Chapter 4

Identification of the extended MHC paralogues in the human genome

4.1 Introduction

Following the analysis of the proposed paralogous region on chromosome 9 and the release of the first assembled draft human genome sequence it became possible to search for MHC paralogues genome-wide. Previous studies had been criticised for being misleading as they concentrated only on the MHC paralogous genes located in the clusters on 1, 9 and 19 but did not consider paralogues elsewhere in the genome in as much detail. In the words of Hughes and Pontarotti (2000) 'there is no reason to believe that genes in the MHC region are any more likely to have paralogues on these three chromosomes than on any three chromosomes chosen at random from the genome'.

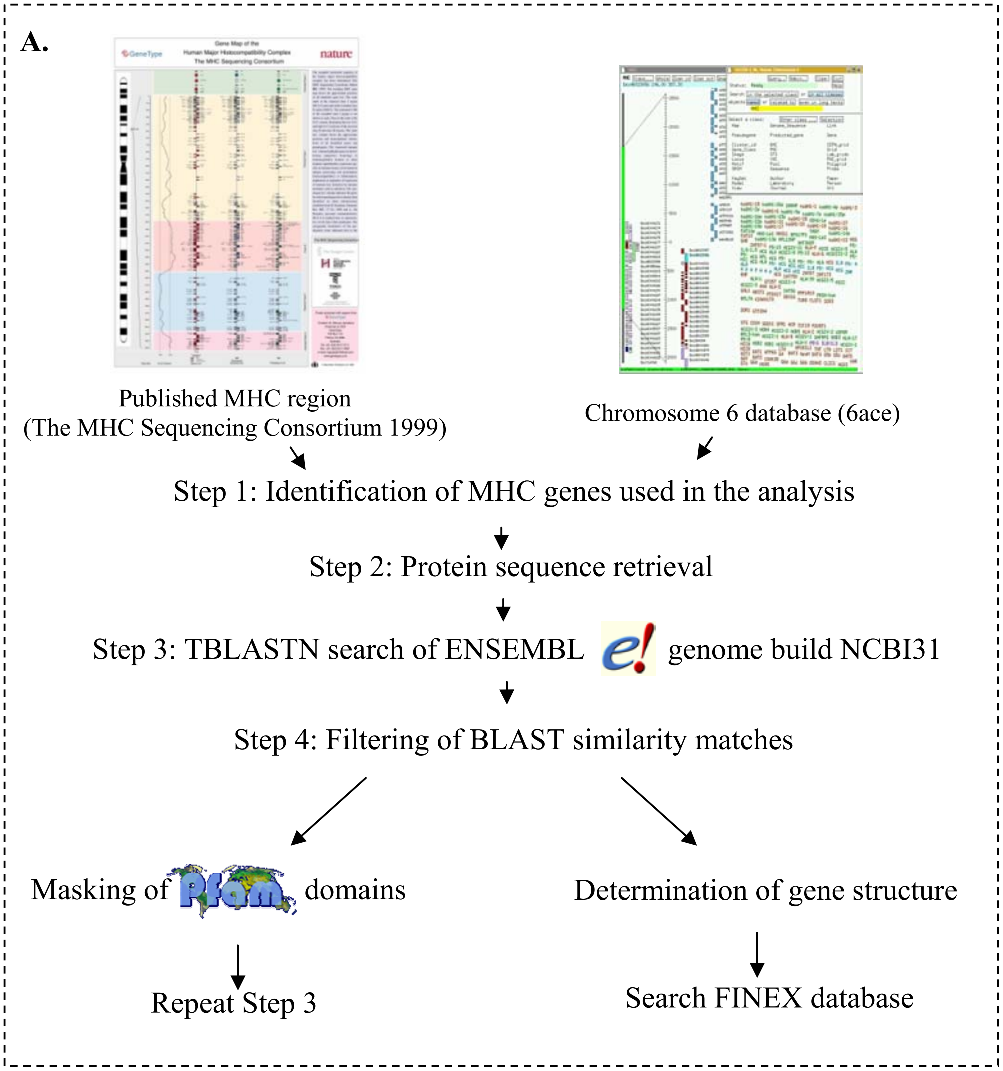
With the increasing amount of genomic sequence data putative MHC paralogues have been identified outside the proposed paralogous regions on 1, 9 and 19; including 12p11-p13, 5q13.1, and 21q22.3 (reviewed by Flajnik and Kasahara, 2001). However, no comprehensive study of the entire genome has been published and the location of the majority of loci cited are based on mapping information available in UNIGENE or generated using cytogenetic mapping techniques, which are not precise. With the advent of the human genome sequence it was now possible to determine the exact location of the proposed MHC paralogues as well as identify novel paralogues which were previously not detected. The purpose of this chapter is to present the findings of a comprehensive and unbiased survey of the human genome with the aim to identify all the MHC paralogous genes and determine their exact location.

4.2 Strategy used to identify MHC paralogues

Previous studies investigating the paralogous gene clusters on chromosomes 1, 9 and 19 identified the putative paralogues using BLAST sequence similarity searches of each available MHC gene (for an example refer to Kasahara, 1999a). Conserved sequence similarity is a feature of homologous gene families and is a good indicator for paralogous genes. In this analysis I use sequence similarity as the initial criterion to identify paralogues but add confidence by using additional sequence features, such as conserved gene structure (intron/exon boundary phases). The approach taken to identify MHC paralogues with increasing levels of confidence in this chapter is outlined in figure 4.1 and discussed in more detail in sections 4.2.1- 4.2.2.

4.2.1 MHC genes used in whole-genome survey

The extended MHC is defined as the sequence on chromosome 6 between HFE (the hereditary haemochromatosis locus), in the extended class I region, and KIFC1 (formerly KNSL2) in the extended class II region (The MHC Sequencing Consortium, 1999).


A.
Published MHC region
(The MHC Sequencing Consortium 1999)
Chromosome 6 database (6ace)
Step 1: Identification of MHC genes used in the analysis
Step 2: Protein sequence retrieval
Step 3: TBLASTN search of ENSEMBL genome build NCBI31
Step 4: Filtering of BLAST similarity matches
Masking of Pfam domains
Determination of gene structure
Repeat Step 3
Search FINEX database

B.

L0-paralogues No detectable sequence similarity outside known protein domains and gene structure not conserved.

L1-paralogues Sequence similarity outside known protein domains.

L2-paralogues Conserved gene organisation detectable by FINEX.

L3-paralogues Sequence similarity outside known protein domains and conserved gene organisation detectable by FINEX.

Figure 4.1 (A) Overview of the strategy used to identify MHC paralogues with increasing levels (L0 to L3) of confidence and definitions (B).

In total, 222 protein-coding gene loci were identified in the region by amalgamating the information published by The MHC Sequencing Consortium (1999) and the up-to-date annotated sequence data available in the Chromosome 6 database, '6ace' (table 4.1). Of the 222 protein-coding genetic loci, 128 were used in the whole-genome survey (table 4.1).

Table 4.1 Distribution of genes in the extended MHC region

	Number of gene loci	***Number of genes****	***Number of genes used in the analysis***	***Size (Mb)***	***~ Gene loci density***
Extended class I	151	93	15	~3.6	1 gene per 24 kb
Class I	111	36	23	1.8	1 gene per 16 kb
Class III	58	58	56	0.7	1 gene per 12 kb
Class II	35	18	18	0.8	1 gene per 23 kb
Extended class II	23	17	16	0.3	1 gene per 13 kb
Total	**378**	**222**	**128**	**~7.2**	**1 gene per 19 kb**

* Genes known, or predicted, to encode a protein.

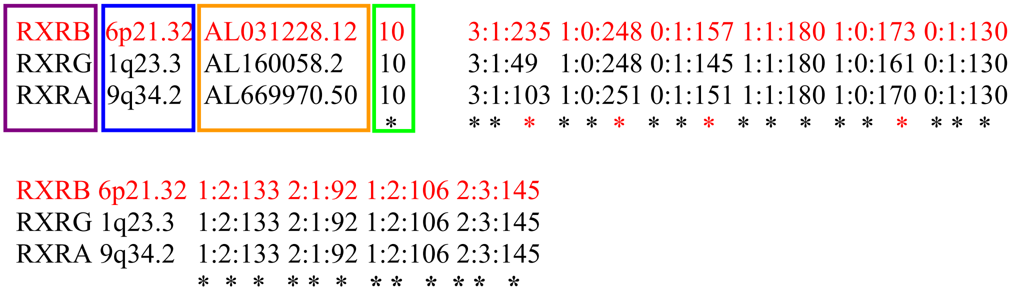
The 128 genes were selected in order to represent each gene family found within the MHC region. In the case of the gene families found within the extended MHC generally only one of the genes was chosen to represent the family. For example, only one of the three Heat Shock Proteins was used in the analysis as they all have one coding exon and share very high sequence similarity. In other cases more than one member of the family was used to attain a full representation of the MHC genes. To date no HLA class II paralogues have been identified in the human genome, therefore in order to ensure any paralogues were detected, all the expressed HLA class II genes were included in the analysis.

Some multigene families that are known to have undergone large-scale expansion have been excluded from the detailed analysis. For example, the extended class I region contains multiple members of the zinc finger, ribosomal and olfactory receptor multigene families that each have up to 1000 paralogues throughout the genome.

4.2.2 Identification of MHC paralogues with increasing levels of confidence

The protein sequences encoded by the 128 MHC gene loci were extracted from either the annotated EMBL database entry of the genomic clone or retrieved from the SWISSPROT or SPTREMBL databases (Bairoch and Apweiler, 1997) and used to identify its paralogues in the human genome (as described in section 2.16). The protein sequence was preferred over the DNA sequence as protein sequence similarity searches increase the likelihood of identifying paralogues which have diverged. DNA sequences are far less conserved particularly as many changes in DNA sequences (third-base changes) do not alter the encoded protein but do change the level of DNA sequence conservation, therefore, lowering the chances of detection by sequence similarity searches. It is generally accepted that if the biological sequence of interest encodes a protein, protein sequence comparison is always the method of choice.

Two sequence features were used to filter the BLAST search results in order to identify paralogues with increasing levels of confidence: these were (1) the exon fingerprints and (2) known protein domains. As described in section 2.16.2, the exon fingerprints of the MHC genes were generated using the CDS features of the annotated genomic clones in the EMBL database and used to search the FINEX database. In addition, the exon fingerprints were deduced for all putative paralogues identified by the initial BLAST similarity search and used to search the FINEX database (summarised in figure 4.2).

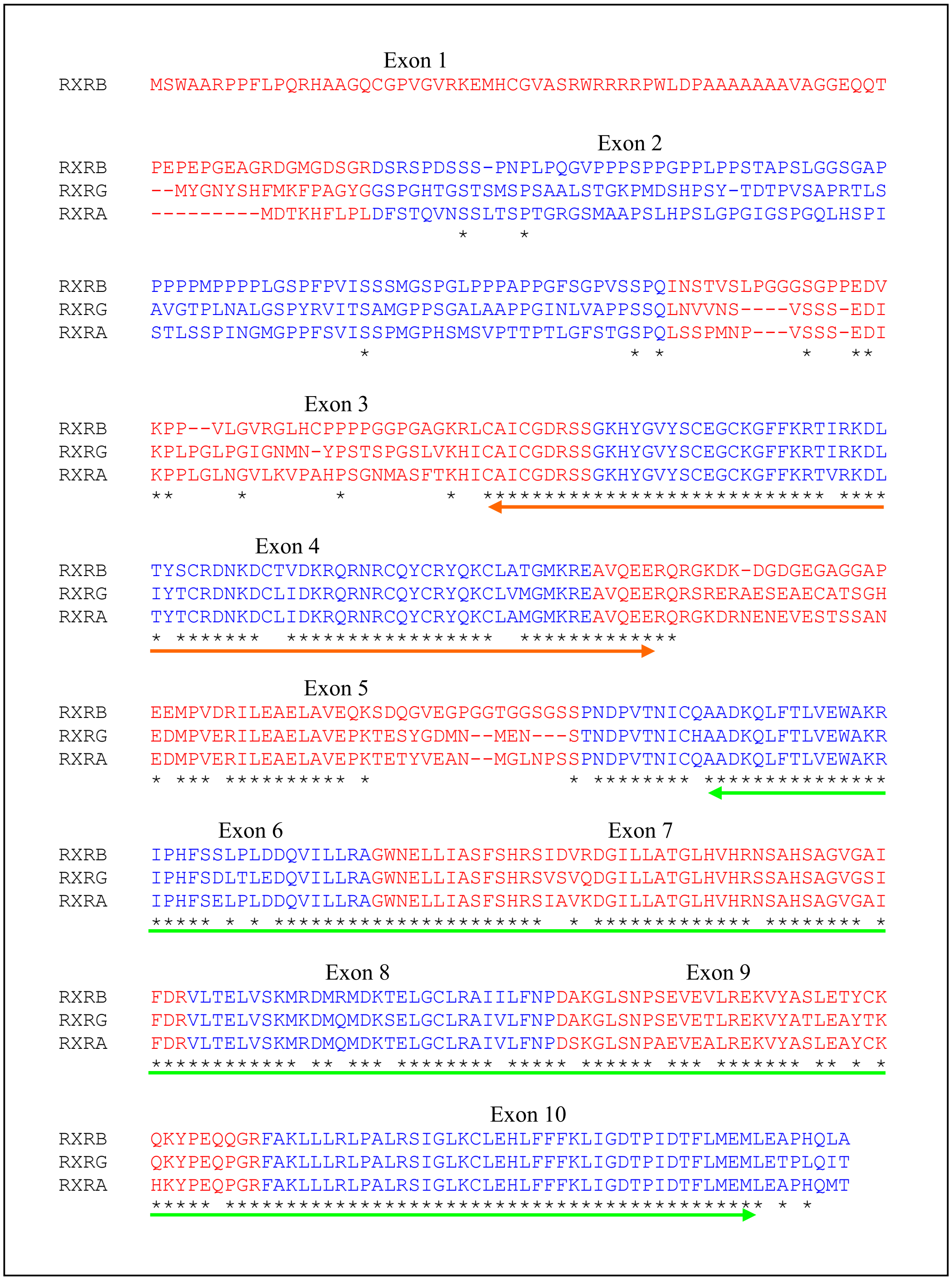


```
RXRB 6p21.32 AL031228.12 10    3:1:235 1:0:248 0:1:157 1:1:180 1:0:173 0:1:130
RXRG 1q23.3  AL160058.2  10    3:1:49  1:0:248 0:1:145 1:1:180 1:0:161 0:1:130
RXRA 9q34.2  AL669970.50 10    3:1:103 1:0:251 0:1:151 1:1:180 1:0:170 0:1:130
                          *    * *  *  * * *  * * *  * *  *  * *  *  * * *

RXRB 6p21.32 1:2:133 2:1:92 1:2:106 2:3:145
RXRG 1q23.3  1:2:133 2:1:92 1:2:106 2:3:145
RXRA 9q34.2  1:2:133 2:1:92 1:2:106 2:3:145
             * * *  * * *  * *  * * *  *
```

