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.

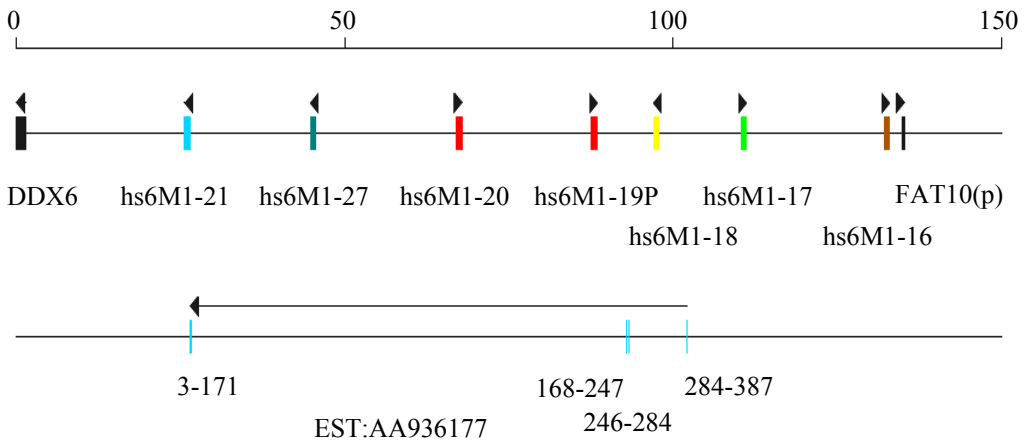


Figure 6.1: EST supported splicing in hs6M1-21. The top track shows the scale in Kb, whilst the middle track shows the exons found within this region of the human extended MHC. Exons are coloured differently according to which subfamily the OR gene belongs. (See figure 4.6). The third track shows positions 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.

