Chapter 7

Conclusions and future work

7.1 Conclusions

This thesis presents the first systematic, unbiased survey of the entire human genome sequence to identify MHC paralogous genes with increasing levels of confidence and to determine their distribution. The genome-wide survey identified 791 MHC paralogous genes in the human genome with increasing levels (L0>L1>L2>L3) of confidence; of which 618 are L0-paralogues, 91 are L1-paralogues, 38 are L2-paralogues and 44 are L3-paralogues. It was found that over two-thirds of the MHC genes used in this study have paralogues located throughout the human genome and a total of one-third have paralogues with the highest level of confidence (L2- and L3-paralogues). The MHC genes with L2- and L3-paralogues are not restricted to just one region of the MHC and span almost the entire length of 6p22.2-p21.3, including genes within the most telomeric and centromeric regions; the extended class I and extended class II regions, respectively. Thus, indicating that the entire MHC region has been involved in the events giving rise to paralogous genes.

The study of the distribution of the MHC paralogous genes has confirmed that there are clusters of MHC paralogues located in the previously proposed regions on human chromosomes 1, 9 and 19. Almost 50% of the L2- and L3-paralogues are located within the regions 1q21.2-q25.3, 9q32-q34.3 and 19p13.3-p13.11. No further clusters of MHC paralogous genes were identified in the human genome, as postulated by Hughes and Pontarotti (2000). However, one of the most interesting, and novel, findings of this thesis is that the MHC paralogous genes are not confined just to these regions but there are paralogues scattered throughout the human genome.

In order to understand the relationship between the MHC and the other chromosomal regions containing clusters of MHC paralogues the region 9q32-q34.3 was mapped, sequenced and analysed. The characterisation of 9q32-q34.3 presented in this thesis represents the largest genomic region containing MHC paralogous genes to be characterised to-date. The comparison of 9q32-q34.3 and the MHC region has revealed a number of features common to both chromosomal segments. In total, 322 genes were identified within the 9q32-q34.3 region, which spans almost 24 Mb, corresponding to approximately one gene per 73 kb. The gene dense nature of 9q32-q34.3 is comparable to that of the MHC region. But this is just one feature these regions share. Other features shared by both regions include; they are associated with a number of diseases, the presence of structurally and functionally different genes and high GC content. One of the key differences between 9q32-q34.3 and the MHC region is that, although paralogues of 25 gene families located within the MHC region were identified on 9q32-q34.3, no HLA class I or class II-like genes were identified. Characterisation of the 1.7 Mb region of the paralogous region on 1q21-q22 confirmed that there was a cluster of HLA class I-like genes, termed the CD1 gene cluster (Shiina *et al,* 2001).

The existence of chromosomal regions containing clusters of duplicated genes is indicative of a common origin by large-scale duplication of either the whole-genome or of a block. In this thesis, I have identified three regions containing clusters of genes paralogous to those found within the MHC region, which is indicative of at least two rounds of large-scale duplication events. This is in support of the 2R hypothesis which, in its simplest form, assumes two rounds of whole-genome duplication early in the vertebrate lineage; the first in the common ancestor of all vertebrates and the second in a common ancestor of jawed vertebrates after its separation from jawless fish (reviewed by Wolfe, 2001). It is also in support of two rounds of duplication of a chromosomal segment, or block duplication. Either way, if they did have a common origin, it is expected that the regions are syntenic, which is not strictly obeyed.

Analysis of the gene order of the 40 MHC genes and the corresponding paralogues on 9q32-q34.3 revealed that the overall gene order is not conserved. Thus, if they did descend from a common ancestral region then they have experienced numerous rearrangements caused by evolutionary mechanisms, such as duplications, inversions, deletions and translocations, after its inception. There is evidence of the dynamic natures of the two regions, particularly of gene and segmental duplications, which would explain the observed differences in gene order. Another factor that would have had an impact on the present-day structure of the paralogous regions is the amount of time which has passed since their emergence. Thus, it is expected that the more time that has elapsed the more time evolution has had to act upon the sequence, which could result in a number of differences between the regions. In order to understand how and when the MHC paralogous genes emerged, the phylogenetic relationships of the MHC paralogues and orthologues were investigated.