Figure 4.2 Alignment of the exon fingerprints of the extended MHC class I gene RXRB (in red) and its paralogues, RXRA and RXRG, identified in the genome survey (discussed in section 4.4.1). The gene names corresponding to the exon fingerprints are boxed in purple and the genomic location in blue. The genomic clone, in which the gene is located, is boxed in orange. In the case of RXRB, the gene is located in the genomic clone, RP5-1033B10, with the EMBL accession number AL031228, and it is the 12th annotated gene with more than one exon within the EMBL entry (hence '.12'). The number boxed in green corresponds to the numbers of coding exons of the gene. The fingerprint of each of the 10 exons follows the same pattern and is represented by a set of three numbers separated by two colons. For example, in the case of the RXRB gene, the first number '3' indicates that this is the start of the gene, characterised by the start codon 'ATG'. The second number '1' indicates that the first intron interrupts a codon and lies between the first and second base. The third number corresponds to the size of the exon, thus, the first exon of RXRB has 235 nucleotides. The asterix indicate whether the phases or exons aligned are identical (black) or different (red).

The known protein domains were identified by searching the PFAM database as described in section 2.16.2. The protein domains of the MHC extended class II encoded protein RXRB and the corresponding regions in the two paralogues, RXRG and RXRA, are shown in figure 4.3.



```
                           Exon 1
RXRB    MSWAARPPPFLPQRHAAGQCGPVGVRKEMHCGVASRWRRRRPWLDPAAAAAAAVAGGEQQT


                                            Exon 2
RXRB    PEPEPGEAGRDGMGDSGRDSRSPDSSS-PNPLPQGVPPPSPPGPPLPPSTAPSLGGSGAP
RXRG    --MYGNYSHFMKFPAGYGGSPGHTGSTSMSPSAALSTGKPMDSHPSY-TDTPVSAPRTLS
RXRA    ---------MDTKHFLPLDFSTQVNSSLTSPTGRGSMAAPSLHPSLGPGIGSPGQLHSPI
                                 *    *

RXRB    PPPPMPPPPLGSPFPVISSSMGSPGLPPPAPPGFSGPVSSPQINSTVSLPGGGSGPPEDV
RXRG    AVGTPLNALGSPYRVITSAMGPPSGALAAPPGINLVAPPSSQLNVVNS----VSSS-EDI
RXRA    STLSSPINGMGPPFSVISSPMGPHSMSVPTTPTLGFSTGSPQLSSPMNP---VSSS-EDI
                         *                     * *           *   **

                    Exon 3
RXRB    KPP--VLGVRGLHCPPPPGGPGAGKRLCAICGDRSSGKHYGVYSCEGCKGFFKRTIRKDL
RXRG    KPLPGLPGIGNMN-YPSTSPGSLVKHICAICGDRSSGKHYGVYSCEGCKGFFKRTIRKDL
RXRA    KPPLGLNGVLKVPAHPSGNMASFTKHICAICGDRSSGKHYGVYSCEGCKGFFKRTVRKDL
        **     *       *        *  **************************** ****

                Exon 4
RXRB    TYSCRDNKDCTVDKRQRNRCQYCRYQKCLATGMKREAVQEERQRGKDK-DGDGEGAGGAP
RXRG    IYTCRDNKDCLIDKRQRNRCQYCRYQKCLVMGMKREAVQEERQRSRERAESEAECATSGH
RXRA    TYTCRDNKDCLIDKRQRNRCQYCRYQKCLAMGMKREAVQEERQRGKDRNENEVESTSSAN
        * *******  *****************  *************

                    Exon 5
RXRB    EEMPVDRILEAELAVEQKSDQGVEGPGGTGGSGSSPNDPVTNICQAADKQLFTLVEWAKR
RXRG    EDMPVERILEAELAVEPKTESYGDMN--MEN---STNDPVTNICHAADKQLFTLVEWAKR
RXRA    EDMPVERILEAELAVEPKTETYVEAN--MGLNPSSPNDPVTNICQAADKQLFTLVEWAKR
        * *** ********** *                * ******** ***************

             Exon 6                               Exon 7
RXRB    IPHFSSLPLDDQVILLRAGWNELLIASFSHRSIDVRDGILLATGLHVHRNSAHSAGVGAI
RXRG    IPHFSDLTLEDQVILLRAGWNELLIASFSHRSVSVQDGILLATGLHVHRSSAHSAGVGSI
RXRA    IPHFSELPLDDQVILLRAGWNELLIASFSHRSIAVKDGILLATGLHVHRNSAHSAGVGAI
        ***** * * **********************  * ************* ******** *

                      Exon 8                         Exon 9
RXRB    FDRVLTELVSKMRDMRMDKTELGCLRAIILFNPDAKGLSNPSEVEVLREKVYASLETYCK
RXRG    FDRVLTELVSKMKDMQMDKSELGCLRAIVLFNPDAKGLSNPSEVETLREKVYATLEAYTK
RXRA    FDRVLTELVSKMRDMQMDKTELGCLRAIVLFNPDSKGLSNPAEVEALREKVYASLEAYCK
        ************ ** *** ******** ***** ****** *** ******* ** * *

                                   Exon 10
RXRB    QKYPEQQGRFAKLLLRLPALRSIGLKCLEHLFFFKLIGDTPIDTFLMEMLEAPHQLA
RXRG    QKYPEQPGRFAKLLLRLPALRSIGLKCLEHLFFFKLIGDTPIDTFLMEMLETPLQIT
RXRA    HKYPEQPGRFAKLLLRLPALRSIGLKCLEHLFFFKLIGDTPIDTFLMEMLEAPHQMT
        ***** ******************************************** * *
```

Figure 4.3 Protein sequence alignment of the extended MHC class II encoded protein, RXRB, and its two paralogues, RXRA and RXRG. The amino acid residues encoded by the 10 exons are shown in red and blue alternatively. The regions indicated by the orange and green arrows correspond to the zinc-finger and ligand binding domains respectively predicted by PFAM, which were masked as described in section 2.16.2. The asterix indicate whether the aligned amino acid residues are identical (black).

4.3 Definitions

The MHC paralogues were identified using the method described in chapter 2 and summarised in section 4.2, and have been defined according to the level of confidence determined by the filtering methods. The terminology used to define the paralogues in this analysis is described in the sections below.

4.3.1 L0-paralogues

L0-paralogues are paralogues that have the lowest level of support. They have been identified by the BLAST similarity search of the ENSEMBL human genome assembly (Hubbard *et al*, 2002) using the TBLASTN executable. They correspond to the BLAST matches with a P-value of less than 10^{-5} and have no other levels of support. A TBLASTN, or translated database search against the human genome is a very productive way to identify paralogous proteins. It is especially suited to working with error prone data like draft genomic sequence because it combines BLAST statistics for hits to multiple reading frames and thus is robust to frame shifts introduced by sequencing or assembly error, which were prevalent in the early genome assemblies.

4.3.2 L1-paralogues

L1-paralogues are paralogues with a moderate level of confidence and level 1 support. They were initially detected by the TBLASTN search of the human genome sequence and have a P-value less than 10^{-5}. In addition, they also have sequence similarity outside the protein domains detected by a TBLASTN search of the domain-masked protein sequence using an expected (E) value of 10 (as described in section 2.16.2.1). In brief, the domains for each protein were identified using the PFAM database and the corresponding residues masked with X's. This method is similar in principle to Repeatmasker which identifies a repeat sequence and substitutes the corresponding nucleotide with either an X or an N (Smit and Green, unpublished). The domain-masked protein sequences were then BLAST searched against the ENSEMBL human genome sequence assembly using the TBLASTN executable.

4.3.3 L2-paralogues

L2-paralogues are paralogues with a higher level of confidence and level 2 support. They were initially detected by the TBLASTN search of the human genome sequence and have a P-value less than 10^{-5} but also have conserved gene structure (FINEX z-value greater than 3.0; as described in section 2.16). In this analysis the FINEX alignment tool was used to compare the exon fingerprints of the MHC encoded gene and the L0-paralogues against the FINEX database (Brown *et al*, 1995). It has been shown for the HLA class II and other gene families that similarities in intron phases and exon fingerprints can be used to define a paralogous gene family (Beck *et al*, 1992a; Radley *et al*, 1994). In addition, MHC proteins encoded by a single exon (for which an exon fingerprint can not be generated) with BLAST similarity matches to paralogues with only one coding exon are also termed L2-paralogues.

4.3.4 L3-paralogues

L3-paralogues are paralogues with the highest level of confidence and level 3 support. They were initially detected by the TBLASTN sequence similarity search of the human genome sequence and have a P-value less than 10^{-5}. They also have conserved sequence identity outside the protein domains and conserved gene structure determined by the two filtering steps.

In summary, the paralogues were identified with varying levels of confidence in order to gain better understanding of the true relationship between the MHC genes and their paralogues. The two filtering methods used to classify the paralogues, detected by the initial sequence similarity search, give an indication of this relationship. The domain-masking filter identifies the paralogues with sequence similarity beyond the domain regions. This filtering step also identifies the paralogues that could be false positives and have only been detected because of a shared domain. These are likely to be members of the same superfamily and are more distantly related. By independently generating the exon fingerprint of the MHC genes, and the paralogues identified in the initial TBLASTN search, the paralogues with conserved gene structures, regardless of sequence similarity, can be distinguished. In addition, the level of conservation of the exons and introns can be determined. Conservation of gene structure and protein sequence indicates that these features are likely to be important for its current day function.

4.4 Results

4.4.1 Identification of MHC paralogues: RXRB as an example

The RXRB gene, also known as retinoic acid receptor beta, is located within the MHC extended class II region on chromosome 6. This gene belongs to the nuclear hormone receptor superfamily and two putative paralogues have previously been identified in the paralogous regions on chromosomes 1 and 9 based on sequence similarity alone. This gene was selected as one of the first genes to be used to identify the paralogues with increasing levels of confidence. The superfamily the gene belongs to is large and includes a number of types of receptors. The receptors share known protein domains and, therefore, sequence identity with RXRB and by applying the filtering steps the paralogues with the highest level of confidence were identified.

The initial TBLASTN sequence similarity search using the RXRB protein sequence identified a total of 48 BLAST sequence similarity matches in the human genome (figure 4.4), of which 27 had a P-value less than 10^{-5}. The 27 BLAST matches, termed paralogues, were then classified (as defined in section 4.3) according to the level of confidence based on the results of two separate filtering steps. One filtering step involved the identification and masking of the protein domains. The RXRB protein contains two PFAM predicted domains; a zinc finger, C4 type spanning from amino acid residue 203 to 278, which is the DNA binding domain of a nuclear receptor (PF00105) and a ligand binding domain spanning from residue 344 to 526, involved in binding the hormone (PF00104). The amino acid residues of the two domains were masked with a series of X's and the masked protein sequence used to BLAST search the human genome using the TBLASTN executable. Two paralogues were identified by this filtering step. They corresponded to two of the 27 paralogues identified by the initial TBLASTN search; the RXRA gene on chromosome 1 and the RXRG gene on chromosome 19 (figure 4.4).

