1 in mouse), but the other 5 groups show 1 to 1 or 1 to 2 relationships suggesting duplication or deletion mechanisms have not acted as strongly as they have in the MHC-linked cluster. The order of OR genes appears to have been conserved between the 2 syntenic clusters, as it has in the MHC-linked cluster.

The analysis of the human and mouse OR clusters located next to the β-globin gene clusters (Bulger *et al.*, 2000) also suggests there has been less duplication or deletion within this cluster compared to the MHC-linked cluster. In this case, of the 6 orthologous groups, 4 groups have a 1 to 1 relationship, whilst 1 group has 1 human OR gene to 2 mouse OR genes and another has 2 human OR genes to 1 mouse OR genes. The relationship of a cluster of OR genes located on mouse chromosome 7 to a syntenic cluster on human chromosome 11p15.4 (Lane *et al.*, 2001) also suggests that 1 to 1 relationships are prevalent: 6 groups were found with this relationship, whilst another orthologous group contains 2 genes on human chromosome 11 and 4 genes on mouse chromosome 7. This paper, however, does provide evidence for an expanded mouse repertoire as there are 2 additional groups containing 7 OR genes for which no ortholog was found. An analysis of the OR cluster located next to the mouse and human T-cell receptor alpha/delta loci was also performed (Lane *et al.*, 2002). Five orthologous groups with a 1 to 1 relationship were identified; a sixth group had 1 human OR gene to 2 mouse OR genes. As in the three other studies, the order of these olfactory receptor genes has been conserved between species.

Reviewing the mouse-human orthologous OR cluster literature, therefore, it appears that the MHC-linked major OR cluster has undergone a more severe process of duplication or deletion compared to other syntenic clusters. However, this conclusion should be treated with caution as these studies provide a snapshot of syntenic clusters rather than a comprehensive picture (especially given the small sizes of the regions and the lack of complete sequencing across regions). The functional repertoire of mouse OR genes has been suggested to be 50% greater than the human repertoire (Young *et al.*, 2002) and so clearly other syntenic regions may have a similar degree of expansion in the mouse lineage or contraction in the human lineage to that in the MHC-linked major OR cluster (26 human ORs compared to 56 mouse ORs suggests an increase of 54.6 % in the mouse lineage, or a decrease of 54.6% in the human lineage). The high

number of mouse OR genes, therefore, suggests a number of clusters may have undergone duplications or deletions similar to that observed by the MHC-linked major OR cluster, as it appears that many OR genes do not have a single clear identifiable ortholog (Young and Trask, 2002).

In conclusion, therefore, comparing the mouse orthologous major and minor MHC-linked OR clusters suggested there had been significantly more duplications or deletions within the major cluster. It was hypothesized that this could be explained by the proximity of the major cluster to the MHC but although small scale studies might support this, the genome wide distribution of mouse OR genes suggests that local duplications occurred across the genome to create a larger mouse repertoire of olfactory receptor genes. At the same time, however, there are examples of MHC-linked OR genes that do not have orthologs within the human genome and so there may be mouse OR genes within the 1500 that have been lost from the human genome. Further characterisations of mouse and human OR clusters are required to support the idea that there has been a larger amount of duplication within the mouse major OR cluster compared to other mouse OR clusters.

5.17. Conservation of orthologs in non MHC-linked OR clusters

Two clusters of OR genes from chromosome 2, 1 syntenic to chromosome 9 and another syntenic to chromosome 11 were considered in more detail to check the amount of mouse OR duplication in both clusters. The results from this are shown in Figure 5.17. These results are based on unfinished sequence and extra genes may be identified and the gene order may be altered as the sequence is finished but in spite of these problems, both clusters show a large duplication in the mouse lineage producing 7 or 5 mouse OR genes in comparison to 2 related OR genes in the

human clusters. These large duplications suggest that the MHC-linked OR gene cluster is not exceptional in the amount of duplication there has been in the mouse lineage.

