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/*).

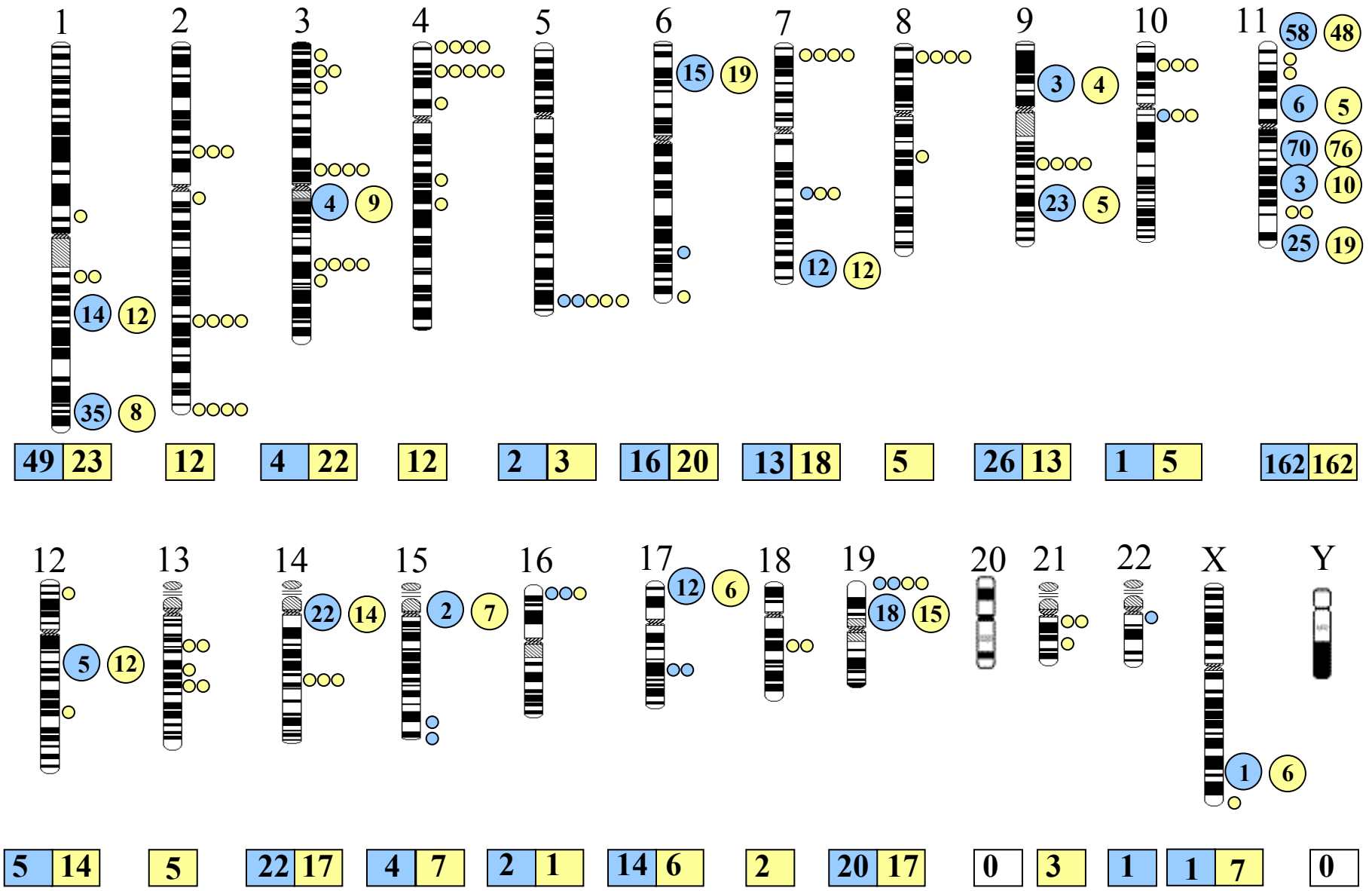
8.2. The human OR database ('ROLF') and the genomic location of OR genes.