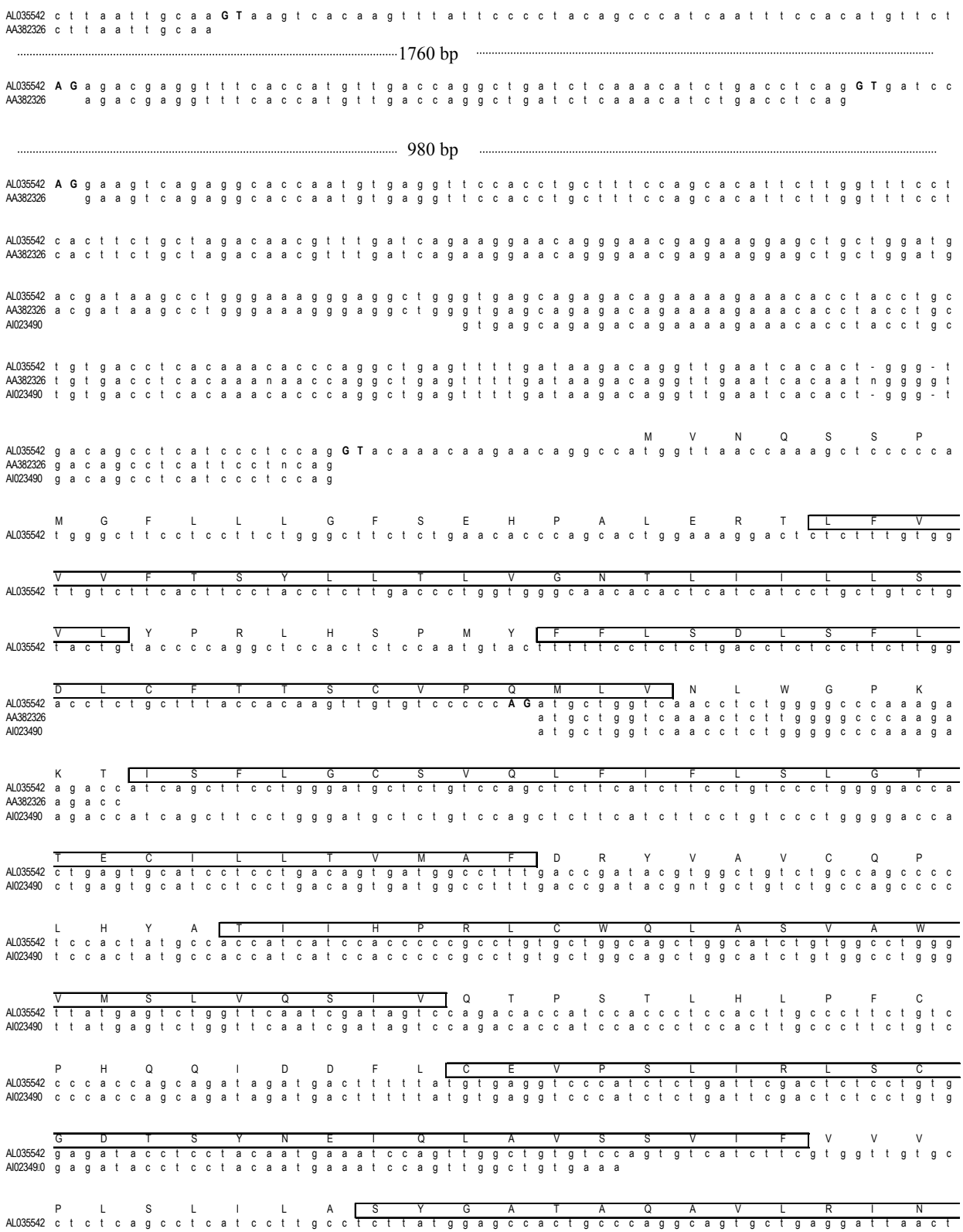

AL035542
AA382326
1760 bp
980 bp
AI023490

Figure 6.3a

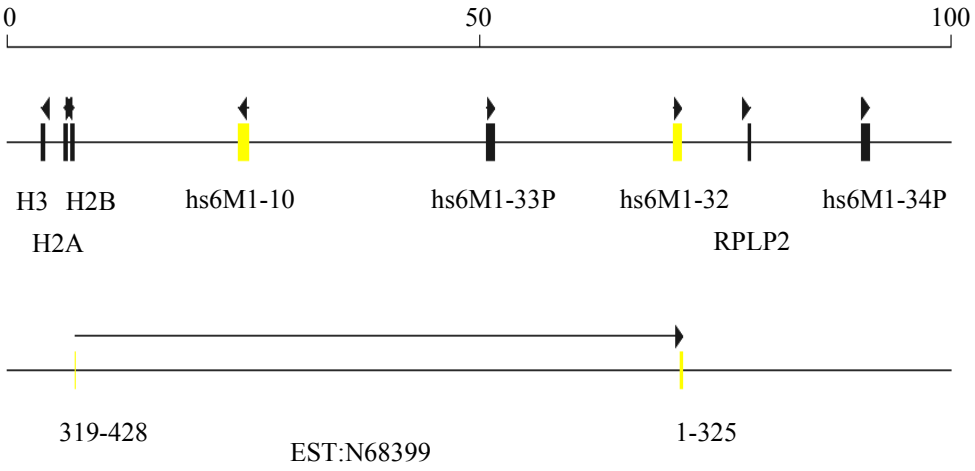


Figure 6.3b

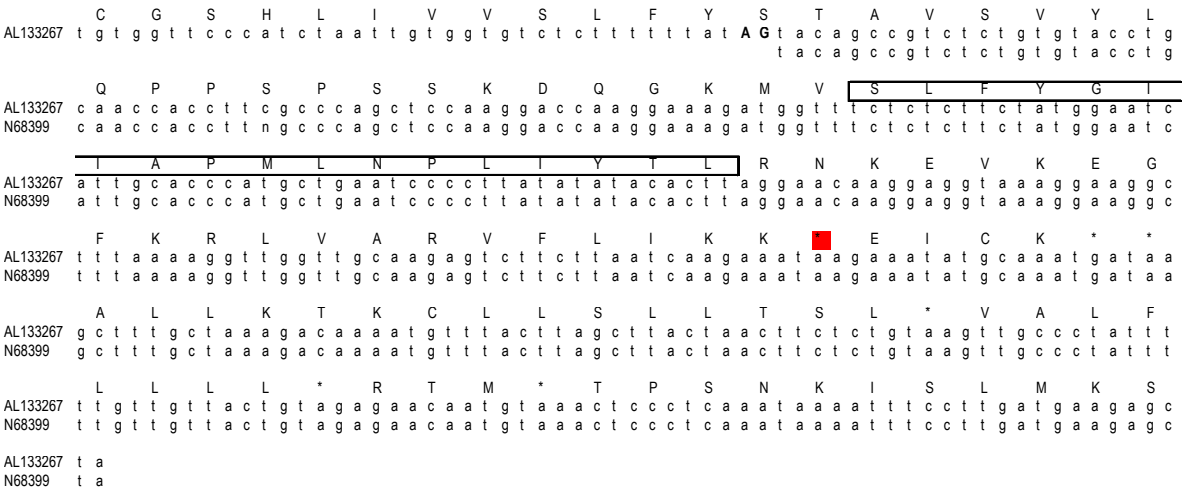


Figure 6.3a: EST associated with hs6M1-32. The top track shows the scale in Kb, whilst the middle track shows the exons found within this region of the human extended MHC. Exons are coloured differently according to which subfamily the OR gene belongs. (See figure 4.6). The third track shows positions 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

