Identification and analysis of MHC-linked olfactory receptor genes

Ruth M Younger

A thesis submitted in partial fulfilment of the requirements of Cambridge University for the degree of Doctor of Philosophy.

Newnham College, Cambridge University.

August 2002

This dissertation does not exceed the length limit set by the Biology Degree Committee.

Abstract/Summary of thesis

Olfactory receptor genes (ORs) are members of the largest known human multigene family of 7-transmembrane receptors; over the past few years, 900-1000 have been identified, existing largely within clusters located on the majority of chromosomes. As their name implies, they are involved in odour perception within the major olfactory epithelium (MOE), although their expression in other tissues, notably the testis, lungs and kidney seems to suggest some non-olfaction related role. The cluster of olfactory receptors located next to the Major Histocompatibility Complex (MHC) were considered to be especially interesting, given that the association between this cluster and the MHC has also been conserved in mouse and rat, and it has been suggested that these genes are involved in detecting odours that control MHC-disparate mate selection.

The aim of this thesis was to identify all olfactory receptor genes located distal to the class I region of the Major Histocompatibility Complex in both the human genome and in the mouse genome, and to study phylogeny, regulation, expression and polymorphism to develop a better understanding of the structure, function and evolution of these genes. In total, 34 MHC-linked ORs were identified in human; within the mouse genome, a larger number of MHC-linked ORs (56) were also identified. Comparing the 2 species, orthologous groups can be identified: groups appear to have undergone both deletions and duplications since mouse-human divergence. An investigation of regulatory elements within the human MHC-linked OR cluster revealed no specific regulatory elements, although a putative locus control region has been identified. Compared to the HLA genes, these OR genes show a limited amount of polymorphism, although existing polymorphisms act to alter the functional repertoire of these genes between individuals and one specific OR gene appears to have an unexpectedly high level of polymorphism.

Acknowledgements

I am indebted to my supervisor at the Sanger Centre, Stephan Beck, whose knowledge, enthusiasm, patience and resourcefulness have seen me through whenever mine have failed. Many thanks also go to my supervisor at Cambridge University, John Trowsdale. I am also very grateful to a number of people with whom I have worked over the course of this project: Armin Volz and Anke Ehlers at Institut für Immungenetik, Universitätsklinikum Charité, Humboldt-Universität, Berlin, Simon Forbes at Cambridge University, and Claire Amadou at CNRS, Toulouse – it has been both a highly informative and a highly enjoyable experience. Special thanks must also go to Andreas Ziegler (Institut für Immungenetik, Universitätsklinikum Charité, Humboldt-Universität, Berlin) – your frequent phone calls were a source of inspiration, motivation and (I admit it, occasionally) fear ! Many thanks also to Kirsten Fischer-Lindahl (HHMI, Dallas, TX) – I have appreciated all your input throughout the course of this thesis.

Moving on to the Sanger Centre, I would like to thank all past and present members of Team 50: Katie Evans, Roger Horton, Melanie Stammers, Vikki Rand, Karen Novik, and Karen Halls – thanks for all the input and advice and the cups of tea and moaning (mine, generally)… I am also very grateful to members of the now (non-existent) Team 33 who did much of the sequencing of this region, as well as members of the Chromosome 6 and mouse project groups at the Sanger Centre. Others at the Sanger Centre who have helped me over the course of this project are too many to name in full - I am privileged to have worked in an environment where help has been offered so freely – but special mention should go to Alex Bateman, Kevin Howe, Matt Craig, Graeme Bethel, Carol Edwards and Andy Mungall. Thanks also to Jill Williamson and Denise Sheer of the Human Cytogenetics Laboratory (ICRF) for their help with the FISH analysis.

On a personal note, thanks must go to the various football teams I have played for and supported whilst doing this thesis – you have provided me with a philosophical approach that being associated with more successful clubs just would not be able to give me and distracted me when I most and least needed it. Thanks to my friends who have allowed me to become even less responsible and cut me huge amounts of slack over the last few months – your understanding and patience is always a lot more than I deserve. I am also grateful to my own source of sunshine, even though it got cloudy, the cloud breaks were great !! Many thanks also go to my parents, Michael and Glynis, for their pillar-like support, unswerving belief and love and the many sandwiches they have made me, especially in the last few weeks. Finally, thanks to the force that has moved in a most mysterious way to shape the big and small events in the universe – I am never grateful enough for everything I have.

Table of contents

List of tables

List of figures

List of abbrieviations

AOB	accessory olfactory bulb
(c)AMP	(cyclic) adenosine 5´-monophosphate
ATP (dATP, ddATP)	adenosine 5'-triphosphate (deoxy-, dideoxy-)
BAC	bacterial artificial chromosome
BLAST	basic local alignment search tool
bp	base pair
BTN	butyrophilin
°C	degrees Celsius
cDNA	complementary deoxyribonucleic acid
chr	chromosome
cm	centimeter
CNS	central nervous system
CpG	cytidyl phosphoguanosine dinucleotide
CTP (dCTP, ddCTP)	cytidine 5'-triphosphate (deoxy-, dideoxy-)
dbEST	database of expressed sequence tags
DNA	deoxyribonucleic acid
dNTP	2'-deoxyribonucleoside 5'-triphosphate
DTT	dithiothreitol
EDTA	ethylenediamine tetra-acetic acid
EMBL	European Molecular Biology Laboratory
EST	expressed sequence tag
FAT10	HLA-F associated transcript 10
FISH	fluorescence 'in situ' hybridisation
FPC	fingerprinting contig
GABA	gamma-aminobutyric acid
GDP	guanine diphosphate
GPCR	G-protein coupled receptor
G-protein	GTP-binding protein
GPX	glutathione peroxidase
GTP (dGTP, ddGTP)	guanine 5'-triphosphate (deoxy-, dideoxy-)
HEK	human embryonic kidney
HFE	Hereditary haemochromatosis locus
HGMP	Human Genome Mapping Resource Centre
HGP	Human Genome Project
HLA	human leukocyte antigen
HMM	Hidden Markov Model
HUGO	Human Genome Organisation
Ig	Immunoglobulin
IHGSC	International Human Genome Sequencing Consortium
Kb	kilobase pairs
l	litre
LB	Luria-Bertani
LD	linkage disequilibrium
LINE	long interspersed nuclear element
M	molar
Mb	megabase pairs
µg	microgram
µl	microlitre

μM	micromolar
min(s)	minute(s)
mg	milligram
MHC	major histocompatibility complex
ml	millilitre
mm	millimetre
mM	millimolar
MOE	major olfactory epithelium
NCBI	National Centre for Biotechnology Information
ng	nanogram
OB	olfactory bulb
OMIM	On-line Mendelian Inheritance in Man
OR	Olfactory receptor
OSN	Olfactory sensory neuron
PAC	P1-derived artificial chromosome
PCR	polymerase chain reaction
PFAM	protein family database
RFP	ret finger protein
RNA (mRNA, rRNA, tRNA)	ribonucleic acid (messenger-, ribosomal-, transfer-)
RNase A	ribonuclease A
rpm	revolutions per minute
RP	ribosomal protein
RT-PCR	reverse transcription polymerase chain reaction
SDS	sodium dodecyl sulphate
sec(s)	second(s)
SINE	short interspersed nuclear element
snoRNA	small nucleolar RNA
SNP	single nucleotide polymorphism
STS	sequence tagged site
TEMED	N, N, N', N'-tetramethylethylenediamine
TM	Transmembrane domain
TrEMBL	Translated EMBL database
Tris	tris(hydroxylmethyl)aminomethane
U	unit
UTR	untranslated region
uv	ultraviolet
V	volt
v/v	volume/volume
VNO	vomeronasal organ
VR	pheromone receptor
W	watt
w/v	weight/volume
WGS	whole genome shotgun
YAC	yeast artificial chromosome
ZNF	zinc finger protein

Publications

Publications describing parts of this thesis:

Ehlers, A., Beck, S., Forbes, S., Trowsdale, J., Uchanska-Ziegler, B., Volz, A., **Younger, R.**, and Ziegler, A. (2000) MHC-linked olfactory receptor loci exhibit polymorphism and define extended HLA/OR haplotypes *Genome Res* 10(12):1968-78.

Ziegler, A., Ehlers, A., Forbes, S., Trowsdale, J., Uchanska-Ziegler, B., Volz,A., **Younger, R.**, and Beck, S. (2000) *Polymorphic olfactory receptor genes and HLA loci constitute extended haplotypes.* In Kasahara, M. (ed) *Major Histocompatibility Complex: Evolution structure, and function*

Ziegler, A., Ehlers, A., Forbes, S., Trowsdale, J., Uchanska-Ziegler, B., Volz,A., **Younger, R.**, and Beck, S. (2000) Polymorphisms in olfactory receptor genes: a cautionary note. *Hum Immunol* 61(12):1281-4

Younger, R., Amadou, C., Bethel, G., Ehlers, A., Fischer Lindahl, K., Forbes, S., Horton, R., Milne, S., Mungall, A., Trowsdale, J., Volz, A., Ziegler, A., and Beck, S. (2001) Characterisation of clustered MHC-linked Olfactory Receptor Genes in Human and Mouse *Genome Res* 11(4):519-30

Ziegler, A., Beck, S., Ehlers, A., **Younger, R.** and Volz, A. (2002) *Testicular transcripts of HLA-linked olfactory receptor genes exhibit unorthodox features* In Hansen, J.A. and Dupont, B. (ed) *HLA 2002 Immunobiology of the Human MHC*

Chapter 1

Introduction

1.1. The olfactory system

In the battle for survival, how an organism responds to various stimuli, such as food sources, potential mates and potential predators, is highly important in determining the evolutionary fitness of a particular individual. Sight, smell and hearing are all vital in an individual's struggle to survive, mate and successfully pass their selfish genes into the next generation. Of these senses the sense of smell, or olfaction, as the perception of odours is more properly known, is considered to be the most primitive. Lower organisms, for example *Caenorhabditis elegans*, are heavily reliant on their olfactory system to perceive the world and make the kind of choices likely to enhance their reproductive fitness. Even in mammals where hearing and sight have evolved, a functioning olfactory system enhances an individual's chances of survival; dogs and mice rely on odours to locate food, recognize territory, identify kin, and find a receptive mate.

The vertebrate olfactory system is well adapted for the detection and recognition of small odorant molecules. It is considered to be able to detect as many as 10,000 different odorants, and it has the ability to detect both subtle differences between chemical stereoisomers and the vastly different chemical structures that odorant molecules may possess. It is also able to detect some odorants at very low airborne concentrations of less than several parts per trillion (Snyder *et al.*, 1988). This ability to deal with very different olfactory inputs in vertebrates is located in one major neural structure, the major olfactory epithelium (MOE) which is a specialized neuroepithelium located in the posterior cavity of the nose. A secondary olfactory organ, the vomeronasal organ (VNO), is located further towards the front of the nose: this organ is considered to be only involved in the detection of pheromones, chemical signals conveying social and sexual information.

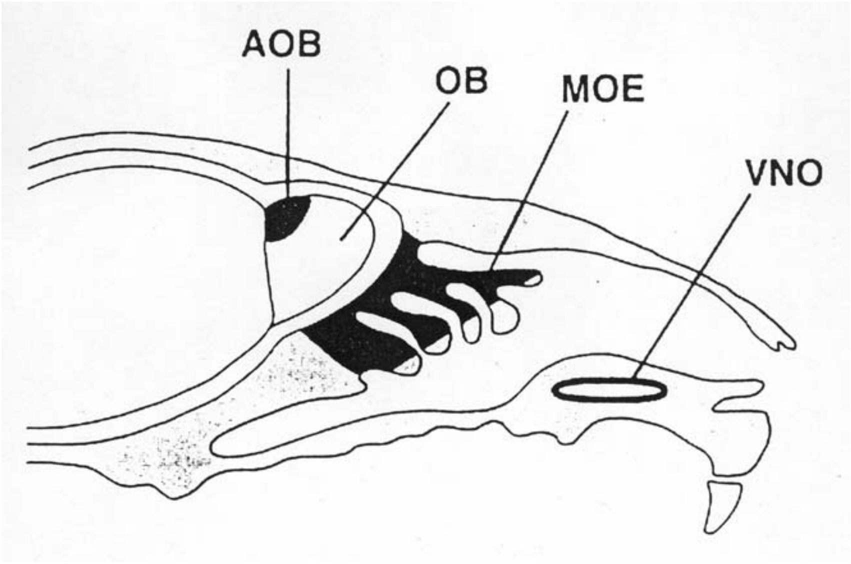


Figure 1.1: Anatomy of the olfactory system. A sagittal view of the rat head, showing the location of the vomeronasal organ (VNO) and the major olfactory epithelium (MOE) within the rat nose. The two structures project their axons to different structures within the rat brain: the sensory neurons of the MOE projecting axons to the olfactory bulb (OB), whilst sensory neurons of the VNO project to the accessory olfactory bulb (AOB).

Figure from Buck and Axel (1991).

1.2. The major olfactory epithelium: anatomical organization and signal transduction

The sensory epithelium of the MOE is made up of three cell types: olfactory sensory neurons (OSNs), their precursor basal cells, and sustentacular cells (which have glia-like, supportive functions within the epithelium). Olfactory sensory neurons are bipolar cells that project a single unmyelinated axon to the olfactory bulb located in the anterior part of the skull. A single dendrite is projected to the epithelial surface where it terminates in a dendritic knob and projects specialized cilia into the nasal lumen. These cilia, which lie in the thin layer of mucus that cover the tissue, provide a large surface area where olfactory receptors are exposed to a stream of warmed, moistened and possibly concentrated odorants. Throughout an individual's lifetime, OSNs appear to continuously undergo cell death and be regenerated from the precursor basal-cell population (Levy *et al.*, 1991).

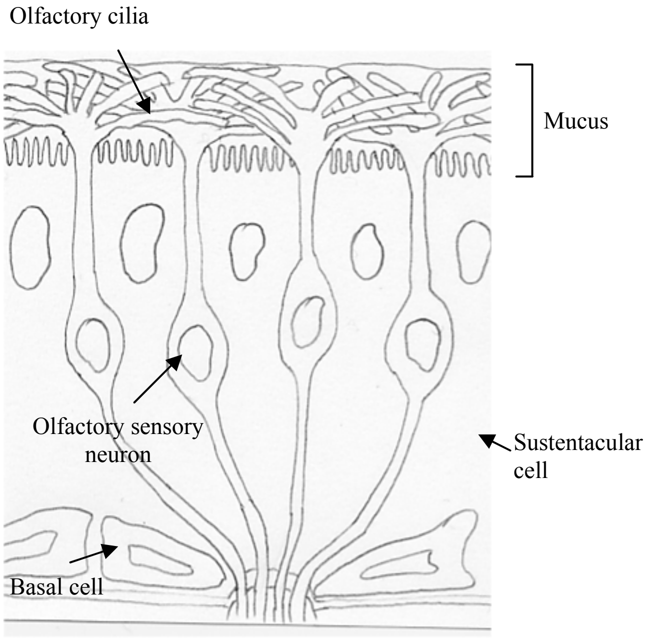


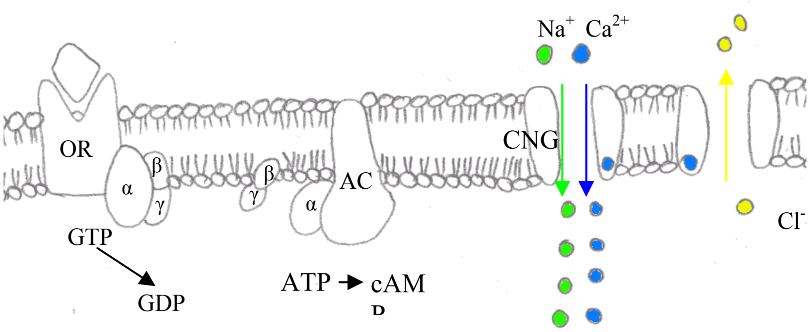
Figure 1.2: The structure of the major olfactory epithelium: odorant molecules enter the nose and dissolve in the layer of mucus covering the OSN cilia, and odorant binding occurs. OSNs are regenerated from basal precursor cells, whilst sustentacular cells have a supportive function within the structure.

Adapted from Firestein (2001).

Within the MOE, the process of olfaction signal transduction is believed to occur as follows. Firstly, odorants enter the nasal cavity where they dissolve in the mucus that covers the luminal surface of the olfactory epithelium. Once dissolved, the odorant molecules are able to bind to specific olfactory receptors (ORs) that are located on the cilia of the dendrites of olfactory sensory neurons. When the odorant molecule ligand has bound to the OR, these cell-surface transmembrane receptors undergo a conformational change, which facilitates interaction with a guanine triphosphate (GTP) binding protein (G protein). G proteins are composed of three subunits (α, β, and γ). The interaction between an olfactory receptor protein and a G protein (likely to be the olfactory specific G_{olf}) results in the exchange of GDP for GTP on the α subunit, and a subsequent dissociation of this subunit from the β and γ subunits. The activated α subunit interacts with a second messenger enzyme, which in the case of G_{olf} is adenylyl cyclase. The cyclase converts adenine triphosphate (ATP) into cyclic adenine monophosphate (cAMP).

G-protein-dependent elevation of cyclic AMP (cAMP) leads to cAMP binding to the intracellular face of an ion channel, allowing the channel to conduct cations such as Na^+ and Ca^{2+}. Membrane depolarization triggers an action potential which is projected to the olfactory bulb. In addition to this pathway, identified using genetically altered mice lacking the various components of this transduction cascade (G_{olf}, adenylyl cyclase and cyclic nucleotide-gated channel (Brunet *et al.*, 1996, Belluscio *et al.*, 1998, Wong *et al.*, 2000)), OSNs also have an amplification mechanism. The calcium ions trigger the opening of an ion channel that allows negatively charged chlorine ions to leave the cell, increasing the net positive charge across the membrane (Kleene and Gesteland, 1991). As well as increasing the membrane depolarization, the calcium ions also act in the negative feedback pathway, acting (probably with calmodulin) on the ion channel to decrease its sensitivity to cAMP (Kurahashi and Menini, 1997). Additional down regulators of the signal include a RGS (regulator of G-protein signalling) protein which acts on adenylyl cyclase to decrease its activity (Sinnarajah *et al.*, 2001), and a kinase that phosphorylates activated receptors sending them into a desensitized state (Dawson *et al.*, 1993, Schleicher *et al.*, 1993).

Information about these action potentials generated in the OSNs is transmitted to the olfactory bulb. In the olfactory bulb, axons of olfactory sensory neurons form synapses with the dendrites of secondary neurons (mitral cells and tufted cells) and interneurons (periglomerular cells) within structures known as glomeruli. Mitral and tufted cells integrate the input from the neurons and local inhibitory currents before relaying this information to the olfactory cortex and other central brain areas via the lateral olfactory tract.



Adapted from Firestein (2001).

Figure 1.3: Sensory transduction in olfactory sensory neurons (OSNs). The OR ligand binds triggering disassociation of the GPCR and the α-subunit activates an adenylyl cyclase (AC). The cyclase converts ATP to cAMP which acts to open the cyclic nucleotide gated (CNG) ion channel. Calcium ions flowing into the cell trigger the opening of an ion channel allowing chlorine ions to leave the cell, increasing the net positive charge across the membrane.

1.3. The molecular basis of the olfactory system.

In spite of all the knowledge about the olfactory system, the molecular basis of olfaction remained elusive for many years. It was, then, a pioneering piece of work from Buck and Axel (1991) that identified a large family of genes considered to encode olfactory receptor genes. Buck and Axel's work on olfactory receptor genes was based on three assumptions. Firstly, the likely involvement of G proteins in the olfactory systems suggested ORs were likely to share structural and sequence similarities with the G-protein coupled receptor (GPCR) superfamily. The superfamily of GPCRs have highly conserved regions within their 7 transmembrane domain structure, allowing degenerate primers to be designed based on these conserved regions. Secondly, Buck and Axel hypothesized that ORs were members of a multigene family of considerable size and diversity because a large number of chemicals with differing structures can be detected. Thirdly, it was suggested that ORs were likely to be expressed only in olfactory sensory neurons.

Buck and Axel, therefore, used degenerate oligonucleotides known to anneal to conserved regions of G-protein coupled receptor genes together with cDNA sequences from rat olfactory epithelium to identify various putative olfactory receptor genes. Looking for a multigene family, they focused on a pair of primers that appeared to amplify a large number of genes. Further experiments on this multigene family revealed that it fitted the third specification: Northern blot analysis revealed that members of the multigene family were expressed exclusively in the olfactory epithelium of rat.

The 18 olfactory receptor proteins isolated in these experiments were aligned, and an additional observation seemed to support the idea that these were olfactory receptors. This observation was that, although the proteins were found to share structural and sequence similarities with the G-protein coupled receptor superfamily of neurotransmitters and hormone receptors, including seven hydrophobic stretches considered to represent seven transmembrane domains, in contrast to other GPCRs where maximum sequence conservation is seen within the transmembrane domains, these novel genes showed striking divergence within the third, fourth and fifth transmembrane domains. This divergence, within the transmembrane domains considered to be involved in ligand binding in other 7 transmembrane proteins (Kobilka, 1992), was consistent with the theory that maximal diversity between olfactory receptors would be expected to be found in the ligand binding regions, allowing the multigene family to interact with a large number of odorant molecules.

A further similarity with some other 7 transmembrane protein genes was that these putative olfactory receptor genes were found to have no introns within their coding sequences. This lack of introns allowed an estimate of the number of OR genes within the genome to be made. Southern blots were probed with OR genes that did not cross-hybridize, and it was found that each gene hybridized to 1 to 17 bands. In all a total of 70 bands were detected which led to an idea that the rat genome contained about 100 to 200 OR genes. Later estimates produced a much higher number of suggested OR genes: in 1992, for example, Buck suggested the size of the multigene family could be estimated as between 500 to 1000 genes (Buck, 1992). These type of estimates suggested the OR gene family would be the largest family in the mammalian genome, with 0.8-1.6% of the 60,000 mammalian genes likely to be OR genes.

The initial identification of OR genes, therefore, revealed they were genes consisting of small (generally less than 1 Kb), intronless open reading frames. Consensus amino acid motifs could be highlighted: LHTPMY in intracellular loop 1, MAYDRYVAIC at the end of the third predicted transmembrane domain, SY at the end of transmembrane domain 5, FSTCSSH in transmembrane domain 6, and PMLNPF in transmembrane 7. Hypervariable regions, which could correspond to ligand binding sites were highlighted in transmembrane domains 3, 4 and 5. From initial experiments it was also known that a given OR gene would cross-hybridize to a small set of related OR genes. Sets of related OR genes suggested the OR gene repertoire could be subdivided into subfamilies (Lancet *et al.*, 1993).

The idea that transmembrane domains 3, 4 and 5 represent the hypervariable binding region of the protein has been supported by data from three dimensional models of other GPCR proteins (Pilpel and Lancet, 1999) which suggest that these three α-helical barrels arrange themselves into a pocket, one third of the way into the membrane. Studies on adrenergic receptors (Kobilka *et al.*, 1988) and the binding site of retinal in rhodopsin (Palczewski *et al.*, 2000) also suggest that these three domains are likely to be the binding site of these proteins.

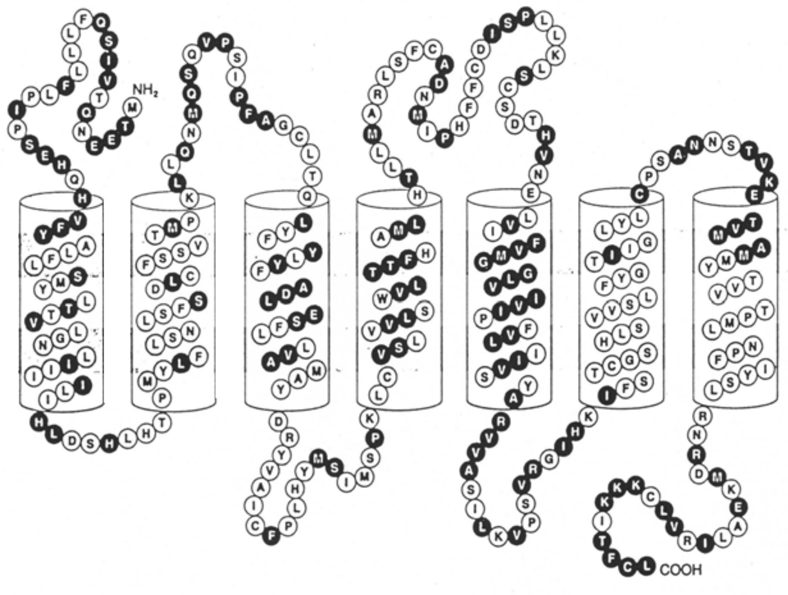
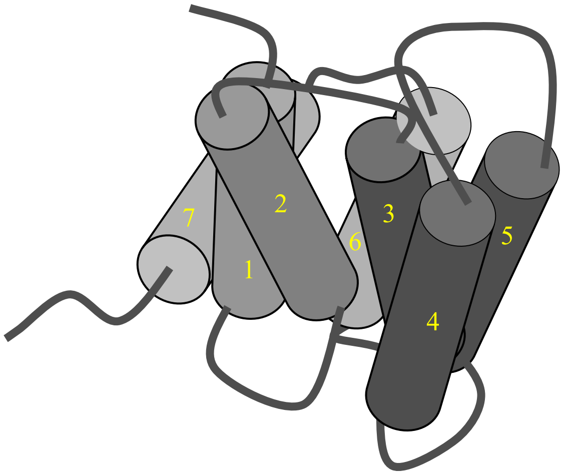


Figure 1.4a: Schematic diagram of the protein structure of an olfactory receptor: amino acids are represented as balls and the 7 predicted transmembrane regions are enclosed within cylinders. Conserved amino acids are shown in white, whilst highly divergent positions are black. This diagram shows the hypervariability of transmembrane regions, 3, 4 and 5, compared to the other transmembrane regions.

Figure from Buck and Axel (1991).

Figure 1.4b: Proposed three dimensional structure of an olfactory receptor protein, with degree of conservation within the transmembrane regions plotted according to conservation of amino acids shown in figure 1.3a. (Dark grey barrels suggest highly divergent region of protein, whilst lighter grey barrels suggest a greater degree of conservation.)

Adapted from Firestein (2001).



1.4. Cloning of further olfactory receptor genes in a variety of species

Following Buck and Axel’s work, OR genes were identified in a number of species. Generally, degenerate PCR primers have been used to amplify related genes from genomic DNA, or (more rarely, since cDNA from olfactory mucosa is difficult to obtain) from olfactory epithelium cDNA libraries. In rat, for example, further OR genes were identified through degenerate PCR of genomic DNA (Levy *et al.*, 1991, Raming *et al.*, 1993, Strotmann *et al.*, 1994b, Drutel *et al.*, 1995, Thomas *et al.*, 1996) and through degenerate PCR of testis cDNA (Vanderhaeghen *et al.*, 1997b). Rat OR genes were also identified using mRNA from the axon terminals of olfactory sensory neurons in the olfactory bulb (Singer *et al.*, 1998). In addition to OR genes identified in rat, OR genes were identified in various fish species: catfish (*Ictalarus punctatus*) (Ngai *et al.*, 1993b), zebrafish (*Danio rerio*) (Barth *et al.*, 1996, Byrd *et al.*, 1996, Weth *et al.*, 1996, Barth *et al.*, 1997), lamprey (*Lampetra fluviatilis*) (Berghard and Dryer, 1998), and mudpuppy (*Necturus maculosus*) (Zhou *et al.*, 1997). Chicken OR genes were cloned to provide an insight into OR genes in birds (Leibovici *et al.*, 1996), whilst information from the frog (*Xenopus laevis*) genome (Freitag *et al.*, 1995) suggests the OR gene repertoire in amphibians consists of some OR genes more closely related to those found in fish species (named Class I ORs by Freitag *et al.*, 1995), and some more closely related to those found in mammals (named class II ORs by the same group Freitag *et al.*, 1995). Information about the OR gene repertoire of other mammalian species has also been accumulated: there is a large amount of data about the OR gene family in mouse (Nef *et al.*, 1992, Ressler *et al.*, 1993, Asai *et al.*, 1996, Mombaerts *et al.*, 1996a, Sullivan *et al.*, 1996, Kubick *et al.*, 1997, Vanderhaeghen *et al.*, 1997b, Qasba and Reed, 1998), and a small amount of data about dog (Parmentier *et al.*, 1992, Issel-Tarver and Rine, 1996, Issel-Tarver and Rine, 1997, Vanderhaeghen *et al.*, 1997b), pig (Matarazzo *et al.*, 1998, Velten *et al.*, 1998) and various primates (Sharon *et al.*, 1999).

In humans, OR genes were initially found in a testis cDNA library: Parmentier *et al.* (1992) discovered that the orphan 7 TM receptor genes they had previously identified were orthologous to the rat OR genes identified by Buck and Axel. These genes were all likely to be functional, with complete, non-pseudogenic open reading frames, but later work on human OR genes revealed that a large number of the repertoire were pseudogenes (Selbie *et al.*, 1992) Some of these pseudogenes were identified from a cDNA library made from olfactory tissue (Crowe *et al.*, 1996), suggesting some olfactory sensory neurons may express non-functional OR genes.

1.5. The genomic organization of olfactory receptor genes.

Initial work on the genomic organization of olfactory receptor genes focused on a cluster of OR genes located on chromosome 17p (Ben-Arie *et al.*, 1994), where the first OR gene had been mapped (Schurmans *et al.*, 1993). 16 human OR genes located in a 350 Kb cluster were cloned (Ben-Arie *et al.*, 1994), but when one of the cosmids predicted to contain 6 of these OR genes was sequenced (Glusman *et al.*, 1996), only 3 OR genes mapping to this cosmid were found. This highlighted the potential problem involved in sequencing OR genes from cloned PCR products: it may be that many of the sequences could have been artificially generated by recombination between highly related nucleotide sequences (Meyerhans *et al.*, 1990, Mombaerts, 1999). Further work on this region of the genome lead to the characterization of a 412 Kb contiguous sequence, containing 17 OR genes, 10 with intact open reading frames, and 7 of which were predicted to be pseudogenes (Glusman *et al.*, 2000).

3 further human olfactory receptor genes were generated through sequencing a 36 Kb cosmid located on chromosome 19 (Trask *et al.*, 1998). Further work on human OR genes involved mapping genes to various locations within the human genome. Fan *et al.* (1995) mapped 3 OR genes to the Major Histocompatibility Complex class I region on chromosome 6p21, whilst human OR genes orthologous to 4 dog genes mapped to chromosomes 7q35, 11q11 and 19p13 (Issel-Tarver and Rine, 1997, Carver *et al.*, 1998). OR genes found within testis cDNA libraries were located to chromosomes 11p22, 17q21 and 19p13-19p31 (Vanderhaeghen *et al.*, 1997a), whilst the 36 Kb sequence block containing 3 OR genes and located on chromosome 19 was used as probe to discover further OR gene regions. Using this sequence, Trask *et al.* (1998) found the

block was duplicated on chromosomes 3q and 15q, and it also appears that additional copies of the block are present in different individuals (a human genome can contain between 7 to 11 copies of this block).

An attempt to try and locate all OR gene-containing sites within the human genome was made by Rouquier *et al.* (1998): a pool of OR gene fragments from degenerate PCR on genomic DNA was used as a probe in a series of fluorescence *in situ* hybridization (FISH) experiments on metaphase chromosomes. These experiments revealed that all chromosomes, except chromosome 20 and chromosome Y, contain OR genes, and suggested OR genes were present in between 25-53 locations within the human genome.

1.6. The expression of olfactory receptor genes within the olfactory system

The initial discovery of OR genes within the olfactory epithelium suggested that these genes were expressed in olfactory sensory neurons. Subsequent work confirmed that many OR genes are expressed in mature OSNs: *in situ* hybridization experiments suggested that specific OR genes were expressed in a subset of OSNs that have a characteristic bilateral symmetry (Nef *et al.*, 1992, Strotmann *et al.*, 1992, Raming *et al.*, 1993). Further work on expression of OR genes in the olfactory epithelium in mouse (Ressler *et al.*, 1993) and the rat (Vassar *et al.*, 1993) revealed that OSNs expressing a certain specific OR gene were located in one of several nonoverlapping zones. Within the specific zone, a random expression pattern with bilateral symmetry can be observed. Initial work suggested there were 3 zones of expression within the olfactory epithelium, but later studies suggested 4 expression zones in the mouse olfactory epithelium (Sullivan *et al.*, 1995, Sullivan *et al.*, 1996). Expression at the protein level, using an antiserum against a rat OR, also provides evidence for the zonal pattern of expression of OR genes (Koshimoto *et al.*, 1994). No physiological basis for this zonal organization has been uncovered, however, it is clear that the zones can be discerned from the earliest embryonic stages at which OR genes are expressed (Sullivan *et al.*, 1995, Menco and Jackson, 1997). A large number of OR probes have all provided support for the zonal expression hypothesis (Ressler *et al.*, 1993, Vassar *et al.*, 1993, Strotmann *et al.*, 1994a, Strotmann *et al.*, 1994b, Sullivan *et al.*, 1996, Kubick *et al.*, 1997). However, one group of genes (OR37 group) does not show this zonal expression pattern; instead, they appear to be expressed in a 'patch' on the tips of some turbinates and not in the septum of the rat, mouse and guinea pig olfactory epitheliums (Strotmann *et al.*, 1992, Strotmann *et al.*, 1994a, Strotmann *et al.*, 1994b, Strotmann *et al.*, 1995). Zonal expression of OR genes in other species has not been demonstrated so clearly: in zebrafish, no observations regarding zones were made at either adult (Barth *et al.*, 1996, Byrd *et al.*, 1996) or embryonic (Vogt *et al.*, 1997) stages. Early observations in catfish also suggested that the expression of a particular OR gene was constant across the olfactory epithelium (Ngai *et al.*, 1993a), however, a later quantitative analysis suggested the existence of 3 or 4 expression domains (Weth *et al.*, 1996). In *Xenopus laevis* the two different classes of ORs appear to be expressed in different regions (Freitag *et al.*, 1995): this could represent a functional adaptation with one region specialized for detecting waterborne odors and the other airborne odors, or it could represent a zonal separation of expression similar to that observed in rat and mouse.

The expression of OR mRNA in the axon terminals of OSNs in the olfactory bulb means it is also possible to consider the distribution of neurons expressing a particular OR gene within the olfactory bulb. Probing the olfactory bulb with rat OR probes revealed that neurons expressing an OR gene appear to converge to a few discrete sites within the bulb (Ressler *et al.*, 1994, Vassar *et al.*, 1994). The location of these converging axons is bilaterally symmetric, and locations appear to be very similar between members of the same species. These experiments, however, did not provide single axon resolution: it may have been that individual axons expressing the same OR projected to different locations. These neurons would not have been detected because expression of the OR can only be detected where there are a number of axons converging. A single axon approach, using a mutant OR gene together with the marker taulacZ, however, suggested all axons expressing this particular allele, converged to fixed, symmetric locations within the olfactory bulb (Mombaerts, 1996, Mombaerts *et al.*, 1996a, Mombaerts *et al.*, 1996b). This targeting of OSNs expressing a specific OR gene to specific glomeruli appears to be partially controlled by the OR gene. Evidence for this type of mechanism is available, for example, in one experiment a mutation introduced into an OR gene meant compared to the non-mutant form, a different glomeruli was targeted (Mombaerts *et al.*, 1996a). Further work on this mouse locus confirmed that OR genes are one component in this guidance process, although it appears there are other important components (Wang *et al.*, 1998).

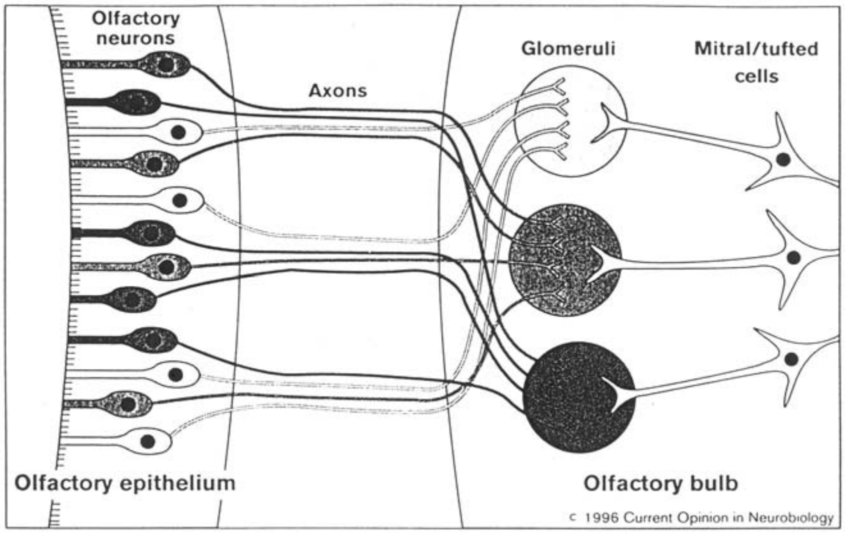


Figure 1.5: OR targeting within the olfactory bulb. OSNs expressing the same OR gene target the same glomerulus within the olfactory bulb, for example, all neurons expressing the 'white' OR within the MOE project to the 'white' glomerulus within the OB.

From Mombaerts (1996).

The fixed location of glomeruli relating to one specific OR suggests odor perception is encoded by a combination of activated glomeruli. This is supported by physiological observations in mammals (Hildebrand and Shepherd, 1997) and by observations from the zebrafish olfactory bulb where 80 defined glomeruli (Baier and Korsching, 1994) show stereotyped patterns of glomerular activation in response to various odorant molecules (Friedrich and Korsching, 1997).

OR genes, therefore, are expressed in olfactory sensory neurons, and this expression can be detected in the major olfactory epithelium, and also, at a much lower level in the olfactory bulb. Both the olfactory epithelium and the olfactory bulb show distinct patterns of expression; within the olfactory epithelium, OSNs expressing a specific OR gene are located in a constant expression zone, whilst in the olfactory bulb, OSNs expressing the same OR gene converge to specific glomeruli. The expression of OR genes, however, does not appear to be restricted to the olfactory system: there is evidence that olfactory receptor genes are expressed in other tissues.

1.7. The expression of olfactory receptor genes outside the olfactory system

Expression of OR genes within the testis has been observed since the initial identification of olfactory receptor genes; the human counterparts to the Buck and Axel's rat OR genes were cloned from a testis cDNA library (Parmentier *et al.*, 1992). In addition to the expression of OR genes in human testis, OR genes have been reported to be expressed in the testis of dog, mouse and rat (Vanderhaeghen *et al.*, 1993, Walensky *et al.*, 1995, Vanderhaeghen *et al.*, 1997a, Vanderhaeghen *et al.*, 1997b, Walensky *et al.*, 1998). OR genes also appear to be expressed in spermatids and spermatozoa: RNA has been found in postmeiotic round spermatids (Parmentier *et al.*, 1992, Walensky *et al.*, 1998), whilst antisera detected OR protein in late round and elongated spermatids, and on the tail midpiece of mature spermatozoa (Vanderhaeghen *et al.*, 1993, Walensky *et al.*, 1995). Expression of OR genes within the testis may result from the aberrant regulation of transcription. Alternatively, there may be some functional reason for the expression of these genes within the testis, for example, they may regulate sperm maturation, sperm motility or sperm attraction to oocytes. The signal transduction machinery associated with ORs is present in the cells of rat testis (G-protein receptor kinase 3, β-arrestin (Walensky *et al.*, 1995), adenylyl cyclase III (Gautier-Courteille *et al.*, 1998)) suggesting that ORs may have some biological role within the testis.

Expression of OR genes has also been reported in other non-olfactory tissues, for example, the heart of rat (Drutel *et al.*, 1995), the notochord of chick (Nef and Nef, 1997), the brainstem of rat (Raming *et al.*, 1998) and mouse (Conzelmann *et al.*, 2000), and the erythroid cells of human and mouse (Feingold *et al.*, 1999). Expressed sequence tag (EST) libraries have also suggested that OR genes are expressed in other non-olfactory organs, such as the lungs, heart, liver, placenta, colon, ovaries, as well as sperm cells. This could be due to genomic contamination of EST libraries, or it could be due to erroneous transcriptional processes. Alternatively, it may be that these OR genes have been recruited for a non-olfactory purpose in different tissues. One non-olfactory purpose was proposed by Dreyer (1998) who suggested that olfactory receptors may be important in providing part of a molecular addressing code required during development. This model advances the idea that cells assemble organisms through a so-called ‘area code’ that functions like the country, area, regional and local portions of the telephone dialing system. The olfactory receptor genes can be seen as the last digits in this cell ‘area code’, an idea that fits in with the predicted large size of the olfactory receptor gene family and the apparent widespread tissue distribution of these genes.

1.8. The regulation of olfactory receptor genes

The size of OR gene repertoire has been estimated to be about 500-1000 genes in rat (Buck, 1992), 400 in dog (Parmentier *et al.*, 1992), 600 in mouse (Qasba and Reed, 1998), and up to 100 in catfish and zebrafish (Ngai *et al.*, 1993b, Barth *et al.*, 1996, Weth *et al.*, 1996). This large number of OR genes means it is difficult to determine the number of OR genes expressed per olfactory sensory neuron, but *in situ* hybridization of the olfactory epithelium with OR probes has not shown any colocalization of OR genes, so it is generally considered that an olfactory sensory neuron expresses a single or a very small number of olfactory receptor genes (Mombaerts, 1999). Expression of OR genes within olfactory sensory neurons, however, is known to be restricted so only one allele of a given OR gene is expressed (Chess *et al.*, 1994). In a cross between mice containing 2 different allelic forms of 2 OR genes, it was found that the offspring expressed either the maternal or paternal allele in OSNs. Regulation of OR genes, therefore, must be tightly controlled in order to express 1 allele of 1 (or very few) OR genes, but little is known about the mechanisms controlling this regulation.

1.9. Ligands of olfactory receptor genes

In spite of the large amounts of data about OR genes that has been accumulated, the vast majority of OR genes remain orphan receptors: very little is known about ligand-receptor interactions. This scarcity of knowledge is due to the fact that ORs are very difficult to express on the surface of heterologous cells (Mombaerts, 1999). Whilst several 7 transmembrane proteins, such as the opsins and the β-2-adrenergic receptor have been successfully expressed on cell surfaces, OR proteins tend not to be incorporated in the plasma membrane, instead they are retained, nonfunctional, in intracellular compartments (McClintock *et al.*, 1997).

Expression of 1 rat OR protein in insect Sf9 cells has been achieved using a baculovirus vector; some odorants caused transient increases in intracellular second messengers but others produced no observable response (Raming *et al.*, 1993). Zebrafish OR genes were also successfully expressed in human embryonic kidney (HEK293) cells by fusing a N-terminal membrane import sequence of a guinea pig serotonin receptor and an artificial c-myc epitope tag to the ORs (Wellerdieck *et al.*, 1997). These cells showed transient increases in intracellular calcium (detected with the calcium-sensitive dye fura-2) when exposed to fish food, but there was no response to amino acids, bile acids or progesterone (physiologically relevant odorants for fish).

A more successful approach to detecting ligand-receptor interactions has been to use OSNs in order to achieve expression. Zhao *et al.* (1998), for example, used an adenoviral vector to drive expression of 1 OR and the green fluorescent protein (GFP) in a number of rat olfactory sensory neurons. Transepithelial potentials across the olfactory epithelium (electroolfactograms, EOGs) were measured in order to measure the response of the OR to 74 ligands, and it was found that octyl aldehyde raised the amplitude of the EOG above control levels. There was also a smaller response to 3 other aldehydes, heptyl aldehyde, nonyl aldehyde and decyl aldehyde. These responses were also measured in single-cell studies of the infected neurons.

Only one human OR gene has been functionally characterized: this gene, OR17-40, was expressed in HEK 293 cells and *Xenopus laevis* oocytes, and was found to be the ligand for helional and another structurally related molecule, heliotroplyacetone (Wetzel *et al.*, 1999).

1.10. The accessory olfactory system

The identification of genes coding for receptors in the olfactory epithelium was followed by the identification of more genes involved in olfaction. However, these genes were found in a functionally and anatomically distinct tissue, the vomeronasal organ (VNO). The VNO is located in a more anterior position within the nose than the MOE, and although it too is involved in olfaction, it sends information via a separate pathway of neuronal projections. In rats and mice, the VNO is typically involved in stimulating innate behavioural responses upon the detection of pheromones, chemical signals conveying social and sexual information (Keverne, 1999). For example, experiments in rodents have linked the VNO to mating and aggressive behaviours in males (Clancy *et al.*, 1984, Meredith, 1986) and sexual development and onset of oestrus in females (Johns *et al.*, 1978, Reynolds and Keverne, 1979, Lomas and Keverne, 1982). In the VNO, work has revealed three families of receptors; Dulac and Axel (1995) identified seven novel seven transmembrane domain receptor sequences (named V1Rs), whilst a number of groups (Herrada and Dulac, 1997, Matsunami and Buck, 1997, Ryba and Tirindelli, 1997) identified another subfamily of pheromone receptors (V2Rs), and Pantages and Dulac (2000) found a third potential group (V3Rs). These three families are not closely related to each other, nor do they appear to be related to the family of OR genes. The V1R genes are predicted to be part of a family containing 30-100 genes, and expression of the family appears to be restricted to the apical cell layer within the VNO, whilst the V3R family (with approximately 100 members) are expressed in a distinct subset of VNO neurons. The V2R genes meanwhile, correspond to a family of about 100 genes that are related to the metabotropic glutamate receptors, the extra-cellular calcium-sensing receptor, and, more distantly to the GABA-B receptor. These similarities suggest the V2R proteins may function in a different way to the smaller OR, V1R and V3R proteins, with binding of ligands occurring in the large extracellular domain rather than somewhere within the 7 transmembrane structure. V2R genes appear to be expressed in a small fraction of vomeronasal sensory neurons (VSNs) in the basal region of the VNO. As with olfactory sensory neurons, expression of V1R, V2R, and V3R genes seems to be tightly controlled with only one or a few receptors expressed per VSN. The organization of expression of different classes of VRs in different regions within the VNO is also reflected within the accessory olfactory bulb (AOB) organisation: apical VSNs project to the rostral AOB whilst basal VSNs project to the caudal AOB.

The role of the accessory olfactory system in the human olfactory system is considered to be minimal. In contrast to rodents, there is only one generally accepted pheromone-like effect in humans, namely the synchronization of the menstrual cycle between women who live in close proximity to each other (McClintock, 1971), and whilst there is evidence that a VNO-like organ develops in the embryo, it is thought that the organ becomes vestigial in adults (Keverne, 1999). In addition, in rodents, one-third of V1R repertoire are predicted to be pseudogenes (Del Punta *et al.*, 2000), and two-thirds of the V2R gene family are predicted to be pseudogenes (Herrada and Dulac, 1997, Matsunami and Buck, 1997). Although functional examples of V1Rs (Rodriguez *et al.*, 2000) and V3Rs (Pantages and Dulac, 2000) have been found in the human genome, it has therefore been suggested that the majority of the VR genes found in the human genome are pseudogenes (Giorgi *et al.*, 2000).

1.11. The Major Histocompatibility Complex (MHC)

The Major Histocompatibility Complex (MHC) is an extended complex of gene clusters located on human chromosome 6p21.3 that is considered to contain a remarkably high number of genes with immunological function. The region is of major biomedical importance, owing to its role in tissue transplantation rejections and its role in influencing susceptibility to a variety of autoimmune diseases, such as insulin dependent diabetes mellitus, multiple sclerosis, systemic lupus erythmatosus and rheumatoid arthritis (Thomson, 1995). A complete map of the human MHC was published by the MHC sequencing consortium in 1999: at this time 224 gene loci (128 of which were predicted to be expressed) had been identified (Figure 1.6).

The MHC has traditionally been divided into three areas: class I, class II, and class III. The most telomeric region is the class I region which contains, amongst other genes, the classical class I genes, HLA-A, HLA-B, and HLA-C. These classical class I genes encode the heavy (α) chain, which together with β2 microglobulin (β2m, locus found on chromosome 15) make up MHC class I molecules. These MHC class I molecules are expressed by virtually all nucleated cells within the human body, and they play a role in the immune system through their ability to bind peptides (antigens) and present these peptides to T-cells. Peptides loaded into class I molecules are generally derived from endogeneous source by the proteasome, of which LMP2 and LMP7 (both found within the MHC, (Driscoll *et al.*, 1993)) are subunits. These peptides are then transported in the endoplasmic reticulum by the TAP1/TAP2 molecule (also encoded by genes within the MHC, Ortmann *et al.*, 1994) where they are bound to the MHC class I molecules which proceed to the cell surface via the Golgi apparatus. At the cell surface the MHC class I molecule-peptide complex is accessible to T cells possessing the CD8 surface antigen. If a T cell is expressing a T cell receptor (TCR) that recognizes the specific peptide an immune response may be initiated: this generally results in the lysis of the cell expressing the peptide.

Class II genes include the LMP2, LMP7, TAP1 and TAP2 genes which have functions as detailed above. The MHC class II region, however, takes its name from the classical and non-classical MHC class II genes. The non-classical class II genes include HLA-DOA and HLA-DOB

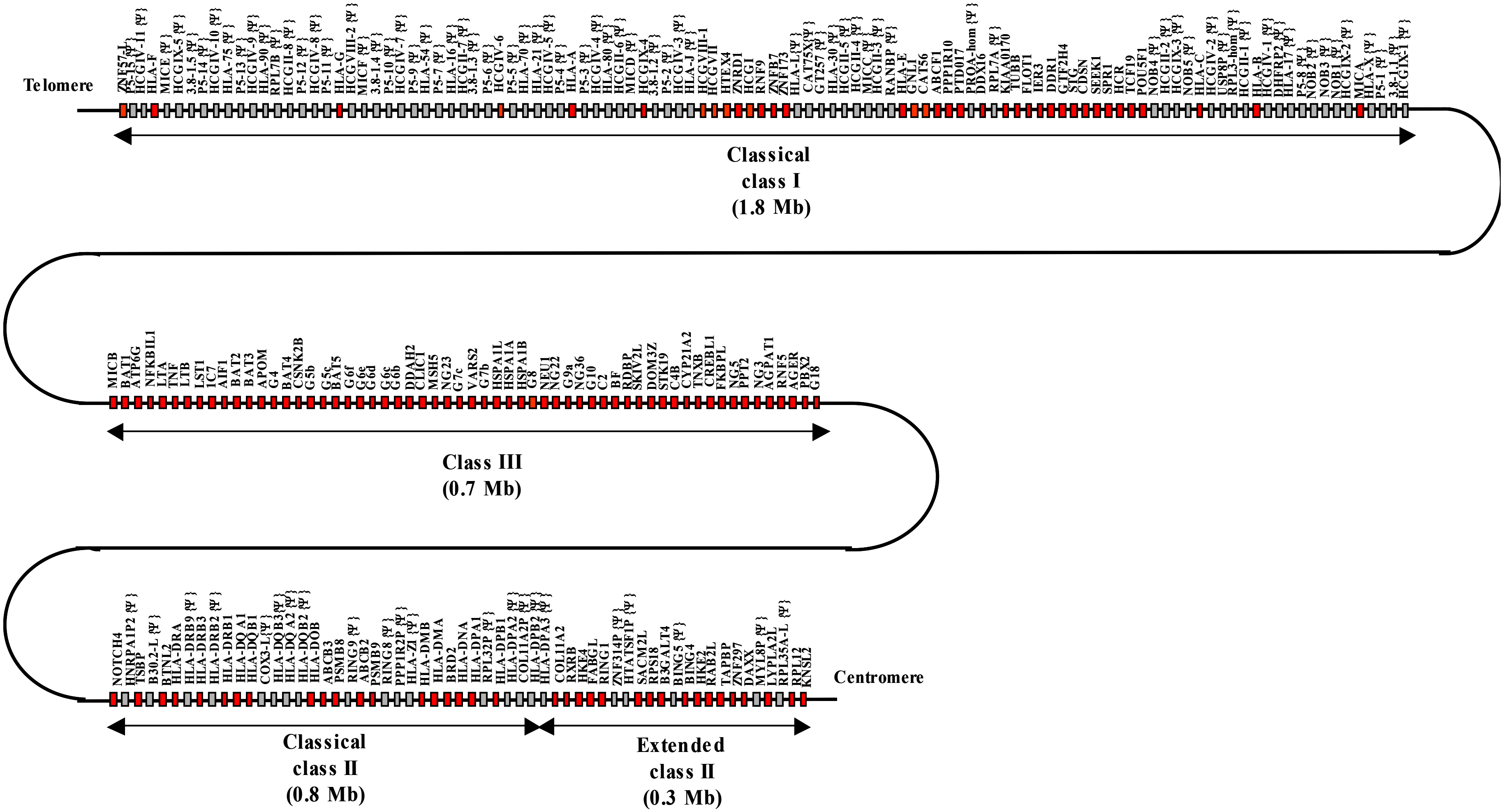


Figure 1.6: Gene map of the human MHC. (Adapted from The MHC Sequencing Consortium, 1999.) Expressed genes are highlighted in red, whilst pseudogenes are indicated by the Ψ symbol. The gene names are taken from the 1999 map: several have been changed subsequently by HUGO.

which produce protein products that combine to form the molecule HLA-DO. This molecule acts to suppress the ability HLA-DM has to facilitate peptide loading (Weber *et al.*, 1996). The classical class II genes (HLA-DP, HLA-DQ, HLA-DR) either encode proteins with 2 α chains (HLA-DPA, HLA-DQA, HLA-DRA) or proteins with 2 β chains (HLA-DPB, HLA-DQB, HLA-DRB): the corresponding α and β chains combine to form class II MHC molecules. MHC class II molecules differ from MHC class I molecules in that the groove of the peptide binding region (PBR) is open-ended, allowing longer peptides (generally 12-24 amino acids) to be bound. In contrast to MHC class I molecules, class II molecules are expressed only on a limited number of cell types, known as antigen presenting cells (APCs). These include B lymphocytes, macrophages, dendritic cells and activated T lymphocytes. Class II molecules bind predominantly to peptides from extracellular sources. Prior to peptide binding, these molecules are assembled in the endoplasmic reticulum (ER) with a membrane-bound chaperone protein (known as the MHC class II-associated invariant chain or γ chain) acting to stabilize the complex. This γ chain is degraded by proteases in the trans-Golgi reticulum, with the exception of a small fragment (the class II associated invariant chain peptide, CLIP) which is buried in the PBR. CLIP is only displaced just prior to binding: this reaction is catalysed by the product of 2 other non-classical class II genes, HLA-DM (Denzin and Cresswell, 1995, Sloan *et al.*, 1995). After binding the MHC class II molecule-peptide complex is transported to the cell surface where, if it is recognized by a specific T lymphocyte carrying the CD4 surface antigen, an adaptive immune response is triggered.

The MHC class III region is located between class I and class II. It is the most gene dense area of the MHC, containing a number of genes involved in the complement cascade of natural immunity, the interferon-inducible heat shock proteins, and a number of genes involved in the inflammation response.

Between species, there is a conservation of some of the basic genes within all 3 regions suggesting there is an evolutionary advantage in conserving the MHC as an unit. This MHC ‘unit’ can be observed in species evolving after the divergence of the jawless vertebrates (for example, hagfish or lamprey). The three regions of the human MHC appear to have been subject to different evolutionary mechanisms: whilst MHC class II and class III genes often appear to have direct orthologs, the MHC class I appears to have expanded and contracted in different species. This is discussed in more detail in the introduction to Chapter 5.

Work on the three ‘classical’ regions of the MHC revealed that sequence conservation and possibly linkage disequilibrium extended further than the three classical regions; immediate flanking regions were termed the extended class I and extended class II regions of the MHC (Stephens *et al.*, 1999). The extended class I region of the MHC is where Fan *et al.* (1995) located a cluster of OR genes; these genes were also found to be conserved in the syntenic region in mice and rats (Szpirer *et al.*, 1997). The conservation of this OR cluster within the MHC across three species suggested there may be some functional reason for this conservation.

1.12. MHC genetic diversity and reproductive selection

The most prominent hypothesis as to why an OR cluster appears to be conserved in its position next to the MHC across several species is the idea that there is some connection between MHC genetic diversity, reproductive selection, and olfaction. Products of genes within the MHC have critical roles during immune recognition, binding self and foreign peptide fragments for presentation to T lymphocytes (Babbitt *et al.*, 1985, Bjorkman *et al.*, 1987). The ‘red queen’ hypothesis, namely that pathogens will constantly evolve to defeat immune defences, means that the MHC has been forced to generate a huge amount of genetic diversity to be able to deal with the constant pathogenic onslaught. This genetic diversity can be generated through preferential selection of heterozygotes over homozygotes (heterozygote-advantage or overdominance) or through preferential selection of relatively rare genotypes (negative frequency-dependent selection), or, more likely, through some combination of the two (Doherty and Zinkernagel, 1975, Hughes and Nei, 1988, Potts and Wakeland, 1990, Takahata and Nei, 1990, Slade and McCallum, 1992). In addition to these selection pressures generated through pathogen-driven selection, however, genetic diversity can also be generated through disassortative mating preferences (Potts and Wakeland, 1993). Offspring produced through disassortative mating choices are likely to be genetically fitter than other offspring since they will have a reduced inbreeding load and they will have an increased resistance to genetic disease arising from their increased MHC heterozygosity.

Within mice, experiments have shown that both inbred male mice and and outbred male mice preferentially mate with females with dissimilar MHC genotypes (Yamazaki *et al.*, 1976), and further experiments revealed that the detection of this dissimilarity is primarily through the mice distinguishing genotypic identity through smelling the odour of conspecific’s urine (Yamaguchi *et al.*, 1981). Selective mating preference appears to be acquired through comparison of a potential mate’s odour with remembered familial odours, since mice raised by foster parents will mate with other mice with an odour dissimilar to that of the family nest, rather than with mice dissimilar to themselves (Yamazaki *et al.*, 1988).

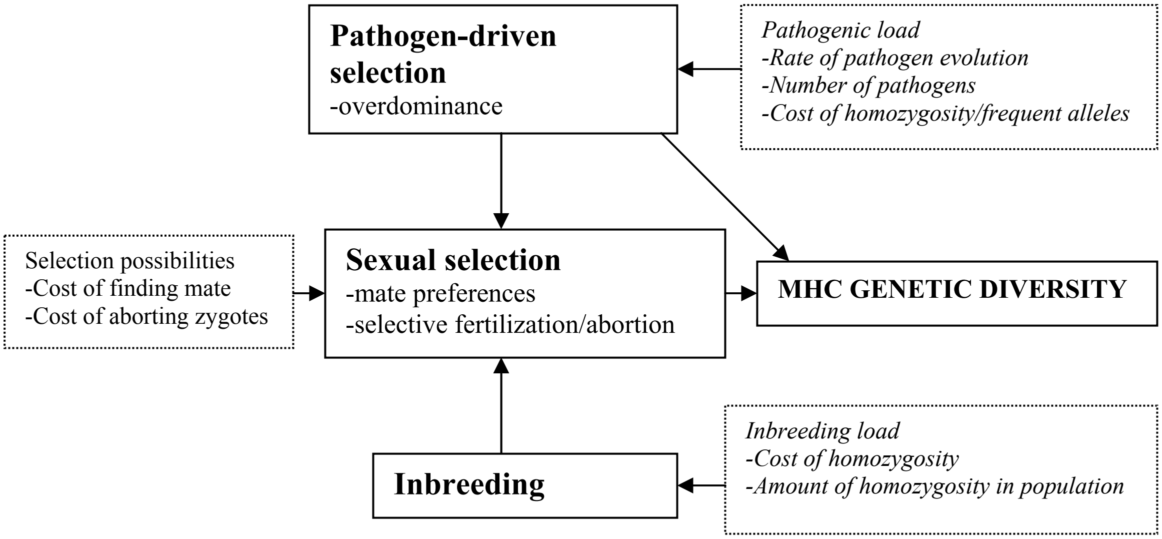

Pathogen-driven selection
-overdominance
Pathogenic load
-Rate of pathogen evolution
-Number of pathogens
-Cost of homozygosity/frequent alleles
Selection possibilities
-Cost of finding mate
-Cost of aborting zygotes
Sexual selection
-mate preferences
-selective fertilization/abortion
MHC GENETIC DIVERSITY
Inbreeding
Inbreeding load
-Cost of homozygosity
-Amount of homozygosity in population

Figure 1.7: Factors contributing to MHC genetic diversity. Selective forces are shown in bold, with some of the variables shown in italics. Adapted from Potts and Wakeland (1993).

Another form of sexual selection has been observed in mice. In female mice, maintenance of pregnancy is dependent on the odour type to which the female mouse is exposed in early pregnancy. Female mice carrying an embryo with a similar MHC genotype will selectively abort the embryo if exposed to a dissimilar odour type (Yamazaki *et al.*, 1983). The exact contribution of this mechanism to the generation of genetic diversity is debatable, since there would appear to be a high cost involved in aborting a foetus, but nonetheless this is a clearly observed behaviour in laboratory mice.

Whether these mechanisms of sexual selection exist within other vertebrates in something that requires further work. Within humans, there is little conclusive evidence suggesting a link between MHC genotype and sexual selection (Tiwari and Terasaki, 1985), and there is contradictory evidence over whether similar MHC couplings produce a higher rate of abortions or have a lower fecundability than dissimilar couplings (Christiansen *et al.*, 1989, Ober *et al.*, 1992, Pennesi *et al.*, 1998).

The idea that there is a connection between sexual selection for dissimilar MHC genotypes and olfaction, therefore, makes the conservation the physical linkage between a cluster of OR genes and the MHC very interesting. It is possible that a combined MHC-olfactory haplotype is inherited. This combined haplotype could confer on an organism an increased sensitivity to detect (through olfaction) MHC alleles within their haplotype.

1.13. Aims of this thesis

The aims of this thesis were, firstly, to identify all OR genes located telomeric of the human MHC classical class I region. The target region was delineated by the HLA-F locus and the HFE locus: this region is known as the ‘extended MHC class I’ region in human. In addition to analyzing the sequence in the human extended class I region, another aim of this thesis was to map, sequence and identify the syntenic OR cluster region in mouse. Here, the target region was demarcated by the Gabbr1 locus and the breakpoint in synteny that occurs telomeric of this locus on mouse chromosome 17 (well centromeric of the HFE locus on human chromsome 6).

After identification of these genes in the human and mouse target regions, the syntenic regions were compared to identify orthologs and other conserved segments of sequence that could have functional roles. Analysis of the regulatory regions and expression profiles of the human MHC-linked ORs, using both experimental and *in silico* approaches was also performed. The MHC-linked OR genes are located just telomeric of the most highly polymorphic region of the human genome, the MHC, so polymorphism within these genes was also considered. A phylogenetic analysis comparing the MHC-linked ORs against other ORs within the human genome was also considered essential to reveal whether the MHC-linked ORs can be considered unique within the human genome. In addition to the MHC-linked OR genes, a number of pheromone receptor (VR) pseudogenes were also found to be located telomeric of the MHC. These pseudogenes, which once formed part of the mammalian olfactory system, were analysed in order to investigate further the relationship between the pheromone and olfactory receptor genes.

These various strands of the thesis were expected to provide an unique insight into the function and evolution of the MHC-linked olfactory receptor genes. By combining data from these pillars of the thesis, a number of key issues were considered including:

- The relationship between the MHC, the ‘extended MHC class I’ region and the ORs. (Chapter 3 and Chapter 4)
- The syntenic relationships of the MHC-linked ORs in mouse and human (Chapter 5)
- The regulation of the MHC-linked ORs (Chapter 6)
- The polymorphisms of (some) MHC-linked ORs (Chapter 7)
- The relationship between MHC-linked ORs and other ORs in the human genome (Chapter 8)
- The relationship between the MHC-linked VR and OR genes (Chapter 9).

Chapter 2

Materials and Methods

2.1. Materials

The majority of chemical reagents were bought from Sigma unless stated in the text; similarly restriction enzymes were largely bought from New England Biolabs unless stated elsewhere. A number of kits from various companies were used: these are specified in the text. PCR was generally performed using Amplitaq and the supplied PCR buffer from Perkin Elmer unless otherwise stated. All primers used in this thesis are listed in Appendix 1.

In the list of materials, where materials have an ambiguous name the numbers in brackets refer to the section for which the material is required.

2.1.1. Solutions

[2.2] Solution I (GTE/GET): 50 mM Glucose, 25 mM Tris, 1 mM EDTA (2.3 ml 20% glucose, 5.0 ml 0.1M EDTA, 1.3 ml 1M Tris (pH 7.4), 42 ml water)

[2.2] Solution II (NaOH/SDS): 0.2 M NaOH, 1% SDS (2.5 ml 4M NaOH, 2.5 ml 20% SDS, 45 ml water.)

[2.2] Solution III: 3.5 M KOAc (pH 5.5) (147.21g potassium acetate, 57.5 ml glacial acetic acid, water to 500 ml)

[2.2] Solution IV: TE (10:0.1) with RNase (10 µg/ml)

Sodium acetate/EDTA solution: 49.218g sodium acetate, 2 ml 0.1 M EDTA, water to 200 ml.

[2.3] Hybridisation solution: 50% formamide, 2 x SSC (pH 7.0), 10% dextran sulphate, 1% Tween 20

SSCTM: 4 x SSC (pH 7.0), 0.05% Tween 20, 5% low-fat dried milk

Buffered phenol: 1 ml phenol, 200 µl 1M Tris-hydrogen chloride (Shaken and placed on ice for 5 minutes, spun, top layer removed and discarded, 200 µl TE (10:0.1) added, mix shaken and spun. Kept on ice until required.)

[2.5.1] *EtOH/NaOAC mix:* 100 ml sodium acetate, 1600 ml ethanol 96%, 300 ml water.

RNase A: 200 mg RNAse A, 100 µl Tris (pH 7.4), 150 µl sodium chloride, water to 10 ml.

1 mM Tris-HCl (pH 8.5): 0.0606 g Tris, made up to 500 ml with water, and adjusted to pH 8.5 with hydrochloric acid.

[2.7] *Denaturing solution:* 25 ml of 10 M sodium hydroxide, 150 ml of 5 M sodium chloride, made up to 500 ml with water.

[2.7] 10 x Neutralization solution: 250 ml of 1 M Tris-chloride, 150 ml of 5 M sodium chloride, made up to 500 ml with water.

IPTG (0.1 M): 0.238 g in 10 ml of water. Sterilized by filtration, then stored at -20°C.

Xgal: 20 mg/ml dissolved in dimethyl sulfoxide (DMSO).

Trypsin-EDTA: 6.4 g sodium chloride, 0.16 g potassium chloride, 0.92 g sodium phosphate, 0.16 g potassium dihydrogen phosphate, 0.16 g sodium-EDTA, 0.5 g trypsin, made up to 1 litre with water. Stored at -20°C.

PBS: 10 g sodium chloride, 0.25 g potassium chloride, 1.44g sodium hydrogen phosphate (dibasic), 0.25 g potassium dihydrogen phosphate, made up to 1 litre with water and made to pH 7.4 with sodium hydroxide. Stored at 4°C.

Cresol Red Solution: 84.5 mg Cresol red sodium salts (Aldrich) in 100 ml TE (10:0.1). Stored at -20°C.

34.6% Sucrose: 121.1 g sucrose dissolved in 350 ml water. Stored at -20°C

10 mM dNTP: 1000 µl of each 100 mM dNTP (Amersham Pharmacia), 6 ml water. Stored at -20°C.

20 x SSC: 175.3 g sodium chloride, 88.2 g sodium citrate, made up to 1 litre with water.

2 x SSC, 0.1% SDS: 100 ml 20 x SSC, 1 g SDS, made up to 1 litre with water.

0.2 x SSC, 0.1% SDS: 10 ml 20 x SSC, 1 g SDS, made up to 1 litre with water.

2.1.2. Media

Circlegrow: from QBiogene

*2 x TY:*15 mg/ml bacto-tryptone, 10 mg/ml bacto-yeast extract, 5 mg/ml NaCl (pH 7.4), water up to 1 litre.

SOB: 20 g tryptone, 5 g yeast extract, 10 ml 1M sodium chloride, 0.5 g potassium chloride, water added up to 1 litre.

SOC: SOB + 200 µl 20% glucose.

TYE plates: 8 g tryptone, 5 g yeast extract, 8 g sodium chloride, 12 g agar, water up to 1 litre.

TYE/Amp plates: 2 ml of 25 mg/ml ampicillin was added to 1 ml TYE autoclaved solution which was allowed to cool to 48°C before addition. (Final concentration of 50 µg/ml).

H-Top: 8 g Bacto Agar, 10 g Bacto Tryptone, 8 g sodium chloride, made up to 1 litre with water.

LB medium: 10 g Bacto-tryptone, 5 g Bacto-yeast extract, 10 g sodium chloride, made up to 1 litre with water, and pH 7.5 with sodium hydroxide.

LB/ampicillin/IPTG/X-Gal plates: 10 g Bacto-tryptone, 5 g Bacto-yeast extract, 10 g sodium chloride, 15 g Bacto-Agar, made up to 1 litre with water. 2 µl ampicillin (25 mg/ml) was added after the agar had cooled to 48°C. 1 ml of 0.1 mM IPTG and 2 ml Xgal (40 ug/ml) were also added to the cooled media, before plating out.

2.1.3. Loading dyes

Blue dextran formamide dye: 9.8 ml deionised formamide, 200 µl of 0.5 M EDTA, 0.01 g of blue dextran

[2.4] Loading dye: 5 mg bromophenol blue, 0.5 g Ficoll 400, 0.5 ml 10 x TBE, 4.5 ml water.

Ficoll dye: 0.5 g Ficoll 400, 100 µl 50 x TAE, 10 mg bromophenol blue, 4.9 ml water.

Sequencer Loading dye: 25 mM EDTA (pH 8.0), 50 mg/ml Blue dextran, deionised formamide (5:1 formamide: EDTA/Blue dextran.)

2.1.4. Buffers

TE (10:1): 2 ml Tris (pH 7.4), 2 ml 0.1 M EDTA, water to 200 ml.

TE (10:0.1): 2 ml Tris (pH 7.4), 200 µl 0.1 M EDTA, water to 200 ml.

10 x TBE buffer, pH 8.8: 162 g Tris base, 27.5 g boric acid, 9.2 g EDTA (disodium salt – Na_2EDTA), made up to 1 litre with water.

[2.4] Loading buffer: 10 µl 10 x TBE, 20 µl loading dye, 50 µl water.

Mung bean nuclease buffer: 100 µl 3 M sodium acetate, 250 µl 2 M sodium chloride, 10 µl 1 M zinc chloride, 140 µl water, 500 µl mung bean nuclease (NEB), 500 µl glycerol.

50 x TAE buffer: 121 g Tris base, 12.2 g EDTA (disodium salt – Na_2EDTA), 28.55 ml acetic acid, made up to 500 ml with water.

10 x Rxn buffer: 4.5 ml 1M Tris-HCl (pH 8.8), 5 ml cresol red solution, 0.15 ml water, 0.35 ml 1 M magnesium chloride (BDH), 0.1454 g ammonium sulphate (GIBCO).

[2.12] Dilution buffer: 100 ml water, 50 ml TE (10:0.1), 0.8125 ml Cresol Red solution, 50 µl 4 M sodium hydroxide. Stored at -20°C.

2.1.5. Markers

[2.4] λ Hind III marker: 8 µl λ DNA- Hind III digest (NEB), 60 µl TBE buffer, 252 µl water. (Mixture was incubated at 65°C for 5 minutes before being rapidly chilled on ice.

80 µl of loading dye was then added.)

[2.4] pBR322 marker: 4 µl pBR322 DNA-BstNI digest (NEB), 60 µl 10 x TBE buffer, 256 µl water, 80 µl loading dye.

1 Kb ladder marker: 5 µl 1 Kb ladder mixed with 1 µl 50 x TAE, 10 µl Ficoll dye, and 34 µl water

[2.6.2] λBst11071 ladder marker: Prepare λBst11071 digest: 2 µl λ DNA (1 µg), 5 µl 10x buffer, 1 µl Bst11071 enzyme. Incubate at 37°C for 2 hours.

[2.6.2] Ladder marker: 2 µl 1 Kb ladder DNA (Gibco BRL, 1 µg/µl), 12 µl 50 x TAE, 50 µl λBst11071 digest, 100 µl Ficoll dye, 436 µl water.

2.1.6 Sequencing gel

Denaturing acrylamide gel (6%): 30 g urea was placed in a 250 ml beaker, together with 9 ml acrylamide/bisacrylamide solution, 4 ml 10x TBE and 37 ml water. The urea was dissolved by heating (60°C) and stirring, and the solution was made up to 60 ml with water. The solution was then placed in a dessicator for 4 minutes and just before pouring the gel 138 µl of 25% ammonium persulphate and 138 µl TEMED were added. The gel mix was then syringed between the glass gel plates whilst tapping the glass gently to get rid of the bubbles. The gel was left to set for at least 90 minutes prior to use.

2.1.7. Antibiotics

Ampicillin (stock solution of 25 mg/ml in water stored at -20°C): add to final concentration of 50 µg/ml

Chloramphenicol (34 mg/ml in 100% ethanol stored at -20°C): add to final concentration of 30 µg/ml

Kanamycin (25 mg/ml in water stored at -20°C): add to final concentration of 50 µg/ml

Tetracycline (12.5 mg/ml dissolved in 50% ethanol, stored in the dark at -20°C): add to final concentration of 15 μg/ml (used with media without magnesium salts).

500x Penicillin/Streptomycin (Boehringer Mannheim): Penicillin (50000 IU/ml), Streptomycin (50 mg/ml). Stored at -20°C.

2.1.8. Web addresses and other *in silico* resources

A number of the methods described require the use of various websites. In the text this is shown by the program name being in italics: websites used in the course of this thesis are listed in Appendix 3. A number of UNIX-based programs were also used in the course of this thesis: these are referenced in the text or listed in Appendix 4. A number of the programs listed in Appendix 4 were written by me in the Perl programming language.

2.2. Fluorescent Fingerprinting

With the advent of large-scale genomic sequencing, it was necessary to develop a method for producing large, deep contig maps allowing an optimal tiling path to be chosen. One method used to achieve this type of contig map is restriction digest fluorescent fingerprinting: this is based on using fluorescently tagged dideoxy ATPs to label the HindIII termini created in a double digest of the clone with HindIII and Sau3AI and using restriction patterns to assess the degree of overlap (Gregory *et al.*, 1997). The mouse MHC-linked OR contig was fingerprinted using this fluoresecent fingerprinting method.

2.2.1. Preparation of DNA for fluorescent fingerprinting (based on Birnboim and Doly, (1979))

1. 500 µl of 2 x TY (containing the appropriate antibiotic, kanamycin for PACs and chloramphenicol for BACS) was dispensed into four 96-well 1 ml Beckman boxes.
2. From the colonies that had been grown, clones were picked into these wells using sterilised toothpicks, and boxes were placed in a shaker at 300 rpm, 37°C for 12-18 hours.
3. 250 µl of each of the cultures were transferred to a round-bottomed 96-well Corning plate using a 50- to 250-multichannel pipette (Finnpipette) and cells were pelleted by centrifugation at 2500 rpm, 20°C for 4 minutes.
4. The supernatant was discarded from the cell pellets and pellets were resuspended in 25 µl of solution I and 25 µl of solution II. (Mixed by tapping the plates gently and then leaving the plates for 5 minutes.)
5. The supernatant was discarded once again and 25 µl of chilled solution III was added before leaving the plate for 5 minutes.
6. Well contents from the plates was transferred to filter plates (Millipore) taped to a round-bottomed Corning plate containing 100 µl isopropanol.
7. These 2 plates were then spun at 2500 rpm, 20°C for 2 minutes to ensure all liquid had been transferred from the filter plate to the lower plate; the filter plate was then discarded.
8. After separation from the filter plate, the lower (Corning) plate was left at room temperature for 30 minutes before being centrifuged at 3200 rpm, 20°C for 10 minutes.
9. The supernatant was discarded from the plate and the DNA pellet was briefly dried before being washed with 100 µl of 70% ethanol, centrifuging at 3200 rpm 20°C for 10 minutes.

10. Finally, the supernatant was discarded and the DNA pellet was dried before being resuspended in 5 µl of solution IV.

2.2.2. Generation of fluorescent marker

1. To make the fluorescent marker, 1.5 µg of lambda DNA (500 ng/µl) was placed in a 1.5 ml eppendorf tube with 7.5 U of BsaJI (2.5 U/µl), 16 U of TaqFS (8 U/µl), 2 µl of ROX ddC (5.08 µM), 5 µl of NEB2 buffer, and 35 µl of TE (10:0.1) (pH 7.4).
2. This tube was then incubated at 60°C for an hour before adding 50 µl of sodium acetate (0.3M) and 200 µl of 96% ethanol, mixing the solutions by vortexing briefly, and then leaving the tube (in the dark) at room temperature for 15 minutes.
3. The tube was then left at -20°C for a further 20 minutes before centrifugation at 14000 rpm for 20 minutes.
4. The supernatant was discarded, and the pellet was air-dried before 100 µl of 70% ethanol was added and the tube was again spun at 14000 rpm for 20 minutes.
5. After spinning, the supernatant was discarded and, after drying, the pellet was resuspended in 60 µl of TE (10:0.1) (pH 7.4) and 60 µl of blue dextran formamide dye.

2.2.3.The restriction digest reaction

1. A mix for the fluorescent labelling reaction was made up. The constitution of this was calculated by considering the amount of chemicals required per fluorescent labelling reaction, namely: (i) 2.8 U of HindIII (20 U/µl), (ii) 3 U of Sau3AI (30 U/µl), (iii) 3.7 U

of ThermoSequenase (32 U/µl), (iv) 0.14 µl of fluorescent ddA (10 µM; either HEX, TET or NED), (v) 0.8 µl of TE (10:0.1), (vi) 1 x NEB2 buffer.

2. Using a Hamilton repeat dispenser, 20 µl of this reaction mix was added to wells containing 5 µl of DNA suspended in TE (10:0.1), and plates were incubated at 37°C for 1 hour.
3. 7 µl of sodium acetate (0.3 M) and 40 µl of 96% ethanol was added to each sample before spinning at 3200 rpm, 20°C for 20 minutes.
4. The supernatant was then discarded, and the pellet was dried before adding 100 µl of 70% ethanol and centrifuging at 3200 rpm, 20°C for 10 minutes.
5. The supernatant was discarded and the pellet was dried and resuspended in 5 µl of TE (10:0.1).
6. Prior to loading on a 377 ABI Automated DNA sequencer, 2 µl of the fluorescent marker was added to each well and samples were denatured by placing them for 10 minutes on a block heated to 80°C.

2.2.4. Data analysis

Data from the ABI sequencer is processed by the 'Image' program (Production Software Group, The Sanger Centre, unpublished, based on Sulston *et al.* (1988). 'Image' automatically tracks lanes on a gel, distinguishes bands and normalizes bands against a marker lane, although the first two steps required checking and normally some form of manual alteration. 'FPC' (Soderlund *et al.*, 1997, Soderlund *et al.*, 2000) takes as input a set of clones and the bands, corresponding to restriction fragments, called by the 'Image' package for each clone. 'FPC' calculates the probability of two clones overlapping based on the similarity of their fragments, using an algorithm based on the probability of coincidence score. Contigs consisting of two or more clones can be built according to these scores, and ‘FPC’ then builds a consensus band (CB) map for each contig. The CB map is used to assign coordinates to the clones based on their position with regard to this consensus map, providing a detailed picture of how much clones overlap within the contig. Some degree of manual editing is generally required: clones can be removed and added to contigs, clone coordinates can be refined, and contigs can be merged, split or deleted.

2.3. Fluorescent *in situ* hybridisation (FISH) mapping

FISH analysis of the mouse clone, bM573K1, was performed in order to confirm that the clone could be mapped to mouse chromosome 17. This work was carried out in collaboration with the Human Cytogenetics Laboratory, ICRF, London (Denise Sheer and Jill Williamson). The protocol was adapted from Senger *et al.* (1993): it is based on labelling the probe using nick-translation.

1. 1 µg of the clone was labelled with biotin-14-dATP using the Bionick labelling system (Gibco-BRL). The reaction was incubated for 1 hour at 15°C.
2. In order to consider the efficacy of the reaction, 100ng of the nick-labelled DNA was loaded onto a 1% agarose gel, and fragments were compared against a 100 bp marker (Gibco-BRL). Fragments of 500 bp are an ideal size for FISH but a range of 200-1000 bp was considered acceptable.
3. As fragments of the required size were produced, the labelling reaction was halted by adding 5 µl of EDTA.
4. 100 ng of the labelling reaction was mixed with 2 µg human Cot-1 DNA (Gibco-BRL), 2 µl 3M sodium acetate (pH 5.6), and 50 µl 100% ethanol. This mixture was placed on dry ice for 1 hour to precipitate the probe.

5. The mixture was centrifuged at 14000 rpm, 4°C for 15 minutes and the pellet was dried in a speed vacuum, before resuspension in 11 μl of hybridisation solution.
6. The resuspended DNA was denatured at 85°C for 5 minutes. Incubation at 37°C for 30 minutes followed in order to compete out the repetitive elements.
7. The metaphase slides were denatured in 100 μl of 70% formamide, 2 x SSC (pH 7.0) at 75°C for 2.5 minutes.
8. The slides were dehydrated through exposure to 3 concentrations of ethanol: slides were shaken for 3.5 minutes with 70% ethanol, 95% ethanol and finally 100% ethanol.
9. The slides were air dried to evaporate the remaining ethanol.
10. The hybridisation mix was placed onto the denatured slides and there were covered by 20 x 20 mm cover slips. Rubber cement was used to seal the cover slips and the slides were incubated in a moist chamber at 37°C for 24 hours.
11. After hybridisation, slides were washed 3 times at 42°C in 50% formamide, 2 x SSC (pH 7.0).
12. Further washing was performed 3 times at 42°C in 2 x SSC (pH 7.0). Finally, the slide was briefly washed in SSCTM solution.
13. For detection of the biotinylated probes, the slide was incubated with a 1:500 dilution of avidin-FITC (Vector laboratories) in SSCTM solution. Incubation was at 37°C for 30 minutes.
14. The slide was counterstained with 0.06 μg/ml of DAPI (4',6'-diamidino-2-phenylindole hydrochloride) in Citifluor AF1 (an antifadent contained in a glycerol PBS which maintains fluorescence).
15. Slides were analysed using a Zeiss Axioscop fluorescence microscope equipped with a CCD camera. Separate images of the DAPI staining of the chromosomes and the biotinylated probes were merged using Smartcapture software (Vysis, UK).

2.4. The production of PUC and M13 shotgun libraries (Bankier *et al.*, 1987)

Using the large-scale maps produced by FPC fingerprinting methods, the minimal number of clones that would cover the region were selected for shotgun sequencing. Shotgun sequencing involves generating a random set of fragments which are then assembled so overlapping fragments of sequence provide the complete sequence across the clone. Typically, to declare a clone sequence accurate, there is a requirement for each section of the clone to be covered by 6-8 separate fragments; this redundancy allows for the resolution of sequencing errors. As part of the large-scale sequencing effort, human PACs and BACs (from the libraries RPCI1 and RPCI-11.1 constructed at the Roswell Park Cancer Institute by the group of Pieter de Jong) were subcloned at the Sanger Centre using the method described below; mouse BACs (from the Research Genetics CITB-CJ7-B library) and PACs (from library RPCI-21, also constructed at Roswell Park Cancer Institute) were supplied by Claire Amadou (Amadou *et al.*, 1999) and subcloned by me.

2.4.1. Isolation of PAC/BAC DNA using the caesium chloride procedure

1. A 500 ml sterile plugged flask containing 200 ml 2x TY and the appropriate selective agent (0.6 ml kanamycin (25 mg/ml) for PACs, 0.1 ml chloramphenicol (25 mg/ml) for BACs) was inoculated with a single colony, and incubated (with shaking at 300 rpm) at 37°C for 18-24 hours.
2. The cells and medium were then transferred to a 250 ml bottle and spun at 6000 rpm for 5 minutes (Sorvall centrifuge).
3. The supernatant was discarded, and the pellet was dried briefly by draining the bottle onto tissue, before being fully resuspended (by drawing the mixture up and down the pipette) in 50 ml GET solution.

4. 50 ml of NaOH/SDS solution was added to the bottle, which was inverted gently 3-4 times, and then left for 5 minutes at room temperature.
5. 50 ml of potassium acetate solution was added, and the bottle was inverted 10-12 times before being placed in an ice-bath for 20 minutes, and then centrifuged at 12000 rpm, 4°C for 20 minutes.
6. After centrifugation, the supernatant was filtered through a piece of sterile cheese cloth into a new 250 ml bottle.
7. 90 ml isopropanol was added to the supernatant, which was then left at room temperature for 5 minutes before spinning at 9000 rpm, 4°C for 15 minutes.
8. The supernatant was discarded and the pellet was washed with 25 ml 70% ethanol, centrifuging at 9000 rpm for 5 minutes.
9. The pellet was dried by removing all traces of the supernatant and then placing the bottle in a vacuum-drier for 2-3 hours.
10. To resuspend the pellet 2.9 ml of TE (10:1) was added and the bottle was swirled gently.
11. 3 g caesium chloride were added to a 50 ml falcon tube, into which the DNA solution was then transferred.
12. After the caesium chloride had dissolved, 290 ml ethidium bromide was added to the tube and the solution was spun at 3000 rpm, 20°C for 12 minutes before being transferred into a TL100 tube.
13. The TL100 tube was heat-sealed and centrifuged at 70000 rpm, 20°C for 16-24 hours.
14. From the TL100 tube, the lower (supercoiled) band of DNA was removed using a 1 ml syringe with a 20G needle. The amount of DNA recovered (about 200-300 μl) was placed in a 1.5 ml eppendorf tube.
15. 0.3 ml of water and 0.5 ml of isobutanol were added to the 1.5 ml eppendorf tube.

16. After mixing, this produced one immiscible (pink) layer which was discarded, leaving 0.4-0.5 ml of solution in the tube.
17. 2 volumes of ethanol (0.8-1.0 ml) was added to the tube and it was placed at 4°C for 5 minutes, before being spun at 1500 rpm for a further 5 minutes.
18. The supernatant was discarded and the pellet was resuspended in 400 μl of TE (10:0.1) by vortexing the tube, chilling at 4°C for 15 minutes, and then vortexing again.
19. 40 μl of sodium acetate/EDTA solution was added, followed by 2 volumes of ethanol (0.8-1.0 ml). Mixing followed the addition of both solutions; the tube was then placed at -20°C for at least 30 minutes.
20. The tube was spun at 1500 rpm for 5 minutes, and the supernatant was discarded.
21. 1 ml of 80% ethanol was added and the tube was centrifuged at 1500 rpm for 5 minutes. The supernatant was removed and final traces of ethanol were left to drain out of the tube for 5 minutes, before vacuum-drying for 5-10 minutes.
22. Resuspension was performed, using 20 μl of TE (10:0.1).