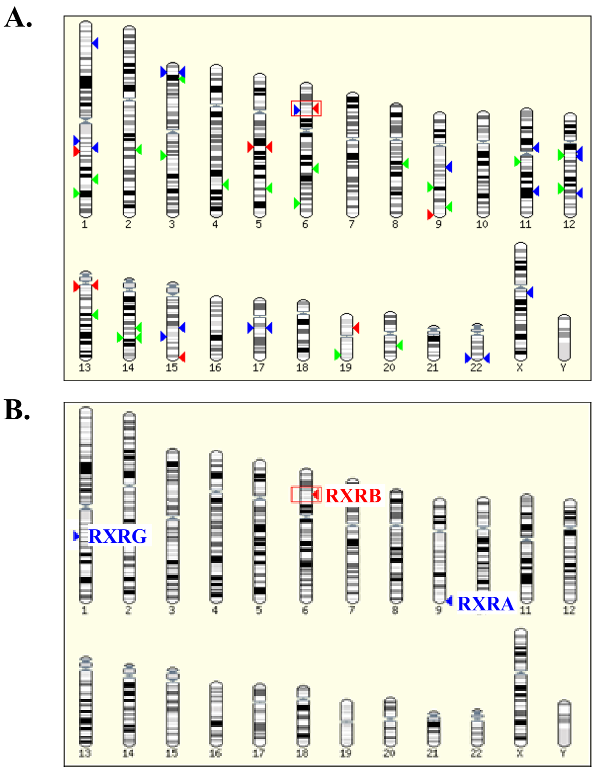
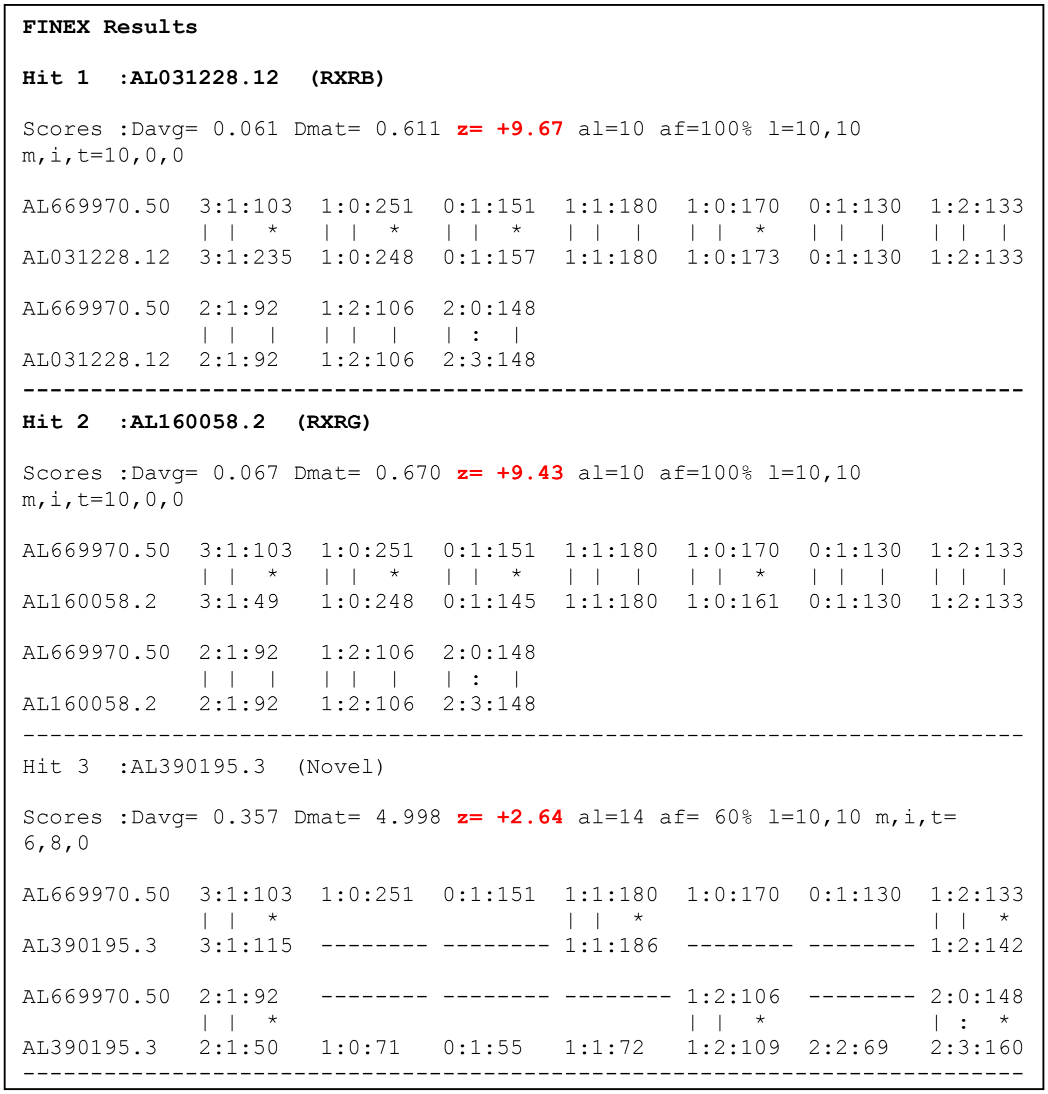


Figure 4.4 Summary of the results of the initial (A) and domain-masked (B) TBLASTN search of the human genome using the RXRB protein sequence. The coloured arrows correspond to the ENSEMBL BLAST score which roughly corresponds to the P-value. In short, a green arrow implies low score and high P-value, a blue arrow indicates moderate score and P-value and a red arrow indicates a high score and low P-value. The location of the RXRB gene and its paralogues RXRA and RXRG are indicated in B.

The second filtering step used gene architectural information to identify paralogues with a higher level of confidence. The intron positions and phases were determined and used to search the FINEX database (described in section 4.2.2 and 2.16). As the FINEX database derived from the EMBL database release 73 contains only 12,282 fingerprints and is not non-redundant the database does not contain the fingerprints for all 30,000 genes in the human genome. This is because the fingerprint database is compiled using annotated coding sequence (CDS feature) information of the EMBL database entry and not all the tiling path clones of the human genome are yet annotated. In order to counteract this, the fingerprints of all 27 paralogues identified by the initial TBLASTN search of the human genome were manually derived and used to search the FINEX database. The paralogues identified the corresponding MHC locus (figure 4.5), and the paralogues were classified based on all three lines of evidence.



```
FINEX Results

Hit 1  :AL031228.12  (RXRB)

Scores :Davg= 0.061 Dmat= 0.611 z= +9.67 al=10 af=100% l=10,10
m,i,t=10,0,0

AL669970.50  3:1:103  1:0:251  0:1:151  1:1:180  1:0:170  0:1:130  1:2:133
             | |  *   | |  *   | |  *   | |  |   | |  *   | |  |   | |  |
AL031228.12  3:1:235  1:0:248  0:1:157  1:1:180  1:0:173  0:1:130  1:2:133

AL669970.50  2:1:92   1:2:106  2:0:148
             | |  |   | |  |   | :  |
AL031228.12  2:1:92   1:2:106  2:3:148
--------------------------------------------------------------------------
Hit 2  :AL160058.2  (RXRG)

Scores :Davg= 0.067 Dmat= 0.670 z= +9.43 al=10 af=100% l=10,10
m,i,t=10,0,0

AL669970.50  3:1:103  1:0:251  0:1:151  1:1:180  1:0:170  0:1:130  1:2:133
             | |  *   | |  *   | |  *   | |  |   | |  *   | |  |   | |  |
AL160058.2   3:1:49   1:0:248  0:1:145  1:1:180  1:0:161  0:1:130  1:2:133

AL669970.50  2:1:92   1:2:106  2:0:148
             | |  |   | |  |   | :  |
AL160058.2   2:1:92   1:2:106  2:3:148
--------------------------------------------------------------------------
Hit 3  :AL390195.3  (Novel)

Scores :Davg= 0.357 Dmat= 4.998 z= +2.64 al=14 af= 60% l=10,10 m,i,t=
6,8,0

AL669970.50  3:1:103  1:0:251  0:1:151  1:1:180  1:0:170  0:1:130  1:2:133
             | |  *                     | |  *                     | |  *
AL390195.3   3:1:115  -------- -------- 1:1:186  -------- -------- 1:2:142

AL669970.50  2:1:92   -------- -------- -------- 1:2:106  -------- 2:0:148
             | |  *                              | |  *            | :  *
AL390195.3   2:1:50   1:0:71   0:1:55   1:1:72   1:2:109  2:2:69   2:3:160
--------------------------------------------------------------------------
```

Figure 4.5 Summary of the FINEX search using the RXRA fingerprint (AL669970.50). The RXRA gene identified the RXRB and RXRG genes (in bold) with a z-score greater than 3.0 (as described in section 2.16.2.2; highlighted in red).

To summarise, by combining all three sets of results, or lines of evidence, it was found that, of the 27 paralogues identified by the initial sequence similarity search the RXRB gene has 25 L0-paralogues, no L1-paralogues, no L2-paralogues and two L3-paralogues. The two L3-paralogues, or paralogues with the highest level of confidence, are the RXRG and RXRA genes located on 1q23.3 and 9q34.2, respectively.

4.4.2 Identification of all the MHC paralogues in the human genome

Over two-thirds of the 128 MHC genes investigated in the genome survey have paralogues in the human genome with, at least, the lowest level of support (88/128); the remaining third have no paralogues detectable by this method. In summary, 30% of the MHC genes with identified paralogues have L3-paralogues (26/88), 16% have L2- paralogues (14/88), 18% have L1-paralogues (16/88) and the remaining 36% have L0-paralogues (32/88). The results are summarised in table 4.2 and figure 4.6.

Table 4.2 Summary of the MHC genes with paralogues of increasing levels with support

MHC Region	*L0-paralogues*	*L1-Paralogues*	*L2-Paralogues*	*L3-Paralogues*	*Total*
Extended class I	3	5	4	3	15
Class I	5	4	2	4	15
Class III	6	4	5	13	28
Class II	11	2	2	3	18
Extended class II	7	1	1	3	12
Total	**32**	**16**	**14**	**26**	**88**

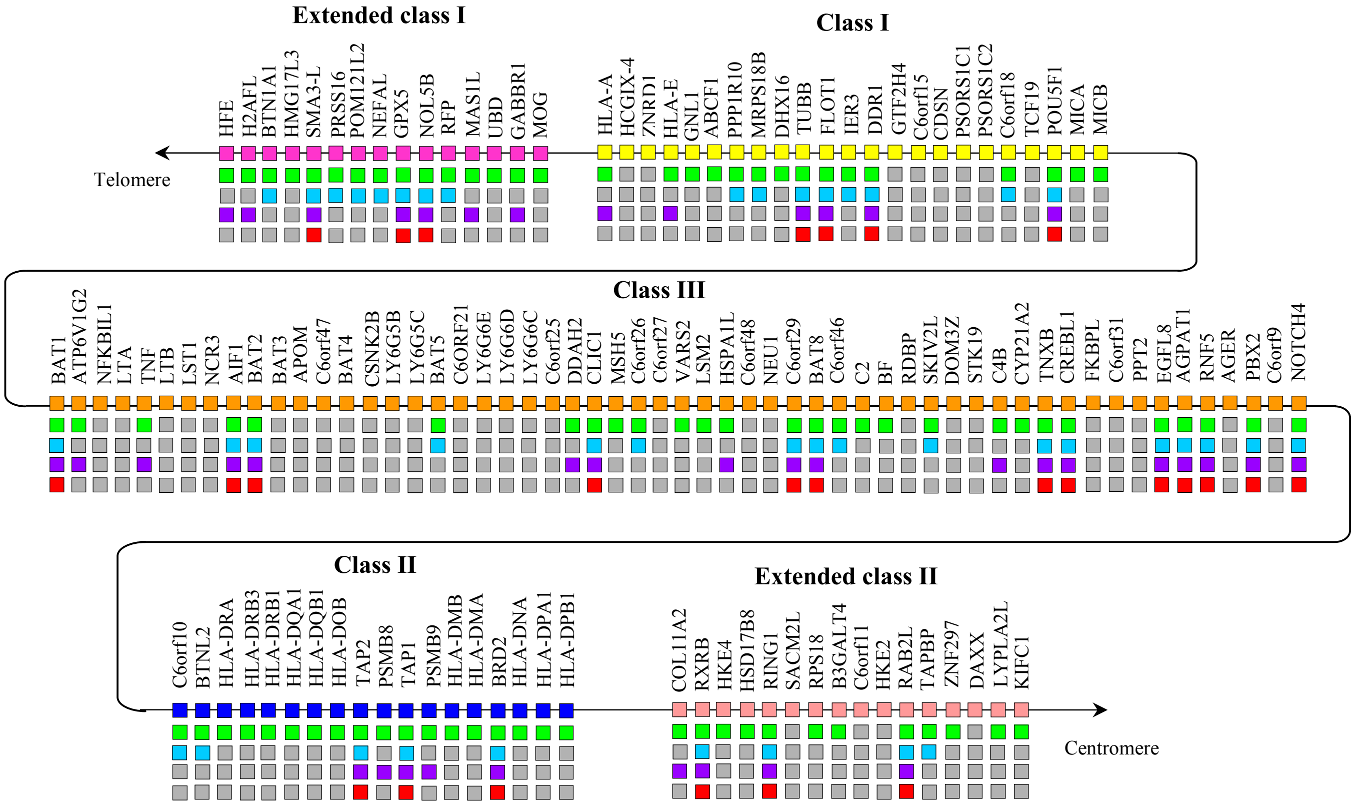


Figure 4.6 Summary of the results of the whole-genome survey using 128 MHC genes. The MHC region is divided into five classes and the genes within each class are represented by coloured boxes; extended class I are cerise, class I yellow, class III orange, class II blue and extended class II are pink. Green filled boxes in row 2 indicate that paralogues were detected by the initial BLAST similarity search, turquoise filled boxes in row 3 indicate that paralogues were detected by the domain-masked BLAST search, purple filled boxes in row 4 indicate that paralogues were detected by FINEX and red filled boxes in row 5 represent genes that have paralogues with the highest level of confidence (L3-paralogues) in the human genome. Grey filled boxes indicate that no results were obtained by the corresponding analysis.