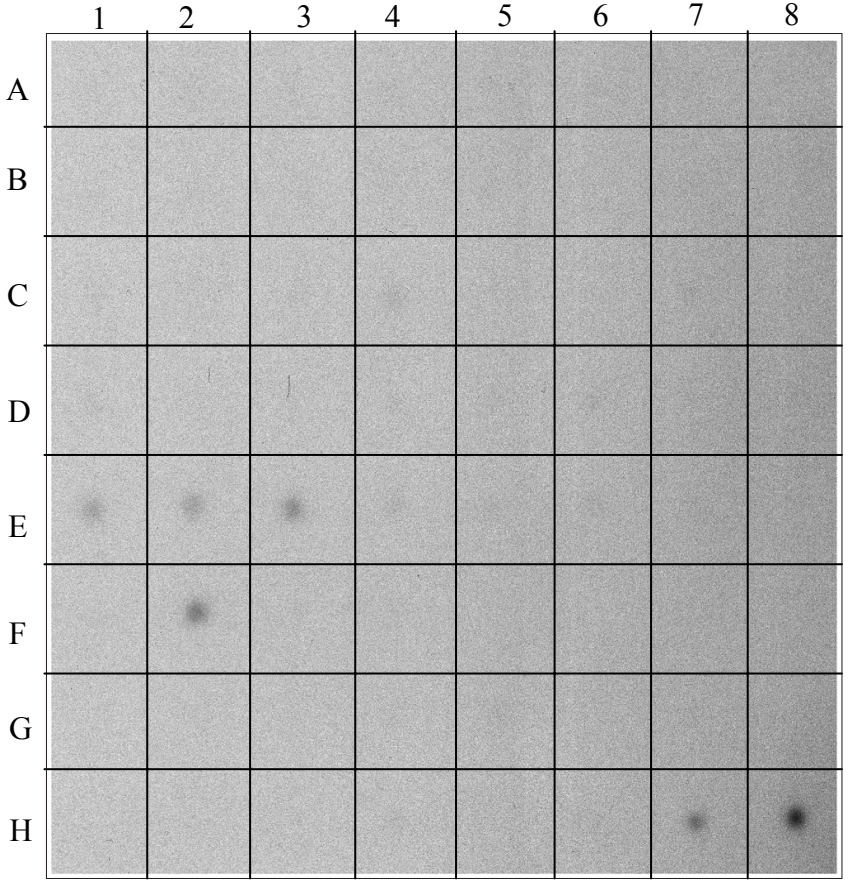


Figure 6.4b

	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

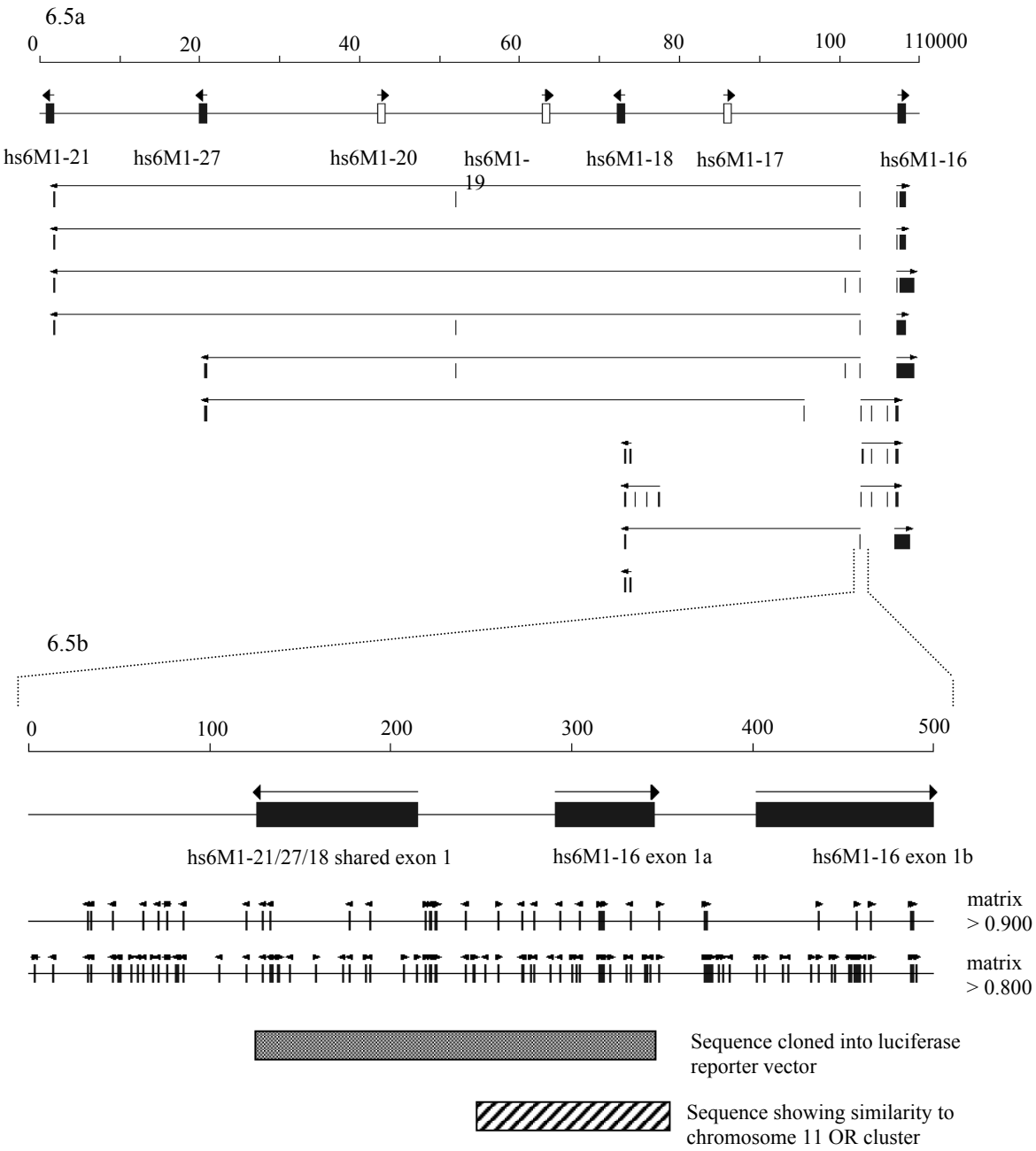


Figure 6.5a: Alternative splice forms of hs6M1-21, -27, -18 and –16. Other olfactory receptor genes within the region were not analysed by 5' RACE (hs6M1-19P, hs6M1-17) or were analysed by 5' RACE but no splicing was observed (hs6M1-20).
Figure 6.5b: Enlarged section showing shared exons and putative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

