Chapter 3

Characterisation of 9q32-q34.3

3.1 Introduction

The extended Major Histocompatibility Complex (MHC), on 6p22.2-p21.3, is a gene rich region that has taught us a great deal about the evolutionary dynamics of a chromosomal segment (The MHC Sequencing Consortium, 1999; Beck and Trowsdale, 2000). It has been proposed that the three chromosomal regions 1q21-q25, 9q33-q34 and 19p13.3-p13.1 in the human genome are paralogous to the MHC region (Sugaya *et al*, 1994; Kasahara *et al*, 1996a; Katsanis *et al*, 1996; Sugaya *et al*, 1997; Kasahara, 1999a; Kasahara, 1999b; Kasahara *et al*, 2000; Flajnik and Kasahara, 2001).

MHC paralogy was first observed by Sugaya and co-workers in 1994 during the analysis of three MHC class III genes in the human genome. Two years later, in 1996, the proteasome Z subunit (PSMB7), a paralogue of the PSMB8 and PSMB9 genes, was mapped to mouse chromosome 2, which is syntenic to 9q34 in humans (Kasahara *et al*, 1996a). PSMB7 is involved in the generation of cytosolic peptides by MHC class I molecules and, on closer inspection of the mouse loci adjacent to the region containing this gene, ten more paralogues representing MHC gene families were identified; including ABCA2, a putative paralogue of the TAP1 and TAP2 genes that are also involved in MHC class I peptide processing. Independently, Katsanis and colleagues (1996) found that the MHC and 9q33-q34 regions are paralogous and also identified two additional regions in the human genome, 1q21-q25 and 19p13.3-p13.1,

containing clusters of genes with related copies in the MHC.

The combined list of genes from both studies indicates that there are ten MHC genes with paralogues in one, two or three of the proposed paralogous regions on chromosomes 1, 9 and 19 (table 3.1).

Table 3.1 Summary of the first MHC paralogues identified in three other regions of the genome

Chromosome 6	***Chromosome 9***	***Chromosome 1***	***Chromosome 19***
BAT2	BAT2 exon	-	-
COL11A2	-	COL11A1	-
HSPA1A/B/L	-	HSPA6/HSPA7	-
NOTCH4	NOTCH1	NOTCH2	NOTCH3
PBX2	PBX3	PBX1	-
RXRB	RXRA	RXRG	-
TNX	TBC	TNR	-
C4	C5	C3	-
TAP1/2	ABC2	-	-
LMP2/7	PSMB7	-	-

The paralogues of MHC genes and their genomic locations were initially identified using mapping data and it was important to clarify these findings using sequence data. Compared with the MHC region the proposed regions on chromosomes 1, 9 and 19 containing the MHC paralogues are much less characterised and, in order to truly understand the evolution of these proposed paralogous regions and the human genome, it is important to have finished, contiguous genomic sequence. With this in mind, and the progress of the mapping and sequencing of chromosome 9, the initial focus of the project was on the characterisation of the proposed paralogous region on 9q32-q34.3. This chapter describes my contribution to the mapping, sequencing and characterisation of 9q32-q34.3 and compares this chromosomal region with the extended MHC region.

3.2 Results

3.2.1 Identification of genes on 9q32-q34.3

The localisation of the MHC paralogues and other genes to 9q32-q34.3 was initially determined using the physical and genetic mapping data available for chromosome 9. The mapping and sequencing of chromosome 9 was carried out by the Chromosome 9 Mapping and Sequencing groups at the Wellcome Trust Sanger Institute in collaboration with the chromosome 9 community. The chromosome 9 project followed the clone-by-clone approach where bacterial clones were initially isolated by screening the human BAC libraries RPCI 11 and 13. The clones were mapped to this region using a landmark map consisting of approximately 15 markers per Mb, first constructed using whole genome radiation hybrid mapping (Walter *et al*, 1994; Hudson *et al*, 1995), incorporating available markers, including STSs (sequence tagged sites), ESTs (expressed sequence tags), polymorphic microsatellites and gene based markers from GeneMap99 (Deloukas *et al*, 1998).