A total of 1057 BLAST similarity matches to the 128 MHC genes were identified with a P-value less than 10^{-5}. Of the 1057 BLAST matches, 128 correspond to the MHC genes used in the analysis and a further 138 loci are located within the MHC region. The 138 loci represent the paralogous genes within the MHC region itself, for example the HIST1H2AC, HLA-A and HLA-E genes are all members of multigene families that share high sequence similarity and, therefore, BLAST sequence similarity search detected the other family members. These 138 loci have been removed from the analysis.

In total, 791 MHC paralogues have been identified outside the MHC region, of which 618 are L0-paralogues, 91 are L1-paralogues, 38 are L2-paralogues and 44 are L3-paralogues (summarised in figure 4.7).

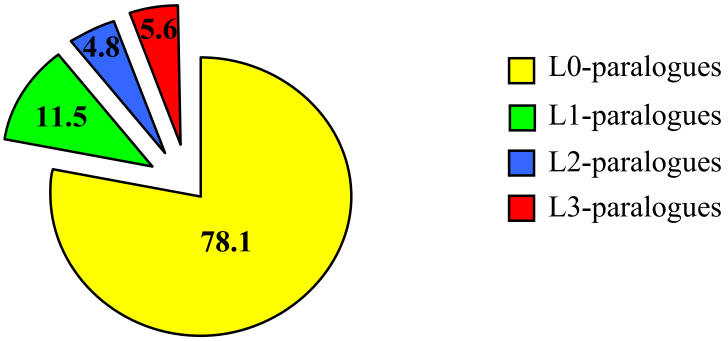


Figure 4.7 Summary of the proportion (%) of BLAST hits corresponding to the paralogues with different levels of confidence.

The paralogues classified as either L2- or L3-paralogues have conserved gene structure, whereas the L0- and L1-paralogues have been identified by sequence similarity alone and may represent distantly related genes rather than paralogues, this will be discussed in section 4.4.6. In total, 44 L3-paralogues have been identified in this analysis. Figure 4.8 summarises the number of paralogues with the different levels of confidence for each MHC gene used in the genome survey.

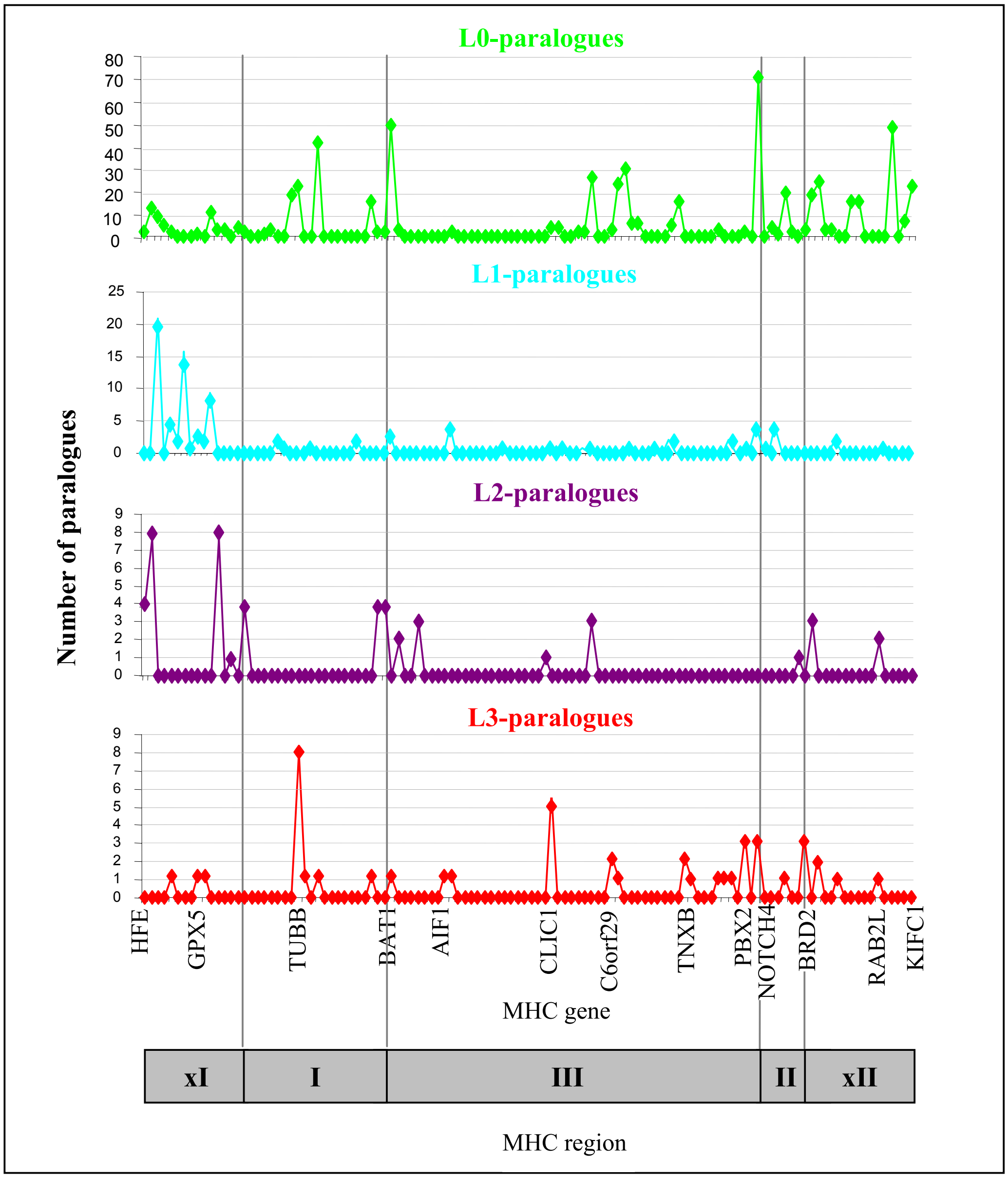


Figure 4.8 Summary of the MHC genes with L0- to L3-paralogues. The L0-paralogues are shown in green, the L1-paralogues are in turquoise, the L2-paralogues are shown in purple and the L3-paralogues are in red. The y-axis on each graph represent the number of paralogues identified with each level of confidence (note scales differ) and the x-axis represent the MHC genes used in the analysis plotted (from left to right) in order from the most telomeric in the extended class I region (xI) to the most centromeric in the extended class II region (xII).

The MHC genes with L3-paralogues are not restricted to one region of the MHC and span almost the entire length of the region, including genes within the telomeric extended MHC class I region and the centromeric extended MHC class II region. Analysis of the distribution of the genes within the MHC region with L3-paralogues reveal 'hotspots' of genes with paralogues; one in particular is located towards the centromeric end of the class III region bordering the class II region. The genes located within this 'hotspot' include EGFL8, TNXB and NOTCH4 which have two, one and three paralogues in the human genome, respectively. There are also 'cold-spots' of MHC genes with no paralogues; namely surrounding the Ly6 gene family in the MHC class III region.

Figure 4.9 summarises the percentage of MHC genes with different numbers of L0-, L1-, L2- and L3-paralogues in the human genome. In general, the MHC genes do not have paralogues with the highest level of confidence; however, there are gene families with two or more L3-paralogues. For example, the C6orf29 gene has two L3-paralogues and the BRD2 gene has three. In the extreme, the TUBB gene has seven L3-paralogues located in the human genome and the CLIC1 gene has five.

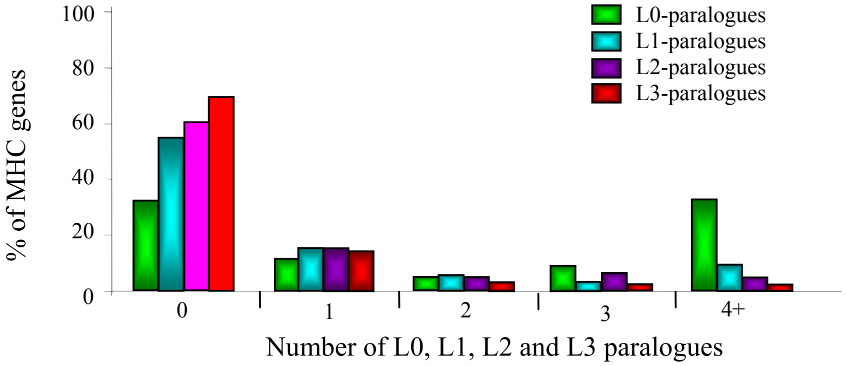


Figure 4.9 Summary of the percentage (%) of MHC genes with no, 1, 2, 3, 4 or more L0, L1, L2 and L3-paralogues in the human genome

4.4.3 Distribution of MHC paralogues in the human genome

In order to determine the distribution of the MHC paralogues in the human genome the L0- to L3-paralogues were plotted on an ideogram of all 24 chromosomes (figure 4.10). The frequency of the paralogues per chromosome is summarised in table 4.3. Interestingly, the chromosomes with the highest number of L3-paralogues correspond to the chromosomes proposed to contain paralogous gene clusters. In contrast, chromosomes 2, 3, 4, 8 and Y do not contain any L2- or L3- paralogues, but do harbour paralogues with lower levels of support.

Table 4.3 Summary of the distribution of MHC paralogues in the human genome.

Chromosome	***L3-paralogue***	***L2-paralogue***	***L1-paralogue***	***L0-paralogue***	***Total***
1	**12**	**10**	11	62	95
2	0	0	6	37	43
3	0	0	2	36	38
4	0	0	7	22	29
5	**2**	**1**	15	28	46
6	**3**	**1**	2	29	35
7	**0**	**2**	8	31	41
8	0	0	2	23	25
9	**12**	**5**	5	33	55
10	**1**	**0**	1	27	29
11	**0**	**8**	10	34	52
12	**1**	**1**	3	25	30
13	**0**	**1**	3	19	23
14	**0**	**1**	1	13	15
15	**0**	**1**	3	26	30
16	**1**	**0**	1	21	23
17	**1**	**1**	0	31	33
18	**1**	**0**	1	2	4
19	**6**	**5**	0	48	59
20	**1**	**0**	2	17	20
21	**1**	**0**	0	10	11
22	**1**	**0**	8	12	21
X	**1**	**1**	0	30	32
Y	0	0	0	2	2
Total	**44**	**38**	**91**	**618**	**791**