2.4.2. Sonication and subfragment end repair of plasmid DNA

1. In order to estimate the concentration of DNA in the BAC/PAC, a 0.5% agarose mini-gel was run on a 10 x dilution of the sample. A gel was prepared using 50 ml 1 x TBE and 0.25 g of agarose, and samples were run alongside λHindIII/pBR322 markers. Samples were visualized by soaking the gel in 500 ml of 1 x TBE containing 25 μl ethidium bromide (10 mg/ml).
2. From this gel picture, the amount of DNA required to obtain 10 μg was taken for sonication. Water was added so the total volume of water and DNA was 54 μl. 6μl of mung bean buffer was added to this and the mixture was vortexed.

3. The tube was placed in the cup horn containing ice cold water inside the sonicator (in a cold room). The tube was positioned about 1 mm away from the face of the probe.
4. An output of approximately 12% on the 400 watt Virsonic 300 sonicator was used for 10 seconds in order to produce fragments of the required length. (Required outputs/ time vary according to the specifications of the sonicator, and the size of fragments which are desired).
5. If no movement and cavitation of the cup and tube could be observed, sonication was performed again. The mixture was briefly centrifuged at 10000 rpm.
6. 1 μl of sonicated DNA was mixed with 4 μl of loading buffer and the sample was run alongside λHindIII/pBR322 markers on a 0.8% minigel. (0.4g agarose, 50 ml 1 x TBE).
7. The DNA was checked after sonication: the ideal outcome was a smear with no sign of a band of high molecular weight DNA. Near complete sonication was also observed (a smear with a faint band of high molecular weight), and unsonicated samples showing only faint smearing with a substantial band of high molecular weight were also present.
8. Unsonicated samples were sonicated again as above. A second check gel was run to see if these samples had been fragmented. Samples showing incomplete sonication were sonicated for 5 further seconds.
9. The ends of the sonicated DNA fragments were repaired by adding 0.3 μl of mung bean nuclease buffer to the DNA. This mixture was placed in a 30°C water bath for 10 minutes.
10. The volume in the tube was made up to 200 μl with water, and 20 μl of 1 M sodium chloride, 550 μl of ice cold 100% ethanol, and 1 μl of pellet paint were added to the DNA.
11. In order to precipitate the DNA, it was left overnight (or for at least 2 hours) at -20°C and then centrifuged for 30 minutes at 4°C, 13000 rpm.

12. The supernatant was removed from the tube, leaving the DNA pellet which was washed in 1 ml 100% ethanol by centrifugation for 10 minutes at 4°C, 13000 rpm.
13. The ethanol was removed and the pellet was dried in a vacuum dryer for 10-15 minutes.

2.4.3. Selection of suitably sized DNA fragments for subcloning

1. A 0.8% TAE gel (0.4 g agarose, 50 ml 1 x TAE, 2 µl ethidium bromide) was made and was placed in a gel tank containing 500 ml of 1 x TAE and 20 µl of ethidium bromide (10 mg/ml).
2. The pellet was resuspended for loading in 6.25 µl of TE (10:0.1), 0.75 µl 10 x TAE, and 2 µl of loading dye. Care was taken to ensure all the DNA pellet was incorporated in this mixture.
3. All (9 µl) of this mix was loaded alongside λHindIII/pBR322 markers, and the gel was run at 35 mA, 50-60 V for approximately 2 hours.
4. On the long wave ultra violet transilluminator, bands corresponding to the 1.4-2 Kb (ideal) size were cut out. Additional bands of 0.6-1 Kb, 1-1.4 Kb and 2-4 Kb were also cut from the gel: these were stored in case they were needed at a later stage.
5. The pieces of gel were weighed so gel volumes could be estimated.
6. The 1.4-2 Kb gel fragment was placed in a tube and incubated at 65 °C for 5-10 minutes.
7. 4 µl of AgarACE (Promega) was added to the tube in a 42°C waterbath. The molten gel was incubated at 42°C for 15 minutes.
8. 30 µl of sodium chloride, 200 µl of buffered phenol and 196 µl (corresponding to the weight of the gel piece) of TE (10:1) buffer were added to the tube.
9. The tube was spun down for 3 minutes at 13000 rpm and the upper (aqueous) phase (about 230 µl) was extracted and added to a new tube.

10. 100 μl of TE (10:1) was added to the old mixture which was respun at 13000 rpm for 3 minutes. The upper layer (about 100 μl) was extracted and added to the 230 μl extracted earlier, whilst the organic phase was discarded.
11. 130 μl of isobutanol was added to the new tube which was spun at 13000 rpm for 1 minute. The aqueous layer was extracted and discarded.
12. 1 μl of pellet paint (Novagen) and 700 μl 100% ethanol were added to the tube which was placed at -20°C overnight (or for a minimum of 30 minutes).
13. The tube was spun at 4°C, 13000 rpm for 30 minutes, and the ethanol was decanted out of the tube.
14. The pellet was resuspended in 1 ml of ethanol and spun at 4°C, 13000 rpm for 10 minutes.
15. Ethanol was removed from the pellet, which was vacuum dried for 5-10 minutes before resuspension in 5 μl of TE (10:0.1).
16. To check for successful elution, 0.5 μl of DNA with 4.5 μl of loading dye was run out on a 0.8% TBE agarose gel with λHindIII/pBR322 markers.

2.4.4. Ligation and transformations

2.4.4.i. Ligation into pUC18 vector

1. A premix of pUC18 (SmaI/CIP, Amersham) and buffer, consisting of 0.05 μl of pUC18 per reaction and 0.1 μl of buffer (supplied with the pUC18) was prepared by vortexing and placing the tube on ice.
2. 0.15μl of the pUC18-buffer mix was dispensed into the 0.5 ml tubes set-up for each reaction.
3. 0.7 μl of DNA was added to each tube. In addition 3 control tubes were set-up with the following: (a) 0.7 μl water (b) 0.7 μl water (c) 0.7 μl Φx174/HaeIII (1.4 ng)
4. 5 μl of mineral oil was added to each tube.

5. With the exception of tube (b), 0.15 µl T4 DNA ligase was dispensed to each tube, aiming for the 'bubble' under the oil, and the tubes were mixed and centrifuged for a few seconds.
6. Tubes were transferred to a 16°C incubator and left overnight to allow ligation to occur.
7. Tubes were heated to 65°C for 7 minutes, before being left at room temperature for 5 minutes, and centrifuged briefly.
8. 49 µl of water was added to each reaction, and tubes were stored at -20°C until transformations were performed.

2.4.4.ii. Ligation into M13 vector

1. A premix consisting of 0.2 µl M13mp18 (SmaI/CIP, Amersham) per reaction and 0.2 µl buffer (supplied with the vector) was made up.
2. 0.4 µl of this mix and 1.4 µl of DNA was dispensed into each tube.
3. As with the pUC18 ligations, 3 controls were set-up: (a) 1.4 µl water (b) 1.4 µl water (c) 1.0 µl phix174/HaeIII (2.0 ng)
4. With the exception of tube (b), 0.2 µl T4 ligase was added to each tube, and tubes were shaken and centrifuged gently.
5. Tubes were transferred to a 16°C incubator overnight.
6. Tubes were heated to 65°C for 7 minutes, before being left at room temperature for 5 minutes, chilled on ice, and then centrifuged briefly.
7. 18 µl of water was added to each reaction and tubes were stored at -20°C.

2.4.4.iii. Transformations of pUC18 vectors

1. 1 µl of ligated DNA was aliquoted into 15 ml glass test-tubes, and 500 µl of SOC was added to each 1 ml Eppendorf tube.
2. TG-1 cells (Invitrogen, maintained in 10% glycerol and stored at -70°C) were removed from the freezer and 150 µl 10% glycerol was added to each tube of cells which were then left on ice.
3. Cells and glycerol were mixed using a P200 Gilson pipette, and 40 µl of this mixture was added to the ligated DNA in the Eppendorf tube.
4. The cells, glycerol and DNA were aliquoted into a cuvette placed on ice
5. The SOC solution was warmed in a water bath (20-30°C) and the solution was taken up in a Pasteur pipette.
6. The cuvette containing the DNA ands cells was placed in an electroporator, which was set to deliver a pulse in the range 3.8-5.0. (This range had been optimised by control experiments assessing the efficiency of transformation at a range of electric pulses.)
7. The cuvette was removed from the electroporator and 400 µl SOC was added to the cuvette: the mixture of SOC, cells and DNA was taken up and ejected into a test-tube.
8. Test-tubes were incubated in a shaker at 30°C for 1 hour with agitation.
9. TYE/Amp plates (90 mm) were placed at room temperature.
10. Test-tubes were removed from the shaker and 50 µl IPTG (40 mg/ml) and 50 µl Xgal (50 mg/ml) were added to each tube.
11. 125 µl of the solution was dispensed onto one TYE/Amp plate and 250 µl was dispensed onto a second plate.
12. A sterile spreader was used to make the solution cover the plate in an even manner.
13. Plates were placed in a 37°C incubator overnight.

2.4.4.iv. Transformations of M13 vectors

1. TYE/AMP plates were placed in a 37°C incubator, and 1 litre of H-Top agar was melted in a microwave.
2. 0.2 µl of ligated DNA was dispensed into a 1 ml eppendorf tube which was placed in a heated rack.
3. 3 ml of H-Top agar was added to one glass test-tube for each reaction, along with 25 µl IPTG (40 mg/ml) and 25 µl Xgal (25mg/ml).
4. Each tube of TG-1 cells was mixed with 150 µl 10% glycerol, and 40 µl of this mixture was added to the ligated DNA in the Eppendorf tube.
5. This mixture was added to a cuvette placed on ice.
6. Plates were removed from the incubator, and warmed SOC was taken up in a Pasteur pipette.
7. The cuvette was placed in the electroporator, and a pulse of 4.4-4.6 was delivered to the cells.
8. The warmed SOC (400 µl) was used to dilute the mixture in the cuvette which was then transferred to a test-tube containing the H-Top, IPTG and Xgal.
9. The contents of the test-tube was mixed by rolling the tube once between the palms.
10. The mixture was then emptied onto a TYE/Amp plate and the plate was swirled until an even coverage was obtained.
11. Once the H-Top had set, plates were placed in a 37°C incubator overnight.

2.5. Shotgun sequencing.

In the high throughput system operated at the Sanger Centre, successful ligations were stored at -20°C until clones were selected for shotgun sequencing. Upon selection, a number of plates (5 pUC, 5 M13) were produced from the ligations. Colonies from this plates were picked into 96 well plates containing the appropriate growth media, and after growth, DNA was prepared for sequencing. Prior to the preparation of this DNA, cells from each well were transferred into a 96 well plate (Corning) containing glycerol; this provided a stock of cells allowing inserts from a specific well to be regrown if necessary. These back-up stocks were useful if DNA from a specific insert was required to assembly a clone.

2.5.1. Preparation of template DNA in M13 vector (based on Mardis (1994)).

1. 5 ml of TG-1 cells was added to 500 ml of 2 xTY media containing ampicillin, and 1 ml of this solution was aliquoted into each well of a 96 well Beckman box. (Either by hand or by using an automated plate filler).
2. Colonies from TYE/Amp plates were picked into these wells. (Either by hand or using an automated picking machine.)
3. Boxes were sealed and lids were pierced for aeration, before boxes were placed in a 37°C incubator at 360 rpm for 12.5 hours.
4. 100 µl of the cells were removed from each well and added to a 96 well plate (Corning) containing 50 µl 10% glycerol. These plates were sealed and stored at -70°C.
5. Boxes were centrifuged at 4000 rpm for 2 minutes.
6. New Beckman boxes, with each well containing 145 µl of 20% PEG 8000, were set-up.
7. After centrifugation, 580 µl of the supernatant was transferred from the Beckman boxes containing cells into the Beckman boxes containing the PEG.

8. Boxes were sealed and shaken (by hand), and left for 20 minutes at room temperature.
9. Boxes were centrifuged at 4000 rpm for 20 minutes, and the supernatant was discarded with boxes being drained onto paper towels.
10. Boxes were spun upside down on towels at 300 rpm for 2 minutes to remove lingering traces of supernatant.
11. 20 μl of triton was added per well and boxes were sealed with silver foil, before being strongly vortexed, briefly spun to 1000 rpm, and vortexed and spun once more.
12. Boxes were placed in a 80°C water bath for 10 minutes.
13. Boxes were centrifuged to 1000 rpm, 40 μl water was added, and boxes were spun to 1000 rpm again.
14. 96 well microtitre plates (Serocluster), containing 170 μl EtOH/NaOAC mix per well, were set-up.
15. The contents of each well (60μl) of the Beckman box were transferred into the prepared microtitre plate, and solutions were mixed by pipetting up and down.
16. Microtitre plates were centrifuged at 4000 rpm for 60 minutes.
17. The supernatant was decanted from the plates, which were drained on towels before adding 200 μl of ice-cold 70% ethanol.
18. Plates were centrifuged at 4000 rpm for 15 minutes, the supernatant was decanted and plates were drained on towels.
19. Finally, plates were placed in a 37°C oven for 30-60 minutes in order to dry the pellets which were then resuspended in 60 μl 0.1 mM EDTA.

2.5.2. Preparation of template DNA in pUC18 vector

1. 1 ml of circlegrow containing ampicillin was aliquoted into each well of a 96 well Beckman box, and separate colonies were picked into each of these wells.
2. Boxes were sealed and the lids were pierced before boxes were placed in a 37°C incubator and left to grow for 22 hours.
3. After growth, 100 µl of the cells were removed from each well and added to a 96 well plate (Corning) containing 50 µl 10% glycerol. These plates were sealed and stored at -70°C.
4. Boxes were spun for 5 minutes at 4000 rpm, the supernatant was discarded and boxes were placed upside down on towels for 20 minutes to dry.
5. 250 µl GET solution (Solution 1) was added to each well and cells were vortexed for 2 minutes.
6. Boxes were spun at 4000 rpm for 5 minutes to pellet cells.
7. The supernatant was discarded and boxes were left to drain, before 250 µl GET solution was added and boxes were vortexed for 2 minutes.
8. Microtitre plates (Serocluster) containing 4 µl RNase A (20 mg/ml) were set-up.
9. From each well, 60 µl of the resuspended cells were transferred to these Serocluster plates.
10. 60 µl NaOH/SDS solution was added to each well and plates were sealed with 3M plate sealers (Scotch), before solutions were mixed by inversion (10 times).
11. Plates were left at room temperature for 10 minutes.
12. 60 µl potassium acetate (3 M) was added, plates were sealed and solutions were mixed by inversion (10 times).

13. Plates were left at room temperature for 10 minutes, plate sealers were removed and plates were placed in a 90°C oven for 30 minutes.
14. The plate was placed on ice for 5 minutes.
15. Filter plates were prepared by taping a Millipore 96 well filter plate on top of a Falcon 96 well plate.
16. The contents of each well was transferred from the Serocluster plate into the filter plate, and filter plates were spun at 3300 rpm for 2 minutes.
17. The top filter plate was discarded and 110 μl isopropanol was added to the filtrate.
18. Plates were sealed and mixed by inversion (twice), before spinning at 3750 rpm for 30 minutes.
19. The supernatant was discarded and 200 μl ice-cold 70% ethanol was added to the plates which were spun for 5 minutes at 3750 rpm.
20. The supernatant was discarded and plates were allowed to drain on towels.
21. When the pellet was totally dry, it was resuspended in 35 μl of 1 M Tris-HCl, 0.1 mM EDTA.

2.5.3. The sequencing reaction.

2.5.3.i .Using dye primers:

1. 2 μl of DNA was aliquoted into 4 wells, and 8 μl of the specific dye primer ready reaction mix (ABI) was added to each well. (Each of the 4 wells should contain only one of the 4 dye primers.)
2. This mixture was spun and placed on a thermocycler with the following program: (i) 92°C for 15 seconds (ii) 50°C for 15 seconds (iii) 70°C for 1 minute, (iv) repeat (i) – (iii) for 20 cycles (v) 4°C until stopped.

3. The DNA from all 4 wells was combined in one well of a Serocluster plate, the plate was spun to 1000 rpm, and 10 µl 3M sodium acetate (pH 4.8) and 160 µl 96% ice-cold ethanol were added.
4. The plate was spun at 4°C, 4000 rpm for 90 minutes, and the ethanol was decanted.
5. 200 µl of 70% ice-cold ethanol was added, and the plate was spun for 15 minutes at 4°C, 4000 rpm.
6. The ethanol was removed and the pellet was dried for sequencing. Alternatively, plates were stored at -20°C until sequencing space became available.

2.5.3.ii. Using dye terminators:

1. 3 µl of DNA was added to 9 µl of a mix made up of 1 µl primer (6 pM), 4 µl dye terminator ready reaction mix (ABI) and 4 µl water.
2. The mixture was spun and placed on a thermocycler with the following program:

   (i) 96°C for 10 seconds (ii) 50°C for 5 seconds (iii) 60°C for 4 minutes, (iv) repeat (i) – (iii) for 25 cycles (v) 4°C until stopped.
3. The DNA was transferred to a Serocluster plate and 10 µl 3 M sodium acetate (pH 4.8) and 160 µl 96% ice-cold ethanol were added.
4. The plate was spun at 4°C, 4000 rpm for 60 minutes, and the ethanol was decanted.
5. Steps 5-6 as described above (for the dye primers) were performed.

2.5.4. Sequencing instrumentation.

Clones sequenced by me were loaded on either an ABIPRISM 373 sequencer or an ABIPRISM 377 sequencer (ABI, Foster City, USA).

2.5.4.i. ABI-373 set-up:

1. 3 µl of sequencer loading dye was added to each well, and samples were briefly centrifuged.
2. The gel was inserted into the machine, and after cleaning the glass plate around the laser, the machine was plate-checked: if the glass plate appeared clear then the upper buffer chamber was put in place and both upper and lower chambers were filled with 1 x TBE buffer before pre-running the machine for 30 minutes.
3. Samples were denatured by heating at 80°C for 10 minutes before loading.
4. The comb was removed from the gel and wells were washed out before samples (36 at most) were loaded using a Gilson pipette.
5. Data was collected over a run-time of 8 hours.

2.5.4.ii. ABI-377 set-up:

1. 2 µl of loading dye was added to each well, and samples were briefly centrifuged.
2. The gel was inserted into the machine, and after cleaning the area of the gel around the laser, the machine was plate-checked.
3. The upper buffer chamber was put in place, along with the heat plate that clipped onto the front of the gel, and the machine was pre-run for 30 minutes.
4. Buffer chambers were filled with TBE, samples were denatured as above, and wells were washed out before samples were loaded (48-60 samples) and run for 4 hours.

2.6. Data analysis of shotgun sequencing reactions and clone assembly.

After a basic level of analysis, data produced from the ABI sequencers was transferred to the UNIX system where a number of programs have been developed for the analysis of this data. The first procedure involved in analysing a sequencing gel is to establish the position of each sample on a gel. This lane tracking is automatically performed by the program 'Gelminder' (Platt and Mullikin, unpublished) but manual checking and in some cases, repositioning is required. After manual checking of the lane tracking, 'Gelminder' moves onto to call the bases. Data from each sample is then passed into the 'Automated Sequence Preprocessor (ASP)' program (Hodgson, unpublished) which cuts off sequence according to whether it is cloning or sequencing vector, *E.coli* contamination or sequence of an unacceptably poor quality. Clipped good quality sequences are then passed into the 'Phrap2Gap' program (Mott and Dear, unpublished). This program is a modified version of 'Phred' and 'Phrap' (Gordon *et al.*, 1998), which are base calling programs and sequence assembly programs respectively. 'Phrap2Gap' allows phrap-assembled reads to be transferred into the 'GAP' editing package. The 'GAP' sequence assembly program was developed as part of the Staden package (Bonfield *et al.*, 1995, Staden *et al.*, 2000, Staden *et al.*, 2001); over the years versions have been updated from 'xGAP' to 'GAP' to 'GAP4' to 'GAP4.new'. Clones assembled as part of this project were largely assembled using 'GAP4' and 'GAP4.new' packages (The human clone AL031983 and the mouse clone AL078630 were both assembled by me using this software).

Generally, upon transfer of clone DNA into a 'GAP' package, the clone was not a contiguous piece of sequence and a number of steps were required in order to produce a 'finished' clone, defined as a contiguous piece of sequence with both cloning vector arms present. A 'finished' clone also required that all the sequence was 'double stranded', which refers to the idea that all the clone should be covered by at least two individual reads. Assembling a clone, therefore, required the use of a number of pieces of software, resequencing certain subclones and generating specific segments of DNA using the PCR reaction.

2.6.1. PCR reaction used in clone assembly

1. 1 µl (40 nM) of forward primer and 1 µl (40 nM) reverse primer were dispensed into a 96 well plate (Costar), spun briefly and dried down in a 90°C oven.
2. A master mix was made: per sample, 5 µl PCR buffer (AmpliTaq), 2 µl 4 x dNTPs, 2 µl AmpliTaq, 33 µl water.
3. 2 µl of the appropriate DNA (taken from DNA stock plates) was added to the well, along with 47 µl of the master mix.
4. Samples were placed on the thermocycler with the following program: (i) 94°C for 1 minute, (ii) 55°C for 1 minute, (iii) 72°C for 3 minutes, (iv) steps (i)-(iii) repeated 25 times, (v) 4°C until program stopped.
5. 50 µl MgCl PEG was added to each sample, and samples were well mixed.
6. Plates were sealed and left at -20°C for 1 hour.
7. Samples were spun for 1 hour at 4°C, 4000 rpm, and the supernatant was discarded.
8. Plates were spun upside on tissue for 2 minutes at 250 rpm and 50 µl of ice-cold 96% ethanol was added to the samples.
9. Ethanol was discarded and plates were spun upside down on tissue for 2 minutes at 250 rpm.
10. After drying, 50 µl water was added to resuspend the DNA pellet.
11. Sequencing protocols were performed as above, using the appropriate primer(s).

After a clone was contiguous and double stranded, the virtual restriction digest of the clone was checked against fragments generated by 3 actual restriction digests. This involved generating the real digests (described below) and generating the virtual digests. Virtual digests were generated by the program ‘Confirm’ (Production Software Group, The Sanger Centre, unpublished) which also has a graphical display showing the real and virtual digests alongside each other.

2.6.2. Restriction digests used to check veracity of clone assembly

1. 30-40 ng DNA, estimated from the gel run prior to subcloning, was diluted to make a total volume of 2.5 µl.
2. 2.5 µl DNA + water, 2.5 µl of the restriction enzyme buffer (supplied with the enzyme) and 0.3 µl of the restriction enzyme (BAMHI, EcoRI, or HindIII) were placed in a well in a Costar 96 well plate.
3. The plate was spun to 1000 rpm and placed on a thermocycler at 37°C for 2 hours.
4. Reactions were placed briefly at -20°C, and then in the oven set at 60°C for 10 minutes.
5. Samples were mixed with 1 µl Ficoll dye and 7 µl of the mixture was run on a 0.7% agarose gel (1 x TAE, 200 ml gel) with 6 µl of 1 Kb ladder marker and 6 µl of λBst11071 ladder marker.
6. This gel was run overnight before staining with Vistra Green (Vistra systems): 5 ml 1 M Tris-HCl, 500 µl 0.1 M EDTA and 50 µl Vistra Green were mixed before being added to 500 ml 1 x TAE in a gel-staining tank.
7. The gel was placed in the tank containing the stain which was sealed and gently shaken for 1 hour.

2.7. Construction of mouse filters for hybridisation

In an attempt to extend the mouse contig, a number of membrane filters spotted with the clones screened during fluorescent fingerprinting were produced.

1. Fresh LB plates (8 x 12 cm, rectangular) containing the appropriate anitibiotic (kanamycin for PACs and chloramphenicol for BACs) were set-up.
2. A 8 x 12 cm piece of nylon membrane filter (Hybond N+, Amersham) was labelled with a permanent pen, and was gently placed on the surface of the agar avoiding trapping any air bubbles between the filter and the agar surface.
3. Glycerol stocks of the clones to be plated were allowed to thaw, whilst the robotic gridding system was set-up with stations containing 95% ethanol bath for sterilization, a sonication bath containing water and 0.1% Decon disinfectant, and a bath of indelible ink (autoclaved 1% Higgins black ink).
4. The thawed glycerol plate and the agar plate with the filter were placed at the appropriate positions, and the sterile ink was spotted in the pattern programmed into the robotic system.
5. Pins were cleaned in the sonication bath for 1 minute.
6. A gridding cycle, consisting of immersing the pins in ethanol for 10 seconds, followed by air drying for 10 seconds and then inoculating the culture onto the filter, was performed.
7. This step was repeated until all the clones were plated out on top of the ink spots generated in step 4; the pins were then sonicated for 1 minute.
8. Plates were incubated upside down at 37°C for 12-16 hours.
9. After 12-16 hours growth should be circular and generally uniform in size across the array: if this was achieved the next step was the lysis of the bacterial colonies.

10. Using forceps, membrane filters were removed from the agar surface, and these were placed colony-side up onto Whatman 3MM paper saturated with 10% SDS for 4 minutes.
11. Membrane filters were then placed colony-side up onto Whatman 3MM paper saturated with denaturing solution for 10 minutes.
12. Membrane filters were placed colony-side up onto Whatman 3MM paper, and allowed to air-dry for 10-20 minutes.
13. Filters were submerged in an excess of 10 x neutralizing solution for 5 minutes, with intermittent agitation. (Repeated separately for each filter).
14. Filters were submerged in an excess of 1 x neutralizing solution for 5 minutes, with intermittent agitation.
15. Filters were submerged in 2 x SSC/0.1% SDS wash solution for 5 minutes, agitating intermittently.
16. Filters were submerged in an excess of 2 x SSC for 5 minutes, again with some agitation.
17. Membrane filters were placed in an excess of 50 mM Tris-Cl and agitated intermittently (Repeated separately for each filter).
18. Membrane filters were placed colony-side up onto Whatman 3MM paper, and allowed to air-dry. These filters can be stored at room temperature for several years.
19. Prior to hybridization the DNA was cross-linked to the filters. (Amount of time for cross-linking calibrated by using a control repeated hybridisation with filters stripped for different lengths of time.)

2.8. Construction of mouse olfactory receptor gene vectors for *in situ* hybridisations

11 mouse OR genes were amplified by PCR and cloned into the pGEM T-Easy vector (Promega) system so these vectors could be used in *in situ* hybridisations of rat and mouse tissue. The pGEM

T-Easy vector system is advantageous for the cloning of PCR products, since vectors are prepared by cutting with EcoRV and adding a 3' terminal thymidine to both ends.

The amount of PCR product to be used was calculated as follows:

$$\frac{\text{ng of vector*kb size of insert}}{\text{kb size of vector}} * \frac{\text{insert}}{\text{vector}} \text{ (ratio)}$$

2.8.1. pGEM T-Easy ligations and transformations

1. The pGEM-T Easy vector, control insert DNA and PCR products were centrifuged, and the rapid ligation buffer was vortexed.
2. The reactions were set up in 0.5 ml tubes, as follows: 5 µl 2x rapid ligation buffer, T4 ligase; 1 µl (50 ng) pGEM-T Easy vector; 1 µl (3 U) T4 DNA ligase; 36 ng of PCR product in 3 µl of water / 2 µl control insert and 1 µl water (positive control) / 3 µl water (negative control).
3. The reactions were mixed by pipetting, and were incubated overnight at 4°C.
4. Tubes containing the ligation reactions were centrifuged, and 2 µl of each reaction was transferred to a sterile 1.5 ml tube on ice. Another sterile tube, containing 0.1 ng of an uncut plasmid (pGEM) was also set-up.
5. Tubes of JM109 High Efficiency Competent cells (Promega) were thawed in an ice bath (for about 5 minutes), and mixed by gently flicking the tube.
6. 50 µl of cells were transferred into each 1.5 ml tube, the contents were mixed by flicking the tubes and tubes were then placed on ice for 20 minutes.
7. Cells were heat-shocked for 45-50 seconds in a water bath at exactly 42°C, before tubes were returned to ice for 2 minutes.

8. 950 µl of SOC medium was added to tubes containing cells transformed with ligation reactions, and 900 µl was added to the tube containing the uncut plasmid.
9. Tubes were then shaken and incubated for 1.5 hours at 37°C (150 rpm).
10. 100 µl of each culture was plated out onto 2 LB/ampicillin/IPTG/X-Gal plates.
11. Plates were incubated for 16-24 hours at 37°C.

2.8.2. pGEM T-Easy Preparation of DNA

1. Colonies were picked and used to inoculate 3 ml of LB containing ampicillin. The broth was then incubated overnight at 37°C.
2. Tubes were centrifuged for 5 minutes at 10000 rpm to pellet bacteria and the medium was discarded.
3. 250 µl of cell resuspension solution was added and the pellet was resuspended by vortexing or pipetting, before resuspended cells were transferred to a sterile 1.5 ml tube.
4. 250 µl cell lysis solution was added and the tube was inverted 4 times. The mixture was incubated at room temperature until the cell solution cleared (1-5 minutes).
5. 10 µl of alkaline protease solution was added and the tube was inverted 4 times, before incubation for 5 minutes.
6. 350 µl of neutralization solution was added and the tube was again inverted 4 times.
7. The lysate was centrifuged at 14000 rpm for 10 minutes.
8. The cleared lysate (~850µl) was transferred to the spin column (avoiding the white precipitate).
9. The supernatant was centrifuged at 14000 rpm, left for 1 minute at room temperature, and the flowthrough was discarded.
10. 750 µl column wash solution (diluted with 95% ethanol) was added.

11. The tube was centrifuged at 14000 rpm for 1 minute and the flowthrough was discarded.
12. 250 µl column wash solution was added and the tube was centrifuged at 14000 rpm for 2 minutes.
13. The spin column was transferred to a new tube, and 100 µl of nuclease free water was added before centrifugation at 14000 rpm for 1 minute.
14. The DNA was precipitated by adding 50 µl of 7.5 M ammonium acetate, 375 µl 95% ethanol, and centrifuging the sample at 14000 rpm for 15 minutes.
15. The pellet was washed in 250 µl 70% ethanol and centrifuged at 14000 rpm for 5 minutes.
16. The pellet was dried and resuspended in 10-25µl nuclease free water, before storage at -20°C.

2.9. Construction of 'olfactory promoter region' pGL3 luciferase reporter vectors

Having found a putative promoter region for the MHC-linked olfactory receptor gene cluster, this region was cloned into the pGL3 luciferase reporter vector (Promega). This vector allows inserts to be analysed for their ability to regulate mammalian gene expression. The pGL3 vector contains a modified coding region for firefly (*Photinus pyralis*) luciferase that has been optimised for monitoring transcriptional activity in transfected eukaryotic cells. Together with the pGL3 basic vector, the pGL3 control vector (containing promoter and enhancer), and the pGL3 promoter vector (containing the promoter but no enhancer), constructed reporter vectors were transfected into Odora and HEK293 cell-lines using the the SuperFect transfection reagent (Qiagen).

2.9.1. pGL3 reporter vector restriction digests, ligations and transformations

BglII digestion: 20 µl DNA (50 ng/µl), 20 µl 'red' buffer (ABgene), 1 µl BglII (ABgene), 9 µl water. Digested at 37°C for 2 hours.

BglII/XhoI digestion: 10 µl DNA(50 ng/µl), 20 µl 'red' buffer (ABgene), 1 µl BglII (ABgene), 1 µl XhoI (ABgene), 18 µl water. Digested at 37°C for 2 hours.

1. Tubes containing 2 µl pGL3 reporter vector DNA (50 ng/µl), 1 µl ligase buffer, 0.5 µl T4 DNA ligase, and 6.5 µl of water containing approximately 40 ng of purified PCR product were set-up.
2. The reaction was incubated for 3 hours at room temperature.
3. 5 µl of each reaction was transferred to a sterile 1.5 ml tube on ice, and tubes of JM109 High Efficiency Competent cells (Promega) were thawed in an ice bath (for about 5 minutes), and mixed by gently flicking the tube.
4. 50 µl of cells were transferred into each 1.5 ml tube, the contents were mixed by flicking the tubes and tubes were then placed on ice for 20 minutes.
5. Cells were heat-shocked for 45-50 seconds in a water bath at 42°C, and tubes were returned to ice for 2 minutes.
6. 950 µl of SOC medium was added to tubes containing cells transformed with ligation reactions, and tubes were then shaken and incubated for 1.5 hours at 37°C (150 rpm).
7. 100 µl of each culture was plated out onto 2 LB/ampicillin/IPTG/X-Gal plates which were incubated for 16-24 hours at 37°C.
8. Colonies were picked into 1 ml of LB broth containing ampicillin and grown overnight at 37°C.

2.9.2. Preparation and sequencing of pGL3 reporter vector DNA

1. Cultures were prepped using the plasmid miniprep purification kit (Qiagen), and the DNA pellet was eluted in 30 µl TE buffer.
2. A 0.8% check gel was loaded with 4 µl of the sample and 4 µl loading dye and run against a 4 µl sample of the unligated vector.
3. A number of samples from each type of PCR product (forward / reverse) were sequenced (see earlier section) in order to check the ligation had been successful.

2.9.3. Transfections of pGL3 reporter vector DNA

HEK293 (Graham *et al.*, 1977) and Odora cell (Murrell and Hunter, 1999) growth medium (with serum, protein and antibiotics): 500 ml Dulbecco's Modified Eagle Medium (DMEM high glucose without L-glutamine and sodium pyruvate, GIBCO), 10% fetal calf serum (FCS, GIBCO), 5 ml 100x L-glutamine, 1 ml 100x penicillin/streptomycin. Stored at 4°C.

1. A cell culture with 70-80% confluence was taken, the growth medium was aspirated using a Pasteur pipette, and the cells were washed twice with 10 ml PBS.
2. 4 ml trypsin-EDTA was added to the cell plate and the plate was swirled gently until cells were detached (for approximately 4 minutes).
3. 6 ml growth medium was placed in a 50 ml tube (Falcon) and the cell suspension was transferred to the tube and mixed with the medium.
4. Cells were centrifuged at 1000 rpm for 5 minutes.
5. The supernatant was removed with a Pasteur pipette and cells were washed with PBS before being resuspended in 10 ml growth medium.

6. 250 µl , 500 µl and 1 ml cell suspension were added to 3 100mm cell culture plates filled with 10 ml growth medium.
7. Plates were swirled gently and put in a 37°C incubator, 5% carbon dioxide for 2 days.
8. Cells were counted using a haemocytometer.
9. 5 x 10^5 cells were seeded in 5 ml growth medium (DMEM + serum, protein and antibiotics) in 60 mm dishes.
10. Cells were incubated at 37°C and 5% carbon dioxide for 2 days (until they had reached 40-80% confluence).
11. 5 µg vector DNA (contained in no more than 50 µl of TE buffer, ie. transfection required a concentration of greater than 0.1 µg/µl) was placed in a 5 ml tube (Eppendorf) with cell growth medium (DMEM containing no serum, proteins or antibiotics) making the total volume up to 150 µl.
12. 20 µl SuperFect transfection reagent was added to the tube, and solutions were mixed by pipetting up and down 5 times.
13. Tubes were incubated at 5-10 minutes room temperature to allow transfection complex formation.
14. The growth medium from cells in the 60 mm dishes was aspirated and cells were washed once with 4 ml PBS.
15. 1 ml cell growth medium (DMEM containing serum and antibiotics) was added to the 5 ml tube containing the vector DNA.
16. Solutions were mixed by pipetting up and down twice, before the total volume was transferred to the cells in the 60 mm dishes.
17. The cells were incubated for 2-3 hours at 37°C, 5% carbon dioxide.
18. The medium was removed by aspiration and the cells were washed 4 times in 4 x PBS.

19. Fresh cell growth medium (DMEM with serum and antibiotics) was added and cells were incubated for 48 hours.

2.10. The pGL3 reporter vector assay

Luciferase activity was determined on the cells in medium with Bright-Glo reagent (Promega). This reagent contains beetle luciferin that is coverted into oxyluciferin by firefly luciferase, releasing a large amount of light energy that can be measured using a luminometer. The Bright-Glo Luciferase Assay system has been designed for use with cells in their growth medium.

1. Cells were counted using a haemocytometer, and the amount of medium containing 2 x 10^4 cells was transferred to a 1.5 ml Eppendorf tube.
2. The volume in the tube was made up to 100 µl using growth medium, and cells were left to equilibrate to room temperature.
3. The Bright-Glo reagent was prepared by transferring the contents of one bottle of Bright-Glo Buffer into one bottle of Bright-Glo substrate, and mixing the solutions by inversion until the substrate was fully dissolved.
4. For each sample, a measurement of luminometer (Sirius) activity was taken prior to the addition of the reagent to control for background luminescence and differences in transparency of tubes and media in each sample.
5. For each sample, the 100 µl of cells was combined with 100 µl of Bright-Glo reagent (Promega), and after 2 minutes to allow cell lysis, a second luminometer measurement was taken.

6. Measurements of background luminescence were subtracted from the second measure of luminescence, and activity values were normalized to the average activity of the pGL3 control vector (which contains both Promoter and Enhancer sequences).
7. For each reaction, 2 separate experiments were performed (new transfections and new luminometer readings.)

2.11. Polymorphism analysis

Cell lines were derived from different donors, representing different HLA haplotypes and different ethnic origins. Eight of the 10 cell lines were HLA homozygous, whereas two (BM19.7, BM28.7) were HLA hemizygous (Ziegler *et al.*, 1985, Volz *et al.*, 1992). All cell lines were grown in RPMI 1640 medium containing antibiotics and 10% fetal calf serum.

1. For each gene, 6 primers were designed (Appendix1) and the appropriate PCRs were set-up (C→D, C→F, C→H, E→D, E→H, G→D) as follows: 2 µl genomic DNA (200 ng), 2 µl primer 1 (50 pmol), 2 µl primer 2 (50 pmol), 4 µl dNTPs (2.5 mMol/1000 ml), 5 µl 10 x PCR buffer (AmpliTaq), 0.5 µl AmpliTaq, 34.5 µl water.
2. Reactions were placed on a thermocycler with the following program: (i) 95°C for 1 minute, (ii) 94°C for 30 seconds, (iii) annealing temperature for 45 seconds, (iv) 72°C for 3 minutes, (v) repeat (ii) – (iv) 45 times, (vi) 72°C for 5 minutes, (vii) 4°C until stopped.
3. 4 µl of the reaction mixture was run out on a 0.8% agarose gel with a 100 bp ladder marker: samples containing DNA products of the correct size were then purified using the PCR purification kit (Qiagen).
4. Reactions were sequenced with the appropriate PCR primers using the protocol described earlier.

5. Sequence from these reactions was assembled in a ‘GAP4’ database with several reads covering each region of the gene.

2.12. Mouse RT-PCR.

Expression of OR genes in various mouse tissues and organs was investigated using RT-PCR. This was performed using unique primers from the 3’ end of the gene.

1. The master-mix was made up as follows: 7.2 µl 34.6% sucrose, 0.187 µl 1 in 10 fresh β-mercaptoethanol (in TE (10:0.1)), 1 µl 10 mM dNTP, 2 µl 10 x Rxn buffer, 3.49 µl dilution buffer, 0.125 µl AmpliTaq polymerase (Perkin Elmer).
2. 2 µl of a mix of both primers (at 50 ng/µl of each primer) and 5 µl of cDNA were aliquoted into the well of a Costar 96 well plate.
3. 14 µl of the master mix was dispensed into each well, and the plate was placed on a thermocycler with the following program: (i) 92°C for 2 minutes, (ii) 92°C for 30 seconds, (iii) 55 / 57.5 / 60 °C for 90 seconds, (iv) 72°C for 1 minute, (v) Repeat (ii) – (iv) 35 or 45 times, (vi) 72°C for 10 minutes, (vii) 4°C until stopped.
4. Prescreening was done at a range of temperatures with either 35 or 45 cycles.
5. Samples were run on a 2.5% agarose gel in 1 x TBE at 200 mA for 40 minutes.

2.13. Human RNA dot-blot hybridisations.

In order to consider whether expression of olfactory receptors was found outside the olfactory epithelium, a human RNA dot-blot (Clontech) was hybridised with probes generated from the 3’ end of the olfactory receptor.

1.15 ml of ExpressHyb (Clontech) was prewarmed in a water bath at 50°C and 150 μl sheared salmon testes DNA (1.5 mg) was heated at 95-100°C for 5 minutes before being chilled quickly on ice.

1. The heat-denatured salmon testes DNA was mixed with the prewarmed ExpressHyb.
2. The master blot was placed in the hybridisation container and 10 ml of the ExpressHyb and salmon testes DNA was added to the container.
3. The master blot was prehybridized for (at least) 30 minutes with continuous agitation at 65°C.
4. The probe was labelled using the High Prime Kit (Boehringer-Mannheim):
    - i. 25 ng DNA in 11 μl water was combined on ice with 4 μl High Prime, and 5 μl (50μCi) [α^{32}P] dCTP, and the mixture was incubated at 37°C for 10 minutes.
    - ii. Free nucleotides were removed using the QiaQuick Nucletide Removal Kit (Qiagen) as follows:
    - iii. 200 μl of buffer PN was added to the reaction which was transferred to a spin column and DNA was bound by centrifuging at 6000 rpm for 1 minute.
    - iv. The flowthrough was discarded and 500 μl buffer PE was added, and the tube was centrifuged at 6000 rpm for 1 minute.
    - v. Step (iv) was repeated again, and the flowthrough was discarded, before centrifugation at 13000 rpm for 1 minute.
    - vi. DNA was eluted with 100 μl buffer EB and by centrifugation at 13000 rpm.
5. 30 μl (30 μg) human CotI DNA, 15 μl salmon sperm DNA, 50 μl 20 x SSC and 5 μl of water was added to the DNA probe.
6. The probe mixture was added to the remaining 5 ml of ExpressHyb and salmon testes DNA solution.

7. The pre-hybridization solution was discarded from the container with the master blot and the solution containing the probe was added to this container.
8. The blot was hybridised with constant agitation at 65°C overnight.
9. Six washes were performed: the first 4 at 65°C using 2 x SSC, 1 % SDS wash solution with the last two at 55°C, using 0.1 x SSC, 0.5% SDS wash solution.
10. The blot was wrapped in a heat sealed plastic bag and exposed to an X-ray film (Fuji) for 2 days at room temperature.

2.14. Titering the olfactory epithelium phage library

1. The bacterial glycerol stock (XL1-Blue MRF' strain) were streaked onto LB-tetracycline agar plates and the plates were incubated overnight at 37°C.
2. Tubes containing 10 ml LB-tetracycline medium supplemented with 10 mM $MgSO_4$ and 0.2% (w/v) maltose were inoculated with a single colony, and grown at 37°C for 4-6 hours (with agitation), or overnight at 30°C (with agitation).
3. Cells were spun at 500 x g for 10 minutes and the supernatant was discarded.
4. Cells were gently resuspended in 5 ml sterile 10 mM $MgSO_4$.
5. Cells were diluted to an OD_{600} of 0.5 with sterile 10 mM $MgSO_4$ and the bacteria were used immediately.
6. 500 µl of the phage library was diluted in 500 µl of SM buffer, and a series of dilutions were set-up and added to the host bacteria in 15 ml tubes (Falcon): (i) 1 µl phage/SM + 200 µl host cells, (ii) 1 µl of 1:10 phage/SM:SM dilution + 200 µl host cells (iii) 1 µl of 1:100 phage/SM:SM dilution + 200 µl host cells, (iv) 1 µl of 1:1000 phage/SM:SM dilution + 200 µl host cells.

7. The phage and bacteria were incubated for 15 minutes at 37°C to allow the phage to attach to cells.
8. The following was added to each tube (15 ml Falcon tube): 2-3 ml of NZY top agar (melted and cooled to ~48°C), 15µl of 0.5 M IPTG, 50µl of X-gal [250 mg/ml (in DMF)]
9. After adding the cooled NZY top agar, the mixture was plated immediately onto NZY agar plates and the plates were allowed to set for 10 minutes and plates were inverted and then incubated overnight at 37°C.
10. Around 1 x 10^5 pfu/µg of background plaques (blue) should be produced by this procedure, with the number of recombinant plaques (white) 10-100 fold higher than the background.

2.15. PCR amplification of the olfactory epithelium phage library

1. The PCR reaction was set-up as follows:0.5 µl of the phage, 2 µl (2.5µM) primer 1, 2 µl (2.5µM) primer 2, 4 µl dNTPS (2.5 mMol/1000 ml), 5 µl 10x PCR buffer, 0.5 µl AmpliTaq, 36 µl water.
2. The thermocycler program was as follows (i) 95°C for 1 minute, (ii) 94°C for 30 seconds, (iii) annealing temperature for 45 seconds, (iv) 72°C for 3 minutes, (v) repeat (ii) – (iv) 45 times, (vi) 72°C for 5 minutes (vii) 4°C until stopped.
3. PCR was set-up using pooled phage suspensions set up in the following way:
    a. The host bacteria were grown as described above.
    b. 1 ml aliquots of bacteria and medium were dispensed into 50 ml tubes, and 1 x 10^5 pfu of phage were diluted in 1 ml of SM buffer.
    c. After 15 minutes in a 37°C waterbath, 18 ml of LB broth containing 10 mM Mg SO_4 was added to the tube.

d. 1 ml of this mixture was aliquoted into a 96 well Beckman box, the box was sealed and incubated at 37°C for 5-6 hours.

e. PCR was performed on 0.5 µl of each of the phage suspensions diluted in 0.5 µl water as follows:

 2µl (2.5µM) primer 1, 2µl (2.5µM) primer 2, 4 µl dNTPS (2.5 mMol/1000 ml), 5 µl 10x PCR buffer, 0.5 µl AmpliTaq, 35.5 µl water.

 Thermocycler program: (i) 95°C for 1 minute, (ii) 94°C for 30 seconds, (iii) annealing temperature for 45 seconds, (iv) 72°C for 3 minutes, (v) repeat (ii) – (iv) 45 times, (vi) 72°C for 5 minutes

2.16. *In silico* analysis: gene finding programs

After assembly, the clone DNA was analysed using a variety of programs. Two major methods for doing this were the use of 'NIX' at the *HGMP* and the use of 'AceDB' databases at the Sanger Centre. 'NIX' runs the sequence through a suite of programs. The initial step involves disregarding the repeats in the sequence through running the sequence through '*RepeatMasker'* (Smit and Green). This program screens DNA sequences against a library of interspersed repeats and low complexity DNA sequence (*Repbase*, (Jurka, 2000)). A file produced by 'RepeatMasker' is then run through a number of gene finding programs, including '*Grail',' Genefinder', 'Genemark', 'Fex', 'Hexon',* and *'Fgene'*. 'NIX' also performs 'BLAST' searches against a number of databases: TrEMBL, Swissprot, Unigene, mRNA, EST, EMBL, HTG, GSS, STS, Ecoli, and Vector. 'BLAST' is the acronym for the basic local alignment search tool: this program has become widely used in DNA and protein database searches. It is based on measuring local similarity between sequences, calculated by the maximal segment pair (MSP) score (Altschul *et al.*, 1990). Databases searched by 'NIX' are generally maintained by *EMBL-EBI* or *NCBI*. 'NIX' (Figure 2.1) produces a graphical output of the information produced by these programs which can be used to provide a guide to features contained within the DNA clone. Similar types of programs are used by the 'AceDB' analysis package at the Sanger Centre (Figure 2.2). One additional program used for gene prediction is '*GENSCAN*', another prediction program which predicts gene structures based on profiles of the basic transcriptional, translational and splicing signals, as well as length distributions and compositional features of exons, introns and intergenic regions (Burge and Karlin, 1997, Burge and Karlin, 1998).

Using a combination of gene finding programs, together with results from 'BLAST' searches is the best way of searching for genes, since predictions from any gene finding program taken in isolation are likely to be lacking in both sensitivity and specificity (Guigo *et al.*, 2000). Gene finding programs also tend to struggle with predicting smaller exons, which was sometimes a problem with predicting olfactory receptor genes, and another limitation of these programs is that they often predict incorrect joins between predicted exons (as can be seen with the OR genes in Figure 2.1 and Figure 2.2).

2.17. Identification of genes

Genes were identified in the majority of cases by homology to proteins already present in the public databases. Identification of the start and stop positions was performed by considering the open reading frame, where this existed. In the case of pseudogenes, the start and end of the gene tended to be defined by where the similarity with other proteins started and finished rather than any open reading frame. A number of novel genes were also defined: these showed matches to unidentified proteins, cDNAs or ESTs.

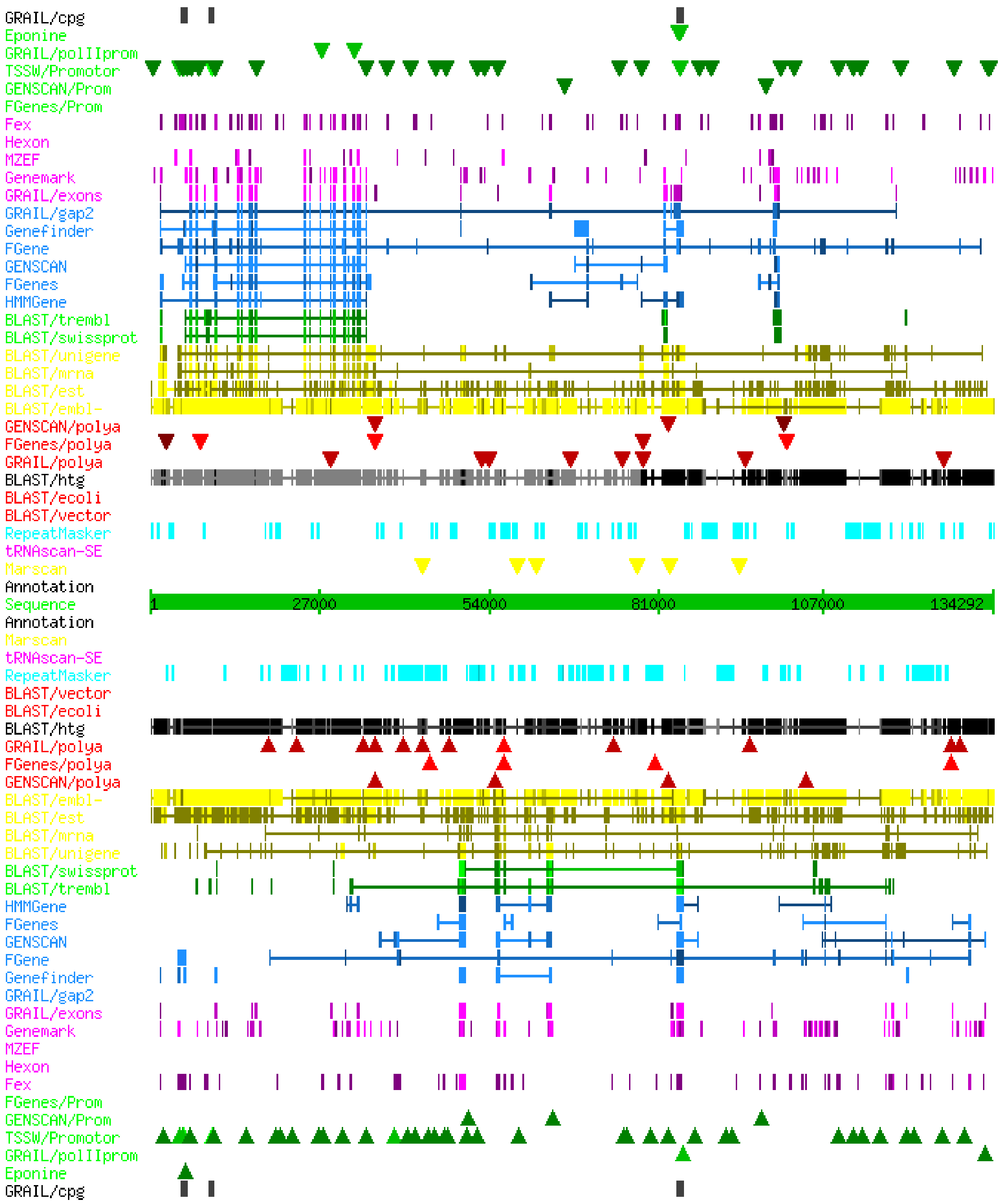
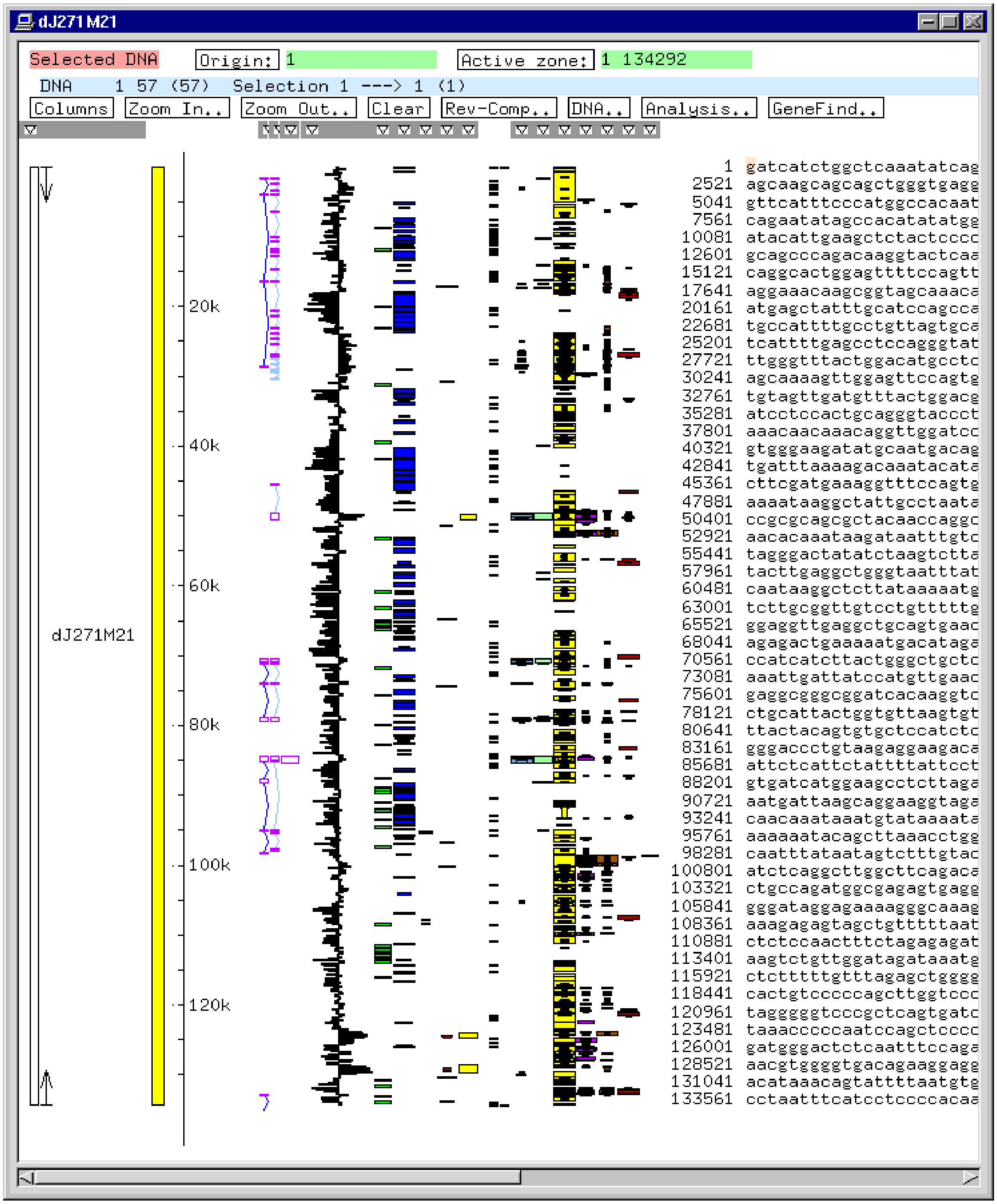


Figure 2.1: ‘NIX’ display of AL031983. Analyses are run on both the forward (above the green sequence line) and reverse strands (below the green line) of the clone. The column to the left shows the name of all the programs run on this piece on sequence: results appear as boxes or triangles on the line corresponding to a particular program.



```
Selected DNA   Origin: 1          Active zone: 1 134292
 DNA      1 57 (57)   Selection 1 ---> 1 (1)
```

```
     1 gatcatctggctcaaatatcag
  2521 agcaagcagcagctgggtgagg
  5041 gttcatttcccatggccacaat
  7561 cagaatatagccacatatatgg
 10081 atacattgaagctctactcccc
 12601 gcagcccagacaaggtactcaa
 15121 caggcactggagttttccagtt
 17641 aggaaacaagcggtagcaaaca
 20161 atgagctatttgcatccagcca
 22681 tgccattttgcctgttagtgca
 25201 tcattttgagcctccagggtat
 27721 ttgggtttactggacatgcctc
 30241 agcaaaagttggagttccagtg
 32761 tgtagttgatgtttactggacg
 35281 atcctccactgcagggtaccct
 37801 aaacaacaaacaggttggatcc
 40321 gtgggaagatatgcaatgacag
 42841 tgatttaaaagacaaatacata
 45361 cttcgatgaaaggtttccagtg
 47881 aaaataaggctattgcctaata
 50401 ccgcgcagcgctacaaccaggc
 52921 aacacaaataagataatttgtc
 55441 tagggactatatctaagtctta
 57961 tacttgaggctgggtaatttat
 60481 caataaggctcttataaaaatg
 63001 tcttgcggttgtcctgtttttg
 65521 ggaggttgaggctgcagtgaac
 68041 agagactgaaaaatgacataga
 70561 ccatcatcttactgggctgctc
 73081 aaattgattatccatgttgaac
 75601 gaggcgggcggatcacaaggtc
 78121 ctgcattactggtgttaagtgt
 80641 ttactacagtgtgctccatctc
 83161 gggaccctgtaagaggaagaca
 85681 attctcattctattttattcct
 88201 gtgatcatggaagcctcttaga
 90721 aatgattaagcaggaaggtaga
 93241 caacaaataaatgtataaaata
 95761 aaaaaatacagcttaaacctgg
 98281 caatttataatagtctttgtac
100801 atctcaggcttggcttcagaca
103321 ctgccagatggcgagagtgagg
105841 gggataggagaaaagggcaaag
108361 aaagagagtagctgtttttaat
110881 ctctccaactttctagagagat
113401 aagtctgttggatagataaatg
115921 ctctttttgtttagagctgggg
118441 cactgtcccccagcttggtccc
120961 taggggggtcccgctcagtgatc
123481 taaaccccccaatccagctcccc
126001 gatgggactctcaatttccaga
128521 aacgtggggtgacagaaggagg
131041 acataaacagtattttaatgtg
133561 cctaatttcatcctccccacaa
```

Figure 2.2: 'AceDB' display of AL031983. Columns to the left of the scale show the following (left to right). Fgenes predictions, GENSCAN predictions, GC content, RepeatMasker SINEs, RepeatMasker LINEs, RepeatMasker inverted repeats, RepeatMasker tandem repeats, CpG islands, BLASTX (SwissProt) matches, GRAIL1.3 predictions, BLASTN matches (EMBL), EST matches, GSS/STS matches and polyA signals.

2.18. Identification of olfactory receptor genes

Olfactory receptor genes were defined as genes with a coding region of approximately 1 Kb. Early work produced the idea that five key protein motifs could be identified as something shared by all olfactory receptor genes: these were (i) FILLG (ii) LHTPMYFFLSHLS (iii) MAYDRYVAIC (iv) KAFSTCGSHLSVV and (v) MLNPFIYSLRN. These motifs were refined further as work progressed. Olfactory receptor pseudogenes (P), meanwhile, were defined as loci with a large number of these motifs still intact in the correct order within an approximate 1 Kb region. In contrast to functional genes, however, pseudogenes were distinguished by stops and/or frame shifts that disrupted the coding region of the gene. Genes lacking a methionine upstream of the FILLG motif (or variation) were also classified as pseudogenes, as were genes lacking any of the well-conserved motifs. Another class of olfactory receptor defined was ‘pseudogene fragments’ (PF). These loci were distinguished as having some of the conserved protein motifs, but they were generally less than 700 bp in length, and did not have 2-3 of the core motifs. Finally, a fourth class of olfactory receptor genes was the ‘fragment’ class (F): these appeared to be functional but disappeared into gaps that were present in the unfinished sequence.

2.19. The assembly of DNA clone sequences

Overlaps between clones were found using two programs, ‘cross_match’ (Green, unpublished) and ‘Dotter’ (Sonnhammer and Durbin, 1995). ‘Cross_match’ utilizes the Smith-Waterman sequence alignment algorithm (Smith and Waterman, 1981) to look for regions of similar sequence. ‘Dotter’, meanwhile, is a graphical dotplot program for detailed comparison of two sequences that compares every residue in one sequence against every residue in the other sequence. The two sequences are plotted on the X and Y axis and high scores are indicated by dots at the appropriate position.

2.20. Storage of information in sequence feature files

Information about the gene content, repeat content, and miscellaneous features (such as CpG islands, promoter predictions) was stored in files formatted so they could be used to generate postscript plots of the area. The format of these files is shown in Appendix 2. Files were either generated manually (".genes" file), or generated through a number of scripts that were developed to produce these files, for example, 'seeclone' which generated these type of files from EMBL entries, 'rptmgenerator' which generated ".rptm" files from RepeatMasker output files, and 'eponinehits' which converted output from the Eponine promoter prediction program into ".misf" files. A number of other programs were also developed to aid in the maintenance of these files: 'reverseclone' alters co-ordinates so the clone can be plotted in the reverse orientation, whilst 'addclone' allows the three files from each clones to be added to the files from another clone. The advantage of storing information in this format was that information could be quickly replotted with additional details or to a different scale. The program 'plotclone' was developed to take input from these three files and produce a postscript plot of features listed in these files. It also allows an optional GC line to be plotted under the other information. (Appendix 4 contains more information about these programs).

2.21. The nomenclature of olfactory receptor genes and the 'ROLF' database

In recent years, a number of nomenclatures for olfactory receptor genes have been published (Zhao and Firestein, 1999, Glusman *et al.*, 2000, Zozulya *et al.*, 2001). These nomenclatures all have different advantages and disadvantages, as does the nomenclature that was used throughout this project. All of the genes described in this project follow a private system of nomenclature that was devised at the start of the project. This nomenclature (for example, hs6M1-6*01) consists of a two letter abbreviation of the species (hs, *Homo sapiens*), followed by the number giving the chromosomal location (in this case chromosome 6, although "U" was used where the location is unknown). This is followed by a letter and number indicating whether the gene shares amino acid identity with the olfactory receptors from the MOE (M1), or with one of the three types of pheromone receptor (V1, V2 or V3). This description is followed by a dash and an arbitrary gene identity number (here, -6). Finally, where the gene has alleles, an asterix and an additional number (*01) describes the corresponding allele, and if the gene is a pseudogene (P), fragment (F), or pseudogene fragment (PF) this is indicated by the appropriate letter. The database created as part of this project utilizes this new nomenclature but links to the public databases are maintained via accession numbers and the position of the gene within the accession number, where applicable. Links to other olfactory receptor gene databases were not built in, owing to time constraints, but where a gene has been extensively studied it has been referred to by the name given in the original literature.

The database of olfactory receptor genes was initially set-up through a search of the EMBL nucleotide database using 'SRS' (Sequence Retrieval Server) and searching for key words. A BLAST-searchable database was created using either the 'gcgtoblast' program (from the 'GCG' package of computer programs (Womble, 2000)) or the 'formatdb' program (part of the 'blastall' suite of programs). A small script ‘Rolfp’ was used to access the database, producing a results file in the directory from which the search was made.

2.22. Gene-finding approach for olfactory receptor genes

Analysis of DNA through running the sequence through the ‘AceDB’ system or the ‘NIX’ suite of programs is one way of identifying genes, but in the large scale analysis of olfactory receptor genes within the human genome, this approach for all clones considered to contain an OR gene would have been impractical. The method adopted for identifying OR genes was therefore to take clones that had been identified as containing olfactory receptor-like sequences (screened using ‘BLAST’) and using the dot-matrix program ‘Dotter’ to identify regions that were positive for OR genes. In the case of OR genes, regions of approximately 2 Kb were identified as positive where a row of high scores ran diagonally across a region of sequence.

In order to extract OR genes from regions of sequence, the ‘olfgrab’ program was developed. This program performs a number of steps: (i) extracts the region of sequence from the relevant clone, (ii) creates 6 files containing each translation of this piece of sequence, (iii) takes the 6 protein files and uses BLAST to compare these files against the database of olfactory receptor genes, (iv) displays the scores from these 6 BLAST searches and (v) creates a file containing the DNA and the highest scoring protein translation. (Appendix 2). This translation and DNA sequence was then manually edited to show only the olfactory receptor gene, and the program ‘olfproducer’ was used to produce a file containing the protein and the nucleotide sequence from this file.

2.23. Analysis of olfactory receptor protein structure

Putative protein features were identified using the 'PIX' suite of programs at the *HGMP*. 'PIX' is similar to 'NIX' in that it runs the protein sequence through a number of programs and outputs all the results from these programs in a graphical form. 'PIX' uses programs searching other protein databases (*'Pfam', 'BLOCKS', 'PRINTS', 'PROSITE'*), predicting protein sorting signals (*'Psort'*), predicting protein secondary structure (*'DSC', 'Simpa96', 'Phd', 'Predator'*) and predicting coiled coils regions (*'COILS'*). In addition it also uses programs predicting transmembrane domains (*'Tmpred', 'Tmap', 'DAS'*), helix-turn-helix motifs ('HTH'), cleavage sites (*'Signal', 'Sigcleave'*), antigenic sites (*'Antigenic'*) and proteolytic enzyme sites (*'Digest'*). Owing to the lack of data about proteins, 'PIX' is less useful in annotating proteins than 'NIX' is in annotating DNA sequence. It did, however, provide estimates for the placement of transmembrane domains within the olfactory receptor proteins. For the consensus olfactory receptor sequence, the placement of transmembrane domains was calculated according to the placement of these domains in each individual protein. This method for predicting transmembrane domains is subject to a number of limitations, notably that these type of prediction programs have a range of 26% to 69% in accurately predicting all transmembrane domains within a protein (Moller *et al.*, 2001). The predictions that were made for OR proteins, however, can be considered to be generally correct since an evaluation of protein prediction programs revealed that one of the prediction programs used in the analysis accurately predicted transmembrane domains in 85% of G-protein coupled receptors tested (Moller *et al.*, 2001).

2.24. Other programs/scripts used throughout the thesis

In addition to the scripts and programs already described, a number of other scripts and programs were used on a regular basis. From EMBOSS suite of programs (Rice *et al.*, 2000), a number of programs were used, including ‘seqret’, a reformatting program; ‘revseq’, which complements a sequence; ‘transeq’, which translates a specific frame and ‘est2genome’ which aligns a set of spliced nucleotide sequences to an unspliced genomic DNA sequence were all used regularly. ‘Water’, a program that uses a modified Smith-Waterman algorithm to produce a local alignment between 2 sequences (DNA or protein) was also used frequently. In addition, I developed a number of scripts, listed in Appendix 4. (This also includes 5 scripts developed by Roger Horton (Chromosome 6 database curator, The Sanger Centre)).

2.25. Genome browsers

The dramatic increase in the amount of sequence data in the public domain led to a number of new bioinformatic resources being developed to provide places to store and annotate this data. Two of these resources were used in this project to provide SNP data for the human MHC extended class I region, and mouse sequence from OR clusters. *‘Ensembl’* is a joint project between EMBL-EBI and the Sanger Institute (Hubbard *et al.*, 2002) and the *UCSC human genome project* is based at UC Santa Cruz (Kent et al, unpublished, Kent and Haussler, 2001).

2.26. Analysis of regulatory regions within OR clusters

2.26.1 ‘Promoter Inspector’, ‘Eponine’ and ‘Transfac’

‘Promoter Inspector’ (Scherf *et al.*, 2000) is an algorithm that relies on the assumption that promoters are embedded into a common genomic context. It assumes that this context can be detected by classifying varying oligonucleotide sequences as ‘promoter’ or ‘non-promoter’-like. The classification was made possible by an original training period in which the algorithm was given access to eukaryotic promoter sequences (from the eukaryotic promoter database (EPD) V

60.0 (Cavin Perier *et al.*, 1998)) and vertebrate exon and intron sequences (randomly extracted from GenBank). 'Promoter Inspector' is currently considered to be 43% accurate in its prediction of promoter sites, and has a dramatically lower rate of false positives compared to older promoter prediction programs. 'Eponine' (Down and Hubbard, 2002) also relies on the detection of a common genomic environment. Classification models that distinguish between promoter and non-promoter-like sequence were trained on mammalian promoters from the EPD (positive sequences) and on an equal number of random sequence fragments from human chromosome 20 that were not annotated as promoters. 'Eponine' has been predicted to have a detection sensitivity of 40%. The false positive rate is also comparable to the rate estimated for 'Promoter Inspector': generally about 45-55% of hits can be considered false positives.

'TRANSFAC' is a comprehensive database containing transcription factors, their genomic binding sites and DNA-binding profiles (Wingender *et al.*, 2000). It can be accessed using programs such as '*MatInspector'* (Quandt *et al.*, 1995), which works with binding profiles generated from the 'TRANSFAC' matrix. 'MatInspector' generates two scores for a sequence matching to a transcription factor, a value for similarity to the matrix and a value for similarity to the core (which is made up of the four consecutive bases within the matrix which show the highest amount of conservation).

2.26.2. DNA Block Aligner ('DBA')

'DBA' (Jareborg *et al.*, 1999) is a program that aligns 2 sequences assuming that sequences share blocks of conservation separated by large and varied lengths of DNA within the 2 sequences. The conserved blocks may have 1 or 2 gaps in them and the amount of conservation can be specified (ranging from 65% upwards). This approach was designed for comparing the upstream regions of genes both between and within species: conserved blocks may be important in regulating the gene.

2.27. Analysis of comparative regions in the human and mouse OR clusters: ‘PipMaker’

‘*PipMake*r’ (Schwartz *et al.*, 2000) is a program that is able to align two long DNA sequences to identify conserved sequences. These conserved sequences are shown in a graphical form as a percentage identity plot (PIP). PIPs are plotted with one sequence along the horizontal axis: this sequence can be labelled with features such as exons and repeats. Sequences conserved within the horizontal sequence compared to the other input sequence are indicated by a vertical line positioned at the appropriate place along the horizontal axis. The extent of conservation is indicated by the position of the vertical line along the vertical axis: conservation ranges from 50% (bottom of vertical axis) to 100% (top of vertical axis). Files containing information about these conserved sequences in textual form are also produced by ‘PipMaker,’ along with the corresponding dot-matrix plot of the region (similar to plots produced by the ‘Dotter’ program).

2.28. Phylogenetic analysis

2.28.1. Protein alignments

All protein alignments were performed using the ‘*ClustalW*’ program (Thompson *et al.*, 1994). This is a progressive multiple sequence alignment method that assigns individual weights to each sequence in a partial alignment in order to down-weight near-duplicate sequences and up-weight the most divergent ones. It also varies amino acid substitution matrices at different alignment stages according to the divergence of the sequences to be aligned, and residue-specific gap penalties and locally reduced gap penalties in hydrophilic regions encourage new gaps in potential loop regions rather than regular secondary structure. After a gap has been opened, locally reduced gap penalties are applied to positions around this gap. In the case of most olfactory receptor alignments, the high number of closely related sequences means that alignments produced by this program tend to be very close to ideal, although all alignments were examined in the program 'belvu' (Sonnhammer, unpublished) before they were used. Alterations to protein alignments were made using the program 'jalview,' (Clamp, unpublished) or alterations were made in Excel after converting a alignment FASTA file into a tabbed alignment using the program 'ProAlnExcel' (Appendix 4).

2.28.2. Phylogenetic tree production

With the exception of the large tree of 716 OR proteins, all phylogenetic trees were constructed using the program 'phylo_win' (Galtier *et al.*, 1996). This is a graphical interface for molecular phylogenetic inference, which can perform neighbor-joining and parsimony methods.

2.28.2.i. The neighbor-joining method and distance calculation methods.

The neighbor-joining (NJ) method (Saitou and Nei, 1987) is a distance based method which uses an algorithm to convert pairwise distances between sequences into a matrix, from which branching order and branch lengths are computed. Advantages of this method include that it is a relatively computationally light process and that only one tree is produced. The disadvantages of this approach stem from these two advantages: producing only one tree means other trees are not evaluated, and the algorithm may not provide an accurate depiction of historical events. Methods to calculate distances for the neighbor-joining method also have a number of problems associated with them. Distance calculation methods available for use in the 'phylo_win' package include the observed divergence, which is simply the observed percent of differences between the compared sequences, the Poisson Correction which attempts to corrects for multiple substitutions according to a one-parameter model, and the PAM distance which corrects for multiple substitutions according to Dayhoff's PAM matrix (Dayhoff, 1976).

In the trees created in this project, the PAM matrix was used as this provides the most complex model of protein evolution, in providing a measure of probability calculating how likely the amino acid in one sequence is likely to change in the amino acid in the other sequence. These probabilities were based on a subset of closely related proteins that were organized into a phylogenetic tree, and the frequency of change from each amino acid to another was determined by adding up the changes at each evolutionary step. This matrix is based on a number of assumptions which will all cause problems when applying this matrix in creating a phylogenetic tree: (i) each amino acid site is equally mutable (ii) the frequency of amino acid changes that would require two nucleotide changes is higher than would be expected by chance (iii) the matrix is based on a small set of closely related proteins (there are other updated matrices based on more protein sequences that could be used (Gonnet *et al.*, 1992, Jones *et al.*, 1992)). The advantage of the PAM matrix, however, is that the frequency of changes was averaged across conserved and non-conserved blocks within the protein (Henikoff and Henikoff, 1992). In view of the fact that olfactory receptor proteins are made up of conserved and non-conserved blocks of protein sequence, therefore, it seemed appropriate to use a matrix based on both conserved and non-conserved region of sequence. Ideally, an olfactory receptor-specific matrix should be used to generate phylogenetic trees but time constraints did not allow for the development of this matrix. In ‘phylo_win’, the PAM matrix is applied using the algorithm of the program ‘PROTDIST’ (Felsenstein 1993, unpublished) of the ‘PHYLIP’ package (Felsenstein 1989).

2.28.2.ii. The (maximum) parsimony method

The second method used to generate phylogenetic trees was parsimony, which is a method that uses the multiple alignment directly rather than using an algorithm based on a calculated distance. Parsimony (Fitch, 1971) is based on the assumption that the most likely tree is the tree that requires the smallest number of changes to explain the data in the alignment. This assumes all the data in the alignment share a common evolutionary origin, and the smallest number of evolutionary steps required to produce the data have been performed. The method calculates a number of trees that could be produced from the data and chooses the tree requiring the lowest possible number of changes to explain the data. In an ideal type situation all possible trees would be calculated and evaluated, however, in reality it would too computationally intensive to generate all possible trees and so parsimony methods generally rely on a heuristic strategy, in which an initial tree is selected and rearrangements are then made to this tree to find the most parsimonious tree. Maximum parsimony has a number of advantages: it does not reduce sequence information to a single number, it tries to provide information on the ancestral sequences and it evaluates different trees. Disadvantages to maximum parsimony include that it is slow in comparison with distance methods, it does not use all the sequence information (only informative sites are used), and it does not correct for multiple mutations that may have occurred in the evolutionary process. Parsimony can also be biased in dealing with among-site variation, it can group sequences with a higher number of changes together rather than considering relatedness, and it can generate a number of trees that are equally parsimonious. Whenever used, these limitations should be considered especially with distantly related protein sequences or sequences of a different function. In the case of olfactory receptor genes, all protein sequences are considered to be derived from a common ancestor and it is assumed that these sequences possess a similar function. The assumed equality of mutational and selectional pressures on olfactory receptor genes means that some of the limitations of this method are less acute when performing reconstructions of this family's evolutionary history. In 'phylo_win' the 'PROTPARS' program from the 'PHYLIP' package is used. (Felsenstein 1989).

2.28.2.iii. Bootstrapping phylogenetic trees

Boot-strapping (Felsenstein, 1985) is a method available in 'phylo_win' that allows the reliability of groupings within a tree to be evaluated. This method involves taking each site within a protein and rearranging sites to create a number of 'pseudoalignments.' These pseudoalignments are then used to recreate a number of trees which are compared to the original tree. Groupings obtained in the original tree are then given a percentage expressing how many times they are recreated in the 'pseudoalignment' trees. Bootstrap values of over 70% were considered to represent reliable groupings, whilst groupings defined with low bootstrap values at their branch points were generally disregarded, although low bootstrap values at interior branches do not necessarily mean the entire phylogenetic tree is worthless. Every phylogenetic tree is the best tree obtainable using a specific method, and computer simulations have shown that that branching patterns of an inferred tree may be correct even if they are not supported by high bootstrap values (Nei and Kumar, 2000).

Chapter 3

The human MHC extended class I region

3.1. Introduction

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

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

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

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

3.2. Sequence assembly and gene content

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

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

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

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

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

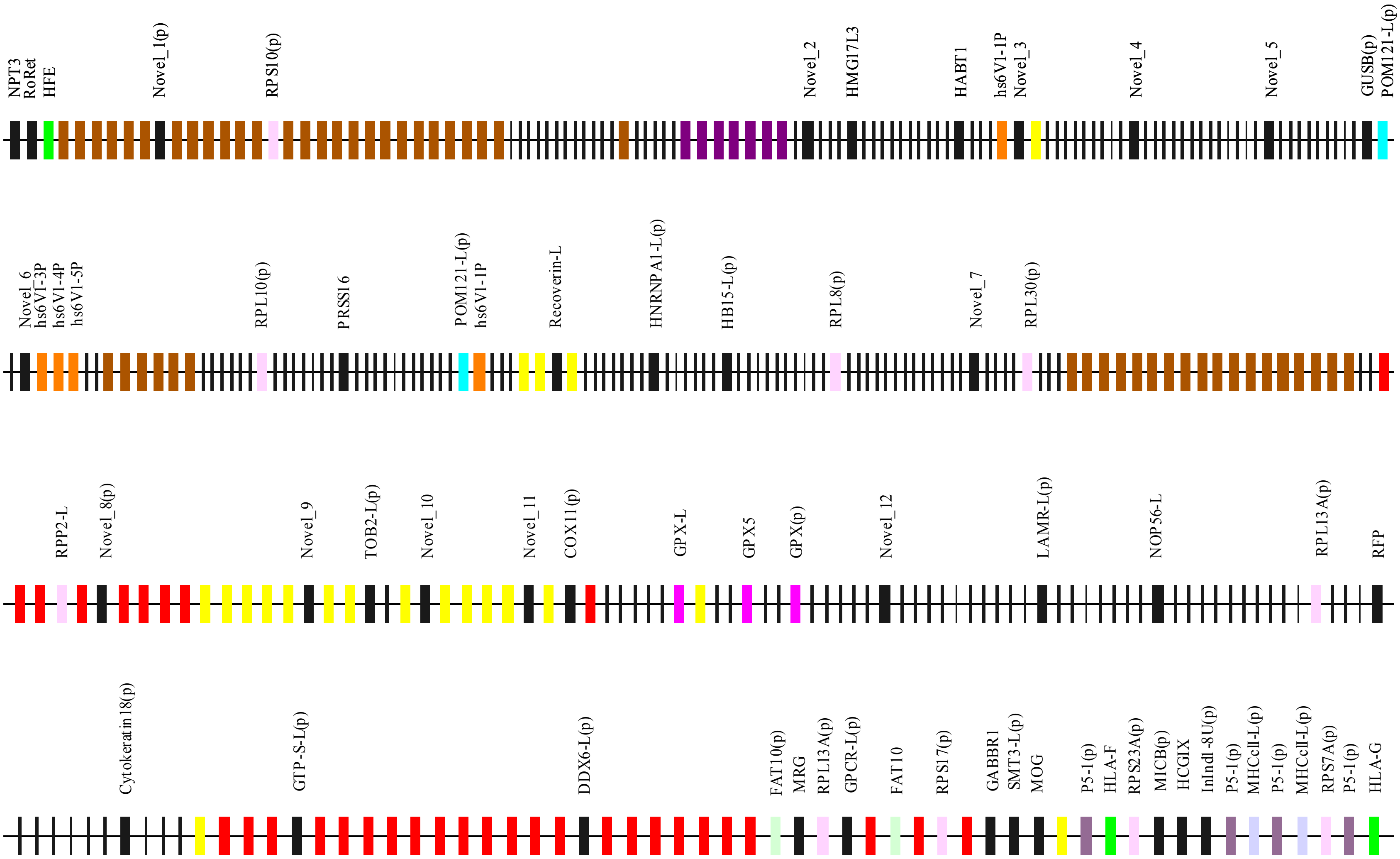


Figure 3.1: Gene content of human MHC extended class I region. 1 Mb of sequence is displayed per track, and genes are coloured according to the family to which they belong: HLA genes (green), histones (brown), ribosomal proteins (light pink), butyrophilins (purple), pheromone receptors (orange), POM121-like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

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

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

3.3.1. The histone cluster

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

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

3.3.2. The zinc finger protein cluster

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

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

3.3.3. The ribosomal protein cluster.

In addition to zinc finger proteins, which have been implicated in the regulation of transcription, and histone genes, which are implicated in allowing transcriptional factors to reach certain areas of chromosome, the MHC extended class I region also contains a number of proteins implicated in the control of translation, the ribosomal proteins (Warner and Nierras, 1998). These ribosomal proteins combine with other ribosomal proteins and 4 species of RNA in order to produce functional ribosomes.

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

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

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

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

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

3.3.4. The butyrophilin cluster.

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

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

3.3.5. The tRNA cluster.

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

tRNA type	No. in extended class I
tRNA-Ala	37
tRNA-Arg	12
tRNA-Asn	1
tRNA-Asp	3
tRNA-Gln	12
tRNA-Glu	1
tRNA-Gly	3
tRNA-His	5
tRNA-Ile	17
tRNA-Leu	11
tRNA-Lys	8
tRNA-Met	22
tRNA-Phe	9
tRNA-Pro	2
tRNA-Ser	18
tRNA-Thr	10
tRNA-Trp	2
tRNA-Tyr	4
tRNA-Val	17

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

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

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

3.4.1. The glutathione peroxidase loci.

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

3.4.2. The FAT 10 gene and pseudogene.

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

3.4.3. The GPCR loci.

In addition to the olfactory receptors and the pheromone receptors which are members of the G-protein coupled receptor superfamily, the human extended class I region contains 3 other genes that are G-protein coupled receptors. Beyond this shared classification, the GTP-SARA-related gene shows little similarity to the other 2 receptors, and the 2 GPCR genes located within 5 KB of sequence (MRG and GPCR-L) also share little similarity on the protein level. This suggests that all 3 loci have different evolutionary histories.

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

3.4.4. The POM121-like loci.

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

3.4.5. The RoRet gene and the Ret Finger Protein gene

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

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

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

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

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

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

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

The pseudogenes found in the extended class I region were compared against functional versions to consider the potential ‘old’ functions of these loci. The HNRNPA1 gene, for example, would code for a heterogeneous nuclear ribonucleoprotein (hnRNPs) (Biamonti *et al.*, 1994). These proteins are a large family of nucleic acid binding proteins that are often found in, but not restricted to, the 40S-ribonucleoprotein particle. HNRNP1A is a polypeptide that appears to be involved in binding nascent hnRNA in the nucleus to form the so called hnRNP complexes which are involved in pre-mRNA processing and in the export of mRNA from the nucleus to the cytoplasm. After exportation in the complex, the hnRNPs are immediately re-imported back into the nucleus (Weighardt *et al.*, 1995). The existence of this pseudogene within the extended MHC is interesting given the number of ribosomal protein pseudogenes within this sequence; it could be suggested that the functional version of this gene formed part of a transcription-translation complex of genes that existed in this region of the genome at one stage of evolution.