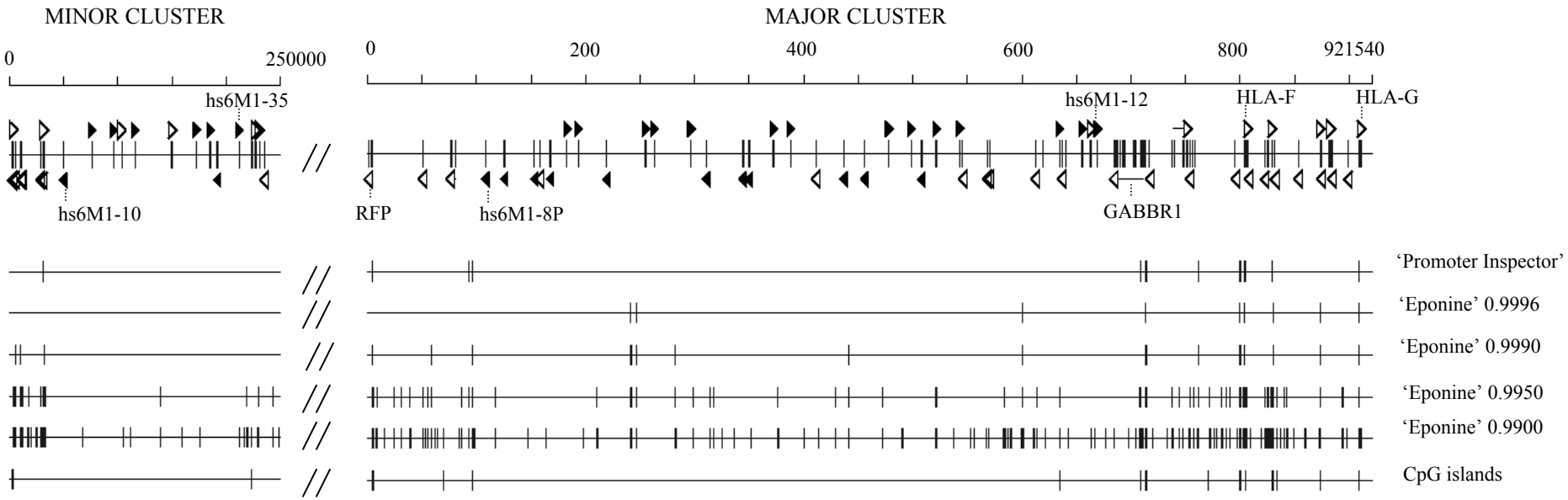


Figure 6.6: Promoter prediction within MHC-linked OR clusters (orientated in a telomere to centromere direction). The minor cluster is located approximately 1200 Kb telomeric to the major cluster. Olfactory receptor genes within the region are marked by black arrows. Genes referred to within the text, 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.