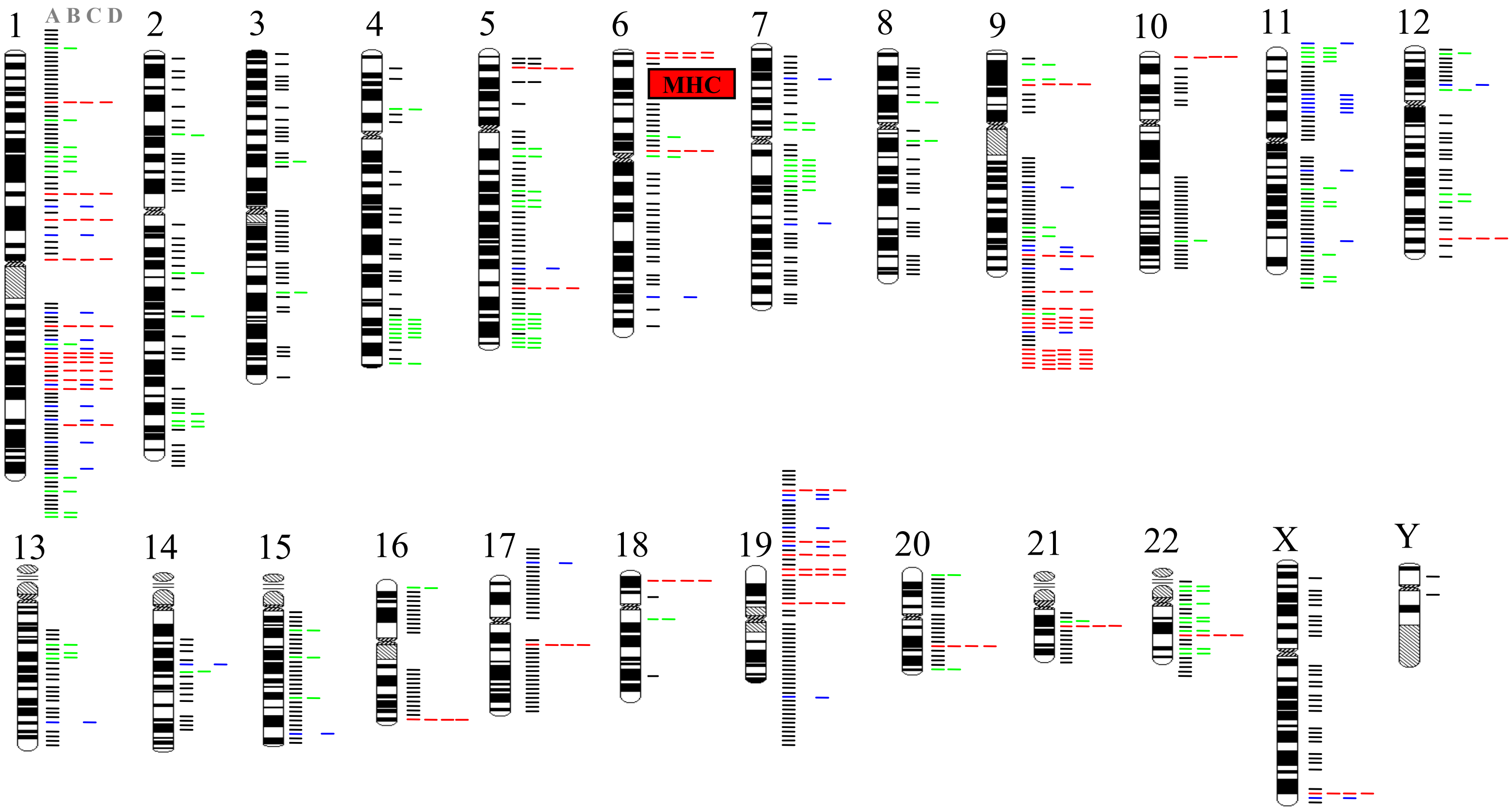


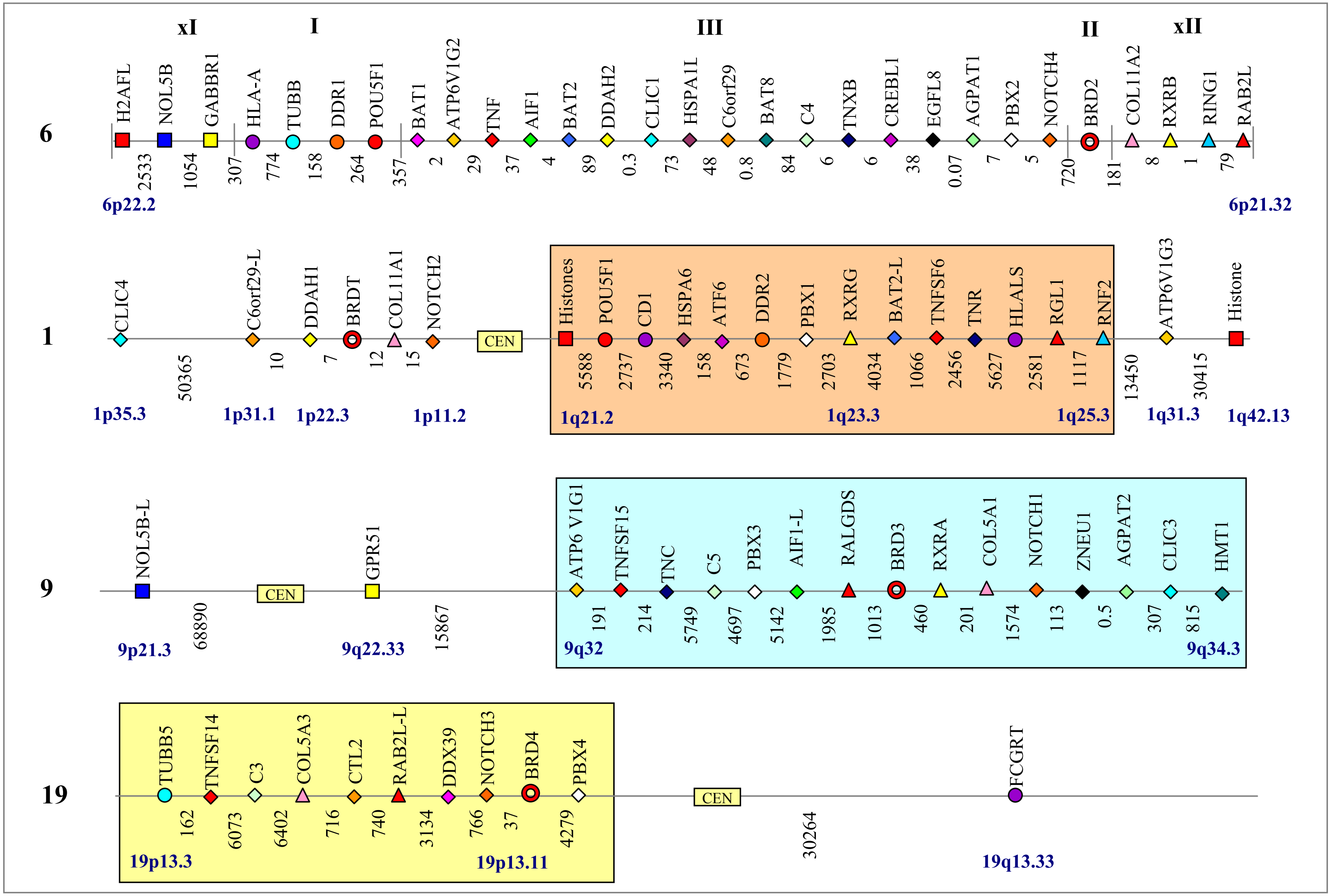
Figure 4.10 Distribution of MHC paralogues in the human genome. Column A represents all BLAST similarity matches with a P-value less than 10^{-5}. Column B represents the BLAST matches still detected after the domain-masking filtering step, column C represents BLAST matches still detected after the FINEX filtering step. The final column (D) represents the BLAST matches which passed both filtering steps and represent the L3 paralogues. The lines correspond to the paralogues and are colour-coded according to type: black represent L0-paralogues, green L1-paralogues, blue L2-Paralogues and red L3-paralogues. The data used to generate this figure is summarised in Appendix 2.

In total, 82 L2- and L3-paralogues of genes located within the MHC region have been identified elsewhere in the genome, and correspond to 29 MHC gene families. Almost 50% (40/82) of these paralogues are confined to the paralogous regions on 1q21.2-q25.3, 9q32-q34.3 and 19p13.3-p13.11 and the remaining 51% are scattered throughout the genome. In total, 38 of the 82 L2- and L3-paralogues are novel findings.

4.4.4 MHC paralogues located on chromosomes 1, 9 and 19

The whole genome survey has confirmed that there are clusters of paralogues on chromosomes 1, 9 and 19. The distribution of the L2- and L3-paralogues on these three chromosomes and the corresponding MHC genes is summarised in figure 4.11. Each of the 29 MHC genes and their respective paralogue(s) are represented by coloured symbols and the distance separating the genes along each chromosome is given. In order to compare the findings of the whole genome survey with previous publications, a comprehensive list of the 78 putative paralogues already described in the literature was obtained by combining the gene lists published by Kasahara (1999a; 1999b) and Flajnik and Kasahara (2001). Each of the chromosomes will be discussed individually in the following sections.

Figure 4.11 Summary of MHC paralogues on chromosomes 1, 9 and 19. The MHC genes on chromosome 6 and corresponding paralogues are represented by coloured symbols. The distance, in kilobases, between paralogues is shown below the gene track. The cytogenetic loci are given in blue text below the gene track for some paralogues for orientation purposes. The paralogous regions are boxed and shaded peach 1q21.2-q25.3, blue for 9q32-q34.3 and yellow corresponds to 19p13.3-p13.11.


xI
I
III
II
xII
6
H2AFL
NOL5B
GABBR1
HLA-A
TUBB
DDR1
POU5F1
BAT1
ATP6V1G2
TNF
AIF1
BAT2
DDAH2
CLIC1
HSPA1L
C6orf29
BAT8
C4
TNXB
CREBL1
EGFL8
AGPAT1
PBX2
NOTCH4
BRD2
COL11A2
RXRB
RING1
RAB2L
2533
1054
307
774
158
264
357
2
29
37
4
89
0.3
73
48
0.8
84
6
6
38
0.07
7
5
720
181
8
1
79
6p22.2
6p21.32
1
CLIC4
C6orf29-L
DDAH1
BRDT
COL11A1
NOTCH2
CEN
Histones
POU5F1
CD1
HSPA6
ATF6
DDR2
PBX1
RXRG
BAT2-L
TNFSF6
TNR
HLALS
RGL1
RNF2
ATP6V1G3
Histone
50365
10
7
12
15
5588
2737
3340
158
673
1779
2703
4034
1066
2456
5627
2581
1117
13450
30415
1p35.3
1p31.1
1p22.3
1p11.2
1q21.2
1q23.3
1q25.3
1q31.3
1q42.13
9
NOL5B-L
CEN
GPR51
ATP6 V1G1
TNFSF15
TNC
C5
PBX3
AIF1-L
RALGDS
BRD3
RXRA
COL5A1
NOTCH1
ZNEU1
AGPAT2
CLIC3
HMT1
68890
15867
191
214
5749
4697
5142
1985
1013
460
201
1574
113
0.5
307
815
9p21.3
9q22.33
9q32
9q34.3
19
TUBB5
TNFSF14
C3
COL5A3
CTL2
RAB2L-L
DDX39
NOTCH3
BRD4
PBX4
CEN
FCGRT
162
6073
6402
716
740
3134
766
37
4279
30264
19p13.3
19p13.11
19q13.33

4.4.4.1 Chromosome 1 paralogues

There are a total of 49 putative paralogues located on the long arm of chromosome 1 spanning from 1q21.1 to 1q44, of which, 28 are L0-paralogues, five are L1-paralogues, eight are L2-paralogues and eight are L3-paralogues. The L2- and L3-paralogues are summarised in table 4.4.

Table 4.4 Summary of the L2- and L3-paralogues on chromosome 1. The paralogues shown in red are novel and the paralogous region is shaded orange.

	MHC gene	*MHC Region*	*Paralogue*	*Locus*	*Confidence*
Chromosome 1	CLIC1	III	CLIC4	1p35.3	3
	C6orf29	III	NM_152697	1p31.1	3
	DDAH2	III	DDAH1	1p22.3	2
	BRD2	II	BRDT	1p22.1	3
	COL11A2	xII	COL11A1	1p21.1	2
	NOTCH4	III	NOTCH2	1p11.2	3
	Histone cluster	xI	Histone cluster	1q21.2	2
	POU5F1	I	Q9BZW0	1q22	3
	HFE	xI	CD1A	1q23.1	2
	HSPA1L	III	HSPA6	1q23.3	2
	CREBL1	III	ATF6	1q23.3	3
	DDR1	I	DDR2	1q23.3	3
	PBX2	III	PBX1	1q23.3	3
	RXRB	xII	RXRG	1q23.3	3
	BAT2	III	BAT2-ISO	1q24.3	3
	TNF	III	TNFSF6	1q24.3	2
	TNXB	III	TNR	1q25.1	3
	HLA Class I/II	I/II	HLALS	1q25.3	2
	RAB2L	xII	RGL1	1q25.3	2
	RING1	xII	RNF2	1q25.3	3
	ATP6V1G2	III	ATP6V1G3	1q31.3	2
	Histone cluster	xI	H2-like	1q42.13	2

The chromosome 1 paralogous region is defined by a histone cluster at the most centromeric end (1q21.2) and the RNF2 paralogue of RING1, spanning approximately 35 Mb (summarised in figure 4.11). The centromeric histone cluster is reminiscent of the histone cluster located in the extended MHC class I region. In addition to the histone gene cluster (considered here as a single entity), there are eight L3-paralogues six L2- paralogues, one L1-paralogue and 11 L0-paralogues located within this region. Within the paralogous region there is a small cluster of L2- and L3-paralogues spanning from the CREBL1 paralogue, ATF6 (most centromeric), located on 1q23.3 to the TNXB paralogue, TNR, located on 1q25.1 (telomeric). This cluster contains seven MHC paralogues with level 2 or 3 confidences encompassing 13.4 Mb.