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

The TOB2-L pseudogene is classified according to its similarity to the gene located on chromosome 22 that codes for the TOB2 protein. TOB (Transducer of ErbB-2) proteins interact with the c-erbB-2 gene product p185erbB2 and they are considered to have a role in negatively regulating cell proliferation (Matsuda *et al.*, 1996). The interactions with p185 could negatively regulate the TOB-mediated antiproliferative pathway, resulting possibly in growth stimulation by p185. TOB has been found to be expressed in primary peripheral blood T lymphocytes and it has to be downregulated for T-cell activation (Tzachanis *et al.*, 2001). TOB2 is also involved in the negative regulation of cell proliferation, inhibiting cell cycle progression from the G0/G1 to S phase of the cell cycle. A high level of expression of TOB2 in oocytes suggested a role for TOB2 in oogenesis (Ikematsu *et al.*, 1999).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

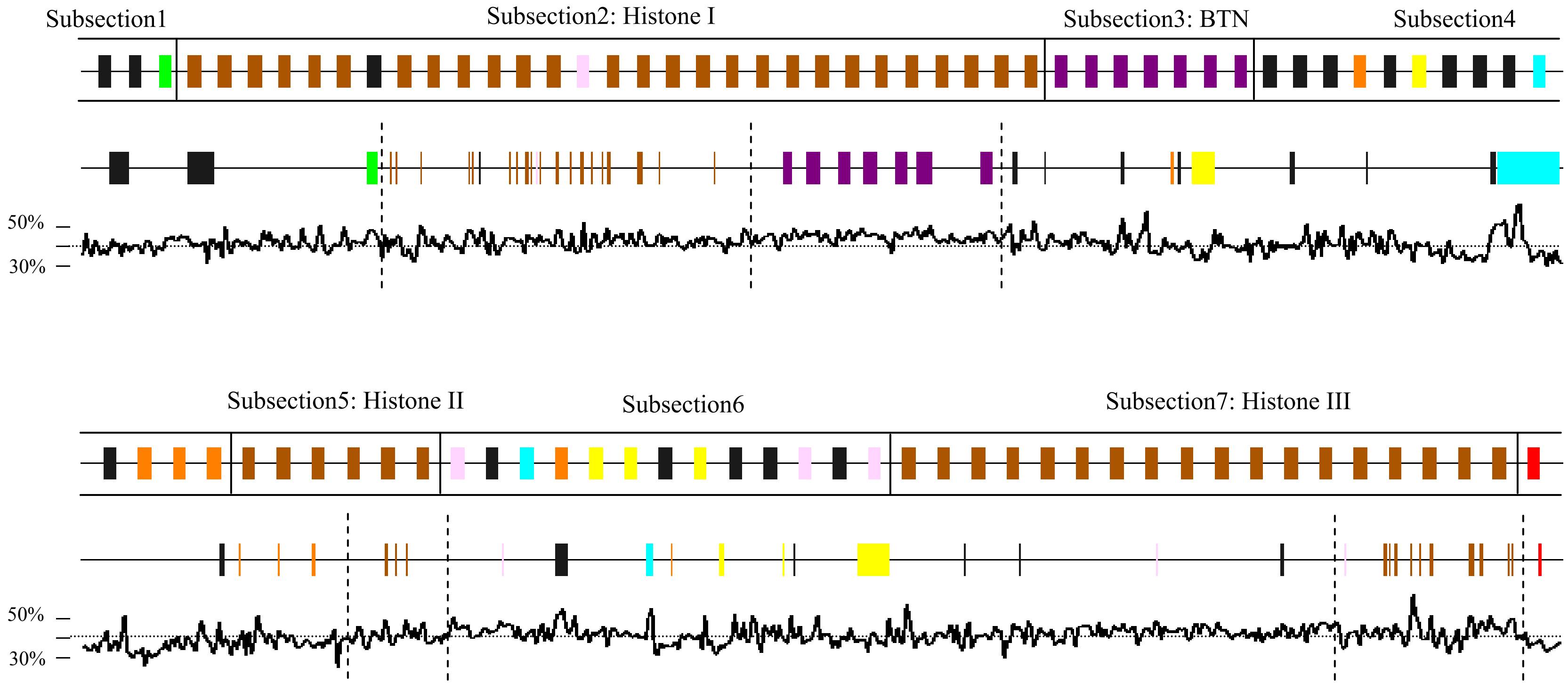
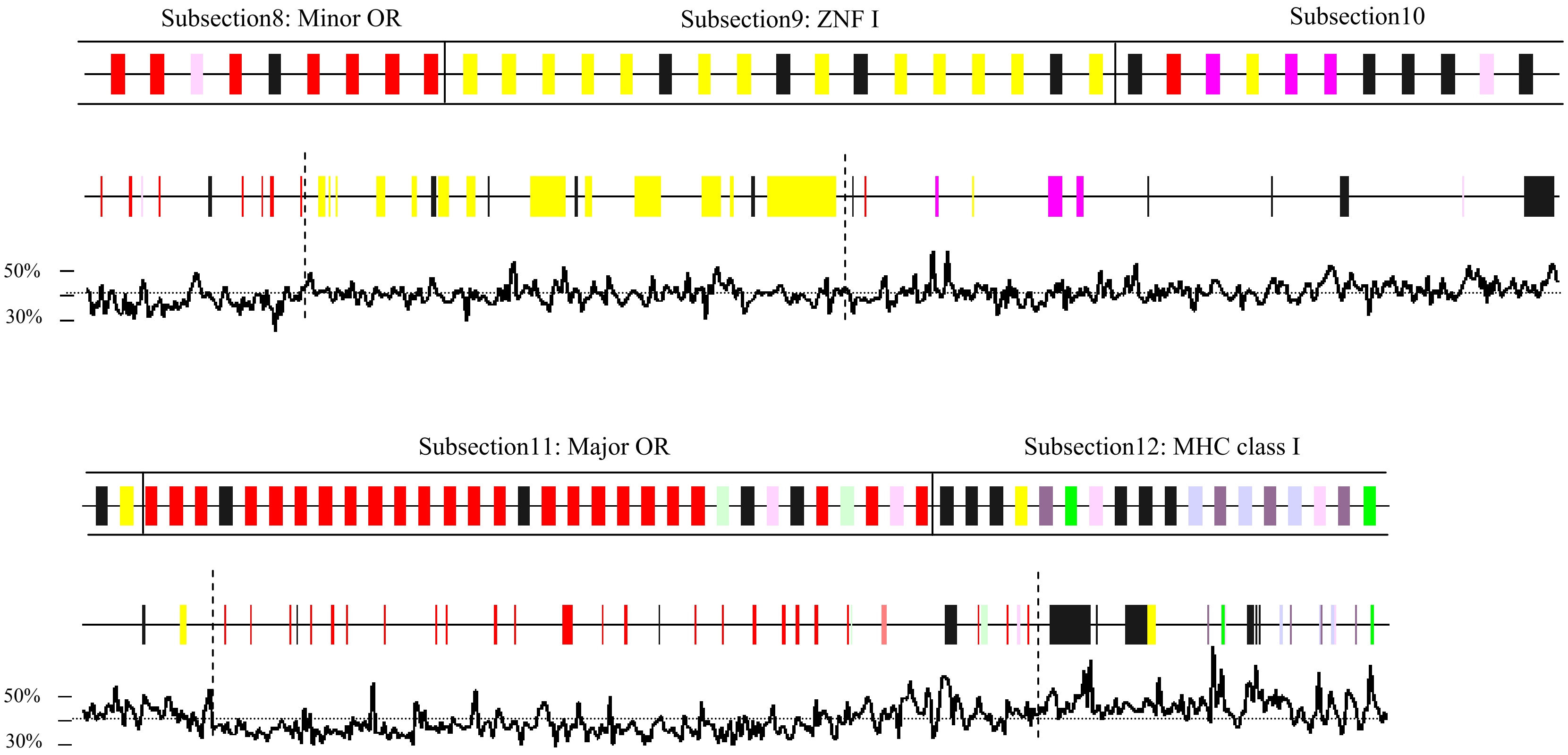


Figure 3.2 (cont. overleaf): Schematic diagram showing subsections of the extended MHC class I. The 3913358 bp sequence has been divided into 12 subsections, defined according to majority gene content. The first track shows a schematic represenation of the genes, for names refer to Figure 3.1. The second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


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

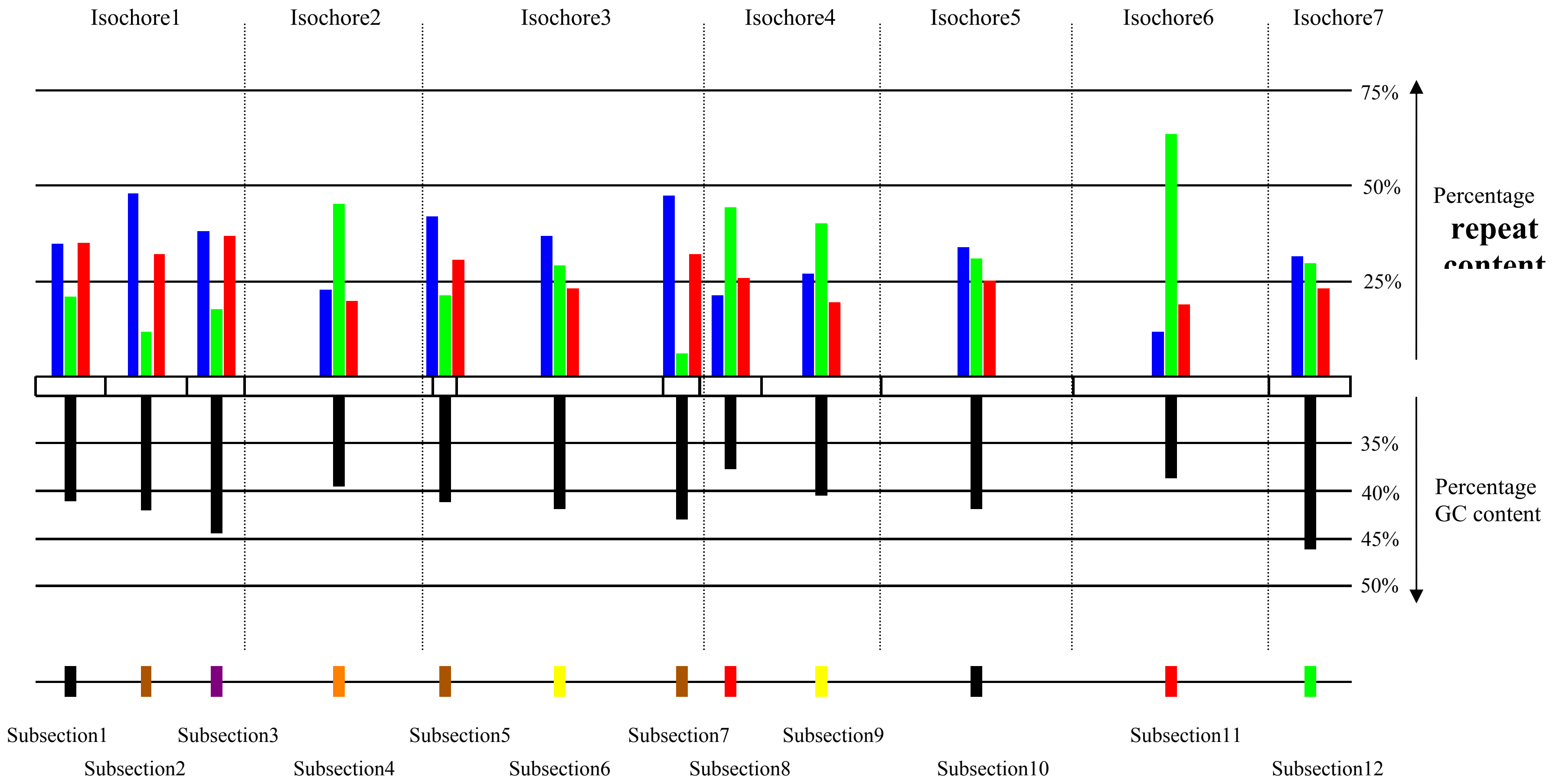


Figure 3.3: Isochores of the MHC extended class I region. Schematic diagram showing sequence containing the genes from HFE to HLA-G divided into 12 subsections, defined according to gene content. Each subsection is plotted to scale, showing how much of the sequence is composed of the various clusters.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

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

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

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

Across the sequence as a whole, an analysis of the repeat content (figure 3.3) and GC content (figure 3.2) suggests the sequence can be divided into 7 domains that may represent isochores (figure 3.3, indicated by dotted lines). Isochores were defined in 1976 (Macaya *et al.*, 1976) (although they were named in 1981 (Cuny *et al.*, 1981)) as long regions of DNA (longer than 300 Kb) that are fairly homogeneous in terms of base composition compared to the heterogeneity present in other (non-satellite) DNA in the human genome. The idea was formalised by Bernardi *et al.* (1985) who suggested that warm-blooded vertebrates had a 'mosaic' genome consisting of 5 distinct types of isochore (L1, L2, H1, H2 and H3 with GC contents of <38%, 38-42%, 42-47%, 47-52% respectively.) In contrast to this view of the vertebrate genome, the draft sequence paper published by the International Human Genome Sequencing Consortium (IHGSC, 2001) suggested there was no evidence of the existence of compositional homogeneous isochores, and suggested that owing to the heterogeneity of the genome, "isochores do not appear to deserve the name 'iso.'" In response to this Bernardi *et al.* (2001) argued that the genome does contain large regions of distinctive GC content that can be used to partition chromosomes, since the denial of a compositionally discontinuous sequence organization results in the denial of "a fundamental level of genome organization" (Eyre-Walker and Hurst, 2001).

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

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

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

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

Alternatively, some type of excision process may operate to remove certain types of repeat from certain isochores. For example, isochores may be low in SINEs because SINEs in this isochore are excised by an enzymatic mechanism. This mechanism may also act to disrupt functional genes by excising valuable sequence around the unwanted repeat. (This is something that is not supported by the higher percentage of pseudogenes in low-SINE, high-LINE areas of the MHC extended class I region). In spite of this potential disruption of functional units, however, negative selectional pressures are favoured as the mechanism by a number of authors who suggest that Alu sequences have not been fixed within human populations for a long enough period of time for positive selection to act upon these sequences (Brookfield, 2001).

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

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

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

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

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

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

3.8. Duplications within the human MHC extended class I region

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

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

3.9. Conclusions.

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

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

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

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.

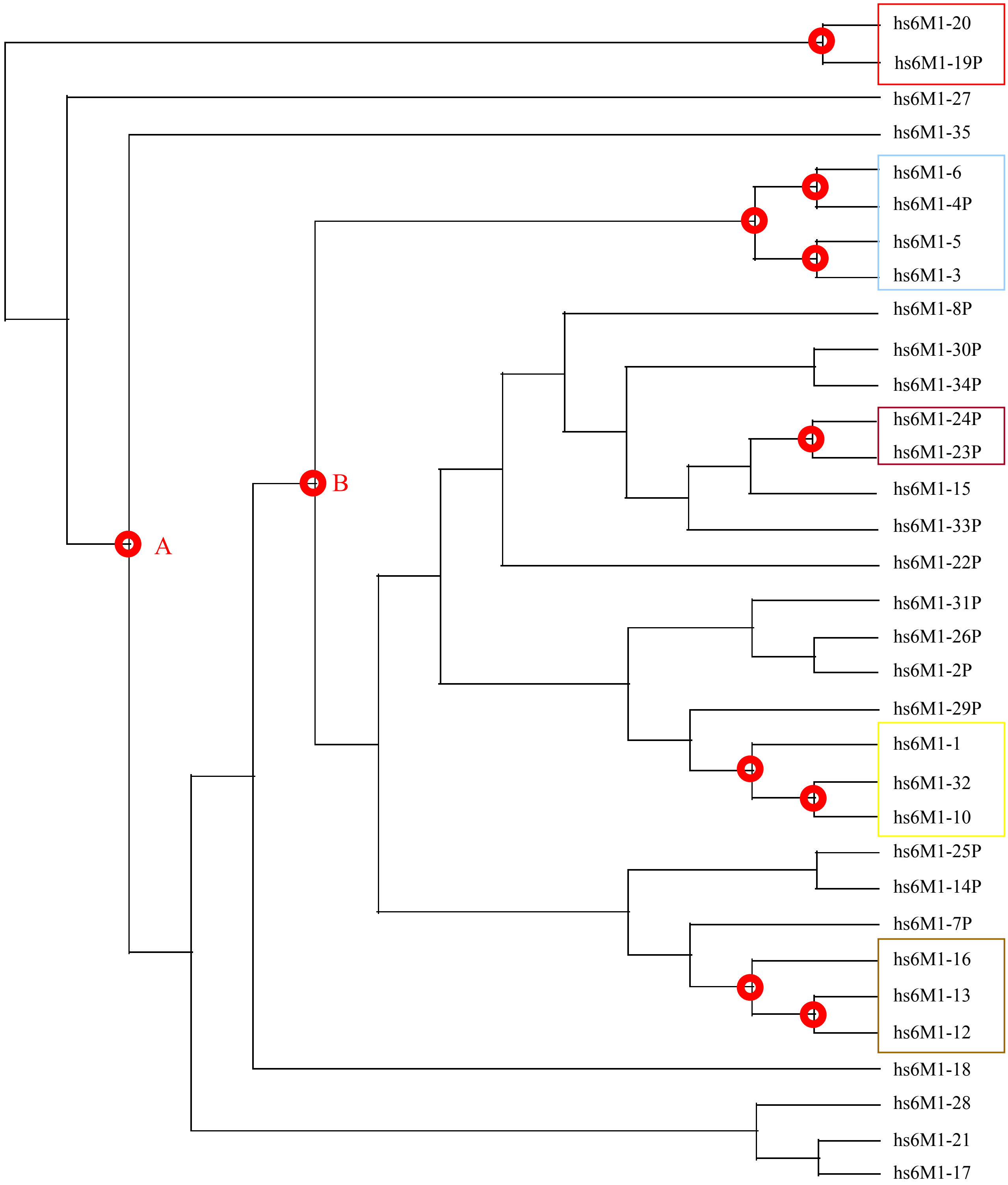


Figure 4.1: Phylogenetic tree (parsimony method) of human MHC-linked OR genes. 175 sites were used and 250 bootstrap replicates were performed. Subfamilies are boxed in different colours, whilst the red rings at branch points indicate where bootstrap values are over 70%.

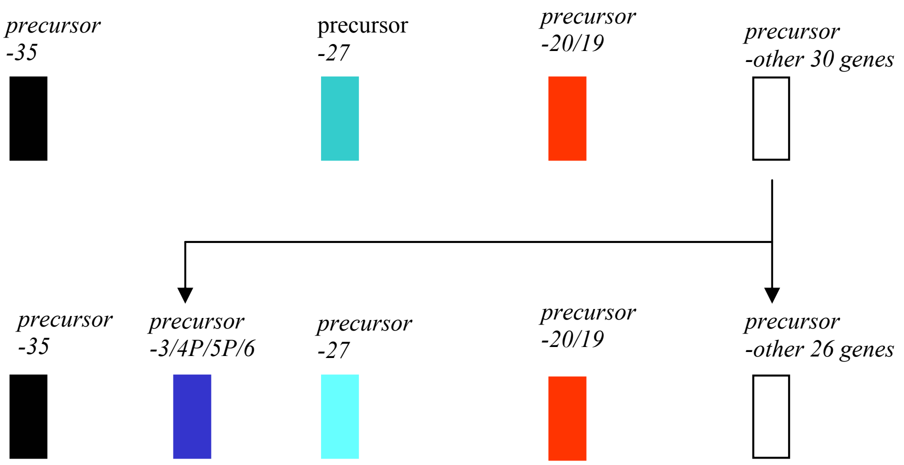


Figure 4.2: A proposed model of evolution for the MHC-linked OR genes, based on significant branch points from the phylogenetic tree (figure 4.1).

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

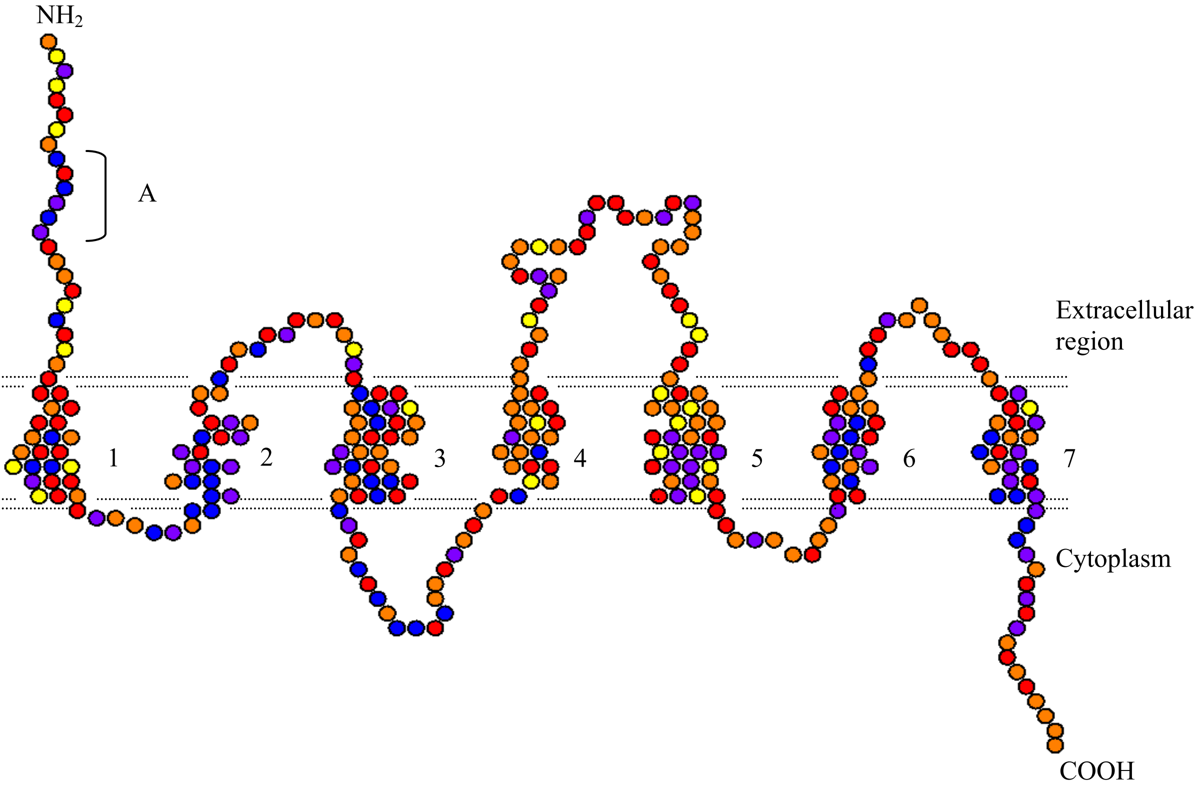


Figure 4.3: Amino acid conservation in human MHC-linked ORs. Schematic diagram showing the conservation of amino acids at predicted positions within a consensus human MHC-linked olfactory receptor protein. The degree of conservation ranges from 90%+ (blue), 75-90% (purple), 50-75% (red), 25-50% (orange)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

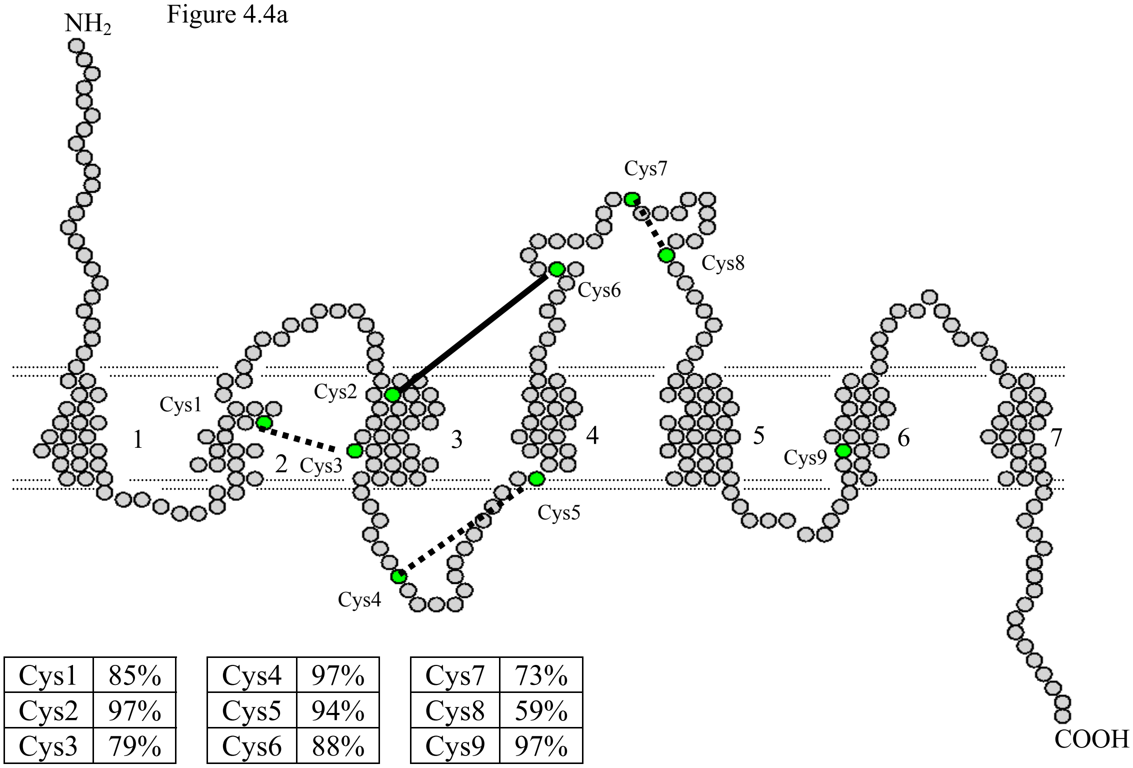
nuclear magnetic resonance (NMR) and x-ray crystallography to transmembrane proteins. These difficulties stem from the hydrophobic nature of transmembrane proteins: they do not easily dissolve, and any structure obtained in solution is likely to be different from the structure the protein adopts in the 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%
Cys2	97%
Cys3	79%

Cys4	97%
Cys5	94%
Cys6	88%

Cys7	73%
Cys8	59%
Cys9	97%

Figure 4.4b

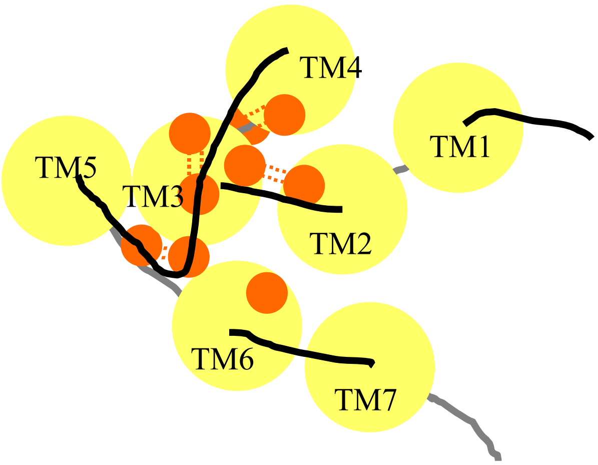


Figure 4.4a. Cysteine conservation in the human MHC-linked ORs. Schematic diagram showing where cysteines are conserved within a consensus human MHC-linked olfactory receptor protein. The percentage of OR proteins containing a cysteine in this position is shown in the table underneath. The disulphide 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.

30P L G L G W Q
ctg ggc ct g gg tgg caa

ttg gga ctc g ggg gga gtg

31P L G L G G V

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

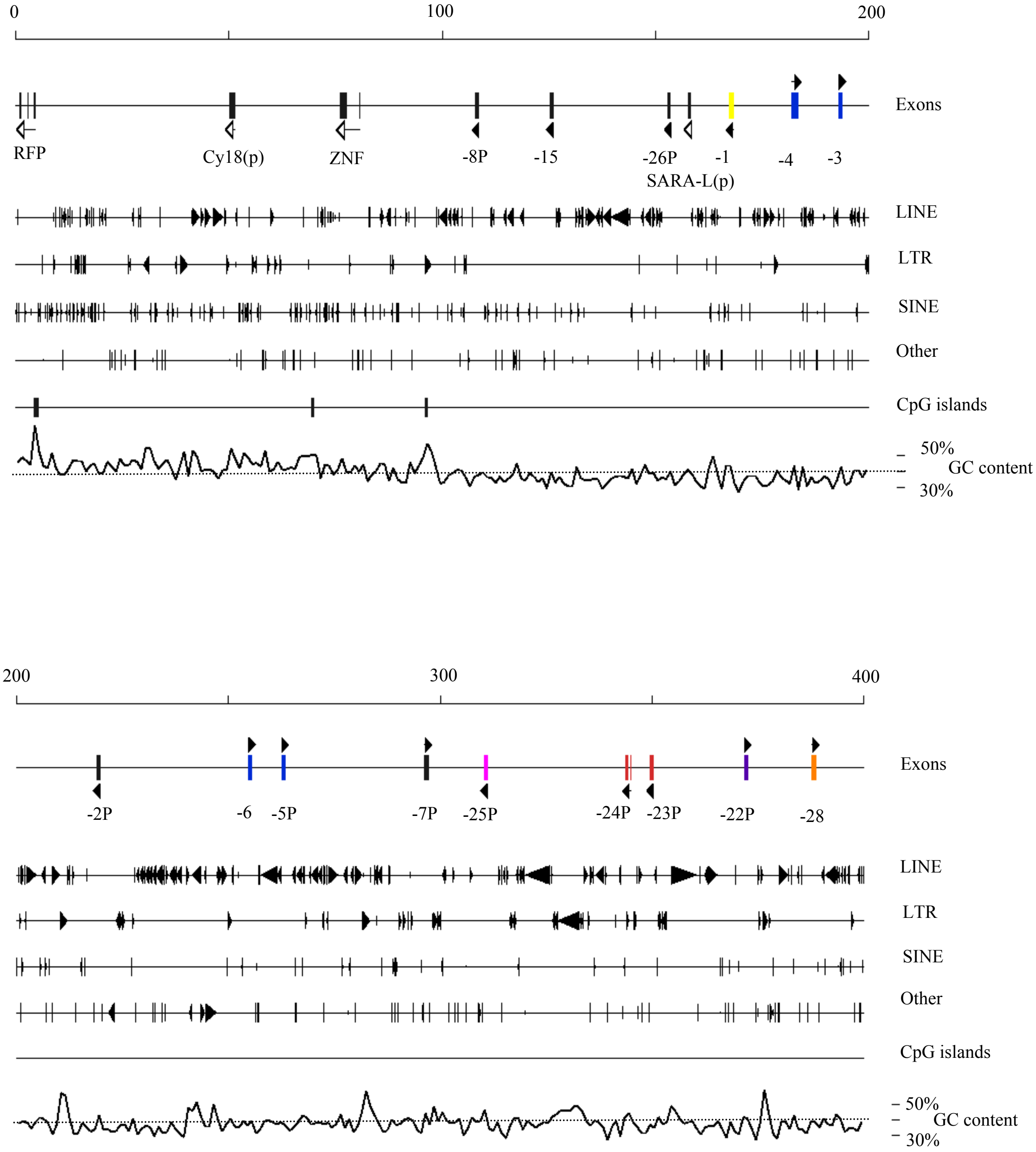
Figure 4.6 shows the genomic environment of the human MHC-linked olfactory receptor genes in the major cluster and the minor cluster. In contrast to some OR clusters, where it was been reported that OR sections of the genome form an exclusive environment, containing no other non-OR genes (Glusman *et 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).

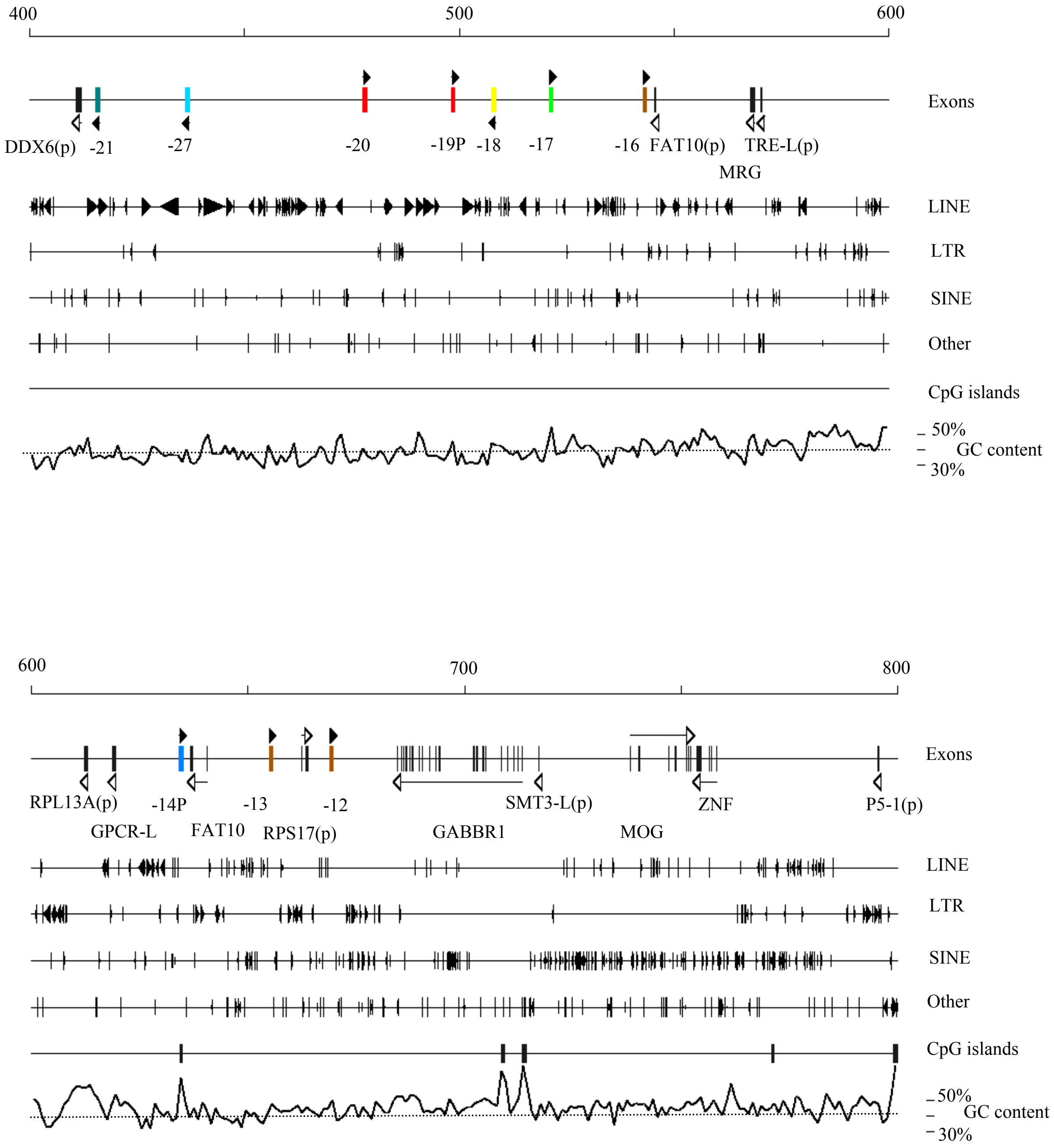
As has been previously discussed (Chapter 3), the human MHC-linked OR genes are located in distinct isochores associated with a low GC content (generally the GC content does not rise above 40%) and a high number of long interspersed nuclear elements (LINEs). This trend is particularly pronounced within 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
SARA-L(p)
-4
-3
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)
GPCR-L
-14P
FAT10
-13
RPS17(p)
-12
GABBR1
SMT3-L(p)
MOG
ZNF
P5-1(p)
LINE
LTR
SINE
Other
CpG islands
50%
GC content
30%

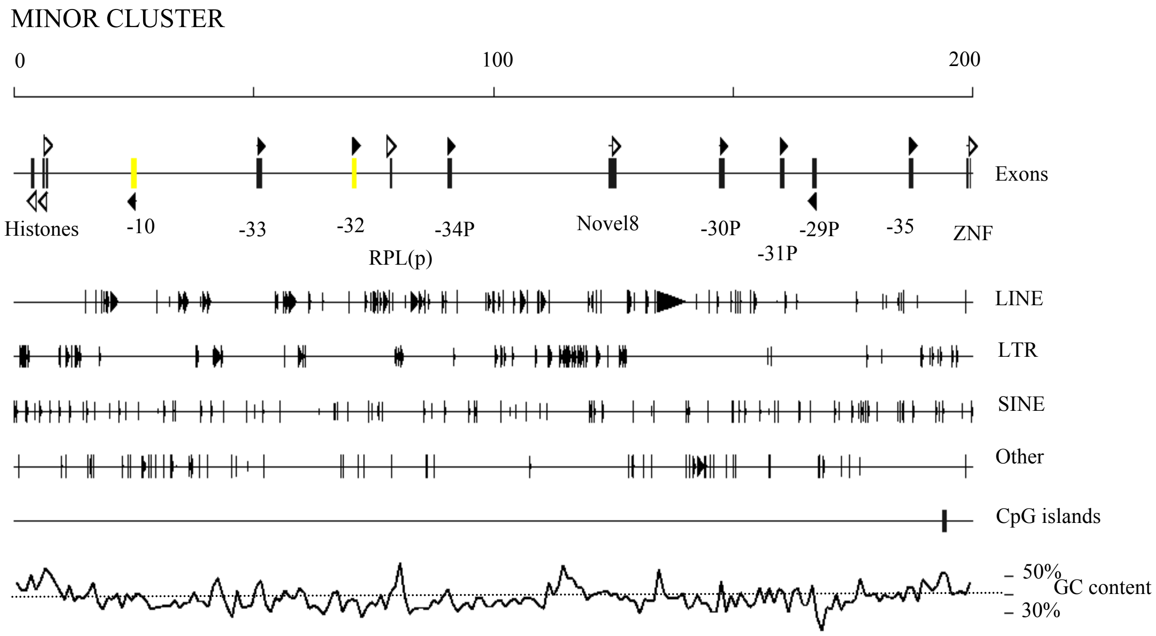


Figure 4.6: The genomic organisation of the human MHC-linked major and minor OR clusters. Predicted exons are shown as boxes on the first track. The orientation of genes is indicated by an arrow either above or below the gene, with a line indicating exons that belong to the same gene. OR genes are indicated 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).

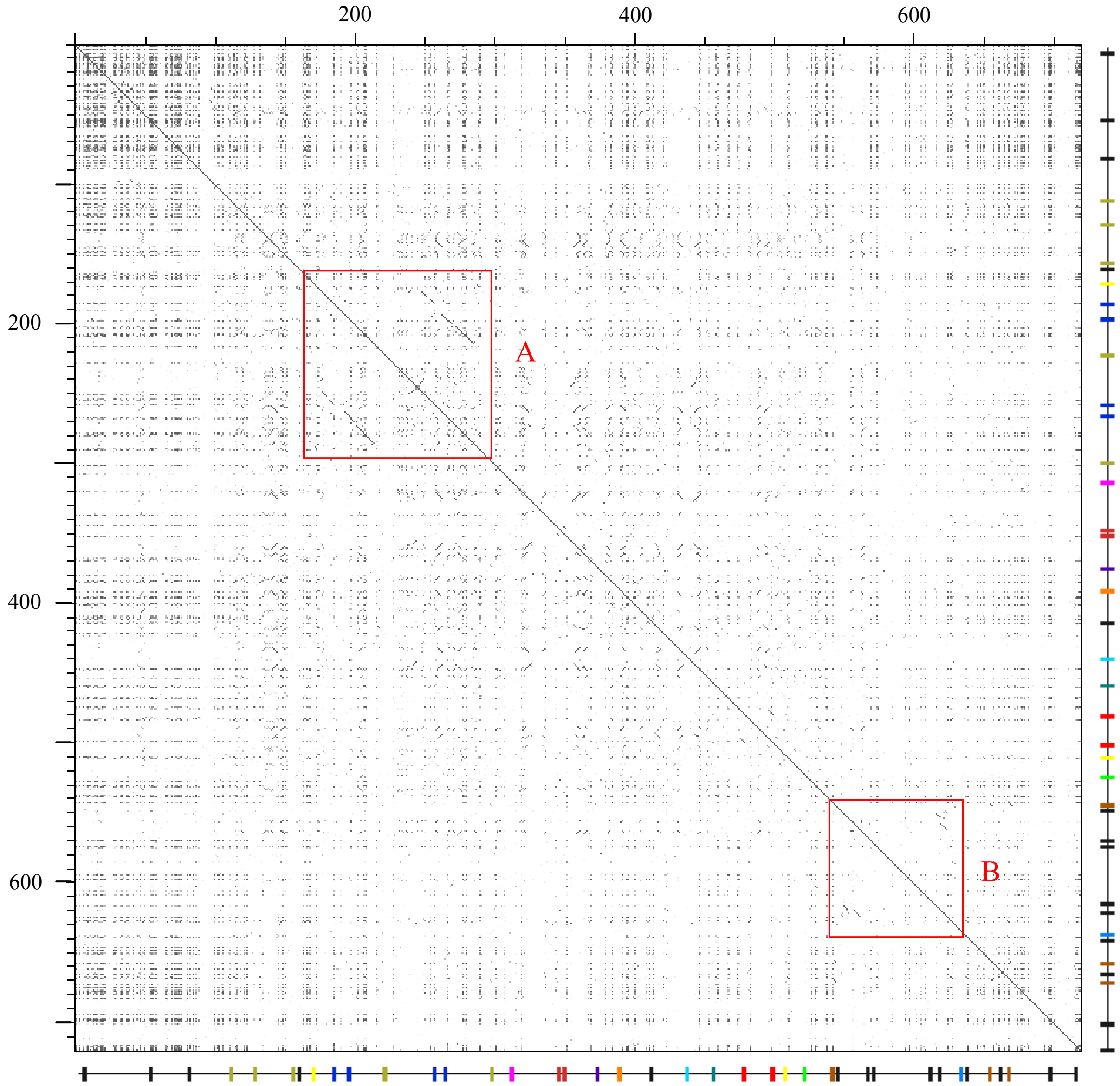


Figure 4.7: A dot-matrix plot showing the major OR cluster plotted against itself. The gene line shows the positions of exons within the sequence. OR subfamilies are shown in colours corresponding to those in figure 4.6. Other ORs are indicated by the light green colour, with non-ORs black.

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.

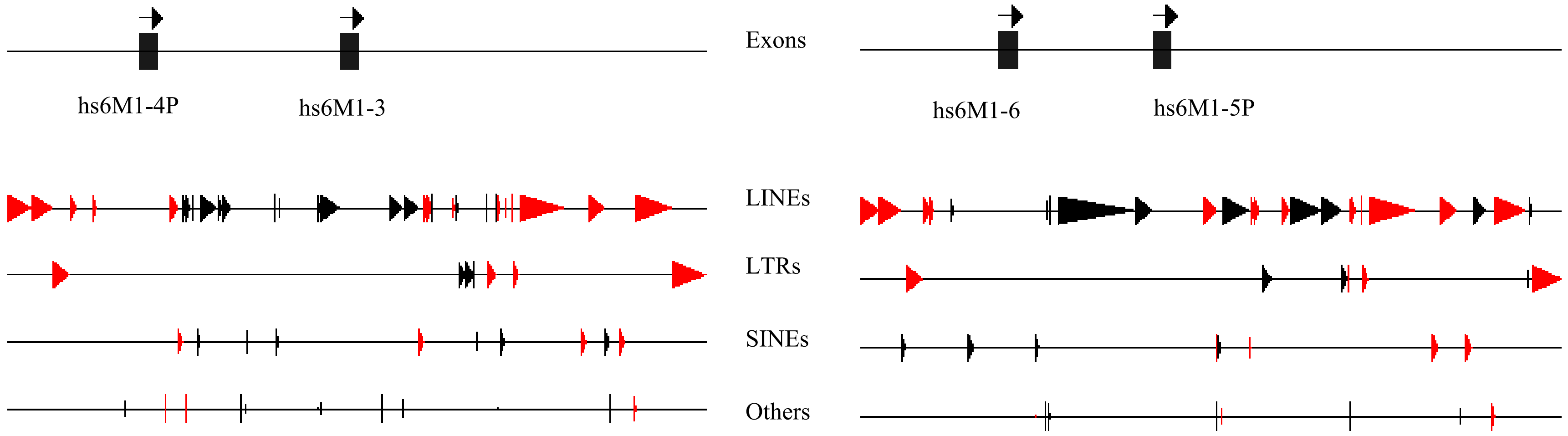


Figure 4.8: Local duplication within the major MHC-linked OR cluster creating 2 new OR genes. The block that has duplicated to create 4 OR genes from 2 original OR loci is shown above. Pairs of OR genes are 96% identical on the DNA level, suggesting this was a recent evolutionary event. Repeats conserved 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.

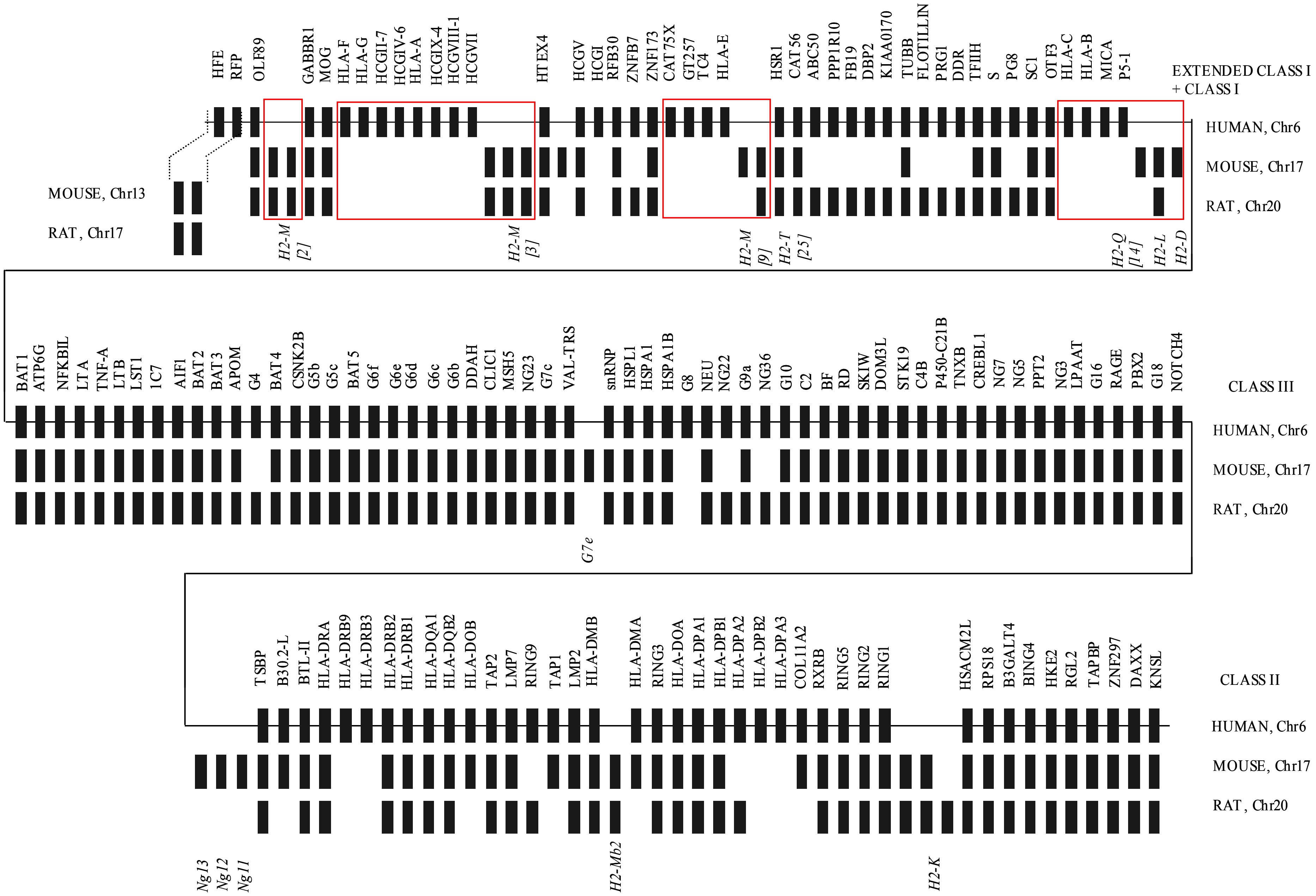
Chapter 5

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).

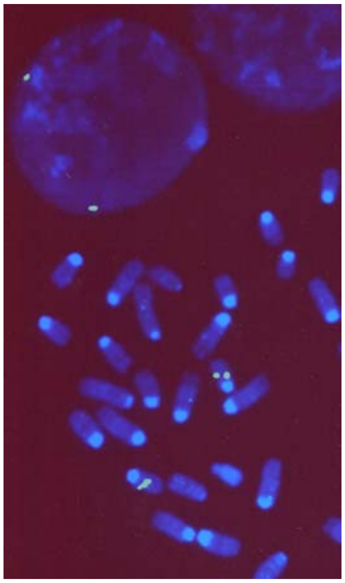

EXTENDED CLASS I
+ CLASS I
HFE
RFP
OLF89
GABBR1
MOG
HLA-F
HLA-G
HCGII-7
HCGIV-6
HLA-A
HCGIX-4
HCGVIII-1
HCGVII
HTEX4
HCGV
HCGI
RFB30
ZNF87
ZNF173
CAT75X
GT257
TC4
HLA-E
HSR1
CAT56
ABC50
PPP1R10
FB19
DBP2
KIAA0170
TUBB
FLOTILLIN
PRG1
DDR
TFIIH
S
PG8
SC1
OTF3
HLA-C
HLA-B
MICA
P5-1
HUMAN, Chr6
MOUSE, Chr17
RAT, Chr20
MOUSE, Chr13
RAT, Chr17
H2-M [2]
H2-M [3]
H2-M [9]
H2-T [25]
H2-Q [14]
H2-L
H2-D
CLASS III
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
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
HUMAN, Chr6
MOUSE, Chr17
RAT, Chr20
G7e
CLASS II
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
HLA-DMA
RING3
HLA-DOA
HLA-DPA1
HLA-DPB1
HLA-DPA2
HLA-DPB2
HLA-DPA3
COL11A2
RXRB
RING5
RING2
RING1
HSACM2L
RPS18
B3GALT4
BING4
HKE2
RGL2
TAPBP
ZNF297
DAXX
KNSL
HUMAN, Chr6
MOUSE, Chr17
RAT, Chr20
Ng13
Ng12
Ng11
H2-Mb2
H2-K

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).



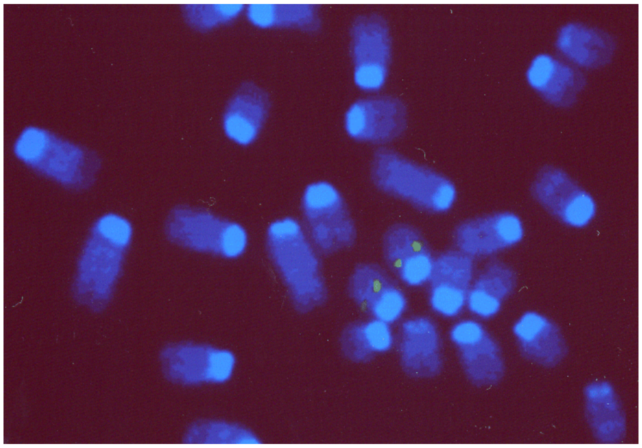
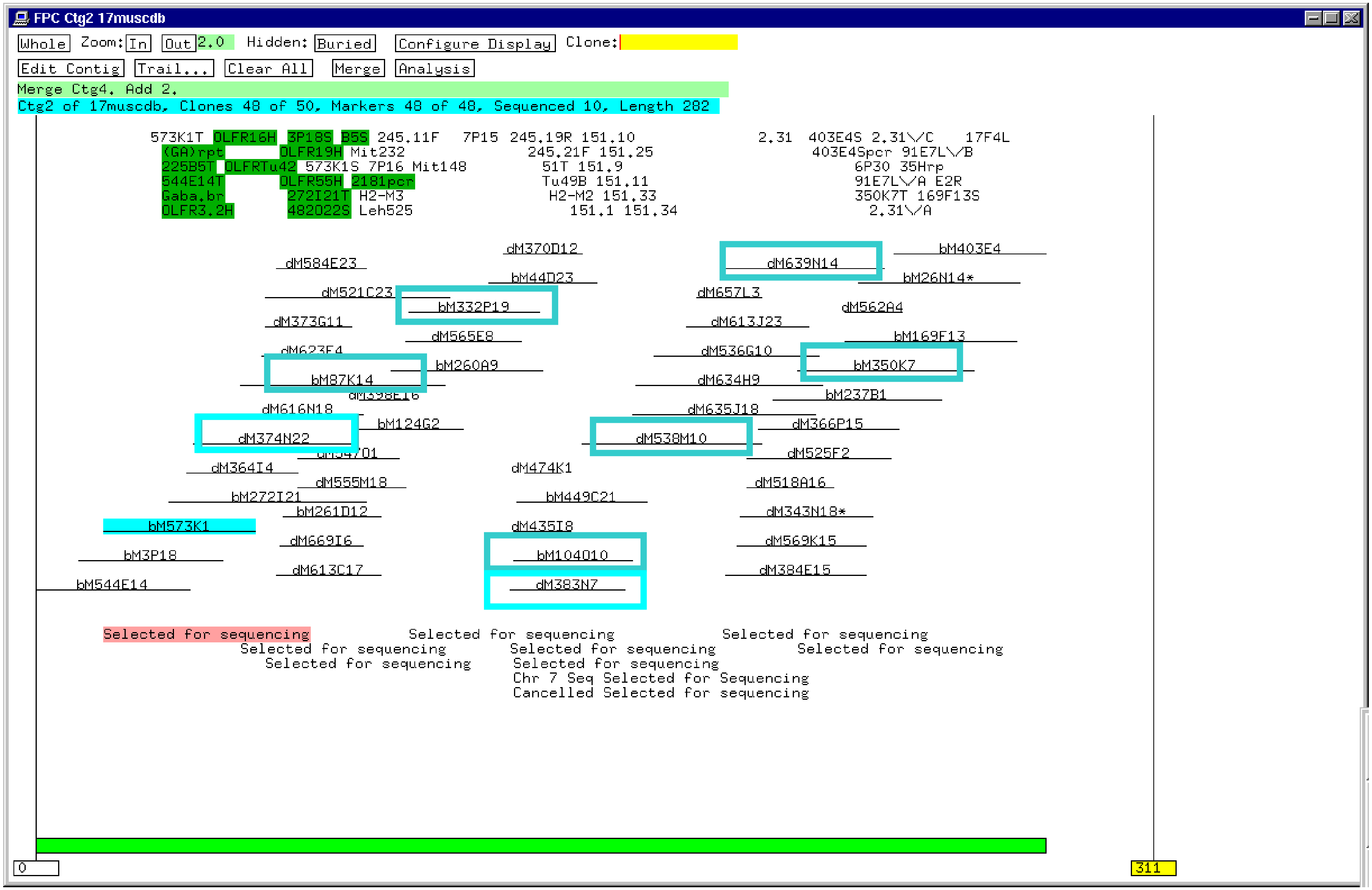


Figure 5.2: FISH analysis of bM332P19. This analysis mapped clone bM332P19 to chromosome 17 in the region C-D

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.



Merge Ctg4. Add 2.

Ctg2 of 17muscdb. Clones 48 of 50. Markers 48 of 48. Sequenced 10. Length 282

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

Selected for sequencing

Selected for sequencing

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.

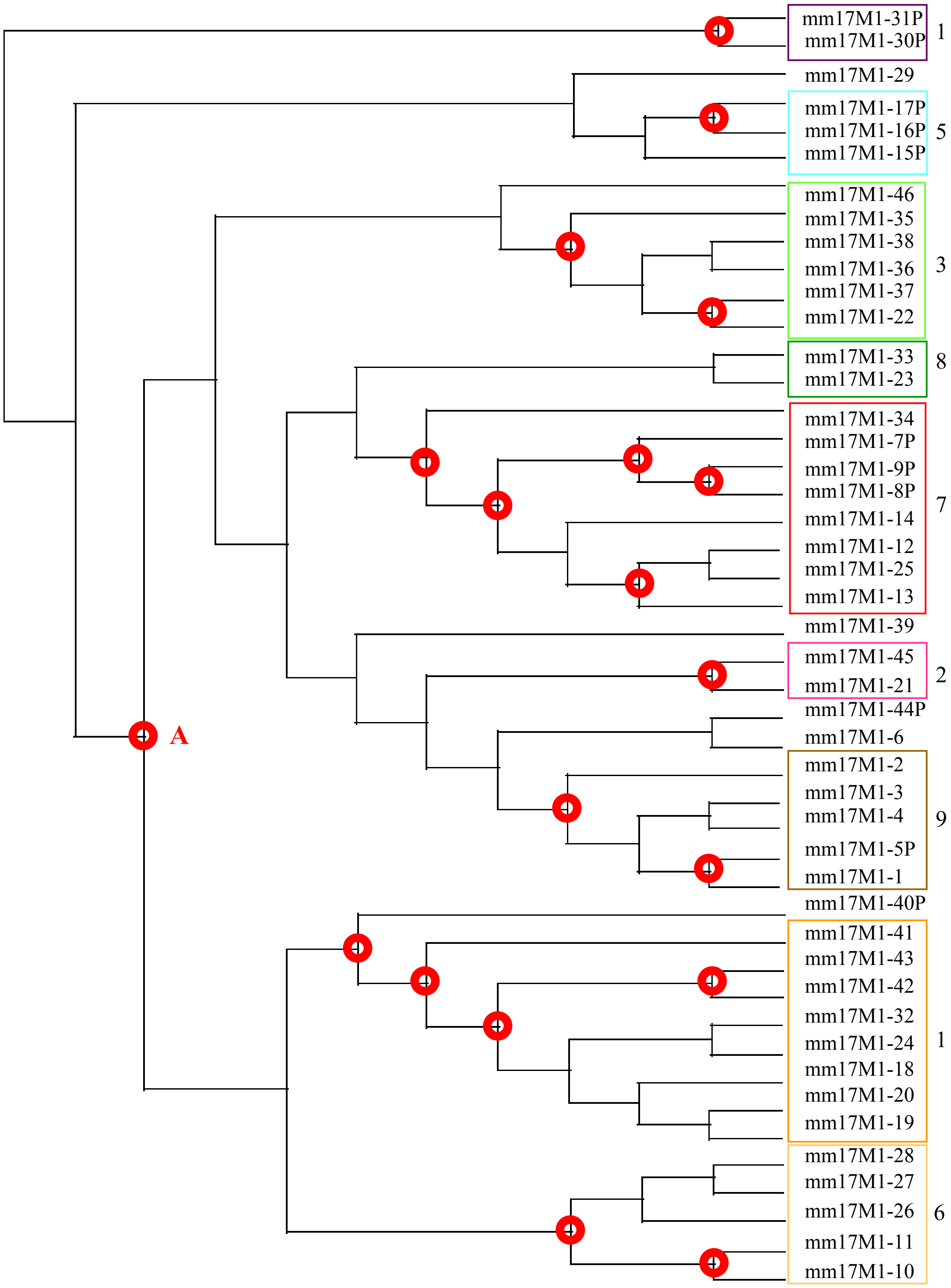
Subfamily	Genes in subfamily					
1	mm17M1-43	mm17M1-42	mm17M1 –41	mm17M1-32	mm17M1-24	mm17M1-18
	mm17M1-19	mm17M1-20				
2	mm17M1-45	mm17M1-21				
3	mm17M1-38	mm17M1-37	mm17M1-36	mm17M1-35	mm17M1-22	mm17M1-46
4	mm17M1-23	mm17M1-33				
5	mm17M1-17P	mm17M1-16P	mm17M1-15P			
6	mm17M1-10	mm17M1-11	mm17M1-28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

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.

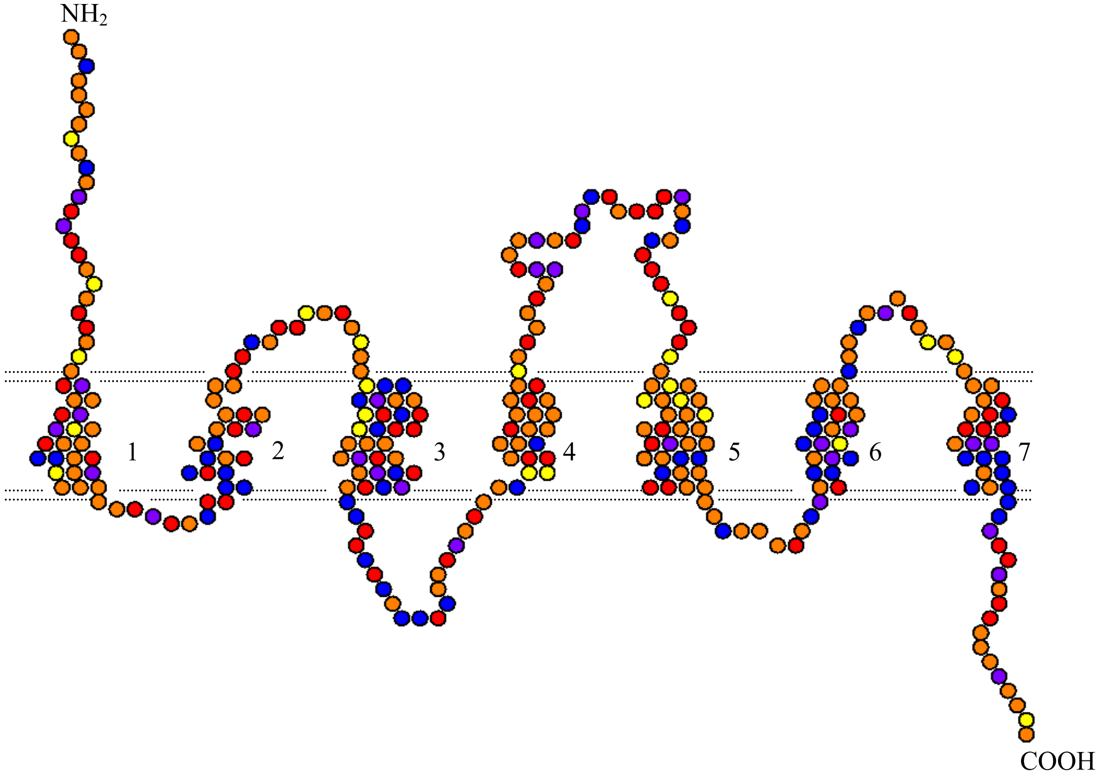


Figure 5.5: Conserved amino acids in mouse MHC-linked OR proteins. Schematic diagram showing the conservation of amino acids at predicted positions within a consensus mouse MHC-linked olfactory receptor protein. The degree of conservation ranges from 90%+ (blue), 75-90% (purple), 50-75% (red), 25-50%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

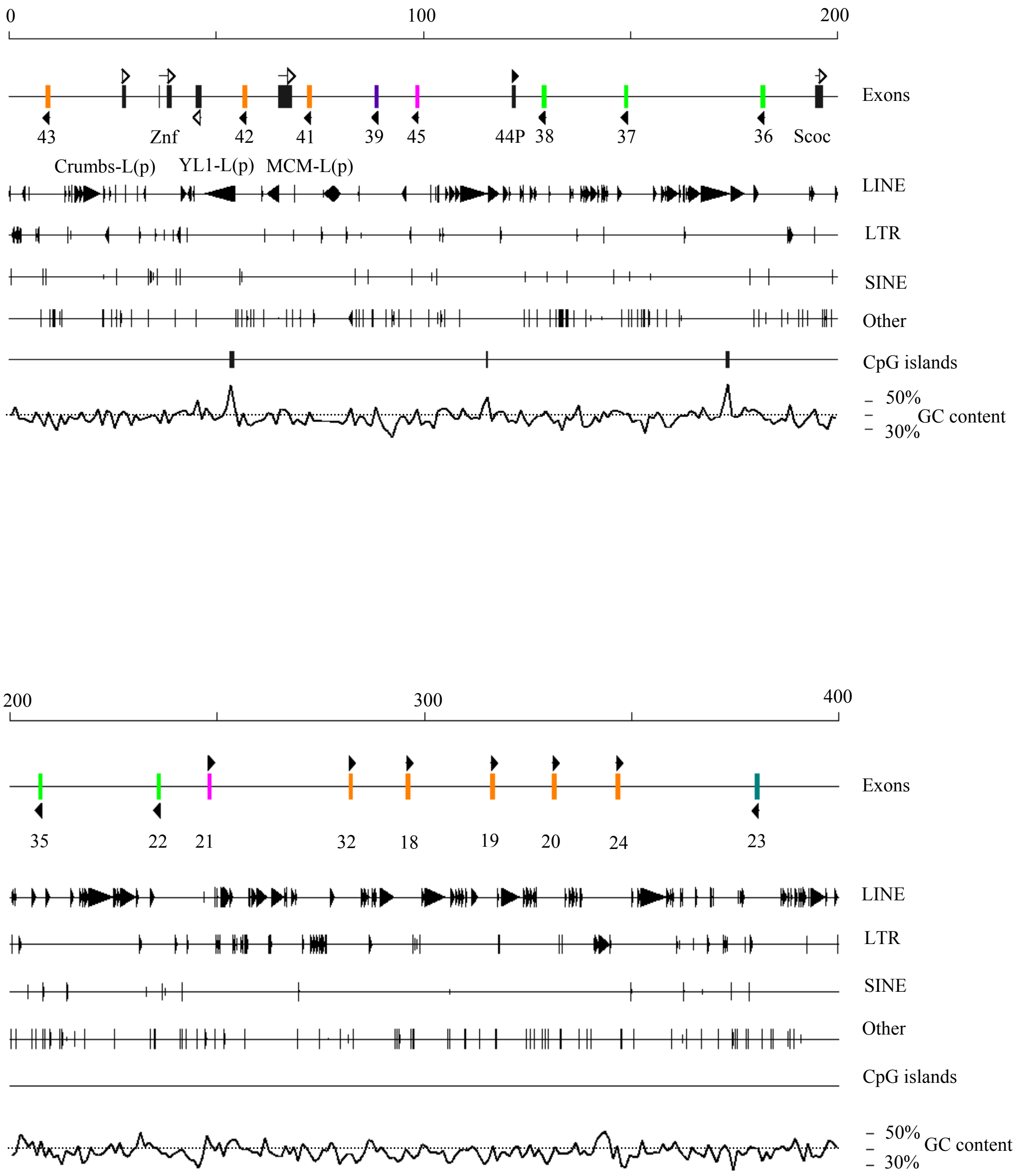
```
S   A   V     L   V  C
tct gct gtc c tta gtt tgc     mm17M1-30P

 F   A        L   S   K
ttt gct       c ctc tcc aag    mm17M1-16/17P
```

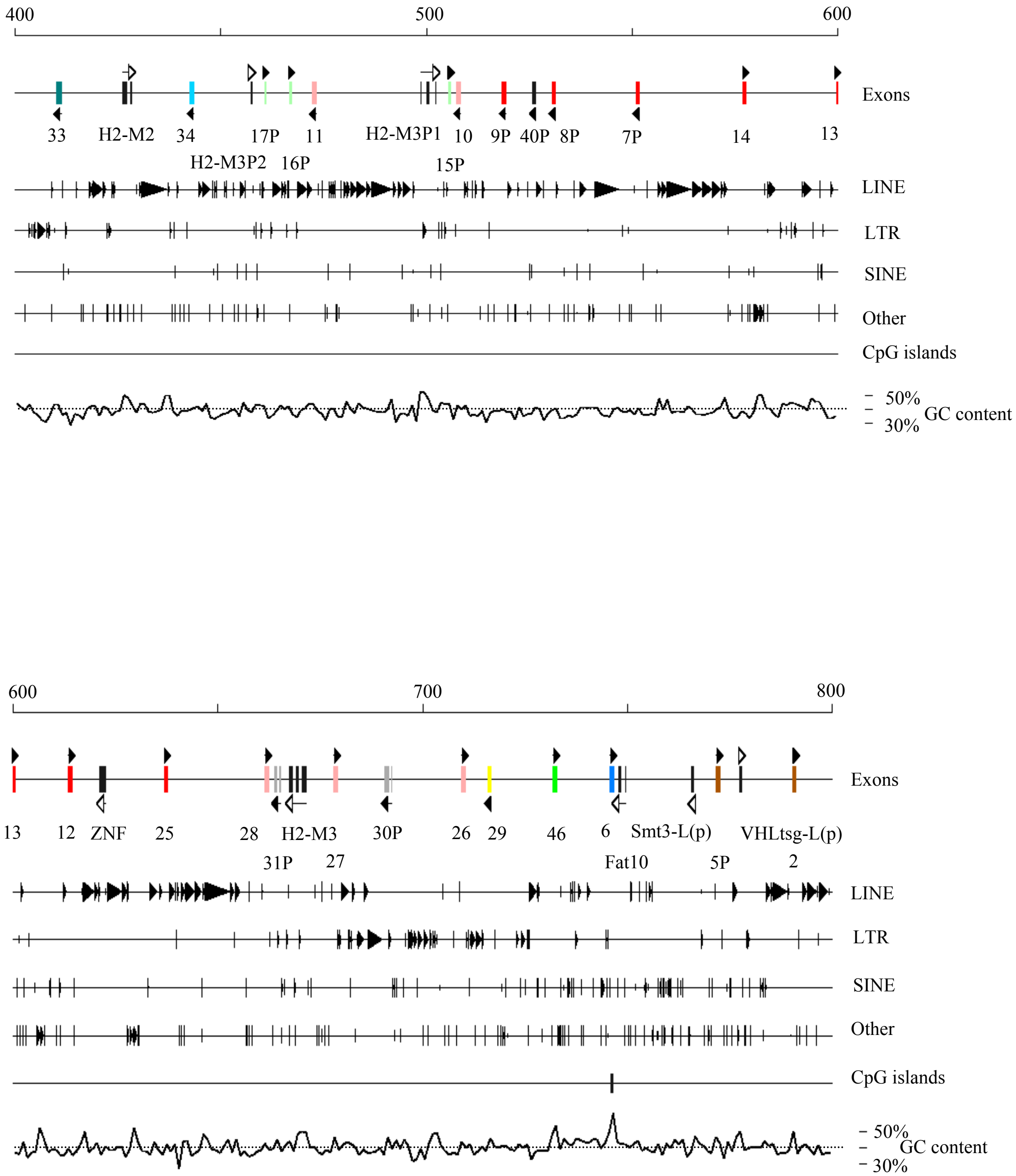
Figure 5.6: A base insertion in mouse pseudogenes mm17M1-30P and mm17M1-16/17P. An insertion of a cytosine has occurred in a similar position in both the mm17M1-30P and mm17M1-16/17P pseudogene, causing a frameshift.

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


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%

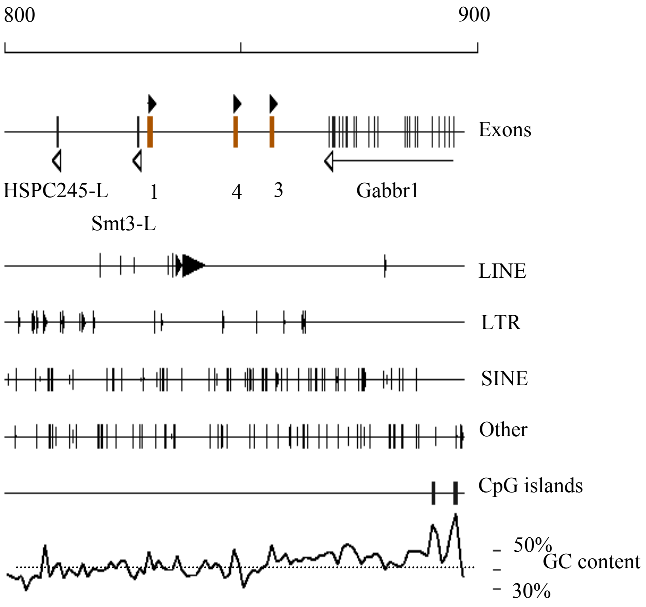


Figure 5.7: The genomic organisation of the mouse MHC-linked OR cluster. Predicted exons are shown as boxes on the first track. The orientation of the gene is indicated by an arrow either above or below the gene, with the line indicating exons that belong to the same gene. OR genes are indicated by filled 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%).

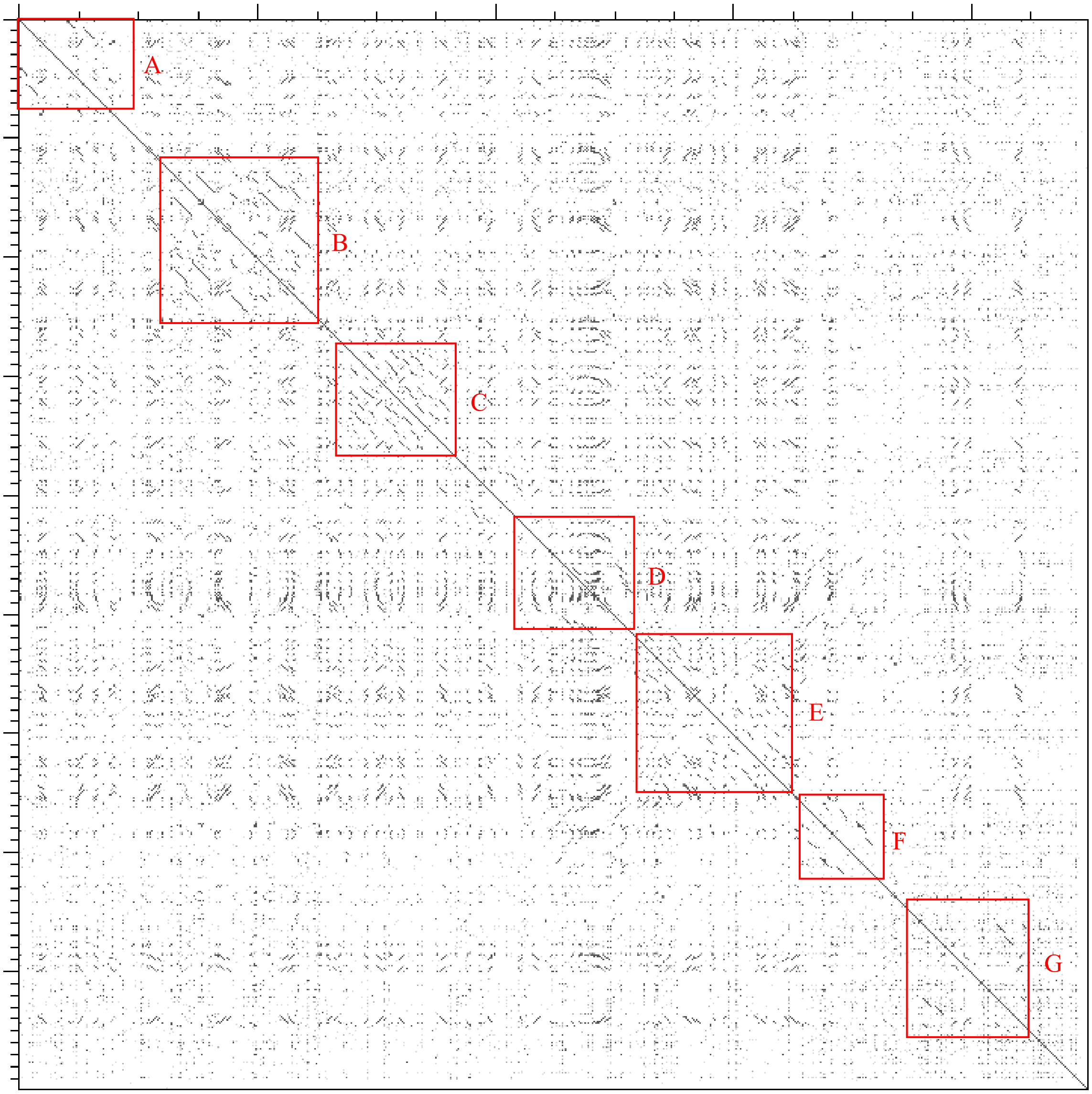


Figure 5.8: Dot-matrix plot of the mouse OR contig. The 897213 bp region was plotted against itself, revealing several duplications indicated by the red boxes which are indicated by a designated letter.

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.

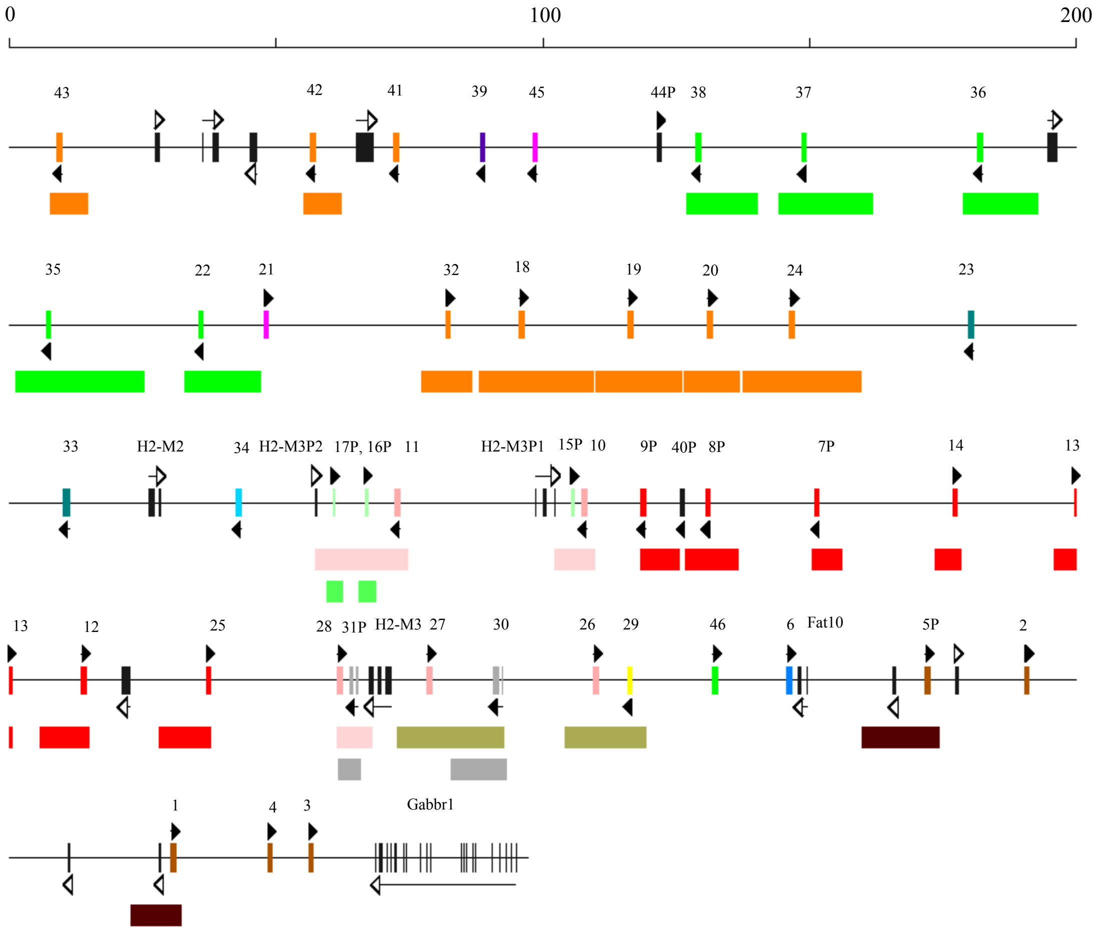


Figure 5.9: Block duplications within the mouse MHC-linked OR cluster. OR genes are named, and the colours represent the subfamily to which the OR genes belong (referred to previously in Table 5.3). Putative blocks are coloured according to the OR genes they are involved in duplicating, with the exception 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.

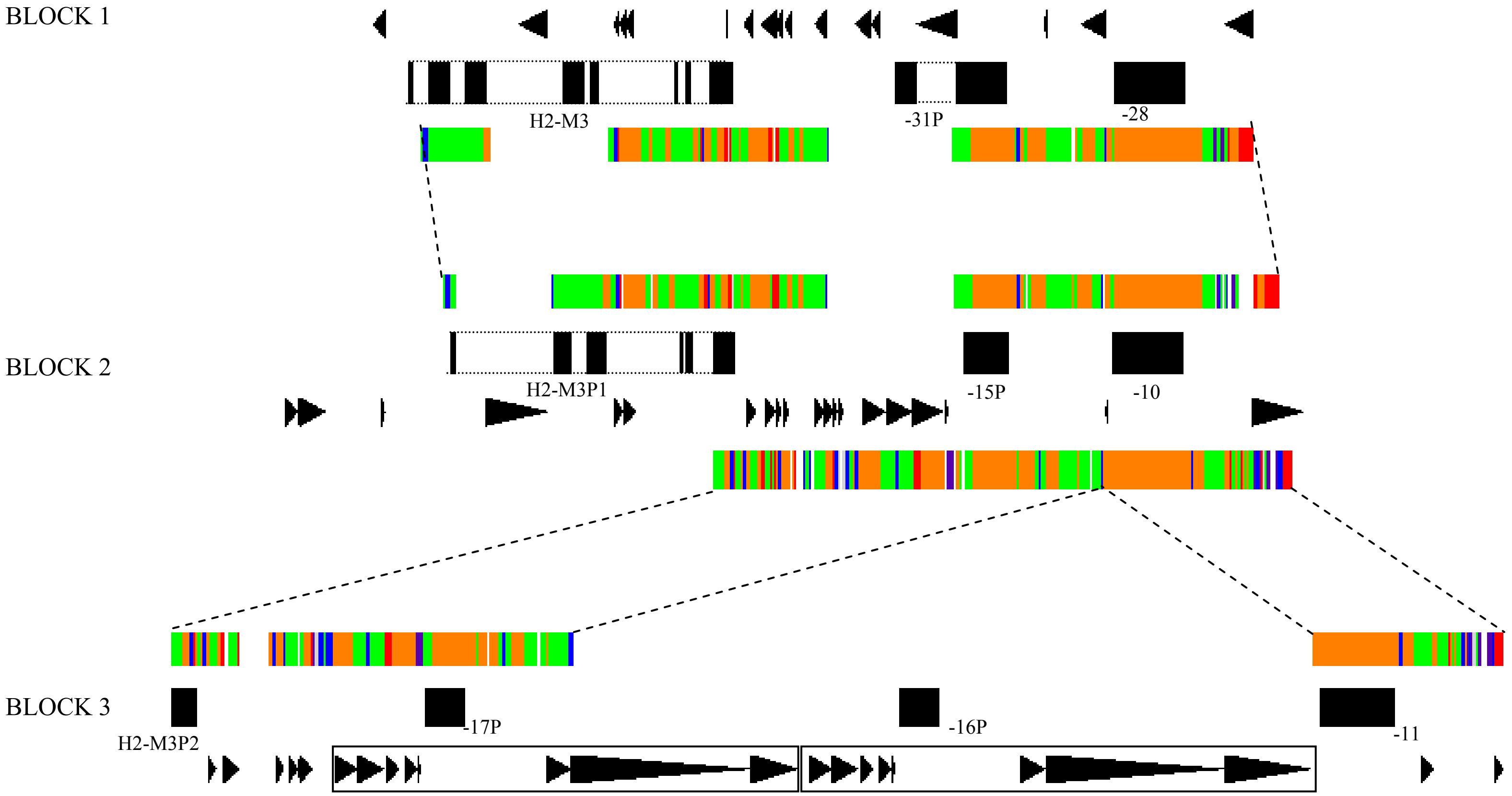


Figure 5.10: Duplications associated with the H2-M3 loci in the mouse MHC-linked OR cluster. The exon content and the repeat content of each block is indicated by the square blocks (exons) and triangles (repeats). Conserved blocks of sequence are indicated by coloured blocks which represent shared nucleotide 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.

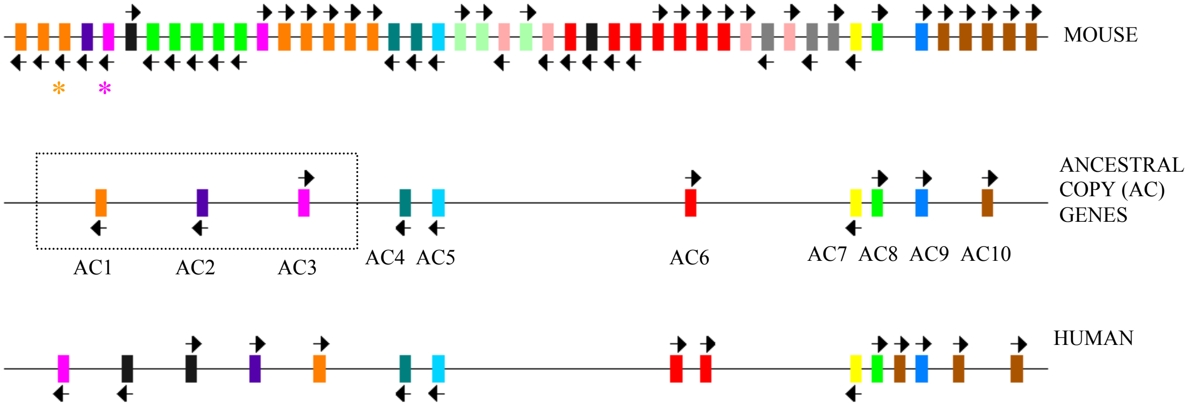


Figure 5.11: Orthologous olfactory receptor gene within the MHC-linked OR clusters in human and mouse (not to scale, other genes within the region not shown). A putative ancestral arrangement of OR genes is represented on the middle line: the boxed OR genes could be in either order as it appears an inversion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

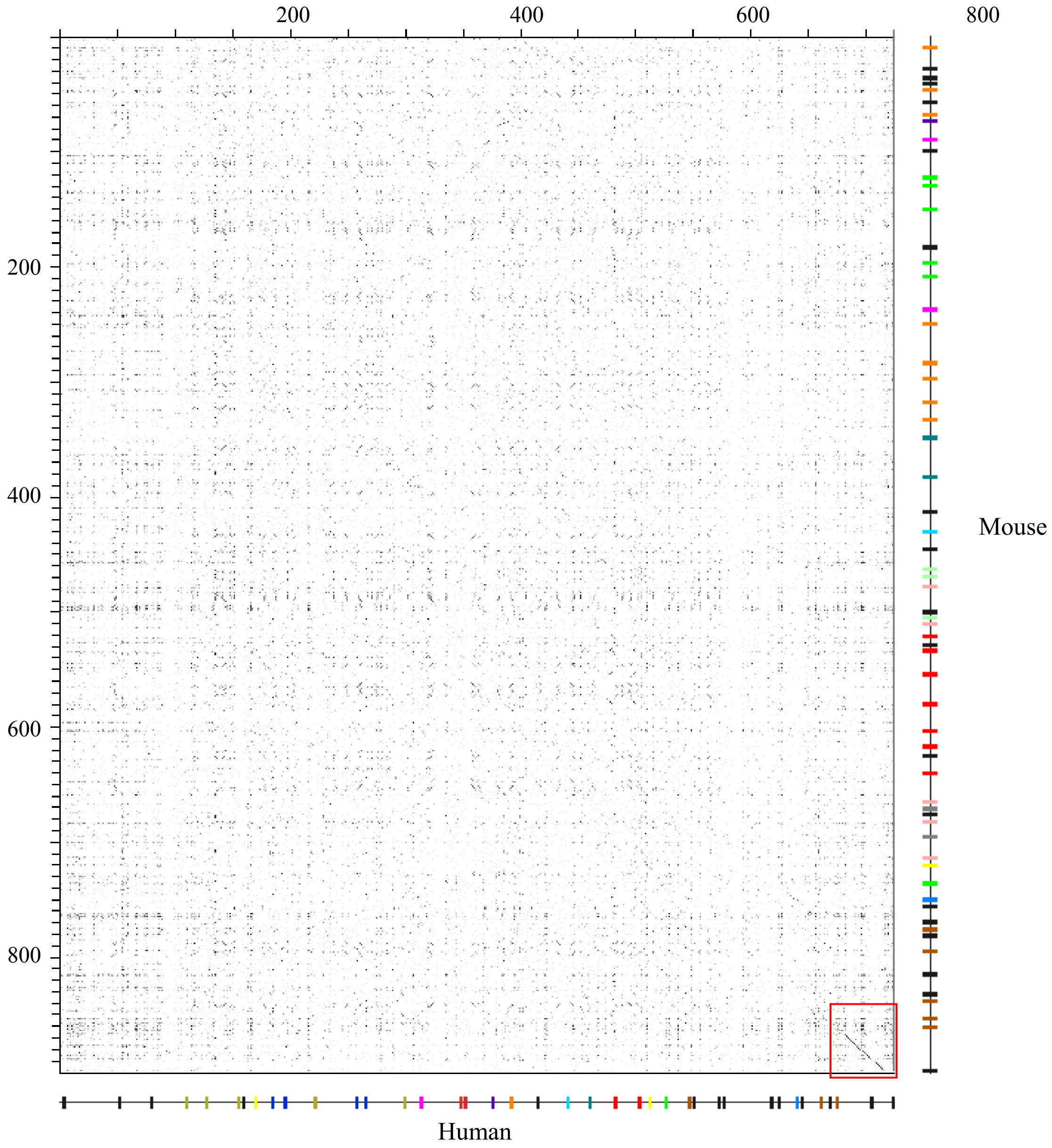


Figure 5.12 (previous page): Dot-matrix plot of human major MHC-linked OR cluster against the mouse MHC-linked cluster. The mouse sequence is plotted on the vertical axis whilst the human sequence is plotted on the horizontal axis. Bars under/to the right of the plot show the gene content of the region: 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

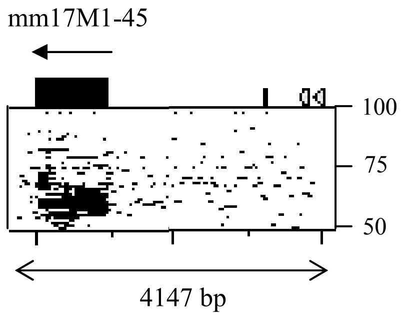


Figure 5.13b

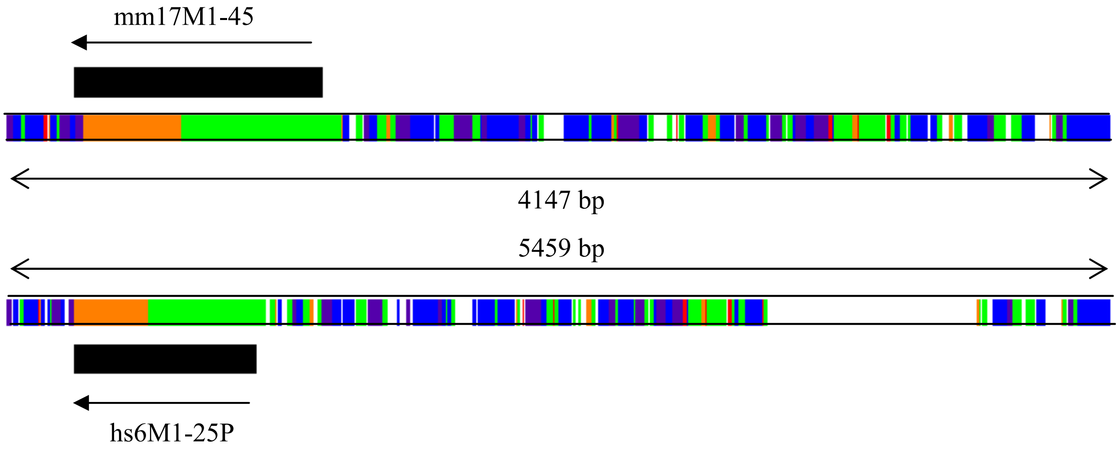


Figure 5.13a: mm17M1-45 PIP identity plot. Coding exons (shaded box) and repeats (triangles) are plotted on top of the box which contains lines showing segments of sequence conserved in the human MHC-linked OR cluster. The vertical position of these lines shows the similarity of these segments which 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).