Figure 5.17a

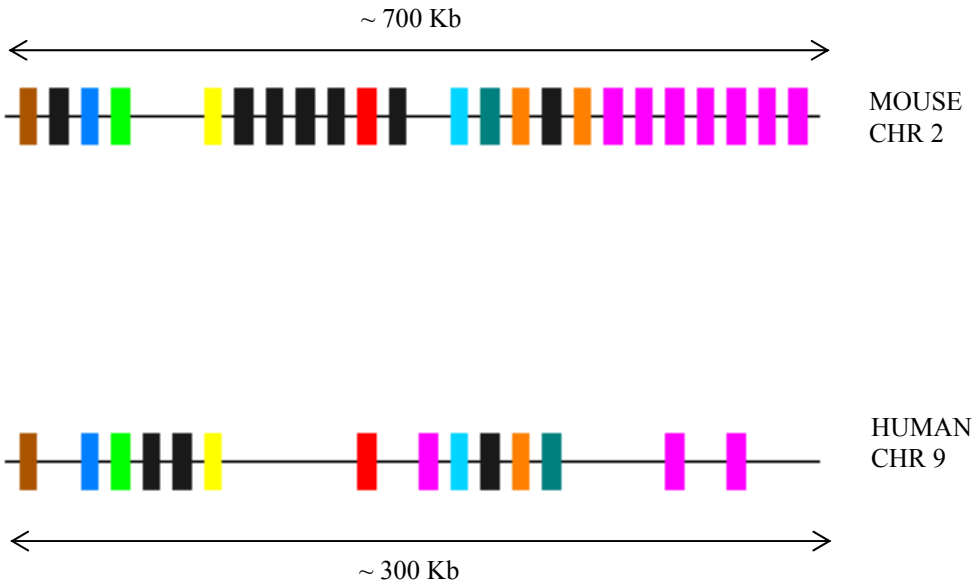


Figure 5.17b

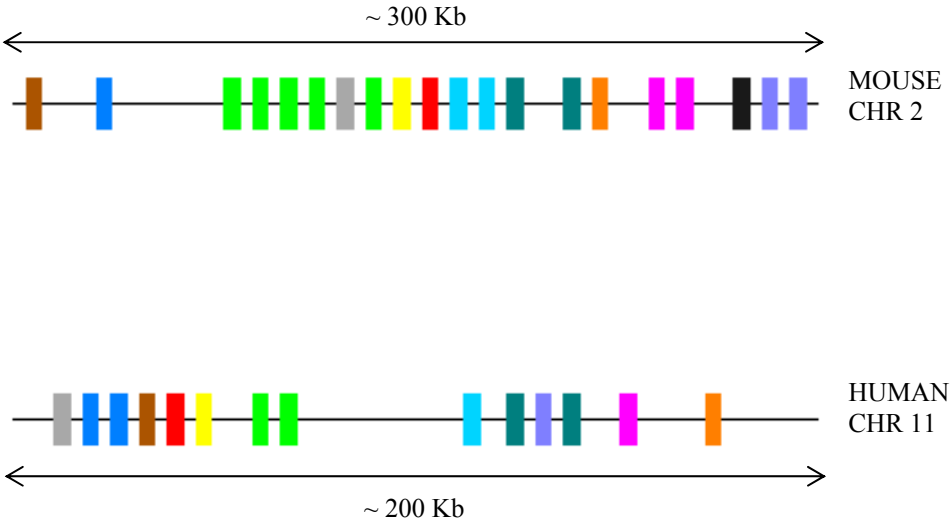


Figure 5.17: Orthologous clusters of OR genes on mouse chromosome 2 and human chromosomes 9 (a) and human chromosome 11 (b) . Orthologous groups were defined by a protein identity > 70%: they are indicated by the various colours in the 2 figures.

5.18. Conclusions.

In order to analyse the mouse MHC-linked OR cluster, a clone contig was assembled and an efficient tiling path was chosen for sequencing. The sequence was assembled into a 897213 bp contig that was analysed and was found to contain 46 olfactory receptor gene loci, 36 of which were considered to be functional. These olfactory receptor genes can be divided into several subfamilies, and an analysis of the amino acid conservation across the cluster revealed that they share several structural features with the human MHC-linked OR genes with regard to hypervariable regions and putative disulphide bridges. As is the case with the human MHC-linked OR genes, the mouse MHC-linked OR pseudogenes also offer limited support for the idea of mutational hotspots within OR genes as 2 independently evolved ORs appear to have a mutation at the same relative position. The genomic environment of the mouse OR genes is also similar to that identified for the human cluster, although across the cluster only 3 non-OR genes, GABBR1 FAT10 and SMT3H2 are found in both species.

One major difference between the mouse and human MHC-linked OR clusters is the presence of MHC class I genes within the mouse cluster. In the case of the H2-M3 genes and pseudogenes, these MHC class I genes appear to have duplicated alongside OR genes, suggesting an old association between the MHC and OR genes. Other extensive duplications have been involved in the proliferation of OR genes throughout this region of the mouse genome.

A more detailed analysis of the mouse and human region identified 10 orthologous groups of olfactory receptor genes, suggesting the common ancestor may have contained 10 ‘framework’ genes. Analysis of upstream regions, however, suggested the situation was more complex than this, with conservation of upstream regions common amongst those OR genes located nearest to the MHC. This appears to suggest a number of these genes may have been present in the common

ancestor: this is supported by the lack of observable block duplications around this region. OR genes telomeric of mm17M1-23 show less upstream conservation between the two species; in addition, they also appear to have duplicated fairly recently as evidenced by the conservation of identity between blocks of mouse sequence (Figure 5.8). Different evolutionary pressures therefore appear to have acted on these two regions, with OR genes nearest the MHC marked by gene loss since human-mouse divergence, and genes further away from the MHC marked by gene duplication since human-mouse divergence.

The availability of mouse draft sequence allowed the comparative analysis to be extended further: an additional 10 mouse ORs were found on chromosome 17, together with an additional 13 on mouse chromosome 13. From this unfinished sequence it appears that the synteny breakpoint is located near a cluster of OR genes orthologous to hs6M1-2P. This region is also the only region in which local duplication could be deduced in the human lineage.

The relationship between the mouse OR cluster and the human OR cluster was compared with other orthologous mouse clusters (4 from the literature and 2 that were identified using unfinished sequence). There was a high degree of duplication in the mouse MHC-linked OR clusters compared to other OR clusters taken from the literature, however, comparing 2 clusters on mouse chromosome 2 with their orthologous clusters on human chromosome 9 and human chromosome 11, results suggested that there was no significant difference in the amount of duplication that could be observed.