In addition to the paralogues located within the region on the q-arm of chromosome 1, there are six L2- and L3-paralogues located on the short arm of chromosome 1, of which some have previously been cited as being part of the paralogous region on chromosome 1 (reviewed by Kasahara, 1999b). It is believed that the paralogous gene cluster was split onto both arms as a result of the insertion of the centromere or by a perincentromeric inversion of chromosome 1 (reviewed by Kasahara, 1999b). Thus, the four L3-paralogues may have been part of the original paralogous gene cluster on the q-arm and have since been separated.

In total, three new paralogues have been identified in the genome survey, which have previously not been cited in the literature. The three novel paralogues are a C6orf29-like gene (NM_152697) on 1p31.1, a POU5F1-like gene (Q9BZW0) and the ATP6V1G3 gene, which is a paralogue of the MHC class III gene ATP6V1G2. A paralogue of the MHC class I gene, POU5F1, has previously been cited in the literature on 1p34.1 (POU3F1) but this was not the paralogue identified in this analysis.

The L2- and L3-paralogues located on both arms of chromosome 1 correspond to 21 MHC gene families. However, 31 paralogues corresponding to 26 MHC gene families have previously been identified on chromosome 1 (see Flajnik and Kasahara (2001) for most recent gene list). In this analysis 18 have been identified as L2- and L3-paralogues, three were identified with the lowest level of support and nine were not identified at all. Of the ten paralogues not identified, four are paralogues of MHC genes that were excluded from the analysis for reasons given in section 4.2. The remaining five paralogues were not detected in the genome survey because of low protein sequence identity and is discussed in more detail in section 4.4.4.4.

The NTRK1 gene, located on 1q23.1, has been cited as a paralogue of the MHC class I gene DDR1 (Flajnik and Kasahara, 2001) and, in this survey of the human genome, was identified as a paralogue with the lowest level of support (L0-paralogue). The NTRK1 gene was identified by the BLAST sequence similarity search but, once the known domains were masked, it did not have conserved sequence identity beyond these regions. The gene structure is also very different to that of the DDR1 gene and shows high conservation with the NTRK2 gene located on 9q21.33. The NTRK2 gene has also been cited as a paralogue of DDR1 but was only identified as an L0-paralogue in this analysis. Evidence, based on gene structure and protein sequence similarity, indicates that NTRK1 is paralogous to NTRK2 but is more distantly related to DDR1. Therefore, the DDR2 gene located on 1q23.3 (an L3-paralogue) represents the only true paralogue of DDR1 in the human genome. Thus demonstrating how the genome-wide survey presented in this chapter has enabled errors to be corrected.

4.4.4.2 Chromosome 9 paralogues

Chromosome 9 harbours 55 paralogues, of which 17 are L2- and L3-paralogues (summarised in table 4.5).

Table 4.5 Summary of the L2- and L3-paralogues on chromosome 9. The paralogues shown in red text are novel and the paralogous region is shaded blue.

	MHC gene	*MHC Region*	*Paralogue*	*Locus*	*Confidence*
Chromosome 9	NOL5B	xI	NOL5B-L	9p21.3	3
	GABBR1	xI	GPR51	9q22.33	2
	ATP6V1G2	III	ATP6V1G1	9q32	2
	TNF	III	TNFSF15	9q32	2
	TNXB	III	TNC	9q33.1	3
	C4	III	C5	9q33.2	2
	PBX2	III	PBX3	9q33.3	3
	AIF1	III	NM_031426	9q34.12	3
	RAB2L	xII	RALGDS	9q34.2	3
	BRD2	II	BRD3	9q34.2	3
	RXRB	xII	RXRA	9q34.2	3
	COL11A2	xII	COL5A1	9q34.3	2
	NOTCH4	III	NOTCH1	9q34.3	3
	EGFL8	III	ZNEU1	9q34.3	3
	AGPAT1	III	AGPAT2	9q34.3	3
	CLIC1	III	CLIC3	9q34.3	3
	BAT8	III	HMT1	9q34.3	3

There is only one L3-paralogue located on the p-arm of chromosome 9, NOL5B-L, which is a novel finding. To-date, no paralogues of the extended MHC class I encoded gene, NOL5B, have been discussed in the literature. The L3-paralogue is actually a 'Novel' protein and, has been termed NOL5B-L in this thesis. The paralogous region encompasses the regions 9q32 to 9q34.3 (refer to chapter 3 for more detail; also see figure 4.11) spanning from the ATP6V1G2 paralogue, ATP6V1G1, to the BAT8 paralogue, HMT1 (approximately 24 Mb). Within this region there are 28 putative paralogues; 15 L2- and L3-paralogues, one L1- paralogue and 12 L0-paralogues. There are two small clusters located within the defined boundaries; cluster 1 spans from the AIF1-L paralogue (9q34.12) to COL5A1 (9q43.3) encompassing approximately 4 Mb and the second cluster spans approximately 14.8 Mb from NOTCH1 (9q34.3) to HMT1 (9q34.3). There is an additional L3-paralogue located on 9q22.33, almost 16 megabases centromeric of the

ATP6V1G2 gene defining the paralogous gene cluster on 9q32-q34.3. This is the previously published GPR51 gene, which is paralogous to the GABBR1 gene.

In total, 30 putative paralogues have been identified in the literature, corresponding to 27 MHC gene families, and are cited as being located within the paralogous region on chromosome 9. The whole genome survey has identified 15 as L2- and L3-paralogues, 1 as a pseudogene (TUBB2) and two as L0-paralogues. In total, nine of the 31 putative paralogues were not identified in the genome survey presented in this chapter. Two of these putative paralogues are paralogous to MHC genes not used in the genome survey, for reasons discussed in section 4.2, and the remaining five were not identified because they share low sequence similarity with the corresponding MHC encoded protein (discussed in more detail in section 4.4.4.4).

One gene of interest is the BAT2 gene located within the MHC class III region. The KIAA0515 gene located on chromosome 9 has been cited as a putative paralogue of BAT2, but it was not identified as a paralogue in my analysis. However, an L1-paralogue has been identified, which is the neighbouring gene of KIAA0515 in the genome. In addition to the new NOL5B paralogue identified on the p-arm of chromosome 9, a novel paralogue of the EGFL8 gene, ZNEU1, has been discovered on 9q34.3. This MHC gene was previously not identified as being part of the published MHC paralogous group.

4.4.4.3 Chromosome 19 paralogues

Sixteen putative paralogues have previously been identified on the short arm of chromosome 19. The genome-wide survey presented in this chapter identified nine of these as L2- or L3-paralogues (table 4.6). The seven remaining putative paralogues were not identified at all; three were not identified because they are paralogous to MHC genes not used in this analysis (for the reasons given in section 4.2), and four share low sequence similarity with the corresponding MHC encoded protein (discussed in section 4.4.4.4).

Table 4.6 Summary of the L2- and L3-paralogues on chromosome 19. The paralogues shown in red text are novel and the paralogous region is shaded yellow.

	MHC gene	***MHC Region***	***Paralogue***	***Locus***	***Confidence***
Chromosome 19	TUBB	I	TUBB5	19p13.3	3
	TNF	III	TNFSF14	19p13.3	2
	C4B	III	C3	19p13.3	2
	COL11A2	xII	COL5A3	19p13.2	2
	C6orf29	III	CTL2	19p13.2	3
	RAB2L	xII	Q8TEP0	19p13.2	2
	BAT1	III	DDX39	19p13.13	3
	NOTCH4	III	NOTCH3	19p13.12	3
	BRD2	II	BRD4	19p13.12	3
	PBX2	III	PBX4	19p13.11	3
	HLA Class I	I	FCGRT	19q13.33	2

The paralogous region spans approximately 13.6 Mb (figure 4.11) from the TUBB5 gene at the telomere to the PBX4 gene towards the centromere. In total, 25 paralogues are located within this region, of which, six are L3-paralogues, four are L2-paralogues and 15 are L0-paralogues. Within this region there is a smaller cluster of paralogues spanning almost 9.9 Mb from the COL5A3 gene (19p13.2) to the PBX4 (19p13.11) gene encompassing seven L2- and L3-paralogues. In addition there is an HLA class I like gene, FCGRT, located on the q-arm of chromosome 19.

Two new paralogues were identified within the paralogous region on 19p13.3-p13.11. The TNFSF14 gene is paralogous to the tumour necrosis factor (TNF) gene located within the MHC class III region. Although, other members of the TNF family have been identified in the literature as putative TNF paralogues, this paralogue is a novel finding. The second new paralogue extends the RAB2L paralogous gene family from two to three members, and the family now has members located in the MHC extended class II region, on 1q25.3 (RGL1) and 19p13.2 (Q8TEP0).

4.4.4.4 Putative paralogues not identified in the genome-wide survey

Of the 78 putative paralogues presented in the literature, 32% were not identified in the whole genome survey presented in this thesis (summarised in table 4.7). The strategy I used to identify paralogues relies on sequence similarity (as described in section 4.2). Therefore, if the protein sequence similarity is too low it is either not detected by a BLAST similarity search using the parameters described in section 2.16 or has a P-value greater than 10^{-5} and is filtered from the BLAST results because it is regarded as either insignificant or a distant relative (Lesk, 2002). This is exemplified by the tumour necrosis factor genes, LTA, TNF and LTB, located in the MHC class III region. In total, seven putative paralogues of these three genes have been discussed in the literature; however, only two were identified in my genome-wide analysis. This is because they share less than 20% protein sequence identity which will probably not be detectable by the BLAST algorithm used in this analysis, WU-BLAST2 (discussed in more detail in section 4.4.7; Brenner *et al*, 1998).

Table 4.7 Summary of the putative MHC paralogues not identified in my genome-wide survey. The putative paralogues of MHC genes not used in the analysis are shaded in lilac and are in italics.

MHC region	***MHC Gene***	***Published paralogue***	***Published locus***
xI	*HMG17L3*	*HMG17*	*1p36.5-p35*
xI	PRSS16	DPP7	9q34
xI	*ZNF184*	*ZNF85*	*19p12-p13.1*
xI	*ZNF184*	*ZNF91*	*19p12-p13.1*
xI	GPX5	GPX4	19p13.1
xI	*OR cluster*	*OR cluster*	*9q21-q22, 9q34*
xI	*OR cluster*	*OR cluster*	*19p13.1*
I	*KIAA0170*	*PRG4*	*1q25-q31*
III	TNF/LTA/LTB	TNFSF18	1q23
III	TNF/LTA/LTB	TNFSF4	1q25
III	TNF/LTA/LTB	TNFSF8	9q33
III	TNF/LTA/LTB	TNFSF9	19p13
III	TNF/LTA/LTB	TNFSF7	19p13
III	AIF1	AIF1-L	1p33-p34
III	HSPA1L	HSPA5	1q23.3
III	C6orf29	CTL1	9q31.1
III	C6orf46	KIAA1572	9q33.3
III	C6orf46	KIAA0414	9q33.3
III	C6orf46	ZNF91	19p13.1
III	PPT2	PPT1	1p32
II	TAP2/1	ABCA2	9q34
II	PSMB8/9	PSMB7	9q34.11-q34.12
xII	*RPS18*	*RPS18-like*	*1q22-q23*
xII	*LYPLA2L*	*LYPLA2*	*1p36.12-p35.1*
xII	*RPL12L*	*RPL12*	*9q34*

4.4.4.5 Comparison of the order of L2- and L3-paralogues located on chromosomes 1, 9 and 19

Now that the MHC paralogues have been identified in the proposed paralogous regions on chromosomes 1, 9 and 19 it is interesting to compare the gene order between chromosomes (summarised in figure 4.12).