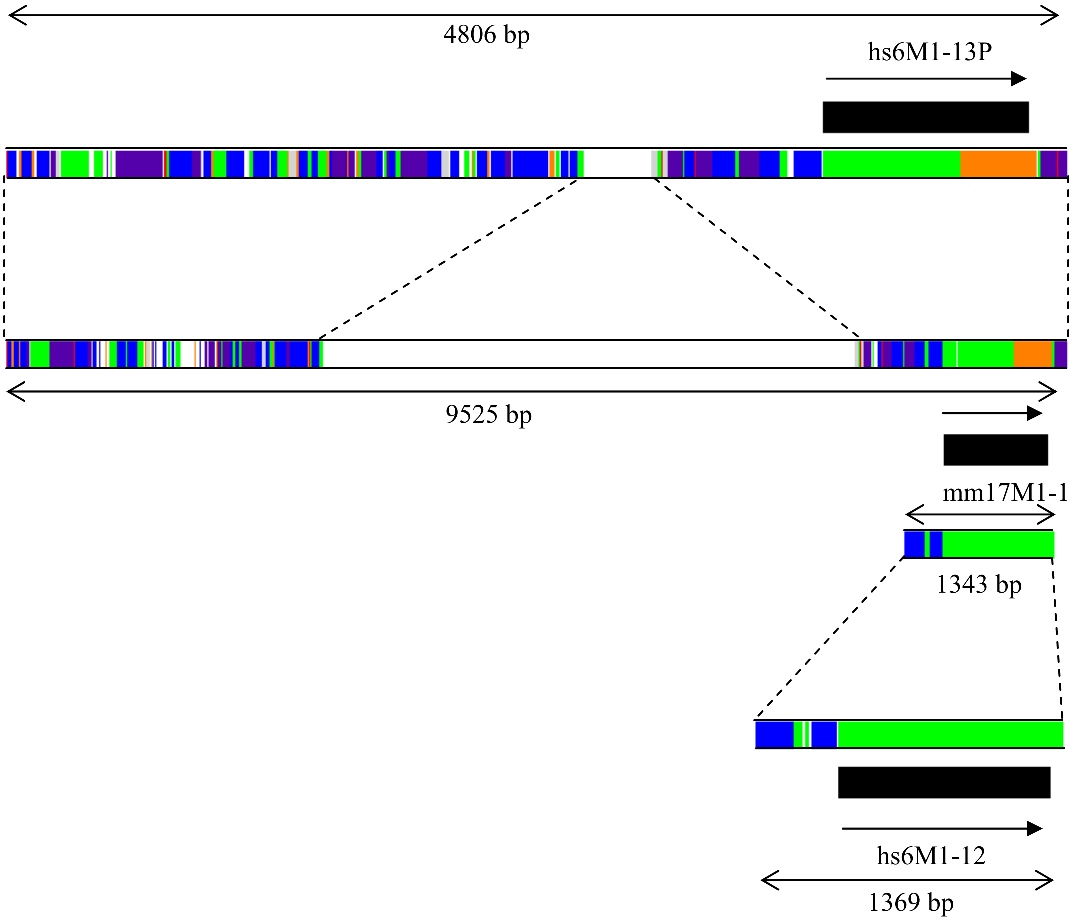


Figure 5.14: mm17M1-1 plotted against hs6M1-12 and hs6M1-13P. Conserved blocks of sequence are plotted in their position relative to the olfactory receptor genes. The colours represent different levels of conservation ranging from >90% (red), >80% (orange), >70% green, >60% (blue), > 50% (purple) and > 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.

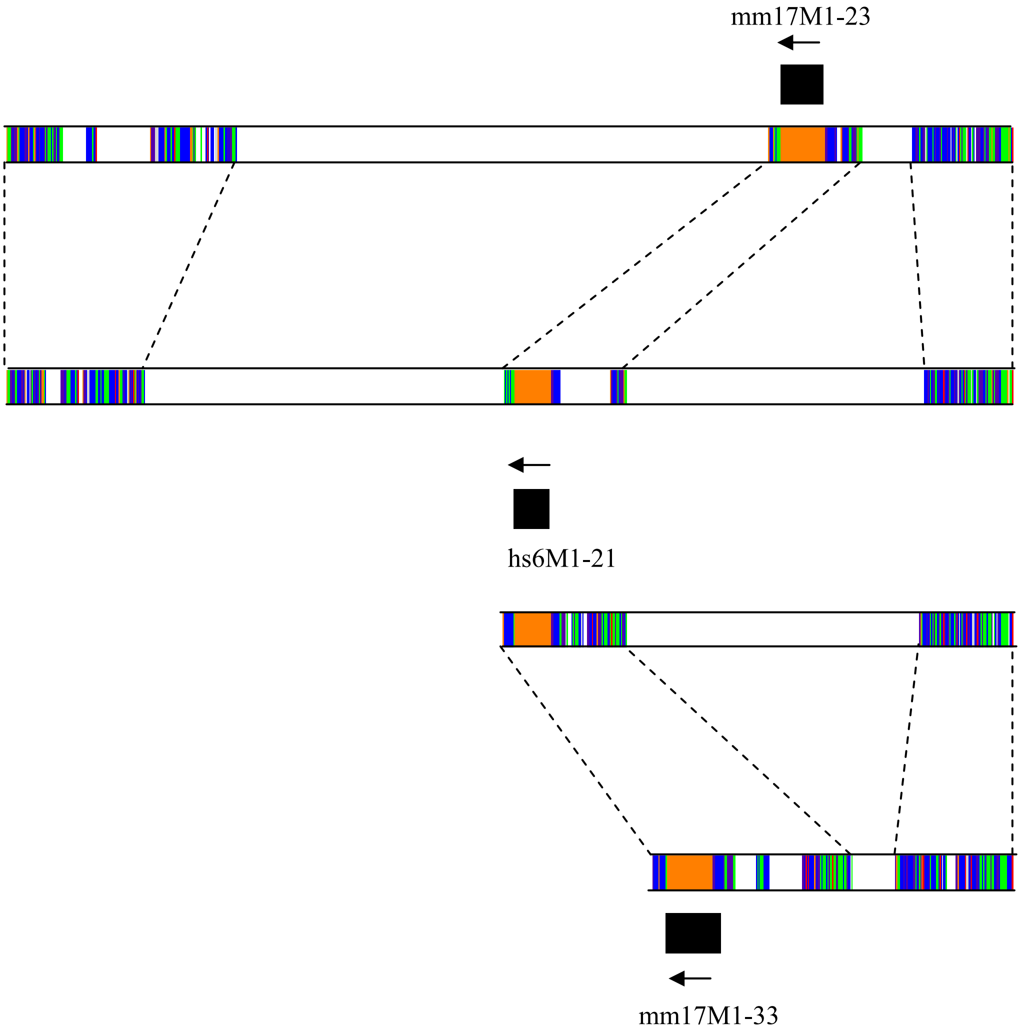


Figure 5.15: hs6M1-21 plotted against mm17M1-23 and mm17M1-33. Conserved blocks of sequence are plotted in their position relative to the olfactory receptor genes. The colours represent different levels of conservation ranging from >90% (red), >80% (orange), >70% green, >60% (blue), > 50% (purple) and > 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

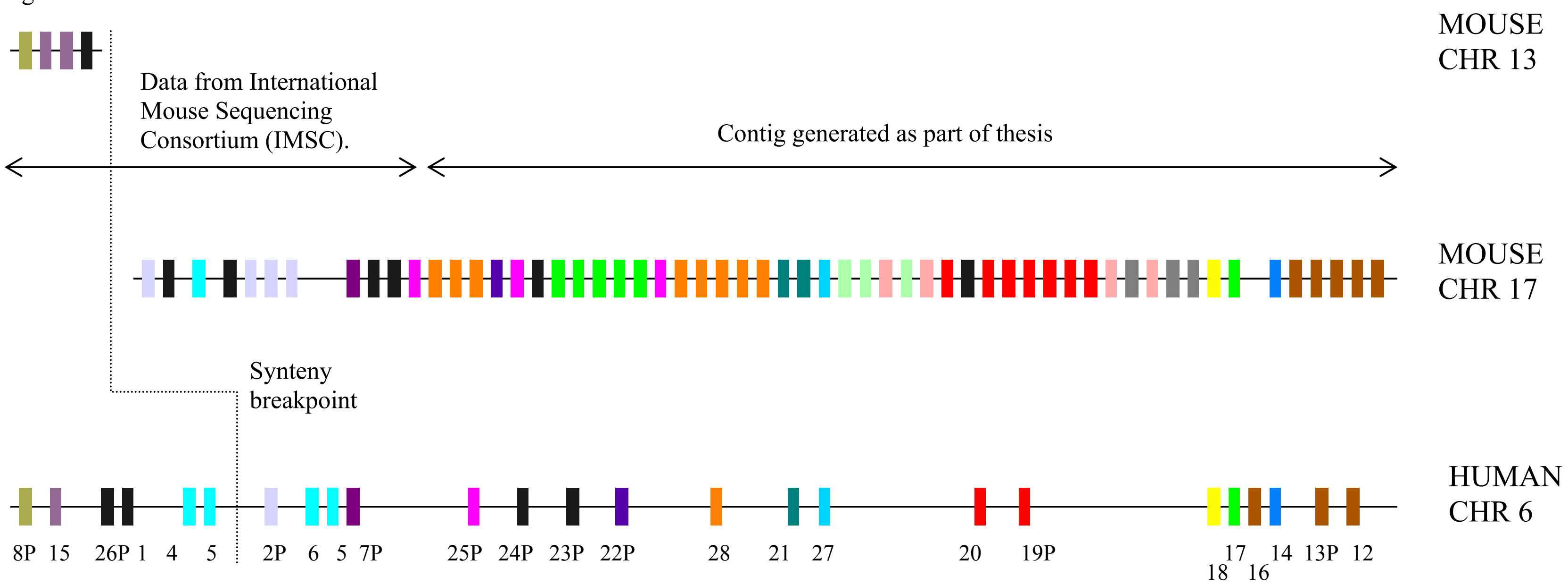


Figure 5.16b

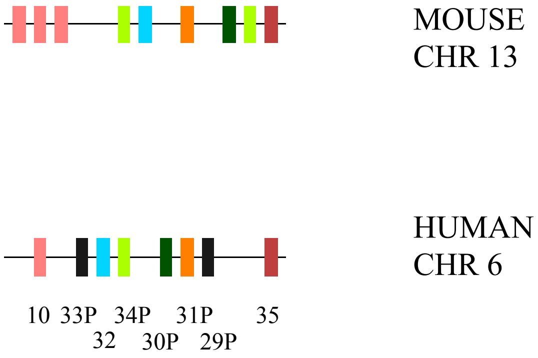


Figure 5.16: Orthologous ORs within the extended MHC major OR cluster (a) and within the extended MHC minor cluster (b). The genes are coloured according to their relationship to OR genes within the other species. The breakpoint in synteny appears to occur within the major OR cluster, around the hs6M1-2P 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

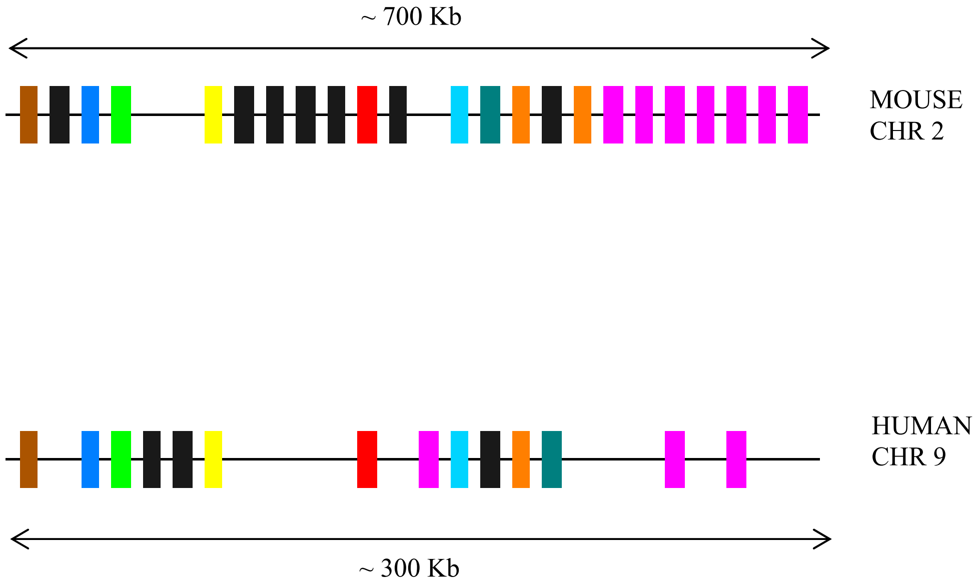


Figure 5.17b

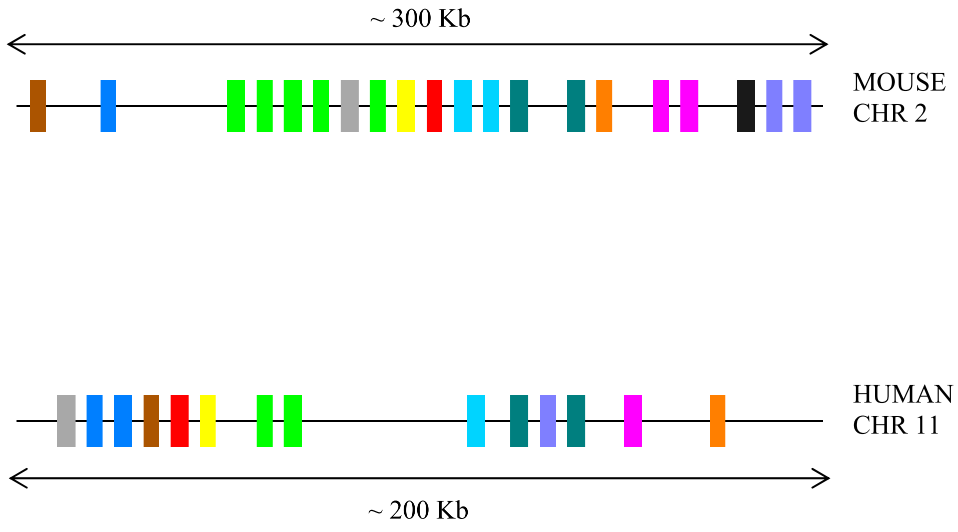


Figure 5.17: Orthologous clusters of OR genes on mouse chromosome 2 and human chromosomes 9 (a) and human chromosome 11 (b) . Orthologous groups were defined by a protein identity > 70%: they are indicated by the various colours in the 2 figures.

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.

Chapter 6

The expression and regulation of the MHC-linked OR genes

6.1. Introduction

The MHC-linked ORs, as is the case for the majority of OR genes within the human genome, are located within a cluster. The origin of these clusters is likely to be due to local duplication mechanisms increasing the number of OR genes within a specific chromosomal region. Clustering of these genes, however, has also been suggested to be functionally significant with regard to possibly controlling how these genes are regulated (Kratz *et al.*, 2002). Other multigene families are clustered within the genome and this clustering appears to be functionally important. Within the Hox transcription factor gene clusters, for example, the position of a gene within the cluster is important in controlling the amount of transcription, and the time and place of transcription (Duboule, 1998). The Hox gene cluster are an extreme example of this in that they are arranged so genes located at the 3' extremity of the cluster are activated first in early embryonic domains, such as the hindbrain, whilst 5' genes are transcribed later in more caudal areas (Lewis, 1978, Duboule and Dolle, 1989, Gaunt *et al.*, 1989, Graham *et al.*, 1989). The β-globin gene family is another example of a multigene family where gene clustering is implicated in the expression of these genes. Expression of β-globin genes is controlled according to developmental phase of the organism (Magram *et al.*, 1985); activation of these genes is thought to be regulated through the influence of locus control regions (LCRs) (Grosveld *et al.*, 1987, Grosveld, 1999).

The regulation of OR genes appears to be tightly controlled since a number of approaches, such as single cell PCR and *in situ* hybridisation, suggested that each olfactory sensory neuron in the olfactory epithelium expresses only a single allele of a single OR gene (Chess *et al.*, 1994, Buck, 2000). The mechanism(s) involved in this control of expression, therefore, determines to which range of odorants an OSN will respond.

The regulation of OR genes is also responsible for another process within the olfactory system. The choice of olfactory receptor gene determines which glomerulus within the olfactory bulb the OSN targets, as well as controlling which set of odorants produce a response in the OSN. The specific mechanism controlling this targeting is unknown but it is clear that OSNs expressing the same OR project to the same glomerulus in the olfactory bulb (Wang *et al.*, 1998, O'Leary *et al.*, 1999).

The regulation of the expression of olfactory receptor genes in olfactory sensory neurons, therefore, must be under a number of constraints in order to produce a functional olfactory epithelium. The idea of OR gene expression being highly restricted was actually used as a criteria for finding this superfamily: olfactory receptor genes were initially defined as genes that were likely to only be expressed in the olfactory epithelium (Buck and Axel, 1991). However, subsequent work on these genes suggested that they were expressed in the canine testis tissue (Parmentier *et al.*, 1992), and the developing rat heart (Drutel *et al.*, 1995). A systematic study of olfactory-like ESTs also provided evidence for the non-exclusive expression of OR genes. OR-like ESTs were found in a number of tissues, including colon, kidney, liver, placenta and testis (Dreyer, 1998). The expression of OR-like sequences in tissues that are not involved in the olfactory system suggests that OR genes may have a role outside the olfactory system.

The expression of the human MHC-linked olfactory receptor genes was therefore investigated to see if there was any evidence that some MHC-linked OR genes were expressed outside the olfactory system. A number of approaches were taken. Firstly, *in silico* analysis involved screening of publicly available expressed sequence tag (EST) databases. Secondly, results obtained from the *in silico* analysis were compared against results produced by hybridising specific probes against commercially bought RNA dot-blots. Thirdly, as the highest level of expression would be expected to be within olfactory epithelium tissue, specific primers were used for PCR on a cDNA library made out of this tissue, and mouse probes for use in *in situ* hybridisation experiments were developed.

The regulation of the MHC-linked olfactory genes was also investigated using a variety of methods, ranging from large scale analysis of the cluster using promoter prediction software through to experimental analysis of the ability of a small segment of sequence from the region to promote expression within a luciferase reporter vector. Analysis of the upstream region of the human MHC-linked olfactory receptor genes against each other and against their mouse orthologs was also performed.

6.2. *In-silico* transcript analysis of human MHC-linked OR genes

The screening of all human MHC-linked ORs against publicly available expressed sequence tag (EST) databases, produced hits as summarised in Table 6.1. The overall low hit rate is not surprising as there are no public EST data available from MOE tissue. Only 5 out of the 35 MHC-linked ORs show any matches to ESTs with greater than 90% similarity. These matches, however, confirm that some ORs are likely to be transcribed in non-MOE tissue such as lung, kidney, colon, prostate, testis and germ cell tumour and, therefore, may be involved in non-olfaction associated function.

OR gene	EST	Length	Location	Pos. in EST	Clone	Pos. in clone	%age
hs6M1-21	AA936177	387	Pooled library	3-171	AL096770	64684-64852	100
				168-247	AL035542	33742-33821	100
				246-284	AL035542	34162-34200	100
				284-387	AL035542	42999-43102	100
hs6M1-16	AI023490	477	Testis	4-370	AL035542	73504-73138	99
				367-477	AL035542	72888-72778	100
	AA382326	352	Testis	1-11	AL035542	69698-69708	100
				12-63	AL035542	71542-71593	100
				60-319	AL035542	72630-72888	97
				317-352	AL035542	73139-73174	88
hs6M1-24	AA922169	385	Pooled library	3-157	AL050339	43645-43491	100
				158-385	AL050339	40621-40394	99
hs6M1-32	N68399	428	fetal liver spleen	1-325	AL133267	21789-21464	99
				319-428	Z98744	57722-57623	97
hs6M1-14	AW071655	457	Germ cell tumors	1-457	AL031983		100
	AI912965	534	Kidney	1-534	AL031983		100
	AI763023	527	Kidney	1-527	AL031983		99
	AI304583	435	Colon	1-435	AL031983		100
	AI813634	580	Lung	1-580	AL031983		100
	AI476350	491	Pooled library	1-491	AL031983		99

Table 6.1: ESTs matching MHC-linked ORs. MHC-linked OR genes were screened against publicly available collections of ESTs. OR genes with matches are listed above, alongside their matching EST(s), the length of the EST and information about the origin of the EST. The ESTs were then mapped back to genomic DNA: columns show the EST positions that correspond to positions in clones contributing to the genomic sequence, and the percentage identity these sequences share. 'Pooled' libraries (location) contained ESTs from fetal lung, testis and B cells.

Alignment of these ESTs to the genomic sequence reveals unusual splicing in the 5'-UTRs of several ORs. For instance, the alignment for *hs6M1-21* reveals three 5'-UTR exons and indicates that the transcription start site is located some 80 kb upstream of the *hs6M1-21* ATG start codon (Figure 6.1). The predicted transcript spans four other OR loci, two of which are in the same (*hs6M1-18, 27*) and two of which are in the opposite (*hs6M1-19P, 20*) transcriptional orientation. This splicing around genes could suggest that long transcripts such as this one may play a role in controlling the expression of clustered ORs through mechanisms such as alternative splicing or antisense regulation.

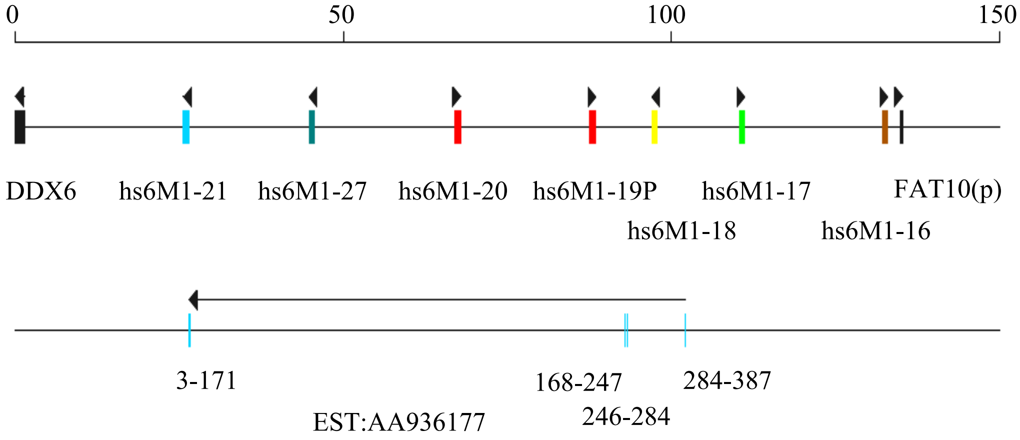


Figure 6.1: EST supported splicing in hs6M1-21. The top track shows the scale in Kb, whilst the middle track shows the exons found within this region of the human extended MHC. Exons are coloured differently according to which subfamily the OR gene belongs. (See figure 4.6). The third track shows positions at which exons of the EST AA936177 find a match. All 4 exons show the expected acceptor/donor (AG/GT) splice sites. The exons appear to splice around 4 OR genes, -27, -20, -19P and –18.

In the case of *hs6M1-16,* the alignment with 2 ESTs (both from testis) also reveals 3 exons in the 5'-UTR but only up to 3 kb upstream of the predicted ATG start codon. Interestingly, both ESTs splice around the expected start codon to the third methionine (amino acid position 79) within the single coding exon of *hs6M1-16*, producing a predicted protein lacking the first 78 amino acids, and therefore, the first two transmembrane domains (Figure 6.2).

A similar scenario exists with reference to the EST that aligns with hs6M1-24P. This EST splices 6 amino acids into the predicted open reading frame of the OR gene. In contrast to the other ESTs that splice into MHC-linked olfactory receptor genes, however, this EST does not appear to have conserved splice sites: a CT dinucleotide is present at the donor site and an AT exists at the acceptor site. This change to the recognised splice sites suggests this EST could be an artifact generated when this EST library was made.

The idea that splicing can create OR proteins that differ from those predicted according to open reading frames is also supported by observations from hs6M1-32 (Figure 6.3a). In this case, the first half of the EST matches to a presumed non-coding sequence in PAC 193B13 (Z98744) and the second half matches to PAC 408B20 (AL133267) and splices into amino acid position 254 of hs6M1-32. This results in a 5'-UTR of approximately 70 kb. As is the case with the EST from hs6M1-21 splicing occurs around other OR genes; hs6M1-10 which is in the same subfamily as hs6M1-32 but has a different transcriptional orientation, and hs6M1-33P a predicted pseudogene with the same orientation as hs6M1-32. Using the first in-frame methionine, this splice form would appear to produce a protein of only 41 amino acids, which possibly contains 1 transmembrane domain (Figure 6.3b). The two examples of hs6M1-32 and hs6M1-16 suggest alternative splicing may exist within the single coding OR exon.

Figure 6.2 (next page): Alignment of ESTs to hs6M1-16. AG/GT splice sites are highlighted in bold. Large introns are not shown but their sizes are indicated. Predicted transmembrane domains are boxed. Dashes were introduced in places to maximise the alignment.

```
AL035542  c t t a a t t g c a a G T a a g t c a c a a g t t t a t t c c c c t a c a g c c c a t c a a t t t c c a c a t g t t c t
AA382326  c t t a a t t g c a a
          ................................................ 1760 bp ................................................
AL035542  A G a g a c g a g g t t t c a c c a t g t t g a c c a g g c t g a t c t c a a a c a t c t g a c c t c a g G T g a t c c
AA382326      a g a c g a g g t t t c a c c a t g t t g a c c a g g c t g a t c t c a a a c a t c t g a c c t c a g

          ................................................ 980 bp ................................................
AL035542  A G g a a g t c a g a g g c a c c a a t g t g a g g t t c c a c c t g c t t t c c a g c a c a t t c t t g g t t t c c t
AA382326      g a a g t c a g a g g c a c c a a t g t g a g g t t c c a c c t g c t t t c c a g c a c a t t c t t g g t t t c c t

AL035542  c a c t t c t g c t a g a c a a c g t t t g a t c a g a a g g a a c a g g g a a c g a g a a g g a g c t g c t g g a t g
AA382326  c a c t t c t g c t a g a c a a c g t t t g a t c a g a a g g a a c a g g g a a c g a g a a g g a g c t g c t g g a t g

AL035542  a c g a t a a g c c t g g g a a a g g g a g g c t g g g t g a g c a g a g a c a g a a a a g a a a c a c c t a c c t g c
AA382326  a c g a t a a g c c t g g g a a a g g g a g g c t g g g t g a g c a g a g a c a g a a a a g a a a c a c c t a c c t g c
AI023490                                                        g t g a g c a g a g a c a g a a a a g a a a c a c c t a c c t g c

AL035542  t g t g a c c t c a c a a a c a c c c a g g c t g a g t t t t g a t a a g a c a g g t t g a a t c a c a c t - g g g - t
AA382326  t g t g a c c t c a c a a a n a a c c a g g c t g a g t t t t g a t a a g a c a g g t t g a a t c a c a a t n g g g g t
AI023490  t g t g a c c t c a c a a a c a c c c a g g c t g a g t t t t g a t a a g a c a g g t t g a a t c a c a c t - g g g - t

                                                                                        M     V     N     Q     S     S     P
AL035542  g a c a g c c t c a t c c c t c c a g G T a c a a a c a a g a a c a g g c c a t g g t t a a c c a a a g c t c c c c c a
AA382326  g a c a g c c t c a t t c c t n c a g
AI023490  g a c a g c c t c a t c c c t c c a g

          M     G     F     L     L     L     G     F     S     E     H     P     A     L     E     R     T     L     F     V
AL035542  t g g g c t t c c t c c t t c t g g g c t t c t c t g a a c a c c c a g c a c t g g a a a g g a c t c t c t t t g t g g

          V     V     F     T     S     Y     L     L     T     L     V     G     N     T     L     I     I     L     L     S
AL035542  t t g t c t t c a c t t c c t a c c t c t t g a c c c t g g t g g g c a a c a c a c t c a t c a t c c t g c t g t c t g

          V     L     Y     P     R     L     H     S     P     M     Y     F     F     L     S     D     L     S     F     L
AL035542  t a c t g t a c c c c a g g c t c c a c t c t c c a a t g t a c t t t t t c c t c t c t g a c c t c t c c t t c t t g g

          D     L     C     F     T     T     S     C     V     P     Q     M     L     V     N     L     W     G     P     K
AL035542  a c c t c t g c t t t a c c a c a a g t t g t g t c c c c c A G a t g c t g g t c a a c c t c t g g g g c c c a a a g a
AA382326                                                                  a t g c t g g t c a a a c t c t t g g g g c c c a a g a
AI023490                                                                  a t g c t g g t c a a c c t c t g g g g c c c a a a g a

          K     T     I     S     F     L     G     C     S     V     Q     L     F     I     F     L     S     L     G     T
AL035542  a g a c c a t c a g c t t c c t g g g a t g c t c t g t c c a g c t c t t c a t c t t c c t g t c c c t g g g g a c c a
AA382326  a g a c c
AI023490  a g a c c a t c a g c t t c c t g g g a t g c t c t g t c c a g c t c t t c a t c t t c c t g t c c c t g g g g a c c a

          T     E     C     I     L     L     T     V     M     A     F     D     R     Y     V     A     V     C     Q     P
AL035542  c t g a g t g c a t c c t c c t g a c a g t g a t g g c c t t t g a c c g a t a c g t g g c t g t c t g c c a g c c c c
AI023490  c t g a g t g c a t c c t c c t g a c a g t g a t g g c c t t t g a c c g a t a c g n t g c t g t c t g c c a g c c c c

          L     H     Y     A     T     I     I     H     P     R     L     C     W     Q     L     A     S     V     A     W
AL035542  t c c a c t a t g c c a c c a t c a t c c a c c c c c g c c t g t g c t g g c a g c t g g c a t c t g t g g c c t g g g
AI023490  t c c a c t a t g c c a c c a t c a t c c a c c c c c g c c t g t g c t g g c a g c t g g c a t c t g t g g c c t g g g

          V     M     S     L     V     Q     S     I     V     Q     T     P     S     T     L     H     L     P     F     C
AL035542  t t a t g a g t c t g g t t c a a t c g a t a g t c c a g a c a c c a t c c a c c c t c c a c t t g c c c t t c t g t c
AI023490  t t a t g a g t c t g g t t c a a t c g a t a g t c c a g a c a c c a t c c a c c c t c c a c t t g c c c t t c t g t c

          P     H     Q     Q     I     D     D     F     L     C     E     V     P     S     L     I     R     L     S     C
AL035542  c c c a c c a g c a g a t a g a t g a c t t t t t a t g t g a g g t c c c a t c t c t g a t t c g a c t c t c c t g t g
AI023490  c c c a c c a g c a g a t a g a t g a c t t t t t a t g t g a g g t c c c a t c t c t g a t t c g a c t c t c c t g t g

          G     D     T     S     Y     N     E     I     Q     L     A     V     S     S     V     I     F     V     V     V
AL035542  g a g a t a c c t c c t a c a a t g a a a t c c a g t t g g c t g t g t c c a g t g t c a t c t t c g t g g t t g t g c
AI023490  g a g a t a c c t c c t a c a a t g a a a t c c a g t t g g c t g t g a a a

          P     L     S     L     I     L     A     S     Y     G     A     T     A     Q     A     V     L     R     I     N
AL035542  c t c t c a g c c t c a t c c t t g c c t c t t a t g g a g c c a c t g c c c a g g c a g t g c t g a g g a t t a a c t
```

Figure 6.3a

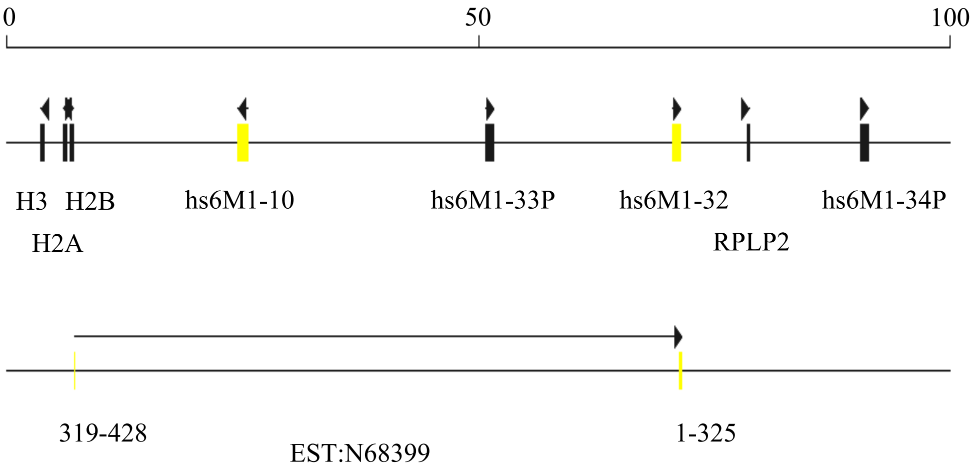


Figure 6.3b

```
           C     G     S     H     L     I     V     V     S     L     F     Y     S     T     A     V     S     V     Y     L
AL133267 t g t g g t t c c c a t c t a a t t g t g g t g t c t c t t t t t t a t A G t a c a g c c g t c t c t g t g t a c c t g
                                                                                     t a c a g c c g t c t c t g t g t a c c t g

           Q     P     P     S     P     S     S     K     D     Q     G     K     M     V     S     L     F     Y     G     I
AL133267 c a a c c a c c t t c g c c c a g c t c c a a g g a c c a a g g a a a g a t g g t t t c t c t c t t c t a t g g a a t c
N68399   c a a c c a c c t t n g c c c a g c t c c a a g g a c c a a g g a a a g a t g g t t t c t c t c t t c t a t g g a a t c

           I     A     P     M     L     N     P     L     I     Y     T     L     R     N     K     E     V     K     E     G
AL133267 a t t g c a c c c a t g c t g a a t c c c c t t a t a t a t a c a c t t a g g a a c a a g g a g g t a a a g g a a g g c
N68399   a t t g c a c c c a t g c t g a a t c c c c t t a t a t a t a c a c t t a g g a a c a a g g a g g t a a a g g a a g g c

           F     K     R     L     V     A     R     V     F     L     I     K     K     *     E     I     C     K     *     *
AL133267 t t t a a a a g g t t g g t t g c a a g a g t c t t c t t a a t c a a g a a a t a a g a a a t a t g c a a a t g a t a a
N68399   t t t a a a a g g t t g g t t g c a a g a g t c t t c t t a a t c a a g a a a t a a g a a a t a t g c a a a t g a t a a

           A     L     L     K     T     K     C     L     L     S     L     L     T     S     L     *     V     A     L     F
AL133267 g c t t t g c t a a a g a c a a a a t g t t t a c t t a g c t t a c t a a c t t c t c t g t a a g t t g c c c t a t t t
N68399   g c t t t g c t a a a g a c a a a a t g t t t a c t t a g c t t a c t a a c t t c t c t g t a a g t t g c c c t a t t t

           L     L     L     L     *     R     T     M     *     T     P     S     N     K     I     S     L     M     K     S
AL133267 t t g t t g t t a c t g t a g a g a a c a a t g t a a a c t c c c t c a a a t a a a a t t t c c t t g a t g a a g a g c
N68399   t t g t t g t t a c t g t a g a g a a c a a t g t a a a c t c c c t c a a a t a a a a t t t c c t t g a t g a a g a g c

AL133267 t a
N68399   t a
```

Figure 6.3a: EST associated with hs6M1-32. The top track shows the scale in Kb, whilst the middle track shows the exons found within this region of the human extended MHC. Exons are coloured differently according to which subfamily the OR gene belongs. (See figure 4.6). The third track shows positions at which exons of the EST N68399 find a match. As is the case for exons associated with hs6M1-21, the exons of the EST appear to splice around 2 OR genes, -10, and –33P.

Figure 6.3b: Alignment of EST N68399 to genomic sequence, including hs6M1-32. The predicted transmembrane domains are boxed, and the stop codon is indicated by the red box. The predicted protein extends 330 amino acids upstream of this point, although the EST splicing seems to suggests an alternative transcript is produced.

The hypothesis of alternative splicing or alternative use of ATG start codons may also explain one of the differences observed between mouse and human ORs. Hs6M1-14P, for example, is considered a pseudogene since it misses the first 78 amino acids compared to its murine ortholog, mm17M1-6. It is, however, the only OR matching a comparatively large number of ESTs all between 99-100% similarity and from non-olfaction associated tissues (Table 6.1). Although the position of sequence divergence coincides perfectly with the presence of an acceptor splice site, several ESTs span the position, indicating that this splice site is not used, at least not in the tissues from which the ESTs were derived (data not shown). This may mean, that as is predicted for hs6M1-16, hs6M1-14P could make use of an alternative ATG start codon, most likely the one corresponding to the methionine mentioned above for hs6M1-16, resulting again in a protein product without the first two transmembrane domains of an OR protein.

An analysis of the MHC-linked OR protein sequences reveals that potentially this alternative splicing or use of alternative ATG start codons may be quite common, as the methionine at amino acid position 79 is conserved in 61% of the MHC-linked ORs. Of these, nine (hs6M1-2P, -7P, -8P, -9P, -15, -16, -21, -22P, -24P) have apparently functional acceptor splice sites which would allow expression from this methionine as for hs6M1-16. The splicing would effectively avoid the frameshift mutations in hs6M1-7P and hs6M1-22P, making these two pseudogenes potentially expressable as proteins. In all examples discussed here, the AGGT splice consensus motif has been preserved and the corresponding splice phases are matching.

The *in silico* transcript analysis also suggests that some ORs (including ORs currently classified as pseudogenes) may be expressed in a truncated, yet functional form. Alternative splicing of OR genes has been reported, although the distances are much shorter than those that are suggested for hs6M1-21 (Asai *et al.*, 1996, Walensky *et al.*, 1998). The expression of olfactory receptor-like sequences coding for proteins containing less than 7 transmembrane domains is also a finding that has not been reported before in the literature. There are at least 3 possible ways to interpret this finding. Firstly, it may be that as a result of alternative splicing these genes are translated as proteins containing less than 7 transmembrane domains. The deletion of the first two transmembrane domains (as in the case of hs6M1-16) has been shown not to affect the functional expression of other members of the 7TM G-coupled protein receptor gene family (Ling *et al.*, 1999).

An alternative interpretation is that segments of different OR genes may recombine to produce novel proteins. This mechanism (somatic recombination) may involve a process similar to that involved in generating diversity within the T-cell receptor family, where arrays of V (variable) gene segments can recombine with members of arrays of D (diversity) and J (joining) gene segments. This process produces a new exon coding for the antigenV–binding pocket of immunoglobulins or T-cell receptors (Lieber, 1996). This hypothesis would explain the high conservation of genes that appear to lack complete open reading frames: conserved gene segments may be involved in recombination events, however, there is currently a lack of expression data supporting this hypothesis.

Thirdly, these EST expression data may represent artifacts. Owing to protocols that ensure the rapid generation of ESTs, it is known that ESTs sometimes contain sequence and annotation inaccuracies, and little manual editing of these single read sequences is performed (Hillier *et al.*, 1996, Wolfsberg and Landsman, 1997). In addition, pairs of ESTs that have been reported as being derived from the same gene have in some cases failed to align to the sequence of the same gene suggesting the presence of artifacts in EST databases. The possibility of genomic contamination in EST libraries is suggested by the existence of ESTs matching to 2 predicted pseudogenes, hs6M1-14P and hs6M1-24P. Expression of olfactory receptor pseudogenes, however, has been observed in the olfactory tissue suggesting that this EST data may not be artifactual; it may represent the fact that some olfactory receptor pseudogenes are being transcribed (Crowe *et al.*, 1996).

The balance of evidence from the analysis of *in silico* transcripts suggests that some MHC-linked olfactory receptor genes are expressed in tissues other than olfactory tissue. It also appears that there is a certain amount of splicing that could contribute to diversity within these genes through the alternative splicing of 5'UTRs or through the alternative splicing within the coding region of the gene. Across the genome as a whole, alternative splicing is very common. Studies have generally suggested alternative splicing takes place in at least 35% of genes in the TIGR human gene index (Mironov *et al.*, 1999) and at least 34% of proteins in the SwissProt database (Hanke *et al.*, 1999). As, on average, ESTs only cover 50% of a gene, these estimates may be underestimates (Hanke *et al.*, 1999).

6.3. Experimental analysis of expression in MHC-linked OR genes

In order to confirm whether expression occurs in tissues outside the olfactory system, probes from the 3' UTR of several MHC-linked olfactory receptor genes were prepared and these were hybridised against a multiple tissue RNA dot-blot. This confirmed the expression of hs6M1-16 in tissues such as the kidney, liver, small intestine, and lung. There was also some support for expression of this gene in the colon (Figure 6.4).

Expression in the testis which appears to exist according to the EST data, however, could not be detected. This discrepancy may be due to the fact that the probe used in this hybridisation came from the 3' end of hs6M1-16 in order to produce a probe that differentiated between hs6M1-16, and the 2 other OR genes in the subfamily hs6M1-12 and hs6M1-13P. The positive hybridisation of the probe to RNA from the kidneys, liver, small intestine and colon does support the idea for an additional function for the hs6M1-16 gene outside the olfactory organs.

This non-olfactory expression, however, was not found for 2 other MHC-linked OR genes. Probes from both hs6M1-15 and hs6M1-20 failed to hybridise to any RNA on the dot-blot, although probes did hybridise to the human genomic control dot on the RNA blots (Figure 6.4b, data not shown). Expression of MHC-linked ORs therefore appears to be variable, although as the probe was designed within the 3’ UTR in order to allow unique primers to be designed, the lack of expression that was detected may be due to the gene possessing a 3’ untranslated region that is alternatively spliced. Work by my collaborators in Berlin (Andreas Ziegler, Armin Volz and Anke Ehlers, Institut für Immungenetik, Universitätsklinikum Charité, Humboldt-Universität zu Berlin) suggested expression in non-olfactory tissues of a number of other MHC-linked OR genes: hs6M1-10, hs6M1-6, hs6M1-1, hs6M1-17 and hs6M1-18. This expression was detected by using probes from the middle of the gene which means there may have been some cross-reactivity with other MHC-linked ORs, for example, probes for hs6M1-6 were likely to hybridise to hs6M1-3 and hs6M1-4P.

In order to consider MHC-linked OR gene expression in the mouse, RT-PCR was performed using primers from mm17M1-1, mm17M1-2, mm17M1-3, mm17M1-4, and mm17M1-6. Results from these experiments were inconsistent, but expression was detected in testis and in a pool containing cDNAs from lung, kidney, stomach and heart. Expression of mouse olfactory receptor genes, therefore, also appears to be something that is not restricted to the olfactory epithelium, although more systematic work is required to confirm this observation. Expression of OR genes in the olfactory epithelium in mouse was also investigated through developing a number of OR constructs that could be used in *in situ* hybridisation experiments, however, hybridisation experiments failed to produce any conclusive results.

Figure 6.4a

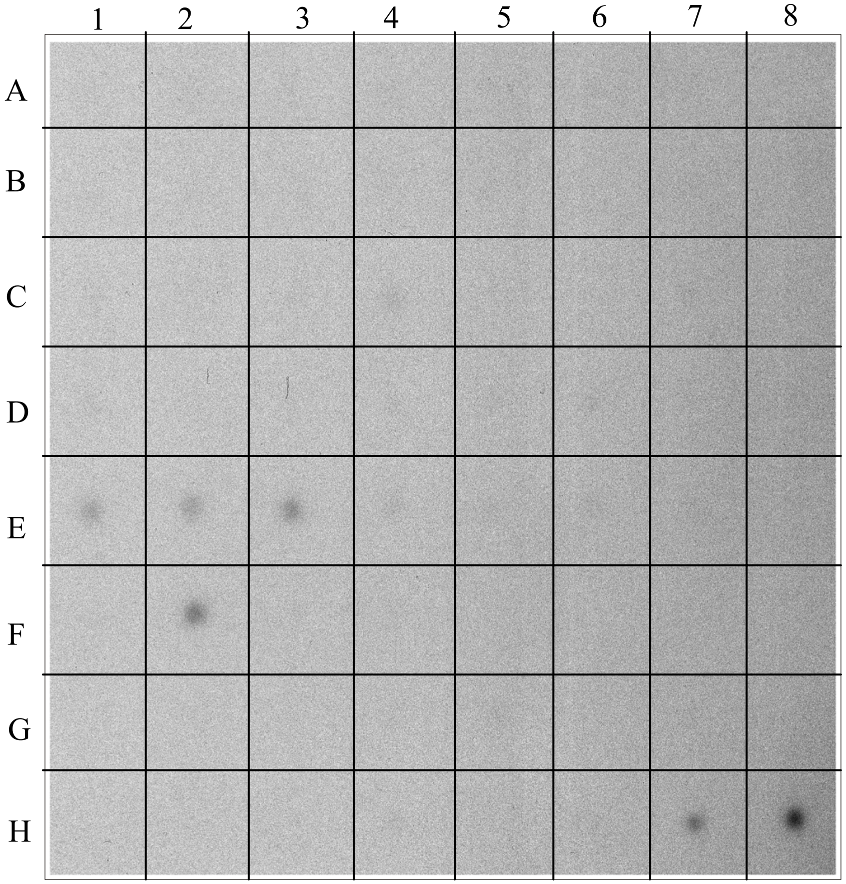


Figure 6.4b

	1	2	3	4	5	6	7	8
A	Whole brain	Amygdala	Caudate nucleus	Cere-bellum	Cerebral cortex	Frontal lobe	Hippo-campus	Medulla oblongata
B	Occipital lobe	Putamen	Substan-tia nigra	Temporal lobe	Thalamus	Sub-Thalamic nucleus	Spinal cord	
C	Heart	Aorta	Skeletal muscle	Colon	Bladder	Uterus	Prostate	Stomach
D	Testis	Ovary	Pancreas	Pituitary gland	Adrenal gland	Thyroid gland	Salivary gland	Mammary gland
E	Kidney	Liver	Small intestine	Spleen	Thymus	Peripheral leukocyte	Lymph-node	Bone marrow
F	Appendix	Lung	Trachea	Placenta				
G	Fetal brain	Fetal heart	Fetal kidney	Fetal liver	Fetal spleen	Fetal thymus	Fetal lung	
H	Yeast Total RNA 100 ng	Yeast tRNA 100 ng	E.coli rRNA 100 ng	E.coli DNA 100 ng	Poly r(A) 100 ng	Human Cot1 DNA 100 ng	Human DNA 100 ng	Human DNA 500 ng

Figure 6.4a: Dot-blot of hs6M1-16 shows expression in the kidney, liver, small intestine and lung after an exposure time of 48 hours.
Figure 6.4b: Key to the dot-blot.

The evidence that hs6M1-16 is expressed outside the olfactory epithelium, supported by additional cDNA data from my Berlin collaborators (Volz *et al.*, unpublished), meant that this gene was chosen as the gene that should be investigated with relation to expression in the olfactory epithelium. The alternative splicing that was observed in the testis EST could be a splice form that would only be expressed in non-olfactory tissue, whilst the 7 transmembrane domain protein could be restricted to the olfactory epithelium. Primers were designed in a number of positions across the predicted open reading frame and the corresponding probes were successfully amplified from genomic DNA. A cDNA library made from the olfactory epithelium and specifically enriched for MHC-linked OR genes was kindly provided by Ian Connerton (University of Nottingham, Crowe *et al.*, 1996). Titering of this library indicated that the number of plaques produced would be insufficient for hybridisation experiments to proceed. In order to consider alternative splicing, therefore, primers that had been used to amplify the probes for hybridisation and primers from 5'UTR exons were used to set-up PCR reactions from the phage stock and from pooled phage suspensions. These methods were both successful in amplifying transmembrane regions from the hs6M1-16 gene, but amplification of 5' untranslated regions (using primers designed from 5' UTR exons observed in testis) was unsuccessful. This could be due to alternative splicing between testis and olfactory epithelium, but since genomic controls also failed to amplify these regions, this conclusion cannot be drawn.

Expression of some MHC-linked olfactory receptor genes was therefore confirmed to take place outside the olfactory system (kidney, liver, small intestine, testis, lung, and colon). Confirmation of expression of hs6M1-16 in the olfactory epithelium was also produced. The EST evidence provided evidence for the possibility of alternative splicing in different tissues within humans: this alternative splicing may be involved in regulating the differential expression, and therefore the (presumbably) different roles of olfactory receptors in these tissues. Attempts to investigate alternative splicing of these genes in different tissues through looking at expression in the olfactory epithelium, were made but the existence of different splice forms between the testis and the olfactory epithelium could not be confirmed.

6.5. Regulation of the generation of alternative transcripts in the human MHC-linked ORs

Additional results from my collaborators in Berlin (Andreas Ziegler, Armin Volz and Anke Ehlers, Institut für Immungenetik, Universitätsklinikum Charité, Humboldt-Universität zu Berlin) confirmed the existence of olfactory transcripts in a number of tissues apart from the olfactory epithelium (Volz *et al.*, unpublished). They also provided evidence for the alternative splicing of a number of MHC-linked olfactory receptor genes. Focussing on genes that had EST data attached, they found several splice forms of hs6M1-21 and hs6M1-16. They also found several splice forms of 2 genes located between hs6M1-21 and hs6M1-16, hs6M1-27 and hs6M1-18. Figure 6.5 shows the integration of their results with my work. These splice forms were identified through a number of 5' RACE experiments using testis cDNA libraries. 9 alternative splice forms of hs6M1-16 were identified, including 2 that splice within the gene (the first 79 base pairs are spliced out, confirming the EST data for hs6M1-16), and 3 that have 5' UTR (non-coding) exons that are located within 200 base pairs of a 5' UTR exon that is shared by hs6M1-18, -21 and –27 (which are transcribed in the opposite orientation). Hs6M1-21 and hs6M1-27 also share another 5' UTR exon which could have some implications in how these genes are transcribed. Data confirming the EST splicing from hs6M1-21 was not produced, although the high number of observed alternative transcripts means this EST data cannot be totally discounted.

The finding of various alternative transcripts which revealed that several splice forms of hs6M1-21, -27, -18 and –16 all appear to have 5' UTR exons that are located very closely together suggested that the region between these exons is involved in the regulation of these genes. This region, the site of a putative promoter, was investigated through (i) searching for transcription factor binding sites within the region (using the TRANSFAC database in conjunction with the 'MatInspector' program (Wingender *et al.*, 2000)), (ii) searching for sequence similarity within other olfactory receptor gene clusters, and (iii) cloning the region (position 126-346, positions relate to scale in Figure 6.5b) in both orientations into a luciferase reporter vector to test for promoter activity.

The TRANSFAC database, accessed using the 'MatInspector' program, was used to search for transcription factor binding sites within a 500 base pair region containing hs6M1-18/21/27 exon 1 and 2 alternative starting exons of hs6M1-16. A number of matches were observed (Figure 6.5b), but only matches lying within the putative promoter region between the first exon of hs6M1-18/21/27 and hs6M1-16 (position 215-291, related to Figure 6.5b) were analysed in detail. Results from the region (Table 6.2) show significant matches to three groups of transcriptional factors: fork head related activators, SRY-related factors and AP1 transcription factors. These binding motifs are all common within the genome, and even using a program such as 'FastM' (Klingenhoff *et al.*, 1999) which allows a model of a putative promoter region to be developed through predicting two binding sites, their strand orientation, their sequential order, and the allowed distance between binding sites, nothing distinctive about this collection of transcription factor binding sites could be discerned. The frequency of fork head related activators located within 30-50 bases of AP1 transcription factor binding sites is fairly high within the genome.

The sequence containing the first two 5' UTR exons of hs6M1-18/21/27 and hs6M1-16 (position 126-346, related to Figure 6.5b) was compared against other regions of the human genome, using the 'BLAST' program, to see whether this is unique sequence, or whether it exists in other OR clusters. Analysis revealed that the first half (position 126-247) of this sequence is unique. However, the second half (position 248-346) of the sequence was found to be similar to several

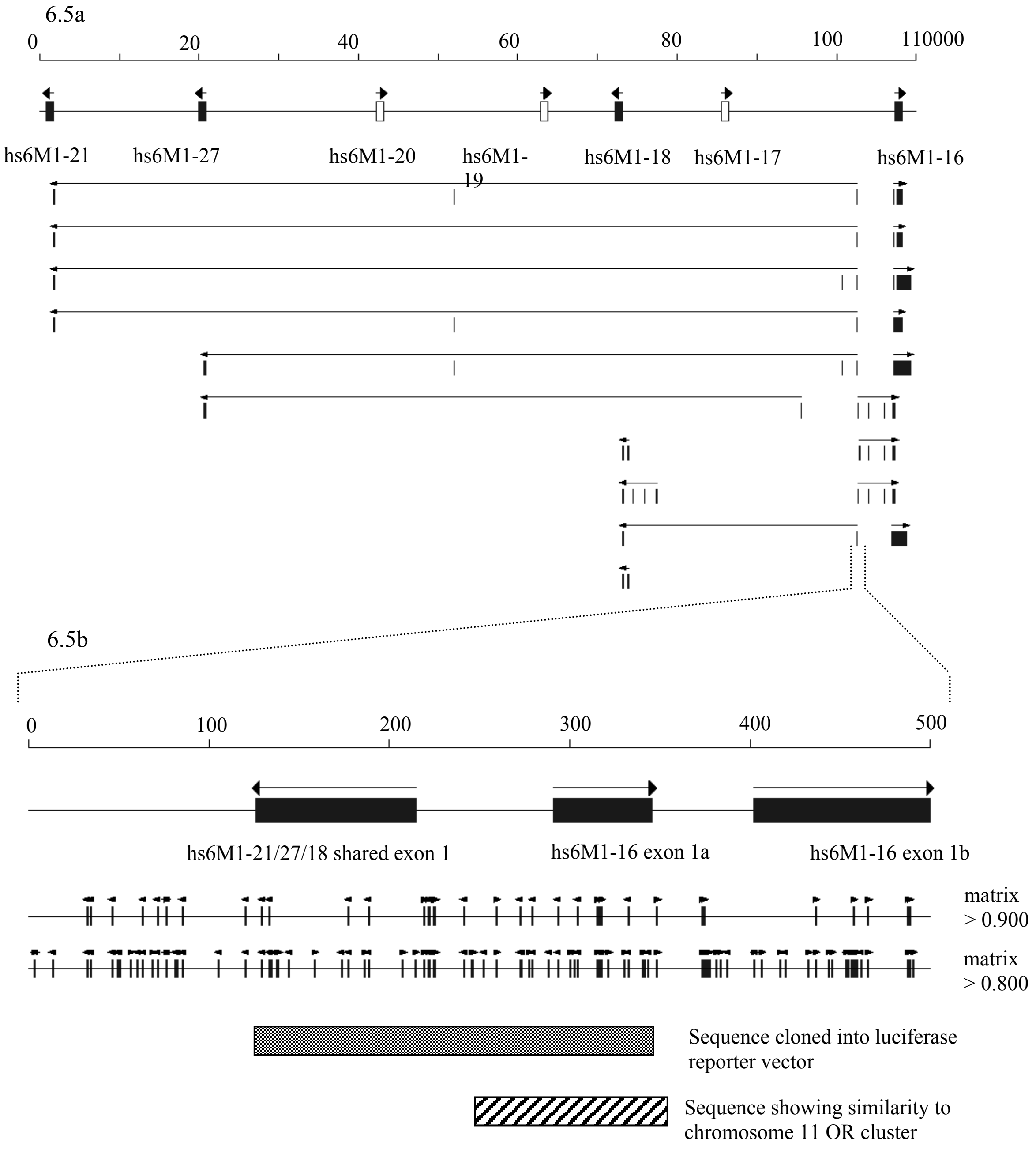


Figure 6.5a: Alternative splice forms of hs6M1-21, -27, -18 and –16. Other olfactory receptor genes within the region were not analysed by 5' RACE (hs6M1-19P, hs6M1-17) or were analysed by 5' RACE but no splicing was observed (hs6M1-20).

Figure 6.5b: Enlarged section showing shared exons and putative promoter sequence. Plotted below are matches from the TRANSFAC database, with matrix similarity >0.900 or >0.800. All matches plotted have >0.900 similarity to the core of the matrix. Below the lines 2 boxes show (i) the region of sequence cloned into the luciferase reporter vector and (ii) the region of sequence showing >67% similarity to a sequence located within a chromosome 11 OR cluster.

Start position in sequence	TRANSFAC accession no.	Description
219	T02465, T02472	Fork head related activators
221	T02474,T02294	Fork head related activator Xenopus fork head domain
222	T02288	Fork head domain
224	T00997	SRY, sex determining region on Y.
225	T01429	SRY-related HMG-box gene 5
241	T00027	AP1 transcription factor
259	T00027	AP1 transcription factor
272	T00027	AP1 transcription factor
279	T01470	Ikaros, lymphoid specific transcription factor.

Table 6.2: TRANSFAC matches found in the ‘olfactory promoter’ region. Matches have a similarity to the matrix of over 0.900 and a similarity to the core of over 0.900, and are found between the two 5’ UTR exons of hs6M1-16 exon 1a and hs6M1-18/21/27 exon 1. Positions relate to the scale used in Figure 6.5b.

other regions of sequence within the genome. One of these similar sequences (with a 69% shared base pair identity) is located in an OR cluster on chromosome 11q12.2, but as this region of the genome is currently unfinished, further work is needed to confirm whether this shared sequence is located in a similar putative regulatory region in the chromosome 11q12.2 cluster.

Computational approaches, therefore, suggested there were few significant features within the putative promoter region that could distinguish this sequence as a putative OR promoter. In spite of these approaches, however, the experimental evidence from the 5’ RACE experiments which suggests transcription is initiated in both orientations from the gap between the exons seems to point to a putative OR promoter that can trigger the transcription of four OR genes, hs6M1-16, hs6M1-18, hs6M1-21 and hs6M1-27, being located in this region (position 215-291, Figure 6.5b). To investigate this further, functional analysis of the region was carried out: this involved cloning the candidate promoter region (position 126-346, Figure 6.5b) into a pGL3 luciferase reporter vector in both the forward and reverse orientation and transfecting this vector and other control vectors into two cell types. Odora cells, from rat olfactory sensory neurons where ORs can be expressed were transfected along with human embryonic kidney cells (HEK293) as there is some evidence that some ORs are expressed in the kidney (EST, dot-blot data). After transfection both sets of cells were assayed for luminescence (Table 6.3). In both cases, control signals were strong, but cells transfected with the test vectors revealed no promoter activity. In the case of the ‘OLFOP(F)’ (the region of interest in the forward orientation) vectors this may be due to a failure in transfecting the vector (cell luminescence is below that observed for samples where there are only cells) but the ‘OLFOP(R)’ (the region of interest in the reverse orientation) vector appears to have been successfully transfected and the activity of this region is still low. In conclusion, the functional approach using HEK293 and Odora cell lines also provided little evidence for this region alone habouring a promoter for the four OR genes (hs6M1-16, hs6M1-18, hs6M1-21 and hs6M1-27). More cell lines need to be transfected to confirm whether or not this region does have some kind of promoter activity.

2a: Odora	Relative luminescence, % (2 separate experiments)	
Cells	0.003	0.004
Cells + basic vector	0.521	0.234
Cells + promoter vector	114.949	85.051
Cells + control vector	1.966	1.505
Cells + OLFOP(F)	0.004	0.003
Cells + OLFOP(R)	0.026	0.060

2b: HEK293	Relative luminescence, % (2 separate experiments)	
Cells	0.017	0.002
Cells + basic vector	0.738	0.767
Cells + promoter vector	65.889	58.174
Cells + control vector	98.782	101.217
Cells + OLFOP(F)	0.007	0.006
Cells + OLFOP(R)	0.111	0.096

Table 6.3: Results from pGL3 reporter vector assay. Relative luminescence (%) after transfection of various constructs into (a) rat olfactory sensory neuron cells (Odora) and (b) human embryonic kidney cells (HEK293) calculated by comparison of samples from 2 experiments against average value of lumiscence of cells with the control vector (over the 2 experiments). Samples: cells (only), cells + basic vector (without promoter or enhancer sequence), cells + promoter vector (without enhancer), cells + control vector (with promoter and enhancer), cells + OLFOP(F) (basic vector + putative olfactory promoter sequence in the forward orientation), and cells + OLFOP(R) (basic vector + putative olfactory promoter sequence in the reverse orientation).

6.5. Regulation of expression within the MHC-linked OR cluster

In order to try and locate putative regulatory regions within the MHC-linked OR cluster on a larger scale, the entire sequence of the MHC-linked OR cluster was analysed using two promoter prediction programs, ‘Promoter Inspector’ (Scherf *et al.*, 2000) that predicts regions of the genome containing promoter-like elements and ‘Eponine’ (Down and Hubbard, 2002) that predicts transcription start sites. Regions immediately flanking such predicted start sites are considered putative promoter regions. Both of these programs rely on the assumption that promoters share a common genomic context that can be detected by an algorithm that has been trained using promoter and non-promoter sequences. These programs are a significant improvement on older promoter predictions, leaving the user with fewer false positives, and a much improved detection sensitivity of 40-45%.

The region analysed here included the minor and major MHC-linked OR clusters and flanking sequences located on chromosome 6. As summarised in Figure 6.6, ‘Promoter Inspector’ identified 6 putative promoter regions, which can be considered to be associated with zinc finger protein 311, zinc finger protein 57, RFP, GABBR1, HLA-F and HLA-G. ‘Eponine’ was used in conjunction with 4 threshold values, ranging from 0.9900 to 0.9996; it also predicted promoter regions associated with genes outside the OR cluster (HLA-G, HLA-F, GABBR1) but there are additional regions predicted within the OR cluster.

Within the region, 3 sequences corresponding to the promoters of RFP, HLA-F and HLA-G have been experimentally confirmed. These confirmed promoters were used in order to test the validity of the two promoter programs. For RFP, the experimental evidence places a promoter for this gene at position 4991 (all positions relate to figure 6.5a) (Iwata *et al.*, 1999). The two programs predict this promoter very accurately: there are matches at position 4881-5081 (‘Promoter

Inspector') and position 4956-4991 ('Eponine', threshold <0.9996). Both algorithms also have some success in predicting the location of the promoters for the nonclassical MHC class I loci, HLA-F and HLA-G. Promoters for these genes, which consist of two modules, one consisting of the enhancer A and ISRE (interferon-stimulated response element), and the other consisting of the SXY module, are located at position 804216-804317 and position 908668-908769 (Gobin and van den Elsen, 2000). 'Eponine' has prediction clusters at position 804209-804222 and position 908752-908762 and 'Promoter Inspector' predicts blocks at position 804229-804456 and position 908897-909112; these positions can be considered to relate to experimentally confirmed promoters for HLA-F and HLA-G.

The ability of these two algorithms to independently predict promoters in these cases where experimental evidence is available suggests that searching for olfactory receptor promoters using these two approaches is a valid approach. However, as can be seen from Figure 6.6, 'Promoter Inspector' does not predict any promoters that could be considered to regulate the transcription of olfactory receptor genes. The lack of predictions that could relate to olfactory receptor genes is probably due to the fact that the algorithm has not been trained on any OR promoters, and the fact that the genomic environment of OR genes is very different from most of the genomic environments of known promoters. OR genes are typically located in areas of low GC content whilst promoter regions have typically been found in areas with a high GC content.

In contrast to 'Promoter Inspector', 'Eponine' does predict a number of putative promoter regions within the olfactory cluster, although at the highest threshold value of 0.9996, there are only 3 that are predicted within the major cluster. At this threshold value, however, the experimentally confirmed RFP promoter is not predicted. The RFP promoter, however, is predicted at the lower threshold of 0.9990 and so it can be hypothesized that the 6 predictions within the major OR cluster might represent putative promoter regions. The highly controlled regulation of olfactory

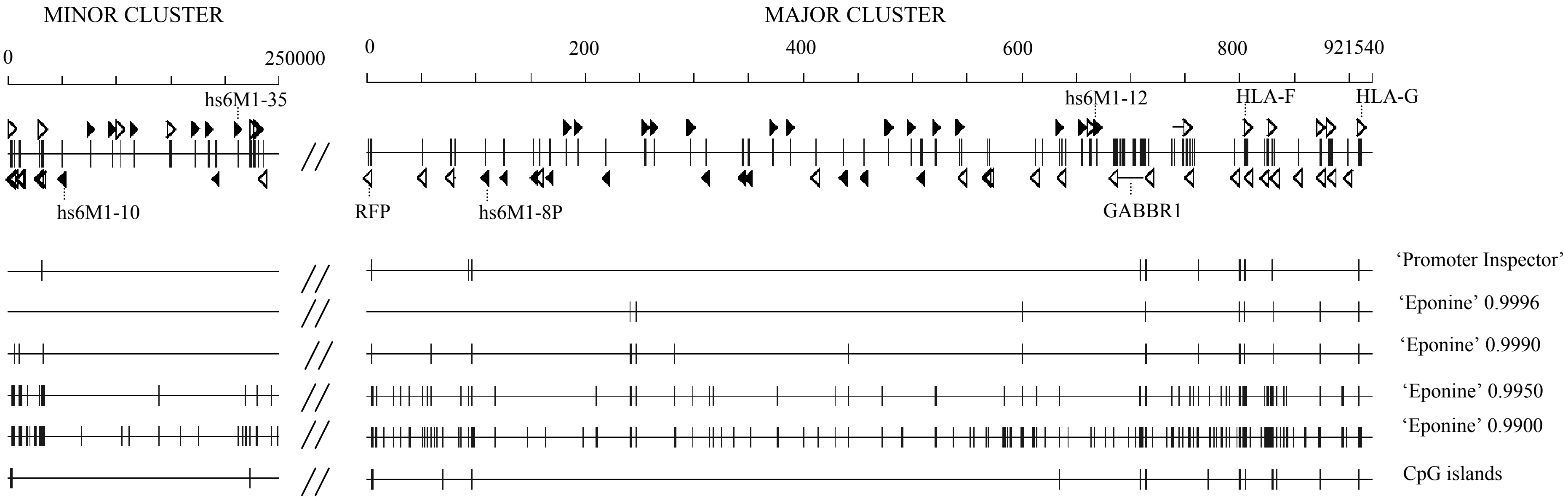


Figure 6.6: Promoter prediction within MHC-linked OR clusters (orientated in a telomere to centromere direction). The minor cluster is located approximately 1200 Kb telomeric to the major cluster. Olfactory receptor genes within the region are marked by black arrows. Genes referred to within the text, and the starting and ending genes of the 2 OR clusters are labelled, for an enlarged diagram of this region showing all gene names see figure 4.6. Plotted below the gene line are the results from the two promoter prediction programs, 'Promoter Inspector' (used at default settings), and 'Eponine' (used at four different thresholds as indicated). The positions of CpG islands are indicated on the bottom line.

receptor genes (only one allele of one OR gene is expressed per olfactory neuron (Chess *et al.*, 1994)) suggests that between OR promoter regions there might be some identifiable form of shared sequence motif. Analyses of these putative promoter regions (which was taken to be the region highlighted by 'Eponine' plus 100 bp upstream and downstream), however, failed to reveal any shared sequence motifs, the only similarity appears to be that these promoter regions are found in areas of high GC-content (ranging fom 54.21-72.07%).

The results from the *in silico* promoter analysis of the MHC-linked OR cluster therefore provided evidence that it appears to be very difficult to predict promoters that could regulate the expression of olfactory receptor genes within clusters. The lack of predictions within these regions is probably due to the fact that no experimental evidence about OR promoters is currently available, which makes it impossible to train the software to detect this type of promoter. The problem is compounded by the fact that these genes are located in areas of the genome that appear to differ from other areas in terms of their genomic environment. (OR genes are typically associated with areas of low GC content). Olfactory receptor gene clusters, therefore, appear to be promoted by regions that bare little resemblance to any other currently known promoters within the human genome.

6.6. Comparison of upstream regions of MHC-linked OR genes

Methods to identify MHC-linked OR promoters are therefore problematic for a number of reasons. On a local scale, alternative transcripts appear to suggest a specific region of sequence could act as a potential bi-directional promoter, but conclusive results indicating that this sequence could act to initiate transcription were not forthcoming. On a large scale across the cluster, the lack of computational predictions can be explained given the lack of olfactory receptor gene promoters in the public databases. Another approach to consider putative promoter regions was therefore developed. This involved extracting a 4 Kb region upstream of the predicted start codon for each MHC-linked OR gene and comparing these regions against the 33 other upstream regions using the alignment program, DNA block aligner ('DBA') which contains an algorithm designed to find conserved blocks of sequence that are flanked by nonconserved sequences of varying lengths (Jareborg *et al.*, 1999). The majority of OR genes for which information about splicing is available have 5' UTR exons located within 4 Kb of their start codon, although there are exceptions to this rule, such as hs6M1-18, -21, -27 and –32. In general, however, OR genes might be expected to have a promoter located within 4 Kb of their start codon, and it was hypothesized that a shared promoter might be identified through shared sequence similarity.

Results from all 34 MHC-linked olfactory receptor genes, however, suggested there was no common element conserved upstream of all these genes: regions with shared nucleotide identity tended to be repeat sequences or represent blocks of sequence that had been duplicated alongside MHC-linked ORs. Figure 6.7 shows the results for hs6M1-16. This shows that there is a high number of MHC-linked OR genes with upstream sequences similar to the sequence found –1500 bp to –1000 bp upstream of the hs6M1-16 gene. This region of sequence, however, contains a 2 repeat elements (an AluSq and a MER42c element) suggesting that this similarity is due to the upstream regions containing repeat elements. The similarity that can be observed between the upstream regions of hs6M1-16 and hs6M1-12, and hs6M1-16 and hs6M1-13P can be attributed to duplication events forming this subfamily: although this does not preclude these sequences having a regulatory function, the lack of conservation of these sequences in the upstream region of other OR genes suggests there is no regulatory sequence motif found upstream of all MHC-linked OR genes.

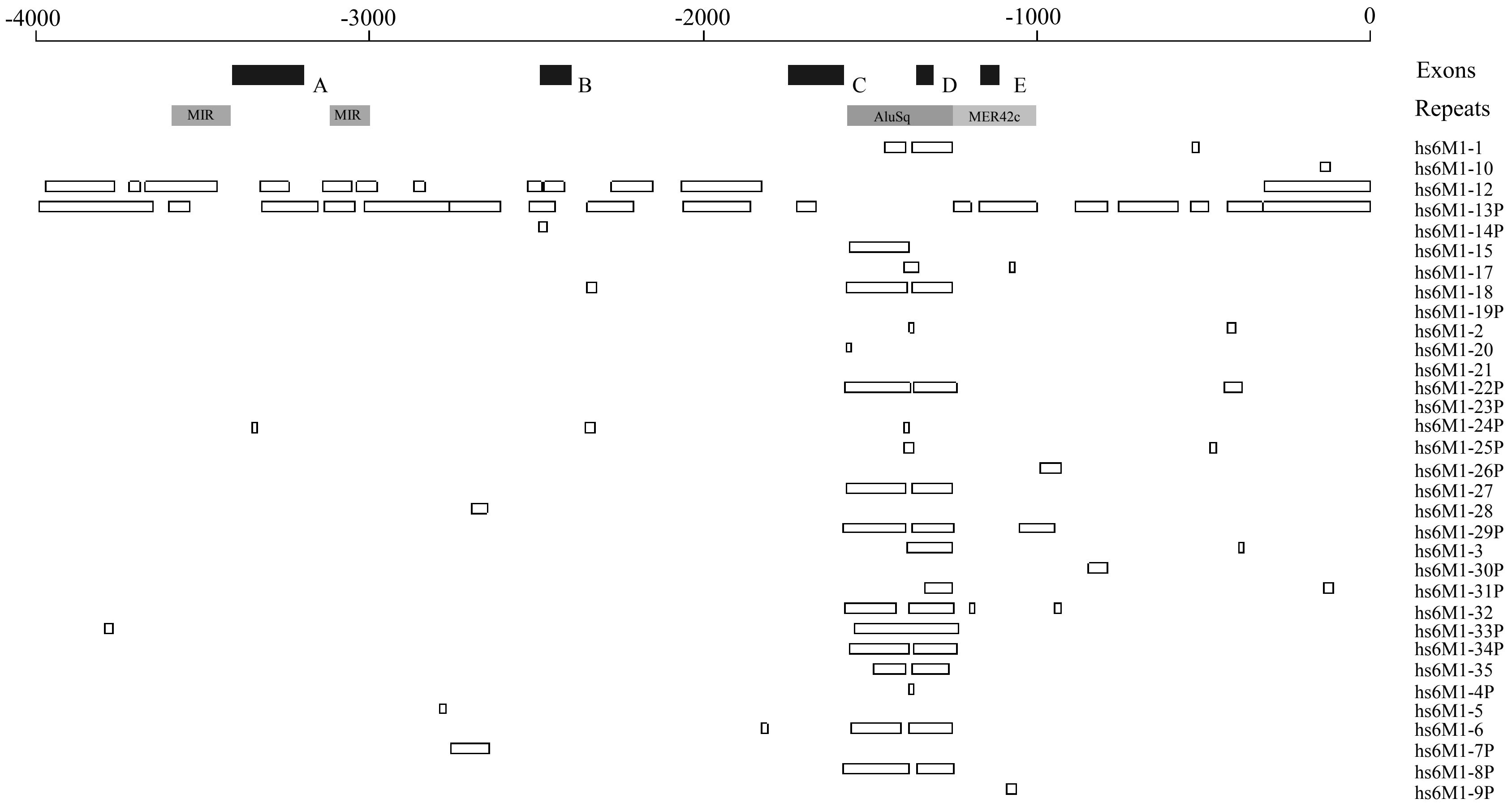


Figure 6.7: ‘DBA’ alignment of the 4 Kb upstream region of hs6M1-16. The scale shows base pairs distance from the proposed start codon of hs6M1-16. Exons and the repeats present within the 4 Kb region are plotted below the scale line. White blocks beneath the repeat line show blocks of sequence that are conserved in upstream regions of other MHC-linked ORs. (Positions are plotted according to where these blocks are found upstream of hs6M1-16, not according to where the block is located upstream of the other gene.)

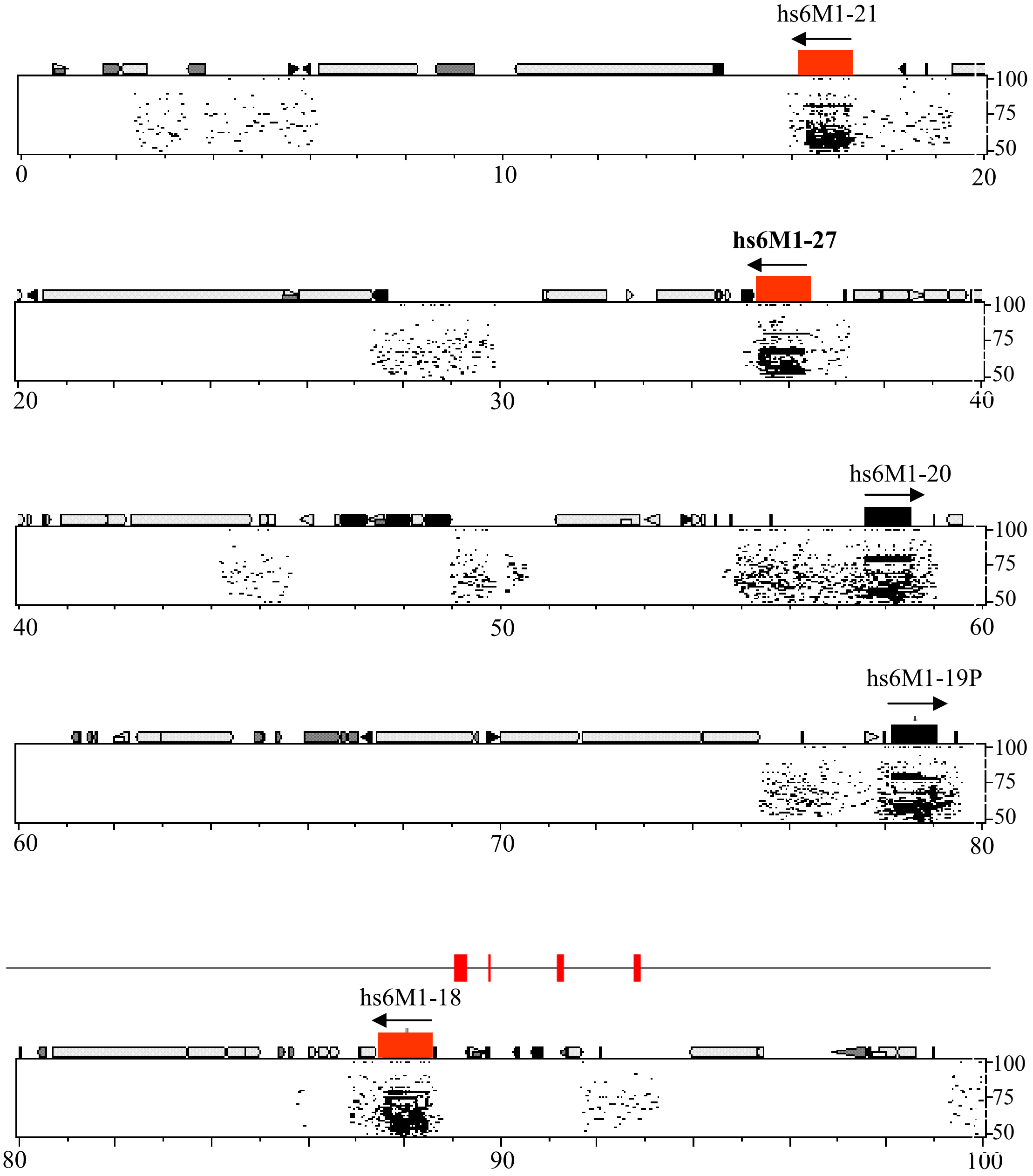
Hs6M1-16 was taken as an example result because information about the 5' UTR exons was available from the group in Berlin. This information meant 5 upstream exons could be compared to see if there was any conservation of these upstream of other MHC-linked ORs. As Figure 6.7 shows, there is little conservation of these exons, with the exception of exon D. However, as this exon appears to be located within an AluSq repeat, it is difficult to consider whether this is a significant observation or whether the sequence similarity is owing to the presence of repeat sequences in upstream regions of the MHC-linked ORs.

6.7. Comparison of upstream regions of human and mouse MHC-linked OR genes

Another approach to consider putative promoter elements for MHC-linked olfactory receptor genes was to compare upstream human sequences against sequences taken from the upstream areas of mouse orthologous genes. Comparative analyses of the mouse and human genomes are expected to identify regulatory sequences, as the 2 species have diverged enough so potential coding sequences can be distinguished from non-coding sequences, but not too much for regulatory sequences to become unrecognisably dissimilar (Koop and Hood, 1994, Baxendale *et al.*, 1995, Hardison *et al.*, 1997, Ansari-Lari *et al.*, 1998). Percentage identity plots (PIPs) (Hardison *et al.*, 1997) generated by comparing 200 Kb stretches of the human MHC-linked major OR cluster against 200 Kb stretches of the mouse MHC-linked cluster were therefore used to identify putative promoter regions.

Figure 6.8 shows a section of the PIP plot produced for the human extended MHC class I region. There are a number of conserved regions around the various olfactory receptor genes that could be involved in some form of regulation. However, this conservation does not necessarily imply function, as comparing the position of clustered conserved elements with untranslated exons within this region it is apparent that the majority of these untranslated exons are absent in mouse.

The possible bi-directional promoter region (Figure 6.8, indicated by the pale yellow box at around 118 Kb), which appears to have the potential to trigger transcription in hs6M1-16, hs6M1-18, hs6M1-21 and hs6M1-27 is also conspicuous in not being conserved within the mouse MHC-linked OR cluster.



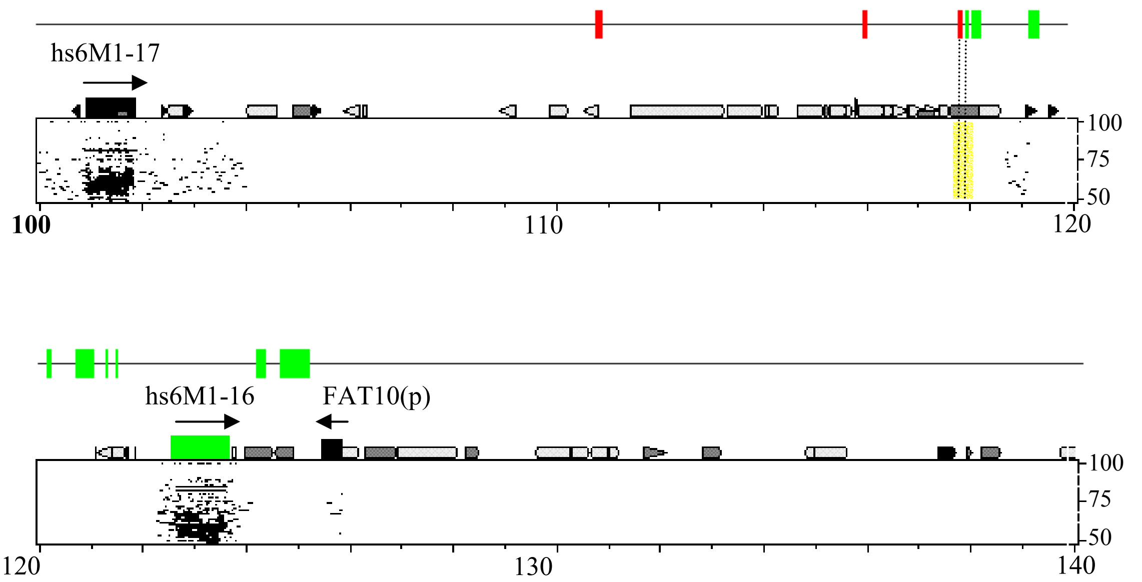


Figure 6.8: PIP plot of a region of the MHC-linked OR cluster. Coding exons and repeats are plotted at the top of each box. Information about untranslated exons (where known) is plotted in a separate track above the box. Untranslated exons that are part of the hs6M1-16 gene structure are plotted in green, whilst exons that are found in hs6M1-18, -21 and –27 transcripts are indicated in red. Within the box, lines plotted at various heights represent segments of sequence found in the mouse MHC-linked OR cluster that are similar to sequences within the human sequence. The height reflects the similarity of the 2 sequences, which can range from 50 to 100%. (Scale on right of box.) The pale yellow box indicates the position of the proposed bi-directional promoter.

Comparison between the mouse and human regions, therefore, produces a number of potential regulatory elements but the lack of conservation in the untranslated exons suggests the situation is more complicated than might be expected from the theory that conservation relates to function. This region, however, may not be the best region to consider as hs6M1-16 appears to lack a true ortholog within the mouse genome. The largest amount of information about exons, therefore, relates to a gene lacking a true ortholog, and it may be that additional information about splicing in hs6M1-17, hs6M1-20, and possibly hs6M1-19P, may reveal a function for the conserved regions located around these OR genes. Information that is currently available about alternative transcripts in hs6M1-21, hs6M1-27 and hs6M1-18, however leaves a large number of conserved regions unaccounted for, and more work is required to elucidate what role these conserved regions play in regulation of OR genes, or indeed, whether they have a role.

6.8. Conclusions

Results from this chapter clearly indicate that a number of the MHC-linked olfactory receptor genes are expressed in tissues other than the olfactory epithelium. This non-specificity of expression has been observed for a number of olfactory receptor genes within the genomes of various organisms (Parmentier *et al.*, 1992, Vanderhaeghen *et al.*, 1993, Vanderhaeghen *et al.*, 1997, Dreyer, 1998). The role olfactory receptor genes perform outside the olfactory system is unknown: one proposal is that ORs are the ‘last digits’ in an area code required for embryo- or organogenesis (Dreyer, 1998). Alternatively, in testis they could be involved in sperm development, sperm kinetics and/ or chemotaxis between sperms and oocytes (Ziegler *et al.*, 2002).

Whatever the role of OR genes in different tissues, clearly, some mechanism is required so the genes can be expressed correctly according to the role they are required to play. Alternative splicing has been observed in a number of MHC-linked olfactory receptor genes, suggesting that alternative versions of olfactory receptor genes with different 5’ untranslated regions may be involved in regulating expression. Alternative usage of 5’ UTR exons has been demonstrated for a mouse olfactory receptor gene (MOR23) where transcription is initiated at 2 different sites (Asai *et al.*, 1996). The MHC-linked OR genes, however, also show alternative splicing within the coding frame. This has not been observed for other OR genes, and how widespread this phenomenon is within the genome is unknown. The suggestion of transcripts that appear to produce proteins lacking transmembrane domains is intriguing: it may be that these shorter proteins are translated and play different roles. Alternatively, short transcripts may be spliced together to form a novel OR protein. This type of somatic recombination mechanism (possibly similar to that of the immunoglobulins) may explain why there are reports of expressed OR pseudogenes, such as hs6M1-24P and hs6M1-14P and Crowe *et al.* (1996).

The regulation of the MHC-linked olfactory receptor genes, therefore, appears to involve several transcriptional start sites, and this may explain difficulties in distinguishing promoter regions for these genes. One olfactory receptor promoter, the Olf-1 site that binds a transcription factor (EBF) expressed solely in OSN and early B-cells has been reported (Wang and Reed, 1993), but the role of this factor in OR expression is debatable since mice lacking the EBF transcription factor develop a morphologically normal olfactory epithelium (Lin and Grosschedl, 1995). This Olf-1 site has been proposed to have a role in the regulation of the chromosome 17 cluster of olfactory receptor genes, alongside 2 other transcription factor sites but experimental evidence supporting this data has not been produced (Sosinsky *et al.*, 2000).

Promoters for the MHC-linked OR gene cluster appear to be elusive. Data from alternate transcripts were used to predict a bi-directional promoter but no promoter activity could be detected, nor could any distinctive characteristics of this sequence be discerned. Comparisons of upstream regions of human-human and mouse-human genes also did not produce data suggesting a discernible transcription start site motif. The lack of sequence similarity between the mouse and human sequences that appear to suggest upstream untranslated exons and transcriptional start sites may have diverged is something that was also reported from a study of the murine P2 cluster (Lane *et al.*, 2001) and from an analysis of the OR cluster flanking the β-globin gene cluster (Bulger *et al.*, 2000).