Prior to my genome survey, the proposed genomic distribution of the MHC paralogous genes in the regions on 1, 9 and 19 was considered as evidence of past large-scale duplication (Kasahara, 1997; 1999a; 1999b). It was presumed that the regions emerged as part of two rounds of whole-genome duplication believed to have occurred early in the history of vertebrates, approximately 500 million years ago. The results of the phylogenetic studies presented in this thesis indicate that some of the MHC paralogues did emerge via large-scale duplication events early in the vertebrate lineage. This is consistent with the extensive evidence emerging in the literature to show that there was a burst of gene duplication during early chordate evolution (Pépusque *et al*, 1998; Wang and Gu, 2000; Miyata and Suga, 2001; Escriva *et al*, 2002; McLysaght *et al*, 2002; Panopoulou *et al*, 2003).

The paralogous gene families (BRD, PBX, and NOTCH) with four members all showed the expected (A,B)(C,D) tree topology that would be the result of two rounds of duplication, as proposed by the 2R hypothesis. Three member families (complement, RXR and tenascin) also indicate that there were two rounds of duplication, accompanied by gene loss. The timings of the duplication events as suggested by the 2R hypothesis are supported by the clustering of ‘key’ organisms in the phylogenetic trees. Thus, a single amphioxus orthologue is positioned at the base of each tree and there is at least one hagfish or lamprey orthologue clustered with the mammalian counterparts. Preliminary analysis of the MHC paralogous gene families, including PBX and tenascin, in the hagfish genome suggests that jawless fish have at least two paralogues (Flajnik and Kasahara, 2001). It is therefore expected that, upon complete sequencing of the hagfish genome, two paralogues should be identified for each MHC paralogue, thus supporting the proposed 2R hypothesis.

The only two member family (AIF) studied also shows that there was at least one round of duplication in the vertebrate lineage after the emergence of amphioxus. The existence of only two paralogues indicates that this family was only involved in one round of genome duplication (supporting a 1R hypothesis) or two paralogues have been lost after two genome duplication events (supporting the 2R hypothesis). It has been calculated that the average time before silencing of one member of a duplicate gene pair is approximately four million years in animals (Lynch and Conery, 2000), therefore the likelihood of these paralogues being lost since their emergence, approximately 500 million years ago, is high.

Phylogenetic studies of paralogous gene families with more than four members have shown the evolution of the MHC paralogous genes is much more complex. The MHC paralogous genes have emerged via recent duplication events that have resulted in the expansion of the paralogous gene families. For example, members of the CLIC paralogous gene family have been involved in a triplication event along with at least two other gene families not associated with the MHC region and a number of duplication events have given rise to members of the β-tubulin paralogous gene families within the last 25 million years of catarrhine (New World Monkeys and humans) evolution. These events have all resulted in MHC paralogues located outside the chromosome 1, 9 and 19 paralogous regions.

Gene and, potentially, genome duplication have played a central role in the evolution of the MHC paralogues. It has been shown that gene duplications are frequent events in the mammalian genome with an average duplication rate of approximately 1% per gene per million years (Lynch and Conery, 2000). But what happens after gene or genome duplication? The classical model predicts that one copy will be maintained under purifying selection whereas the other will accumulate mutations which will generally lead to the loss of function of that gene copy. In rare cases, new functions will be created and both duplicate genes will be conserved. In contrast, under the sub-functionalisation model both duplicates are preserved due to the partition of different functions between duplicates: the development of new functions is also possible. In order to understand how the paralogues have evolved since duplication the first step was to determine the function(s) of these genes in humans. This was addressed in this thesis by generating the expression profiles of the members of nine MHC paralogous gene families. The profiles were generated in a range of tissues corresponding to the major systems of the human body using several different approaches.

Comparison of the expression profiles of the MHC paralogous genes revealed that, in most cases, the paralogues have distinct expression profiles. However, there is still some overlap in the expression patterns of some members of the same paralogous gene family indicating levels of genetic redundancy. The absence of a modified or 'scoreable' phenotype following gene knock-out studies has alerted biologists to the presence of genes with overlapping, or redundant, functions. As paralogues have arisen from the same ancestral gene, and therefore may still have conserved gene structure, sequence, protein structure etc., they may act as a 'back-up' system in order to protect an organism against phenotypic changes and any deleterious effects of gene loss. However, further functional studies are necessary to confirm this prediction.