At the start of this project, in October 1999, the sequencing of the human genome was still in its early stages and less than 5% of the genomic sequence was available. The region 9q32 to 9q34.3, at the telomere of chromosome 9, had less than 2% draft sequence coverage and was split into 13 contigs of various sizes. As the region was largely unfinished, the chromosome 9 mapping data available in the Sanger in-house ACE database '9ace' was interrogated using 103 genes, identified in LocusLink, HUGO and GeneMap99, known to map to this region. Initial searches anchored 60% of known genes, including all 10 proposed paralogues to chromosome 9q32-q34.3 clones and contigs. This was done using electronic PCR (as described in section

2.15.1) on available cDNA sequences and BLAST sequence similarity searches against the genomic databases. In addition, several genes had previously been mapped to this region and were already incorporated into the chromosome 9 database and were therefore identified by querying '9ace'.

3.2.2 Mapping of the Olfactory Receptor gene cluster to 9q33.1-q34.12

Sequence similarity search identified the BAC clone, bA465F21 (AC006313), as containing several olfactory receptor genes (ORs). The clone was being sequenced by another sequencing centre, the Whitehead Institute, and draft sequence was publicly available. Initial mapping information mapped this clone to 9q34 according to sequence similarity with three STS markers (stSG69605, stSHGC-9207 and stAFMa239xe9). As an olfactory receptor gene cluster is located in the most telomeric region of the MHC, the MHC extended class I region, but no cluster had previously been identified in the paralogous regions it was important to determine the exact location of this clone. In addition, identification of an OR cluster within this region may help determine the boundaries of the paralogous region.

There are clusters of olfactory receptor genes located throughout the genome (Rouquier *et al*, 1998), therefore, it was important to confirm that bA465F21 mapped to 9q34 and determine the precise chromosomal location. Using fluorescent *in situ* hybridisation (FISH) of metaphase chromosomes, low resolution mapping of bA465F21 chromosome 9 was achieved. The genomic clone was fluorescently labelled and hybridised to metaphase chromosomes revealing the approximate location of this fragment of genomic sequence as shown on figure 3.1.

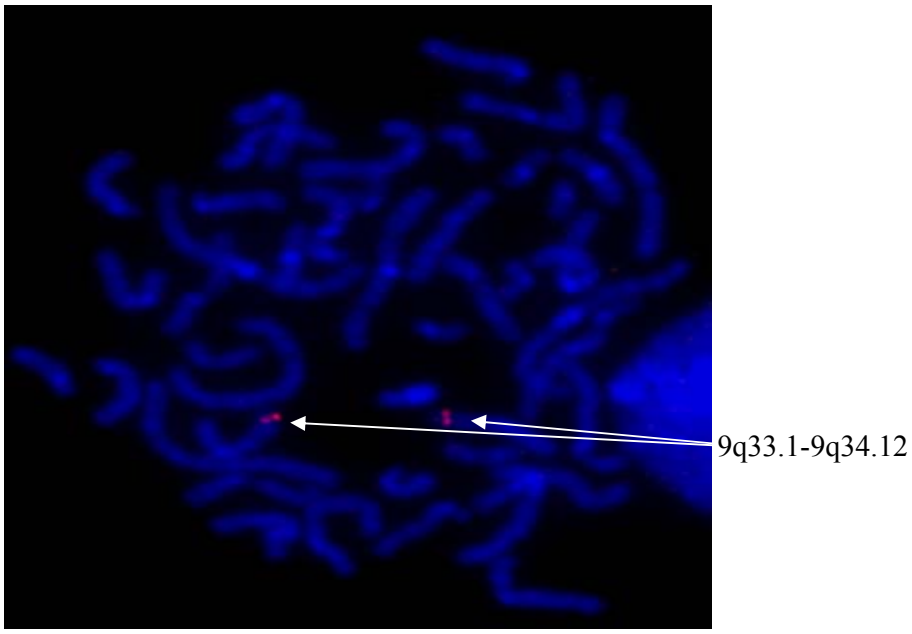

Figure 3.1 FISH analysis of bA465F21. Arrows indicate the location of the clone on the two copies of chromosome 9.

Upon comparison of the clone position with standard chromosome bands the clone was anchored to 9q33.1-q34.12. However, the precise location within the clone-contig map could not be determined. This was achieved using *Hind* III restriction digest fingerprinting (as described in section 2.4). The restriction pattern of bA465F21 was created (figure 3.2A), compared against other 9q33.1-q34.12 clones, and the degree of overlap between clones calculated according to shared restriction sites (figure 3.2B). Based on the comparison of restriction patterns, clone bA465F21 was localised to contig 100 on 9q33.2 overlapping clones bA64P14 and bA163B6 (figure 3.2C).