In conclusion, therefore, MHC-linked olfactory receptor genes are expressed in a highly controlled manner in both the olfactory epithelium, and in other non-olfactory tissues. Within the olfactory epithelium, some form of control must act to ensure that out of around 900 OR genes only 1 allele of one gene produces a functional product, and the regeneration of olfactory sensory neurons within an organism's lifetime means this process must be repeated numerous time. This chapter has presented evidence for alternative splicing that may have some role to play in this process. Promoters for these alternative start sites, however, remain enigmatic, and answers may lie in the control of chromatin structure or in epigenetic mechanisms (such as methylation) rather than in detectable sequence motifs. Lane *et al.* (2001) have suggested a mechanism of control that is based on the idea that each olfactory sensory neuron contains a single OR transcription complex that can only stably accommodate 1 OR gene. This would be similar to the "expression site body" (ESB) observed in *Trypanosoma brucei* (Navarro and Gull, 2001). An active ESB (there are several ESBs but only one is ever active) contains an expression site (ES) to which one from hundreds of variant surface glycoprotein (VSG) genes is transposed. A similar structure that controlled the expression of OR genes would explain the elusive quality of OR promoters and it may also explain the lack of luciferase promoter activity, since the sequence within the construct may not have the conventional properties of a promoter.

Chapter 7

Polymorphism of human MHC-linked ORs.

7.1. Introduction

Within the human species, both the ability to sense various chemicals and how these chemicals are perceived varies widely. The chemical androsterone, for example, cannot be smelt by some individuals, whilst for those who can smell it, the smell is either considered to be similar to urine or sandalwood. Although this difference in perception may be caused by non-genetic factors (for example, a traumatic head injury or an infection of the nasal mucosa can produce damage leading to anosmias), studies of differences in the perception of androsterone and other odorants between fraternal and identical twins have indicated that there is a genetic component of this observed variation (Wysocki and Beauchamp, 1984, Wysocki *et al.*, 1989, Gross-Isseroff *et al.*, 1992). This genetic component of differences in olfactory perception could be located at a number of steps within the olfactory pathway: for example, odorant binding proteins appear to be involved in facilitating the transfer of odorants across the mucus layer to olfactory receptors (Tegoni *et al.*, 2000), so allelic differences in these genes may play a role in differing perceptions of odorants. Similarly, allelic variations in G-proteins (Rana *et al.*, 2001) which are coupled to olfactory receptors and trigger an increase in cyclic AMP when an odorant binds to an olfactory receptor may also be important in individual's differing olfactory abilities.

Within the olfactory system as it is currently understood, however, the large number of olfactory receptor genes within the human genome compared to the number of odorant binding proteins or G-proteins implicated in the olfactory system suggests that the majority of differences in smelling ability caused by genetic factors can be predicted to be due to genetic variations within the repertoire of olfactory receptor genes. This genetic variation within the olfactory receptor family appears to be able to take 2 forms: firstly, there may be variation within the sequence of olfactory receptors (Gilad *et al.*, 2000, Sharon *et al.*, 2000), or secondly, there may be variation in the number of olfactory receptor genes at certain chromosomal sites within the genome (Trask *et al.*, 1998, Linardopoulou *et al.*, 2001).

In order to assess genetic variation that could potentially have a role in some anosmias, two MHC-linked olfactory receptor genes (hs6M1-17 and hs6M1-20) were resequenced in 10 different cell lines. These cell lines were derived from different donors, representing different HLA haplotypes and different ethnic origins. Eight of the 10 cell lines were HLA homozygous, whereas two (BM19.7, BM28.7) were HLA hemizygous (Ziegler *et al.*, 1985, Volz *et al.*, 1992). Other MHC-linked ORs were resequenced by collaborators in Berlin (Anke Ehlers and Armin Volz) and in Cambridge (Simon Forbes).

7.2. Alleles of hs6M1-17

For hs6M1-17, sequence variations were observed at 10 positions within the gene (Table 7.1). Of these 10 variations, all but 2 are predicted to affect the protein sequence of the OR. The most notable variation is at amino acid position 55 where one cell line (BM19.7 (East African)) has a stop codon as opposed to the functional CAG codons found at this position in the other 10 DNA samples surveyed (including DNA from the Human Genome Project). This stop codon, which can be predicted to make the OR gene non-functional, may explain why there have been a number of other changes within this allele.

Allele	1	2	3	4	5
DNA 163 AA 55	CAG Glu	CAG Glu	CAG Glu	CAG Glu	TAG stop
DNA 183 AA 61	TTC Phe	TTC Phe	TTC Phe	TTC Phe	TTT Phe
DNA 265 AA 89	CGC Arg	CGC Arg	CGC Arg	CGC Arg	AGC Ser
DNA 361 AA 121	CGC Arg	CGC Arg	CGC Arg	CGC Arg	TGC Cys
DNA 412 AA 138	CGG Arg	CGG Arg	CGG Arg	CGG Arg	TGG Try
DNA 478 AA 160	CCT Pro	CCT Pro	CCT Pro	TCT Ser	TCT Ser
DNA 521 AA 174	CCG Pro	CCG Pro	CAG Glu	CCG Pro	CCG Pro
DNA 736 AA 246	GTG Val	ATG Met	ATG Met	ATG Met	GTG Val
DNA 762 AA 254	GCA Ala	GCA Ala	GCA Ala	GCA Ala	GCC Ala
DNA 929 AA 310	ATG Met	ATG Met	ATG Met	ATG Met	AGG Arg
Cell lines	BM 28.7 LG2 Genomic	SA H2LCL WT51 YAR	KR3598 OLGA	AMAI	BM19.7

Table 7.1: Polymorphisms observed in hs6M1-17. The differences in the DNA sequences and the amino acid found in the OR protein of the 5 alleles are listed, alongside cell lines found to carry a specific allele.

In addition to the change at amino acid (AA) 55, 5 changes that are only found within this allele can be observed (AA 89: Arg → Ser, AA 61 Phe → Phe (synonymous mutation), AA 121 Arg → Cys, AA 138 Arg → Try, AA 310 Met → Arg). In contrast to the variation that has been generated within this allele, comparing the other 4 alleles only 3 nonsynonymous mutations can be observed (AA 160 Pro → Ser, AA 174 Pro → Glu, AA position 246 Val → Met). The high variation in the allele containing the stop codon compared to the other alleles of hs6M1-17 suggests that selectional forces conserving the structure of OR genes acted less strongly to maintain the DNA sequence of the BM19.7 allele after the mutation at DNA position 163 rendered the allele non-functional. This idea, that the selectional forces acting to conserve the amino acids of the OR protein have been relaxed since the stop codon mutation partially mediates against the hypothesis that OR pseudogenes or fragments or these pseudogenes play a functional role within the human genome (Chapter 6).

7.3. Alleles of hs6M1-20

Resequencing hs6M1-20 in the 10 individuals led to 6 different combinations of alleles being found (Table 7.2). 8 substitutions at the DNA level were observed. These consisted of 1 silent mutation at amino acid position 255 and 7 changes predicted to code for different amino acids. 3 of these changes appear to be nonpolar amino acid for nonpolar amino acid (Val → Phe, Phe → Leu, Val → Ile), whilst the other 4 changes can be predicted to have more of an impact upon the protein structure (Leu → Pro, Phe → Ser, Leu → Arg, Ser→ Cys).

In 3 samples, this gene appears to be heterozygous. The DNA sample from population KR3598, for example, has 2 alleles that differ at DNA position 362. The difference between the alleles in the SA and OLGA cell lines is even greater: they differ at 5 and 4 positions respectively.

Allelic combination	1	2	3	4	5	6
DNA 139 AA 47	GTC Val	GTC Val	TTC Phe	TTC Phe	GTC Val	GTC Val
DNA 167 AA 56	CTT Leu	CTT Leu	CCT Pro	CCT Pro	CTT/CCT Leu/Pro	CTT/CCT Leu/Pro
DNA 311 AA 104	TTC Phe	TCC Ser	TTC Phe	TTC Phe	TTC Phe	TTC Phe
DNA 339 AA 113	TTC Phe	TTC Phe	TTG Leu	TTG Leu	TTC/TTG Phe/Leu	TTC/TTG Phe/Leu
DNA 359 AA 120	CTC Leu	CTC Leu	CGC Arg	CGC Arg	CTC/CGC Leu/Arg	CTC/CGC Leu/Arg
DNA 362 AA 121	TCT Ser	TCT Ser	TGT Cys	TGT/TCT Cys/Ser	TGT/TCT Cys/Ser	TGT/TCT Cys/Ser
DNA 475 AA 159	GTA Val	GTA Val	ATA Ile	ATA Ile	GTA/ATA Val/Ile	ATA Ile
DNA 765 AA 255	CTT Leu	CTT Leu	CTC Leu	CTC Leu	CTC Leu	CTC Leu
Cell lines	BM28.7 LG2 AMAI Genomic	WT51	BM19.7 H2LCL YAR	KR3598	SA	OLGA

Table 7.2: Polymorphisms observed in hs6M1-20. The differences in the sequences of the 5 alleles are listed, alongside cell lines found to carry a specific allele.

7.3. Alleles of other MHC-linked OR genes.

The polymorphisms observed in hs6M1-17 and hs6M1-20 were compared against other polymorphisms in MHC-linked OR genes (data generated by Anke Ehlers and Armin Volz, Berlin, and Simon Forbes, Cambridge, summarised in table 7.3 (Ehlers *et al.*, 2000, Ziegler *et al.*, 2000)). In all 52 point mutations were detected. On the nucleotide level, the majority of these changes are transitions (C → T, A → G) rather than transversions (C → A, C → G, G → T, T → A). Contrary to what might be expected, however, these point mutations appear to be largely equally distributed throughout codon positions: in fact there are slightly more mutations that alter the first nucleotide of a codon than the other 2 nucleotides within a codon. The apparent lack of a selective pressure producing more mutations in the third and second nucleotide positions than in the first nucleotide position means that the majority of nucleotide mutations produce nonsynonymous changes within the OR protein.

Hs6M1-4 is similar to hs6M1-17 in having a functional and a non-functional allele found within different cell lines. (Non-functional alleles are both disrupted by a stop codon.) In contrast to hs6M1-17, however, the non-functional allele has not amassed a number of mutations that are not found in the other alleles of this gene. This suggests the selective pressure is stronger on hs6M1-4. Reasons for this include the idea that this could be a more recent mutation that has not been around for a long enough period of time to accumulate the same number of mutations as hs6M1-17. Alternatively, the pseudogene allele of hs6M1-4P could be functional outside the olfactory system whereas the pseudogene allele of hs6M1-17 could be totally non-functional.

Although, not included in table 7.3, functional and non-functional alleles were observed in hs6M1-19P. At this locus in addition to the pseudogene form found in the genomic sequence, there appears to be a functional form without the 16 base pair deletion that renders hs6M1-19 a pseudogene.

Across the cluster, the number of alleles of MHC-linked OR genes ranges from 2 (hs6M1-1, hs6M1-10, hs6M1-18) up to as many as 7 (hs6M1-17, hs6M1-20), although the average is 3-4. It is interesting to note that members of the same subfamily appear to contain a similar number of point mutations, for example, hs6M1-1 and hs6M1-10 both have 1 point mutation, whilst hs6M1-12 and hs6M1-16 have 4 and 3 respectively. This may reflect similar evolutionary pressures acting on members of a subfamily, although these events occur at different amino acid positions in the 2 proteins in both cases.

Name	No. of	No. of	Position in	AA change	DNA	Codon
	point	alleles	consensus seq.		change	position
	mutations					
hs6M1-1	1	2	105	Leu → Leu	**A → G**	3
hs6M1-3	4	4	108	Ala → Thr	**G → A**	1
			221	Gln → Arg	**A → G**	2
			223	Val → Ile	**G → A**	1
			256	Ile → Met	**A → G**	3
hs6M1-6	6	3	71	Tyr → His	**T → C**	1
			108	Ala → Thr	**G → A**	1
			117	Ser → Ser	**G → A**	3
			143	Val → Ala	**T → C**	2
			211	Leu → Leu	C → G	1
			215	Ala → Thr	**G → A**	3
hs6M1-10	1	2	232	Gln → Arg	**A → G**	2
hs6M1-12	5	4	19	Pro → Pro	**A → G**	3
			29	Phe → Leu	**T → C**	1
			37	Leu → Leu	**A → G**	3
			48	Ala → Val	**C → T**	2
			78	Gln → Gln	**A → G**	3
hs6M1-15	3	3	79	Met → Val	**A → G**	1
			277	Thr → Thr	**C → T**	3
			294	Asp → Asn	**G → A**	1
hs6M1-16	3	3	62	Ser → Ser	**C → T**	3
			63	Asn → Asp	**A → G**	1
			208	Pro → Pro	**C → T**	3
hs6M1-18	1	2	162	Ala → Thr	**G → A**	1

Name	No. of point mutations	No. of alleles	Position in consensus seq.	AA change	DNA change	Codon position
hs6M1-20	8	7	47	Val → Phe	G → T	1
			56	Leu → Pro	**T → C**	2
			104	Phe → Ser	C → G	2
			113	Phe → Leu	C → G	3
			120	Leu → Arg	T → G	2
			121	Ser → Cys	C → G	2
			159	Val → Ile	**G → A**	1
			254	Leu → Leu	**T → C**	3
hs6M1-21	4	4	21	Leu → Trp	T → G	2
			104	Phe → Phe	**C → T**	3
			231	Gly → Arg	**G → A**	1
			236	Phe → Phe	**T → C**	3
hs6M1-4	6	5	11	Ile → Leu	A → C	1
			81	Val → Val	**G → C**	3
			96	Thr → Thr	A → G	3
			179	*Val → Ala*	***C → T***	*2*
			193	*Gln → Stop*	***C → T***	*1*
			203	Ile → Ile	T → A	3
hs6M1-17	10	7	*54*	*Gln → Stop*	***C→ T***	*1*
			60	*Phe → Phe*	***T → C***	*3*
			88	*Arg → Ser*	*C → A*	*1*
			120	*Arg → Cys*	***C → T***	*1*
			137	*Arg → Trp*	***C → T***	*1*
			159	Pro → Ser	**C → T**	1
			173	Pro → Gln	C → A	2
			245	Val → Met	**G → A**	1
			253	*Ala → Ala*	*A → C*	*3*
			309	*Met → Arg*	*T → G*	*2*

Table 7.3: Summary of all polymorphisms found within MHC-linked OR genes. Transitions are indicated in bold, whilst italics are used to indicate where changes are only observed in the non-functional allele of the 2 loci, hs6M1-4 and hs6M1-17.

There are, however, 4 pairs of olfactory receptor genes where polymorphisms are found at the same position relative to the consensus protein sequence (see Chapter 4). At consensus sequence position 104, for example, hs6M1-20 and hs6M1-21 both show polymorphisms (although in –21 it is silent, whilst in –20 a phenylalanine becomes a serine). Other shared positions for polymorphisms include 108, where both hs6M1-3 and hs6M1-6 have a G to A transition which changes an alanine into a threonine. Hs6M1-17 and hs6M1-20 have 2 positions in common where polymorphisms are located, although the changes (at position 120, Leu → Arg and Arg → Cys; at position 159, Val → Ile and Pro → Ser) involve different nucleotide changes.

The distribution of the polymorphic amino acid sites is shown in Figure 7.1. From this it can be seen that the largest number of polymorphisms are found within the first half of the olfactory receptor protein consensus sequence. 5 regions show a significant amount of polymorphism: cytoplasmic region 1 , transmembrane region 2, extracellular region 2 and transmembrane region 3 (which are all located next to each other), and cytoplasmic region 3. These results are surprising in the light of the conservation profile of the MHC-linked OR proteins (Chapter 4): transmembrane region 2 was found to be highly conserved in this conservation profile, and so it might be expected to show a lower percentage of polymorphic sites.

7.4. Single nucleotide polymorphism (SNP) large scale analysis

Data about single nucleotide polymorphisms (SNPs) were extracted from the Ensembl database (Chapter 2) and mapped onto the detailed plots of the major and minor human MHC-linked OR clusters. This revealed a total of 561 SNPs within the major MHC-linked OR cluster (561 per 718800 bp = density of 1 SNP per 1281 bp) and 207 SNPs within the minor MHC-linked OR cluster (207 per 200000 bp = density of 1 SNP per 966 bp). These figures are

Figure 7.1a

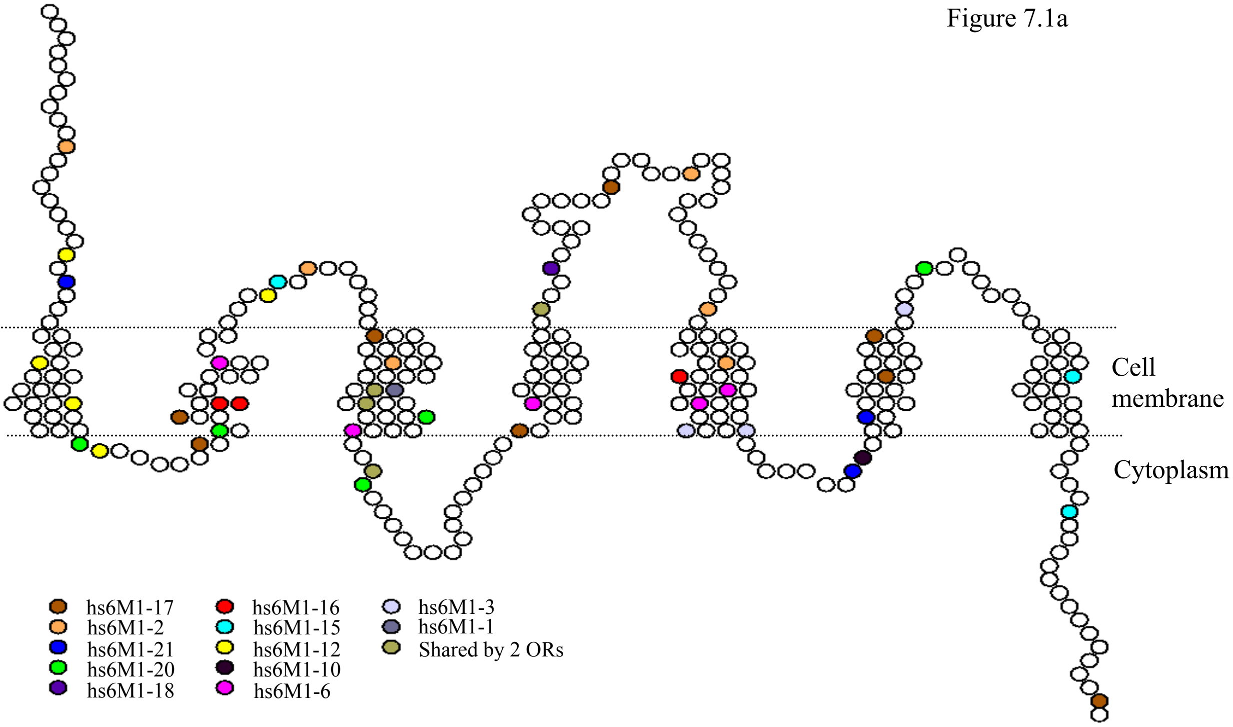


Figure7.1b

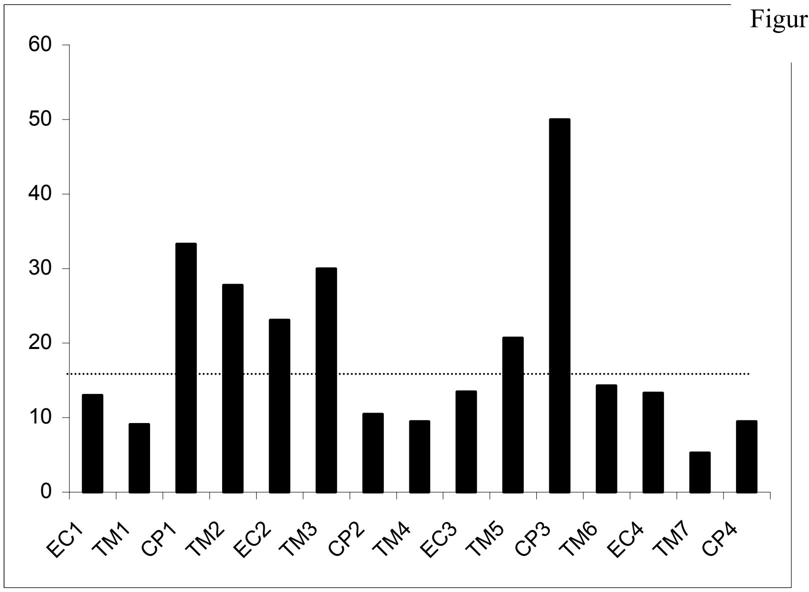


Figure 7.1a: Polymorphisms found in human MHC-linked ORs. The positions of the polymorphisms are displayed with reference to the position of the polymorphism in the consensus protein sequence. The polymorphisms found in different genes are indicated by different coloured residues. The light green colour indicates the 4 positions where 2 ORs both have a polymorphism (details in text).

Figure 7.1b: Percentage of polymorphisms per regions of the consensus MHC-linked OR protein. This shows the number of polymorphic residues with respect to the number of other amino acid residues within the defined section of the protein. The dotted line shows the average figure found across the protein of 16.9%.

slightly higher then the average reported figure of 1 SNP per 1910 bp for the human genome (Sachidanandam *et al.*, 2001), but this difference is likely to reflect the steady accumulation of SNP data since publication of this paper rather than a higher rate of SNPs per base pairs in these regions of the genome. This estimate of the number of SNPs in the human genome was, in any case, fairly conservative, since other studies have suggested the figure may be higher (1 SNP per 721 bp, generalised from 2 Mb of sequence tagged sites (Wang *et al.*, 1998), 1 SNP per 100-300 bp (dbSNP database, *http://www.ncbi.nlm.nih.gov/SNP/.*))

The position of SNPs within the region is shown in Figure 7.2. The distribution within the major OR cluster is striking in its inequality: SNPs appear to be concentrated in a 260 Kb region located at the centromeric end of the cluster, with far fewer SNPs located where the bulk of the olfactory receptor genes are found. In the minor cluster, the distinction in SNP frequency is less pronounced, but there do seem to be fewer SNP in the middle of the cluster. Whether these differences in the distribution of SNPs are significant is debatable, especially since the coverage of SNPs in the public database may represent a partial rather than a complete picture of SNPs within the human genome. With regard to the major cluster, however, it is interesting that there is much higher number of SNPs in the region nearest the MHC, one of the regions within the human genome with the highest amount of variation between individuals (Horton *et al.*, 1998).

The vast majority of these SNPs are located within non-coding sequence. In the major MHC-linked OR cluster, 507 (90.4%) are predicted to exist outside gene loci (both OR genes and other genes within the region), whilst a higher percentage of 95.4% are associated with pseudogene loci and non-coding sequence. The true percentage of SNPs not implicated with affecting coding sequence is likely to be between these 2 figures, since some of the olfactory genes seem to be coding in some individuals and pseudogenes in other individuals.

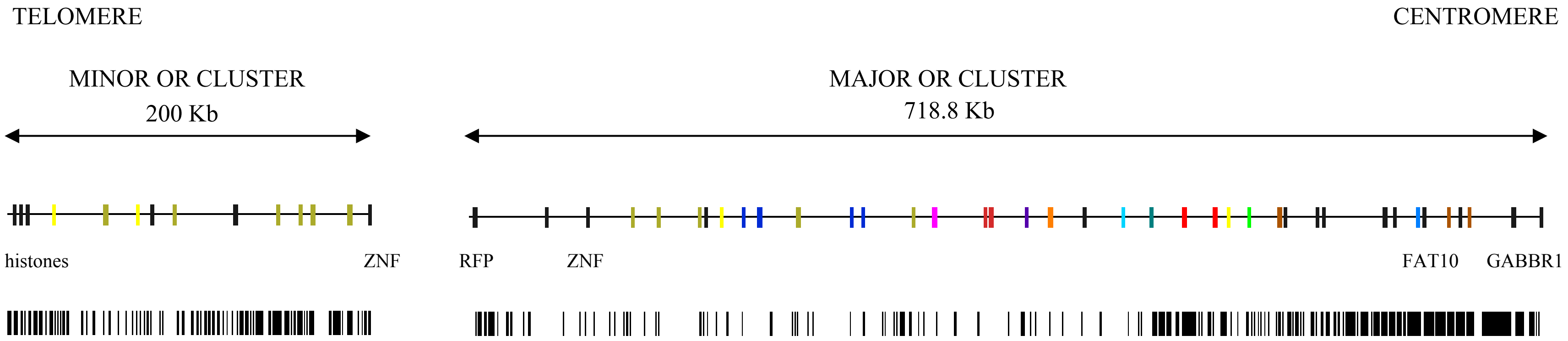


Figure 7.2: Distribution of single nucleotide polymorphisms (SNPs) across the major and minor MHC-linked OR clusters (extracted from the Ensembl database). The major and minor OR clusters were both analysed: OR genes are coloured according to their subfamily designation, or where they do not belong to a subfamily they are coloured pale green. The majority of genes are not labelled: for gene names refer to Figure 4.6. The track underneath the gene name shows the SNP distribution across the region. Within the major OR cluster, this distribution appears to be more dense towards the centromeric end of the cluster. This centromeric end is also the nearest end to the MHC, a well-characterised variable region within the human genome.

Within the minor cluster, 94.7% (196) of the SNPs are found outside gene loci (both OR genes and other genes within the region), the percentage rises to 97.6% (202) assuming pseudogenes are non-coding. Across the two clusters, 32 of the 768 SNPs are associated with olfactory receptor loci (both genes and pseudogenes). These SNPs were analysed further to see if they added to the allelic diversity already described or to see whether these SNPs can be confirmed by the data already described.

7.5. SNPs of MHC-linked ORs

Analysing the SNP data (Table 7.4) associated with the MHC-linked ORs suggested that the resequencing of many of the MHC-linked alleles had found the large majority of polymorphic sites within the OR genes. Generally, the resequencing strategy found more polymorphic sites than the SNP genome-wide approach, and SNPs detected were found at the same sites as those that had already been identified. For example, in hs6M1-4P, resequencing found 6 point mutations whereas the number of SNPs that were identified was 3. These 3 SNPs had already been uncovered by the resequencing approach.

Hs6M1-17 was the exception to this rule: 3 extra point mutations were identified using the SNP data, meaning this OR gene contains at least 13 point mutations. This high number of point mutations could be attributed to the pseudogene status of hs6M1-17 in some haplotypes. However, it appears that hs6M1-17 has a higher number of point mutations than many of the pseudogenes, as the 8 pseudogenes with identified SNPs have on average 1-2 point mutations. Taking hs6M1-17 as the model for how successfully the SNP data manages to identify point mutations, (assuming 13 mutations, 6 of which were identified by the SNP analysis), the number of point mutations per pseudogene can be estimated as 2-4 suggesting hs6M1-17 has a much higher mutation rate than some OR genes within the cluster.

Name	No. of point mutations	Position in consensus seq.	AA change	DNA change	Codon position
hs6M1-8P	1	130	Tyr → Tyr	T → C	3
hs6M1-35	2	253	Ile → Asn	T→ A	2
		305	Arg → Arg	A → G	3
hs6M1-13P	1	257	Tyr → Tyr	C → T	3
hs6M1-14	4	107	Ser → Ser	C → G	3
		149	Ser → Ser	T → C	3
		220	Ala → Ala	C → G	3
		225	Cys → Cys	C → T	3
hs6M1-32	1	144	Ala → Ala	T → C	3
hs6M1-22P	1	124	Ile → Ile	A → T	3
hs6M1-2P	1	40	Asn → Asn	C → T	3
hs6M1-29P	4	55	Asn → Thr	A → C	2
		117	Ala → Val	C → T	2
		224	Val → Glu	T → A	1
		228	Ser → Leu	C → T	2
hs6M1-31P	1	127	Pro → Ser	C → T	1
hs6M1-30P	1	230	Ala → Thr	G → A	1
hs6M1-19P	1	252	Pro → Arg	C → G	2
hs6M1-17	+3	85	Phe → Leu	T →C	1
		188	Phe → Leu	C → A	3
		227	Pro → Pro	A → G	3

Table 7.4: SNPs in MHC-linked ORs found by searching the public databases.

This high mutation rate of hs6M1-17 suggests selective pressures are more relaxed on hs6M1-17 than any other OR identified in this analysis. One tentative explanation for this could be that, as the non-functional allele of hs6M1-17 begins to propagate throughout the population, the protein is no longer expressed and so the selective pressure on hs6M1-17 is lost. This allows the number of point mutations within different haplotypes to increase dramatically, producing the situation that can currently be observed The contrast between hs6M1-17 and pseudogenes that appear to be pseudogenic in all haplotypes can be explained by hypothesizing that these ‘pseudogenes’ have been recruited for other purposes within the genome: for example, they may be expressed as a 5 transmembrane domain protein, or they may form other genomic structures, for example, CpG islands, like one of the OR genes from the chromosome 17 cluster (Glusman *et al.*, 2000), or nuclear matrix attachment regions (Gimelbrant and McClintock, 1997). Alternatively, as has been suggested for loci within the MHC, it may be that these pseudogenes are maintained as they are involved in generating new alleles through gene conversion (Haino *et al.*, 1994, The MHC Sequencing Consortium, 1999).

The SNP data also provide tentative support for the non-pseudogenic status of hs6M1-14. 4 point mutations were observed in this gene, but these are all present in the third nucleotide of the codons producing the 4 amino acids and these mutations produce no changes to the predicted amino acid that will be translated. This higher rate of codon conservation in the first and second coding positions suggests some form of selective pressure is acting upon this locus. This apparent selective pressure, alongside the high conservation of this locus compared to the mouse OR gene, mm17M1-6 (Chapter 5) appears to imply that this gene may have a functional role in spite of its lack of open reading frame.

A point mutation was also observed in hs6M1-19P. This suggests at least 3 alleles of hs6M1-19P are present within the human species: 2 non-functional alleles, and 1 functional allele observed in the resequencing study. The hs6M1-19P data remain the only data that suggests insertions and deletions may also be present within the MHC-linked olfactory receptor cluster: SNP data do not include these type of mutation events.

7.6. Conclusions

In total, 73 point mutations (52 from resequencing OR genes, 21 from SNP survey) have been described. This produces an average value of 2.2 point mutations per olfactory receptor locus, although there are genes for which no mutations have been reported which brings this average value down. This value can, however, be compared with the figure of 1.7 point mutations per olfactory receptor gene within the chromosome 17 cluster (26 point mutations identified in 15 olfactory receptor genes) (Sharon *et al.*, 2000). The higher frequency within the MHC-linked cluster may reflect the proximity to the MHC, where class I and class II alleles are characterized by an extremely high number of alleles (Bodmer *et al.*, 1999). Proximity to the MHC has been proposed to explain the high variability of the GABBR1 locus (Peters *et al.*, 1998).

The functional implications of these polymorphisms, with the exception of the non-functional alleles caused by stop codons (hs6M1-4 and hs6M1-17) or deletions (hs6M1-19), are difficult to assess. 46 amino acid changes are nonsynonymous in a variety of positions across the protein, variability is not just restricted to the position where the ligand is thought to interact with the protein. The functional importance of these amino acid variations cannot be predicted. Within the transmembrane domains implicated in ligand binding, however, variations can be predicted to be likely to have a very large effect, since even conservative amino acid changes (such as Val → Ile in a mouse OR, transmembrane domain 5) result in different preferences for odorant binding (octanal → heptanal in the mouse OR) (Krautwurst *et al.*, 1998).

Differences in the ability to sense different odorants within the human species are clearly present and it is likely that a large amount of this heterogeneity is caused by variations in olfactory receptor gene repertoires. It is clear, however, that compared to some gene families, such as those involved in immune defence and those involved in development, the olfactory receptor gene family is likely to be under a lower amount of selective pressure. This is because, although the olfactory system will contribute to survival chances, mutations in the immune defence system or developmental processes are likely to have a larger effect on an organisms' survival chances (Trask *et al.*, 1998, Gilad *et al.*, 2000, Sharon *et al.*, 2000). In the light of these lower selectional pressures, it is both possible to imagine non-functional olfactory receptor alleles spreading quickly across the population, and it is also possible to imagine that an olfactory receptor gene could amass a large number of alleles that would produce a number of proteins with no significant difference to an organisms' survival or mate-finding chances. Considering these 2 possibilities, therefore, it appears that the number of polymorphisms within the MHC-linked OR cluster is fairly low. This could be explained by the recruitment of olfactory receptor genes into other biological systems or it could imply that the mutational rate within olfactory receptor gene clusters is lower than that found in other regions of the genome. The mutational rate within OR clusters may be lower than that found within other areas of the genome because, rather than diversity being generated by a high number of alleles within a moderate number of genes, it may be that diversity is generated by a high number of genes.

Chapter 8

Comparison of MHC-linked ORs and other human ORs.

8.1. Introduction

The MHC-linked OR cluster was found to contain 34 loci coding for olfactory receptor genes or pseudogenes. These 34 genes, however, are only a small subset of the entire human OR repertoire, which was estimated as having 500-1000 members (Buck, 1992). These OR genes were found to be largely clustered within the human genome (Ben-Arie *et al.*, 1994, Buettner *et al.*, 1998, Trask *et al.*, 1998, Bulger *et al.*, 1999), and FISH analysis suggested these clusters were spread over most chromosomes (Rouquier *et al.*, 1998). Any analysis of the MHC-linked OR cluster, therefore, must take in the relationship of this cluster to other OR genes located within the human genome.

Major questions that a comparison of the MHC-linked OR genes against other OR genes in the human genome aimed to answer concerned the evolution of the MHC-linked cluster, whether this cluster can be regarded as distinctive from other OR genes within the human genome, and whether the linkage between the MHC and the OR cluster is a recent event or whether it has been maintained over evolutionary time. Two routes for the diversification of OR genes within the genome have been suggested: local duplication (Ben-Arie *et al.*, 1994, Glusman *et al.*, 2000) or intrachromosomal transfers of genetic material (Trask *et al.*, 1998, Mefford *et al.*, 2001). A comparison of the MHC-linked ORs against other ORs within the human genome, therefore, should reveal whether this cluster evolved through local duplications or intrachromosomal transfers. Comparison of the MHC-linked ORs against other ORs should also reveal whether this is a distinct cluster with a distinct MHC-related function. It has been suggested that the MHC-linked ORs might play a role in determining MHC-based odours, and this detection of odours may be important in influencing mate choice (Jacob *et al.*, 2002). MHC odours have been detected in many species, including rats, mice (Carroll *et al.*, 2002), humans (Wedekind *et al.*, 1995), salmons (Landry *et al.*, 2001) and sticklebacks (Reusch *et al.*, 2001). As the linkage between a cluster of OR genes and the MHC has been conserved, there may be a functional reason for this conservation and MHC-linked ORs may be solely responsible for the perception of these MHC odours. Comparison of the MHC-linked ORs against other ORs, therefore, may suggest this cluster is distinct from other ORs, with an evolutionary history tightly connected to that of the MHC.

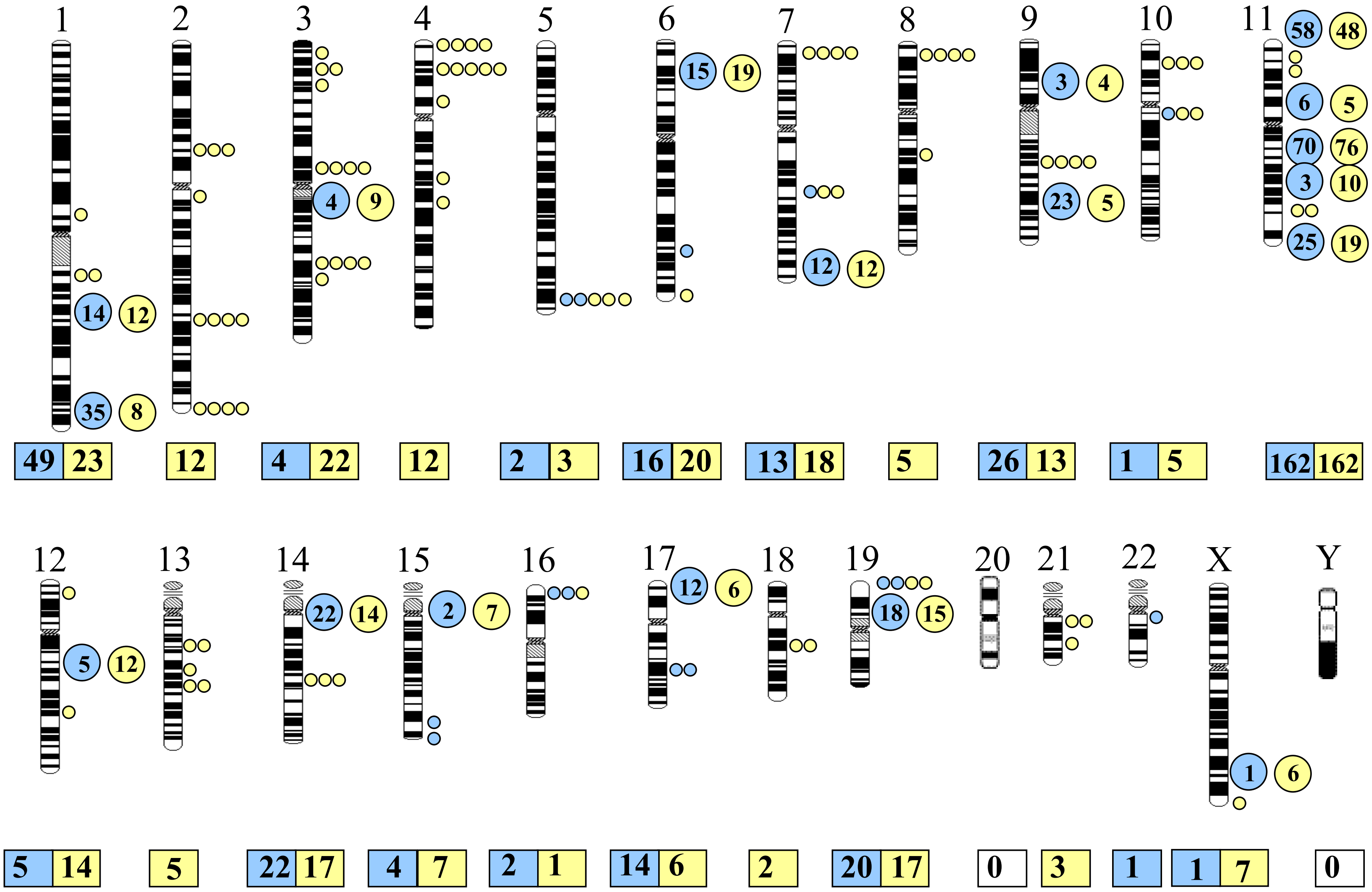
In order to compare the MHC-linked ORs, a database of OR genes was constructed and this database was used to try to resolve these questions. (Another human OR database has been published: this was not used in the following analysis as a large amount of data had already been collected prior to the publication of this article (Glusman *et al.*, 2001), *http://bioinformatics.weizmann.ac.il/HORDE/*).

8.2. The human OR database (‘ROLF’) and the genomic location of OR genes.

The final version of my human OR database (‘ROLF’) represents a comprehensive attempt to extract all human OR genes from genomic sequence. All OR genes within the database are anchored to a position within the genome: there are a number of reported ORs which could not be anchored within the genome, and so these were not included in the analysis. The final version of the ‘ROLF’ database contains 716 olfactory receptor genes, 341 with open reading frames or with fragments of open reading frames. 375 of the loci represent pseudogenes or incomplete pseudogenes. The genomic locations of these olfactory receptor genes are shown in Figure 8.1: location was plotted according to a clone’s position in the latest version of the Ensembl database (release 5.28.1, updated March 2002). These figures for the number of olfactory receptors in the human genome are similar to figures reported in two other studies considering olfactory receptors within the human genome. Glusman et al (2001) reported a figure of 797 olfactory receptors that could be localized within the genome, in addition to 82 which could not be localized, and 27 from nongenomic sources, such as ESTs and mRNAs. 317 of these genes were reported as having complete open reading frames (Glusman *et al.*, 2001). In the second study performed using the genome data, a total of 347 full-length olfactory receptor genes with open reading frames was reported (Zozulya *et al.*, 2001).

The similarity of all these figures suggests an estimate of 341 functional olfactory receptors within the human genome can be made with a high degree of confidence. The number of pseudogenes (not reported by Zozulya et al (2001)) is, however, more problematic, as fragmented pseudogenes with a large number of frameshifts and stop codons would not necessarily have been detected using my method for detecting OR genes. In spite of the slightly lower sensitivity, however, the total figure of 716 OR genes is not vastly dissimilar from the reported total of 797 OR genes. Differences between my database and the 'HORDE' database (*http://bioinformatics.weizmann.ac.il/HORDE,* (Glusman *et al.*, 2001)) were observed after the database was completed: there appear to be a number of discrepancies owing to fragmented OR genes not being detected, but there are also cases where the same OR gene appears in the 'HORDE' database twice. Chromosome 6 provides an example of this: the 'HORDE' database records a total of 55 olfactory receptor genes, whilst my database has 36 OR genes located on chromosome 6. On this chromosome the difference in numbers is due to a large number of genes appearing in the 'HORDE' database twice. The 'HORDE' database also contains a number of

Figure 8.1: The distribution of OR genes within the human genome. Genes with complete open reading frames are shown in blue, whilst pseudogenes are represented by yellow dots. Isolated OR genes are shown as small dots with clusters represented as large circles: the number of genes is shown within the circle. Boxes underneath the chromosomes show the total of genes and pseudogenes per chromosome.


1
14
12
35
8
49
23
2
12
3
4
9
4
22
4
12
5
2
3
6
15
19
16
20
7
12
12
13
18
8
5
9
3
4
23
5
26
13
10
1
5
11
58
48
6
5
70
76
3
10
25
19
162
162
12
5
12
5
14
13
5
14
22
14
22
17
15
2
7
4
7
16
2
1
17
12
6
14
6
18
2
19
18
15
20
17
20
0
21
3
22
1
X
1
6
1
7
Y
0

PCR products that were meant to be amplified from certain chromosomes: these PCR products cannot always be related back to genomic sequence on this specific chromosome.

8.3. Description of OR clusters and local duplications

Chromosome 1 has 2 major clusters of OR genes, with 26 ORs (13 with open reading frames and 13 pseudogenes) located at 1q23.2, and 43 ORs (32 complete, 8 pseudogenes, and 3 fragments) located at 1q43. 3 other OR loci are found on chromosome 1: 1 pseudogene and 1 pseudogene fragment at 1q21.1 and 1 pseudogene at 1p13. In the 2 major clusters, there are groups of related OR genes that appear to be descended from the same ancestral OR that appears specific for a particular cluster: for example, hs1M1-31, -3, -18 and -6 are all located on 1q43 and they all share over 70% protein identity. The independent evolution of the two clusters by local duplications, however, is not supported by other relationships of chromosome 1 olfactory receptor genes. Hs1M1-9, located on 1q23.2 is closely related (84%) to hs1M1-2 on 1q43, whilst hs1M1-34 (1q23.2) is closely related (over 80% protein identity) to a group of OR genes located on chromosome 1q43.

12 olfactory receptor pseudogenes, located in 3 major regions (2p13.2, 2q24.2, and 2 q37.3), were found on Chromosome 2. Of these chromosome 2 OR genes, there are 3 pairs of genes which are closely related (over 80% shared amino acid identity) to each other. Two of these pairs are located on the same chromosomal region, however, the third pair (hs2M1-1P and hs2M1-2P) are found in different sections of the chromosome (2q11.2 and 2p13.2), suggesting these genes did not arise through local duplication, as can be predicted for the other pair of OR genes.

Chromosome 3 is also dominated by OR pseudogenes: 22 are located in 2 major regions, along with 4 functional OR genes. These OR genes can be divided into 5 subgroups, with members sharing a protein sequence identity greater than 60%. The majority of closely related genes within these subgroups are found within the same chromosomal region: for example, hs3M1-21 and hs3M1-20 have an identity of 95% on the amino acid level. As with chromosome 2, however, this shared location for shared identity is not the case for all related ORs: hs3M1-6P is found in the 3p25.3 region, whilst hs3M1-7P is found in the 3p12.3 region, but these 2 ORs are 84% identical on the protein level.

12 pseudogenes are located on chromosome 4. The majority of these (8) are found in the 4p15.33-4p16.2 area, and they are all very similar (pairwise shared protein identities range from 59% to 89%). The other OR loci on the chromosome are 3 fragmented pseudogenes, and there is also 1 complete pseudogene (hs4M1-6P).

2 complete OR genes with open reading frames and 3 pseudogenes are found in the 5q35 region of chromosome 5. Within this small cluster, 1 pair of genes are closely related to each other (87.3%) suggesting a local duplication was responsible for increasing the number of genes at this locus. The rest of the genes are not very similar to other OR genes on this chromosome.

OR genes were found in 3 locations on chromosome 7, 7p22.1-7p21.3, 7q22.1, and 7q34-35. The largest cluster is located on 7q34-35: this region contains 1 pair of OR genes with a shared identity greater than 90%, and 2 groups of OR genes with shared protein identities that are greater than 60%. Another group of OR genes on chromosome also share an identity that is over 60%: these OR genes are located in both the 7p22.1-7p21.3 and the 7q22.1 regions suggesting that there has been an exchange of genetic material between these 2 chromosomal regions.

Chromosome 8 has one region, 8p23.1, containing 4 OR pseudogenes. Three of these pseudogenes are closely related (over 80%), suggesting a local duplication event. The fourth pseudogene is less closely related to these 3 pseudogenes (around 50% protein identity) but it is still very dissimilar from the fragment found in the 8q21.13 region.

Chromosome 9 has OR genes located in 4 regions, 9p13, 9q22.2, 9q31 and 9q33. There are a number of small gene clusters or pairs that appear to have been formed by local duplication events, for example, hs9M1-15, -21 and –16 share over 70% protein identity and are all found in region 9q33. Independent evolution of these clusters, however, is not supported by a large group of OR genes which all share over 60% identity: ORs found in both region 9q31 and region 9p13 are members of this group.

6 OR loci are located in 2 locations on chromosome 10, 10p13 and 10q11.21. Within the 10p13 region, the 3 pseudogenes appear to be slightly similar with a protein identity of > 50% but within the 10q11.21 the 3 genes (1 functional OR, 2 pseudogenes) are not very closely related.

Chromosome 11 is the chromosome on which over 45% of the entire OR repertoire of the human genome is located. With the exception of 4 genes, these OR genes are located within 5 major clusters located at 11p11.2, 11p15.4-5, 11q11, 11q12.1-3, 11q13.4, and 11q24.2. The majority of these ORs located in a cluster are most closely related (over 60%) to ORs from the same cluster or from the adjacent cluster in the case of ORs on 11q12.1-3 and 11q11. There are a number of exceptions to this rule: for example the closest relative of hs11M1-101 (11q24.2) is hs11M1-104 which is located on 11q13.4 but shares a protein identity of 71% with hs11M1-101. In general, however, the closest relative of OR genes on chromosome 11 tend to be found within the same cluster.

Chromosome 12 has one major cluster of OR genes located at 12q13. There are 2 pairs of closely related OR genes (over 70%) located within this cluster, but the rest of the genes appear highly diverged from other genes within the cluster. 8 OR genes of the 19 located on chromosome 12 are more closely related to ORs located on other chromosomes.

5 OR pseudogenes are located on chromosome 13. Two of the pseudogene fragments located near hs13M1-1P are very similar to this gene. The other 2 pseudogenes located on 13q21.32 are not closely related to OR genes on this chromosome: they are closely related (over 90% shared protein identity) to 2 OR genes on chromosome 3q11.2 and 3p26.3.

Chromosome 14 has one major cluster of olfactory receptor genes located in the 14q11.2 region. This cluster contains 22 functional OR genes and 14 genes predicted to be pseudogenes. A number of these genes within the cluster appear to share a common evolutionary history: 27 out of the 36 have a protein identity of greater than 60% with other olfactory receptor genes located in this cluster. The additional 3 OR genes found on the chromosome are located at 14q22.1; they do not appear to be closely related.

11 OR genes are located on chromosome 15. The majority of these are found in 15q11.2, although 2 pseudogenes are located in the 15q26 region. Two pairs of genes within the 15q11.2 region can be predicted to have arisen through local duplication (hs15M1-8P and hs15M1-9P (68.6% identical), and hs15M1-4 and hs15M1-5P (84% identical)), but the rest of the olfactory receptor genes on this chromosome have no distinctive relationship to each other.

Chromosome 16 has 3 OR loci located on 16p13.3. Two of these loci are closely related (84.7% shared protein identity) suggesting one local duplication event. This may have led to the degeneration of one of these loci which is a pseudogene owing to its lack of a starting methionine. The third OR gene is found in the same chromosomal region but it appears unrelated to the other 2 OR loci.

The cluster of olfactory receptor genes located on chromosome 17 is among the best characterised olfactory receptor clusters in the human genome. This cluster is located on 17p13.3 and contains 18 OR loci, 12 of which appear to be functional and 6 appear to be pseudogenes. (In this analysis one of the fused pseudogenes reported is considered to be functional as it possesses an open reading frame.) This cluster contains a number of genes that are closely related to other genes within the cluster, and 4 subfamilies of OR gene sharing greater than 60% protein identity with other subfamily members can be discerned. One pair of OR genes (hs17M1-13 and hs17M1-6) even has a shared protein identity of 98.7% suggesting a very recent duplication has occurred within this cluster, although recent duplications within other less well-characterised regions of the genome may have been discounted as allelic variations rather than considering these variations as coming from 2 different genes. Two additional OR genes on chromosome 17 are located on chromosome 17q23: a shared identity of 79.9% suggests they arose through local duplication.

The 2 ORs located on chromosome 18q11.2 are less than 60% similar to each other. There are, however, similar to 2 genes located on chromosome 21 (hs21M1-2P and hs21M1-3P) and 2 genes located on chromosome 14 (hs14M1-23P and hs14M1-11P). These pairs of genes are located within the same chromosomal region on their respective chromosomes, suggesting these genes may have proliferated by block duplications between chromosomes.

Chromosome 19 has a cluster of OR genes located on 19p13.11-19p13.3. Within this cluster, there is one group of 5 closely related genes all sharing protein identities of greater than 85%. In addition to this closely related group, there are also 4 pairs of OR genes within the cluster with shared protein identities of over 74%.

Chromosome 21 has 3 OR pseudogenes located at 21q21-21q22. None of these pseudogenes appear closely related to each other, although 2 of these genes are implicated in a block duplication also involving chromosome 18 and chromosome 14. Chromosome 22 has 1 functional OR located at 22q11.21, whilst 8 OR genes were found on chromosome X. They are all located at Xq26.2 or Xq28 region, but the protein sequences show little shared identity (all below 60%).

Finally, there are 2 chromosomes in the human genome that completely lack even fragments of OR genes. Chromosome 20 and chromosome Y either have completely lost any trace of their old OR gene repertoire, or OR genes were never present on these chromosomes.

8.4. The genomic environment of OR clusters

Olfactory receptor genes, therefore, are distributed across the human genome. The majority of these genes (87.5%) are located in regions that are defined as having olfactory receptor gene clusters. (A cluster in this case is classified as a region of the genome, defined according to cytogenetic position, that contains 5 or more olfactory receptor loci. Clusters located in adjacent cytogenetic bands were merged with the cluster immediately telomeric of them.) In order to consider the genomic environment of these clusters, clones from the various clusters were analysed for repeat content and GC content using RepeatMasker. Average figures from each cluster are shown in Table 8.1.

These results show the majority of OR gene clusters occupy a similar genomic environment to that observed for the MHC-linked OR cluster. This genomic environment is characterised by a low GC content (typically 38-40%), a high percentage of LINE repeats (typically over 20%) and a low percentage of Alu elements (typically less than 10%). There are, however, clear exceptions to this generalised environmental profile. Clusters located on 4p16.1-2, 5q34-35, 9p13.2-3, 11q13 and 19p13.2-3 all have a GC content of over 42%, together with a LINE content percentage below the average value and an increased number of Alu repeats. In the case of 5q34-5, 9p13.2-3, and 11q13 these values can be attributed to the high number of base pairs per OR gene within these clusters which suggests that these regions may contain a large amount of sequence not associated with OR genes.

Cluster	Gene number	Size of region, bp	bp per OR gene	%age ALU	%age MIR	%age LINE	%age REPEAT	%age GC
1q23.2	26	1091551	41983	3.8	2.27	33.2	48.11	37.38
1q43	43	1584990	36860	4.71	0.85	33.79	51.64	37.82
3q11-12	13	429047	33004	8.53	1.01	24.42	56.88	40.61
4p16.1-2	9	374045	41561	16.03	3.2	11.87	45.79	45.4
5q34-35	5	684155	136831	14.66	1.57	22.98	49.3	46.12
6p21	34	918800	27024	7.73	1.43	27.91	49.56	39.54
7q34-35	24	972036	40502	6.19	1.7	27.95	48.1	39.51
9p13.2-3	7	493964	70566	22.89	0.96	18.31	52.32	42.63
9q31-33	28	843276	30117	9.51	3.48	28.24	51.1	40.01
11p15	106	3609464	34052	6.8	1.93	27.06	48.75	40.21
11p11	11	497947	45268	7.59	2.97	42.62	63.91	40.09
11q11-12	145	4987411	34396	5.52	2.06	28.99	49.71	37.49
11q13	13	1140050	87696	11.96	2.94	17.65	49.8	45.72
11q24	44	727056	16524	4.73	2.49	30.45	46.33	37.73
12q13	17	1295197	76188	7.81	2.26	29.66	50.3	40.27
14q11.2	36	1544059	42891	11.67	1.5	24.77	50.52	40.66
15q11.2	9	332649	36961	4.8	1.3	32.49	45.94	38.26
17p13.3	18	578953	32164	7.81	1.24	42.12	59.4	40.87
19p13.2-3	33	473432	14346	21.63	1.53	20	54.89	44.36
Xq26	7	351732	50247	4.49	2.93	44.31	59.78	38.47
TOTAL/ AVERAGE	628	22929814	36512	7.97	1.96	28.29	50.45	39.85

Table 8.1: The repeat and GC content of OR clusters within the human genome. The number of genes in each cluster was established using the OR database: clones were positioned according to the latest version of the Ensembl database (5.28.1). The size of the region was calculated by adding the clones in each region together: no allowance was made for overlaps, with the exception of the 6p21 region which was analysed in detail (Chapter 4). The percentage repeat content and GC content was taken from the RepeatMasker output.

The clusters located on 4p16.1-2 and 19p13.2-3, however, cannot be considered to contain a large amount of sequence that is not associated with OR genes, since their base pairs per OR gene are either similar to that obtained from other clusters (41561 bp in the case of 4p16.1-2) or well below the average value obtained from other clusters (in the case of 19p13.2-3, 14346 bp per OR gene). The reasons for these two clusters to differ from the others are difficult to discern. It may be that these results are an artifact produced by using unfinished sequence. Alternatively, in the case of the chromosome 4 cluster it may be due to the lack of predicted functional OR genes: this lack of OR genes has reduced selectional pressure on the region, allowing Alu insertions to be maintained. The difference in the genomic environment of the 19p13.2-3 region cannot, however, be explained by the lack of functional genes: it is similar to other clusters in terms of the gene to pseudogene ratio. If these figures are not skewed by using unfinished, non-contiguous sequence, therefore, the cluster on chromosome 19 appears to represent a distinctively different genomic environment for OR genes to be found within. This difference could represent the fact that chromosome 19 is clearly distinct from other chromosomes in the genome: it possesses the most CpG islands (43 per 1 Mb on average), is the most GC-rich chromosome and whilst it makes up 2% of the genome it contains 5% of the Alu content of the genome (IHGSC, 2001). Alternatively, the OR genes on chromosome 19 may have followed a different evolutionary pathway to other OR genes within the human genome or it may be that these genes are regulated in a different manner to other OR genes.

8.5. Phylogenetic analysis of human OR genes

The evolutionary relationship between the MHC-linked ORs and other ORs within the human genome was investigated through constructing a phylogenetic tree of all 716 ORs identified within the human genome. The sheer size of the OR repertoire was problematic in this respect since there was no available alignment program that could handle this amount of data and, similarly, tree-building programs do not generally handle this amount of data. It is true that small sections of sequence from each protein could be aligned and used to construct a tree, but this would have meant ignoring the majority of the data, and it was felt that the majority of information should be used in attempting to reconstruct OR phylogenies. The problems involved in dealing with such a large data set were therefore solved with the assistance of the Pfam protein database team at the Sanger Institute (Kevin Howe and Alex Bateman).

The 716 olfactory receptor proteins were aligned using the method developed by the curators of the Pfam database (Sonnhammer *et al.*, 1997, Sonnhammer *et al.*, 1998). This involves generating a high quality 'seed' alignment from a small representative non-redundant sample of the protein sequences. The remainder of the protein sequences are then aligned using a hidden Markov model (HMM) based on the profile obtained from the 'seed' alignment. The full alignment produced using this process was then assembled into a phylogenetic tree using the program 'QuickTree' (Howe *et al.*, unpublished). The 'QuickTree' program is based on an efficient implementation of the Neighbor-Joining algorithm: it does not improve on the limitations of the Neighbor-Joining Method (Chapter 2), it just allows very computationally heavy processes to be run on a desktop machine.

Figure 8.2 (pullout, at back of thesis) shows the phylogenetic tree produced using this method. 500 bootstraps were performed, but the limitations associated with all Neighbor-Joining Trees also apply to this tree. In order to check the tree, results were compared with trees made for each chromosomal repertoire of OR genes constructed using the 'ClustalW' program for alignments and the maximum parsimony method for phylogenetic reconstruction. Unless otherwise stated, relationships observed in figure 8.2 were all observed in these chromosomal trees (data not shown).

An initial observation that can be made from the phylogenetic tree is that ORs located on the same chromosome (highlighted in the same colour) tend to cluster together. The majority of OR

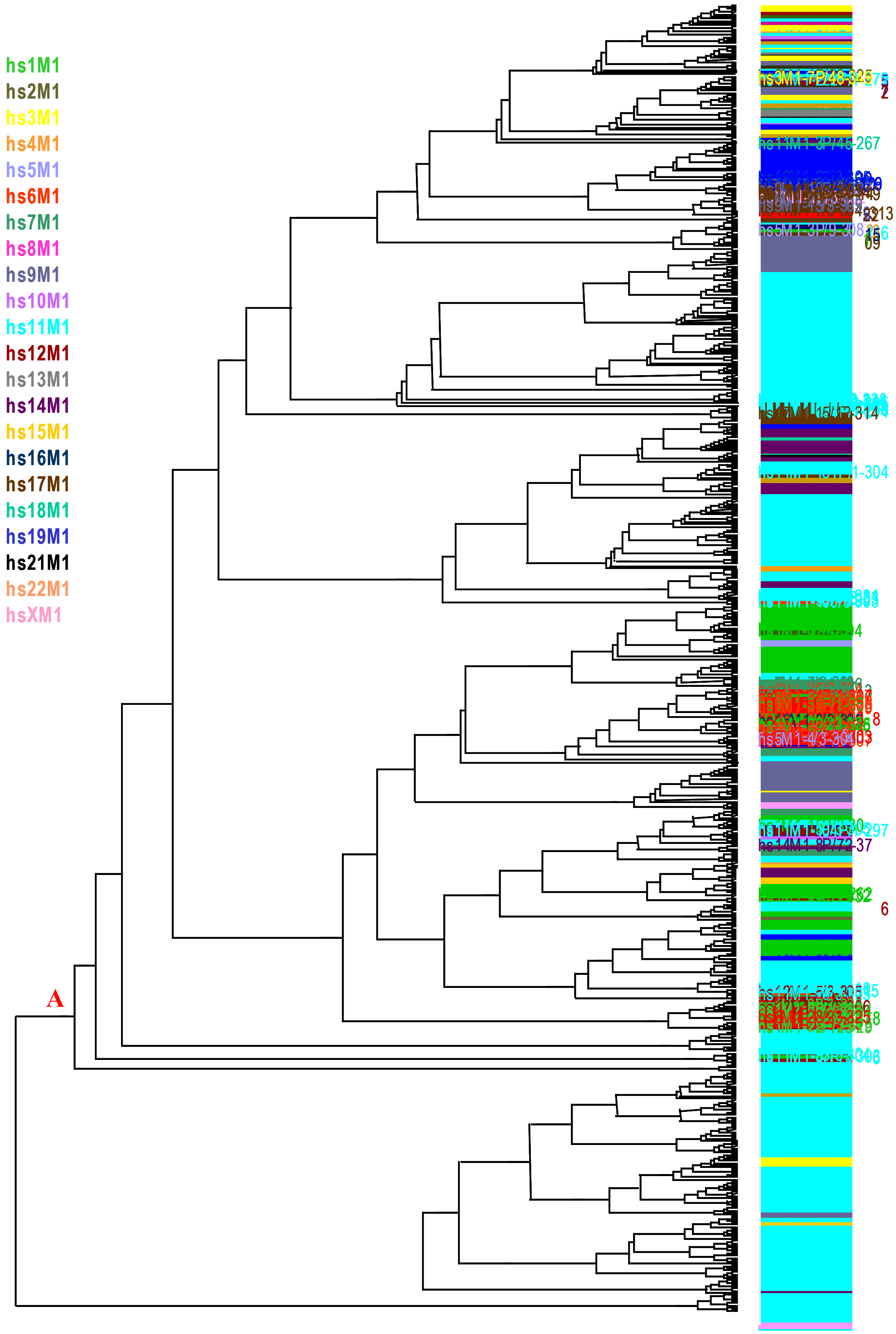


Figure 8.2. Phylogenetic tree of all ORs within the human genome. (Larger version of this tree, showing all gene names is included as a pullout at the back of the thesis.) Tree was constructed using a HMM alignment and the program ‘QuickTree’. ORs on different chromosomes are shown in different colours (key above). Bootstrap values are shown on the larger version of this tree.

genes (500 out of 716 OR genes = 69.8%), therefore, are more closely related to ORs on their chromosome than ORs located on other chromosomes. These small clusters are well supported by the bootstrap values which tend to be greater than 50% where the branch points are relatively recent.

A second observation is that there appears to be an ancient divide between hs11M1-39 and hs11M1-108P (point 'A' in Figure 8.2). This split could represent a proposed ancient event in the evolution of ORs: a split between Class I OR genes (similar to those found in fish, Freitag *et al.* (1995)) and Class II OR genes (tetrapod-specific ORs). The difference between these 2 classes is considered to be due to the specialization of the Class I ORs to detect water soluble odorants and Class II ORs to detect airborne odorants. The closest relatives of the Class I ORs defined by in *Xenopus laevis* by Freitag *et al.* (1995), however, are not found within the region of the phylogenetic tree that would be predicted if this branch point did represent the Class I-Class II split. The division of the phylogenetic tree at this point, therefore, cannot be explained by the Class I-Class II split, and with a bootstrap value of 0% it is not a divide that can be classed as statistically significant.

8.6. The evolutionary origins of the MHC-linked OR cluster

The majority (24 out of 34) of the chromosome 6 MHC-linked OR genes are clustered in 1 group. This group consists of OR genes from both the major and the minor cluster and it also contains a number of OR genes from different chromosomes: 4 from chromosome 1q43, 2 from 5q35.3 and hs16M1-3 from 16p13.3. Three other MHC-linked OR genes, hs6M1-19P, hs6M1-20 and hs6M1-27 are located in another distinct cluster within the phylogenetic tree: they have no clear relationship to any other ORs within the genome, although 2 of their nearest relatives are hs11M1-147 (11q12.1) and another MHC-linked OR, hs6M1-21. The other 7 MHC-linked OR genes are found to be associated with a number of OR genes from other chromosomes. These relationships are more tentative than the clustered relationships described for the other 27 OR genes which are supported by high bootstrap values and shared protein sequence similarity. The large phylogenetic tree, however, agrees with the phylogenetic tree constructed in Chapter 4 (Figure 4.1) in postulating a separate origin for the hs6M1-35 gene. Figure 8.2 also suggests a shared ancestor for hs6M1-19P, hs6M1-20 and hs6M1-27, and suggests separate origins for hs6M1-17, hs6M1-18, hs6M1-21 and hs6M1-28.

Origins of the MHC-linked cluster, therefore, remain elusive. It has been proposed that the origins of this cluster are linked to a group of OR genes located on chromosome 1q43 (Glusman *et al.*, 2001), and the association of a large cluster of MHC-linked ORs with 4 ORs from this region tends to support this. This is likely to have followed a transfer of OR genes from chromosome 11 to create this 1q43 region. Chromosome 11 can be considered to be where the ‘founder cluster’ of OR genes was located. This is based on the idea that a ‘founder cluster’ may have existed on this chromosome for a significantly longer period of time to allow the number of local duplications to produce such a large genomic repertoire. Alternatively, the rate and propensity of local duplications may vary between chromosomes and it may be that chromosome 11 was colonized later and the genomic environment of this chromosome allowed the extreme proliferation of OR genes.

8.7. Paralogous MHC regions and the MHC-linked OR cluster

It has been proposed that the olfactory receptor genes form part of a ‘MHC paralogous region’ on chromosome 1 (Shiina *et al.*, 2001). This is an interesting theory as OR clusters on chromosome 9q31-33 and 19p13.11-p13.2 might also be expected to form part of MHC paralogous regions that have been localised to 9q33-34 and 19p13.1-p13.3 (Kasahara *et al.*, 1997, Kasahara, 1999). If OR genes existed as part of the ‘framework’ MHC that duplicated to form these 4 paralogous regions, it should be possible to see some relationship between framework olfactory receptors that were carried alongside the MHC genes in these duplication events. The phylogenetic tree produced in this analysis provides some evidence for this idea, as there are clusters of ORs within ‘paralogous regions’ that do cluster together (albeit with very low bootstrap values), suggesting an ancient origin for framework OR genes that were duplicated alongside the ‘framework’ MHC. A schematic diagram of this proposal for framework OR genes is shown in Figure 8.3. This would account for 29 of the 34 human MHC-linked OR genes: hs6M1-19P, hs6M1-20, hs6M1-21, hs6M1-27, and hs6M1-28 are independent of the evolution of this paralogous MHC region duplication. The other 3 paralogous MHC-linked OR clusters also have OR genes that are independent of this block duplication event. Figure 8.3 also shows that the framework OR genes have expanded to different degrees in the 4 ‘MHC-linked’ clusters. This suggests duplications of different framework OR genes were maintained in each of the 4 clusters. This would have acted to reduce the redundancy of the OR repertoire, allowing the organism to develop different OR genes in different parts of the genome.

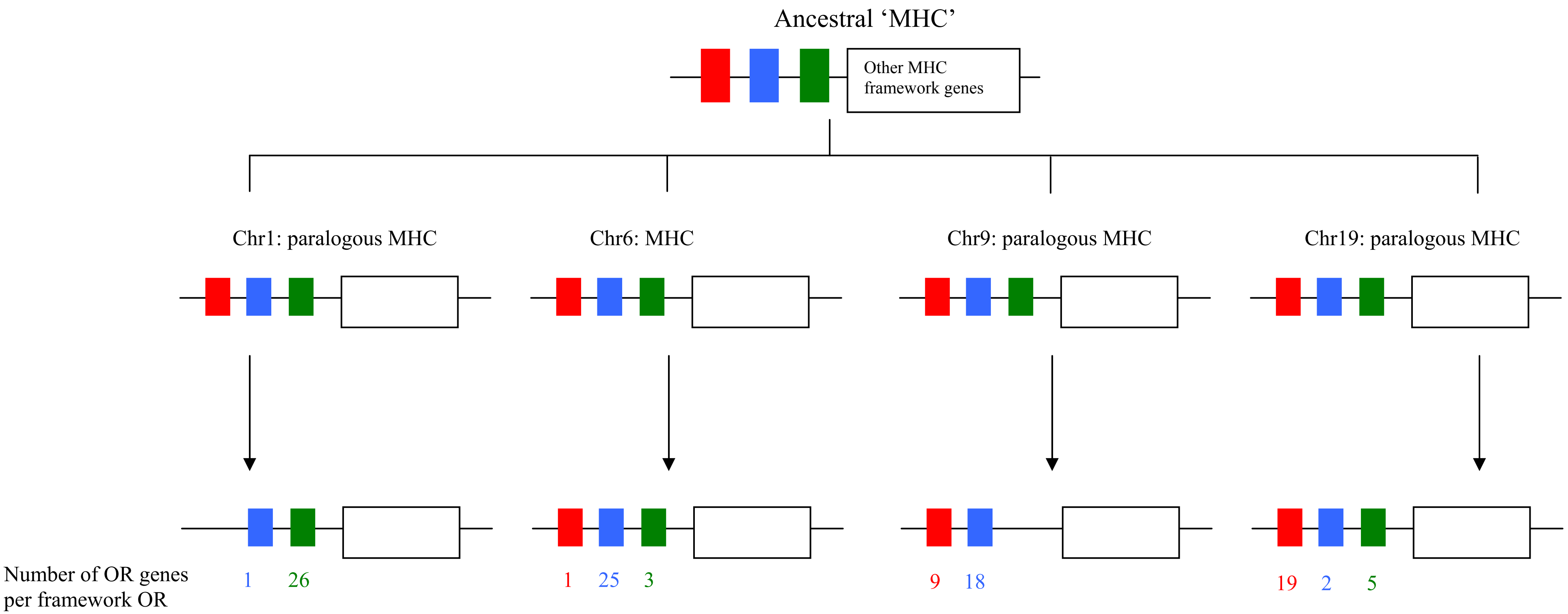


Figure 8.3: Proposed MHC-linked OR evolution on chromosome 1, 6 , 9 and 19. Clustering in the phylogenetic tree suggests 3 framework OR genes were duplicated alongside an ancestral 'framework MHC'. The 3 framework OR genes followed different evolutionary histories in the four chromosomal regions. For example, the red framework gene was lost from chromosome 1, whilst the green framework gene was lost from chromosome 9. The framework OR genes have expanded to different degrees in the 4 clusters.

8.8. OR pseudogenes

374 of the loci that were identified as olfactory receptor genes were considered to be pseudogenes based on one or more stop codon or frameshift, or the lack of an appropriate starting methionine. Work on the chromosome 6 OR cluster, and examples from elsewhere in the genome (notably chromosome 11, data not shown), however, had revealed that some genes that appeared to be pseudogenes in some haplotypes existed in functional form in other haplotypes. This allelic variation means that some OR pseudogenes with only one frameshift or stop codon may exist as functional alleles within the population as a whole. In an analysis of the ORs classified as pseudogenes, approximately 10% of ORs (38 out of 374) contain only one stop codon, whilst 25% (94 out of 374) are disrupted by one frameshift. Assuming all these OR pseudogenes have a functional form, therefore, it appears that another 132 functional ORs could exist within the genome, further increasing the diversity of the OR repertoire.

Other potentially functional forms of OR pseudogene could exist. 16 OR genes were classified as pseudogenic owing to the fact that they do not have a starting methionine after the ‘FILLG’ motif. The alternative splicing that appears to produce a 5 transmembrane version of hs6M1-16 (Chapter 6), however, could be more widespread and it may be that splicing produces functional forms of these OR pseudogenes. The position of frameshifts within all 374 OR pseudogenes also suggests this possibility. Figure 8.4 is a plot showing where OR pseudogenes are disrupted by a frameshift or a stop codon (plotted against the consensus human chromosome 6 OR sequence). This shows a concentration of frameshifts disrupting the sequence just before the ‘FILLG’ motif suggesting there may be less selective pressure on this area of the protein owing to alternative splicing. A similar phenomenon is observed just after the ‘KAFSTCGSHLSVV’ motif. The concentration of frameshifts in this region of the protein is interesting as the alternative splicing

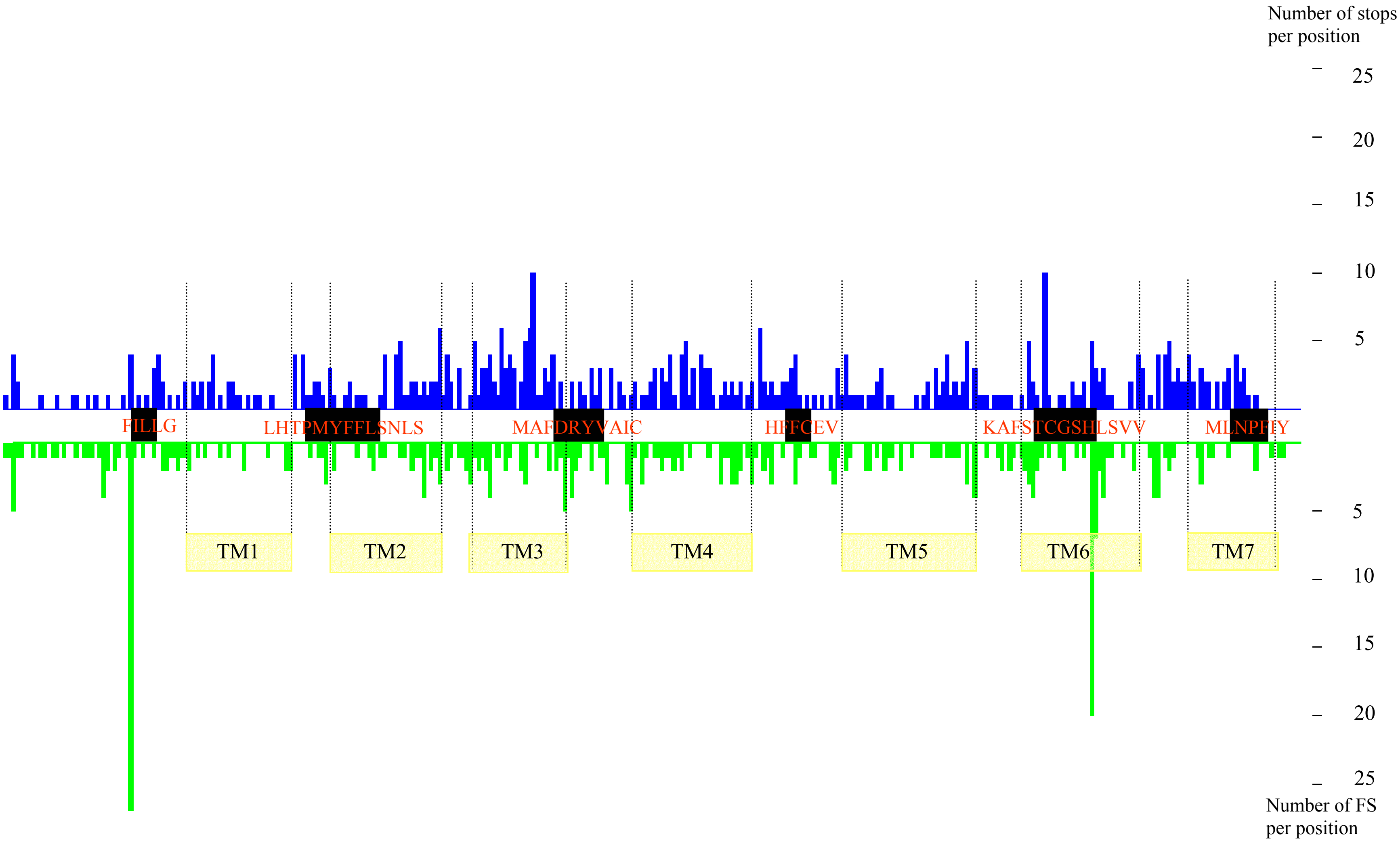


Figure 8.4: Distribution of stops and frameshifts in OR pseudogenes. The position of frameshifts and stop codons in each pseudogene was mapped relative to the position of 6 motifs within the protein. All of these positions were then applied to the chromosome 6 OR consensus protein sequence. The position and number of stops is shown in blue, with the position and number of frameshifts shown in green. The position of the transmembrane domains is indicated by the pale yellow blocks and dotted lines.

of hs6M1-32 (Chapter 6) involved a splice site after this motif. Alternative splicing, and possibly some form of segmental recombination, therefore, is one theory that could be used to explain the high prevalence of frameshifts within these regions of OR pseudogenes. Alternatively, the high number of frameshifts in these positions may be due to higher mutation rates within this section of the protein.

Stop codons disrupting the OR pseudogenes are dispersed slightly more evenly throughout the consensus protein, although there is a high prevalence in the region between motif 2 (‘LHTPMYFFLSNLS’) and motif 3 (‘MAFDRYVAIC’) and a high prevalence around motif 6 (KAFSTCGSHLSVV’) and motif 7 (‘MLNPFIY’). These regions which are located in transmembrane domains 3 and 6, therefore, appear to have higher mutation rates than other regions of the OR protein. This high mutation rate for transmembrane domain 3 correlates with the high variability between MHC-linked OR genes shown in TM3 (Chapter 4) and it also correlates with the high percentage of polymorphisms in alleles of the MHC-linked ORs in TM3 (Chapter 7). The high mutation rate in TM6 was not observed in the MHC-linked ORs, although the cytoplasmic region leading into this transmembrane domain did show a large amount of variability (Chapter 7). Sites of hypermutation and mutational hotspots have been observed in a large number of genes such as the MHC class I and class II genes, the immunoglobulins (Storb, 1996) and venom-derived toxins, such as the conopeptides (Conticello *et al.*, 2001).

In conclusion, therefore, an analysis of the position of frameshifts and stop codons within OR pseudogenes revealed that another 132 genes could be potentially functional within human populations as this is the number of OR pseudogenes only disrupted at one position within the protein sequence. Alternative splicing of mRNA transcripts and/or some form of protein recombination may also render further OR pseudogenes functional: the high number of frameshifts just outside the first and sixth motifs suggests OR genes may function with 5 or fewer transmembrane domains. Evidence for this alternative splicing has been found for the MHC-linked ORs (Chapter 6), and the distribution of frameshifts could imply this is widespread throughout the human OR repertoire. Alternatively, this distribution of frameshifts could be due to a higher rate of mutation at some positions. The distribution of stop codons supports the idea that some positions experience higher mutation rates. These higher mutation rates are present within transmembrane domain 3: this can be explained by the fact that TM3, as a ligand binding region, is a region of hypervariability and it may be advantageous to allow the proteins to diverge at this position, allowing a number of variant proteins and/or alleles to bind with different ligands. Higher mutational pressures or lower selectional pressures also appear to be present in TM6. This region is not predicted to bind to the ligand and so at the present time this higher mutational rate cannot be connected to the function of this part of the OR protein.

8.9. Conclusions

The comparison of MHC-linked ORs against other ORs within the human genome, therefore, resolved a number of issues with regard to the evolution of these genes. Firstly, the majority of human OR genes appear to have evolved through local duplication, although the number of recent duplications does not appear to be that high based on shared protein identities. In addition to this local duplication, however, there are a number of closely related genes found on the same chromosome that are located within cytogenetically distinct bands. Exchange of genetic material between distinct regions on the same chromosome appears to have occurred more frequently than exchange of OR genetic material between 2 different chromosomes, although there are examples where this has occurred (notably between chromosome 18 and chromosome 21).

Within the human genome, OR genes are largely found within clusters. These clusters share a distinctive genomic environment characterised by low GC content, low Alu content and a high

LINE content. Possible reasons for the propensity of OR genes to be found in this type of genomic region were discussed in detail in Chapter 4, although further work on the 2 OR clusters within the genome (4p16.1-2 and 19p13.2-3) that do not conform to this model may provide additional evidence that could be used to refine these explanations.

A phylogenetic tree of all human OR genes was constructed. This supported the idea that local duplications were highly important in the evolution of the OR gene repertoire. It also suggests that the MHC-linked OR cluster does not have an unique position within the human OR genomic repertoire implying (if the sequence-function paradigm holds in this case) that the MHC-linked ORs have no specific functional role that differs from that of other OR genes. The phylogenetic tree also provided evidence for the origins of the MHC-linked OR cluster: 24 of the 34 MHC-linked ORs evolved from one ancestor, 3 (hs6M1-19P, hs6M1-20 and hs6M1-27) evolved from another ancestor, whilst hs6M1-35 appears to have a distinct evolutionary history to the rest of the MHC-linked cluster. The origin of the remaining 6 MHC-linked OR genes remains unclear.

The phylogenetic tree also provides some support for the idea that OR genes were part of a framework MHC that duplicated twice to produce ‘paralogous MHC regions’ within the human genome. ‘Framework’ ORs may have followed different evolutionary pathways in the 4 different chromosomal regions, although relatives of these framework genes are not restricted to these ‘MHC paralogous’ regions: they are found all over the genome.

Finally, an analysis of OR pseudogenes within the human genome suggested an additional 132 genes could be functional across the human species as a whole. Analysis of the position of frameshifts and stops within the human genome suggested that there were positions where these events were most likely to happen. Two explanations for this can be put forward: firstly, there is a higher mutation rate or lower selectional pressures at some positions, and/or, secondly alternative splicing, gene conversion and/or protein recombination act to conserve certain parts of the protein but not others.

Chapter 9

MHC-linked vomeronasal receptor (VR) genes

9.1. Introduction

In addition to the major olfactory epithelium (MOE), where olfactory receptor (OR) genes are expressed, mammals also have another anatomically distinct organ where odorants are perceived. This secondary organ is known as the vomeronasal organ and it is located at the base of the nasal septum connected to the nasal cavity by a small duct (Bargmann, 1997, Keverne, 1999). In rodents, removing the VNO interferes with the detection of pheromones which are defined as chemicals that convey information about reproductive and social status between members of the same species (Wysocki and Lepri, 1991). Changing the ability of the VNO to detect pheromones generally produces changes in mating and aggressive behaviour in rodents.

In contrast to the major olfactory epithelium, where a large amount is known about the pathway from olfactory receptor to olfactory bulb, very little is known about how pheromones are detected in the VNO. Three families of pheromone receptors have been identified, as is the case for ORs, these families all code for proteins belonging to the 7 transmembrane G-protein coupled receptor superfamily. These families, known as V1Rs (Dulac and Axel, 1995), V2Rs (Herrada and Dulac, 1997, Matsunami and Buck, 1997) and V3Rs (Pantages and Dulac, 2000), are all expressed within the rat or mouse VNO and contain about 100-150 members. The V2Rs differ from the other 2 families in that they possess a large extracellular N-terminal domain, similar to that found in extracellular calcium-sensing receptors and metabotropic glutamate receptors.

The existence of functional VR genes within the human species is controversial. Seven different human V1R sequences were identified using PCR and library screening with rodent V1R sequences but this approach failed to produce any functional human V1Rs (Giorgi *et al.*, 2000). The lack of functional human V1Rs is supported by anatomical evidence that suggests although there is a foetal VNO in humans, the adult version is an atrophied, obsolete organ (Tirindelli *et al.*, 1998). On the other side of the controversy, one group reported finding a functional V1R-type pheromone receptor gene expressed in the olfactory mucosa (Rodriguez *et al.*, 2000) suggesting pheromones could be perceived through the main olfactory system as they are in rabbit (Hudson and Distel, 1986) and pig (Rodriguez *et al.*, 2000). This could explain the observation of proposed pheromone-regulated behaviour, such as the synchronization of menstrual cycles among women living together (McClintock, 1971, Stern and McClintock, 1998). There have also been a number of studies that reported finding a structurally intact VNO (Garcia-Velasco and Mondragon, 1991, Moran *et al.*, 1991, Stensaas *et al.*, 1991).

9.2. Identification of VR genes in the human extended MHC class I region

Five pheromone receptor loci of the V1R-type were identified in the human extended MHC. These are found within a genomic region of 662765 bp, separated by a large number of other genes, including a cluster of histone genes (Figure 3.1). In contrast to the olfactory receptor genes located centromeric of this sequence, the VR genes are all pseudogenic, and a large distance separates the loci: hs6V1-5P and hs6V1-1P are separated from the 3 other VR genes by 371 Kb and 250 Kb respectively (Figure 3.1). They are generally dissimilar on the protein level, with shared protein identities typically below 30%. There is, however, evidence that hs6V1-3P and hs6V1-4P may be related as they share a protein identity of 63.9%. hs6V1-2P, the other pheromone receptor located within the core group of 3 is also similar to hs6V1-3P and hs6V1-4P: it has a shared protein identity of greater than 40%. The lack of conservation within these pheromone receptors is not surprising as these genes are all pseudogenic, disrupted by frameshifts and stop codons. The pseudogenic properties of these genes is something that appears to be shared by the majority of the VR type 1 genes within the human genome.

9.3. Identification of VR genes in the human genome

In addition to the 5 MHC-linked VR genes, a further 46 VR genes were identified in the human genome, using a similar method to that used to create the human OR database. This total represents a first attempt to identify human V1R genes; in contrast to the human ORs, it cannot be regarded as a comprehensive identification. Nevertheless, an estimate of the total number of V1R genes within the rat genome suggested a total number between 30 and 100, so the 51 VR loci identified in this project seems likely to represent between 50 to 100% of the human VR type 1 repertoire. In the light of this the number of functional V1R genes within the human genome appears to be very small. Of the 51 VR genes identified, only one, located on chromosome 19 (hs19V1-5), is predicted to produce a functional protein. The majority of the V1 pheromone receptor genes are characterised by pseudogenic reading frames, usually containing several stops and frameshifts.