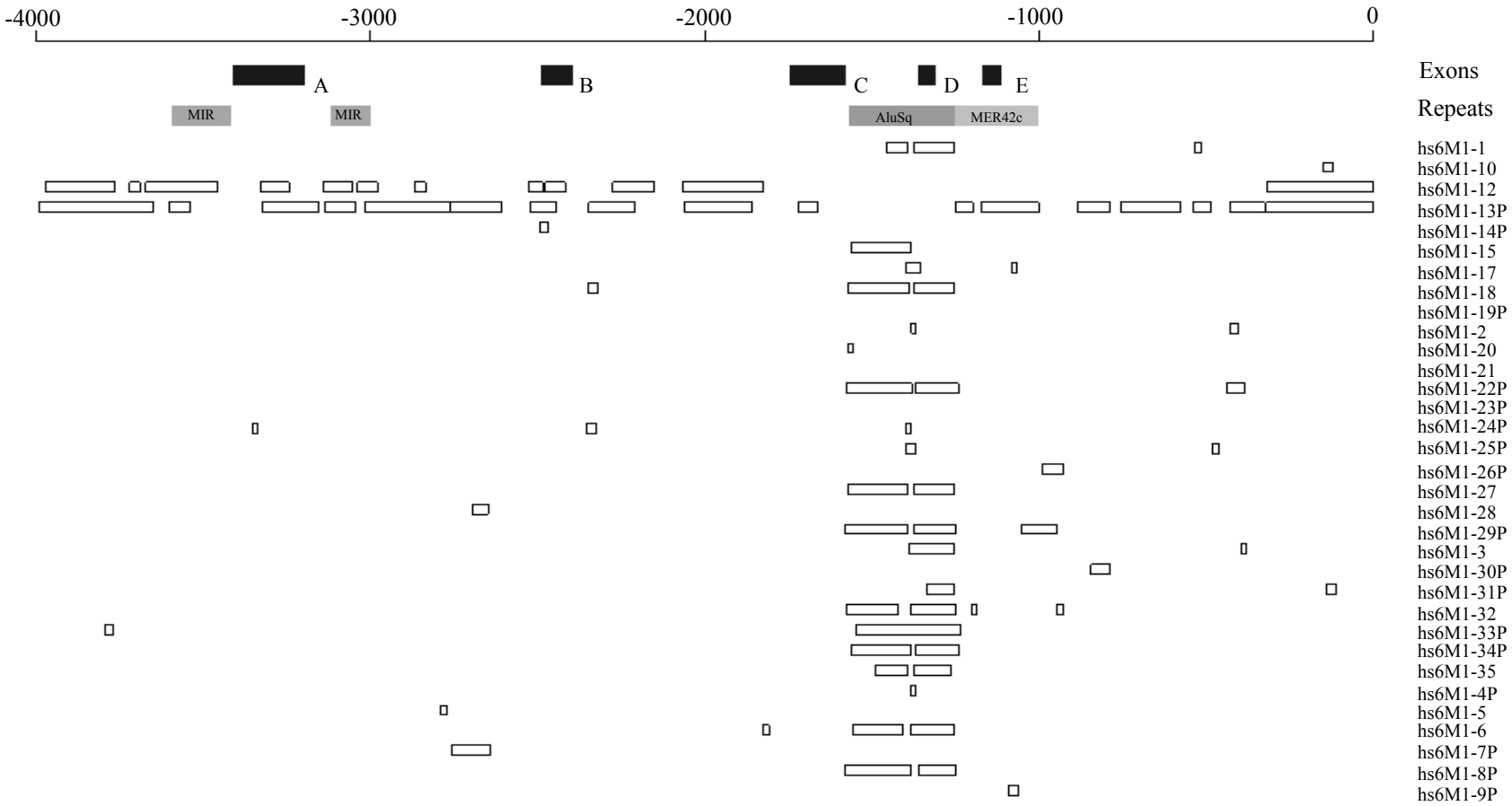


Figure 6.7: 'DBA' alignment of the 4 Kb upstream region of hs6M1-16. The scale shows base pairs distance from the proposed start codon of hs6M1-16. Exons and the repeats present within the 4 Kb region are plotted below the scale line. White blocks beneath the repeat line show blocks of sequence that 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.)