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

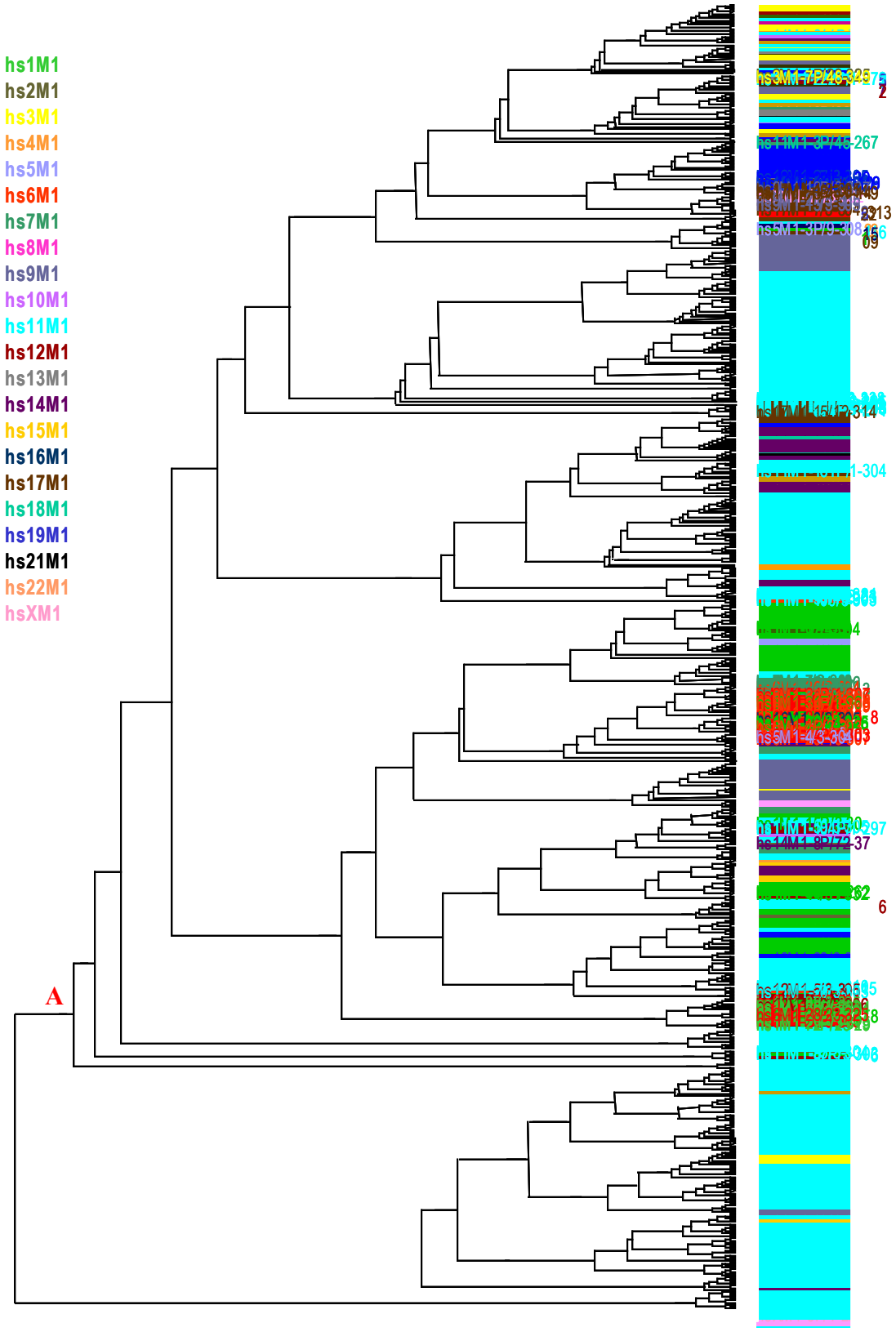


Figure 8.2. Phylogenetic tree of all ORs within the human genome. (Larger version of this tree, showing all gene names is included as a pullout at the back of the thesis.) Tree was constructed using a HMM alignment and the program 'QuickTree'. ORs on different chromosomes are shown in different colours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