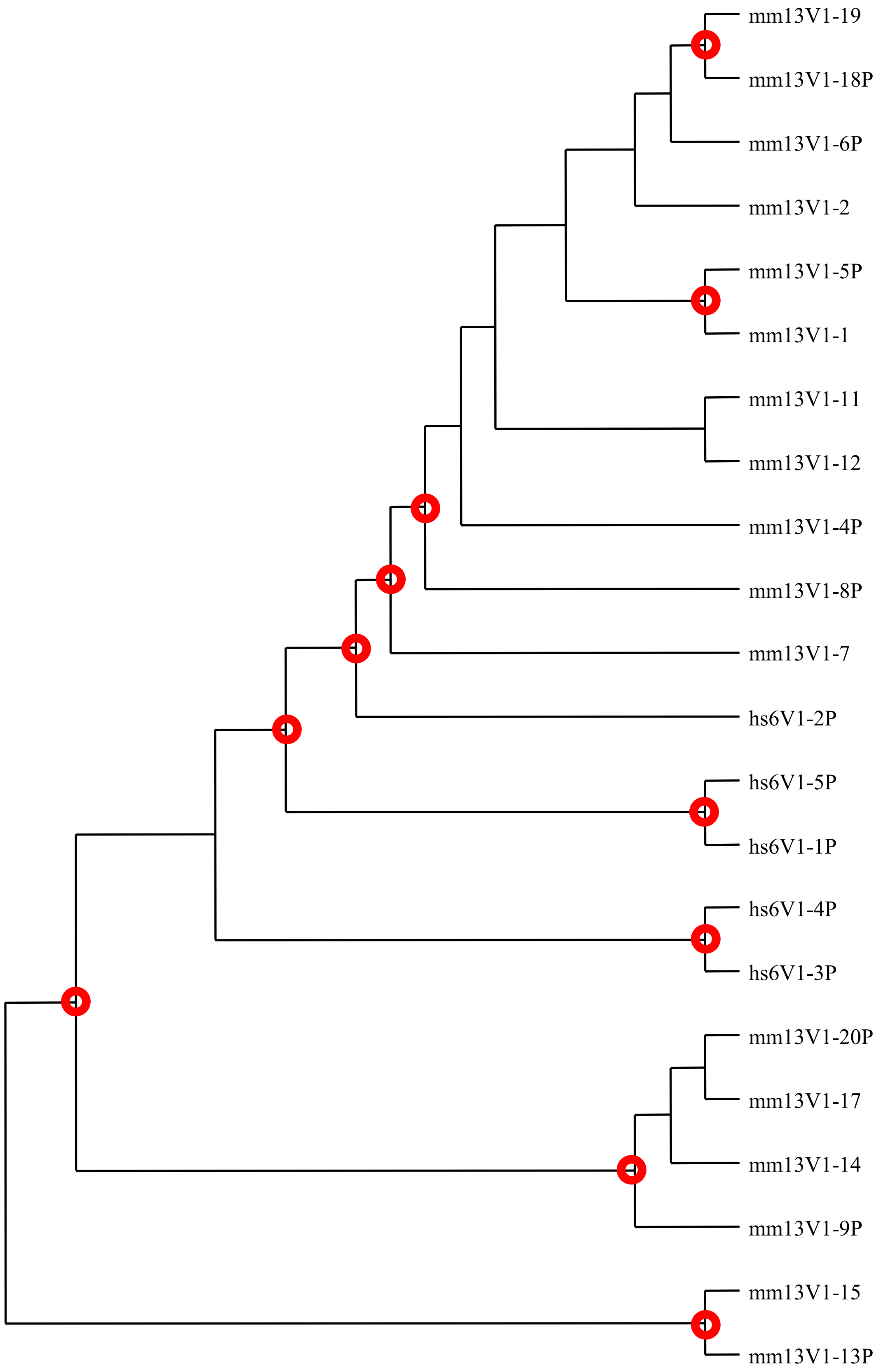
In order to amass this degree of pseudogenicity, therefore, the decline of the VR gene family either started several million years ago, or these genes may have rapidly mutated as the VNO became less important in human evolution. The majority of the loci remaining in the human genome today are found in clusters (36 out of 51 = 70.5%) which suggests the functional VR gene family may have been arranged in clusters, similar to the OR clusters. As with the OR family, the majority of VR genes appear to be more closely related to other VR genes located on the same chromosome than to other VR genes within the human genome, suggesting local duplications may have been the major pathway through which this gene family proliferated.

The proximity of the MHC-linked VR genes to the MHC-linked OR genes suggested that there may be a relationship between clusters of olfactory receptor genes and clusters of pheromone receptor genes. This relationship between VR genes and OR genes can also be observed on other chromosomes, notably chromosome 1 (VR genes: 1q44, OR genes: 1q43), chromosome 3 (VR genes: 3p25.1-2, OR genes: 3p25.3), and chromosome 15 (both VR and OR genes: 15q11.2). This association between VR genes and OR genes could be the remnants of a functionally important association or it could reflect the fact that the 2 gene families both relied on local duplications to provide new members. Clusters of OR genes and VR genes may have ended up in similar genomic locations owing to the ability of these chromosomal regions to promote local duplications.

9.4. Identification of VR genes in the syntenic mouse region (mouse chromosome 13)

Using sequences obtained from the human region, the Ensembl mouse draft sequence was searched for potential orthologs of these human VR genes. With the exception of hs6V1-1P, the most closely related sequences to these genes were found on mouse chromosome 13, localising to a position downstream of a cluster of histone genes and the mouse *Hfe* locus. The mouse VR genes, therefore, appear to have been conserved in a similar position to their human orthologs. Potential orthologs for hs6V1-1P were also localised to this area, but they also appear to be localised to mouse chromosome 7. In contrast to the OR genes, the VR genes are too pseudogenic to accurately predict orthologous relationships, although from the phylogenetic tree (Figure 9.1), one subgroup of VRs are associated with hs6V1-2P, and there are at least 2 additional subgroups not associated with human orthologs.

Figure 9.1 (next page): Phylogenetic tree (maximum parsimony) showing the relationships between the VR genes in the human MHC extended class I and VR genes on mouse chromsome 13. 142 sites were used and 500 bootstrap replicates were performed. The red rings at branch points indicate where bootstrap values are over 70%.


mm13V1-19
mm13V1-18P
mm13V1-6P
mm13V1-2
mm13V1-5P
mm13V1-1
mm13V1-11
mm13V1-12
mm13V1-4P
mm13V1-8P
mm13V1-7
hs6V1-2P
hs6V1-5P
hs6V1-1P
hs6V1-4P
hs6V1-3P
mm13V1-20P
mm13V1-17
mm13V1-14
mm13V1-9P
mm13V1-15
mm13V1-13P

20 mouse VR genes were identified in a segment of sequence approximately 700 Kb in length. 9 of these genes were considered to be functional, with another 11 pseudogenes owing to frameshifts, stop codons and a lack of a starting methionine. The ratio of genes to pseudogenes, therefore, is 0.81 which resembles the gene to pseudogene ratio of the ORs in the human genome (0.8) rather than the gene to pseudogene ratio of the ORs in the mouse genome (3.6). This seems to suggest that the VR family in mouse has undergone or is undergoing the same type of contraction that has been observed for the OR family within the human genome, where the reduction in the fraction of functional OR genes within the OR repertoire has been considered to be due to the reduced functional importance of the sense of smell (Sharon *et al.*, 1999, Rouquier *et al.*, 2000). This is a higher amount of pseudogenicity than that observed by Del Punta *et al.* (2000). If this gene to pseudogene ratio is consistent throughout the mouse genome (and there is no reason that the MHC-linked VR genes cannot be considered representative of this repertoire), it does appear that a higher number of V1R genes than might be predicted are pseudogenes, suggesting a possible decline in the importance of these genes in detecting pheromones. Alternatively, it may be that the poor quality of the draft mouse sequence accounts for a number of these genes lacking open reading frames, although the identification of several complete open reading frames in OR genes identified from the mouse genome sequence seems to suggest this is not the case.

9.5. Conclusions

5 VR pseudogenes were identified within the human extended MHC class I region. These were compared to other VR genes within the human genome, and to VR genes from the mouse syntenic region. The MHC-linked VR genes are similar to other VR genes within the human genome in that the majority of these genes are disrupted by stop codons and frameshifts. There was limited evidence suggesting that some chromosomes have clusters of VR genes and OR

genes located closely together: this could represent the fact that the 2 gene families proliferate or are regulated within the same genomic environment, or it could be due to an ancient relationship between the 2 families.

The mouse MHC-linked VR genes could suggest the 2 families of genes have a shared evolutionary fate. The high number of pseudogenes within the mouse VR repertoire appears to suggest that the VNO could be in the process of being made redundant or at least downsized from the structure it once was. It is tempting to speculate that, as the OR gene family grew in size and function, the VR family was downsized. This suggests the VNO was the primitive chemosensory organ, something that may be supported by observations that a VNO is present in most amphibia, reptiles and nonprimate mammals but absent in birds and adult catarrhine monkeys and apes (Stoddart, 1980). This primitive VNO-non-primitive MOE distinction is, however, too simplistic for three reasons. Firstly, there are a number of other gene families that are VNO receptors and these may not be in decline in the mouse genome. Secondly, the VNO may be variably present in species such as catarrhine primates (Old World monkeys, apes and humans (Smith *et al.*, 2001)), it cannot necessarily be regarded as continually declining throughout evolution. Thirdly, the basis for this suggestion would require a VNO-type structure and VR genes to be more functionally important than a MOE and OR genes within a species that had evolved earlier than mouse or human. There is currently little evidence for VR genes in a species such as *Danio rerio* (zebrafish), although data from the zebrafish genome project (The Sanger Centre in collaboration with the zebrafish community, Zebrafish Workshop 2000) may serve to refute or add evidence to this hypothesis. Primitive or not, it is clear however that similar selective pressures would have acted on both these gene families, and their fate is likely to have been linked in some way.

Chapter 10

Final discussion and conclusions

This thesis has resulted in the identification of 34 MHC-linked olfactory receptor genes in the human MHC extended class I region. These OR genes are located in 2 clusters (the ‘major’ and ‘minor’ cluster) within the MHC extended class I region which contains a number of other gene clusters, including the histones, zinc finger proteins and tRNAs. Both the major and the minor OR clusters are found in LINE-rich, Alu-poor isochores that can be defined across the extended class I region.

A further 56 MHC-linked ORs were identified in the mouse genome. A number of features of both the mouse and human OR genes were investigated. This included their phylogeny, the conservation of structural features within these proteins, and local duplications involved in the creation of new members of this very large gene family. Analysis of the human and mouse MHC-linked OR gene clusters has revealed the syntenic relationships between these genes, allowing the state of the ancestral MHC-linked OR cluster to be considered. Syntenic relationships between OR genes have also revealed putative functional sequences that have been conserved over evolutionary time.

The regulation and expression of the MHC-linked ORs was investigated. MHC-linked ORs were found to be expressed both within and outside the olfactory epithelium. Factors involved in regulating this expression were considered but the variety of approaches only served to suggest that these genes have a mechanism of control not yet observed in any other human multigene family. The polymorphism study of these genes, meanwhile, suggested that the overall level of polymorphism within the MHC-linked ORs is fairly low, although the functional repertoire may vary between individuals as some genes appear to have both functional and non-functional alleles.

A comparison of the MHC-linked ORs against other ORs in the human genome was perfomed. This involved constructing an OR database which suggested that the majority of the diversity had been generated through local duplications rather than exchanges of genetic material between chromosomes. The phylogenetic tree constructed from this data led to the proposal of a 'framework MHC' that included OR genes alongside MHC genes.

Finally, 5 pheromone receptor pseudogenes were identified in the human extended MHC class I region. Putative orthologs of these genes were located in the syntenic region on chromosome 13 in mouse: the majority of these mouse VR genes were pseudogenes, although there were a number of functional loci.

Future work on the MHC-linked OR genes could follow a number of lines:

- Alternative splicing within these genes could be investigated further: a survey across the genome would reveal whether alternative splicing is restricted to MHC-linked ORs or whether it is a genome-wide phenomenon.
- The characterisation of clusters of OR genes in more detail across the genome may give more insight into the regulatory control of these genes: if there are small 'control regions', similar to that observed for the MHC-linked ORs, these may be detectable across clusters.

- Syntenic regions could be investigated in more detail: if these regions are not exons, they may be important in controlling gene expression so they could be assayed for promoter activity.
- Data from the polymorphism study, OR pseudogenes and OR hypervariable regions could be combined to provide an integrated picture of selective forces acting upon OR genes: matrices for OR-specific phylogenetic trees would allow the ancestral relationships of these genes to be investigated further.
- The existence of a 'framework MHC' containing OR genes representative of the 3 groups identified in Chapter 8 could be investigated in an ancient species such as lamprey or hagfish.
- Further work on the pheromone receptor genes in mouse or another species could be performed to investigate the relationship between the 2 olfactory organs in more detail.

Bibliography

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., George, R. A., Lewis, S. E., Richards, S., Ashburner, M., Henderson, S. N., Sutton, G. G., Wortman, J. R., Yandell, M. D., Zhang, Q., Chen, L. X., Brandon, R. C., Rogers, Y. H., Blazej, R. G., Champe, M., Pfeiffer, B. D., Wan, K. H., Doyle, C., Baxter, E. G., Helt, G., Nelson, C. R., Gabor, G. L., Abril, J. F., Agbayani, A., An, H. J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R. M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E. M., Beeson, K. Y., Benos, P. V., Berman, B. P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M. R., Bouck, J., Brokstein, P., Brottier, P., Burtis, K. C., Busam, D. A., Butler, H., Cadieu, E., Center, A., Chandra, I., Cherry, J. M., Cawley, S., Dahlke, C., Davenport, L. B., Davies, P., de Pablos, B., Delcher, A., Deng, Z., Mays, A. D., Dew, I., Dietz, S. M., Dodson, K., Doup, L. E., Downes, M., Dugan-Rocha, S., Dunkov, Dunkov, B. C., Dunn, P., Durbin, K. J., Evangelista, C. C., Ferraz, C., Ferriera, S., Fleischmann, W., Fosler, C., Gabrielian, A. E., Garg, N. S., Gelbart, W. M., Glasser, K., Glodek, A., Gong, F., Gorrell, J. H., Gu, Z., Guan, P., Harris, M., Harris, N. L., Harvey, D., Heiman, T. J., Hernandez, J. R., Houck, J., Hostin, D., Houston, K. A., Howland, T. J., Wei, M. H., Ibegwam, C., Jalali, M., Kalush, F., Karpen, G. H., Ke, Z., Kennison, J. A., Ketchum, K. A., Kimmel, B. E., Kodira, C. D., Kraft, C., Kravitz, S., Kulp, D., Lai, Z., Lasko, P., Lei, Y., Levitsky, A. A., Li, J., Li, Z., Liang, Y., Lin, X., Liu, X., Mattei, B., McIntosh, T. C., McLeod, M. P., McPherson, D., Merkulov, G., Milshina, N. V., Mobarry, C., Morris, J., Moshrefi, A., Mount, S. M., Moy, M., Murphy, B., Murphy, L., Muzny, D. M., Nelson, D. L., Nelson, D. R., Nelson, K. A., Nixon, K., Nusskern, D. R., Pacleb, J. M., Palazzolo, M., Pittman, G. S., Pan, S., Pollard, J., Puri, V., Reese, M. G., Reinert, K., Remington, K., Saunders, R. D., Scheeler, F., Shen, H., Shue, B. C., Siden-Kiamos, I., Simpson, M., Skupski, M. P., Smith, T., Spier, E., Spradling, A. C., Stapleton, M., Strong, R., Sun, E., Svirskas, R., Tector, C., Turner, R., Venter, E., Wang, A. H., Wang, X., Wang, Z. Y., Wassarman, D. A., Weinstock, G. M., Weissenbach, J., Williams, S. M., WoodageT, Worley, K. C., Wu, D., Yang, S., Yao, Q. A., Ye, J., Yeh, R. F., Zaveri, J. S., Zhan, M., Zhang, G., Zhao, Q., Zheng, L., Zheng, X. H., Zhong, F. N., Zhong, W., Zhou, X., Zhu, S., Zhu, X., Smith, H. O., Gibbs, R. A., Myers, E. W., Rubin, G. M. and Venter, J. C. (2000) The genome sequence of Drosophila melanogaster *Science* 287 (5461): 2185-95

Adams, S. M., Helps, N. R., Sharp, M. G., Brammar, W. J., Walker, R. A. and Varley, J. M. (1992) Isolation and characterization of a novel gene with differential expression in benign and malignant human breast tumours *Hum Mol Genet* 1 (2): 91-6

Akao, Y. and Matsuda, Y. (1996) Identification and chromosome mapping of the mouse homologue of the human gene (DDX6) that encodes a putative RNA helicase of the DEAD box protein family *Cytogenet Cell Genet* 75 (1): 38-44

Albig, W. and Doenecke, D. (1997) The human histone gene cluster at the D6S105 locus *Hum Genet* 101 (3): 284-94

Albig, W., Drabent, B., Burmester, N., Bode, C. and Doenecke, D. (1998) The haemochromatosis candidate gene HFE (HLA-H) of man and mouse is located in syntenic regions within the histone gene cluster *J Cell Biochem* 69 (2): 117-26

Allcock, R. J., Martin, A. M. and Price, P. (2000) The mouse as a model for the effects of MHC genes on human disease *Immunol Today* 21 (7): 328-32

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) Basic local alignment search tool *J Mol Biol* 215 (3): 403-10

Amadou, C. (1996) Structure et évolution du bras court du chromosome 6 humain: la region de classe I du Complexe Majeur d'Histocompatibilité et sa partie distale. Ph.D thesis, Université Paul Sabatier, Toulouse III.

Amadou, C. (1999) Evolution of the Mhc class I region: the framework hypothesis *Immunogenetics* 49 (4): 362-7

Amadou, C., Kumanovics, A., Jones, E. P., Lambracht-Washington, D., Yoshino, M. and Lindahl, K. F. (1999) The mouse major histocompatibility complex: some assembly required *Immunol Rev* 167 211-21

Ansari-Lari, M. A., Oeltjen, J. C., Schwartz, S., Zhang, Z., Muzny, D. M., Lu, J., Gorrell, J. H., Chinault, A. C., Belmont, J. W., Miller, W. and Gibbs, R. A. (1998) Comparative sequence analysis of a gene-rich cluster at human chromosome 12p13 and its syntenic region in mouse chromosome 6 *Genome Res* 8 (1): 29-40

Armstrong, L. C., Saenz, A. J. and Bornstein, P. (1999) Metaxin 1 interacts with metaxin 2, a novel related protein associated with the mammalian mitochondrial outer membrane *J Cell Biochem* 74 (1): 11-22

Asai, H., Kasai, H., Matsuda, Y., Yamazaki, N., Nagawa, F., Sakano, H. and Tsuboi, A. (1996) Genomic structure and transcription of a murine odorant receptor gene: differential initiation of transcription in the olfactory and testicular cells *Biochem Biophys Res Commun* 221 (2): 240-7

Babbitt, B. P., Allen, P. M., Matsueda, G., Haber, E. and Unanue, E. R. (1985) Binding of immunogenic peptides to Ia histocompatibility molecules *Nature* 317 (6035): 359-61

Baier, H. and Korsching, S. (1994) Olfactory glomeruli in the zebrafish form an invariant pattern and are identifiable across animals *J Neurosci* 14 (1): 219-30

Bailey, J. A., Carrel, L., Chakravarti, A. and Eichler, E. E. (2000) Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: the Lyon repeat hypothesis *Proc Natl Acad Sci U S A* 97 (12): 6634-9

Bankier, A. T., Weston, K. M. and Barrell, B. G. (1987) Random cloning and sequencing by the M13/dideoxynucleotide chain termination method *Methods Enzymol* 155 51-93

Barth, A. L., Dugas, J. C. and Ngai, J. (1997) Noncoordinate expression of odorant receptor genes tightly linked in the zebrafish genome *Neuron* 19 (2): 359-69

Barth, A. L., Justice, N. J. and Ngai, J. (1996) Asynchronous onset of odorant receptor expression in the developing zebrafish olfactory system *Neuron* 16 (1): 23-34

Bates, E. E., Ravel, O., Dieu, M. C., Ho, S., Guret, C., Bridon, J. M., Ait-Yahia, S., Briere, F., Caux, C., Banchereau, J. and Lebecque, S. (1997) Identification and analysis of a novel member of the ubiquitin family expressed in dendritic cells and mature B cells *Eur J Immunol* 27 (10): 2471-7

Baxendale, S., Abdulla, S., Elgar, G., Buck, D., Berks, M., Micklem, G., Durbin, R., Bates, G., Brenner, S. and Beck, S. (1995) Comparative sequence analysis of the human and pufferfish Huntington's disease genes *Nat Genet* 10 (1): 67-76

Bellefroid, E. J., Marine, J. C., Ried, T., Lecocq, P. J., Riviere, M., Amemiya, C., Poncelet, D. A., Coulie, P. G., de Jong, P., Szpirer, C. and et al. (1993) Clustered organization of homologous KRAB zinc-finger genes with enhanced expression in human T lymphoid cells *Embo J* 12 (4): 1363-74

Bellefroid, E. J., Poncelet, D. A., Lecocq, P. J., Revelant, O. and Martial, J. A. (1991) The evolutionarily conserved Kruppel-associated box domain defines a subfamily of eukaryotic multifingered proteins *Proc Natl Acad Sci U S A* 88 (9): 3608-12

Belluscio, L., Gold, G. H., Nemes, A. and Axel, R. (1998) Mice deficient in G(olf) are anosmic *Neuron* 20 (1): 69-81

Ben-Arie, N., Lancet, D., Taylor, C., Khen, M., Walker, N., Ledbetter, D. H., Carrozzo, R., Patel, K., Sheer, D., Lehrach, H. and et al. (1994) Olfactory receptor gene cluster on human chromosome 17: possible duplication of an ancestral receptor repertoire *Hum Mol Genet* 3 (2): 229-35

Ben-Chetrit, E., Gandy, B. J., Tan, E. M. and Sullivan, K. F. (1989) Isolation and characterization of a cDNA clone encoding the 60-kD component of the human SS-A/Ro ribonucleoprotein autoantigen *J Clin Invest* 83 (4): 1284-92

Berghard, A. and Dryer, L. (1998) A novel family of ancient vertebrate odorant receptors *J Neurobiol* 37 (3): 383-92

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M. and Rodier, F. (1985) The mosaic genome of warm-blooded vertebrates *Science* 228 (4702): 953-8

Bernardi, G. (2001) Misunderstandings about isochores. Part 1 *Gene* 276 (1-2): 3-13

Biamonti, G., Ruggiu, M., Saccone, S., Della Valle, G. and Riva, S. (1994) Two homologous genes, originated by duplication, encode the human hnRNP proteins A2 and A1 *Nucleic Acids Res* 22 (11): 1996-2002

Bignon, C., Roux-Dosseto, M., Zeigler, M. E., Mattei, M. G., Lissitzky, J. C., Wicha, M. S. and Martin, P. M. (1991) Genomic analysis of the 67-kDa laminin receptor in normal and pathological tissues: circumstantial evidence for retroposon features *Genomics* 10 (2): 481-5

Birnboim, H. C. and Doly, J. (1979) A rapid alkaline extraction procedure for screening recombinant plasmid DNA *Nucleic Acids Res* 7 (6): 1513-23

Bjorkman, P. J., Saper, M. A., Samraoui, B., Bennett, W. S., Strominger, J. L. and Wiley, D. C. (1987) The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens *Nature* 329 (6139): 512-8

Blanpain, C., Lee, B., Vakili, J., Doranz, B. J., Govaerts, C., Migeotte, I., Sharron, M., Dupriez, V., Vassart, G., Doms, R. W. and Parmentier, M. (1999) Extracellular cysteines of CCR5 are required for chemokine binding, but dispensable for HIV-1 coreceptor activity *J Biol Chem* 274 (27): 18902-8

Blow, J. J. and Laskey, R. A. (1988) A role for the nuclear envelope in controlling DNA replication within the cell cycle *Nature* 332 (6164): 546-8

Bodmer, J. G., Marsh, S. G., Albert, E. D., Bodmer, W. F., Bontrop, R. E., Dupont, B., Erlich, H. A., Hansen, J. A., Mach, B., Mayr, W. R., Parham, P., Petersdorf, E. W., Sasazuki, T., Schreuder, G. M., Strominger, J. L., Svejgaard, A. and Terasaki, P. I. (1999) Nomenclature for factors of the HLA system, 1998 *Vox Sang* 77 (3): 164-91

Bonfield, J. K., Smith, K. and Staden, R. (1995) A new DNA sequence assembly program *Nucleic Acids Res* 23 (24): 4992-9

Brinon, J. G., Martinez-Guijarro, F. J., Bravo, I. G., Arevalo, R., Crespo, C., Okazaki, K., Hidaka, H., Aijon, J. and Alonso, J. R. (1999) Coexpression of neurocalcin with other calcium-binding proteins in the rat main olfactory bulb *J Comp Neurol* 407 (3): 404-14

Britten, R. J., Baron, W. F., Stout, D. B. and Davidson, E. H. (1988) Sources and evolution of human Alu repeated sequences *Proc Natl Acad Sci U S A* 85 (13): 4770-4

Brookfield, J. F. (2001) Selection on Alu sequences? *Curr Biol* 11 (22): R900-1

Brunet, L. J., Gold, G. H. and Ngai, J. (1996) General anosmia caused by a targeted disruption of the mouse olfactory cyclic nucleotide-gated cation channel *Neuron* 17 (4): 681-93

Brzustowicz, L. M., Lehner, T., Castilla, L. H., Penchaszadeh, G. K., Wilhelmsen, K. C., Daniels, R., Davies, K. E., Leppert, M., Ziter, F., Wood, D. and et al. (1990) Genetic mapping of chronic childhood-onset spinal muscular atrophy to chromosome 5q11.2-13.3 *Nature* 344 (6266): 540-1

Buck, L. and Axel, R. (1991) A novel multigene family may encode odorant receptors: a molecular basis for odor recognition *Cell* 65 (1): 175-87

Buck, L. B. (1992) The olfactory multigene family *Curr Opin Neurobiol* 2 (3): 282-8

Buck, L. B. (1996) Information coding in the mammalian olfactory system *Cold Spring Harb Symp Quant Biol* 61 147-55

Buck, L. B. (2000) The molecular architecture of odor and pheromone sensing in mammals *Cell* 100 (6): 611-8

Bulger, M., Bender, M. A., van Doorninck, J. H., Wertman, B., Farrell, C. M., Felsenfeld, G., Groudine, M. and Hardison, R. (2000) Comparative structural and functional analysis of the olfactory receptor genes flanking the human and mouse beta-globin gene clusters *Proc Natl Acad Sci U S A* 97 (26): 14560-5

Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA *J Mol Biol* 268 (1): 78-94

Burge, C. B. and Karlin, S. (1998) Finding the genes in genomic DNA *Curr Opin Struct Biol* 8 (3): 346-54

Byrd, C. A., Jones, J. T., Quattro, J. M., Rogers, M. E., Brunjes, P. C. and Vogt, R. G. (1996) Ontogeny of odorant receptor gene expression in zebrafish, Danio rerio *J Neurobiol* 29 (4): 445-58

Carrier, A., Nguyen, C., Victorero, G., Granjeaud, S., Rocha, D., Bernard, K., Miazek, A., Ferrier, P., Malissen, M., Naquet, P., Malissen, B. and Jordan, B. R. (1999) Differential gene expression in CD3epsilon- and RAG1-deficient thymuses: definition of a set of genes potentially involved in thymocyte maturation *Immunogenetics* 50 (5-6): 255-70

Carver, E. A., Issel-Tarver, L., Rine, J., Olsen, A. S. and Stubbs, L. (1998) Location of mouse and human genes corresponding to conserved canine olfactory receptor gene subfamilies *Mamm Genome* 9 (5): 349-54

Cavin Perier, R., Junier, T. and Bucher, P. (1998) The Eukaryotic Promoter Database EPD *Nucleic Acids Res* 26 (1): 353-7

Chess, A., Simon, I., Cedar, H. and Axel, R. (1994) Allelic inactivation regulates olfactory receptor gene expression *Cell* 78 (5): 823-34

Chong, J. P., Thommes, P. and Blow, J. J. (1996) The role of MCM/P1 proteins in the licensing of DNA replication *Trends Biochem Sci* 21 (3): 102-6

Christiansen, O. B., Riisom, K., Lauritsen, J. G. and Grunnet, N. (1989) No increased histocompatibility antigen-sharing in couples with idiopathic habitual abortions *Hum Reprod* 4 (2): 160-2

Chu, W. M., Ballard, R., Carpick, B. W., Williams, B. R. and Schmid, C. W. (1998) Potential Alu function: regulation of the activity of double-stranded RNA-activated kinase PKR *Mol Cell Biol* 18 (1): 58-68

Clancy, A. N., Coquelin, A., Macrides, F., Gorski, R. A. and Noble, E. P. (1984) Sexual behavior and aggression in male mice: involvement of the vomeronasal system *J Neurosci* 4 (9): 2222-9

Conzelmann, S., Levai, O., Bode, B., Eisel, U., Raming, K., Breer, H. and Strotmann, J. (2000) A novel brain receptor is expressed in a distinct population of olfactory sensory neurons *Eur J Neurosci* 12 (11): 3926-34

Crowe, M. L., Perry, B. N. and Connerton, I. F. (1996) Olfactory receptor-encoding genes and pseudogenes are expressed in humans *Gene* 169 (2): 247-9

Cuny, G., Soriano, P., Macaya, G. and Bernardi, G. (1981) The major components of the mouse and human genomes. 1. Preparation, basic properties and compositional heterogeneity *Eur J Biochem* 115 (2): 227-33

Dalvi, A. and Rodgers, R. J. (1996) GABAergic influences on plus-maze behaviour in mice *Psychopharmacology (Berl)* 128 (4): 380-97

Davies, B., Feo, S., Heard, E. and Fried, M. (1989) A strategy to detect and isolate an intron-containing gene in the presence of multiple processed pseudogenes *Proc Natl Acad Sci U S A* 86 (17): 6691-5

Dawson, T. M., Arriza, J. L., Jaworsky, D. E., Borisy, F. F., Attramadal, H., Lefkowitz, R. J. and Ronnett, G. V. (1993) Beta-adrenergic receptor kinase-2 and beta-arrestin-2 as mediators of odorant-induced desensitization *Science* 259 (5096): 825-9

Dayhoff, M. O. (1976) The origin and evolution of protein superfamilies *Fed Proc* 35 (10): 2132-8

Dear, T. N., Campbell, K. and Rabbitts, T. H. (1991) Molecular cloning of putative odorant-binding and odorant-metabolizing proteins *Biochemistry* 30 (43): 10376-82

Del Punta, K., Rothman, A., Rodriguez, I. and Mombaerts, P. (2000) Sequence diversity and genomic organization of vomeronasal receptor genes in the mouse *Genome Res* 10 (12): 1958-67

den Hollander, A. I., Ghiani, M., de Kok, Y. J., Wijnholds, J., Ballabio, A., Cremers, F. P. and Broccoli, V. (2002) Isolation of Crb1, a mouse homologue of Drosophila crumbs, and analysis of its expression pattern in eye and brain *Mech Dev* 110 (1-2): 203-7

Denzin, L. K. and Cresswell, P. (1995) HLA-DM induces CLIP dissociation from MHC class II alpha beta dimers and facilitates peptide loading *Cell* 82 (1): 155-65

Dizhoor, A. M., Ray, S., Kumar, S., Niemi, G., Spencer, M., Brolley, D., Walsh, K. A., Philipov, P. P., Hurley, J. B. and Stryer, L. (1991) Recoverin: a calcium sensitive activator of retinal rod guanylate cyclase *Science* 251 (4996): 915-8

Doherty, P. C. and Zinkernagel, R. M. (1975) Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex *Nature* 256 (5512): 50-2

Dong, X., Han, S., Zylka, M. J., Simon, M. I. and Anderson, D. J. (2001) A diverse family of GPCRs expressed in specific subsets of nociceptive sensory neurons *Cell* 106 (5): 619-32

Down, T. A. and Hubbard, T. J. (2002) Computational detection and location of transcription start sites in mammalian genomic DNA *Genome Res* 12 (3): 458-61

Dreyer, W. J. (1998) The area code hypothesis revisited: olfactory receptors and other related transmembrane receptors may function as the last digits in a cell surface code for assembling embryos *Proc Natl Acad Sci U S A* 95 (16): 9072-7

Driscoll, J., Brown, M. G., Finley, D. and Monaco, J. J. (1993) MHC-linked LMP gene products specifically alter peptidase activities of the proteasome *Nature* 365 (6443): 262-4

Drutel, G., Arrang, J. M., Diaz, J., Wisnewsky, C., Schwartz, K. and Schwartz, J. C. (1995) Cloning of OL1, a putative olfactory receptor and its expression in the developing rat heart *Receptors Channels* 3 (1): 33-40

Duboule, D. (1998) Vertebrate hox gene regulation: clustering and/or colinearity? *Curr Opin Genet Dev* 8 (5): 514-8

Duboule, D. and Dolle, P. (1989) The structural and functional organization of the murine HOX gene family resembles that of Drosophila homeotic genes *Embo J* 8 (5): 1497-505

Dulac, C. and Axel, R. (1995) A novel family of genes encoding putative pheromone receptors in mammals *Cell* 83 (2): 195-206

Ehlers, A., Beck, S., Forbes, S. A., Trowsdale, J., Volz, A., Younger, R. and Ziegler, A. (2000) MHC-linked olfactory receptor loci exhibit polymorphism and contribute to extended HLA/OR-haplotypes *Genome Res* 10 (12): 1968-78

Eyre-Walker, A. and Hurst, L. D. (2001) The evolution of isochores Nat Rev Genet 2 (7): 549-55

Fan, W., Cai, W., Parimoo, S., Schwarz, D. C., Lennon, G. G. and Weissman, S. M. (1996) Identification of seven new human MHC class I region genes around the HLA-F locus *Immunogenetics* 44 (2): 97-103

Fan, W., Liu, Y. C., Parimoo, S. and Weissman, S. M. (1995) Olfactory receptor-like genes are located in the human major histocompatibility complex *Genomics* 27 (1): 119-23

Feingold, E. A., Penny, L. A., Nienhuis, A. W. and Forget, B. G. (1999) An olfactory receptor gene is located in the extended human beta-globin gene cluster and is expressed in erythroid cells *Genomics* 61 (1): 15-23

Felsenstein, J. (1985) Confidence-Limits On Phylogenies - an Approach Using the Bootstrap. *Evolution* 39(4):783-791.

Felsenstein, J. (1989) PHYLIP -- Phylogeny Inference Package (Version 3.2). *Cladistics* 5: 164-166.

Firestein, S. (2001) How the olfactory system makes sense of scents *Nature* 413 (6852): 211-8

Fitch, D. H., Bailey, W. J., Tagle, D. A., Goodman, M., Sieu, L. and Slightom, J. L. (1991) Duplication of the gamma-globin gene mediated by L1 long interspersed repetitive elements in an early ancestor of simian primates *Proc Natl Acad Sci U S A* 88 (16): 7396-400

Fitch, W. M. 1971. Toward defining the course of evolution: minimum change for a specified tree topology. *Systematic Zoology* 20: 406-416.

Freitag, J., Krieger, J., Strotmann, J. and Breer, H. (1995) Two classes of olfactory receptors in Xenopus laevis *Neuron* 15 (6): 1383-92

Friedrich, R. W. and Korsching, S. I. (1997) Combinatorial and chemotopic odorant coding in the zebrafish olfactory bulb visualized by optical imaging *Neuron* 18 (5): 737-52

Galtier, N., Gouy, M. and Gautier, C. (1996) SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny *Comput Appl Biosci* 12 (6): 543-8

Gardiner-Garden, M., Ballesteros, M., Gordon, M. and Tam, P. P. (1998) Histone- and protamine-DNA association: conservation of different patterns within the beta-globin domain in human sperm *Mol Cell Biol* 18 (6): 3350-6

Gatewood, J. M., Cook, G. R., Balhorn, R., Bradbury, E. M. and Schmid, C. W. (1987) Sequence-specific packaging of DNA in human sperm chromatin *Science* 236 (4804): 962-4

Gaunt, S. J., Krumlauf, R. and Duboule, D. (1989) Mouse homeo-genes within a subfamily, Hox-1.4, -2.6 and -5.1, display similar anteroposterior domains of expression in the embryo, but show stage- and tissue-dependent differences in their regulation *Development* 107 (1): 131-41

Gautier-Courteille, C., Salanova, M. and Conti, M. (1998) The olfactory adenylyl cyclase III is expressed in rat germ cells during spermiogenesis *Endocrinology* 139 (5): 2588-99

Gehlsen, K. R., Dillner, L., Engvall, E. and Ruoslahti, E. (1988) The human laminin receptor is a member of the integrin family of cell adhesion receptors *Science* 241 (4870): 1228-9

Gilad, Y., Segre, D., Skorecki, K., Nachman, M. W., Lancet, D. and Sharon, D. (2000) Dichotomy of single-nucleotide polymorphism haplotypes in olfactory receptor genes and pseudogenes *Nat Genet* 26 (2): 221-4

Gimelbrant, A. A. and McClintock, T. S. (1997) A nuclear matrix attachment region is highly homologous to a conserved domain of olfactory receptors *J Mol Neurosci* 9 (1): 61-3

Giorgi, D., Friedman, C., Trask, B. J. and Rouquier, S. (2000) Characterization of nonfunctional V1R-like pheromone receptor sequences in human *Genome Res* 10 (12): 1979-85

Glusman, G., Bahar, A., Sharon, D., Pilpel, Y., White, J. and Lancet, D. (2000) The olfactory receptor gene superfamily: data mining, classification, and nomenclature *Mamm Genome* 11 (11): 1016-23

Glusman, G., Clifton, S., Roe, B. and Lancet, D. (1996) Sequence analysis in the olfactory receptor gene cluster on human chromosome 17: recombinatorial events affecting receptor diversity *Genomics* 37 (2): 147-60

Glusman, G., Sosinsky, A., Ben-Asher, E., Avidan, N., Sonkin, D., Bahar, A., Rosenthal, A., Clifton, S., Roe, B., Ferraz, C., Demaille, J. and Lancet, D. (2000) Sequence, structure, and evolution of a complete human olfactory receptor gene cluster *Genomics* 63 (2): 227-45

Glusman, G., Yanai, I., Rubin, I. and Lancet, D. (2001) The complete human olfactory subgenome *Genome Res* 11 (5): 685-702

Gobin, S. J. and van den Elsen, P. J. (2000) Transcriptional regulation of the MHC class Ib genes HLA-E, HLA-F, and HLA-G *Hum Immunol* 61 (11): 1102-7

Goei, V. L., Choi, J., Ahn, J., Bowlus, C. L., Raha-Chowdhury, R. and Gruen, J. R. (1998) Human gamma-aminobutyric acid B receptor gene: complementary DNA cloning, expression, chromosomal location, and genomic organization *Biol Psychiatry* 44 (8): 659-66

Gonnet, G. H., Cohen, M. A. and Benner, S. A. (1992) Exhaustive matching of the entire protein sequence database *Science* 256 (5062): 1443-5

Gordon, D., Abajian, C. and Green, P. (1998) Consed: a graphical tool for sequence finishing *Genome Res* 8 (3): 195-202

Graham, A., Papalopulu, N. and Krumlauf, R. (1989) The murine and Drosophila homeobox gene complexes have common features of organization and expression *Cell* 57 (3): 367-78

Graham, F. L., Smiley, J., Russell, W. C. and Nairn, R. (1977) Characteristics of a human cell line transformed by DNA from human adenovirus type 5 *J Gen Virol* 36 (1): 59-74

Gregory, S. G., Howell, G. H. and Bentley, D. R. (1997) Genome Mapping by Fluorescent Fingerprinting *Genome Research* 7 1162-1168

Grifa, A., Totaro, A., Rommens, J. M., Carella, M., Roetto, A., Borgato, L., Zelante, L. and Gasparini, P. (1998) GABA (gamma-amino-butyric acid) neurotransmission: identification and fine mapping of the human GABAB receptor gene *Biochem Biophys Res Commun* 250 (2): 240-5

Gross-Isseroff, R., Ophir, D., Bartana, A., Voet, H. and Lancet, D. (1992) Evidence for genetic determination in human twins of olfactory thresholds for a standard odorant *Neurosci Lett* 141 (1): 115-8

Grosveld, F. (1999) Activation by locus control regions? *Curr Opin Genet Dev* 9 (2): 152-7

Grosveld, F., van Assendelft, G. B., Greaves, D. R. and Kollias, G. (1987) Position-independent, high-level expression of the human beta-globin gene in transgenic mice *Cell* 51 (6): 975-85

Gruen, J. R., Goei, V. L., Summers, K. M., Capossela, A., Powell, L., Halliday, J., Zoghbi, H., Shukla, H. and Weissman, S. M. (1992) Physical and genetic mapping of the telomeric major histocompatibility complex region in man and relevance to the primary hemochromatosis gene (HFE) *Genomics* 14 (2): 232-40

Guigo, R., Agarwal, P., Abril, J. F., Burset, M. and Fickett, J. W. (2000) An assessment of gene prediction accuracy in large DNA sequences *Genome Res* 10 (10): 1631-42

Gunther, E. and Walter, L. (2001) *The major histocompatibility complex of the rat (Rattus norvegicus)* Immunogenetics 7 (53): 520-42

Haino, M., Hayashida, H., Miyata, T., Shin, E. K., Matsuda, F., Nagaoka, H., Matsumura, R., Taka-ishi, S., Fukita, Y., Fujikura, J. and et al. (1994) Comparison and evolution of human immunoglobulin VH segments located in the 3' 0.8-megabase region. Evidence for unidirectional transfer of segmental gene sequences *J Biol Chem* 269 (4): 2619-26

Hall, L., Williams, K., Perry, A. C., Frayne, J. and Jury, J. A. (1998) The majority of human glutathione peroxidase type 5 (GPX5) transcripts are incorrectly spliced: implications for the role of GPX5 in the male reproductive tract *Biochem J* 333 (Pt 1) 5-9

Hallberg, E., Wozniak, R. W. and Blobel, G. (1993) An integral membrane protein of the pore membrane domain of the nuclear envelope contains a nucleoporin-like region *J Cell Biol* 122 (3): 513-21

Hanke, J., Brett, D., Zastrow, I., Aydin, A., Delbruck, S., Lehmann, G., Luft, F., Reich, J. and Bork, P. (1999) Alternative splicing of human genes: more the rule than the exception? *Trends Genet* 15 (10): 389-90

Hardison, R. C., Oeltjen, J. and Miller, W. (1997) Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome *Genome Res* 7 (10): 959-66

Hellmann-Blumberg, U., Hintz, M. F., Gatewood, J. M. and Schmid, C. W. (1993) Developmental differences in methylation of human Alu repeats *Mol Cell Biol* 13 (8): 4523-30

Henikoff, S. and Henikoff, J. G. (1992) Amino acid substitution matrices from protein blocks *Proc Natl Acad Sci U S A* 89 (22): 10915-9

Henry, J., Mather, I. H., McDermott, M. F. and Pontarotti, P. (1998) B30.2-like domain proteins: update and new insights into a rapidly expanding family of proteins *Mol Biol Evol* 15 (12): 1696-705

Herrada, G. and Dulac, C. (1997) A novel family of putative pheromone receptors in mammals with a topographically organized and sexually dimorphic distribution *Cell* 90 (4): 763-73

Hildebrand, J. G. and Shepherd, G. M. (1997) Mechanisms of olfactory discrimination: converging evidence for common principles across phyla *Annu Rev Neurosci* 20 595-631

Hillier, L. D., Lennon, G., Becker, M., Bonaldo, M. F., Chiapelli, B., Chissoe, S., Dietrich, N., DuBuque, T., Favello, A., Gish, W., Hawkins, M., Hultman, M., Kucaba, T., Lacy, M., Le, M., Le, N., Mardis, E., Moore, B., Morris, M., Parsons, J., Prange, C., Rifkin, L., Rohlfing, T., Schellenberg, K., Marra, M. and et al. (1996) Generation and analysis of 280,000 human expressed sequence tags *Genome Res* 6 (9): 807-28

Horikawa, I., Tanaka, H., Yuasa, Y., Suzuki, M. and Oshimura, M. (1995) Molecular cloning of a novel human cDNA on chromosome 1q21 and its mouse homolog encoding a nuclear protein with DNA-binding ability *Biochem Biophys Res Commun* 208 (3): 999-1007

Horton, R., Niblett, D., Milne, S., Palmer, S., Tubby, B., Trowsdale, J. and Beck, S. (1998) Large-scale sequence comparisons reveal unusually high levels of variation in the HLA-DQB1 locus in the class II region of the human MHC *J Mol Biol* 282 (1): 71-97

Howe, K., Bateman, A. and Durbin, R. (unpublished) QuickTree: A program for building huge Neighbor-Joining trees of protein sequences

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., Mongin, E., Pettett, R., Pocock, M., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I. and Clamp, M. (2002) The Ensembl genome database project *Nucleic Acids Res* 30 (1): 38-41

Huebner, K., Cannizzaro, L. A., Nakamura, T., Hillova, J., Mariage-Samson, R., Hecht, F., Hill, M. and Croce, C. M. (1988) A rearranged transforming gene, tre, is made up of human sequences derived from chromosome regions 5q, 17q and 18q *Oncogene* 3 (4): 449-55

Hughes, A. L. and Nei, M. (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection *Nature* 335 (6186): 167-70

Ikematsu, N., Yoshida, Y., Kawamura-Tsuzuku, J., Ohsugi, M., Onda, M., Hirai, M., Fujimoto, J. and Yamamoto, T. (1999) Tob2, a novel anti-proliferative Tob/BTG1 family member, associates with a component of the CCR4 transcriptional regulatory complex capable of binding cyclin-dependent kinases *Oncogene* 18 (52): 7432-41

Isomura, T., Tamiya-Koizumi, K., Suzuki, M., Yoshida, S., Taniguchi, M., Matsuyama, M., Ishigaki, T., Sakuma, S. and Takahashi, M. (1992) RFP is a DNA binding protein associated with the nuclear matrix *Nucleic Acids Res* 20 (20): 5305-10

Isono, K., McIninch, J. D. and Borodovsky, M. (1994) Characteristic features of the nucleotide sequences of yeast mitochondrial ribosomal protein genes as analyzed by computer program GeneMark *DNA Res* 1 (6): 263-9

Issel-Tarver, L. and Rine, J. (1996) Organization and expression of canine olfactory receptor genes *Proc Natl Acad Sci U S A* 93 (20): 10897-902

Issel-Tarver, L. and Rine, J. (1997) The evolution of mammalian olfactory receptor genes *Genetics* 145 (1): 185-95

Iwata, Y., Nakayama, A., Murakami, H., Iida, K., Iwashita, T., Asai, N. and Takahashi, M. (1999) Characterization of the promoter region of the human RFP gene *Biochem Biophys Res Commun* 261 (2): 381-4

Jabbari, K. and Bernardi, G. (1998) CpG doublets, CpG islands and Alu repeats in long human DNA sequences from different isochore families *Gene* 224 (1-2): 123-7

Jack, L. J. and Mather, I. H. (1990) Cloning and analysis of cDNA encoding bovine butyrophilin, an apical glycoprotein expressed in mammary tissue and secreted in association with the milk-fat globule membrane during lactation *J Biol Chem* 265 (24): 14481-6

Jareborg, N., Birney, E. and Durbin, R. (1999) Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs *Genome Res* 9 (9): 815-24

Johns, M. A., Feder, H. H., Komisaruk, B. R. and Mayer, A. D. (1978) Urine-induced reflex ovulation in anovulatory rats may be a vomeronasal effect *Nature* 272 (5652): 446-8

Jones, D. T., Taylor, W. R. and Thornton, J. M. (1992) The rapid generation of mutation data matrices from protein sequences *Comput Appl Biosci* 8 (3): 275-82

Jones, E. P., Kumanovics, A., Yoshino, M. and Fischer Lindahl, K. (1999) Mhc class I and non-class I gene organization in the proximal H2-M region of the mouse *Immunogenetics* 49 (3): 183-95

Jones, E. P., Xiao, H., Schultz, R. A., Flaherty, L., Trachtulec, Z., Vincek, V., Larin, Z., Lehrach, H. and Lindahl, K. F. (1995) MHC class I gene organization in > 1.5-Mb YAC contigs from the H2-M region *Genomics* 27 (1): 40-51

Jurka, J. (2000) Repbase update: a database and an electronic journal of repetitive elements *Trends Genet* 16 (9): 418-20

Jurka, J. and Smith, T. (1988) A fundamental division in the Alu family of repeated sequences *Proc Natl Acad Sci U S A* 85 (13): 4775-8

Jurka, J., Kapitonov, V. V., Klonowski, P., Walichiewicz, J. and Smit, A. F. (1996) Identification of new medium reiteration frequency repeats in the genomes of Primates, Rodentia and Lagomorpha *Genetica* 98 (3): 235-47

Kagie, M. J., Kleijer, W. J., Huijmans, J. G., Maaswinkel-Mooy, P. and Kanhai, H. H. (1992) beta-Glucuronidase deficiency as a cause of fetal hydrops *Am J Med Genet* 42 (5): 693-5

Kamitani, T., Kito, K., Nguyen, H. P., Fukuda-Kamitani, T. and Yeh, E. T. (1998) Characterization of a second member of the sentrin family of ubiquitin-like proteins *J Biol Chem* 273 (18): 11349-53

Kapitonov, V. and Jurka, J. (1996) The age of Alu subfamilies *J Mol Evol* 42 (1): 59-65

Kapitonov, V. V., Holmquist, G. P. and Jurka, J. (1998) L1 repeat is a basic unit of heterochromatin satellites in cetaceans *Mol Biol Evol* 15 (5): 611-2

Kasture, S. B., Mandhane, S. N. and Chopde, C. T. (1996) Baclofen-induced catatonia: modification by serotonergic agents *Neuropharmacology* 35 (5): 595-8

Kenmochi, N., Kawaguchi, T., Rozen, S., Davis, E., Goodman, N., Hudson, T. J., Tanaka, T. and Page, D. C. (1998) A map of 75 human ribosomal protein genes *Genome Res* 8 (5): 509-23

Kent, W. J. and Haussler, D. (2001) Assembly of the working draft of the human genome with GigAssembler *Genome Res* 11 (9): 1541-8

Kerr, D. I. and Ong, J. (1995) GABAB receptors *Pharmacol Ther* 67 (2): 187-246

Keverne, E. B. (1999) The vomeronasal organ *Science* 286 (5440): 716-20

Kleene, S. J. and Gesteland, R. C. (1991) Calcium-activated chloride conductance in frog olfactory cilia *J Neurosci* 11 (11): 3624-9

Klingenhoff, A., Frech, K., Quandt, K. and Werner, T. (1999) Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity *Bioinformatics* 15 (3): 180-6

Knowles, B. B., Solter, D., Trinchieri, G., Maloney, K. M., Ford, S. R. and Aden, D. P. (1977) Complement-mediated antiserum cytotoxic reactions to human chromosome 7 coded antigen(s): immunoselection of rearranged human chromosome 7 in human-mouse somatic cell hybrids *J Exp Med* 145 (2): 314-26

Kobilka, B. (1992) Adrenergic receptors as models for G protein-coupled receptors *Annu Rev Neurosci* 15 87-114

Kobilka, B. K., Kobilka, T. S., Daniel, K., Regan, J. W., Caron, M. G. and Lefkowitz, R. J. (1988) Chimeric alpha 2-,beta 2-adrenergic receptors: delineation of domains involved in effector coupling and ligand binding specificity *Science* 240 (4857): 1310-6

Kochanek, S., Renz, D. and Doerfler, W. (1993) DNA methylation in the Alu sequences of diploid and haploid primary human cells *Embo J* 12 (3): 1141-51

Koop, B. F. and Hood, L. (1994) Striking sequence similarity over almost 100 kilobases of human and mouse T-cell receptor DNA *Nat Genet* 7 (1): 48-53

Koshimoto, H., Katoh, K., Yoshihara, Y., Nemoto, Y. and Mori, K. (1994) Immunohistochemical demonstration of embryonic expression of an odor receptor protein and its zonal distribution in the rat olfactory epithelium *Neurosci Lett* 169 (1-2): 73-6

Kratz, E., Dugas, J. C. and Ngai, J. (2002) Odorant receptor gene regulation: implications from genomic organization *Trends Genet* 18 (1): 29-34

Krautwurst, D., Yau, K. W. and Reed, R. R. (1998) Identification of ligands for olfactory receptors by functional expression of a receptor library *Cell* 95 (7): 917-26

Ku, N. O., Wright, T. L., Terrault, N. A., Gish, R. and Omary, M. B. (1997) Mutation of human keratin 18 in association with cryptogenic cirrhosis *J Clin Invest* 99 (1): 19-23

Kubick, S., Strotmann, J., Andreini, I. and Breer, H. (1997) Subfamily of olfactory receptors characterized by unique structural features and expression patterns *J Neurochem* 69 (2): 465-75

Kulski, J. K. and Dawkins, R. L. (1999) The P5 multicopy gene family in the MHC is related in sequence to human endogenous retroviruses HERV-L and HERV-16 *Immunogenetics* 49 (5): 404-12

Kurahashi, T. and Menini, A. (1997) Mechanism of odorant adaptation in the olfactory receptor cell *Nature* 385 (6618): 725-9

Ladenburger, E. M., Fackelmayer, F. O., Hameister, H. and Knippers, R. (1997) MCM4 and PRKDC, human genes encoding proteins MCM4 and DNA-PKcs, are close neighbours located on chromosome 8q12-->q13 *Cytogenet Cell Genet* 77 (3-4): 268-70

Laity, J. H., Lee, B. M. and Wright, P. E. (2001) Zinc finger proteins: new insights into structural and functional diversity *Curr Opin Struct Biol* 11 (1): 39-46

Lancet, D., Ben-Arie, N., Cohen, S., Gat, U., Gross-Isseroff, R., Horn-Saban, S., Khen, M., Lehrach, H., Natochin, M., North, M. and et al. (1993) Olfactory receptors: transduction, diversity, human psychophysics and genome analysis *Ciba Found Symp* 179 131-41; discussion 141-6

Landsman, D., Soares, N., Gonzalez, F. J. and Bustin, M. (1986) Chromosomal protein HMG-17. Complete human cDNA sequence and evidence for a multigene family *J Biol Chem* 261 (16): 7479-84

Lane, R. P., Cutforth, T., Young, J., Athanasiou, M., Friedman, C., Rowen, L., Evans, G., Axel, R., Hood, L. and Trask, B. J. (2001) Genomic analysis of orthologous mouse and human olfactory receptor loci *Proc Natl Acad Sci U S A* 98 (13): 7390-5

Lane, R. P., Roach, J. C., Lee, I. Y., Boysen, C., Smit, A., Trask, B. J. and Hood, L. (2002) Genomic Analysis of the Olfactory Receptor Region of the Mouse and Human T-Cell Receptor alpha/delta Loci *Genome Res* 12 (1): 81-7

Lapenta, V., Chiurazzi, P., van der Spek, P., Pizzuti, A., Hanaoka, F. and Brahe, C. (1997) SMT3A, a human homologue of the S. cerevisiae SMT3 gene, maps to chromosome 21qter and defines a novel gene family *Genomics* 40 (2): 362-6

Lapidot, M., Pilpel, Y., Gilad, Y., Falcovitz, A., Sharon, D., Haaf, T. and Lancet, D. (2001) Mouse-human orthology relationships in an olfactory receptor gene cluster *Genomics* 3 (71): 296-306

Latif, F., Tory, K., Gnarra, J., Yao, M., Duh, F. M., Orcutt, M. L., Stackhouse, T., Kuzmin, I., Modi, W., Geil, L. and et al. (1993) Identification of the von Hippel-Lindau disease tumor suppressor gene *Science* 260 (5112): 1317-20

Le Gouill, C., Parent, J. L., Rola-Pleszczynski, M. and Stankova, J. (1997) Role of the Cys90, Cys95 and Cys173 residues in the structure and function of the human platelet-activating factor receptor *FEBS Lett* 402 (2-3): 203-8

Leibovici, M., Lapointe, F., Aletta, P. and Ayer-Le Lievre, C. (1996) Avian olfactory receptors: differentiation of olfactory neurons under normal and experimental conditions *Dev Biol* 175 (1): 118-31

Levy, N. S., Bakalyar, H. A. and Reed, R. R. (1991) Signal transduction in olfactory neurons *J Steroid Biochem Mol Biol* 39 (4B): 633-7

Lewis, E. B. (1978) A gene complex controlling segmentation in Drosophila *Nature* 276 (5688): 565-70

Lieber, M. (1996) Immunoglobulin diversity: rearranging by cutting and repairing *Curr Biol* 6 (2): 134-6

Lin, H. and Grosschedl, R. (1995) Failure of B-cell differentiation in mice lacking the transcription factor EBF *Nature* 376 (6537): 263-7

Linardopoulou, E., Mefford, H. C., Nguyen, O., Friedman, C., van den Engh, G., Farwell, D. G., Coltrera, M. and Trask, B. J. (2001) Transcriptional activity of multiple copies of a subtelomerically located olfactory receptor gene that is polymorphic in number and location *Hum Mol Genet* 10 (21): 2373-83

Lindahl, K. F., Byers, D. E., Dabhi, V. M., Hovik, R., Jones, E. P., Smith, G. P., Wang, C. R., Xiao, H. and Yoshino, M. (1997) H2-M3, a full-service class Ib histocompatibility antigen *Annu Rev Immunol* 15 851-79

Ling, K., Wang, P., Zhao, J., Wu, Y. L., Cheng, Z. J., Wu, G. X., Hu, W., Ma, L. and Pei, G. (1999) Five-transmembrane domains appear sufficient for a G protein-coupled receptor: functional five-transmembrane domain chemokine receptors *Proc Natl Acad Sci U S A* 96 (14): 7922-7

Liu, W. M., Chu, W. M., Choudary, P. V. and Schmid, C. W. (1995) Cell stress and translational inhibitors transiently increase the abundance of mammalian SINE transcripts *Nucleic Acids Res* 23 (10): 1758-65

Liu, Y. C., Pan, J., Zhang, C., Fan, W., Collinge, M., Bender, J. R. and Weissman, S. M. (1999) A MHC-encoded ubiquitin-like protein (FAT10) binds noncovalently to the spindle assembly checkpoint protein MAD2 *Proc Natl Acad Sci U S A* 96 (8): 4313-8

Lomas, D. E. and Keverne, E. B. (1982) Role of the vomeronasal organ and prolactin in the acceleration of puberty in female mice *J Reprod Fertil* 66 (1): 101-7

Loveland, B., Wang, C. R., Yonekawa, H., Hermel, E. and Lindahl, K. F. (1990) Maternally transmitted histocompatibility antigen of mice: a hydrophobic peptide of a mitochondrially encoded protein *Cell* 60 (6): 971-80

Lyon, M. F. (1998) X-chromosome inactivation: a repeat hypothesis *Cytogenet Cell Genet* 80 (1-4): 133-7

Lyon, M. F. (2000) LINE-1 elements and X chromosome inactivation: a function for "junk" DNA? *Proc Natl Acad Sci U S A* 97 (12): 6248-9

Macaya, G., Thiery, J. P. and Bernardi, G. (1976) An approach to the organization of eukaryotic genomes at a macromolecular level *J Mol Biol* 108 (1): 237-54

Magram, J., Chada, K. and Costantini, F. (1985) Developmental regulation of a cloned adult beta-globin gene in transgenic mice *Nature* 315 (6017): 338-40

Malfroy, L., Roth, M. P., Carrington, M., Borot, N., Volz, A., Ziegler, A. and Coppin, H. (1997) Heterogeneity in rates of recombination in the 6-Mb region telomeric to the human major histocompatibility complex *Genomics* 43 (2): 226-31

Malnic, B., Hirono, J., Sato, T. and Buck, L. B. (1999) Combinatorial receptor codes for odors *Cell* 96 (5): 713-23

Mardis, E. R. (1994) High-throughput detergent extraction of M13 subclones for fluorescent DNA sequencing *Nucleic Acids Res* 22 (11): 2173-5

Matarazzo, V., Tirard, A., Renucci, M., Belaich, A. and Clement, J. L. (1998) Isolation of putative olfactory receptor sequences from pig nasal epithelium *Neurosci Lett* 249 (2-3): 87-90

Matsuda, S., Kawamura-Tsuzuku, J., Ohsugi, M., Yoshida, M., Emi, M., Nakamura, Y., Onda, M., Yoshida, Y., Nishiyama, A. and Yamamoto, T. (1996) Tob, a novel protein that interacts with p185erbB2, is associated with anti-proliferative activity *Oncogene* 12 (4): 705-13

Matsunami, H. and Buck, L. B. (1997) A multigene family encoding a diverse array of putative pheromone receptors in mammals *Cell* 90 (4): 775-84

Maxam, A. M. and Gilbert, W. (1977) A new method for sequencing DNA *Proc Natl Acad Sci U S A* 74 (2): 560-4

Maxson R., Cohn R., Kedes L. and Mohun T. (1983) Expression and organization of histone genes. *Annu Rev Genet.* 17:239-77.

McClintock, M. K. (1971) Menstrual synchorony and suppression *Nature* 229 (5282): 244-5

McClintock, T. S., Landers, T. M., Gimelbrant, A. A., Fuller, L. Z., Jackson, B. A., Jayawickreme, C. K. and Lerner, M. R. (1997) Functional expression of olfactory-adrenergic receptor chimeras and intracellular retention of heterologously expressed olfactory receptors *Brain Res Mol Brain Res* 48 (2): 270-8

McPherson, J. D., Marra, M., Hillier, L., Waterston, R. H., Chinwalla, A., Wallis, J., Sekhon, M., Wylie, K., Mardis, E. R., Wilson, R. K., Fulton, R., Kucaba, T. A., Wagner-McPherson, C., Barbazuk, W. B., Gregory, S. G., Humphray, S. J., French, L., Evans, R. S., Bethel, G., Whittaker, A., Holden, J. L., McCann, O. T., Dunham, A., Soderlund, C., Scott, C. E., Bentley, D. R., Schuler, G., Chen, H. C., Jang, W., Green, E. D., Idol, J. R., Maduro, V. V., Montgomery, K. T., Lee, E., Miller, A., Emerling, S., Kucherlapati, Gibbs, R., Scherer, S., Gorrell, J. H., Sodergren, E., Clerc-Blankenburg, K., Tabor, P., Naylor, S., Garcia, D., de Jong, P. J., Catanese, J. J., Nowak, N., Osoegawa, K., Qin, S., Rowen, L., Madan, A., Dors, M., Hood, L., Trask, B., Friedman, C., Massa, H., Cheung, V. G., Kirsch, I. R., Reid, T., Yonescu, R., Weissenbach, J., Bruls, T., Heilig, R., Branscomb, E., Olsen, A., Doggett, N., Cheng, J. F., Hawkins, T., Myers, R. M., Shang, J., Ramirez, L., Schmutz, J., Velasquez, O., Dixon, K., Stone, N. E., Cox, D. R., Haussler, D., Kent, W. J., Furey, T., Rogic, S., Kennedy, S., Jones, S., Rosenthal, A., Wen, G., Schilhabel, M., Gloeckner, G., Nyakatura, G., Siebert, R., Schlegelberger, B., Korenberg, J., Chen, X. N., Fujiyama, A., Hattori, M., Toyoda, A., Yada, T., Park, H. S., Sakaki, Y., Shimizu, N., Asakawa, S., Kawasaki, K., Sasaki, T., Shintani, A., Shimizu, A., Shibuya, K., Kudoh, J., Minoshima, S., Ramser, J., Seranski, P., Hoff, C., Poustka, A., Reinhardt, R. and Lehrach, H. (2001) A physical map of the human genome *Nature* 409 (6822): 934-41

Menco, B. P. and Jackson, J. E. (1997) A banded topography in the developing rat's olfactory epithelial surface *J Comp Neurol* 388 (2): 293-306

Meng, E. C. and Bourne, H. R. (2001) Receptor activation: what does the rhodopsin structure tell us? *Trends Pharmacol Sci* 22 (11): 587-93

Meredith, M. (1986) Vomeronasal organ removal before sexual experience impairs male hamster mating behavior *Physiol Behav* 36 (4): 737-43

Meunier-Rotival, M., Soriano, P., Cuny, G., Strauss, F. and Bernardi, G. (1982) Sequence organization and genomic distribution of the major family of interspersed repeats of mouse DNA *Proc Natl Acad Sci U S A* 79 (2): 355-9

Meyerhans, A., Vartanian, J. P. and Wain-Hobson, S. (1990) DNA recombination during PCR *Nucleic Acids Res* 18 (7): 1687-91

Mironov, A. A., Fickett, J. W. and Gelfand, M. S. (1999) Frequent alternative splicing of human genes *Genome Res* 9 (12): 1288-93

Moller, S., Croning, M. D. and Apweiler, R. (2001) Evaluation of methods for the prediction of membrane spanning regions *Bioinformatics* 17 (7): 646-53

Mombaerts, P. (1996) Targeting olfaction *Curr Opin Neurobiol* 6 (4): 481-6

Mombaerts, P. (1999) Molecular biology of odorant receptors in vertebrates *Annu Rev Neurosci* 22 487-509

Mombaerts, P. (1999) Odorant receptor genes in humans *Curr Opin Genet Dev* 9 (3): 315-20

Mombaerts, P., Wang, F., Dulac, C., Chao, S. K., Nemes, A., Mendelsohn, M., Edmondson, J. and Axel, R. (1996a) Visualizing an olfactory sensory map *Cell* 87 (4): 675-86

Mombaerts, P., Wang, F., Dulac, C., Vassar, R., Chao, S. K., Nemes, A., Mendelsohn, M., Edmondson, J. and Axel, R. (1996b) The molecular biology of olfactory perception *Cold Spring Harb Symp Quant Biol* 61 135-45

Mondadori, C., Jaekel, J. and Preiswerk, G. (1993) CGP 36742: the first orally active GABAB blocker improves the cognitive performance of mice, rats, and rhesus monkeys *Behav Neural Biol* 60 (1): 62-8

Monnot, C., Weber, V., Stinnakre, J., Bihoreau, C., Teutsch, B., Corvol, P. and Clauser, E. (1991) Cloning and functional characterization of a novel mas-related gene, modulating intracellular angiotensin II actions *Mol Endocrinol* 5 (10): 1477-87

Murrell, J. R. and Hunter, D. D. (1999) An olfactory sensory neuron line, odora, properly targets olfactory proteins and responds to odorants *J Neurosci* 19 (19): 8260-70

Musahl, C., Schulte, D., Burkhart, R. and Knippers, R. (1995) A human homologue of the yeast replication protein Cdc21. Interactions with other Mcm proteins *Eur J Biochem* 230 (3): 1096-101

Nagase, T., Ishikawa, K., Suyama, M., Kikuno, R., Miyajima, N., Tanaka, A., Kotani, H., Nomura, N. and Ohara, O. (1998) Prediction of the coding sequences of unidentified human genes. XI. The complete sequences of 100 new cDNA clones from brain which code for large proteins in vitro *DNA Res* 5 (5): 277-86

Nahm, D. H., Lee, Y. E., Yim, E. J., Park, H. S., Yim, H., Kang, Y. and Kim, J. K. (2002) Identification of cytokeratin 18 as a bronchial epithelial autoantigen associated with nonallergic asthma *Am J Respir Crit Care Med* 165 (11): 1536-9

Nakamura, T., Hillova, J., Mariage-Samson, R. and Hill, M. (1988) Molecular cloning of a novel oncogene generated by DNA recombination during transfection *Oncogene Res* 2 (4): 357-70

Navarro, M. and Gull, K. (2001) A pol I transcriptional body associated with VSG mono-allelic expression in Trypanosoma brucei *Nature* 414 (6865): 759-63

Nef, P., Hermans-Borgmeyer, I., Artieres-Pin, H., Beasley, L., Dionne, V. E. and Heinemann, S. F. (1992) Spatial pattern of receptor expression in the olfactory epithelium *Proc Natl Acad Sci U S A* 89 (19): 8948-52

Nef, S. and Nef, P. (1997) Olfaction: transient expression of a putative odorant receptor in the avian notochord *Proc Natl Acad Sci U S A* 94 (9): 4766-71

Nei, M. and Kumar, S. *Molecular Evolution and Phylogenetics* (Oxford University Press, New York, 2000).

Neitzel, H., Kalscheuer, V., Henschel, S., Digweed, M. and Sperling, K. (1998) Beta-heterochromatin in mammals: evidence from studies in Microtus agrestis based on the extensive accumulation of L1 and non-L1 retroposons in the heterochromatin *Cytogenet Cell Genet* 80 (1-4): 165-72

Ngai, J., Chess, A., Dowling, M. M., Necles, N., Macagno, E. R. and Axel, R. (1993a) Coding of olfactory information: topography of odorant receptor expression in the catfish olfactory epithelium *Cell* 72 (5): 667-80

Ngai, J., Dowling, M. M., Buck, L., Axel, R. and Chess, A. (1993b) The family of genes encoding odorant receptors in the channel catfish *Cell* 72 (5): 657-66

Nomura, N., Miyajima, N., Sazuka, T., Tanaka, A., Kawarabayasi, Y., Sato, S., Nagase, T., Seki, N., Ishikawa, K. and Tabata, S. (1994) Prediction of the coding sequences of unidentified human genes. I. The coding sequences of 40 new genes (KIAA0001-KIAA0040) deduced by analysis of randomly sampled cDNA clones from human immature myeloid cell line KG-1 (supplement) *DNA Res* 1 (1): 47-56

O'Hara, P. J., Sheppard, P. O., Thogersen, H., Venezia, D., Haldeman, B. A., McGrane, V., Houamed, K. M., Thomsen, C., Gilbert, T. L. and Mulvihill, E. R. (1993) The ligand-binding domain in metabotropic glutamate receptors is related to bacterial periplasmic binding proteins *Neuron* 11 (1): 41-52

Ober, C., Elias, S., Kostyu, D. D. and Hauck, W. W. (1992) Decreased fecundability in Hutterite couples sharing HLA-DR *Am J Hum Genet* 50 (1): 6-14

Okazaki, S., Tanase, S., Choudhury, B. K., Setoyama, K., Miura, R., Ogawa, M. and Setoyama, C. (1994) A novel nuclear protein with zinc fingers down-regulated during early mammalian cell differentiation *J Biol Chem* 269 (9): 6900-7

Olavesen, M. G., Bentley, E., Mason, R. V., Stephens, R. J. and Ragoussis, J. (1997) Fine mapping of 39 ESTs on human chromosome 6p23-p25 *Genomics* 46 (2): 303-6

O'Leary, D. D., Yates, P. A. and McLaughlin, T. (1999) Molecular development of sensory maps: representing sights and smells in the brain *Cell* 96 (2): 255-69

Ortmann, B., Androlewicz, M. J. and Cresswell, P. (1994) MHC class I/beta 2-microglobulin complexes associate with TAP transporters before peptide binding *Nature* 368 (6474): 864-7

Palczewski, K., Kumasaka, T., Hori, T., Behnke, C. A., Motoshima, H., Fox, B. A., Le Trong, I., Teller, D. C., Okada, T., Stenkamp, R. E., Yamamoto, M. and Miyano, M. (2000) Crystal structure of rhodopsin: A G protein-coupled receptor *Science* 289 (5480): 739-45

Panning, B. and Smiley, J. R. (1993) Activation of RNA polymerase III transcription of human Alu repetitive elements by adenovirus type 5: requirement for the E1b 58-kilodalton protein and the products of E4 open reading frames 3 and 6 *Mol Cell Biol* 13 (6): 3231-44

Parmentier, M., Libert, F., Schurmans, S., Schiffmann, S., Lefort, A., Eggerickx, D., Ledent, C., Mollereau, C., Gerard, C., Perret, J. and et al. (1992) Expression of members of the putative olfactory receptor gene family in mammalian germ cells *Nature* 355 (6359): 453-5

Pennesi, G., Brioli, G., Lulli, P., Mariani, B., Morellini, M., Nicotra, M. and Trabace, S. (1998) HLA and complement factors alleles sharing in Italian couples with recurrent spontaneous abortions *Hum Immunol* 59 (6): 382-6

Peters, H. C., Kammer, G., Volz, A., Kaupmann, K., Ziegler, A., Bettler, B., Epplen, J. T., Sander, T. and Riess, O. (1998) Mapping, genomic structure, and polymorphisms of the human GABABR1 receptor gene: evaluation of its involvement in idiopathic generalized epilepsy *Neurogenetics* 2 (1): 47-54

Petruzzella, V., Tiranti, V., Fernandez, P., Ianna, P., Carrozzo, R. and Zeviani, M. (1998) Identification and characterization of human cDNAs specific to BCS1, PET112, SCO1, COX15, and COX11, five genes involved in the formation and function of the mitochondrial respiratory chain *Genomics* 54 (3): 494-504

Pham-Dinh, D., Della Gaspera, B., Kerlero de Rosbo, N. and Dautigny, A. (1995) Structure of the human myelin/oligodendrocyte glycoprotein gene and multiple alternative spliced isoforms *Genomics* 29 (2): 345-52

Pham-Dinh, D., Mattei, M. G., Nussbaum, J. L., Roussel, G., Pontarotti, P., Roeckel, N., Mather, I. H., Artzt, K., Lindahl, K. F. and Dautigny, A. (1993) Myelin/oligodendrocyte glycoprotein is a member of a subset of the immunoglobulin superfamily encoded within the major histocompatibility complex *Proc Natl Acad Sci U S A* 90 (17): 7990-4

Pilpel, Y. and Lancet, D. (1999) The variable and conserved interfaces of modeled olfactory receptor proteins *Protein Sci* 8 (5): 969-77

Porteros, A., Brinon, J. G., Crespo, C., Okazaki, K., Hidaka, H., Aijon, J. and Alonso, J. R. (1996) Neurocalcin immunoreactivity in the rat accessory olfactory bulb *Brain Res* 729 (1): 82-9

Potts, W. K. and Wakeland, E. K. (1990) The maintenance of MHC polymorphism *Immunol Today* 11 (2): 39-40

Potts, W. K. and Wakeland, E. K. (1993) Evolution of MHC genetic diversity: a tale of incest, pestilence and sexual preference *Trends Genet* 9 (12): 408-12

Prober, J. M., Trainor, G. L., Dam, R. J., Hobbs, F. W., Robertson, C. W., Zagursky, R. J., Cocuzza, A. J., Jensen, M. A. and Baumeister, K. (1987) A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides *Science* 238 (4825): 336-41

Qasba, P. and Reed, R. R. (1998) Tissue and zonal-specific expression of an olfactory receptor transgene *J Neurosci* 18 (1): 227-36

Quandt, K., Frech, K., Karas, H., Wingender, E. and Werner, T. (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data *Nucleic Acids Res* 23 (23): 4878-84

Raasi, S., Schmidtke, G. and Groettrup, M. (2001) The ubiquitin-like protein FAT10 forms covalent conjugates and induces apoptosis *J Biol Chem* 276 (38): 35334-43

Raming, K., Konzelmann, S. and Breer, H. (1998) Identification of a novel G-protein coupled receptor expressed in distinct brain regions and a defined olfactory zone *Receptors Channels* 6 (2): 141-51

Raming, K., Krieger, J., Strotmann, J., Boekhoff, I., Kubick, S., Baumstark, C. and Breer, H. (1993) Cloning and expression of odorant receptors *Nature* 361 (6410): 353-6

Rana, B. K., Shiina, T. and Insel, P. A. (2001) Genetic variations and polymorphisms of G protein-coupled receptors: functional and therapeutic implications *Annu Rev Pharmacol Toxicol* 41 593-624

Ressler, K. J., Sullivan, S. L. and Buck, L. B. (1993) A zonal organization of odorant receptor gene expression in the olfactory epithelium *Cell* 73 (3): 597-609

Ressler, K. J., Sullivan, S. L. and Buck, L. B. (1994) Information coding in the olfactory system: evidence for a stereotyped and highly organized epitope map in the olfactory bulb *Cell* 79 (7): 1245-55

Reynolds, J. and Keverne, E. B. (1979) The accessory olfactory system and its role in the pheromonally mediated suppression of oestrus in grouped mice *J Reprod Fertil* 57 (1): 31-5

Rhodes, D. A., Stammers, M., Malcherek, G., Beck, S. and Trowsdale, J. (2001) The cluster of BTN genes in the extended major histocompatibility complex *Genomics* 71 (3): 351-62

Rice, P., Longden, I. and Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite *Trends Genet* 16 (6): 276-7

Richards, S., Bucan, M., Brorson, K., Kiefer, M. C., Hunt, S. W., 3rd, Lehrach, H. and Lindahl, K. F. (1989) Genetic and molecular mapping of the Hmt region of mouse *Embo J* 8 (12): 3749-57

Roetto, A., Sbaiz, L., Bosio, S., Piperno, A., Fargion, S., Carella, M., Totaro, A., Grifa, A., Gasparini, P. and Camaschella, C. (1997) A recombination event close to HFE gene in hereditary hemochromatosis *Ann Genet* 40 (3): 150-3

Rogers, J. and Bradley, A. (2001) The mouse genome sequence: status and prospects *Genomics* 77 (3): 117-8

Roth, M. P., Malfroy, L., Offer, C., Sevin, J., Enault, G., Borot, N., Pontarotti, P. and Coppin, H. (1995) The human myelin oligodendrocyte glycoprotein (MOG) gene: complete nucleotide sequence and structural characterization *Genomics* 28 (2): 241-50

Rouquier, S., Taviaux, S., Trask, B. J., Brand-Arpon, V., van den Engh, G., Demaille, J. and Giorgi, D. (1998) Distribution of olfactory receptor genes in the human genome *Nat Genet* 18 (3): 243-50

Rubin, C. M., VandeVoort, C. A., Teplitz, R. L. and Schmid, C. W. (1994) Alu repeated DNAs are differentially methylated in primate germ cells *Nucleic Acids Res* 22 (23): 5121-7

Ruddy, D. A., Kronmal, G. S., Lee, V. K., Mintier, G. A., Quintana, L., Domingo, R., Jr., Meyer, N. C., Irrinki, A., McClelland, E. E., Fullan, A., Mapa, F. A., Moore, T., Thomas, W., Loeb, D. B., Harmon, C., Tsuchihashi, Z., Wolff, R. K., Schatzman, R. C. and Feder, J. N. (1997) A 1.1-Mb transcript map of the hereditary hemochromatosis locus *Genome Res* 7 (5): 441-56

Ryba, N. J. and Tirindelli, R. (1997) A new multigene family of putative pheromone receptors *Neuron* 19 (2): 371-9

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J., Willey, D. L., Hunt, S. E., Cole, C. G., Coggill, P. C., Rice, C. M., Ning, Z., Rogers, J., Bentley, D. R., Kwok, P. Y., Mardis, E. R., Yeh, R. T., Schultz, B., Cook, L., Davenport, R., Dante, M., Fulton, L., Hillier, L., Waterston, R. H., McPherson, J. D., Gilman, B., Schaffner, S., Van Etten, W. J., Reich, D., Higgins, J., Daly, M. J., Blumenstiel, B., Baldwin, J., Stange-Thomann, N., Zody, M. C., Linton, L., Lander, E. S. and Altshuler, D. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms *Nature* 409 (6822): 928-33

Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees *Mol Biol Evol* 4 (4): 406-25

Sakmar, T. P. (2002) Structure of rhodopsin and the superfamily of seven-helical receptors: the same and not the same *Curr Opin Cell Biol* 14 (2): 189-95

Sanger, F., Nicklen, S. and Coulson, A. R. (1977) DNA sequencing with chain-terminating inhibitors *Proc Natl Acad Sci U S A* 74 (12): 5463-7

Satoh, T., Tsuruga, H., Yabuta, N., Ishidate, M., Jr. and Nojima, H. (1997) Assignment of the human CDC21 (MCM4) gene to chromosome 8q11.2 *Genomics* 46 (3): 525-6

Scherf, M., Klingenhoff, A. and Werner, T. (2000) Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach *J Mol Biol* 297 (3): 599-606

Schleicher, S., Boekhoff, I., Arriza, J., Lefkowitz, R. J. and Breer, H. (1993) A beta-adrenergic receptor kinase-like enzyme is involved in olfactory signal termination *Proc Natl Acad Sci U S A* 90 (4): 1420-4

Schmid, C. W. (1998) Does SINE evolution preclude Alu function? *Nucleic Acids Res* 26 (20): 4541-50

Scholler, N., Hayden-Ledbetter, M., Hellstrom, K. E., Hellstrom, I. and Ledbetter, J. A. (2001) CD83 is a sialic acid-binding Ig-like lectin (Siglec) adhesion receptor that binds monocytes and a subset of activated CD8+ T cells *J Immunol* 166 (6): 3865-72

Schurmans, S., Muscatelli, F., Miot, F., Mattei, M. G., Vassart, G. and Parmentier, M. (1993) The OLFR1 gene encoding the HGMP07E putative olfactory receptor maps to the 17p13-->p12 region of the human genome and reveals an MspI restriction fragment length polymorphism *Cytogenet Cell Genet* 63 (3): 200-4

Schwartz, S., Zhang, Z., Frazer, K. A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R. and Miller, W. (2000) PipMaker--a web server for aligning two genomic DNA sequences *Genome Res* 10 (4): 577-86

Schwarz, D. A., Barry, G., Eliasof, S. D., Petroski, R. E., Conlon, P. J. and Maki, R. A. (2000) Characterization of gamma -aminobutyric acid receptor GABAB(1e), a GABAB(1) splice variant encoding a truncated receptor *J Biol Chem* 275 (41): 32174-81

Seki, N., Ohira, M., Nagase, T., Ishikawa, K., Miyajima, N., Nakajima, D., Nomura, N. and Ohara, O. (1997) Characterization of cDNA clones in size-fractionated cDNA libraries from human brain *DNA Res* 4 (5): 345-9

Selbie, L. A., Townsend-Nicholson, A., Iismaa, T. P. and Shine, J. (1992) Novel G protein-coupled receptors: a gene family of putative human olfactory receptor sequences *Brain Res Mol Brain Res* 13 (1-2): 159-63

Senger, G., Ragoussis, J., Trowsdale, J., and Sheer, D. (1993) Fine mapping of the human MHC class II region within chromosome band 6p21 and evaluation of probe ordering using interphase fluorescence in situ hybridization *Cytogenet Cell Genet.* 64 (1):49-53

Sharon, D., Gilad, Y., Glusman, G., Khen, M., Lancet, D. and Kalush, F. (2000) Identification and characterization of coding single-nucleotide polymorphisms within a human olfactory receptor gene cluster *Gene* 260 (1-2): 87-94

Sharon, D., Glusman, G., Pilpel, Y., Horn-Saban, S. and Lancet, D. (1998) Genome dynamics, evolution, and protein modeling in the olfactory receptor gene superfamily *Ann N Y Acad Sci* 855 182-93

Sharon, D., Glusman, G., Pilpel, Y., Khen, M., Gruetzner, F., Haaf, T., and Lancet, D. (1999) Primate evolution of an olfactory receptor cluster: diversification by gene conversion and recent emergence of pseudogenes *Genomics* 61 (1): 24-36

Shiina, T., Tamiya, G., Oka, A., Takishima, N., Yamagata, T., Kikkawa, E., Iwata, K., Tomizawa, M., Okuaki, N., Kuwano, Y., Watanabe, K., Fukuzumi, Y., Itakura, S., Sugawara, C., Ono, A., Yamazaki, M., Tashiro, H., Ando, A., Ikemura, T., Soeda, E., Kimura, M., Bahram, S. and Inoko, H. (1999) Molecular dynamics of MHC genesis unraveled by sequence analysis of the 1,796,938-bp HLA class I region *Proc Natl Acad Sci U S A* 96 (23): 13282-7

Simon, M., Le Mignon, L., Fauchet, R., Yaouanq, J., David, V., Edan, G. and Bourel, M. (1987) A study of 609 HLA haplotypes marking for the hemochromatosis gene: (1) mapping of the gene near the HLA-A locus and characters required to define a heterozygous population and (2) hypothesis concerning the underlying cause of hemochromatosis-HLA association *Am J Hum Genet* 41 (2): 89-105

Singer, M. S., Hughes, T. E., Shepherd, G. M. and Greer, C. A. (1998) Identification of olfactory receptor mRNA sequences from the rat olfactory bulb glomerular layer *Neuroreport* 9 (16): 3745-8

Sinnarajah, S., Dessauer, C. W., Srikumar, D., Chen, J., Yuen, J., Yilma, S., Dennis, J. C., Morrison, E. E., Vodyanoy, V. and Kehrl, J. H. (2001) RGS2 regulates signal transduction in olfactory neurons by attenuating activation of adenylyl cyclase III *Nature* 409 (6823): 1051-5

Slade, R. W. and McCallum, H. I. (1992) Overdominant vs. frequency-dependent selection at MHC loci *Genetics* 132 (3): 861-4

Sloan, V. S., Cameron, P., Porter, G., Gammon, M., Amaya, M., Mellins, E. and Zaller, D. M. (1995) Mediation by HLA-DM of dissociation of peptides from HLA-DR *Nature* 375 (6534): 802-6

Sly, W. S., Quinton, B. A., McAlister, W. H. and Rimoin, D. L. (1973) Beta glucuronidase deficiency: report of clinical, radiologic, and biochemical features of a new mucopolysaccharidosis *J Pediatr* 82 (2): 249-57

Smaglik, P. and Abbott, A. (2000) Project offers free mouse sequence *Nature* 407 (6805): 663-4

Smit, A. F. (1996) The origin of interspersed repeats in the human genome *Curr Opin Genet Dev* 6 (6): 743-8

Smit, A. F. (1999) Interspersed repeats and other mementos of transposable elements in mammalian genomes *Curr Opin Genet Dev* 9 (6): 657-63

Smith, L. M., Sanders, J. Z., Kaiser, R. J., Hughes, P., Dodd, C., Connell, C. R., Heiner, C., Kent, S. B. and Hood, L. E. (1986) Fluorescence detection in automated DNA sequence analysis *Nature* 321 (6071): 674-9

Smith, T. F. and Waterman, M. S. (1981) Identification of common molecular subsequences *J Mol Biol* 147 (1): 195-7

Snyder, S. H., Sklar, P. B. and Pevsner, J. (1988) Molecular mechanisms of olfaction *J Biol Chem* 263 (28): 13971-4

Soderlund, C., Humphray, S., Dunham, A. and French, L. (2000) Contigs built with fingerprints, markers, and FPC V4.7 *Genome Res* 10 (11): 1772-87

Soderlund, C., Longden, I. and Mott, R. (1997) FPC: a system for building contigs from restriction fingerprinted clones *Comput Appl Biosci* 13 (5): 523-35

Solovyev, V. and Salamov, A. (1997) The Gene-Finder computer tools for analysis of human and model organisms genome sequences *Proc Int Conf Intell Syst Mol Biol* 5 294-302

Sonnhammer, E. L. and Durbin, R. (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis *Gene* 167 (1-2): GC1-10

Soriano, P., Macaya, G. and Bernardi, G. (1981) The major components of the mouse and human genomes. 2. Reassociation kinetics *Eur J Biochem* 115 (2): 235-9

Soriano, P., Meunier-Rotival, M. and Bernardi, G. (1983) The distribution of interspersed repeats is nonuniform and conserved in the mouse and human genomes *Proc Natl Acad Sci U S A* 80 (7): 1816-20

Sosinsky, A., Glusman, G. and Lancet, D. (2000) The genomic structure of human olfactory receptor genes *Genomics* 70 (1): 49-61

Speleman, F., Vervoort, R., van Roy, N., Liebaers, I., Sly, W. S. and Lissens, W. (1996) Localization by fluorescence in situ hybridization of the human functional beta-glucuronidase gene (GUSB) to 7q11.21 --> q11.22 and two pseudogenes to 5p13 and 5q13 *Cytogenet Cell Genet* 72 (1): 53-5

Staden, R., Beal, K. F. and Bonfield, J. K. (2000) The Staden package, 1998 *Methods Mol Biol* 132 115-30

Staden, R., Judge, D. P. and Bonfield, J. K. (2001) Sequence assembly and finishing methods *Methods Biochem Anal* 43 303-22

Stephens, R., Horton, R., Humphray, S., Rowen, L., Trowsdale, J. and Beck, S. (1999) Gene organisation, sequence variation and isochore structure at the centromeric boundary of the human MHC *J Mol Biol* 291 (4): 789-99

Stoesser, G., Baker, W., van den Broek, A., Camon, E., Garcia-Pastor, M., Kanz, C., Kulikova, T., Leinonen, R., Lin, Q., Lombard, V., Lopez, R., Redaschi, N., Stoehr, P., Tuli, M. A., Tzouvara, K. and Vaughan, R. (2002) The EMBL Nucleotide Sequence Database *Nucleic Acids Res* 30 (1): 21-6

Strotmann, J., Beck, A., Kubick, S. and Breer, H. (1995) Topographic patterns of odorant receptor expression in mammals: a comparative study *J Comp Physiol [A]* 177 (6): 659-66

Strotmann, J., Wanner, I., Helfrich, T., Beck, A. and Breer, H. (1994a) Rostro-caudal patterning of receptor-expressing olfactory neurones in the rat nasal cavity *Cell Tissue Res* 278 (1): 11-20

Strotmann, J., Wanner, I., Helfrich, T., Beck, A., Meinken, C., Kubick, S. and Breer, H. (1994b) Olfactory neurones expressing distinct odorant receptor subtypes are spatially segregated in the nasal neuroepithelium *Cell Tissue Res* 276 (3): 429-38

Strotmann, J., Wanner, I., Krieger, J., Raming, K. and Breer, H. (1992) Expression of odorant receptors in spatially restricted subsets of chemosensory neurones *Neuroreport* 3 (12): 1053-6

Sullivan, S. L., Adamson, M. C., Ressler, K. J., Kozak, C. A. and Buck, L. B. (1996) The chromosomal distribution of mouse odorant receptor genes *Proc Natl Acad Sci U S A* 93 (2): 884-8

Sullivan, S. L., Bohm, S., Ressler, K. J., Horowitz, L. F. and Buck, L. B. (1995) Target-independent pattern specification in the olfactory epithelium *Neuron* 15 (4): 779-89

Sulston, J., Mallett, F., Staden, R., Durbin, R., Horsnell, T. and Coulson, A. (1988) Software for genome mapping by fingerprinting techniques *Comput Appl Biosci* 4 (1): 125-32

Szpirer, C., Szpirer, J., Riviere, M., Tazi, R. and Pontarotti, P. (1997) Mapping of the Olf89 and Rfp genes to the rat genome: comparison with the mouse and human and new insights into the evolution of the rodent genome *Cytogenet Cell Genet* 78 (2): 137-9

Takahashi, K., Tsuchida, K., Tanabe, Y., Masu, M. and Nakanishi, S. (1993) Role of the large extracellular domain of metabotropic glutamate receptors in agonist selectivity determination *J Biol Chem* 268 (26): 19341-5

Takahata, N. and Nei, M. (1990) Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci *Genetics* 124 (4): 967-78

Tang, M. X., Redemann, C. T. and Szoka, F. C., Jr. (1996) In vitro gene delivery by degraded polyamidoamine dendrimers *Bioconjug Chem* 7 (6): 703-14

Tegoni, M., Pelosi, P., Vincent, F., Spinelli, S., Campanacci, V., Grolli, S., Ramoni, R. and Cambillau, C. (2000) Mammalian odorant binding proteins *Biochim Biophys Acta* 1482 (1-2): 229-40

The International Human Genome Sequencing Consortium [IHGSC]: Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J. and Szustakowki, J. (2001) Initial sequencing and analysis of the human genome *Nature* 409 (6822): 860-921

The MHC sequencing consortium (1999) Complete sequence and gene map of a human major histocompatibility complex. *Nature* 401 (6756): 921-3

Theodosiou, A. M., Morrison, K. E., Nesbit, A. M., Daniels, R. J., Campbell, L., Francis, M. J., Christodoulou, Z. and Davies, K. E. (1994) Complex repetitive arrangements of gene sequence in the candidate region of the spinal muscular atrophy gene in 5q13 *Am J Hum Genet* 55 (6): 1209-17

Thomas, M. B., Haines, S. L. and Akeson, R. A. (1996) Chemoreceptors expressed in taste, olfactory and male reproductive tissues *Gene* 178 (1-2): 1-5

Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice *Nucleic Acids Res* 22 (22): 4673-80

Thomson, G. (1995) HLA disease associations: models for the study of complex human genetic disorders *Crit Rev Clin Lab Sci* 32 (2): 183-219

Tiwari, J. L. and Terasaki, P. I. *HLA and Disease Associations* (Springer-Verlag, New York, 1985).