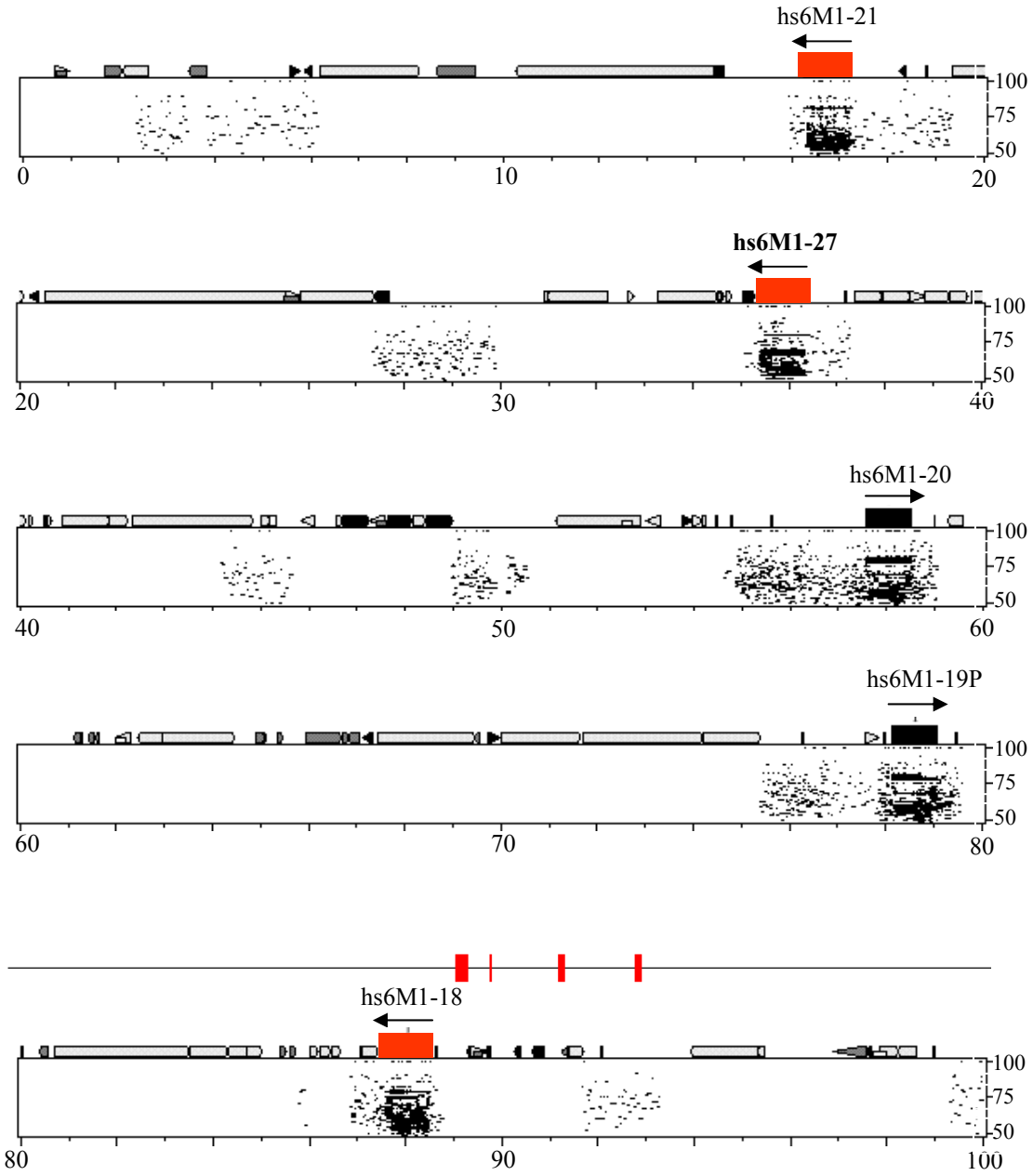
Hs6M1-16 was taken as an example result because information about the 5' UTR exons was available from the group in Berlin. This information meant 5 upstream exons could be compared to see if there was any conservation of these upstream of other MHC-linked ORs. As Figure 6.7 shows, there is little conservation 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.