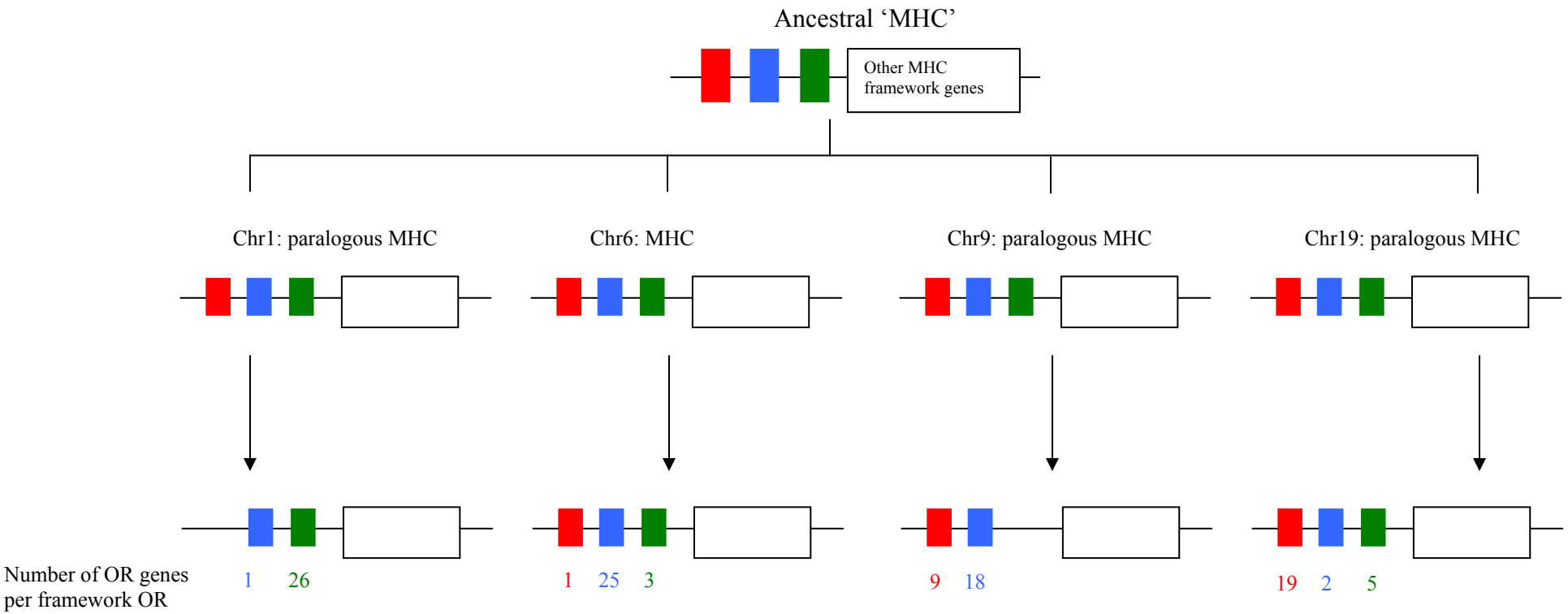


Figure 8.3: Proposed MHC-linked OR evolution on chromosome 1, 6 , 9 and 19. Clustering in the phylogenetic tree suggests 3 framework OR genes were duplicated alongside an ancestral 'framework MHC'. The 3 framework OR genes followed different evolutionary histories in the four chromosomal regions. For 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