Tohgo, A., Takasawa, S., Munakata, H., Yonekura, H., Hayashi, N. and Okamoto, H. (1994) Structural determination and characterization of a 40 kDa protein isolated from rat 40 S ribosomal subunit *FEBS Lett* 340 (1-2): 133-8

Trappe, R., Doenecke, D. and Albig, W. (1999) The expression of human H2A-H2B histone gene pairs is regulated by multiple sequence elements in their joint promoters *Biochim Biophys Acta* 1446 (3): 341-51

Trask, B. J., Friedman, C., Martin-Gallardo, A., Rowen, L., Akinbami, C., Blankenship, J., Collins, C., Giorgi, D., Iadonato, S., Johnson, F., Kuo, W. L., Massa, H., Morrish, T., Naylor, S., Nguyen, O. T., Rouquier, S., Smith, T., Wong, D. J., Youngblom, J. and van den Engh, G. (1998) Members of the olfactory receptor gene family are contained in large blocks of DNA duplicated polymorphically near the ends of human chromosomes *Hum Mol Genet* 7 (1): 13-26

Trask, B. J., Massa, H., Brand-Arpon, V., Chan, K., Friedman, C., Nguyen, O. T., Eichler, E., van den Engh, G., Rouquier, S., Shizuya, H. and Giorgi, D. (1998) Large multi-chromosomal duplications encompass many members of the olfactory receptor gene family in the human genome *Hum Mol Genet* 7 (13): 2007-20

Twist, C. J., Beier, D. R., Disteche, C. M., Edelhoff, S. and Tedder, T. F. (1998) The mouse Cd83 gene: structure, domain organization, and chromosome localization *Immunogenetics* 48 (6): 383-93

Tzachanis, D., Freeman, G. J., Hirano, N., van Puijenbroek, A. A., Delfs, M. W., Berezovskaya, A., Nadler, L. M. and Boussiotis, V. A. (2001) Tob is a negative regulator of activation that is expressed in anergic and quiescent T cells *Nat Immunol* 2 (12): 1174-82

Uberbacher, E. C. and Mural, R. J. (1991) Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach *Proc Natl Acad Sci U S A* 88 (24): 11261-5

Vanderhaeghen, P., Schurmans, S., Vassart, G. and Parmentier, M. (1993) Olfactory receptors are displayed on dog mature sperm cells *J Cell Biol* 123 (6 Pt 1): 1441-52

Vanderhaeghen, P., Schurmans, S., Vassart, G. and Parmentier, M. (1997a) Molecular cloning and chromosomal mapping of olfactory receptor genes expressed in the male germ line: evidence for their wide distribution in the human genome *Biochem Biophys Res Commun* 237 (2): 283-7

Vanderhaeghen, P., Schurmans, S., Vassart, G. and Parmentier, M. (1997b) Specific repertoire of olfactory receptor genes in the male germ cells of several mammalian species *Genomics* 39 (3): 239-46

Vassar, R., Chao, S. K., Sitcheran, R., Nunez, J. M., Vosshall, L. B. and Axel, R. (1994) Topographic organization of sensory projections to the olfactory bulb *Cell* 79 (6): 981-91

Vassar, R., Ngai, J. and Axel, R. (1993) Spatial segregation of odorant receptor expression in the mammalian olfactory epithelium *Cell* 74 (2): 309-18

Velten, F., Rogel-Gaillard, C., Renard, C., Pontarotti, P., Tazi-Ahnini, R., Vaiman, M. and Chardon, P. (1998) A first map of the porcine major histocompatibility complex class I region *Tissue Antigens* 51 (2): 183-94

Vernet, P., Faure, J., Dufaure, J. P. and Drevet, J. R. (1997) Tissue and developmental distribution, dependence upon testicular factors and attachment to spermatozoa of GPX5, a murine epididymis-specific glutathione peroxidase *Mol Reprod Dev* 47 (1): 87-98

Vogt, R. G., Lindsay, S. M., Byrd, C. A. and Sun, M. (1997) Spatial patterns of olfactory neurons expressing specific odor receptor genes in 48-hour-old embryos of zebrafish Danio rerio *J Exp Biol* 200 (Pt 3) 433-43

Volz, A., Ehlers, A., Younger, R. M., Forbes, S., Trowsdale, J., Beck, S. and Ziegler, A. (unpublished) Complex transcriptional control of MHC-linked olfactory receptor genes includes long distance and extensive alternative splicing, exon sharing and premature polyadenylation

Volz, A. and Ziegler, A. (1996) Physical mapping of a 6 Mbp region directly telomeric of the HLA-complex *DNA Seq* 7 (1): 61-2

Volz, A., Fonatsch, C. and Ziegler, A. (1992) Regional mapping of the gene for autosomal dominant spinocerebellar ataxia (SCA1) by localizing the closely linked D6S89 locus to 6p24.2----p23.05 *Cytogenet Cell Genet* 60 (1): 37-9

Walensky, L. D., Roskams, A. J., Lefkowitz, R. J., Snyder, S. H. and Ronnett, G. V. (1995) Odorant receptors and desensitization proteins colocalize in mammalian sperm *Mol Med* 1 (2): 130-41

Walensky, L. D., Ruat, M., Bakin, R. E., Blackshaw, S., Ronnett, G. V. and Snyder, S. H. (1998) Two novel odorant receptor families expressed in spermatids undergo 5'-splicing *J Biol Chem* 273 (16): 9378-87

Wang, C. R., Loveland, B. E. and Lindahl, K. F. (1991) H-2M3 encodes the MHC class I molecule presenting the maternally transmitted antigen of the mouse *Cell* 66 (2): 335-45

Wang, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M. S., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T. J., Lander, E. S. and et al. (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome *Science* 280 (5366): 1077-82

Wang, F., Nemes, A., Mendelsohn, M. and Axel, R. (1998) Odorant receptors govern the formation of a precise topographic map *Cell* 93 (1): 47-60

Wang, M. M. and Reed, R. R. (1993) Molecular cloning of the olfactory neuronal transcription factor Olf-1 by genetic selection in yeast *Nature* 364 (6433): 121-6

Warner, J. R. and Nierras, C. R. (1998) Trapping human ribosomal protein genes *Genome Res* 8 (5): 419-21

Weber, D. A., Evavold, B. D. and Jensen, P. E. (1996) Enhanced dissociation of HLA-DR-bound peptides in the presence of HLA-DM *Science* 274 (5287): 618-20

Weighardt, F., Biamonti, G. and Riva, S. (1995) Nucleo-cytoplasmic distribution of human hnRNP proteins: a search for the targeting domains in hnRNP A1 *J Cell Sci* 108 (Pt 2) 545-55

Wellerdieck, C., Oles, M., Pott, L., Korsching, S., Gisselmann, G. and Hatt, H. (1997) Functional expression of odorant receptors of the zebrafish Danio rerio and of the nematode C. elegans in HEK293 cells *Chem Senses* 22 (4): 467-76

Weth, F., Nadler, W. and Korsching, S. (1996) Nested expression domains for odorant receptors in zebrafish olfactory epithelium *Proc Natl Acad Sci U S A* 93 (23): 13321-6

Wetzel, C. H., Oles, M., Wellerdieck, C., Kuczkowiak, M., Gisselmann, G. and Hatt, H. (1999) Specificity and sensitivity of a human olfactory receptor functionally expressed in human embryonic kidney 293 cells and Xenopus Laevis oocytes *J Neurosci* 19 (17): 7426-33

Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., Meinhardt, T., Pruss, M., Reuter, I. and Schacherer, F. (2000) TRANSFAC: an integrated system for gene expression regulation *Nucleic Acids Res* 28 (1): 316-9

Wolfsberg, T. G. and Landsman, D. (1997) A comparison of expressed sequence tags (ESTs) to human genomic sequences *Nucleic Acids Res* 25 (8): 1626-32

Womble, D. D. (2000) GCG: The Wisconsin Package of sequence analysis programs *Methods Mol Biol* 132 3-22

Wong, S. T., Trinh, K., Hacker, B., Chan, G. C., Lowe, G., Gaggar, A., Xia, Z., Gold, G. H. and Storm, D. R. (2000) Disruption of the type III adenylyl cyclase gene leads to peripheral and behavioral anosmia in transgenic mice *Neuron* 27 (3): 487-97

Wysocki, C. J. and Beauchamp, G. K. (1984) Ability to smell androstenone is genetically determined *Proc Natl Acad Sci U S A* 81 (15): 4899-902

Wysocki, C. J., Dorries, K. M. and Beauchamp, G. K. (1989) Ability to perceive androstenone can be acquired by ostensibly anosmic people *Proc Natl Acad Sci U S A* 86 (20): 7976-8

Xu, Y., Mural, R., Shah, M. and Uberbacher, E. (1994) Recognizing exons in genomic sequence using GRAIL II *Genet Eng (N Y)* 16 241-53

Yamagata, K., Goto, K., Kuo, C. H., Kondo, H. and Miki, N. (1990) Visinin: a novel calcium binding protein expressed in retinal cone cells *Neuron* 4 (3): 469-76

Yamaguchi, M., Yamazaki, K., Beauchamp, G. K., Bard, J., Thomas, L. and Boyse, E. A. (1981) Distinctive urinary odors governed by the major histocompatibility locus of the mouse *Proc Natl Acad Sci U S A* 78 (9): 5817-20

Yamazaki, K., Beauchamp, G. K., Kupniewski, D., Bard, J., Thomas, L. and Boyse, E. A. (1988) Familial imprinting determines H-2 selective mating preferences *Science* 240 (4857): 1331-2

Yamazaki, K., Beauchamp, G. K., Wysocki, C. J., Bard, J., Thomas, L. and Boyse, E. A. (1983) Recognition of H-2 types in relation to the blocking of pregnancy in mice *Science* 221 (4606): 186-8

Yamazaki, K., Boyse, E. A., Mike, V., Thaler, H. T., Mathieson, B. J., Abbott, J., Boyse, J., Zayas, Z. A. and Thomas, L. (1976) Control of mating preferences in mice by genes in the major histocompatibility complex *J Exp Med* 144 (5): 1324-35

Yoon, S. J., LeBlanc-Straceski, J., Ward, D., Krauter, K. and Kucherlapati, R. (1994) Organization of the human keratin type II gene cluster at 12q13 *Genomics* 24 (3): 502-8

Yoshino, M., Xiao, H., Amadou, C., Jones, E. P. and Lindahl, K. F. (1998a) BAC clones and STS markers near the distal breakpoint of the fourth t-inversion, In(17)4d, in the H2-M region on mouse chromosome 17 *Mamm Genome* 9 (3): 186-92

Yoshino, M., Xiao, H., Jones, E. P. and Fischer Lindahl, K. (1998b) BAC/YAC contigs from the H2-M region of mouse Chr 17 define gene order as Znf173-Tctex5-mog-D17Tu42-M3-M2 *Immunogenetics* 47 (5): 371-80

Yoshino, M., Xiao, H., Jones, E. P., Kumanovics, A., Amadou, C. and Fischer Lindahl, K. (1997) Genomic evolution of the distal Mhc class I region on mouse Chr 17 *Hereditas* 127 (1-2): 141-8

Young, J. M. and Trask, B. J. (2002) The sense of smell: genomics of vertebrate odorant receptors *Hum Mol Genet* 11 (10): 1153-60

Young, J. M., Friedman, C., Williams, E. M., Ross, J. A., Tonnes-Priddy, L. and Trask, B. J. (2002) Different evolutionary processes shaped the mouse and human olfactory receptor gene families *Hum Mol Genet* 11 (5): 535-46

Younger, R. M., Amadou, C., Bethel, G., Ehlers, A., Lindahl, K. F., Forbes, S., Horton, R., Milne, S., Mungall, A. J., Trowsdale, J., Volz, A., Ziegler, A. and Beck, S. (2001) Characterization of clustered MHC-linked olfactory receptor genes in human and mouse *Genome Res* 4 (11): 519-30

Zeng, F. Y. and Wess, J. (1999) Identification and molecular characterization of m3 muscarinic receptor dimers *J Biol Chem* 274 (27): 19487-97

Zerial, M., Salinas, J., Filipski, J. and Bernardi, G. (1986) Gene distribution and nucleotide sequence organization in the human genome *Eur J Biochem* 160 (3): 479-85

Zhang, M. Q. (1997) Identification of protein coding regions in the human genome by quadratic discriminant analysis *Proc Natl Acad Sci U S A* 94 (2): 565-8

Zhang, Q. H., Ye, M., Wu, X. Y., Ren, S. X., Zhao, M., Zhao, C. J., Fu, G., Shen, Y., Fan, H. Y., Lu, G., Zhong, M., Xu, X. R., Han, Z. G., Zhang, J. W., Tao, J., Huang, Q. H., Zhou, J., Hu, G. X., Gu, J., Chen, S. J. and Chen, Z. (2000) Cloning and functional analysis of cDNAs with open reading frames for 300 previously undefined genes expressed in CD34+ hematopoietic stem/progenitor cells *Genome Res* 10 (10): 1546-60

Zhang, X. and Firestein, S. (2002) The olfactory receptor gene superfamily of the mouse *Nat Neurosci* 5 (2): 124-33

Zhao, H. and Firestein, S. (1999) Vertebrate odorant receptors *Cell Mol Life Sci* 56 (7-8): 647-59

Zhao, H., Ivic, L., Otaki, J. M., Hashimoto, M., Mikoshiba, K. and Firestein, S. (1998) Functional expression of a mammalian odorant receptor *Science* 279 (5348): 237-42

Zhou, L. J., Schwarting, R., Smith, H. M. and Tedder, T. F. (1992) A novel cell-surface molecule expressed by human interdigitating reticulum cells, Langerhans cells, and activated lymphocytes is a new member of the Ig superfamily *J Immunol* 149 (2): 735-42

Zhou, Q., Hinkle, G., Sogin, M. L. and Dionne, V. E. (1997) Phylogenetic analysis of olfactory receptor genes from mudpuppy (Necturus maculosus) *Biol Bull* 193 (2): 248-50

Ziegler, A., Dohr, G. and Uchanska-Ziegler, B. (2002) Possible roles for products of polymorphic MHC and linked olfactory receptor genes during selection processes in reproduction *Am J Reprod Immunol* 48 (1): 34-42

Ziegler, A., Beck, S., Ehlers, A., Younger, R. and Volz, A. (2002) *Testicular transcripts of HLA-linked olfactory receptor genes exhibit unorthodox features* In Hansen, J.A. and Dupont, B. (Ed) *HLA 2002 Immunobiology of the Human MHC*

Ziegler, A., Ehlers, A., Forbes, S., Trowsdale, J., Volz, A., Younger, R. and Beck, S. (2000) Polymorphisms in olfactory receptor genes: a cautionary note *Hum Immunol* 61 (12): 1281-4

Ziegler, A., Muller, C., Heinig, J., Radka, S. F., Kompf, J. and Fonatsch, C. (1985) Monosomy 6 in a human lymphoma line induced by selection with a monoclonal antibody *Immunobiology* 169 (5): 455-60

Zozulya, S., Echeverri, F. and Nguyen, T. (2001) The human olfactory receptor repertoire *Genome Biol* 2 (6): RESEARCH0018

Appendix 1: List of non-standard primers used

Construction of mouse OR gene vectors for *in situ* hybridisation [2.8]

MP1A	GTCTCATCCTGCCAACTAGACA	59.36	>
MP1.2.5B	TATGGATTGGACCAAACCTATA	56.11	<
MP1.2.5C	TATAGGTTTGGTCCAATCCATA	56.11	>
MP1D	CAAATTCATCTTAATTTGGAAA	53.88	<
MP2A	GTCTCATTCCTCCAACTAGATA	52.76	>
MP2D	AAGAAACTTTGTAATTCAGAAA	50.61	<
MP3A	GTCTCATCCCTCCAACTAGATA	55	>
MP3B	TACTGACTCCACTAGGCCAATC	58.34	<
MP3C	GATTGGCCTAGTGGAGTCAGTA	58.34	>
MP3D	TGCAGATTTCCCATTCAGCTTA	61.94	<
MP4A	CTTCCAGTTGGATTCATAGCAG	58.86	>
MP4B	TATGGATTGAAGCAGTCCCATC	61.2	<
MP4C	GATGGGACTGCTTCAATCCATA	61.2	>
MP4D	ATCTGGCCTTGAACTTACAGCA	61.16	<
MP5A	GTCTCATCCTGCCGACTAGACA	61.74	>
MP5D	TTCTAATTCATCTTAATTTGGA	51.28	<
MP6A	GGTTAGCTAGCGATAGACTTAG	52.48	>
MP6B	TGGCACCACACTAGTGGTGAAA	63.77	<
MP6C	TTTCACCACTAGTGTGGTGCCA	63.77	>
MP6D	AGAGAAAAGAATCTCGCTACTA	52.31	<
MP7A	TATTGCCCATAAGACTTATGGA	56.39	>
MP7B	ACGTCAATATGGAGTGAACCAG	58.98	<
MP7C	CTGGTTCACTCCATATTGACGT	58.98	>
MP7D	ACATTGATTTTCACAATAACTG	52.06	<
MP8A	CATGAAATTATTGCCTAGTAA	50.15	>
MP8B	ATTGAGACGTCAATATAGAGTG	51.25	<
MP8C	CACTCTATATTGACGTCTCAAT	51.25	>
MP8D	ATCATAAACAATTGCAGAATC	51.99	<
MP9A	TATTGACTTTAAAGATGATGGA	51.86	>
MP9B	ATTGAGATGTCAATATGGAGTG	54.17	<
MP9C	CACTCCATATTGACATCTCAAT	54.17	>
MP9D	ATCATACACAATTGCAGAATC	52.64	<
MP10A	GCTATCATAATTTCCATTTCTC	52.56	>
MP10B	CTCACTGAGGTGAGTGTCTGAG	57.64	<
MP10C	CTCAGACACTCACCTCAGTGAG	57.64	>
MP10D	ATGAGCCATGGTATTAACTGAC	55.78	<
MP11A	GCGTTCTAGTTTCTATTACTCA	51.14	>
MP11B	CTCATTGAGGTGAGTGTCAGAA	57.93	<
MP11C	TTCTGACACTCACCTCAATGAG	57.93	>
MP11D	CTGAGCAACAGTGTTAACTGAC	55.22	<
MP12.13B	ACAGAAGAAATGATGGACATGA	57.11	<
MP12.13C	TCATGTCCATCATTTCTTCTGT	57.11	>
MP12A	GTCAACATGCAGGGTAGTATT	54.27	>

MP12D	CTGGTTATTTGCACTGAAGAAG	56.76	<
MP13A	GTTAACATGCAGGGTAGTATT	51.22	>
MP13D	CTGGTTATTTACACCAAAGCAG	56.61	<
MP14A	CTGGTTGCATTCGACAAACTAT	59.16	>
MP14B	ACAGAAGAAATGATGAATATGA	51.54	<
MP14C	TCATATTCATCATTTCTTCTGT	51.54	>
MP14D	CTGGTTATTTACACCTCAGTAG	51.1	<
MP17A	TAGATAGTGTGAGGTCTATGCA	52.7	>
MP17B	TACACTTTCTGGGACTCATGAT	56.26	<
MP17C	ATCATGAGTCCCAGAAAGTGTA	56.26	>
MP17D	AAACCAATACATTGTCTTTCAT	53.04	<
MP18A	TTGATAGTGTGAGATGTATGGA	53.16	>
MP18.24B	TGCACTTTCTGGGACTCATGAT	61.97	<
MP18.24C	ATCATGAGTCCCAGAAAGTGCA	61.97	>
MP18D	ACATTAGACTTTGAAGTCATCC	52.67	<
MP19A	GACTGGATAGCTTGAGATCTAA	53.09	>
MP19B	TACACTTTCTGGGACACATGAT	57.05	<
MP19C	ATCATGTGTCCCAGAAAGTGTA	57.05	>
MP19D	ACATTAGAAATTGAAGGAATCT	51.64	<
MP20A	TGGATAGCTTGTGATCTATGTA	52.69	>
MP20B	TGATGACTTCATAGTGAAGTGG	55.32	<
MP20C	CCACTTCACTATGAAGTCATCA	55.32	>
MP20D	ACATTAGAGTTTGAAGGAAACC	54.3	<
MP21A	CACCTTTAGGGTCCATGTCTCT	59.5	>
MP21B	GATGAAGTGATAAATTTTGTGG	53.97	<
MP21C	CCACAAAATTTATCACTTCATC	53.97	>
MP21D	CATGTGGATCAACACCTTCATT	59.72	<
MP22A	ACATGGTAAGTCTTCTGCCTGA	58.86	>
MP22B	CATTCGGGTTGAGTAATGGAGT	60.24	<
MP22C	ACTCCATTACTCAACCCGAATG	60.24	>
MP22D	GGCAGACAAAAGAATGTACCAG	58.77	<
MP23A	CATTATTGGTACTAGCACTTGC	54.44	>
MP23B	GATTGTTACCACAGAAGGGCAG	60.9	<
MP23C	CTGCCCTTCTGTGGTAACAATC	60.9	>
MP23D	TTCTAATGTAAGATCATTGACC	51.23	<
MP24A	TATGAGGAAATAGTCTGAGATG	51.3	>
MP24D	ACATTACAGTTTGAAGGAATCC	54.93	<

Construction of ‘olfactory promoter region’ pGL3 luciferase reporter vectors [2.9]

OLFOP-F	CCTCGAGATCTTCCCTTACCTTTGCATGG	59.56	>
OLFOP-R	GATAGATCTTTCTATGGGCCAGCAATTC	60.04	<

Polymorphism analysis [2.11]

17C	TTGTCTTTCTGACAGGCTGG	59.01	>
17D	AGGGAGATCTAGTGCTGCGA	60.12	<
17E	CATAGAAGAGGGAGACCACGAT	59.6	>
17F	GTATCCTCGTTACCATCTTCCG	59.86	<
17G	AGAAGAACATCTGGAGAGCACA	59.09	>
17H	GATTGGCTATACGTCTGTCACG	59.66	<

20C	TTTTTCTTTTATCCAGTTGCCTC	58.85	>
20D	GGTGAGAAAATTCTGAGCCG	59.81	<
20E	TGATTAAGCTCAGTGTTCCCAC	59.24	>
20F	CCCCTTCTTTCTGACACTTCTC	59.36	<
20G	CAGATAGCCACAGAGAGGTCAA	59.5	>
20H	GGAAACCTGTCCTACCTGGATA	59.35	<

22C	CTATTGGGATTTTCTGACCGTC	59.84	>
22D	TATGCATTCAGTAGAACCCAGG	59.11	<
22E	GGTTGTGTTGCCCAACTCTATA	59.02	>
22F	TTCACAGAAGAAATGGTCCAAC	59.08	<
22G	TTGTCCAGTCCACTCTCACAGT	59.8	>
22H	ACAGGTATTGAAAGCCTTCCAG	59.65	<
22J	TCAGGCAGTAATGAGAATCTGC	59.48	>
22K	ACTGTTCCCTCATCAATTCCAT	59.7	<
22L	CCATGCATGTAGAGTTGATGGT	59.88	>
22M	ACCTCACGATAGTACTGGCATG	59.15	<
22N	CATTGATGATGTTGTCTCCACA	59.42	>
22P	ACCTTACAAAAGCATTTGAGGG	59.54	<
22Q	CCAAAGCATCTCACTGTTCATC	59.74	<
22R	CCACTGAGTTTGTGTCCCTAAA	59.14	<
22S	GCACATCCTCCTGTAGTCTTCA	59.36	>
22T	CTGTTGCCACCATTACATATCC	59.23	>

Mouse RT-PCR [2.12]

1a	CCTGACTTCCTCTCAGCCAC	59.99	>
1b	GGAGCACAAGACAACGACAA	59.88	<
2a	CTGAGAAGCACAAAAAGGGC	59.99	>
2b	TTTCCAGTTGTGGGTGTTCA	59.98	<
3a	CACCACACTCCTGGTTTCCT	60	>
3b	TTTCCAGTTGTGGGTGTTCA	59.98	<
4a	GGATGCACCTGTTCCTTGTT	59.97	>
4b	TCCAGTTGTGGGTGTTCAGA	60.13	<
6a	CAACGCCCACTTTTAAGGAA	60.1	>
6b	CCAGGTCTGTGTTCCTGGTT	60	<

Human RNA dot-blot hybridisations [2.13]

15A	GCAAGTAATTCCTAGGTCATGGA	59.51	>
15B	TGATCAGGGACATAAAGGAGC	59.14	<
16A	GCGAGCACTCAGGAGGTTAC	60.02	>
16B	CAGGTGTATGTGGGAAGGGT	59.7	<
20A	TAGCCATGGAACCCTAATGC	59.92	>
20B	CATGGGCAATATAGGCAAAA	58.51	<

PCR amplification of olfactory epithelium phage library [2.15]

16EST1	ACGAGGTTTCACCATGTTGA	>	52.4
16EST2a	ATCAGAAGGAACAGGGAACGA	>	55.07
16EST2b	TTTTCTGTCTCTGCTCACCCA	<	54.2
16EX3	TCAGACTCTCTTCACGGCCTT	>	55.14
16EX4a	TGTTTGTACCTGGAGGGATGA	<	54.6
16EX4b	AAGTCAGAGGCACCAATGTGA	>	54.01
16C	CAGGTGTATGTGGGAAGGGT	>	59.70
16D	CACCCAGGCTGAGTTTTGAT	<	60.11
16E	TAGAAGAGGGTGACCACAGTGA	>	59.76
16F	AGGCAGTGCTGAGGATTAACTC	<	59.91
16G	ACAGGAAGATGAAGAGCTGGAC	>	59.88
16H	CAAAGAAGACCATCAGCTTCCT	<	59.89

Appendix 2: Sequence feature files and output of ‘olfgrab’

“.genes” file

```
> (or) <        230     7898    genename       CDS (or) mRNA
            (tab)    (tab)   (tab)         (tab)


                230     560                 ⎫
              (tab)   (tab)                 ⎬
                6560    6989                ⎭      (exons)
              (tab)   (tab)
                7854    7898
              (tab)   (tab)
```

etc. for all genes present in the clone.

The >/< determines whether the gene is present on the forward or reverse strand, whilst the gene name should ideally consist of one word (although - , _ ,* or numbers are acceptable).

“.rptm” file

```
68      145     repeatname     +(or)C         repeattype
   (tab)    (tab)         (tab)        (tab)
575     4260    repeatname     +(or)C         repeattype
   (tab)    (tab)         (tab)        (tab)
```

etc. for all repeats present in the clone.

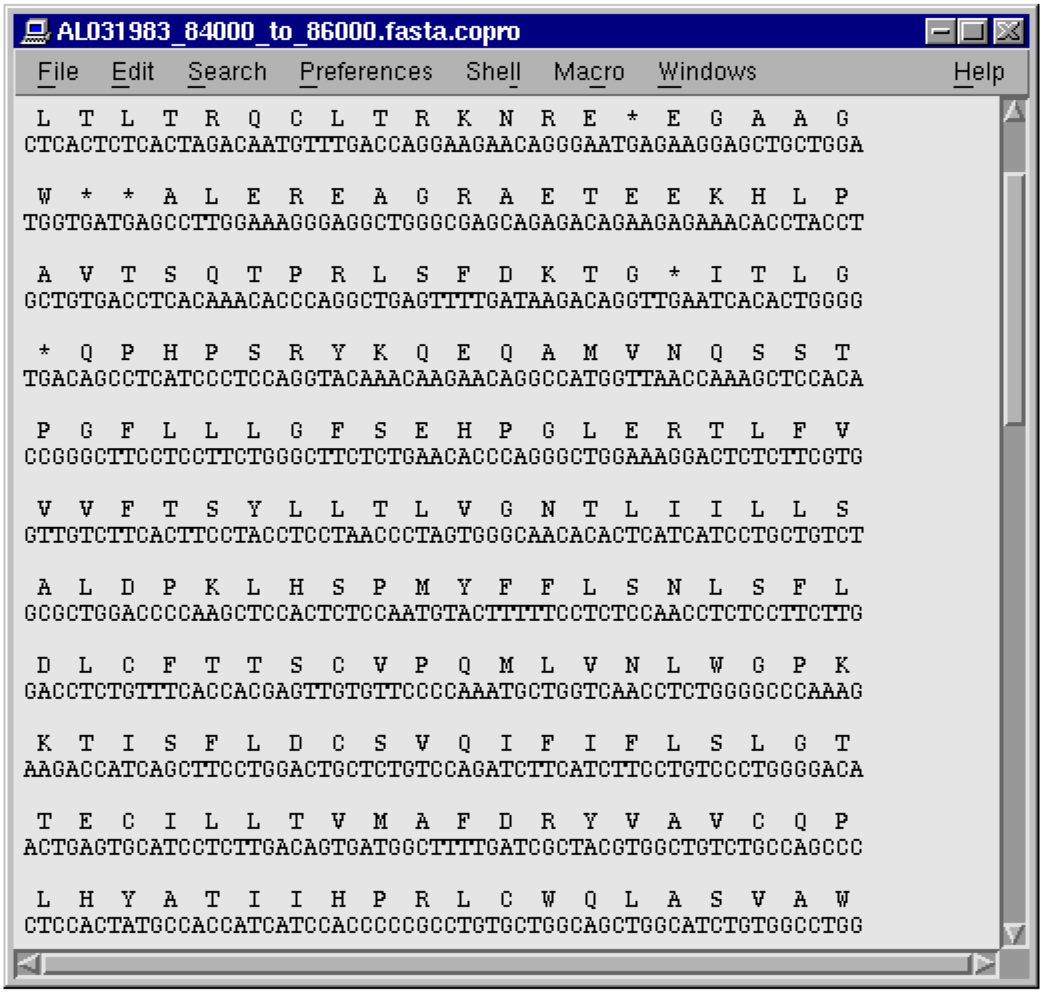
The +/C states whether the repeat match is on the forward or complement strand.

“.misf” file

12235 12325 CpG

(tab) *(tab)*

etc. for all CpG islands/other features in the clone.



```
 L  T  L  T  R  Q  C  L  T  R  K  N  R  E  *  E  G  A  A  G
CTCACTCTCACTAGACAATGTTTGACCAGGAAGAACAGGGAATGAGAAGGAGCTGCTGGA

 W  *  *  A  L  E  R  E  A  G  R  A  E  T  E  E  K  H  L  P
TGGTGATGAGCCTTGGAAAGGGAGGCTGGGCGAGCAGAGACAGAAGAGAAACACCTACCT

 A  V  T  S  Q  T  P  R  L  S  F  D  K  T  G  *  I  T  L  G
GCTGTGACCTCACAAACACCCAGGCTGAGTTTTGATAAGACAGGTTGAATCACACTGGGG

 *  Q  P  H  P  S  R  Y  K  Q  E  Q  A  M  V  N  Q  S  S  T
TGACAGCCTCATCCCTCCAGGTACAAACAAGAACAGGCCATGGTTAACCAAAGCTCCACA

 P  G  F  L  L  L  G  F  S  E  H  P  G  L  E  R  T  L  F  V
CCGGGCTTCCTCCTTCTGGGCTTCTCTGAACACCCAGGGCTGGAAAGGACTCTCTTCGTG

 V  V  F  T  S  Y  L  L  T  L  V  G  N  T  L  I  I  L  L  S
GTTGTCTTCACTTCCTACCTCCTAACCCTAGTGGGCAACACACTCATCATCCTGCTGTCT

 A  L  D  P  K  L  H  S  P  M  Y  F  F  L  S  N  L  S  F  L
GCGCTGGACCCCAAGCTCCACTCTCCAATGTACTTTTTCCTCTCCAACCTCTCCTTCTTG

 D  L  C  F  T  T  S  C  V  P  Q  M  L  V  N  L  W  G  P  K
GACCTCTGTTTCACCACGAGTTGTGTTCCCCAAATGCTGGTCAACCTCTGGGGCCCAAAG

 K  T  I  S  F  L  D  C  S  V  Q  I  F  I  F  L  S  L  G  T
AAGACCATCAGCTTCCTGGACTGCTCTGTCCAGATCTTCATCTTCCTGTCCCTGGGGACA

 T  E  C  I  L  L  T  V  M  A  F  D  R  Y  V  A  V  C  Q  P
ACTGAGTGCATCCTCTTGACAGTGATGGCTTTTGATCGCTACGTGGCTGTCTGCCAGCCC

 L  H  Y  A  T  I  I  H  P  R  L  C  W  Q  L  A  S  V  A  W
CTCCACTATGCCACCATCATCCACCCCCGCCTGTGCTGGCAGCTGGCATCTGTGGCCTGG
```

File created by the ‘olfgrab’ program. Key OR motifs can be identified (for example, MYFFL, MAFDRY), and the file can be edited and then run through the ‘olfproducer’ program to produce nucleotide and protein files.

Appendix 3: List of websites

HGMP	*http://www.hgmp.mrc.ac.uk*
RepeatMasker	*http://ftp.genome.washington.edu/cgi-bin/RepeatMasker*
Repbase	*http://www.geospiza.com/products/tools/repbase.htm*
GRAIL1	*http://compbio.ornl.gov/Grail-1.3*
Genfinder	*http://argon.cshl.org/genefinder/human.htm*
Genemark	*http://www.ebi.ac.uk/genemark/*
Fex	*http://genomic.sanger.ac.uk/gf/gf.shtml*
Fgene	*http://genomic.sanger.ac.uk/gf/gf.html*
Genscan	*http://genes.mit.edu/GENSCAN.html*
EBI	*http://www.ebi.ac.uk*
NCBI	*http://www.ncbi.nlm.nih.gov*
Pfam	*http://www.sanger.ac.uk/Software/Pfam/search.shtml*
BLOCKS	*http://blocks.fhcrc.org/blocks/blocks_search.html*
PRINTS	*http://bioinf.man.ac.uk/cgi-bin/dbbrowser/fingerPRINTScan/muppet/FPScan.cgi*
PROSITE	*http://www.expasy.ch/prosite/*
Psort	*http://psort.nibb.ac.jp/*
DSC	*http://www.hgmp.mrc.ac.uk/Registered/Option/dsc.html*
Simpa96	*http://npsa-pbil.ibcp.fr/cgibin/npsa_automat.pl?page=/NPSA/npsa_simpa96.html*
Phd	*http://www.public.iastate.edu/~pedro/pprotein_query.html*
Predator	*http://www.embl-heidelberg.de/cgi/predator_serv.pl*
COILS	*http://dot.imgen.bcm.tmc.edu:9331/seq-search/struc-predict.html*
Tmpred	*http://www.ch.embnet.org/software/TMPRED_form.html*
Tmap	*http://bioweb.pasteur.fr/seqanal/interfaces/tmap.html*
DAS	*http://www.sbc.su.se/~miklos/DAS/*

HTH	*http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_hth.html*
Signal	*http://www.cbs.dtu.dk/services/SignalP/*
Sigcleave	*http://bioweb.pasteur.fr/seqanal/interfaces/sigcleave.html*
Antigenic	*http://bioweb.pasteur.fr/seqanal/interfaces/antigenic.html*
Digest	*http://www2.no.embnet.org/Pise/digest-simple.html*
Dotter	*http://www.cgr.ki.se/cgr/groups/sonnhammer/Dotter.html*
Ensembl	*http://www.ensembl.org*
UCSC	*http://genome.ucsc.edu*
PipMaker	*http://bio.cse.psu.edu/pipmaker*
Promoter Inspector	*http://www.genomatix.de/cgi-bin/promoterinspector/promoterinspector.pl*
TRANSFAC	*http://transfac.gbf.de/TRANSFAC/*

Appendix 4: List of scripts/programs

Script	Author	Description	Input	Additional information
AAbreak	rmy	Pulls out co-ordinates of olfactory domains, and also pulls out position of stops (*) and frameshifts (X)	1 protein file	Asks "Domain Information only ?", N/n -will show entire protein with amino acids numbered.
accmatch	rmy	Takes 2 lists of accession numbers and compares them to see if there are any numbers that match.	2 lists of accession numbers.	
addclone	rmy	Adds 2 clones together to produce new ".genes", ".rptm" and ".misf" files	2 ".genes" files 2 ".rptm" files 2 ".misf" files.	Asks for lengths of clones (defaults will be taken from EMBL) and for the overlap of the clone(default=100)
addclonegenes	rmy	Adds certain length to ".genes" file, produces ".ADD" file.	1 ".genes" file	Asks for length to be added to file.
addclonemisf	rmy	Adds certain length to ".misf" file, produces ".ADD" file.	1 ".misf" file	Asks for length to be added to file.
addclonerptm	rmy	Adds certain length to ".rptm" file, produces ".ADD" file.	1 ".rptm" file	Asks for length to be added to file.
blastcut	rmy	Given a blast results file and an accession number, chops out the first result associated with the accession number.	BLAST output file, accession number.	Output file: ".blast"
blastextract	rmy	Produces a list of BLAST hits, showing accession number and start and stop positions.	BLAST output file	Writes to the screen.
changename	rmy	Changes the name within an OR file to a new name.	Name of OR file, new name.	Produces a .bak file showing old contents of file.
chopseq	rmy	Takes a piece of sequence and produces files containing sequence of a specified length.	1 FASTA format file	Produces file.001, file.002 etc.
chopuprptm	rmy	Pulls out repeats from a given region in a ".rptm" file.	1 ".rptm" file, a start co-ordinate an a stop co-ordinate.	File is named .rptm_start_to_stop

clonebreak	rmy	Takes a sequence file and produces files of each piece of sequence divided from the other by 5 or more 'n''s	1 unfinished sequence file (containing 'n''s)	Files are given increasing numbers as they are produced.
cloneends	rmy	Given a sequence file pulls out the first and last 100 bases and puts these in ".start" and ".end" files.	1 finshed fasta file.	
count	rog	Counts the number of bases in a sequence file, and calculates GC content.	1 sequence file	
eponinehits	rmy	Takes hits from eponine output file and converts to ".misf" format.	1 eponine output file	
estlocate	rmy	Takes an EST accession number and BLAST searches available genomic databases.	1 EST accession number	
fembl	rog	Takes accession number and produces file of fasta sequence of accession number	1 accession number	Output file: ".fasta"
gc_table	rog	Calculates gc content across a specified number of bases in a fasta file.	1 sequence file	Prompts for length of bases over which content is to be calculated. Produces "gc_out" file.
getgeneSEQ	rmy	Gets sequence for a gene specified in a gene file. (Name of file corresponds to EMBL sequence file)	1 EMBL-based ".genes" file	
getspecies getspecieslots	rmy	Gets species an accession number is derived from.	1 accession number List of accession numbers	
hembl	rog	Takes an accession number and produces file showing descriptive part of EMBL file.	1 accession number	Output file: ".head"
listlook	rmy	Takes a file and pulls out a list of accession numbers featured in file.	File containing accession numbers.	
localareagrab	rmy	Takes list of accession numbers and stop and start positions and grabs sequence -500 to +500 upstream and downstream of these positions.	List of files containing accession numbers and start and stop positions.	Produces a number of ".geneseq" files.

olfcheck	rmy	Takes protein file, strips it of '*' and 'X' characters and compares edited protein against other protein files in the directory (using the Emboss program 'water'). Reports hits over 1500.	1 protein file (Other proteins in directory named ".pro")	Output file: ".hits"
olfchromosome	rmy	Takes accession number and looks in EMBL file to see if chromosome is attached to accession number.	1 accession number	
olfclonesift	rmy	Takes a list of accession numbers and compares list of OR/clones, pulling out ORs on particular clones.	List of accession numbers List of OR/clones (on same line).	
olfdba	rmy	Runs 'DBA' program on list of sequences.	List of genes (with sequences contained in files), and required match for 'DBA' program.	Produces a list.dba file. Match for DBA: MatchA 0.65 MatchB 0.75 MatchC 0.85 MatchD 0.95
olfdba2draw	rmy	Plots graphical representation of 'DBA' results.	1 ".dba" file, 1 ".rptm" file.	Produces postscript file as output.
olfgrab	rmy	Uses 'wombl' or 'pullout' to get a sequence, translates this in 6 frames and BLAST searches against the OR database. The highest hit(s) are then placed in a ".copro" file, created using the 'translator' program.	1 sequence file or accession number, start and stop positions.	Produces ".copro" file which can be edited and used in 'olfproducer' program.
olflocate	rmy	When given an OR accession number, prints where the number is in the genome (if known).	1 accession number	Location manually programmed in using information from Ensembl database.
or_nip	rmy	Looks in a fasta file for a sequence pattern specified in a PATTERNS/file.pat file.	1 fasta file, 1 pattern file (in correct directory)	Uses the 'NIP' program on the command line
olfproducer	rmy	From a ".copro" an OR protein and nucleotide file is produced.	1 ".copro" file	Prompts for name of OR and any additional comments.
PIPlocaldraw	rmy	Plots a graphical represenation of how 2 sequences are related to	1 edited PIP output file, position of start of region, position of end of region.	Produces a postscript file.

		each other, based on a PIP file.		
proteingrab	rmy	Creates a protein file of an accession number (if file found in protein databases)	1 EMBL accession number.	
pullout pulloutmany	rmy	Given a sequence file and start and stop positions pulls out the piece of sequence between these positions. Given a list start ands stop positions uses pullout to get required sequence	1 sequence file, start and stop positions. 1 sequence file, list file containing start and stop positions.	Produces a sequencefile_start_to_stop.fasta file.
rcin	rmy	Reverses a piece of sequence inputted on the command line	1 sequence on the command line	
reformatXLSclustal	rmy	Reformats to a clustalw fasta format file from an Excel format (“.tab”, generated by proalnXLS)	1 “.tab” file, created by proalnXLS.	
repembl	rmy	Calculates percentage type repeats from a EMBL file.	1 EMBL accession number	
relateolf	rmy	Takes OR and BLAST searches against the OR database, finding the top 5 hits, and the top hit on chromosome 6.	1 protein file	
reverseclone	rmy	Reverses “.genes”, “.misf” and “.rptm” files.	1 filename (“.genes”, “.misf” and “.rptm” files corresponding to this name). Length of clone	If the size of the clone is not entered the length is given as the size in EMBL.
rolfp	rmy	Taking a protein file as input, BLAST searches against the OR database.	1 protein file	Produces a “.pblast” output file
rmpars	rog	Calculates percentage type repeats from a RepeatMasker output file.	1 RepeatMasker output file.	
rptcalc	rmy	Takes a “.rptm” file and calculates the percentage of repeats belonging to each class.	1 “.rptm” file	
rptmgenerate	rmy	Takes a RepeatMasker out file and produces a correctly formatted “.rptm” file.	File from repeatmasker containing co-ordinates and type of repeat.	
seeclone	rmy	Produces “.genes”, “.rptm” and “.misf” files from an EMBL accession number, also produces “.ps” postscript file.	1 accession number	
SNPdefine	rmy	Takes gene file and list of SNPs and produces output	1 “.genes” file List of SNP position	Output file: list “.assn”, SNP ,

		file showing whether a SNP is found in an exon		blank if SNP does not appear in exon
translator	rmy	Takes a DNA sequence file and produces an output file showing protein translation above (forward) and below (reverse) the DNA.	(i) 1 DNA file (ii) Optional – R or r will only plot reverse strand and reverse translation, F or f will only plot forward strand and forward translation	
transfachits	rmy	Takes hits from transfac output file and converts to ".misf" format.	1 transfac output file	
waterall	rmy	Uses the Emboss program 'water' to compare one sequence against a number of others	(i) 1 sequence file (ii) file containing a list of sequences to be compared against.	
wombl	rog	Takes an accession number and 2 positions, and pulls out sequence corresponding to that between these positions.	1 accession number 2 positions found within the accession number.	Output file: "file_pos1_to_pos2.fasta"

Appendix 5: List of clones assembled to produce MHC extended class I region

Accession number	Size	Position of start in consensus sequence	Overlap with neighbouring clone	Orientation
U91328	246282	1	-	rev
AL353759	101099	198877	47405	for
AL031777	89301	299877	100	for
AL021917	170001	389078	100	for
AL050330	9015	558979	100	for
AL121936	122979	567894	100	for
AL513348	109476	690773	100	for
AL591044	133561	800248	2000	for
AL596216	432	931809	255	for
AL133255	152782	932058	100	for
AL590062	65119	1084585	100	for
AL021807	89016	1149604	100	for
AL121934	49261	1238520	100	for
AL021808	154066	1287681	102	for
AL031118	82456	1441647	101	for
AL021918	159506	1524002	100	for
AL031229	61450	1683407	100	for
AL009179	139904	1744757	100	for
AL049822	24706	1884561	100	for
Z98744	100375	1909167	101	for
AL133267	44788	2009442	100	for
AL121944	130279	2054130	100	for
AL358933	55480	2184309	101	for
AL022393	85654	2239689	100	rev
AL390721	13994	2325243	101	for
AL021997	97847	2339137	100	rev
AL358785	17464	2436884	100	for
Z98745	128779	2454248	100	for
AL049543	95594	2582927	100	rev
AL121932	85952	2678421	100	for
AL390196	43042	2764273	101	for
AL133258	73666	2807215	100	for
Z84474	107527	2880781	101	for
AL139329	3812	2988207	100	for
Z84476	112659	2991918	100	for
AL035402	47216	3104477	100	for
AL022727	144868	3151593	100	for
AL050339	68105	3296361	100	for
AL096770	97392	3364366	100	for
AL035542	114868	3461658	100	for
AL031983	134292	3576426	100	for
AL050328	54106	3710618	100	for
AL022723	148834	3764624	100	for

Appendix 6: List of human MHC-linked ORs

Name	Clone		Position in clone	State	Length	Additional information
hs6M1-1	AL022727	<	7626..8567	C	942	
hs6M1-2	AL022727	<	59199..60135	P	937	1 bp insertion at 430 > frameshift
hs6M1-3	AL022727	>	33215..34150	C	936	
hs6M1-4	AL022727	>	22275..23207	P	933	Substitution at 574,575 or 576 > stop codon
hs6M1-5	AL022727	>	102837..103772	P	936	Substitution at 565,566 or 567 > stop codon
hs6M1-6	AL022727	>	94963..95901	C	939	
hs6M1-7	AL022727	>	136590..137533	P	944	1 bp deletion at 95 > frameshift
hs6M1-8	Z84476	>	4055..4982	P	928	1 bp deletion at 810 > frameshift
hs6M1-9	Z98745	<	74312..74816	PF		No starting methionine Lacking 2 end motifs 1bp deletion at 263 > frameshift Substitution at 282,283 or 284 > stop codon
hs6M1-10	Z98744	<	74988..76061	C	1074	
hs6M1-12	AL031983	>	84339..85277	C	939	
hs6M1-13	AL031983			P		1 bp deletion at 114 > frameshift 1 bp insertion at 618 > frameshift 1 bp insertion at 679 > frameshift
hs6M1-14	AL031983			P	942	No starting methionine
hs6M1-15	AL035402	<	12652..13614	C	963	
hs6M1-16	AL035542			C	951	
hs6M1-17	AL035542	>	51151..52089	C	939	
hs6M1-18	AL035542	<	37829..38776	C	948	
hs6M1-19	AL035542	>	28417..29363	P	947	16 bp deletion at 552 > frameshift
hs6M1-20	AL035542	>	7847..8770	C	924	
hs6M1-21	AL096770	>	32625..33590	C	966	
hs6M1-22	AL096770 AL050339	< <	97167..97392 1..817	P	943	1 bp insertion at 82 bp > frameshift
hs6M1-23	AL050339	>	22125..23084	P	960	Substitution at 550,551 or 552 > stop codon Substitution at 553, 554 or 555 > stop codon
hs6M1-24	AL050339	> >	27467..28053 28552..28912	P	956	1 LTR19A insertion at 587 1 bp deletion at 871 > frameshift
hs6M1-25	AL050339	>	61300..62293	P	994	1 bp deletion at 244 > frameshift

						1 bp deletion at 599 > frameshift
hs6M1-26	AL035402	<	40196..40877	P	682	Missing 1st 2 motifs 1 bp insertion at 160 > frameshift Substitution at 290, 291 or 292 > stop codon Substitution at 305, 306 or 307 > stop codon
hs6M1-27	AL096770	>	13518..14468	C	951	
hs6M1-28	AL096770	<	81166..82197	C	1032	3 potential starting methionines
hs6M1-29	AL121944	<	72007..72944	P	938	1 bp deletion at 76 > frameshift
hs6M1-30	AL121944	>	52652..53634	P	943	1 bp insertion at 28 > frameshift 1 bp deletion at 316 > frameshift 1 bp insertion at 448 > frameshift Substitution at 782,783 or 784 > stop codon
hs6M1-31	AL121944	>	65213..66146	P	982	1 bp insertion at 322 > frameshift
hs6M1-32	AL133267	>	20708..21649	C	942	
hs6M1-33	AL133267	>	920..1870	P	951	Substitution at 1 > Valine not methionine start
hs6M1-34	AL133267	>	40617..41542	P	926	1 bp deletion at 539 > frameshift
hs6M1-35	AL121944	>	92095..93108	C	1014	

Appendix 7: Alignment of human MHC-linked ORs

```
hs6M1-17      1 ----------------------MSANTSMVTEFLLLGFSHLADLQG-LLFSVFLTIYLLT
hs6M1-21      1 ---------------------MERKNQTAITEFIILGFSNLNELQF-LLFTIFFLTYFCT
hs6M1-28      1 -MHFLPTVFGFLNRVTLGIFRETMVNLTSMSGFLLMGFSDERKLQI-LHALVFLVTYLLA
hs6M1-35      1 ---------------------MEGKNQTNISEFLLLGFSSWQQQQV-LLFALFLCLYLTG
hs6M1-19P     1 -----------------------MLNTTSVTEFLLLGVTDIQELQP-FLFVVFLTIYFIS
hs6M1-20      1 -----------------------MLNTTSVTEFLLLGVTDIQELQP-FLFVVFLTIYFIS
hs6M1-27      1 -----------------------MENVTTMNEFLLLGLTGVQELQP-FFFGIFLIIYLIN
hs6M1-18      1 -------------------MEIVSTGNETITEFVLLGFYDIPELHF-LFFIVFTAVYVFI
hs6M1-12      1 -----------------------MVNQSSTPGFLLLGFSEHPGLERTLFVVVF-TSYLLT
hs6M1-13P     1 -----------------------MVNQSSAPGFLLLGFSEHPALERTLFVVVF-TSYLLT
hs6M1-16      1 -----------------------MVNQSSPMGFLLLGFSEHPALERTLFVVVF-TSYLLT
hs6M1-7P      1 ---------------------MIIICNDSHSDFILLGFSNKPHLEKILFGSFLXIFYFLT
hs6M1-14P     1 ------------------------ANYSAEERFLLLGFSDWPSLQPVLFALVL-LCYLLT
hs6M1-25P     1 MANTLSSLNSCNVFLLVLNRVMGMTNSSVKGDFILVGFSHQPHLEKILFVAVL-ISYLLT
hs6M1-10      1 ---------------------MNWVNKSVPQEFILLVFSDQPWLEI-PPFVMFLFSYILT
hs6M1-32      1 ---------------------MNWVNDSIIQEFILLGFSDRPWLEF-PLLVVFLISYTVT
hs6M1-1       1 ---------------------MNWENESSPKEFILLGFSDRAWLQM-PLFVVLLISYTIT
hs6M1-22P     1 ---------------------MWINNQSSLDDFILLGFSDRPWLET-PLXVIFLVAYIFS
hs6M1-2P      1 ---------------------MPLTNESHPEEFILLGFADRPWLEL-PLFTSLLIMYPIA
hs6M1-9P      1 --------------------SVKYLNESFPEDFILMGFVKYPWLDF-LLFCVLLTFYMFT
hs6M1-29P     1 ---------------------MDQKNGSSFTGFILLGFSDRPQLELXSPLCGFLIFYIFT
hs6M1-31P     1 ---------------------MERANDSTFSGFILLGFSNRPQLET-ALFVVILIIYFLS
hs6M1-3       1 ------------------MNDDGKVNASSEGYFILVGFSNWPHLEV-VIFVVVLIFYLMT
hs6M1-5       1 ------------------MNDDGKVNASSEGYFILVGFSNWPYLEV-VLFVVILIFCLMT
hs6M1-4P      1 ----------------------MKKNASFEDFFILLGFSNWPHLEV-VLFVVILIFYLIT
hs6M1-6       1 --------------------MMIKKNASSEDFFILLGFSNWPQLEV-VLFVVILIFYLMT
hs6M1-33P     1 -----------------VAAGVENDNTSSFEGFILVGFSDRPHLEL-IVFVVVLIFYLLT
hs6M1-34P     1 ---------------------MEKSNVSSVYGFILVGFSDRPKLEM-VLFTVNFILYSVA
hs6M1-30P     1 ---------------------MTNQSCPETX-FILLGFSGRPRLEH-VLFVFVLIFYLVT
hs6M1-15      1 ---------------------MDQSNYSSLHGFILLGFSNHPKMEM-ILSGVVAIFYLIT
hs6M1-8P      1 ---------------------METSSVSSGTDFILLGFSDRPQLEH-IISVVVFIIYIVT
hs6M1-23P     1 -----------------------MINDSHFSGFILLGFTGQPQLQM-MISGVVFFFYTIA
hs6M1-24P     1 -----------------------MINDSYFGWLMLLGFPGKPQLEM-IISGVVFFFYAIS


hs6M1-17     38 VAGNFLIVVLVSTDAALQSPMY-FFLRTLSALEIGYTSVTVPLLLHHLLTGRRHISRSGC
hs6M1-21     39 LGGNILIILTTVTDPHLHTPMY-YFLGNLAFIDICYTTSNVPQMMVHLLSKKKSISYVGC
hs6M1-28     59 LTGNLLIITIITVDRRLHSPMY-YFLKHLSLLDLCFISVTVPQSIANSLMGNGYISLVQC
hs6M1-35     39 LFGNLLILLAIGSDHCLHTPMY-FFLANLSLVDLCLPSATVPKMLLNIQTQTQTISYPGC
hs6M1-19P    37 VAGNGAILMIVISDPRLHSPMY-FFLGNLSCLDICYSSVTLPKMLQNFLSAHKAISFLGC
hs6M1-20     37 VTGNGAVLMIVISDPRLHSLMY-FFLGNLSYLDICYSTVTLPKMLQNFLSTHKAISFLGC
hs6M1-27     37 LIGNGSILVMVVLEPQLHSPMY-FFLGNLSCLDISYSSVTLPKLLVNLVCSRRAISFLGC
hs6M1-18     41 IIGNMLIIVAVVSSQRLHKPMY-IFLANLSFLDILYTSAVMPKMLEGFLQEAT-ISVAGC
hs6M1-12     37 LVGNTLIILLSALDPKLHSPMY-FFLSNLSFLDLCFTTSCVPQMLVNLWGPKKTISFLDC
hs6M1-13P    37 PVX-GLIILLSVLDPRLHSPMY-FFLSNLSFLDLCFTISCVPGMLVNLWEPKKTIILLGC
hs6M1-16     37 LVGNTLIILLSVLYPRLHSPMY-FFLSDLSFLDLCFTTSCVPQMLVNLWGPKKTISFLGC
hs6M1-7P     40 LAGNMVIVLVSLKDPKLHIPMY-FFLSNLSLVDLCLTSSCVPQMLINFWGPEKTISYIGC
hs6M1-14P    36 LTGNSALVLLAVRDPRLHTPMY-YFLCHLALVDAGFTTSVVPPLLANLRGPALWLPRSHC
hs6M1-25P    60 LVGNTVIILICSVDPKLKTPMYXFFLTHLSLVDICFTTSIVPQLLWNLKGPDKTITFLGC
hs6M1-10     39 IFGNLTIILVSHVDFKLHTPMY-FFLSNLSLLDLCYTTSTVPQMLVNICNTRKVISYGGC
hs6M1-32     39 IFGNLTIILVSRLDTKLHTPMY-FFLTNLSLLDLCYTTCTVPQMLVNLCSIRKVISYRGC
hs6M1-1      39 IFGNVSIMMVCILDPKLHTPMY-FFLTNLSILDLCYTTTTVPHMLVNIGCNKKTISYAGC
hs6M1-22P    39 LFGNISIILVSHLDPQLDSPMY-FFVSNLSFLDLCYTTSTVPQMLVNLRGPEKTISYGGC
hs6M1-2P     39 VMGNITIILMSRLDSRLHSPMY-FFLTNLSFLDMCYTTSIVPQMLFNLGSSKKTISYMGC
hs6M1-9P     40 LLGNSAIILVSQLDSQLHSPMY-FLLTSLSVLYLCFTTTTVPQMLFNLGGPXKNITIG-C
hs6M1-29P    40 LLGNKTIIVLSHLDPHLHNPMY-FFFSNLSFLDLCYTTGIVPQLLVNLRGADKSISYGGC
hs6M1-31P    39 FLGNGTIILLSIVDPRLHTPMY-FFLSNLSFMDLCLTTCTVPQTLVNFKGKDKTITYGGC
hs6M1-3      42 LIGNLFIIILSYLDSHLHTPMY-FFLSNLSFLDLCYTTSSIPQLLVNLWGPEKTISYAGC
hs6M1-5      42 LIGNLFIIILTYLDSHLHTPLY-FFLSNLSFLDLCYTTSSIPQLLVSLWGVEKTISYAGC
hs6M1-4P     38 LIGNLFIIILSYLDSHLHTPMY-FFLSNLSFLDLCYTTSSIPQLLVNLWGPEKTISYAGC
hs6M1-6      40 LTGNLFIIILSYVDSHLHTPMY-FFLSNLSFLDLCHTTSSIPQLLVNLRGPEKTISYAGC
hs6M1-33P    43 LLGNMTIVLLSALDSRLHTPMY-FFLANLSFLDMCFTTGSIPQMLYNLWGPDKTISYVGC
hs6M1-34P    39 VLGNSTIILVCILDSQLHTPMY-FFLANLSFLDLCFSTSCIPQMLVNLWGPDKTISCAGC
hs6M1-30P    38 LVGNIIIILISHLDPCLHMPMY-FFLTNLSFLDLCFTTSSIPQLLFNLGSPGKTISHTGC
hs6M1-15     39 LVGNTAIILASLLDSQLHTPMY-FFLRNLSFLDLCFTTSIIPQMLVNLWGPDKTISYVGC
hs6M1-8P     39 LVGNTTIILVSYLDTQLHTFMY-FFLSNLSFLDLCYTTSIIPQMLANQWGPKKSITYGGC
hs6M1-23P    37 FMGNMAIILLSFLDDHLQVPMY-FFLRNLAILDLCYTTNIVPQMLVSIWGKDKRITFGGC
hs6M1-24P    37 LMGNMVLILLPLLDKHLQTPIY-FFLRNLAILDLCYTTNIVPQMLVNAWGKDKKITFGGC
```

```
hs6M1-17     97 ALQMFFFLFFG-ATECCLLAAMAYDRYAAICEPLRYPLLLSHRVCLQLAGSAWAC-GVLV
hs6M1-21     98 VVQLFAFVFFV-GSECLLLAAMAYDRYIAICNPLRYSVILSKVLCNQLAASCWAA-GFLN
hs6M1-28    118 ILQVFFFIALA-SSEVAILTVMSYDRYAAICQPLHYETIMDPRACRHAVIAVWIA-GGLS
hs6M1-35     98 LAQMYFCMMFA-NMDNFLLTVMAYDRYVAICHPLHYSTIMALRLCASLVAAPWVI-AILN
hs6M1-19P    96 ISQLHFFHFLG-STEAMLLAVMAFDRFVAICKPLRYTVIMNPQLCTQMAITIWMI-GFFH
hs6M1-20     96 ISQLHFFHFLG-STESMLFAVMAFDLSVAICKPLRYTVIMNPQLCTQMAITIWVI-GFFH
hs6M1-27     96 ITQLHFFHFLG-STEAILLAIMAFDRFVAICNPLRYTVIMNPQVCILLAAAAWLI-SFFY
hs6M1-18     99 LLQFFIFGSLA-TAECLLLAVMAYDRYLAICYPLHYPLLMGPRRYMGLVVTTWLS-GFVV
hs6M1-12     96 SVQIFIFLSLG-TTECILLTVMAFDRYVAVCQPLHYATIIHPRLCWQLASVAWVI-GLVE
hs6M1-13P    95 SVQFFIFLSLG-TTECILLTVMAFDRYMAIFKPLRHATIVHLCLCWQLASVAWVI-GLVE
hs6M1-16     96 SVQLFIFLSLG-TTECILLTVMAFDRYVAVCQPLHYATIIHPRLCWQLASVAWVM-SLVQ
hs6M1-7P     99 AIQLYVFLWLG-ATEYVLLVVMAVDCYVAVCHPLQNTMIMHPKLCLQLAILAWGT-GLAQ
hs6M1-14P    95 TAQLCASLALG-SAECVLLAVMALDRAAAVCRPLRYAGLVSPRLCRTLASASWLS-GLTN
hs6M1-25P   120 VIQLYISLALG-STECVLLAVMAFDRYAAVCKPLHYTAVMNPQLCQALAGVAWLS-GVGN
hs6M1-10     98 VAQLFIFLAL-GSTECLLLAVMCFDRFVAICRPLHYSIIMHQRLCFQLAAASWIS-GFSN
hs6M1-32     98 VAQLFIFLAL-GATEYLLLAVMSFDRFVAICRPLHYSVIMHQRLCLQLAAASWVT-GFSN
hs6M1-1      98 VAHLIIFLAL-GATECLLLAVMSFDRYVAVCRPLHYVVIMNYWFCLRMAAFSWLI-GFGN
hs6M1-22P    98 VAQLYIFLAL-GSTECILLAIMAFDRYAAICKPLHYPVIMNHRRCIHMAAGTWIS-GFAN
hs6M1-2P     98 AVQLYFFHIM-GGTECLLLAIMSFDRYVAICRPLHYTLIMNQRVCILXVSTVWLI-GIIY
hs6M1-9P     98 MAQAYVFHWL-ACIECVLLGIVALDCYVAVCKPPRYTIIIDHKVCLHLSSTAWLI-GLAN
hs6M1-29P    99 VVQLYISLGL-GSTECVLLGVMAFDRYAAVCRPLHYTVVMHPCLYVLMASTSWVI-GFAN
hs6M1-31P    98 VTQLFIALGLXGGVECVLLSAMAYDRYAAVCRPLHYMVSMHPQLCLQLVVTTWLT-GFGN
hs6M1-3     101 MIQLYFVLAL-GTTECVLLVVMSYDRYAAVCRPLHYTVLMHPRFCHLLAVASWVS-GFTN
hs6M1-5     101 MVQLYFFLTL-GTTECVLLVVMSYDRYAAVCRPLHYTVLMHSRFCHLLAVASWVS-GFTN
hs6M1-4P     97 TVQLYFVLAL-GTAECVLLVVMSYDRYAAVCRPLHYTVLMHPRFCRLLAAASWVS-GFTT
hs6M1-6      99 MVQLYFVLAL-GIAECVLLVVMSYDRYVAVCRPLHYTVLMHPRFCHLLAAASWVI-GFTI
hs6M1-33P   102 AIQLYFVLALG-GVECVLLAVMAYDRYAAVCKPLHYTIIMHPRLCGQLASVAWLS-GFGN
hs6M1-34P    98 VVQLFSFLSVR-GIECILLAVMAYDSYAAVCKPLRYLVIMHLQLCLGLMAAAWGS-GLVN
hs6M1-30P    97 AIQLFMFLGLXGWQECILLAAVAYDRFIAICKPLHYSVIMHPQLCWKLVSVARGCXGLLS
hs6M1-15     98 IIQLYVYMWLG-SVECLLLAVMSYDRFTAICKPLHYFVVMNPHLCLKMIIMIWSI-SLAN
hs6M1-8P     98 VLQFFFVLDLG-ATECLLLAVMAYDRYAAVCQPLHY-----TLKCTLSFATAWLS-GLAS
hs6M1-23P    96 AFQLFIDVALY-SVECILLSMMSYDRLNAICKPLHHMTIMNLQLCQGLVVISWVV-GVIN
hs6M1-24P    96 AFQLFTNVTLC-TVECMLLAVMSYDPFNAVCKPLDYMTIMNPQLCQGLVAMTWLI-GVTN
hs6M1-26P     1 -----------GSTKCIILAVTSLDPYIAICKHLRYPAIMHQQLCVLLVAMAWLS-SLAN

hs6M1-17    155 GLGHTPFIFSLPFCGPNTIPQFFCEIQPVLQLVCG----DTSLN-ELQIILATALLILCP
hs6M1-21    156 SVVHTVLTFCLPFCGNNQINYFFCDIPPLLILSCG----NTSVN-ELALLSTGVFIGWTP
hs6M1-28    176 GLMHAAINFSIPLCGKRVIHQFFCDVPQMLKLACS----YEFIN-EIALAAFTTSAAFIC
hs6M1-35    156 PLLHTLMMAHLHFCSDNVIHHFFCDINSLLPLSCS----DTSLN-QLSVLATVGLIFVVP
hs6M1-19P   154 ALLHSLMTSRLNFCGSNRIYHFFCDVKPLLKLSX--------LN-QWLLSTVTGTIAMGP
hs6M1-20    154 ALLHSVMTSRLNFCGSNRIHHFLCDIKPLLKLACG----NTELN-QWLLSTVTGTIAMGP
hs6M1-27    154 ALMHSVMTAHLSFCGSQKLNHFFYDVKPLLELACS----DTLLN-QWLLSIVTGSISMGA
hs6M1-18    157 DGLVVALVAQLRFCGPNHIDQFYCDFMLFVGLACS----DPRVA-QVTTLILSVFCLTIP
hs6M1-12    154 SVVQTPSTLHLPFCPDRQVDDFV-CEVPALIRLS---CEDTSYN-EIQVAVASVFILVVP
hs6M1-13P   153 SVVQTPSTLRLPFCPHQQVDDFV-CEVPALIRLS---CEDTSYN-EIQMAVASVFILAVP
hs6M1-16    154 SIVQTPSTLHLPFCPHQQIDDFL-CEVPSLIRLS---CGDTSYN-EIQLAVSSVIFVVVP
hs6M1-7P    157 SLIQSPATLRLPFCSQRMVDDVV-CEVPALIQLS---STDTTYS-EIQMSIASVVLLVMP
hs6M1-14P   153 SVAQTALLAERPLCAPRLLDHFI-CELPALLKLA---CGGDGDTTENQMFAARVVILLLP
hs6M1-25P   178 TLIQALSPSGFLAVDTDCSNISSXREVPSMIKLA---CVDIHDN-EVQLFVASLVLLLLP
hs6M1-10    156 SVLQSTWTLKMPLCGHKEVDHFFCE-VPALLKLS---CVDTTAN-EAELFFISVLFLLIP
hs6M1-32    156 SVWLSTLTLQLPLCDPYVIDHFLCE-VPALLKLS---CVETTAN-EAELFLVSELFHLIP
hs6M1-1     156 SVLQSSLTLNMPRCGHQEVDHFFCE-VPALLKLS---CADTKPI-EAELFFFSVLILLIP
hs6M1-22P   156 SLVQSTLTVVAPRCGQRVLDHFFCE-VPALLKLA---CIDIRVN-EMELNVLGALLLLMP
hs6M1-2P    156 AVSEATATLQLPLCGLNKLDHLVCE-IPVLIKIA---CGEKGSN-ELTLSVVCIFMLAVP
hs6M1-9P    156 SLLQSTITIQLPL-----------------------------------------------
hs6M1-29P   157 SLLQTVLILLLTLCGRNKLEHFLCE-VPPLLKLA---CVDTTMN-ESELFFVSVIILLVP
hs6M1-31P   157 SVIQTALTMTLPLCDKNQVDHFFCE-VPVMLKLS---CTNTSIN-EAEIFAVSVFFLVVP
hs6M1-3     159 SALHSSFTFWVPLCGHRQVDHFFCE-VPALLRLS---CVDTHVN-ELTLMITSSIFVLIP
hs6M1-5     159 PALHSSFTFWVPLCGHRQIDHFFCE-VPALLLS----FVNTREN-KLTLMITSSIFVLLL
hs6M1-4P    155 SALHSSFTFWIPLCRHRLVDHFFCE-APALLRLS---CVDTAN--ELTLMVMSSIFVLIP
hs6M1-6     157 SALHSSFTFWVPLCGHRLVDHFFCE-VPALLRLS---CVDTHAN-ELTLMVMSSIFVLIP
hs6M1-33P   160 SLIMAPQTLMLPRCGHRRVDHFLCE-MPALIGMA---CVDTMML-EALAFALAIFIILAP
hs6M1-34P   156 AVVMSPLTMTLSRSGRRRVNHFLCEXKPALIKMA---CLDVRAV-EMLAFAFAVLIVLLP
hs6M1-30P   157 SLVMSPVTMKLPRCGRCKLKHFLCE-MPALIKIT---CVDTVAM-ESTVFTLSVVIVLMP
hs6M1-15    156 SVVLCTLTLNLPTCGNNILDHFLCE-LPALVKIA---CVDTTTV-EMSVFALGIIIVLTP
hs6M1-8P    151 ALIVCSLTLKLPRCGHREVDNFFCE-MPALIKMA---CVYSKVI-EIVVFAFGVVFLFVP
hs6M1-23P   154 CIIPSPYATSLPRCRNHHLDHFFVCVKCLQSRFK--IACVDTTAMEVTTFAMCLIIVLVP
hs6M1-24P   154 CMILSPCPVSLPRCGDHHLDHYFCEISAMVKIACGATTVMEETKPYLHCVVVVVFIFLAS
hs6M1-26P    49 STSVIXLAVQLPLGG-NKVDDFLCEVSAMIKISR----FDTTFN--V-SMLSIVRIFSLV
```

```
hs6M1-17    210 FGLILGSYGRILVTIFRIP-SVAGRRKAFSTCSS-HLIVVSLFYGTALFIYIRPKASYDP
hs6M1-21    211 FLCIVLSYICIISTILRIQ-SSEGRRKAFSTCAS-HLAIVFLFYGSAIFTYVRPISTYSL
hs6M1-28    231 LISIVLSYIRIFSTVLRIP-SAEGRTKVFSTCLP-HLFVATFFLSAAGFEFLRLPSDSSS
hs6M1-35    211 SVCILVSYILIVSAVMKVP-SAQGKLKAFSTCGS-HLALVILFYGAITGVYMSPLSNHST
hs6M1-19P   205 FFLTLLSYFYIITHLFFKTHSFSMLRKALSTCAS-HFMVVILLYAPVLFTYIHHASGTSM
hs6M1-20    209 FFLTLLSYFYIITYLFFKTRSCSMLCKALSTCAS-HFMVVILFYAPVLFTYIHPALESFM
hs6M1-27    209 FFLTLLSCFYVIGFLLFKNRSCRILHKALSTCAS-HFMVVCLFYGPVGFTYIRPASATSM
hs6M1-18    212 FGLILTSYARIVVAVLRVP-AGASRRRAFSTCSS-HLAVVTTFYGTLMIFYVAPSAVHSQ
hs6M1-12    209 LSLILVSYGAITWAVLRIN-SAKGRRKAFGTCSS-HLTVVTLFYSSVIAVYLQPKNPYAQ
hs6M1-13P   208 XSLILVSYGAIAWAVLRTN-CKXGQRKAFGTCSS-HLTVVTLFYSSVIAVYLQPKNPYAQ
hs6M1-16    209 LSLILASYGATAQAVLRIN-SATAWRKAFGTCSS-HLTVVTLFYSSVIAVYLQPKNPYAQ
hs6M1-7P    212 LIIILSSSGAIAKAVLRIK-STAGQKKAFGTCIS-HLLVVSLFYGTVTGVYLQPKNHYPH
hs6M1-14P   209 FAVILASYGAVARAVCCMR-FSGGRRRAVGTCGS-HLTAVCLFYGSAIYTYLQPAQRYNQ
hs6M1-25P   234 LVLILLSYGHIAKVVIRIK-SVQAWCKGLGTCGS-HLIVVSLFCGTITAVYIQSNSSYAH
hs6M1-10    211 VTLILISYAFIVQAVLRIQ-SAEGQRKAFGTCGS-HLIVVSLFYGTAISMYLQPPSPSSK
hs6M1-32    211 LTLILISYAFIVRAVLRIQ-SAEGRQKAFGTCGS-HLIVVSLFYSTAVSVYLQPPSPSSK
hs6M1-1     211 VTLILISYGFIAQAVLKIR-SAEGRQKAFGTCGS-HMIVVSLFYGTAIYMYLQPPSSTSK
hs6M1-22P   211 LTLILGTYVFIAQAVMRIC-SAESRWKAFNTCAS-HLLVVSLFYFTAISMYVQPPSSYSH
hs6M1-2P    211 LCLILASYASIGSAVFKIK-SSKGRKKAFGTCSS-HLIVVFLFYGPAISMYLQPPSSISR
hs6M1-29P   212 VALIIFSYSQIVRAVVRIK-SATGQRKVFGTCGSPHLTVVSLFYGTAIYAYLQPGNNYSQ
hs6M1-31P   212 LSLILASYGHITHAVLKIK-SAQGRQKAFGTCGS-HLLVVIIFFGTLISMYLQPPSSYSQ
hs6M1-3     214 LILILTSYGAIVRAVLRMQ-STTGLQKVFGTCGA-HLMAVSLFFIPAMCIYLQPPSGNSQ
hs6M1-5     213 LTLIFTSYGAIAQAVLRMQ-STTGLQKVFGTCGA-HHMVVSLFFIPAMCMYLQPPSGNSQ
hs6M1-4P    209 LILILTSYGAIARAVLSMQ-STTGLQKVLRTCGA-HLMVVSLFFIPVMCMYLQPPSENSQ
hs6M1-6     212 LILILTAYGAIARAVLSMQ-STTGLQKVFRTCGA-HLMVVSLFFIPVMCMYLQPPSENSP
hs6M1-33P   215 LILILISYGYVGGTVLRIK-SAAGRKKAFNTCSS-HLIVVSLFYGTIIYMYLQPANTYSQ
hs6M1-34P   212 LTLILVSYGYIAAAVLSIK-SAARQWKAFHTCSS-HLTVVSLFYGSIIYMYMQPGNSSSQ
hs6M1-30P   212 LCLILISYSYIALAVLRIK-SAAGRRKAFNMCGS-HLTVVSLFYGNIIYMYMQPNNS-SQ
hs6M1-15    211 LILILISYGYIAKAVLRTK-SKASQRKAMNTCGS-HLTVVSMFYGTIIYMYLQPGNRASK
hs6M1-8P    206 LSLILISYGVITQAVMRIK-SATRLQKILNTCGS-HLTVVILFYGTIIYIYMKPQNTISQ
hs6M1-23P   212 LLLILVSYGFIAVAVLKIK-SAAGRQKAFGTCSS-HLVVVSIFCGTVTYMYIQPGNSPNQ
hs6M1-24P   214 LLLILVSYGFIAVAVLKIK-SAAGRQKAFGTCFS-HLIVVSIFYGTVRYMYIEPGNSPSQ
hs6M1-26P   101 LSIIFAYCGFIVATVLRIQ-SSGGKKEVFNTCGS---HIVSLLYGPVISMYVQPSAN-SQ


hs6M1-17    268 ATDPLVS-LFYAVVTPILNPIIYSLRNTEVKAALKRTIQKTVPMEI--------------
hs6M1-21    269 KKDRLVS-VLYSVVTPMLNPIIYTLRNKDIKEAVKTIGSKWQPPISSLDSKLTY------
hs6M1-28    289 TVDLVFS-VFYTVIPPTLNPVIYSLRNDSMKAALRKMLSKEELPQRKMCLKAMFKL----
hs6M1-35    269 EKDSAAS-VIFMVVAPVLNPFIYSLRNNELKGTLKKTLSRPGAVAHACNPSTLGGRGGWI
hs6M1-19P   264 DQDRITA-IMYTVVTPVLNPLIYTLRNKEVKGAFNRAMKRWLWPKEILKNSSEA------
hs6M1-20    268 DQDRIVA-IMYTVVTPVLNPLIYTLRNKEVKGALGRVIRRL-------------------
hs6M1-27    268 IQDRIMA-IMYSAVTPVLNPLIYTLRNKEVMMALKKIFGRKLFKDWQQHH----------
hs6M1-18    270 LLSKVFS-LLYTVVTPLFNPVIYTMRNKEVHQALRKILCIKQTETLDRRX----------
hs6M1-12    267 ERGKFFG-LFYAVGTPSLNPLIYTLRNKEVTRAFRRLLGKEMGLTQS-------------
hs6M1-13P   266 ERGKFFG-LFYAVGTPSLNPLIYTLRNKEVTRAFRRLLAKEMGLIQS-------------
hs6M1-16    267 GRGKFFG-LFYAVGTPSLNPLVYTLRNKEIKRALRRLLGKERDSRESWRAA---------
hs6M1-7P    270 EWGKFLT-LFYTVVTPTLNPLIYTLRNKEVKGALIRLGRRTWDSQNN-------------
hs6M1-14P   267 ARGKFVS-LFYTVVTPALNPLIYTLRNKKVKGAARRLLRSLGRGQAGQ------------
hs6M1-25P   292 AHGKFIS-LFYTVVTPTLNPLIYTLRNNDVKGALRLFNRDLGT-----------------
hs6M1-10    269 DRGKMVS-LFCGIIAPMLNPLIYTLRNKEVKEAFKRLVAKSLLNQEIRNMQMISFAKDTV
hs6M1-32    269 DQGKMVS-LFYGIIAPMLNPLIYTLRNKEVKEGFKRLVARVFLIKK--------------
hs6M1-1     269 DWGKMVS-LFYGIITSMLNSLIYSLRNKDMKEAFKRLMPRIFFCKK--------------
hs6M1-22P   269 DRGKIMA-LFYGIVTPTLNPFIYTLRNKDVKAALRRSLTKEFWIKTR-------------
hs6M1-2P    269 DQPKFMA-LFYGVVTPSLNPFIYTLRNKNVKGALRNLVRSIFSFK---------------
hs6M1-29P   271 DQGKVIS-LFYTIITPMINPLIYTLRNKDVKGALKKVLWKNYDSR---------------
hs6M1-31P   270 DVNKSIA-LFYTLVTPLLNPLIYTLRNKEVKGATKKTSGEDHRCMRKLTQGLQFQTFVH-
hs6M1-3     272 DQGKFIA-LFYTVVTPSLNPLIYTLRNKVVRGAVKRLMGWE-------------------
hs6M1-5     271 DQGKFIA-LFYTVVTPSLNPLIYTLRNKDVRGVVKRLRGWE-------------------
hs6M1-4P    267 DQGKFIA-LFYTVVTPSLNPLIYTFRNKDVRGAVKRLMGWEWGM----------------
hs6M1-6     270 DQGKFIA-LFYTVVTPSLNPLIYTLRNKHVKGAAKRLLGWEWGK----------------
hs6M1-33P   273 DQGKFLT-LFYTIVTPSVNPLIYTLRNKDVKEAMKKVLGKGSAEI---------------
hs6M1-34P   270 DQGKFLT-LFYNLVTPMLNLLIYTLRNKEVKGALKKVLGRQ-------------------
hs6M1-30P   269 DQGKFLT-LFYNLMTPMLNPVIYTLRNKDVKGALKRLVSRKHSDSDCS------------
hs6M1-15    269 DQGKFLT-LFYTVITPSLNPLIYTLRNKDMKDALKKLMRFHHKSTKIKRNCKS-------
hs6M1-8P    264 DEGKFFTXLFYTIITPSLNLPIYTLRNKDVKSALKRILWMKKSSAES-------------
hs6M1-23P   270 NEGKLLS-IFYSIVTPSLNPLIYTVRNKEFKGAMKRLTGKEKDCMEKRGH----------
hs6M1-24P   272 DEGKLLH-IFYSIVTPTLNPXIP-LRNKEFKWAMKRLIGKEKGSGDTI------------
hs6M1-26P   156 DKNKFMS-LFYSLVTPMLNPFIYTLSNRDIKGAMRRLLVFLYHQEENKSNYFYTPHSSYT


hs6M1-35     328 MRSGDRDHPG-------------------- 337
hs6M1-10     328 LTYLTNFSASCPIFVITIENYCNLPQRKFP 357
hs6M1-26P    215 GQKISCSKITC------------------- 225
```

Appendix 8: List of mouse MHC-linked ORs

Name	Clone		Position in clone	State	Length	Comment
mm17M1-1	AL078630	<	65954..66892	C	939	
mm17M1-2	AL078630	<	105931..106869	C	939	
mm17M1-3	AL078630	<	40162..41094	C	933	
mm17M1-4	AL078630	<	47848..48786	C	939	
mm17M1-5	AL078630	<	124623..125561	P	939	Substitution at 292, 293 or 294 > stop codon
mm17M1-6	AL078630	<	150388..151317	C	1161	
mm17M1-7	AL133159	>	84417..85262	P	930	Substitution at 1 > Valine not methionine start
mm17M1-8	AL133159	>	63932..64765	P	930	Substitution at 1 > Valine not methionine start
mm17M1-9	AL133159	>	96478..97311	P	930	Substitution at 1 > Valine not methionine start
mm17M1-10	AL133159	>	107383..108312	C	930	
mm17M1-11	AL133159	>	89150..89942	C	963	
mm17M1-12	AL133159	<	1306..2232	C	927	
mm17M1-13	AL133159	<	15182..16108	C	927	
mm17M1-14	AL133159	<	38006..38947	C	942	
mm17M1-15	AL133159	<	109667..110311	P	595	Missing 1st 2 motifs Substitution at 55, 56, 57> stop codon Substitution at 109, 110, 111 > stop codon 1 bp insertion at 172 > frameshift
mm17M1-16	AL133159	<	148446..148910	P	523	Missing 1st 2 motifs Substitution at 13, 14, 15 > stop codon Substitution at 172, 173, 174 > stop codon Substitution at 205, 206, 207 > stop codon 1 bp insertion at 433 > frameshift
mm17M1-17	AL133159	<	154600..155102	P	523	As above
mm17M1-18	AL136158	<	111592..112530	C	1014	
mm17M1-19	AL136158	<	91202..92140	C	1014	
mm17M1-20	AL136158	<	76271..77185	C	1014	
mm17M1-21	AL136158	<	159485..160438	C	954	
mm17M1-22	AL136158	>	171735..172700	C	1002	
mm17M1-23	AL136158	>	27408..28361	C	954	
mm17M1-24	AL136158	<	60863..61900	C	1038	
mm17M1-25	AL359381	<	80755..81681	C	927	
mm17M1-26	AL359381	<	8047..8979	C	933	
mm17M1-27	AL359381	<	39327..40256	C	930	
mm17M1-28	AL359381	<	56079..56996	C	918	
mm17M1-29	AL359381	>	1739..2680	C	942	

mm17M1-30	AL359381	>	25985..27445	P	942	Substitution at 25, 26, 27 > stop codon LTR Insertion at 262 Substitution at 391, 392, 393 > stop codon Substitution at 577, 578, 579 > stop codon 1 bp insertion at 796 > frameshift
mm17M1-31	AL359381	>	53234..54676	P	917	Substitution at 25, 26, 27 > stop codon LTR Insertion at 262 Substitution at 391, 392, 393 > stop codon Substitution at 571, 572, 573 > stop codon 1 bp deletion at 630 > frameshift
mm17M1-32	AL136158	<	125473..126438	C	966	
mm17M1-33	AL450393	>	35674..36816	C	1143	
mm17M1-34	AL450393	>	3434..4378	C	945	
mm17M1-35	AL365336	>	11562..12527	C	966	
mm17M1-36	AL365336	>	36956..37921	C	966	
mm17M1-37	AL359352	>	13527..14492	C	966	
mm17M1-38	AL359352	>	33373..34338	C	966	
mm17M1-39	AL359352	>	73703..74650	C	948	
mm17M1-40	AL136158	>	89150..89942	P	793	1 bp deletion at 153 > frameshift Substitution at 273, 274, 275 > stop codon Substitution at 435, 436, 437 > stop codon 1 bp deletion at 461 > frameshift 1 bp insertion at 581 > frameshift 1 bp deletion at 722 > frameshift
mm17M1-41	AL359352	>	89991..90953	C	963	
mm17M1-42	AL359352	>	105573..106532	C	960	
mm17M1-43	AL359352	>	153054..154025	C	972	
mm17M1-44	AL359352	<	40711..41542	P	832	Substitution at 124, 125, 126 > stop codon Substitution at 133, 134, 135 > stop codon 1 bp deletion at 204 > frameshift 1 bp insertion at 270 > frameshift Substitution at 298, 299, 300 > stop codon Substitution at 349, 350, 351 > stop codon Substitution at 403, 404, 405 > stop codon Substitution at 421, 422, 423 > stop codon Substitution at 514, 515, 516 > stop codon 1 bp deletion at 730 > FS
mm17M1-45	AL590433	>	64004..64933	C	930	
mm17M1-46	AL590433	<	11889..12827	C	939	