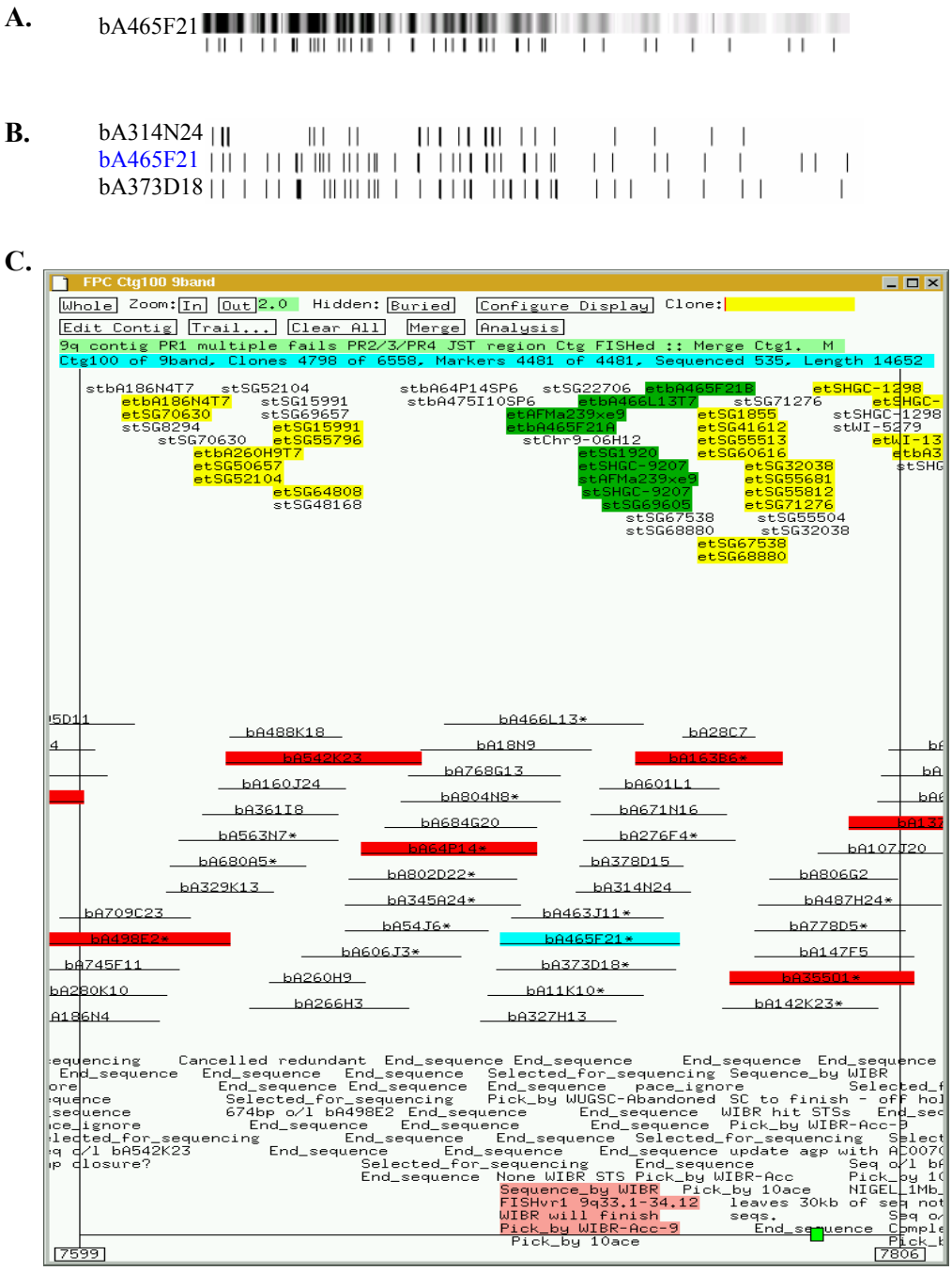

Figure 3.2 Localisation of the clone bA465F21 to the chromosome 9 tiling path. The gel image and the computationally determined fingerprint generated by *Hind* III restriction digest (A) was compared with the fingerprints of other chromosome 9 clones (B). Overlapping bands with clones bA314N24 and bA373D18 localised the clone to contig 100 of chromosome 9 (C). The screen shot of the chromosome 9 FPC database (C) shows the marker information; those highlighted in green correspond to markers present in the clone bA465F21, which is highlighted in blue. Other clones involved in the tiling path across the contig are highlighted in red, and corresponding marker data is highlighted yellow. Useful information regarding bA465F21 is highlighted pink.