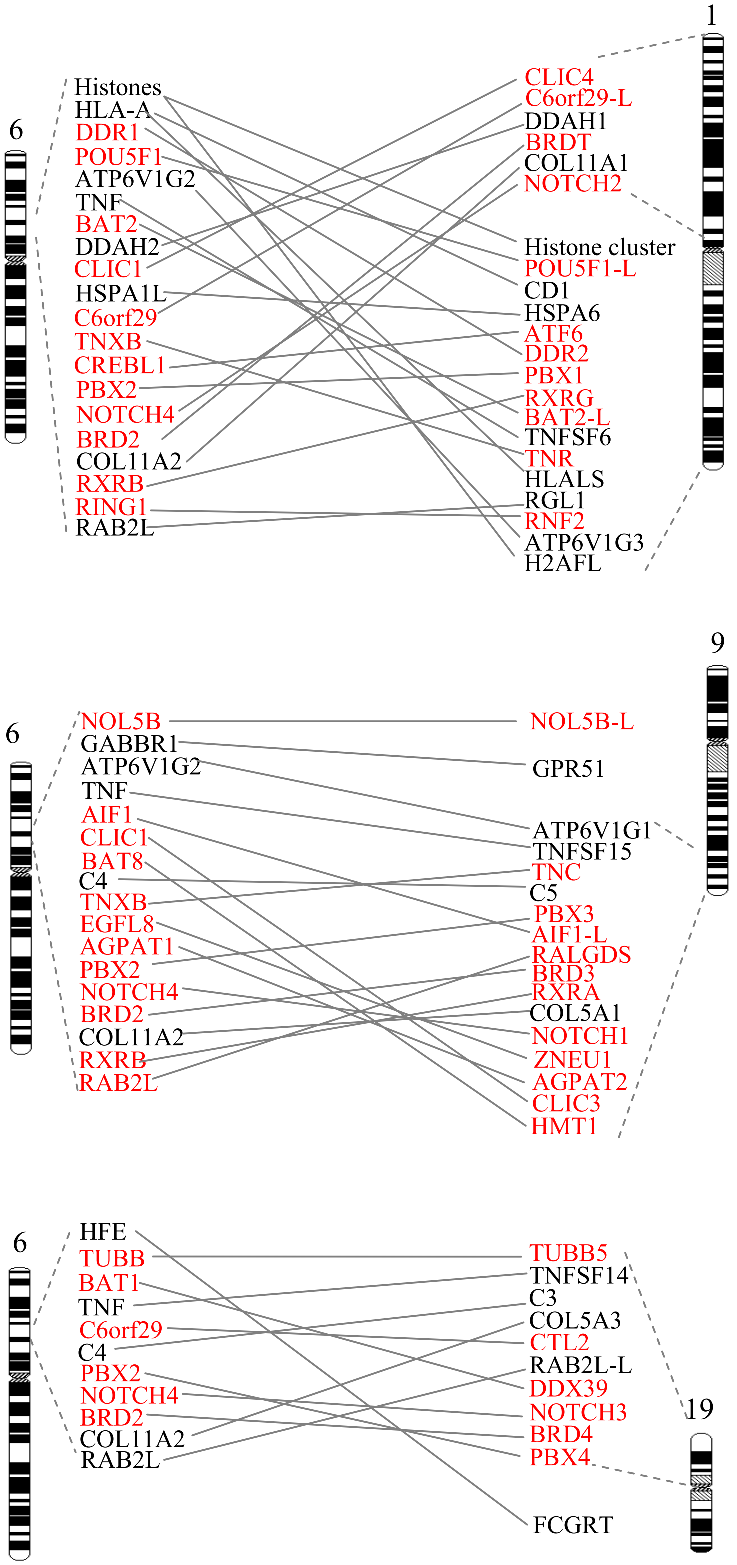


Figure 4.12 Comparison of the order of L2- and L3-paralogues on chromosomes 1, 9 and 19. The gene names in red represent L3-paralogues and the L2-paralogues are shown in black. Continued on next page.

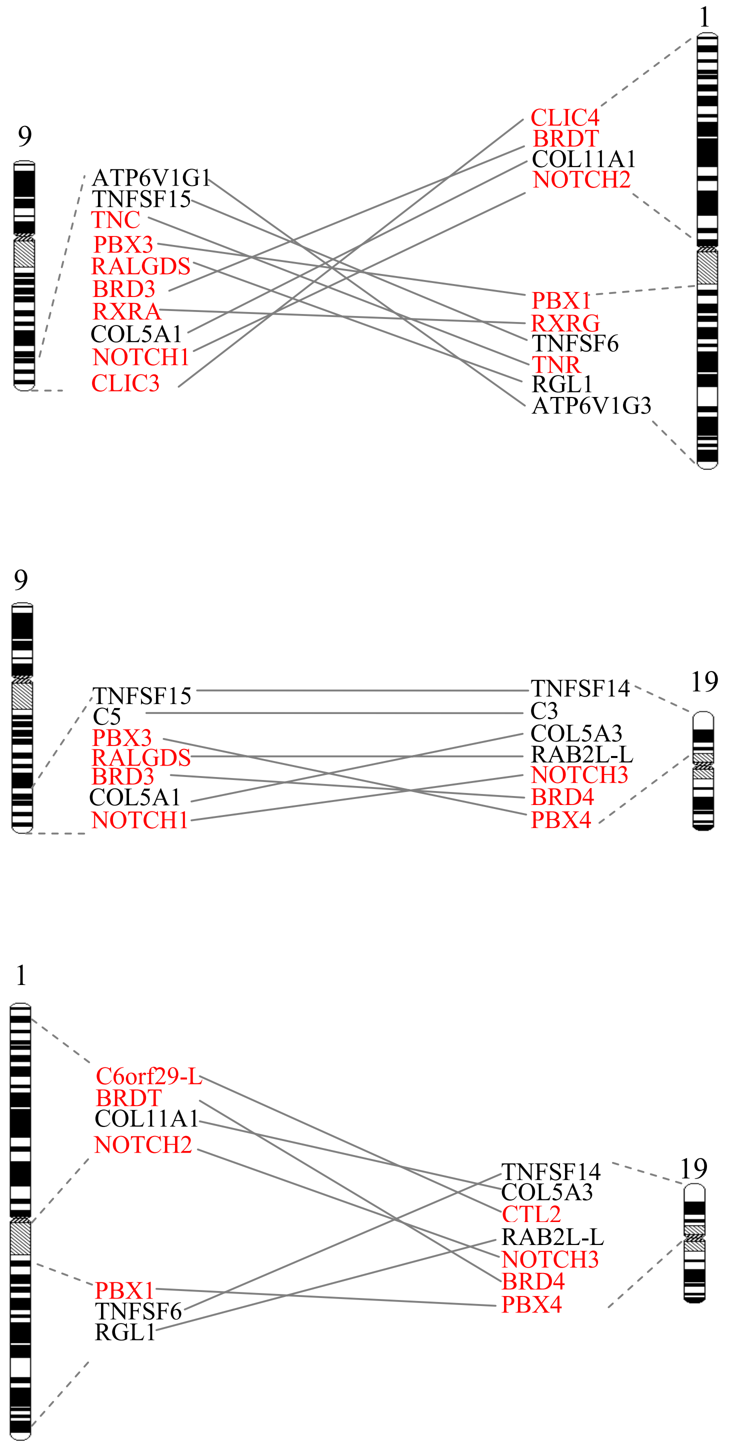


Figure 4.12 Continued. See previous page for legend.

If the four regions did arise by block duplication events, they would be expected to have detectable conservation of gene order (Endo *et al*, 1997). In general, the order of paralogues is not conserved. However, this is not surprising if hundreds of millions of years have passed since their emergence by duplication. The most interesting group of paralogues in this context are the paralogues with copies on all four chromosomes. In total, there are five MHC genes with paralogues that have been conserved on all four paralogous regions; they are NOTCH4, PBX2, COL11A2, BRD2 and RAB2L. Comparison of the gene order of the five genes in each region reveals that they are not strictly conserved (summarised in figure 4.13).

Chromosome 6	TNF-PBX2-NOTCH4-BRD2-COL11A2-RAB2L
Chromosome 1	BRDT-COL11A1-NOTCH2-PBX1-TNFSF6-RGL1
Chromosome 9	TNFSF15-PBX3-RALGDS-BRD3-COL5A1-NOTCH1
Chromosome 19	TNFSF14-COL5A3-RAB2L-L-NOTCH3-BRD4-PBX4

Figure 4.13 Comparison of the MHC paralogues with copies on all four paralogous regions. The MHC genes and corresponding paralogues are represented by the same coloured text.

However, there are pairs of genes that are in the same order on two or more of the chromosomes, such as the paralogues of the chromosomes 6 genes BRD2 and COL11A2 are in the same order on chromosomes 1 and 9. The TNF and PBX2 paralogues on chromosome 9 are also in the same order, whereas the paralogues of PBX2 and NOTCH4 on chromosome 1 are in the reverse orientation. Thus, showing that if the genes did emerge together as part of a series of block duplication events the regions have been subjected to extensive chromosomal rearrangements.

As mentioned in chapter 3, Flajnik and Kasahara (2001) analysed the gene order of all four proposed paralogous regions in the most recent analysis of the MHC paralogues. One of the examples of gene order conservation mentioned in this study involved six paralogues on chromosomes 9 and 19; they are (using chromosome 9 gene symbols) CTL1 (not identified as a true paralogue in this genome survey), TNFSF15, C5, DNM1 (does not have a paralogue in the MHC region therefore not identified in this analysis), BRD3 and NOTCH1. The genome survey presented in this chapter reveals that the order is not conserved overall and that the paralogues of BRD3 and NOTCH1 on chromosome 19 are actually in the reverse order. Therefore, the gene order of the chromosome 19 paralogues (TNFSF14-C3-NOTCH3-BRD4) is identical to the order on chromosome 6 (TNF-C4-NOTCH4-BRD2) rather than chromosome 9 (TNFSF15-C5-BRD3-NOTCH1). In comparison, the order of the equivalent paralogues identified on chromosome 1 (BRDT-NOTCH2-TNFSF6) is actually the reverse order of chromosomes 6 and 19.

4.4.5 Paralogues located outside the paralogous regions

One of the most interesting and novel findings of the whole-genome survey was that not all paralogues are confined to the paralogous regions on chromosomes 1, 9 and 19 but others are scattered throughout the genome (table 4.8).

Table 4.8 Summary of the MHC paralogues located outside the paralogous regions on chromosomes 1, 9 and 19. Cells shaded grey represent paralogues discussed in previous sections.

	MHC gene	*MHC Region*	*Paralogue*	*Locus*	*Confidence*
1	CLIC1	III	CLIC4	1p35.3	3
	C6orf29	III	NM_152697	1p31.1	3
	DDAH2	III	DDAH1	1p22.3	2
	BRD2	II	BRDT	1p22.1	3
	COL11A2	xII	COL11A1	1p21.1	2
	NOTCH4	III	NOTCH2	1p11.2	3
	ATP6V1G2	III	ATP6V1G3	1q31.3	2
	Histone cluster	xI	H2-like	1q42.13	2
2	No L2 or L3 paralogues				
3	No L2 or L3 paralogues				
4	No L2 or L3 paralogues				
5	SMA3L	xI	Novel	5p13.3	3
	Histone	xI	H2AFY	5q31.1	2
	GPX5	xI	GPX3	5q33.1	3
6	TUBB	I	TUBBL	6p25.2	3
	TUBB	I	TUBBL2	6p25.2	3
	CLIC1	III	CLIC5	6p21.1	3
	MAS1L	xI	MAS1	6q25.3	2
7	HSPA1L	III	Genscan prediction	7p21.3	2
	HLA Class I	xI	AZGP1	7q22.1	2
8	No L2 or L3 paralogues				
9	NOL5B	xI	Genscan	9p21.3	3
	GABBR1	xI	GPR51	9q22.33	2
10	TUBB	I	Q8WZ78	10p15.3	3
11	MAS1L	xI	Novel	11p15.4	2
	MAS1L	xI	MRGX3	11p15.1	2
	MAS1L	xI	MRGX4	11p15.1	2
	MAS1L	xI	MRGX1	11p15.1	2
	MAS1L	xI	Novel	11p15.1	2
	MAS1L	xI	MRGX2	11p15.1	2
	MAS1L	xI	Q8TDS7	11q13.3	2
	Histone	xI	H2AFX	11q23.3	2
12	Histone	xI	H2AFJ	12p12.3	2
	TAP2/1	II	ABCB9	12q24.31	3
13	Histones	xI	H2A-like	13q32.3	2
14	HSPA1L	III	HSPA2	14q23.3	2
15	Histones	xI	H2 -like	15q26.1	2
16	TUBB	I	TUBB4	16q24.3	3
17	PSMB9	II	PSMB6	17p13.2	2
	FLOT1	I	FLOT2	17q11.2	3
18	TUBB	I	TUBBL	18p11.32	3
19	HLA Class I	I	FCGRT	19q13.33	2
20	TUBB	I	TUBB1	20q13.32	3
21	CLIC1	III	CLIC6	21q22.12	3
22	RNF5	III	Q96GF1	22q12.2	3
X	CLIC1	III	CLIC2	Xq28	3
	Histones	xI	H2AFB	Xq28	2

In total, there are 43 L2- and L3-paralogues located outside the paralogous regions on 1q21.2-q25, 9q32-q34.3 and 19p13.3-p13.11; corresponding to over 50% of the total number of L2- and L3-paralogues identified. The paralogues located outside the paralogous regions predominately exist as singletons. Singletons are paralogues which are not in clusters or pairs with other paralogues and exist as a single entity in the genome. Nevertheless, there are paralogues located within clusters, for example there is a cluster of paralogues of the MAS1L gene located on chromosome 11p15.1. In addition, another MAS1L paralogue is located on 11q13.3. Chromosome 6 contains four paralogues, of which two are TUBB paralogues located within 70 kb of each other. There is also a CLIC1 paralogue (CLIC4) and a MAS1L paralogue (MAS1) located on the p-arm and q-arm, respectively.