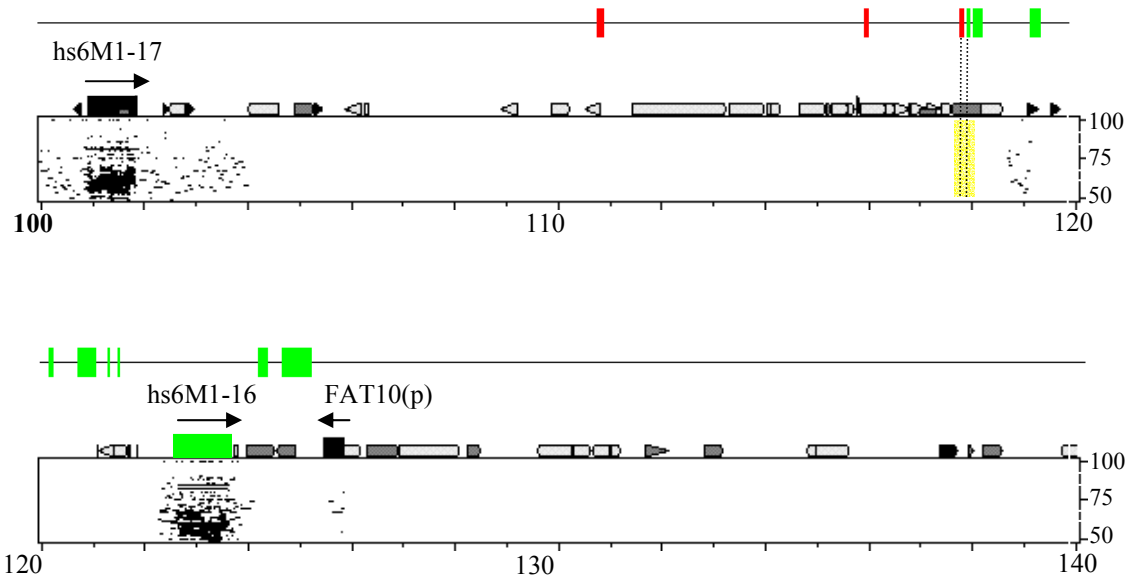


Figure 6.8: PIP plot of a region of the MHC-linked OR cluster. Coding exons and repeats are plotted at the top of each box. Information about untranslated exons (where known) is plotted in a separate track above the box. Untranslated exons that are part of the hs6M1-16 gene structure are plotted in green, 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.