With confirmation that bA465F21 was located within 9q32-q34.3 it was placed in the minimum tiling path covering the region. The 186555 bp sequence of the clone was completely sequenced and finished at the Whitehead Institute. NIX analysis identified seven olfactory receptor genes and, upon investigation of sequences corresponding to the overlapping clones, a further nine olfactory receptor genes were identified. This gene cluster was found to be uninterrupted by any other genes thus forms a novel olfactory receptor gene cluster on 9q32 spanning 324109 bp.

Further analysis of chromosome 9 identified a single olfactory receptor (OR) gene approximately 3.2 Mb centromeric of the 9q32-q34.3 paralogous region boundary (on clone bA386D8). The existence of the single OR gene and a cluster in the paralogous region resembled the arrangement of the two OR gene clusters (a major and a minor one) in the extended class I MHC region. As the extended class I region is one of the flanking regions of the MHC, this arrangement suggested that this could be the boundary of the chromosome 9 paralogous region. However, the identification of an additional cluster of approximately 10 OR genes on 9q31.1, approximately 6.6 Mb centromeric of the single OR gene, revealed that the olfactory receptor genes could not be used as a reliable source when defining the boundaries.

3.2.3 Identification of the Allograft Inflammatory Factor 1 (AIF1) paralogue

Ab initio analysis of the sequence available in draft format for clone bA544A12 identified a putative paralogue of the allograft inflammatory factor 1 (AIF1) gene (figure 3.3).