Analysis of the expression profiles revealed evidence of functional divergence. In particular, paralogues located on chromosome 1 appear to have a more specialised expression profile, for example BRDT and RXRG are restricted to expression in only a few tissues. Further investigation of genes on this chromosome is necessary before any conclusions can be made but data presented in this thesis indicates that genes on chromosomes 1 may have a more specialised function compared with genes elsewhere in the genome. It is apparent from this study that the precise mechanism by which they have evolved could not be determined based on the comparison of expression profiles alone. It is essential that we understand the function(s) of the ancestral genes as well as the functions of the human genes.

In conclusion, the paralogous genes on human chromosomes 1, 6, 9 and 19 emerged together as part of two large-scale duplication events early in the vertebrate lineage. From the emerging evidence in the literature and the findings of this thesis it can be argued that they were part of the two rounds of whole-genome duplication proposed by the 2R hypothesis. Further gene duplication events have also occurred resulting in the present-day genome organisation. The precise mechanism by which the MHC paralogues have functionally evolved is unclear. Overall, the MHC region has offered a unique opportunity for scrutinising genome evolution in vertebrates and, one thing that is clear from this thesis is that, the investigation of gene duplication events remains an exciting field of research. Further investigation of genes, genomes and the encoded proteins are necessary before we have a true understanding of the biological processes that shaped evolution and the complexity of our own species at the molecular level.

7.2 Future work

I would suggest that future work involving the MHC paralogous genes should follow a number of lines:

1. Comparative studies

The MHC paralogous genes have provided an exciting model to study genome evolution. It is also of particular interest regarding the origin of adaptive immunity. The rapidly accumulating information on the genomic organisations of the MHC regions in various model organisms is already providing insights into the long-term dynamics and evolution which have moulded the present day MHC and human genome (reviewed by Flajnik and Kasahara, 2001). Therefore, complete sequencing of the genomes of key species in the vertebrate lineage, namely amphioxus, hagfish and lamprey, will be invaluable for deciphering the evolution of the MHC paralogues and the genome as a whole.

2. Further analysis of the human genome

Initial analysis of the human genome identified significantly less genes than expected and it has been proposed that differential splicing of genes and the different encoded proteins play a crucial role in humans. The analysis of the splice variants identified in this thesis will help understand the role paralogues play and it will also be of interest to determine whether the same splice variants are maintained and used by the different paralogues. It will also be of interest to determine whether the functions of the ancestral genes have been split between the different splice variants of the paralogues and whether they have specific functions. The study of the regulatory features of the paralogues will also provide insight to both the evolution and control of the paralogues.

3. Improved strategy to identify paralogues

It is apparent from this thesis that a number of sequence features can be used to identify paralogous genes. The program, FINEX, used to search for paralogues with conserved gene structures will be invaluable once the EMBL genomic clones are fully annotated. An additional dimension that should be added to the strategy I have employed in this thesis is the emerging 3-dimensional protein structures. It will then be possible to identify novel paralogues that no longer share detectable sequence identity.

4. Functional studies

The determination of the function(s) of the ancestral genes is crucial to understanding the mechanism(s) by which the paralogues have evolved. Functional studies of orthologues in key organisms, such as amphioxus and hagfish, as well as in higher organisms will also shed light on the present day role of the paralogues in our own genome. Genetic redundancy is evident between paralogues and it will be of interest to understand why redundancy has been maintained.

5. Paralogue-specific microarrays

The expression profile analysis using microarrays has highlighted the value of developing genome-wide paralogue-specific microarrays. Furthermore, in order to truly understand the expression pattern of a particular transcript it is important to develop microarrays that are also splice-variant and allele specific.

Final conclusion

The evidence presented in this thesis is concordant with the 2R hypothesis. Phylogenetic analysis showed that the MHC paralogues located in the paralogous regions on human chromosomes 1, 9 and 19 emerged as part of two large-scale, whole-genome duplication events early in the vertebrate lineage; the first prior to the emergence of jawless fish and one shortly after. Furthermore, investigation of MHC paralogues located outside these regions showed that small-scale duplications, both prior to and after the two whole-genome duplication events, have also moulded the present-day human genome. In total, 791 MHC paralogues were identified in the human genome and were classified as L0, L1, L2 or L3-paralogues by applying a number of criteria. I am confident that the majority of MHC paralogues were identified using my method. However, the addition of further information, such as protein structure data, will enable the detection of any MHC paralogues that have significantly diverged in both sequence and structure since their emergence and were undetected in this thesis. In conclusion, if this project were to be repeated I believe that the approach I took is still viable but I would consider modifying my identification method to include other criteria and I would select more MHC paralogous gene families for further investigation to confirm my thesis findings.