Appendix 9: Alignment of mouse MHC-linked ORs

```
mm17M1-6      1 MWLCNKTKTWACAEFIPYWRFLFVVVSGKTGFYYVALAGLELTEISGLCLPAQGPQHCLA
mm17M1-33     1 -----------MRVCTYILRDRYGLCFAKKCFICGSVSNVVGGDIIGTSTCFLVLQTIIN
mm17M1-22     1 --------------------------------------------------------MDLM
mm17M1-19     1 -------------------------------------------------MPQFLSTAFVV
mm17M1-20     1 -------------------------------------------------MPQSHSTAFIV
mm17M1-24     1 -------------------------------------------------MPQCHSTAFIV
mm17M1-18     1 -------------------------------------------------MPQFLSTVFVV
```

```
mm17M1-29     1 ------------MGILSTGNQTVTEFVLLGFHEVPGLHLLFFSVFTILYASIITGNMLIA
mm17M1-30     1 -------------------MEMSPQKTFTESVPLGLSEVPPIYLCIMIYASIITKNMLVL
mm17M1-31P    1 -------------------MGMSPQETFTESVPHGLSEVPPIYLCIMIYASIITRSMLIV
mm17M1-10     1 -----------------MNCSKTPGFILLGLSSDPEKWQPLFNIFLCLYLLGLLGNLLLL
mm17M1-11     1 ------MQISSSIISPRMNCSQAPGFILLGLPREPEKWQHFFIIFLGLYLLGLLGNLLLL
mm17M1-26     1 ----------------MMNCSQAPGFILLGLSSNSEKWQPLFSIFLVLYLLGLLGNLLLL
mm17M1-27     1 -----------------MNCSQAPTFILLGLSSDAEKWQPLFSIFLVLYLLGLLGNLLLL
mm17M1-28     1 -----------------MNCSQAPTLILLGLSSDAEKWQPLFSIFLVLYLLGLLGNLLLL
mm17M1-1      1 ----------------MVNQSSPVVFFLLGFSEHPQLKKVLFVVVLCSYLLTLLGNTLIL
mm17M1-5P     1 ----------------MVNQSSPVVFFLLGFSEHPQLEKVLFVVVLCSYLLTLLGNTLIL
mm17M1-2      1 ----------------MVNQSTPVGFLLLGFSEHPQLEKVLFVVVLCSYLLTLLGNTLIL
mm17M1-4      1 ----------------MVNQSSPVGFLLLGFSEHPQLEKVLFVIVLCSYLLTLLGNTLIL
mm17M1-3      1 ----------------MVNQSSPVGFLLLGFSEHPQLEKVLIVVVLCSYLLTLLGNTLIL
mm17M1-21     1 ---------------MAINKSSGGDFILVGFSDQPQLEKILFVLVLISYLLTLVGNTAII
mm17M1-45     1 ----------------MINSSVSSDFILVGFSDQPQLERRLFIVVLISYLLTLVGNTIII
mm17M1-6     61 LNIFVSPSEPSWSFPPQANHSSAERFLLLGFSDWPSLQPVLFALVLLCYLLTLTGNAALV
mm17M1-39     1 --------------MWINNQSSVDDFILLGFSDRPWLETPLFVIFLVAYIFALFGNISII
mm17M1-13     1 ----------------MSNQTSVTEFLLLGVTDIQELNPILFVIFFTIYFVNITGNGAIL
mm17M1-25     1 ----------------MSNQTSVTEFLLLGVTDVQELNPILFVIFFTIYFVNITGNGAIL
mm17M1-12     1 ----------------MSNQTSVTEFLLLGVTDIQELNPILFVIFFTIYFINITGNGAIL
mm17M1-14     1 ----------------MLNQTSVTEFILLGVRDIQEPQPFLFAIFFTIYFVNITGNGAIL
mm17M1-8      1 ---------------VLLNHTLVTEFLLLGVTDIQELNPILFVTVLAMYFVNVAGNGAIL
mm17M1-9      1 ---------------VLLNQTLVTEFLLLGVTDIQELNPILFVTVLAMYFVNVAGNGAIL
mm17M1-7P     1 ---------------VLLNHTFITEFLLLGVTDIQELNPILFVMVLAMYFINVFGNGAIM
mm17M1-34     1 ----------------MENSTSVDEFLLLGLTSVQKLQPIIFVMFLTIYLLNLVGNGVIL
mm17M1-23     1 --------------MEGKNQTAPSEFIILGFDHLNELQYLLFTIFFLTYICTLGGNVFII
mm17M1-33    50 APVCQIHKGKPADIMEGKNQTAPSEFIILGFDHLNELQYLLFTIFFLTYICTLGGNVFII
mm17M1-22     5 ICCPFFQEMSVNCSLWQENKLSVKHFAFAKFSEVPEECFLLFTLILLMFLVSLTGNALIT
mm17M1-36     1 --------MSVNCSLWQENKLSVKHFAFAKFSEVPEECFLLFTLILLMFLVSLTGNALIT
mm17M1-35     1 --------MSINCSLWQENSLSVKRFAFAKFSEVPGECFLLFTLILLMFLVSLTGNALIA
mm17M1-37     1 --------MSINCSLWQENSLSVKRFAFSKFSEVPGECFLLFTLILLMFLVSLTGNELIA
mm17M1-38     1 --------MSINCSLWQENSLSVKRFAFAKFSEVPGECFLLFTLILLMFLVSLTGNSLIA
mm17M1-46     1 ---------------MGTNSSLVTEFVLVGFSRLVHLQGILFSLFLTVYLLTVAGNLLIV
mm17M1-19    12 LFYFPAATTISVFKMIMENITTMSGFLLMGFSDNHELQILQAVLFLVTYLVGSAGNVIII
mm17M1-20    12 LFYFPAATAISVFKMIVENITTMRGFLLMGFSDNRELQILHALFFLVAYLLGSAGNVIII
mm17M1-24    12 LFYFPAATVIYVFKMIVENITTMSGFLLMGFSDNHELQILQALLFLVTYLVGSAGNVIII
mm17M1-32     1 --------------MTVKNITTMSGFLLMGFSDNRELQILYALLFLLTYLLGSAGNFIII
mm17M1-18    12 LSYFLAATTISVAKMIMENITTMSGFLLMGFSDNRELQILQALLFLVTYLVGSAGNFIII
mm17M1-42     1 --------------MTARNMTTMSGFLLMGFSDNHELQILQALLFLLTYLLGSAGNFIII
mm17M1-43     1 --------------MTPRNMTTVSGFLLMGFSDNHELQILQALLFLVTYLLDSAGNFIII
mm17M1-41     1 -----------------MNVSFKTGFLLMGFSDERNLQILHAVLFLITYLLAIMGNLLII
mm17M1-40P    1 -------------MNVANFTAMTIFLLLMGFSRNSQVEIIFSTLALVVLIGTISIVAVTS
mm17M1-44P    1 ---------------------------------------------------MLTQNTMII
```

```
mm17M1-29    49 VVVVSSQRLHTPMYFFLVNLSFIEIVYTSTVVPKMLEG-----FLQEATISVAGCLLQFF
mm17M1-30    42 VVVNSSQKLPTPMYFFLVSQAFQAHVHSGDQNAGGFSVGKPHSCDBKSLATLAGCLLQFL
mm17M1-31P   42 VMLDSSQRLHTPTHFFLVSQAFQAHVHSGDQNAGGFPVGKHHSCEBKSLVTLPGCLLQFF
mm17M1-10    44 LAIGTDVHLHTPMYFFLSQLSLVDLCFITTTAPKMLEALW----TGDGSISFSGCLTQFY
mm17M1-11    55 LAIGSDVHLHTPMYFFLSQLSLVDLCFITTTAPKTLETWW----TGDGSISFSGCLTQLY
mm17M1-26    45 LAIGTDVHLHTPMYFFLSQLSLVDLCFITTTAPKMLETLW----TGDGSISFSGCLTQLY
mm17M1-27    44 LAIGTDVHLHTPMYFFLSQLSLVDLCFITTTAPKMLEALW----TGDGSISFSGCLTQLY
mm17M1-28    44 LAIGTDVHLHTPMYFFLSQLSFVDLCFITTTAPKMLEALW----TGDGSISFSGCLTQLY
mm17M1-1     45 LLSTLDPRLHSPMYFFLSNLSFLDLCFTTTCVPQMLFNLW----GPAKTISFLGCFVQLF
mm17M1-5P    45 LLSTLDPRLHTPMYFFLSNLSFLDLCFTTTCVPQMLFNLW----GPEKTISFLGCFV-LF
mm17M1-2     45 LLSTLDPRLHSPMYFFLSNLSFLDLCFTTTCVPQMLFNLW----GPTKTISFLGCSVQLF
mm17M1-4     45 LLSTLDPRLHSPMYFFLSNLSFLDLCFTTTCVPQMLFNLW----GPAKTISFLGCSVQLF
mm17M1-3     45 LLSTLDPRLHSPMYFFLSNLSFLDLCFTTTCVPQMLFNLW----GPAKTISFLGCFVQLF
mm17M1-21    46 LVSCLDSALQTPMYYFLTNLSFVDICFSTSIVPQLLWNLH----GPAKTITATGCAIQLY
mm17M1-45    45 LISSIDSKLKTPMYFFLTHLSFVDICFTTSIVPQLLWNLK----GPAKTITAVGCAVQLY
mm17M1-6    121 LLAIRDPRLHTPMYYFLCHLALVDVGFTTSVVPPLLASLR----GSMLQLPRAGCMAQLC
mm17M1-39    47 LVSRLDPQLDSPMYFFVSNLSLLDLCYTTSTVPQMLVNLR----GPEKTISYGGCVAQLY
mm17M1-13    45 MIVILDPRLHSPMYFFLGNLACLDICFSTVTLPKMLQNLL----STNKAISFLGCITQLH
mm17M1-25    45 MIVILDPRLHSPMYFFLGNLACLDICFSTVTLPKMLQNLL----STNKAISFLGCITQLH
mm17M1-12    45 MIVILDPRLHSPMYFFLGNLACLDISYSTVTVPKLLQNLL----STSKAISFLGCITQLH
mm17M1-14    45 MIVILDPRLHSPMYFFLGNLACLDISYSTVTVPKMLENLL----STNKAISLLGCITQLH
mm17M1-8     46 MIVISDPRLHLPMYFFLGNLACLDICFSTVTVPKMLENFF----STSKAISFLGCITQLH
mm17M1-9     46 LIVISDPRLHSPMYFFLGNLACLDICFSTVTVPKILDNFF----STSKAISFLGCITQLY
mm17M1-7P    46 MIVILDSRLYSPMYFFLGNLACLDICFSTVTVPKMLENFF----STSKAISFLGCITQLH
mm17M1-34    45 MIVTLERRLHSPMYFFLGNLSCLDICYSSVTLPKVLINLL----SRRRAISFLGCITQLY
mm17M1-23    47 VVTIADSHLHTPMYYFLGNLALIDICYTTTNVPQMMVHLL----SEKKIISYGGCVTQLF
mm17M1-33   110 VVTIADSHLHTPMYYFLGNLALIDICYTTTNVPQMMVHLL----SEKKIISYGGCVTQLF
mm17M1-22    65 LAICTSPALHTPMYFFLANLSLLEIGYTCSVIPKMLQNLV----SEIRGISREGCVTQMF
mm17M1-36    53 LAICTSPALHTPMYFFLANLSLLEIGYTCSVIPKMLKNLV----TEARGISREGCATQMF
mm17M1-35    53 LVICTNPSLHNPMYFFLANLSLLEIGYTCSVIPKMLQSLV----SEAREISREGCATQMF
mm17M1-37    53 IAICTSPALHTPMYFFLANLSLLEIGYTCSVIPKMLQSLV----SEAREISREGCATQMF
mm17M1-38    53 LAICTSPALHTPMYFFLANLSLLEIGYTCSVIPKMLQSLV----SEARGISREGCATQMF
mm17M1-46    46 ALVSTDAALQSPMYFFLRILSALEICYTSVTVPLLLHHLL----TGRRHISRSGCALQMF
mm17M1-19    72 TITTLDPQLQSPMYYFLKQLSILDLSSLSVTVPQYVDSS----LARSGYISYGQCMLQIF
mm17M1-20    72 TITTLDPQLQSPMYYFLKHLSILDLSSLSVTVPQYVDIC----LTQSGYISYAQCMLQIF
mm17M1-24    72 TITTLDPQLQSPMYYFLKHLSILDLSSLSVTVPQYVDSS----LAQSGYISYAQCMLQIF
mm17M1-32    47 TITTLDPQLQSPMYYFLKHLSILDLSSLSVTVPQYVDSS----LAGSGYISYGQCMLQIF
mm17M1-18    72 TITTLDPQLKSPMYYFLKHLSILDLSSLSVTVPQYVDSS----LARSGYISYEQCMLQIF
mm17M1-42    47 TITTLDPQLQSPMYYFLKQLSTLDLSSLSVTVPQYVASS----LARSGYISYGQCMLQIF
mm17M1-43    47 TITTIDKQLQSPMYYFLKHLSIMDFSSLSVTVPQYVDSS----LARSGYISYGQCMLQVF
mm17M1-41    44 TIITLDQRLHSPMYYFLKHLSFLDLCFISVTVPQSIANS----LMNNGFISLGQCMLQVF
mm17M1-40P   48 LSIVX---LCSLMPFLLIHLFCFDVCYISVMMPKSVCS------SFMYSAYISPNAHCQV
mm17M1-44P   10 LVSFLNSRLQTPMYFFLSNFFFLDLCFMTNVLIVTSKG------PEK---ITHAVQSMST

mm17M1-15P    1 ----------FLLIVMVYDHYLTICHHL-YPFLMGPLWGLGFGLTDLX--FVVDELIVAL
mm17M1-29   104 VFGSLATDECFLLAVMAYDRYLAICHPLRYPHLMGPQWCLGLVLTVWLSGFMVDGLVVAL
mm17M1-16P    1 ------------------DHYLICHPLH-YPLLMGHQWCLGFVLTLQLFGITVDGLVVIL
mm17M1-17P    1 ------------------DHYLICHPLH-YPLLMGHQWCLGFVLTLQLFGITVDGLVVIL
mm17M1-30   102 TFTSLDADEYFLLTLMAHDHCLAIFYSL-YPRLMRPQWCLGLVIIVWLSGFMEAGLVVAL
mm17M1-31P  102 TFSSLYIDEYFLLALMAYDHCLAICYSL-YPRLMRPQWCLGLVLTVWVSGFMEDGLVVAL
mm17M1-10   100 FFAVFADMDNLLLAVMAIDRYAAICHPLFYPFLMTPCRCEVLASGSWGIAHCVSLFYTLL
mm17M1-11   111 FFGVFADMDNLLLAVMAIDRYAAICHPLLYPLLMTPCRCEVLVSGSWGIAHCVSLMYTLL
mm17M1-26   101 FFAVFADMDNLLLAVMAIDRYAAICHPLLYPLLMTPCRCRVLVSGSWGVAHCVSLTHILL
mm17M1-27   100 FFAVFADMDNLLLAVMAIDRYAAICHPLLYPLLMTPCRCRVLVSGSWGVAHCVSLTHTLL
mm17M1-28   100 FFAVFADMDNLLLAVMAIDRYAAICHPLRYSALMTPFRCGVLVSGSWGVTNCVSLTHTLL
mm17M1-1    101 IFMSLGTTECILLTVMAFDRYVAVCQPLHYATKINPHLCRQLAGIAWAIGLVQSIVQTPP
mm17M1-5P   100 IFMSLGTTECILLTVMAFDRYVAVCQPLHYATVINPRLCQQLAGIAWAIGLVQSIVQTPP
mm17M1-2    101 IFMLLGTTECILLTVMAFDRYVAVCQPLHYATIIHPRLCRQLAGVAWAIGLVQSIVQIPP
mm17M1-4    101 IFLSLGTTECILLTVMSFDRYVAVCQPLHYATVIHPRLCWKLAAVAWMMGLLQSIVQTPP
mm17M1-3    101 IFLSLGTTECILLAVMSFDRYVAVCQPLHYATVIHPRLCCQLAAVACTIGLVESVVQTPS
mm17M1-21   102 VSLALGSTECVLLAVMAFDRYAAVCRPLHYATVMHPRLCQSLAGVAWLSGVGNTLIQGTI
mm17M1-45   101 VSLTLGSTECILLAVMAFDRYAAVCKPLHYVAVMNPQLCRALAGISWLSGIGNALIQGTI
mm17M1-6    177 SSLALGSAECVLLAVMALDRAAAVCNPLRYTSLASPLLCRTLAGVSWLGGLANSAAQTAL
mm17M1-39   103 IFLALGSTECILLAIMAFDRFAAICRPLHYPIIMNQKRCIHMATGTWISGFANSLVQSTL
mm17M1-13   101 FFHFLGSTEAMLLPVMAFDRFVAICRPLHYSVIMNHQLCIHMTVTIWTLGFFHALLHSVM
mm17M1-25   101 FFHFLGSTEAMLLPVMAFDRFVAICRPLHYSVIMNHQLCIHMTVTIWTLGFFHALLHSVM
mm17M1-12   101 FFHFLGSTETMLLPVMAFDRFVAICRPLHYSVIMNHQLCIHMTVTIWTLGFFHALLHSVM
mm17M1-14   101 FFHFLGTTESLLLAVMAFDRFVAICRPLHYSVIMNWQVCILMAVTIWTIAFLHALLHSVM
mm17M1-8    102 FFNFLGSTEALLLTVMAFDRFVAICRPLHYPAIMNSQVCIQVAISIWAIPFLHALVHSIL
mm17M1-9    102 FFHLLGSTEALLLAVMAFDRFVAICRPLHYPSIMNGQVCIQVAISIWAIPFVHALVHSIL
mm17M1-7P   102 FFHFLGCTDALLLTVMAFDRFVAICRPLHYPSIMNRQVCIQVAATIWAIPFLHALVHSIL
mm17M1-34   101 FFHFLGSTEAILLAVMAFDRFVAICSPLRYTAIMNPQLCILLAATAWFTSFFYALLHSVM
mm17M1-23   103 AFIFFVGSECLLLAAMAYDRYIAICKPLRYSFIMNKALCSWLAASCWTCGFLNSVLHTVL
mm17M1-33   166 AFIFFVGSECLLLAAMAYDRYIAICKPLRYSFIMNKALCSWLAASCWTGGFLNSVLHTVL
mm17M1-22   121 FFIFFGITECCLLAAMAFDCYMAICSPLHYSTRMSREVCAHLALVSWGMGCIVGLGQTNF
mm17M1-36   109 FFIFFGITECCLLAAMAFDRYMAICSPLHYATRMSREVCAHLAIVSWGMGCIVGLGQTNF
mm17M1-35   109 FFTFFGITECCLLAAMAYDRCMAICSPLHYPTRMSSGVCAHLAIVSWGMGCIVGLGQTNF
mm17M1-37   109 FFTFFGITECCLLAAMAYDRCMAICSPLHYATRMSHGVCAHLAIVSWGMGCIVGLGQTNF
mm17M1-38   109 FFIFFGITECCLLAAMAFDRYMAICSPLHYATRMSRGVCAHLAIVSWGMGCIVGLGQTNF
mm17M1-46   102 FFLFFGATECCLLAAMAYDRYAAICEPLRYQVLLSRRVCVQLAGAAWSCGALVGLGHTSF
mm17M1-19   128 FFTWFAWGEMAILTVMSYDRYIAVCLPLHYEIIMCPRKCRWAVTAVWLSSSIPGTLYLAT
mm17M1-20   128 FFTGFAWGEVAILTVMSYDRYVAVCLPLHYEVIMGPSKCRWAVTAVWLSSVIPGTLYIAS
mm17M1-24   128 FFTAFAWGELAILTVMSYDRYVAICLPLHYEVIMSPRKCTWAVATVWLSGGISGTLYITG
mm17M1-32   103 FFAAFAWGEVAILTVMSYDRYVAICLPLHYEVIMSPRKCTWAVTSVWLSSVIPGTLYIAS
mm17M1-18   128 FFTCFAWDEMAILTVMSYDRYVAVCLPLHYEVIMSPRKCTWALAAVWLSGGVSGTLYTAS
mm17M1-42   103 FFTGLAWSEMATLTVMSYDRYVAICLPLHYEVIMSPRKCTWAVAAVWLSGGISGTLFTAS
mm17M1-43   103 FFTGLAWSEVAILTVMSYDRYVAICLPLHYEVIMSPRKCTWAVAAVWLSGGISGTLFTAS
mm17M1-41   100 FFIALASSEVAILTVMSYDRYVAICRPLQYETIMDPHACKCAVIAVWMAGGLSGLLHTGV
mm17M1-40P   99 FYSQSSYTAMAILTVMSYDCYMAVWHKVITNVSTCIHGVLAVLVNVMN--YLWSYAHXQL
mm17M1-44P   61 LFCDWTXTKCVLLTMMAYNPVTPICWPLXCSPYYTPKICHTPKVSLEASCLGLDLFYGVH
```

```
mm17M1-15P   48 MAQLRFCVPK----QIDHFYYDFSPLVVLAYTDTGLVQVTTFVLFVVFLTVPFG-LVLIS
mm17M1-29   164 MAQLRFCGPN----LVDHFYCDFSPLMVLACSDTQVAQVTTFVLSVVFLTVPFG-LVLIS
mm17M1-16P   42 VAQMWFCGPN----LIDYFYNFSP--IMDLASDTQVFQVITFVLSVVFLTVPFG-LVLIS
mm17M1-17P   42 VAQMWFCGPN----LIDYFYNFSP--IMDLASDTQVFQVITFVLSVVFLTVPFG-LVLIS
mm17M1-30   161 TAQLRFCGPN----LIDHFYCDFSP-LMILACSDTVAQMTTFVLFVVFLPVLSG-LILMS
mm17M1-31P  161 IAQLRFCGPN----LIDHFYCDFS---PLMACFDTVAQMTTFVLSVIFLTVPFGXLVLIS
mm17M1-10   160 LSQFYYHTNQ----GIPHFFCDSRPLLLLSCS-DTHLSEGLMMALSGVLGMSSVLCLVSS
mm17M1-11   171 LSQLYFHTNQ----EIPRFFCDCRPLLLLSCS-DTHLNEVLMMALAGVLGVSAVLCIVSS
mm17M1-26   161 LSQLYFHTNQ----EIPHFFCDFGPLLLLSCS-DAHLNESLMMALAGVLGISALLCIVSS
mm17M1-27   160 FSKLYFHNNQ----EIPHFFCDFGPLLLLSCS-DTYLNESLMMALSGLLAISAFLCIVSS
mm17M1-28   160 LSKLYFHTNQ----EIPHFFCEFGPLLLLSCS-DTHLNKILVIILVGILGISAVLCIVSS
mm17M1-1    161 TLKLPFCSHR----QIDNFLCEVPSLIQLSCG-DTTYNEIQMAVASIFIVVVPLSLILVS
mm17M1-5P   160 TLKLPFCSHR----QIDNFVCEVPSLIQLSCG-DITYNEIQMAVASIFIVVVPLSLILVS
mm17M1-2    161 TLTLPFCSHR----QIDDFLCEVPSLIRLSCG-DTTFNEIQLSVAGVIFLLVPLSLIIVS
mm17M1-4    161 TLKLPFCPHR----QIDDFLCEVPSLIRLSCG-DTTFNEIQLAVSSVILVVVPLSLILVS
mm17M1-3    161 TLRLPFCPHH----QVDDFVCEVPALIRLSCG-DTTYNEIQMAVASVFILVVPLSLILVS
mm17M1-21   162 TLRLPRCGNH----KIYHFICEVPAMIKLACV-DIHANEVQLFMASLVLLLLPLTLILVS
mm17M1-45   161 TLWLPRCGHL----WLHHFFCEVPSMIKLACV-DIHANEVQLFVASLVLLLLPLALILTS
mm17M1-6    237 LAARPLCAPR----CLDHFICELPALLQLACRGGRSATERQMFAARVVILLVPSAVILAS
mm17M1-39   163 TVVAPRCGQR----VIDHFFCEVPALLKLACT-DTSVNEAELNVLGALLLLVPLSLILGT
mm17M1-13   161 TSRLSFCGPN----HVHHFFCDIKPLLDLACG-NTELNLWLLNTVTGTIALTPFFLTFLS
mm17M1-25   161 TSRLSFCGPN----HVHHFFCDIKPLLDLACG-NTELNLWLLNTVTGTIALTPFFLTFLS
mm17M1-12   161 TSRLSFCGPN----HVHHFFCDIKPLLDLACG-NTELNLWLLNTVTGTIALTSFFLIFLS
mm17M1-14   161 TSRLSFCGLN----HIHHFFCDVKPLLELACG-NTELNLWLLNTVTGTIASVPFFLTFLS
mm17M1-8    162 TSQLNFCGSN----HIHHFFCDVKPLLELACG-NTELNRWLLNTLTGTVAIGLFFLTFLS
mm17M1-9    162 TSQLNFCGSN----QIHHFFCDVKPLLELACG-NTELNRWLLNTFTGTFAIGLFFLTFLS
mm17M1-7P   162 TSQLNFCGSN----RIHHFFCDVKPLLELACG-NTELNRWLLNTLAGTIGIGLFFLTFLS
mm17M1-34   161 TAHLNFCHSH----KLSHFFCDVKPLLEVACG-NTVLNQWLLSVVTGSISMGAFLLILLS
mm17M1-23   163 TFHLPFCGNN----QINYFFCDIPPLLILSCG-DTSLNELALLSIGILIGWTPFLCIILS
mm17M1-33   226 TFHLPFCGNN----QINYFFCDIPPLLILSCG-DTSLNELALLSIGILIGWTPFLCIILS
mm17M1-22   181 IFSLNFCGPC----EIDHFFCDLPPVLALACG-DTSQNEAAIFVAVVLCISSPFLLIIYS
mm17M1-36   169 IFSLNFCGPC----EIDHFFCDLPPVLALACG-DTSQNEAAIFVTVVLCISSPFLLIIYS
mm17M1-35   169 IFSLEFCGPC----EIDHFFCDLPPVLALACG-DTSQNEAAIFVAAVLCISSPFLLIIYS
mm17M1-37   169 IFSLNFCGPC----EIDHFFCDLPPVLALACG-DTSQNEAAIFVAAILCISSPFLLIIYS
mm17M1-38   169 IFSLNFCGPC----EIDHFFCDLPPVLALACG-DTSQNEAAIFVAAVLCIFSPFLLIISS
mm17M1-46   162 IFSLPFCGPN----AVPHFFCEIQPVLQLVCG-DTSLNELQIILAAALIILCPFGLILSS
mm17M1-19   188 IFSIRICRAK----IIHQFFCDVPQLLKLSCS-NDYLVIMGVADFLSVIGFACFVGIVIS
mm17M1-20   188 IFSIRFCGDR----IIHQFFCDVPQVLKLSCS-DDYLVTVGVADFLSAVAFACFIGIVNS
mm17M1-24   188 TLFIRFCGDK----IIHQFFCDVPQLLKLSCS-NDHLVIMDMVSFLTAVSFACFTGIVIS
mm17M1-32   163 IFSIRFCRAK----IIHQLFCDVPQLLKLSCS-NDHLVVIGMVSFMTAVAFACFVGIVIS
mm17M1-18   188 TLSIRFCGDR----IIHQFFCDVPQVLKLSCS-NDYLLTIGVANILSAVAFACFIGIVIS
mm17M1-42   163 TLSIRFCGDK----IIHQFFCDIPQLLKLSCS-NDYFGVLEVSTFMSVMAFACFVGIAFS
mm17M1-43   163 TLSIRFCGHK----IIHQFFCDIPQLLKLSCS-NDDFGLLKVSTFIAVMGFACFVGIAFS
mm17M1-41   160 NFSIPLCGKR----IIHQFFCDIPQMLKLACS-YEFINEIAVAAFTTSTAFVCLIAIVFS
mm17M1-40P  157 LRLHLHCG------TSTIRFCDVLLVLKLSFT-NDHVNELESLAX---------------
mm17M1-44P  121 NSDHSGFSIISVPPQNEFFMCEESPLVKITFMDTTSLEKHISVFT--FLAVIPCGEYSII


mm17M1-15P  103 CAQIAVTVLR-VPSRTRRNKAFSTCSSHLDEVSTFYGSLMVWYTEPSAVHS--QILSKVI
mm17M1-29   219 YAQIVVTVLR-VPSGTRRTKAFSTCSSHLAVVSTFYGTLMVLYIVPSAVHS--QLLSKVI
mm17M1-16P   95 YIQIVVTVLR-VLSGDRRTKDFSTCSSHLAVVSTFYRSLMVLYTVPFAPX-----LSKVI
mm17M1-17P   95 YIQIVVTVLR-VLSGDRRTKDFSTCSSHLAVVSTFYRSLMVLYTVPFAPX-----LSKVI
mm17M1-30   215 YAQFVVIVLR-IPSGARRTKAFFTCSSHLAMMFTFYGSLMVWYTAPSAVXLVCTLLSKVI
mm17M1-31P  214 YAQLVVTVLR-ILSGARRTKAFVICSSHLAVVSTLYGTLMGLYTVPFVV--HSQLLTKVI
mm17M1-10   215 YGCIFYAVAR-VPSAQGKRKSLATCSSHLSVVLLFYSTVFATYLKPPST--SHSSAEVVA
mm17M1-11   226 YGCIFYAVAR-VPSAQGKRKALTTCSSHLSVVLLFYSTVFATYLKPPST--SHSSGEVVA
mm17M1-26   216 YGCIFYAVAK-VPSAQGKRKALATCSSHLSVVLLFYSTVFATYLKPPSS--SRSSGEVVA
mm17M1-27   215 YGCIFYAVAK-VPSAQGKRKALATCSSHLSVVLLFYSTVFATYLKPPSS--SHSSQEVVA
mm17M1-28   215 YGCIFYAVAK-VPSAQGKRKALSTCSSHLSVVLLFYSTVFATYLKPPSS--SRSSEEVVA
mm17M1-1    216 YGAIARAVLK-ISSAKGRRKAFGTCSSHLIVVTLFYSSVIAVYLQPKNP--YARERGKFF
mm17M1-5P   215 YGAIARAVLK-ISSAKGRRKAFGTCSSHLIVVTLFYSSVIAVYLQPKNP--YARERGKFF
mm17M1-2    216 YGVIARAVLK-TNSSKGRRKAFGTCSSHLIVVTLFYSSVIAVYLQPKNP--YAQERSKFF
mm17M1-4    216 YGAIARAVMR-INSTEAWKKALRTCSSHLIVVTLFYSSVIAVYLQPKNP--YAQERGKFF
mm17M1-3    216 YGAIARAVLR-ISSAKGRRKAFGTCSSHLIVVTLFYSSVIAVYLQPKNP--YARERGKFF
mm17M1-21   217 YGYIAQALMR-LRSALTWGKALGTCGSHLIVVVLFYGTSTAVYIHPNSS--YAQSQGKFI
mm17M1-45   216 YGHIAKAVIR-IKSSQAWRRALGTCGSHLMVVSLFYGSITAIYIQPNSS--YAHTHGKFI
mm17M1-6    293 YIAVGRAVWG-MHSSSGWRKAASTCGSHLTAVCLFYGSATYTYLQPTHS--YNQGRGKFV
mm17M1-39   218 YVFIAQAVLK-LRSAESRRKAFNTCASHLLVVSLFYFTAISMYVQPPSS--YSHERGKIM
mm17M1-13   216 YFYIITYLFLKTRSCSMLHKALSTCASHFMVVILLYVPVLFTYIRPASG--SSLDQDRII
mm17M1-25   216 YFYIIIYLLLKTRSCSMLHKALSTCASHFMVVFLFYAPVLFIYISPTSG--SSLDQDRII
mm17M1-12   216 YFYIITNLLLKTRSCSMLHKALSTCASHFMVVVLFYAPVLFTYIRPASG--SSLDQDTII
mm17M1-14   216 YFYIITYLFLKTRSCSMLHKALSTCASHFMVVVLFYAPVLFTYIRPTSG--SSLDQDRII
mm17M1-8    217 YFYIVTYLFLKTRSCSMLHKALSTCASHFMVVMIFYAPVLFIYINPDSG--SSLEKDRII
mm17M1-9    217 YFYIITYLFLKTRSCSMLHKALSTCASHFMVVMIFYAPVLFIYINPDSG--SSLEKDRII
mm17M1-7P   217 YFYIVTYLFLKTHSCSMLHKALSTCASHFMVVMIFYAPVLFIYINPDSG--SSLEKDRII
mm17M1-34   216 YFYIIAFLLFKNRSCRMLKKALSTCTSHFMVVCLFYGPVGFTYIRPATASASSMSEDRVV
mm17M1-23   218 YLYIISTILR-IRSSEGRQKAFSTCASHLLIVILYYGSAIFTYVRPISS--YSLEKDRLI
mm17M1-33   281 YLYIISTILR-IRSSEGRHKAFSTCASHLLIVILYYGSAIFTYVRPISS--YSLEKDRLI
mm17M1-22   236 YVRILVAVLV-MPSPEGRHKALSTCSSHLLVVTLFFGSGSITYLRPKSS--HLPGMDKLL
mm17M1-36   224 YVRILFAVLV-MPSPEGRHKALSTCSSHLLVVTLFYGSASITYLRPKSS--HSPGIDKLL
mm17M1-35   224 YVRILVAVLL-MPSPEGRHKALSTCSSHLLVVTMFYGSASITYLRPKSS--HSPGMDKLL
mm17M1-37   224 YVRILVAVLV-MPSPEGRHKALSTCSSHLLVVTLFFGSGSITYLRPKSS--HLPGMDKLL
mm17M1-38   224 YVRILIAVLV-MPSREGRHKALSTCSSHLLVVTLFYGSTSATYLRPKSD--HSPEVDKLL
mm17M1-46   217 YGRILVTIFR-IPSAAGRRKAFSTCSSHLVVVSLFYGTAIFIYIRPKAS--YDPTTDPLL
mm17M1-19   243 YVHIFSTVLR-MPSAESRSKVFSTCLPHLFVVSLFLSTGIFAYLNPTSD--FPTALEFLF
mm17M1-20   243 YVHIFSTVLR-MPSAESRSKVFSTCLPHLFVVLLFLSTGIFAYLNPTSD--SPTALQFLV
mm17M1-24   243 YVHIFSTVLR-MPSAESRSKVFSTCLPHLFVVSLFLSTGAFAYLNLTSD--SSTALEFLL
mm17M1-32   218 YVHIFSTVLR-MPSAESRSKVFSTCLPHLFVVSLFLSTGSCAYLNTSSD--SPTALEFLF
mm17M1-18   243 YVHIFSTVLR-MPSAESRYKVFSTCLPHLFVVSLFLSTSTFAYLNPTAD--SPTALEFLF
mm17M1-42   218 YGQIFSTVLR-MPSAEGRSKVFSTCLPHLFVVSFFLSTGICAYLKPTSD--SPTALDLML
mm17M1-43   218 YCQIFSTVLR-MPSAEGRSKVFSTCLPHLFVVSFFLSTGICAYLKPSSD--SPTALDLML
mm17M1-41   215 YTQIFSTVMR-IPSADSRTKVFSTCLPHLFVVMFFLSAAGFEFLRPPSD--SLSAMDLVF
mm17M1-40P  195 -----SHLWR-------AEPVFLTCLGHVSVGSLFNPPGVFEFLNPYSE---SPTSLDII
mm17M1-44P  179 YLLVLLKVWLKIKFTG-RMKTFGSCGFHLMAIVLFFGNESSVYMVYMYPRANACQYRKFL
```

```
mm17M1-15P   160 ALLYTVVTTIFDPGIYTLRNQEVQQSLRRHLYCKPTEM----------- 197
mm17M1-29    276 ALLYTVVTPIFNPVIYTLRNQEVQQALRRLLYCKPTEM----------- 313
mm17M1-16P   149 ALLYKVVIPIFNHVIYTLRNQEVP------------------------- 172
mm17M1-17P   149 ALLYKVVIPIFNHVIYTLRNQEVP------------------------- 172
mm17M1-30    274 ALLYTVFAPIFNSVIYTLRNLDMQKALRRLLYCKSTEM----------- 311
mm17M1-31P   271 ALLYTVFTPIFNPVIYTLKNQEVQQALRRLLYC---------------- 303
mm17M1-10    272 AVMYTLVTPTLNPFIYSLRNKDVKSSLRKILNMDKFQG----------- 309
mm17M1-11    283 AVMYTLVTPTLNPFIYSLRNKDVKSSLRRVLNIEKSQD----------- 320
mm17M1-26    273 AVMYTLVTPTLNPFIYSLRNKDVKSSLRRILNMVKSQD----------- 310
mm17M1-27    272 AVMYTLVTPTLNPFIYSLRNKDVKSSLRRILNMVKSQD----------- 309
mm17M1-28    272 AVMYSLVTPTLNPFIYSLRNKDVKSSLRRILNME--------------- 305
mm17M1-1     273 GLFYAVGTPTLNPLVYTLRNKEVKRAFWKLLRKDEDSEES--------- 312
mm17M1-5P    272 GLFYAVGTPILNPLVYTLRNKEVKRAFWKLLRKDEDSEES--------- 311
mm17M1-2     273 GLFYAVGTPTLNPLVYTLRNKEVKRAFWRLLGKDAASGRN--------- 312
mm17M1-4     273 GLFYAVGTPTLNPLIYTLRNKEVKRAFWRLLGKDGDSKNT--------- 312
mm17M1-3     273 GLFYAVGTPSLNPLIYTLRNKEVKRAFRRLLWKEVKPS----------- 310
mm17M1-21    274 TLLYTVVIPTLNPLIYTLRNKDVKGALKRLVRKDSSTGKKILSR----- 317
mm17M1-45    273 SLFYTVMTPTLNPLIYTLRNKEVKGALGRLFNRASGV------------ 309
mm17M1-6     350 SLFYTVVTPALNPLIYTLRNKEVKGAALRLLRSLGRP------------ 386
mm17M1-39    275 ALFYGIVTPTLNPFIYTLRNKDVKAALRRALTKEFWVKARQ-------- 315
mm17M1-13    274 AIMYSVVTPALNPLIYTLRNKEVRSALNRKVRRWL-------------- 308
mm17M1-25    274 AIMYSVVTPALNPLIYTLRNKEVRSALNRKLRRWL-------------- 308
mm17M1-12    274 AIMYSVVTPALNPLIYTLRNKEVRSALNRKVRRWL-------------- 308
mm17M1-14    274 AIMYSVVTPALNPLIYTLRNKEVRSALNRKVRRCLLLEEI--------- 313
mm17M1-8     275 AVMYTVVTPALNPLIYTLRNKEVRGALNRKIRILL-------------- 309
mm17M1-9     275 AVMYTVVTPALNPLIYTLRNKEVRGALNRKLRILL-------------- 309
mm17M1-7P    275 AVMYTVVTPALNPLIYALRNKEVRCALNRKLRILI-------------- 309
mm17M1-34    276 AIIYSAVTPVLNPLIYTLRNKEVMLALKKNFGKKLFKGN---------- 314
mm17M1-23    275 SVLYSVFTPMLNPIIYALRNKDIKEAVKAIGRKWQPPVFSSDM------ 317
mm17M1-33    338 SVLYSVVTPMLNPVIYTLRNKDIKEAVKAIGRKWQPPVFSSDI------ 380
mm17M1-22    293 ALFYTAVTSMLNPIIYSLRNKEVKAALRKTLSLKTSRAINR-------- 333
mm17M1-36    281 ALFYTAVTSMLNPIIYSLRNKEVKAALRRTLSLKKPLAINR-------- 321
mm17M1-35    281 ALFYTAVTSMLNPIIYSLRNKEVKAALRKTLSLKKPLAINR-------- 321
mm17M1-37    281 ALFYTAVTSMLNPIIYSLRNKEVKTALRKTLSLKTSRAINR-------- 321
mm17M1-38    281 ALFYTAVTSMLNPIIYSLRNKEVKAALRKTLSLKKVLIMNR-------- 321
mm17M1-46    274 SLFYAVITPILNPVIYSLRNADVKAALKRSIQKMGPSEI---------- 312
mm17M1-19    300 SVFYTVLPPTLNPVIYSLRNDAIKSVVRKLLLSRKFTS----------- 337
mm17M1-20    300 SIFYTVLPPTLNPVIYSLRNETIKSVIRKLLLSSKFTG----------- 337
mm17M1-24    300 SIFYTVLPPTLNPVIYSLRNETIKNVVRKLLLSTKFTVRIIFSCCF--- 345
mm17M1-32    275 SIFYTVLPPTLNPVIYSLRNETIKSVVRKLLLSSKFTVRIICPVATD-- 321
mm17M1-18    300 SILYTVLPPTINPVIYSLRNETIKSVVRKLLSSTKFTV----------- 337
mm17M1-42    275 SIFYTLLPPTLNPVIYSLRNESLKRALKKLLLSEEFIRKKCLFYF---- 319
mm17M1-43    275 SIFYTVLPPTLNPVIYSLRNESLKRAVKKLLLSEEFIGKNYVCSVFSAC 323
mm17M1-41    272 SIFYTVIPPTLNPLIYSLRNEAMKAALRKVLSKEEFSRRMVYVKAIFNL 320
mm17M1-40P   240 VKXVFILPQTLSVEIYSLSNEAIDTA----------------------- 265
mm17M1-44P   238 SXFYMIVTPSINPLIY-LRNKEFRWAVQRLVTRDPS------------- 272
```

Appendix 10: List of mouse ORs from the UCSC assembly

Name	Contig	Comments	Closest MHC-linked human OR	%age similarity to human MHC-linked OR.
mm13M1-1	13.20000001-13.25000000		hs6M1-22P	60
mm13M1-2P	13.20000001-13.25000000		hs6M1-15	85
mm13M1-3	13.20000001-13.25000000		hs6M1-15	86
mm13M1-4	13.20000001-13.25000000		hs6M1-8P	81
mm13M1-5	13.20000001-13.25000000		hs6M1-35	85
mm13M1-6	13.20000001-13.25000000		hs6M1-34P	85
mm13M1-7	13.20000001-13.25000000		hs6M1-31P	81
mm13M1-8	13.20000001-13.25000000		hs6M1-32	81
mm13M1-9	13.20000001-13.25000000		hs6M1-34P	94
mm13M1-10F	13.20000001-13.25000000		hs6M1-10	90
mm13M1-11	13.20000001-13.25000000		hs6M1-10	87
mm13M1-12	13.20000001-13.25000000		hs6M1-10	85
mm13M1-13	13.20000001-13.25000000		hs6M1-30P	83
mm17M1-47	17.40000001-17.45000000		hs6M1-7P	78
mm17M1-48	17.40000001-17.45000000		hs6M1-2P	83
mm17M1-49P	17.40000001-17.45000000	1 FS	hs6M1-2P	82
mm17M1-50P	17.40000001-17.45000000	1 FS	hs6M1-2P	82
mm17M1-51P	17.40000001-17.45000000	1 FS	hs6M1-3 hs6M1-4	81
mm17M1-52P	17.40000001-17.45000000	1 FS, 1 insertion, 4 stops, no methionine	hs6M1-3 hs6M1-6	49
mm17M1-53	17.40000001-17.45000000		hs6M1-2P	83
mm17M1-54	17.40000001-17.45000000		hs6M1-17	54
mm17M1-55	17.40000001-17.45000000		hs6M1-25P	72
mm17M1-56	17.40000001-17.45000000		hs6M1-32	61

Appendix 11: Human ‘ROLF’ database

Name	State	Accession number	Position in acc. No.	Comments
hs1M1-1	C	AL390860	65000..67000	
hs1M1-2	C	AL365440	43000..45000	
hs1M1-3	F	AL356607	59000..61000	goes into gap
hs1M1-4	C	AL356607	68000..70000	
hs1M1-5	P	AL365440	92000..94000	5 FS, 6 stops
hs1M1-6	C	AL513488	143000..145000	
hs1M1-7	C	AC024654	8000..10000	
hs1M1-8	C	AC024654	14000..16000	
hs1M1-9	C	AC024654	24000..26000	
hs1M1-10	P	AC024654	73500..75500	1 stop
hs1M1-11	P	AC024654	104500..106500	1 stop
hs1M1-12	C	AC024654	153500..155500	
hs1M1-13	C	AC025115	20500..22500	
hs1M1-14	P	AC025115	26000..28000	1 stop
hs1M1-15	P	AC025115	66500..68500	1 FS
hs1M1-16	C	AC025115	89000..91000	
hs1M1-17	C	AL356607	92500..94500	
hs1M1-18	C	AL356607	110500..112500	
hs1M1-19	P	AL354713	64000..66000	1 stop
hs1M1-20	P	AL357039	4000..6000	1 FS
hs1M1-21	C	AL357039	15000..17000	
hs1M1-22	C	AL357039	34000..36000	
hs1M1-23	P	AL357039	48000..50000	Missing start
hs1M1-24	C	AL357039	116000..118000	
hs1M1-25	P	AL357039	121000..123000	4 FS,stops
hs1M1-26	C	AL357039	149500..151500	
hs1M1-27	C	AL357039	171000..173000	
hs1M1-28	F	AC026038	126000..128000	goes into gap
hs1M1-29	P	AL358773	11500..13500	1 FS
hs1M1-30	P	AL513488	97500..99500	1 FS
hs1M1-31	C	AL513488	81500..83500	
hs1M1-32	C	AL358773	60500..62500	
hs1M1-33	C	AB045359	156000..158000	
hs1M1-34	P	AB045359	92000..94000	No Met, 5 FS,
hs1M1-34	C	AL513488	59000..61000	
hs1M1-35	C	AB045359	140000..142000	
hs1M1-36	C	AL358874	12000..14000	
hs1M1-38	C	AB045360	41500..43500	
hs1M1-39	P	AB045360	129000..131000	1 FS, 1 stop
hs1M1-40	C	AB045360	152000..154000	
hs1M1-41	P	AB045360	158500..160500	No Met
hs1M1-42	C	AB045360	134000..136000	
hs1M1-43	C	AB045361	14000..16000	2 stops, Alu insertion

hs1M1-44	C	AB045361	26000..28000	
hs1M1-45	C	AB045361	37000..39000	
hs1M1-46	C	AB045361	48000..50000	
hs1M1-47	P	AB045361	80000..82000	1 FS
hs1M1-48	P	AB045365	87500..89500	No Met, 1 FS
hs1M1-49	P	AB045365	122000..124000	1 FS
hs1M1-50	C	AL391534	101500..103500	
hs1M1-51	C	AL358874	84500..86500	
hs1M1-52	P	AL513323	144500..146500	1 FS
hs1M1-53	P	AL513323	159500..161500	1 stop, 1 FS
hs1M1-54	P	AL513323	198500..200500	4 stops, 2 FS
hs1M1-55	P	AC016787	13000..15000	1 stop
hs1M1-56	C	AC016787	134000..136000	
hs1M1-57	C	AL358874	102000..104000	
hs1M1-58	C	AL358874	122500..124500	
hs1M1-59	F	AL358874	133000..135000	goes into gap
hs1M1-60	C	AL358874	149000..151000	
hs1M1-61	C	AL391534	8000..10000	
hs1M1-62	C	AL513488	36000..38000	
hs1M1-63	C	AL513488	1500..3500	
hs1M1-64	C	AL391534	140000..142000	
hs1M1-65	C	AL391534	62000..64000	
hs1M1-66	C	AL450303	10500..12500	
hs1M1-67	C	AL450303	46500..48500	
hs1M1-68	C	AL450303	80000..82000	
hs1M1-69	C	AL450303	102000..104000	
hs1M1-70	C	AL450303	131000..133000	
hs1M1-71	PF	AL391904	40500..42500	No Met, 1 FS, missing motifs
hs1M1-72	P	AL391904	67000..69000	2 FS, 2 stops
hs2M1-1	P	AC009237	70000..72000	no Met
hs2M1-2	P	AC007881	21500..23500	3 FS,2 stop, no Met
hs2M1-3	P	AC007881	4000..6000	1 FS, 3 stops
hs2M1-4	F	AC069348	4000..6000	goes into gap
hs2M1-5	P	AC007040	177000..179000	1 FS,2 stop, no Met
hs2M1-6	P	AC064843	37500..39500	1 FS, 3 stops
hs2M1-7	P	AC064843	105000..107000	5 stops
hs2M1-8	P	AC064843	61500..62500	2 FS,goes into gap
hs2M1-9	P	AC013469	106000..108000	2 FS, stop
hs2M1-10	P	AC013469	21000..23000	3 FS
hs2M1-11	P	AC013469	45000..47000	4 FS
hs2M1-12	PF	AC013469	74000..76000	1 FS, no start
hs3M1-1	F	AC022049	12500..14500	goes into gap
hs3M1-2	P	AFO42089	47297..48332	
hs3M1-3	P	AFO42089	61204..62209	
hs3M1-4	P	AFO42089	69892..70886	
hs3M1-5	P	AC023058	70000..72000	stops,no met

hs3M1-6	P	AC024169	17500..19500	1 FS, 1 stop
hs3M1-7	P	AC024709	89500..91500	2 stop,3 FS
hs3M1-8	P	AC034187	21000..23000	2 stop,1 FS
hs3M1-9	P	AC067827	16000..18000	2 stop,1 FS
hs3M1-10	P	AC069518	41500..43500	3 FS, 1 stop
hs3M1-11	P	AC024892	56500..58500	1 FS
hs3M1-12	PF	AC024892	144000..146000	2 FS,missing start
hs3M1-13	C	AC025942	29500..31500	
hs3M1-14	P	AC025942	43500..45500	1 stop
hs3M1-15	C	AC025942	109000..111000	
hs3M1-16	P	AC025942	15000..17000	3 FS, 1 stop
hs3M1-17	P	AC025942	68000..70000	1 FS
hs3M1-18	P	AC025942	135500..137500	1 FS
hs3M1-19	P	AC025942	150000..152000	2 FS,short tail
hs3M1-20	C	AC074274	104500..106500	
hs3M1-21	C	AC074274	120500..124500	
hs3M1-22	P	AF186996	48000..50000	1 FS, 1 stop
hs3M1-24	P	AF186996	60000..62000	no met, stops
hs3M1-25	P	AF186996	66500..68500	no met, stops
hs3M1-26	P	AF186996	76500..78500	3 FS, stops
hs3M1-27	P	AC069047	92000..94000	1 FS, stops
hs3M1-28	P	AC069047	104500..106500	no met, stops
hs3M1-29	P	AC069047	174000..176000	no met,stops
hs3M1-30	P	AF165423	58000..60000	3 FS, 2 stops
hs3M1-31	P	AF165423	71000..73000	1 FS, no Met, stops
hs4M1-1	P	AC015709	82000..84000	3 FS, 1 stop
hs4M1-2	P	AC022674	33000..35000 stops	
hs4M1-3	P	AC022674	86500..88500	no met, 2 stop, 1 FS
hs4M1-4	P	AC022674	115500..117500	1 FS, no met, stops
hs4M1-5	P	AC068403	149500..151500	2 FS, 1 stop
hs4M1-6	P	AC007310	114000..116000	3 FS,stops, unfinished end
hs4M1-7	PF	AC013359	172000..174000	Alu insertion, no end, 4 stops
hs4M1-8	PF	AC013662	108000..110000	1 FS, 1 stop, missing end
hs4M1-9	P	AC011744	157000..159000	1 FS, 3 stop, no met
hs4M1-10	P	AC011744	175000..177000	2 FS, 2 stop
hs4M1-11	P	AC011744	96000..97000	no met, goes into gap
hs4M1-12	PF	AC008374	6000..8000	goes into gap, 1 FS(100% 16M1-1 no clone overlap)
hs4M1-13	P	AC022674	108000..110000	stops, 3 FS, no met
hs5M1-1	C	AC008620	87000..89000	
hs5M1-2	P	AC008620	119000..121000	1 FS
hs5M1-3	P	AC008454	47500..49500	Alu insertion,no met,4 stops
hs5M1-4	C	AC023255	180000..182000	
hs5M1-5	P	AC025336	160000..162000	1 FS,1 stop

For MHC-linked ORs, refer to Appendix 6

hs6M1-11	P	AL031259	8399..9336	1 FS
hs6M1-36	C	AL135904	21000..23000	

hs7M1-1	C	AC004853	17162..18115	
hs7M1-2	C	AC004853	41900..42853	
hs7M1-3	C	AC004853	85926..86861	
hs7M1-4	C	AC076959	40500..42500	
hs7M1-5	PF	AC076959	11000..12000	1 stop
hs7M1-6	C	AC076959	26000..28000	
hs7M1-7	C	AC076959	57000..59000	
hs7M1-8	P	AC076959	79500..81500	1 stop
hs7M1-9	C	AC073647	108000..110000	
hs7M1-10	P	AC004853	62834..63764	1 FS
hs7M1-11	P	AC004967	113880..114883	2 FS, 1 stop
hs7M1-12	P	AC004967	133062..134010	2 FS, 2 stop
hs7M1-13	P	AC004889	8842..9772	1 FS
hs7M1-14	C	AC004889	84457..83528	
hs7M1-15	P	AC004889	103217..102288	1 stop
hs7M1-16	P	AC076959	94000..96000	No Met, 2 FS
hs7M1-17	C	AC073647	164500..166500	
hs7M1-18	P	AC076959	123000..125000	4 FS, 3 stops
hs7M1-19	F	AC076959	136000..137000	
hs7M1-20	P	AC076959	140000..142000	1 FS
hs7M1-21	P	AC073647	140000..142000	1 stop
hs7M1-22	P	AC079804	89500..91500	3 FS, 1 stop
hs7M1-23	P	AC079804	77000..79000	No Met, 2 FS, 1 stop
hs7M1-24	P	AC079882	68000..70000	No Met, 2 FS, 1 stop
hs7M1-25	P	AC027522	39500..41500	1 FS, 1 stop
hs7M1-26	P	AC079882	207000..209000	3 FS, 1 stop
hs7M1-27	P	AC073264	60000..62000	1 FS
hs7M1-28	C	AC073079	9000..12000	
hs7M1-29	C	AC004977	101000..103000	
hs7M1-30	C	AC073079	69000..71000	
hs7M1-31	P	AC073264	37000..39000	1 FS

hs8M1-1	P	AC025126	94000..96000	4 stops, 1 FS
hs8M1-2	P	AC000385	125709..126761	2 stops, no met
hs8M1-3	P	AC000385	138500..140500	4 stops
hs8M1-4	P	AC000385	145952..147061	2 FS, 4 stops
hs8M1-5	P	AC015480	45000..47000	1 stop, goes into gap

hs9M1-1	C	AC006313	2200..3000	
hs9M1-2	C	AC006313	16000..17000	
hs9M1-3	C	AC006313	49000..50000	
hs9M1-4	C	AC006313	62500..63500	
hs9M1-5	C	AC006313	111000..112000	
hs9M1-6	C	AC006313	136000..137000	3 starts

hs9M1-7	C	AC006313	176000..177000	
hs9M1-8	C	AL135841		
hs9M1-9	C	AL135841		
hs9M1-10	P	AL135841		1 FS
hs9M1-11	P	AL138834		
hs9M1-12	P	AL138834		
hs9M1-13	P	AL138834		1 FS, 2 stops
hs9M1-14	F	AL138834		
hs9M1-15	C	AL353767	196500..198500	
hs9M1-16	C	AL162254	19000..21000	
hs9M1-17	C	AL162254	57500..59500	
hs9M1-18	C	AL162254	67000..69000	
hs9M1-19	P	AL162254	92500..94500	2 FS
hs9M1-20	C	AL162254	115000..117000	
hs9M1-21	C	AL162254	167500..169500	
hs9M1-22	C	AL359512	30000..32000	
hs9M1-23	C	AL354661	61000..63000	
hs9M1-24	P	AL354862	1..2000	1 FS,2 stops
hs9M1-25	P	AC009900	80500..82500	3 FS, 2 stop
hs9M1-26	P	AC009900	177500..179205	no met, stops
hs9M1-27	P	AC010612	118000..120000	3 FS, stops
hs9M1-28	C	AC072059	36500..38500	
hs9M1-29	C	AC072059	58500..60500	
hs9M1-30	C	AC072059	68000..70000	
hs9M1-31	C	AC072059	101500..103500	
hs9M1-32	P	AC072059	122500..124500	1 FS
hs9M1-33	C	AC072059	131000..133000	
hs9M1-34	C	AC072059	137000..139000	
hs9M1-35	C	AC072059	149500..151500	
hs9M1-36	P	AC072059	162000..164500	Alu insertion,1 FS, stops
hs9M1-37	P	AC072059	189000..191000	1 stop
hs9M1-38	C	AC009594	114000..116000	
hs9M1-39	P	AC009594	135932..136865	
hs10M1-1	P	AL358394	199500..201500	1 stop
hs10M1-2	C	AC011879	42500..44500	
hs10M1-3	PF	AC011879	10000..12500	1 FS,1 stop, goes into gap
hs10M1-4	P	AL157391	6000..8000	4 FS,2 stops
hs10M1-5	P	AL157391	14500..16500	stops
hs10M1-6	P	AL360083	173500..175500	1 FS,2 stops, missing last domains
hs11M1-1	P	AC016856	83000..85000	3 FS
hs11M1-2	C	AC016856	108000..110000	
hs11M1-3	P	AC068339	154000..155482	2 stops, goes into gap
hs11M1-4	P	AC022289	30500..32500	5 FS
hs11M1-5	C	AC027239	116500..118500	
hs11M1-6	C	AP001998	11500..13500	
hs11M1-7	P	AC027239	135000..137000	1 stop

hs11M1-8	P	AC023080	500..2500 1 FS, 1 stop	
hs11M1-9	C	AC023080	121000..123000	
hs11M1-10	C	AC023080	51000..53000	
hs11M1-11	PF	AC023080	141000..143000	1 FS, missing start
hs11M1-12	C	AC023080	161500..163500	
hs11M1-13	C	AC026975	38000..40000	
hs11M1-14	C	AC026975	95000..97000	
hs11M1-15	C	AC026975	140000..142000	
hs11M1-16	P	AC026975	158000..160000	3 FS, 1 stop
hs11M1-17	P	AC005729	81000..83000	2 FS, 1 stop
hs11M1-18	P	AC025730	135000..137000	1 FS
hs11M1-19	P	AC021427	11000..13000	1 FS
hs11M1-20	P	AC021427	55500..57500	1 FS
hs11M1-21	C	AC021427	78500..80500	
hs11M1-22	C	AC021427	107500..109500	
hs11M1-23	P	AC022998	73000..75000	1 FS, 1 stop
hs11M1-24	P	AC021427	27500..29500	
hs11M1-25	C	AC044810	53500..55500	
hs11M1-26	C	AC044810	86000..88000	
hs11M1-27	P	AC060812	175500..177500	2 stops, no met
hs11M1-28	P	AP001998	20500..24000	L1 insertion, 10 stops
hs11M1-29	C	AC002555	49657..50631	
hs11M1-30	C	AF137396	5387..6325	AAD29425.1
hs11M1-31	P	AF137396	13990..14962	-
hs11M1-32	C	AF137396	27728..28660	AAD29426.1
hs11M1-33	P	AC002555	26414..27316	
hs11M1-34	P	AC002555	12222..13091	
hs11M1-35	P	AC002555	10218..11147	
hs11M1-36	P	AC000378	55912..56936	
hs11M1-37	C	AC018700	1415..2374	
hs11M1-38	C	AC018700	23342..24409	
hs11M1-39	C	AC018700	31182..32114	
hs11M1-40	P	AC018700	88726..89666	
hs11M1-41	P	AC016856	139500..141500	
hs11M1-42	PF	AC018700	136-138	goes into a gap currently
hs11M1-43	C	AC018700	151916..152893	
hs11M1-44	C	AC018700	200871..201815	
hs11M1-45	P	AC017103	5500..7500	
hs11M1-46	C	AC017103	33500..35500	
hs11M1-47	C	AC017103	44500..46500	
hs11M1-48	C	AC017103	79000..81000	
hs11M1-49	C	AC017103	122500..124500	
hs11M1-50	C	AF321237	49500..51500	
hs11M1-51	P/C	AC019108	10500..12500	1 FS
hs11M1-52	F	AC019108	21000..23000	
hs11M1-53	PF	AC019108	52500..54500	
hs11M1-54	C	AC019108	56500..58500	
hs11M1-55	P	AC019108	108500..110500	

hs11M1-56	C	AC019108	128000..130000	
hs11M1-57	C	AP002407	155000..157000	
hs11M1-58	P	AC019108	152500..154500	1 FS
hs11M1-59	C	AP002407	60000..62000	
hs11M1-60	C	AP002407	57000..59000	
hs11M1-61	C	AP003034	40500..42500	
hs11M1-62	P/C	AC022882	62500..64500	
hs11M1-63	P/C	AC022802	93500..95500	1 FS
hs11M1-64	C	AC022882	100000..102000	
hs11M1-65	P	AC022882	106500..108500	1 FS, no met
hs11M1-66	C	AC022882	170500..172500	
hs11M1-67	C	AC022882	156500..158500	
hs11M1-68	C	AC022882	128500..130500	
hs11M1-69	P	AP003034	122000..124000	1 stop
hs11M1-70	P	AC022802	146500..148500	2 FS, 4 stop
hs11M1-71	P	AC027641	51000..53000	
hs11M1-72	P	AC027641	94000..96000	2 FS, stops
hs11M1-73	P	AC036111	7000..9000	no met
hs11M1-74	C	AC021530	130000..132000	
hs11M1-75	C	AC021530	73000..75000	
hs11M1-76	PF	AC036111	42000..44000	incomplete
hs11M1-77	P	AC036111	66500..68500	3 FS
hs11M1-78	F	AC036111	77000..79000	incomplete
hs11M1-79	C	AC021530	111000..113000	
hs11M1-80	C	AC036111	129000..131000	
hs11M1-81	P	AC021530	97500..99500	
hs11M1-82	C	AC021530	137500..139500	
hs11M1-83	C	AC021530	149000..151000	
hs11M1-84	PF	AC036111	181000..end	incomplete
hs11M1-85	P	AP002517	46500..48500	
hs11M1-86	P	AP001803	36000..38000	
hs11M1-87	P/C	AP001803	25500..27500	1 FS
hs11M1-88	C	AP001803	28500..30500	
hs11M1-89	C	AC037472	88500..90500	
hs11M1-90	P	AC037472	93500..95500	2 FS, stops
hs11M1-91	C	AP002517	62000..64000	
hs11M1-92	P	AP002517	78000..80000	1 FS
hs11M1-93	P	AP002517	82500..84500	1 FS
hs11M1-94	C	AP002517	168500..170500	
hs11M1-95	C	AC037472	170500..172500	
hs11M1-96	C	AP000818	10500..12500	
hs11M1-97	P	AP000818	22000..24000	3 FS
hs11M1-98	C	AP000818	57000..59000	
hs11M1-99	P/C	AP000818	86000..88000	no met
hs11M1-100	P/C	AP000818	120000..122000	
hs11M1-101	C	AP000825	6000..8000	
hs11M1-102	C	AP000868	21500..23500	
hs11M1-103	P	AP000825	45500..47500	

hs11M1-104	P	AP000868	128000..130000	1 FS, no met
hs11M1-105	C	AC083958	127000..129000	
hs11M1-106	C	AC083958	31500..33500	
hs11M1-107	PF	AP000916	16000..18000	start only, no met
hs11M1-108	C	AC083958	154500..156500	
hs11M1-109	P	AC083958	83500..85500	
hs11M1-110	C	AC083958	15000..17000	
hs11M1-111	P	AP000916	155000..157000	1 FS
hs11M1-112	P	AP000916	31000..33000	1 stop
hs11M1-113	C	AP002512	153000..155000	
hs11M1-114	C	AP001112	13500..15500	
hs11M1-115	P	AP002517	93000..95000	1 stop
hs11M1-116	C	AP002512	60000..62000	
hs11M1-117	P	AP001112	60000..62000	1 FS
hs11M1-118	P	AP001112	69500..71500	1 FS, 2 stop
hs11M1-119	P	AP001112	81500..83500	1 FS
hs11M1-120	C	AP001112	92000..94000	
hs11M1-121	P	AP001112	104000..106000	1 stop
hs11M1-122	P	AP001803	75500..77500	
hs11M1-123	P	AP001803	105000..107000	1 FS
hs11M1-124	P	AP001803	87000..89000	
hs11M1-125	C	AP001804	17500..19500	
hs11M1-126	C	AP001804	31500..33500	
hs11M1-127	P	AP001804	37000..39000	3 stops
hs11M1-128	P	AP001804	48000..50000	5 FS, disrputed MAYDRYVAIC
hs11M1-129	P	AP001803	60500..62500	1 stop
hs11M1-130	C	AP001804	132500..134500	
hs11M1-131	C	AP001804	142000..144000	
hs11M1-132	P	AP001804	152000..154000	1 stop
hs11M1-133	P	AP001804	67000..69000	1 FS, 3 stops
hs11M1-134	P	AP001804	127000..129000	2 FS, 2 stops
hs11M1-135	C	AP001884	11000..13000	similar to -57
hs11M1-136	PF	AP001884	90000..92000	1 FS, goes into gap
hs11M1-137	C	AP001884	129000..131000	
hs11M1-138	C	AP002407	132500..134500	
hs11M1-139	P	AC022289	22500..24500	FS, incomplete
hs11M1-140	P	AC027239	178000..180000	no met
hs11M1-141	C	AC022891	1..2000	
hs11M1-142	P	AC022891	16500..18500	5 stops, 1 FS, missing motifs
hs11M1-143	C	AC022891	22500..24500	
hs11M1-144	C	AP003034	17000..19000	
hs11M1-145	C	AC022891	57500..59500	
hs11M1-146	C	AC068339	40000..42000	
hs11M1-147	P	AP003033	140000..142000	1 FS
hs11M1-148	C	AP003033	86500..88500	
hs11M1-149	P	AC068339	79000..81000	no met
hs11M1-150	P	AP003033	1500..3500	1 stop
hs11M1-151	P	AC022891	156500..158500	1 stop

hs11M1-152	C	AC022891	167000..169000	
hs11M1-153	P	AP000723	79000..81000	
hs11M1-154	C	AC026076	42000..44000	
hs11M1-155	P	AP000723	25500..27500	2 FS, no met
hs11M1-156	P	AP000723	58000..60000	1 FS, 2 stops
hs11M1-157	P	AC026076	114500..116500	1 stop
hs11M1-158	C	AP000723	87000..89000	
hs11M1-159	C	AP001998	99000..101000	
hs11M1-160	P	AP001998	116500..118500	no met
hs11M1-161	C	AP001998	148000..150000	
hs11M1-162	C	AC027239	90000..92000	
hs11M1-163	C	AC027239	102500..104500	
hs11M1-164	P	AP001998	36000..38000	1 stop, no met
hs11M1-165	P	AP001998	54500..56500	1 FS
hs11M1-166	P	AP001998	81500..83500	3 stops, 1 FS
hs11M1-167	C	AP002512	86500..88500	
hs11M1-168	P	AP002512	91000..93000	1 FS
hs11M1-169	P	AP002512	110000..113000	2 FS
hs11M1-170	P	AP002512	122000..124000	1 FS, 2 stops
hs11M1-171	C	AP002512	128000..130000	
hs11M1-172	C	AP002517	26000..28000	95% sim to hs11M1-91
hs11M1-173	P	AP003033	30000..32000	3 stops, 1 FS, missing start motifs
hs11M1-174	C	AP003033	68000..70000	Q13606 U56420 OLF1
hs11M1-175	P	AP003033	105000..107000	3 stops
hs11M1-176	C	AP003033	134500..136500	
hs11M1-177	C	AP003033	158500..160500	
hs11M1-178	P	AP003034	70500..72500	1 stop
hs11M1-179	P	AP003034	141000..143000	missing motifs
hs11M1-180	C	AP003034	153000..155000	
hs11M1-181	P	AC019093	46000..48000	1 FS, no met, 1 stop
hs11M1-182	C	AC019093	63000..65000	
hs11M1-183	C	AC019093	119000..121000	
hs11M1-184	P	AC019093	128500..130500	1 stop
hs11M1-185	C	AC019093	154000..156000	
hs11M1-186	P	AC019093	21500..23500	
hs11M1-187	C	AC069371	51000..53000	
hs11M1-188	P	AC019093	57500..59500	1 stop, 1 FS
hs11M1-189	C	AC019093	109000..111000	
hs11M1-190	P/C	AC040925	56000..58000	1 stop
hs11M1-191.1	P	AC040925	120000..122000	1 FS, goes into gap
hs11M1-192	P	AC040925	151500..153500	3 FS, 4 stops
hs11M1-193	P	AP000435	53000..55000	1 FS, no met
hs11M1-194	C	AP000435	119500..121500	
hs11M1-195	P	AP000435	79000..81000	1 FS
hs11M1-196	P	AP000435	105000..107000	7 stops
hs11M1-197	P	AP000435	110000..112000	2 stops, 2 F
hs11M1-198	C	AP000435	28500..30500	
hs11M1-199	C	AP002345	10000..12000	

hs11M1-200	C	AP002345	30000..32000	
hs11M1-201	P	AP002345	45000..47000	3 stops, 1 FS
hs11M1-202	P	AP002345	66500..68500	1 FS
hs11M1-203	C	AP002345	114000..116000	
hs11M1-204	C	AC021809	19000..21000	
hs11M1-205	P	AC021809	35000..37000	4 stops, 2 FS, missing start
hs11M1-206	P	AC021809	117500..119500	1 stop, 1 FS, missing start
hs11M1-207	P	AP002780	97500..98500	1 stop, 1 FS, no end(-208P)
hs11M1-208	P	AP002780	97500..99500	2 FS, 3 stops
hs11M1-209	C	AP002780	7000..9000	
hs11M1-210	P	AP002780	143000..145000	
hs11M1-211	P	AC004923	54500..56500	1 FS, 3 stops, glv type
hs11M1-212	P	AC079973	120000..122000	1 FS
hs11M1-213	F	AC079973	1..1000	goes into gap
hs11M1-214	P	AC079973	184000..186000	no start, 1 FS, (TTC)n repeat insert
hs11M1-215	P	AP000719	55500..57500	7 FS, 2 stops
hs11M1-216	P	AP000719	45500..47500	1 FS, 2 stops
hs11M1-217	P	AP000867	17500..19500	2 stops, 1 FS
hs11M1-218	P	AP000867	79500..81500	3 stops
hs11M1-219	C	AP003175	59000..61000	
hs11M1-220	P	AP003385	15000..17000	5 FS, no met
hs11M1-221	P	AP003385	28000..30000	1 stop, no met
hs11M1-222	P	AC009867	16500..18500	6 FS
hs11M1-223	P	AC009867	36000..38000	2 FS
hs11M1-224	P	AP002965	5000..7000	4 FS, 2 stops
hs11M1-225	P	AP002965	18500..20500	1 FS
hs11M1-226	P	AP002965	36000..38000	4 FS
hs11M1-227	C	AP002965	49500..51500	
hs11M1-228	C	AP002965	101000..103000	
hs11M1-229	P	AP002965	159000..161000	1 stop, short last motif
hs11M1-230	P	AC009545	48000..52000	Disrupted by LTR5
hs11M1-231	P	AC009545	92000..94000	1 Fs, 1 stop
hs11M1-232	C	AC009545	112500..114500	8 potential met
hs11M1-233	C	AC009545	137000..139000	
hs11M1-234	P	AC009545	152500..154500	
hs11M1-235	P	AC009545	160000..162000	
hs11M1-236	C	AC009758	136000..138000	
hs11M1-237	C	AC009758	14500..16500	
hs11M1-238	P	AC009758	28500..30500	
hs11M1-239	C	AC009758	54000..56000	
hs11M1-240	P	AC009758	72000..74000	no met
hs11M1-241	P	AC009758	85500..87500	1 FS, 1 stop
hs11M1-242	P	AC009758	125000..127000	1 FS
hs11M1-243	C	AC009642	50500..52500	
hs11M1-244	P	AC009642	5500..7500	1 FS
hs11M1-245	C	AC009642	31500..33500	
hs11M1-246	PF	AC009642	67500..69500	missing 2 end motifs
hs11M1-247	P	AC009642	132000..134000	1 FS, 1 stop

hs11M1-248	C	AC009642	74000..76000	
hs11M1-249	C	AC009642	139500..141500	
hs11M1-250	C	AC010930	9000..11000	
hs11M1-251	P	AC010930	35000..37000	1 FS
hs11M1-252	C	AC010930	49000..51000	
hs11M1-253	P	AC010930	74500..76500	1 FS, small (simple repeat) insertion
hs11M1-254	C	AC010930	105000..107000	
hs11M1-255	C	AC010930	147000..148300	
hs11M1-256	C	AC011647	30000..32000	
hs11M1-257	P	AC011647	44500..46500	2 FS, 2 stops
hs11M1-258	P	AC011647	50000..52000	2 stops
hs11M1-259	P	AC011647	59500..61500	
hs11M1-260	C	AC011647	67000..69000	
hs11M1-261	P	AC011647	84000..86000	2 FS
hs11M1-262	C	AC011647	123000..125000	
hs11M1-263	P	AC011647	143500..145500	1 FS
hs11M1-264	P	AC026083	1000..3000	2 FS
hs11M1-265	P	AC026083	12500..14500	1 FS
hs11M1-266	C	AC026083	74500..76500	
hs11M1-267	C	AC026083	87000..89000	
hs11M1-268	C	AC026083	121500..123500	
hs11M1-269	C	AC026083	138000..140000	
hs11M1-270	P	AC026083	157500..159500	1 FS
hs11M1-271	F	AC020597	3500..5500	goes into gap
hs11M1-272	P	AC020597	13500..15500	1 FS
hs11M1-273	C	AC020597	25500..27500	
hs11M1-274	P	AC020597	31500..33500	2 FS
hs11M1-275	P	AC020597	55500..57500	1 FS, 2 stops
hs11M1-276	C	AC020597	63000..65000	
hs11M1-277	C	AC020597	76000..78000	
hs11M1-278	C	AC020597	90000..92000	
hs11M1-279	P	AC020597	108000..110000	1 stop
hs11M1-280	P	AC020597	132000..134000	3 stops
hs11M1-281	C	AC020597	141000..143000	
hs11M1-282	P	AC020597	151000..153000	3 FS, 4 stops
hs11M1-283	C	AC020597	160500..162500	
hs11M1-284	P	AC020597	177000..179000	1 stop
hs11M1-285	C	AC011711	2000..4000	
hs11M1-286	C	AC011711	9000..11000	
hs11M1-287	P	AC011711	27000..29000	3 FS
hs11M1-288	C	AC011711	34500..36500	
hs11M1-289	P	AC011711	40000..42000	2 FS
hs11M1-290	C	AC011711	68000..70000	
hs11M1-291	P	AC011711	84000..86000	2 FS
hs11M1-292	C	AC011711	95000..97000	
hs11M1-293	P	AC011711	129500..131500	1 FS
hs11M1-294	C	AC011711	147000..149000	
hs11M1-295	C	AC011711	113000..115000	

hs11M1-296	C	AP002826	86500..88500	
hs11M1-297	P	AP002826	115500..117500	1 FS, 2 STOPS
hs11M1-298	C	AP002826	68500..70500	
hs11M1-299	P	AP002509	18000..20000	1 FS
hs11M1-300	P	AP002509	98000..100000	1 FS
hs11M1-301	P	AP001521	34500..36500	1 FS, 1 STOP
hs11M1-302	C	AP001521	5500..7500	
hs11M1-303	P	AP001521	54000..56000	
hs11M1-304	P	AP000629	51000..53000	4 FS
hs11M1-305	C	AC025249	16500..18500	
hs11M1-306	C	AC025249	25500..27500	
hs11M1-307	C	AC025249	38000..40000	
hs11M1-308	C	AC025249	84000..86000	
hs11M1-309	P	AC025249	139500..141500	1 STOP
hs11M1-310	P	AC025249	157000..159000	1 FS
hs11M1-311	P	AC021935	62000..64000	1 FS
hs11M1-312	P	AC021935	39000..41000	1 FS
hs11M1-313	C	AC021935	80000..82000	
hs11M1-314	C	AC021935	109500..111500	
hs11M1-315	P	AC019088	10000..12000	4 FS
hs11M1-316	C	AC019088	16500..18500	
hs11M1-317	C	AC019088	1500..3500	
hs11M1-318	C	AC019088	34000..36000	
hs11M1-319	C	AC019088	47000..49000	
hs11M1-320	C	AC019088	70000..72000	
hs11M1-321	C	AC019088	77500..79500	
hs11M1-322	C	AC019088	98000..100000	
hs11M1-323	P	AC019088	114500..116500	2 FS
hs11M1-324	P	AC019088	124000..126000	2 FS, 1 STOP, MISSING START DOMAINS
hs11M1-325	C	AC019088	148500..150500	
hs12M1-1	P	AC022207	41500..43500	1 stop
hs12M1-2	C	AC022207	55000..57000	
hs12M1-3	P	AC022207	105500..107500	1 FS,stop
hs12M1-4	P	AC022207	22000..24000	2 FS
hs12M1-5	C	AC022207	163500..165500	
hs12M1-6	P	AC068994	382500..384500	4 FS, 1 stop
hs12M1-7	P	AC009779	124000..126000	1 FS
hs12M1-8	C	AC009779	184500..186500	
hs12M1-9	P	AC008035	142000..144000	fragment, 1 FS, 1 stop
hs12M1-10	P	AC009775	73000..75000	3 FS,2 stops
hs12M1-11	P	AC009775	5500..9500	2 stops, large insertion unique DNA, SINEs)
hs12M1-12	C	AC024257	6500..8500	
hs12M1-13	P	AC083933	15500..17500	1 FS, goes into gap, 2 stops
hs12M1-14	P	AC090115	67500..69500	1 FS,1 stop
hs12M1-15	P	AC090115	76500..78500	no met, 6 stops
hs12M1-16	P	AC090115	106000..108000	3 stops, 2 FS, no final motif
hs12M1-17	P	AC090115	20000..22000	1 stop

hs12M1-18	P	AC078864	128000..130000	3 FS, 1 stop
hs12M1-19	C	AC009779	161000..163000	

hs13M1-1	P	AL354833	1500..3500	2 FS, 2 stops
hs13M1-2	PF	AL138686	64000..66000	1 FS, no start, 2 stops,LINE repeat disrupts
hs13M1-3	PF	AC024458	101000..103000	2 FS, 1 stop, no start
hs13M1-4	P	AL353580	32500..34500	2 stops, 1 FS
hs13M1-5	P	AL353580	41000..43000	2 stops, no met

hs14M1-1	C	AE000658	75000..77000	
hs14M1-2	P	AE000658	107000..109000	1 stop
hs14M1-3	C	AE000658	170500..172500	
hs14M1-4	P	AE000658	175500..177500	1 FS
hs14M1-5	C	AE000658	139000..141000	
hs14M1-6	C	AL157687	145500..147500	
hs14M1-7	C	AL160314	102000..104000	
hs14M1-8	P	AL160314	167000..169000	1 FS, 1 stop
hs14M1-9	P	AC024399	16000..18000	1 FS, no starting motifs
hs14M1-10	C	AC024399	32000..34000	
hs14M1-11	P	AC024399	64000..66000	2 FS
hs14M1-12	C	AC024399	114500..116500	
hs14M1-13	C	AC024399	122500..124500	
hs14M1-14	C	AC024399	136500..138500	
hs14M1-15	P	AC024399	152500..154500	2 FS, 1 stop
hs14M1-16	C	AC024399	67000..69000	
hs14M1-17	C	AC024399	183500..185322	
hs14M1-18	C	AL359218	26500..28500	
hs14M1-19	P	AL163152	10500..12500	3 FS,6 stops
hs14M1-20	C	AL359218	53000..55000	
hs14M1-21	C	AL359218	72500..74500	
hs14M1-22	P	AL359218	84500..86500	1 FS
hs14M1-23	C	AL359218	111000..113000	
hs14M1-24	P	AL359218	130000..132000	1 FS,1 stop
hs14M1-25	C	AL359218	151000..153000	
hs14M1-26	PF	AL163152	60000..62000	no start, 2 FS, 7 stops
hs14M1-27	C	AL163152	84000..86000	
hs14M1-28	C	AL163152	110500..112500	
hs14M1-29	P	AL163152	146000..148000	3 FS
hs14M1-30	P	AL356019	69000..71000	2 stops, 1 FS
hs14M1-31	C	AL356019	83500..85500	
hs14M1-32	P	AL356019	89000..91000	1 stop
hs14M1-33	C	AL356019	102500..104500	
hs14M1-34	C	AL356019	57000..59000	
hs14M1-35	C	AL163636	72500..74500	
hs14M1-36	P	AL132827	58000..60000	3 FS, 5 stops
hs14M1-37	P	AL079307	91500..93500	3 stops
hs14M1-38	P	AL079307	97500..99500	2 stops, 2 FS
hs14M1-39	PF	AL079307	106000..108000	1 FS, only end 2 motifs

hs15M1-1	P	AC010760	25000..27000	1 FS, no Met
hs15M1-2	P	AC010760	37500..39500	1 FS, 2 stops
hs15M1-3	C	AC010760	61500..63500	
hs15M1-4	C	AC010760	75000..77000	
hs15M1-5	P	AC010760	106000..108000	1 stop
hs15M1-6	P	AC010760	150500..152500	1 stop, 1 FS
hs15M1-7	P	AC010760	172000..174000	1 FS
hs15M1-8	P	AC020679	33000..35000	2 FS
hs15M1-9	P	AC020679	41500..43500	no met, 1 FS
hs15M1-10	C	AC025234	70000..72000	
hs15M1-11	C	AC005143	2500..4500	

hs16M1-1	C	AJ003147	163476..164414	
hs16M1-2	P	AJ003147	174780..175778	No Met
hs16M1-3	C	AC068380	147000..149000	

hs17M1-1	P	AC007194	142000..143000	OR17-25
hs17M1-2	C	AC007194	142000..143000	OR17-24
hs17M1-3	PF	AC023106	1000..2000	1 FS, 1 stop, goes into gap
hs17M1-4	C	AC007194	56000..57000	OR17-2
hs17M1-5	C	AC007194	162000..163000	OR17-40
hs17M1-6	C	AC007194	406000..407000	OR17-31
hs17M1-7	C	AC007194	376000..377000	OR17-31
hs17M1-8	P	AC007194	352000..353000	OR17-210
hs17M1-9	C	AC007194	342000..343000	OR17-209
hs17M1-10	P	AC007194	314000..315000	OR17-208
hs17M1-11	C	AC007194	255000..256000	OR17-6
hs17M1-12	C	AC007194	238000..239000	OR17-7
hs17M1-13	C	AC007194	212000..214000	OR17-30
hs17M1-14	P	AC007194	187000..189000	2 FS, OR17-23
hs17M1-15	C	AC007194	175000..176000	OR17-228
hs17M1-16	P	AC007194	67000..69000	2 FS, OR17-1
hs17M1-17	C	AC007194	32000..35000	OR17-201
hs17M1-18	C	AC007194	20000..22000	OR17-93
hs17M1-19	C	AC005962	76500..78500	
hs17M1-20	C	AC005962	91000..93000	

hs18M1-1	P	AC025953	66500..68500	1 stop
hs18M1-2	P	AC025953	79500..81500	2 FS

hs19M1-1	C	AC005255	40752..41771	
hs19M1-2	C	AC011517	16000..18000	
hs19M1-3	C	AC010322	1..2000	
hs19M1-4	C	AC002988	93896..94343	
hs19M1-5	P	AC002988	38600..37677	1 FS
hs19M1-6	P	AC002988	9581..10510	no met
hs19M1-7	C	AC005255	12644..13606	

hs19M1-8	P	AC005255	65502..66414	1 FS
hs19M1-9	C	AC005255	54418..55347	
hs19M1-10	P	AC004659	23355..24339	2 stops,1FS
hs19M1-11	F	AC004659	1..476	goes into gap
hs19M1-12	C	AC004659	37792..38751	
hs19M1-13	C	L78442	4856..5773	
hs19M1-14	P	AC003956	29110..30022	no met, 2 FS
hs19M1-15	C	AC004510	1852..2808	
hs19M1-16	C	AC004597	2990..3937	
hs19M1-17	C	AC004597	16239..17289	
hs19M1-18	C	AC004794	6034..7101	
hs19M1-20	P	AC006271	14410..15367	no met, 1 FS, 1 stop
hs19M1-21	C	AC006271	24021..24957	
hs19M1-22	P	AC006271	45263..46252	
hs19M1-23	C	AC006271	61163..62179	
hs19M1-24	P	AC006271	88552..89488	1 FS, 1 stop
hs19M1-25	P	L78442	10000..12000	no met, 1 FS, 1 stop
hs19M1-26	P	L78442	21000..23000	7 FS, 4 stops
hs19M1-27	C	AC011464	39000..41000	
hs19M1-28	C	AC011464	98000..100000	
hs19M1-29	F	AC011464	103000..105000	missing 2 end motifs
hs19M1-30	C	AC011464	110000..112000	
hs19M1-31	C	AC011464	122000..124000	
hs19M1-32	C	AC011537	7516..8463	
hs19M1-33	C	AC011464	131000..133000	
hs19M1-34	PF	AC011464	142000..144000	1 FS, missing end motifs
hs19M1-35	PF	AC006271	26302..26526	(1 stop)
hs19M1-36	P	AC006271	76494..75515	no met, 1 FS
hs19M1-37	PF	AC006271	71835..71098	missing 2 start motifs, 2 stops
hs21M1-1	P	AP000181	48500..50500	4 FS, 1 stop
hs21M1-2	P	AP001465	1500..3500	2 FS
hs21M1-3	P	AP001465	38500..40500	2 FS
hs22M1-1	C	AP000534	32582..33529	
hsXM1-1	P	AL049734	500..2500	
hsXM1-2	C	AL049734	120000..121000	
hsXM1-3	P	AL109853	26000..28000	
hsXM1-4	P	AL109853	4000..6000	fragment, repeat insert(nnn)
hsXM1-5	P	AL135784	66000..68000	3 FS
hsXM1-6	P	AL135784	10000..12000	
hsXM1-7	P	AL355366	12000..14000	1 L1PA7 insert, 1 FS, 1 stop
hsXM1-8	P	AF277315	139000..141000	1 FS, no met, stops