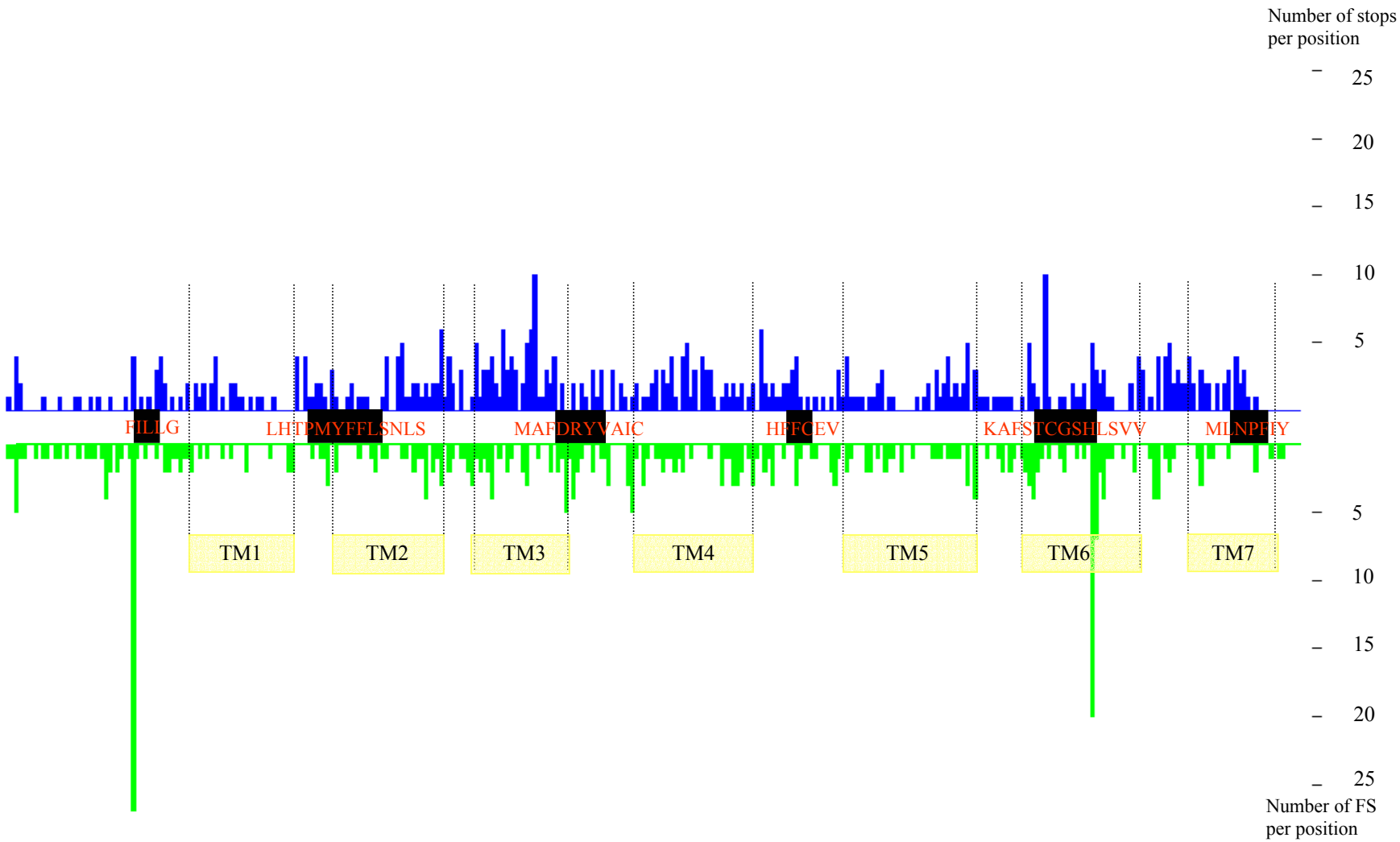


Figure 8.4: Distribution of stops and frameshifts in OR pseudogenes. The position of frameshifts and stop codons in each pseudogene was mapped relative to the position of 6 motifs within the protein. All of these positions were then applied to the chromosome 6 OR consensus protein sequence. The position 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.