Of the 44 L2- and L3-paralogues located outside the paralogous regions, 32 are novel findings. This corresponds to 89% (32/36) of all the new paralogues identified in this analysis. The chromosome harbouring the largest number of paralogues is chromosome 11 with a total of 8, including the MAS1L paralogue gene cluster. The majority of chromosomes only have one L2- or L3-paralogue; however, chromosomes 5, 7, 12 and 17 contain two to three paralogues. Chromosomes 2, 3, 8 and 12 do not contain any L2- or L3-paralogues.

4.4.6 L0- and L1-paralogues

The L0- and L1-paralogues were identified in the genome survey based on sequence similarity alone. Analysis of the 709 paralogues has revealed that they largely represent homologues with shared domains and are members of a protein superfamily.

For example, the DHX6 gene has 19 L0-paralogues and no paralogues have been identified with higher levels of support. The DHX6 gene is a member of the DEAD box helicase protein superfamily, which is a very large family of proteins with over 60 members identified in the human genome (ENSEMBL NCBI 31). Of the 709 paralogues with the lowest level of support it is expected that only the minority will represent paralogues and the majority will be homologues that share similar domains and are distantly related. For examples, the NR5A2 gene on 1q32.1 was detected as an L0-paralogue of the RXRB gene but it is actually a distant relative. Both the RXRB and NR5A2 genes belong to the nuclear receptor gene superfamily and have the same domains. Therefore, the NR5A2 gene was identified as a paralogue because of sequence similarity to the domain regions, but it is actually a more distant relative.

4.4.7 Caveats associated with my strategy

Paralogues are genes that are found within the same genome and have originated through duplication of an ancestral gene. Immediately after duplication the paralogous genes will be identical; they will have the same exon fingerprint, DNA sequence and code for the same protein. These features have been used in my strategy to identify paralogues in the human genome. However, this type of analysis has its limitations. Over time a number of evolutionary processes may act upon the genomic sequence that will result in changes to the DNA, gene structure and, consequently, the encoded protein. Such processes include exon shuffling and mutations that will render the genes undetectable as paralogues by my strategy. Therefore, paralogues do not necessarily have any sequence similarity at all. This is one of the inherent difficulties of this type of research and the main caveat associated with the strategy I have used to identify paralogous genes, exemplified by the HLA class I-like genes.

There are several HLA class I-like genes located outside the extended MHC region in the human genome; the CD1A-E genes (1q22-q23), AZGP1 (7q22.1), FCGRT (19q13.33) and HLALS (1q25.3) and RAET1E-N genes (6q24.2-q25.3). The CD1 genes, AZGP1, FCGRT and HLALS have previously been cited as putative paralogues. However, they were not all identified in the genome survey because they share low sequence similarity with the five HLA class I and class I-like genes, HFE, HLA-A, HLA-E, MICA and MICB, used in the genome survey (summarised in table 4.9).

Table 4.9 Summary of the P-values obtained for the HLA class I-like genes (column 1) from the BLAST similarity search using HFE, HLA-A, HLA-E, MICA and MICB, and the percentage sequence identities (%ID) determined from a global sequence alignment. The four HLA class I-like genes identified as paralogues in the genome survey, and the corresponding P-values and % IDs, are in red. The shaded boxes denote that the HLA class I-like gene was not detected by BLAST search using the MHC encoded protein sequence, therefore no P-values was obtained.

HLA class	***HFE***		***HLA-A***		***HLA-E***		***MICA***		***MICB***	
I-like gene	***P-value***	***%ID***	***P-value***	***%ID***	***P-value***	***%ID***	***P-value***	***%ID***	***P-value***	***%ID***
CD1A	**2.6e-05**	**23.9**		25.10	0.014	24.70	0.054	24.1	0.58	23.0
CD1B		20.8		23.80		25.40		23.8		22.4
CD1C		22.3		23.70		22.40		23.4		22.9
CD1D	0.012	25.6	0.013	23.90	0.12	24.80	0.097	24.5	0.98	22.2
CD1E		22.1		25.80		22.40		24.3		21.2
HLALS	**3.5e-31**	**38.8**	**5.8e-38**	**37.20**	**3.3e-38**	**39.10**	**1.1e-18**	**30.4**	**3.2e-15**	**27.0**
RAET1E		23.6		22.70		18.80	0.039	27.7		28.9
ULBP2	0.013	26.4		24.60	0.999950	28.00		26.9		22.9
ULBP1		24.2		24.70		26.10		22.5		23.0
RAET1L		27.7		24.10		26.90		26.7		24.1
ULBP3		24.7	0.47	26.60	0.31	26.00		26.5		25.9
AZGP1	**8.3e-29**	**37.5**	**1.9e-34**	**38.80**	**3.8e-29**	**35.70**	**5.4e-08**	**30.0**	**2.5e-07**	**29.2**
FCGRT	**8.6e-10**	**29.4**	**2.1e-09**	**31.20**	**1.3e-13**	**31.30**	**7.9e-06**	**27.2**	**0.00059**	**24.3**

The percentage amino acid sequence identities between the class I like protein sequence and the protein sequences used in the analysis (HFE, HLA-A, HLA-E, MICA and MICB) are within the ‘Twilight Zone’ of homology; described as between 15% and 25% amino acid identity (Doolittle *et al*, 1986). The BLAST algorithm used in this analysis, WU-BLAST2, is capable of detecting almost all relationships between proteins whose sequence identities are greater than 30% but is only 50% effective when the proteins have sequence identities between 20 and 30% (Brenner *et al*, 1998). Thus, it is not unexpected that the HLA class I like genes were not detected at all or only detected with a low P-value (summarised in table 4.9). In cases like these, with low sequence similarity, the Position-Specific Iterated BLAST (or PSI-BLAST) program could have been used. This is the most sensitive BLAST program and is designed to detect more distantly related proteins. However, at this time it was not possible to use this program to search the assembled human genome sequence in ENSEMBL.

To summarise, HLALS, AZGP1, FCGRT and CD1A were the only HLA class I-like genes identified as paralogues in this genome survey. Interestingly, other members of the CD1 cluster and members of the RAET1/ULBP gene cluster were found by the BLAST similarity search, although the P-values for members of this gene cluster were more than the designated BLAST cut-off, 10^{-5}, and were therefore eliminated from the analysis. Thus, indicating that, in this case, the search criteria were too strict to detect this family of HLA class I-like genes. Using the gene architectural information the HLALS, AZGP1, FCGRT and CD1A genes were classified as L2-paralogues. No HLA class I-like genes were classified as L3-paralogues as the detectable homology is restricted to the shared immunoglobulin domain only.

The RAET1/ULBP genes are a novel family of HLA class I-like genes located outside the MHC region (Radosavljevic *et al*, 2002). Although they are recognised as being related to the HLA class I genes they have not been identified as paralogues in this, or any other analysis performed to-date. In order to determine the relationship between the HLA class I genes and the RAET1-N genes an independent FINEX analysis was performed using RAET1N (alias ULBP3) which was detected by BLAST analysis of HFE, HLA-A and HLA-E protein sequences, albeit with very high P-values. The highest z-scores, not exceeding 4.5, were to the PROCR gene, located on 20q11.2, and to the HLA class II genes located within the MHC region. The PROCR gene is an endothelial protein involved in the blood coagulation pathway which shares the same tridomain backbone as the HLA class I and class I-like genes. The HLA class II genes are believed to have arisen from the same ancestral gene but have since undergone significant expansion by gene duplication (reviewed by Beck and Trowsdale, 2000), thus the original HLA class I and class II genes are paralogous. In addition, gene structure homology was also detected for members of the CD1 gene cluster with z-scores approximately 3.5, to FCGRT with a z-score of approximately 3.2 and to HLA-A with a z-score of 3.4. Thus, showing that the relationship of this complex family of HLA class I genes and HLA class I-like genes cannot be determined using sequence similarity alone and demonstrates the importance of using additional criteria to detect paralogous relationships, such as exon fingerprints.

4.5 Discussion

The genome-wide survey presented in this chapter shows the true distribution of MHC paralogues in the human genome. Not only has this piece of research confirmed that there are regions on chromosomes 1, 9 and 19 that contain clusters of MHC paralogues but I have also shown that there are paralogues located throughout the human genome. Furthermore, I have also presented a novel method to identify and classify the MHC paralogues, in which the paralogues are initially identified based on sequence similarity, but, by applying additional knowledge of gene structure and domain content, paralogues with increasing levels of confidence (L0>L1>L2>L3) are identified.

In total, 82 L2- and L3-paralogues of genes located within the MHC region were identified in the genome, corresponding to 29 MHC gene families. Almost 50% are located within the paralogous gene clusters on 1q21.2-q25.3, 9q32-q34.3 and 19p13.3-p13.1. Analysis of the paralogous genes within the clusters on 1, 9 and 19 defined the boundaries of these paralogous gene clusters as 1q21.2-q25.3, 9q32-q34.3 and 19p13.3-p13.1. As discussed in the literature (such as Kasahara 1999a) there is also a smaller cluster of six paralogues located on the short arm of chromosome 1, which span over 95 Mb of genomic sequence. The paralogous gene cluster on the q-arm spans from 1q21.2 to 1q25.3 and contains one histone cluster, the CD1 gene cluster and 12 single MHC paralogues and encompasses approximately 35 Mb of genomic sequence. The region 9q32-q34.3 spans approximately 24 Mb and 19p13.3-p13.1 encompasses almost 14 Mb, each cluster contains 15 and 10 MHC L2- or L3-paralogues, respectively. In concordance with my results, McLysaght and co-workers (2002) conducted an analysis of the entire draft human genome sequence in order to identify paralogons, or pairs of regions containing duplicated genes. The most extensive region paired 41 Mb of chromosome 1q, including the tenascin paralogue TNR, with a 20 Mb region of chromosome 9q, including the TNC gene.

The existence of four paralogous gene clusters suggests that they have a common origin by either two rounds of large-scale block duplication or even as part of the whole-genome duplication events originally proposed by Ohno (1970). Interestingly, a single related cluster of genes orthologous to the MHC paralogues located in two or more of the clusters on 1, 6, 9 and 19 has been identified in amphioxus (reviewed by Flajnik and Kasahara, 2001) and linkage between orthologues of MHC region genes has also been observed in *Drosophila* (Danchin *et al*, 2003). The region in amphioxus is believed to be the closest living example of the ancestral region of 1q21.2-q25.3, 6p22.2-p21.3, 9q32-q34.3 and 19p13.3-p13.1, as this organism is ideally situated at the base of the vertebrate lineage and predates the duplication events proposed by the 2R hypothesis. Therefore, once the complete amphioxus genome sequence is available it will be of interest to determine which genes were involved in the genome-duplication events.

If the MHC paralogues within the regions on chromosomes 1, 9 and 19 did have a common origin there should be detectable synteny between them. However, comparison of gene order within the paralogous regions revealed that the order is not strictly conserved. The lack of synteny between the paralogous regions raises a counterpoint to the hypothesis that these four regions arose simultaneously as part of a block or whole-genome duplication event: it may be that they have duplicated individually and are clustered because of a selective reason (Hughes, 1998) or that there has been extensive chromosomal rearrangement since the block/whole-genome duplication events. There is strong evidence to support the latter explanation. For example, duplicons have been identified in both the MHC and 9q32-q34.3, and there is also evidence of a recent pericentromeric inversion on chromosome 1 resulting in the rearrangement of the genes on the chromosome.

One of the most interesting and novel findings of the whole-genome survey was that over 50% of the MHC paralogues are not located within clusters but are scattered throughout the genome, largely as singletons. No further clusters of genes paralogous to different MHC genes were identified, but small clusters of members of the same MHC paralogous gene family were identified, for example there is a cluster of six MAS1L paralogues on the short arm of chromosome 11 suggesting that this gene family has expanded by local duplication events. Of the 44 L2- and L3-paralogues located outside the paralogous regions, 32 are novel findings. This corresponds to 89% (32/36) of all the new paralogues identified in this analysis.

The existence of paralogues located outside the regions on chromosome 1, 9 and 19 suggests a more complex history than that previously proposed - the origin of the paralogues will be addressed in more detail in chapter 5. One thing that is clear is that not all MHC genes have paralogues in the human genome; this corresponds to approximately one-third of the genes used in the analysis (40/128). There are two hypotheses to explain why some genes do not have paralogues, these are; (1) there has been extensive gene loss or silencing since the large-scale duplication of the ancestral region or (2) not all MHC genes were involved in the proposed large-scale duplication events. This issue should be resolved upon analysis of the gene contents of ‘key’ organisms in the vertebrate lineage, such as amphioxus, hagfish and lamprey once the complete genomic sequence is available.