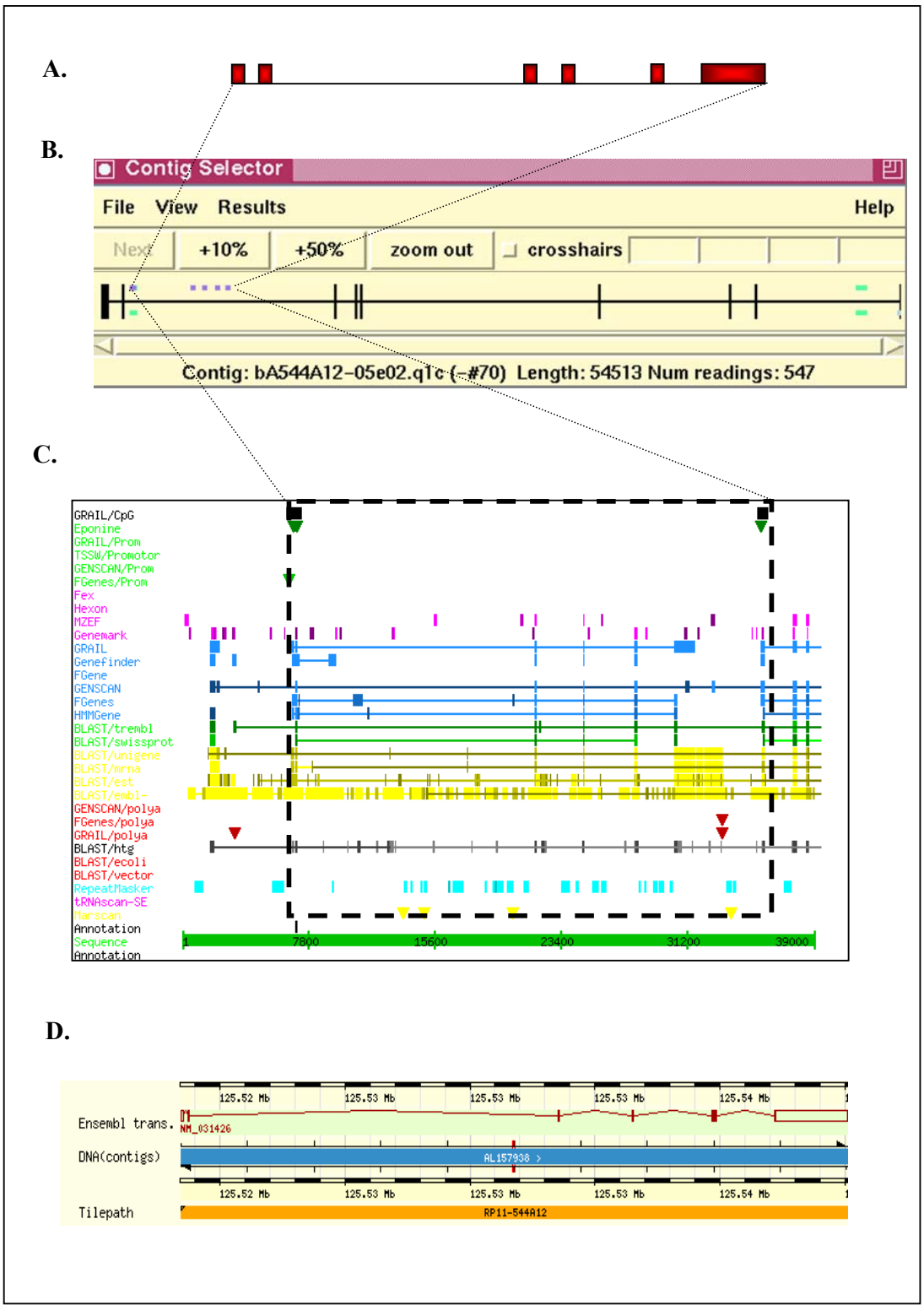

Figure 3.3 Computational identification of the AIF1 paralogue. The six exons of the AIF1 paralogue on chromosome 9, as shown in (A) were determined during sequence assembly (B) and NIX analysis (C). The gene structure has since been confirmed in ENSEMBL (D).

The genomic structure of the AIF1 paralogue (figure 3.3.A) was determined and confirmed using a number of bioinformatics tools (figure 3.3). During sequence assembly the exons and CpG islands (associated with the start of a gene) were predicted within the GAP4 database (figure 3.3.B). In total, six exons were predicted (shown in purple in figure 3.3.B) by the GAP4 software with the start of the gene (characterised by 'ATG') located adjacent to a predicted CpG island (shown in green in figure 3.3B). NIX analysis (figure 3.3.C) determined the coding region, spanning from the start to the stop of the gene, and detected protein sequence identity of 64% to the human (P55008), 61.3% to mouse (O70200), and 62% to rat (P55009) AIF1 genes.

The AIF1 transcript located on 6p22.2-p21.3 has a length of 661 bp, spans 1.79 kb, and encodes for a 147 amino acid protein. In contrast, the AIF1 paralogue located on 9q32-q34.3 encodes for a 150 amino acid protein and has a transcript length of 3381 bp spanning 26.62 kb of the genome. Both paralogues have conserved gene structure with similar exon sizes, but the intron sizes vary greatly with the average intron length of the chromosome 6 paralogue being 0.2 kb compared to 4.6 kb on chromosome 9. The exon and intron sizes are summarised in table 3.2.

Table 3.2 Summary of exon and intron sizes and comparison of splicing phases of the two AIF1 paralogues. The exon and intron sizes are shown in nucleotides.

EXON	***6p22.2-p21.3 AIF1***			***9q32-q34.3 AIF1***		
	Exon	Intron	Phase	Exon	Intron	Phase
1	25	166	1	31	141	1
2	62	87	0	62	14733	0
3	67	367	1	67	2913	1
4	42	198	1	42	3137	1
5	163	309	2	163	2316	2
6	85			88		

The exon and intron boundaries and splice phases were determined by aligning the cDNA sequence with the genomic sequence and were annotated using the GT/AG rule (summarised in table 3.2; Padgett *et al*, 1986). Both genes have identical exon splice phases indicating the importance of conservation of the gene structure. Conservation at the protein level has also been maintained as they share 64% sequence identity and both contain the sequence encoding for an EF-hand domain, which is involved in calcium binding (figure 3.4).

```
                                                                      X
AIF1    MS--QTRDLQGGKAFGLLKAQQEERLDEINKQFLDDPKYSSDEDLPSKLEGFKEKYMEFD
AIF1L   MSGELSNRFQGGKAFGLLKARQERRLAEINREFLCDQKYSDEENLPEKLTAFKEKYMEFD
        **   :. :***********:**.** ***::** * ***.:*:**.** .*********

         Y Z-Y-X -Z
AIF1   LNGNGDIDIMSLKRMLEKLGVPKTHLELKKLIGEVSSGSGETFSYPDFLRMMLGKRSAILK
AIF1L  LNNEGEIDLMSLKRMMEKLGVPKTHLEMKKMISEVTGGVSDTISYRDFVNMMLGKRSAVLK
       **.:*:**:******:**********:**:*.**:.* .:*:** **:.********:**

AIF1    MILMYEEKAREKE-KPTGPPAKKAISELP
AIF1L   LVMMFEGKANESSPKPVGPPPERDIASLP
      :::*:* **.*.. **.***.:: *:.**
```

Figure 3.4 ClustalX sequence alignment of the two AIF1 paralogues. The protein sequence encoded by the six exons are alternatively coloured red and blue. The asterix symbol '*' indicates identical residues, ':' shows highly conserved residues, '.' is used for weakly conserved residues and no symbol indicates no conservation (Chenna *et al*, 2003). The residues corresponding to the EF-hand domain are shown in bold and labeled X, Y, Z, -Y, -X and –Z.

Whole-genome assembly of the human genome sequence in the ENSEMBL genome browser subsequently confirmed the structure of the AIF1 paralogue (figure 3.3.D) and also enabled the annotation of the surrounding genes. The genomic clone bA544A12 was sequenced and finished in its entirety to identify adjacent genes. In

total, 5079 reads, of which 72.8% were of good quality, were used to assemble the 238131 bp sequence of the clone, which was submitted to the EMBL database under accession number AF157938. The clone bA544A12 had previously been mapped to 9q34.12 and was anchored within a contig containing the BAT2 paralogue, which is a neighbouring gene of AIF1 in the MHC class III region. Therefore, it was essential to identify adjacent genes which might be putative paralogues and further examine the degree of shared synteny between these two chromosome regions (6p22.2-p21.3 and 9q32-q34.3).

NIX analysis did not identify any further paralogues on this genomic clone but did identify a 36 exon gene encoding a nuclear pore complex protein (NUP214). The nuclear pore is a large structure that extends across the nuclear envelope and the protein encoded by NUP214 is required for cell cycle progression and nucleocytoplasmic transport. The 3-prime portion of the gene forms a fusion gene with the DEK gene located on chromosome 6 (6p22.3) in a t(6,9) translocation associated with myeloid leukaemia, providing evidence of the fragile nature of this paralogous region. In addition, the most 3-prime exon (approximately 3 kb) of the 28 exon laminin gamma-3 precursor, LAMC3, was identified on this clone. Laminin is a complex glycoprotein consisting of three different polypeptide chains (alpha, beta and gamma) that bind cells via a high affinity receptor and it is thought to mediate the attachment, migration and organisation of cells into tissues during embryonic development.

3.2.4 Problems associated with using mapping data and draft sequence

Investigation of the shared synteny between the MHC region and 9q32-q34.3 using

the available mapping and sequencing data revealed a number of paralogues previously not cited in the literature. For example, a putative paralogue of the extended class I MHC gene GPX5 was localised to clone bA18B16 using mapping data. This clone had been mapped to a contig on chromosome 9q33 centromeric to the olfactory receptor gene cluster on 9q33.2. Therefore, the identification of another extended class I gene paralogue would enable the boundaries of 9q32-q34.3 to be clarified. In order to confirm this prediction, the clone was successfully sub-cloned into pUC vectors and the inserts were sequenced.

In total, 5225 reads were generated and used in the assembly of the genomic clone (AL157702). NIX analysis of the finished sequence to identify genes on this clone did not identify a GPX5 paralogue but did reveal a 'Novel' gene (0.4 kb transcript) with homology to a cDNA clone in mouse (Q921Q2) of unknown function, which did not share significant sequence similarity with GPX5. These findings were also confirmed in the assembly of the whole genome in the ENSEMBL genome browser thus demonstrating the importance of having finished sequence. As sequence became available, computational analysis of the surrounding region did not identify the paralogue on the overlapping clones (figure 3.5).

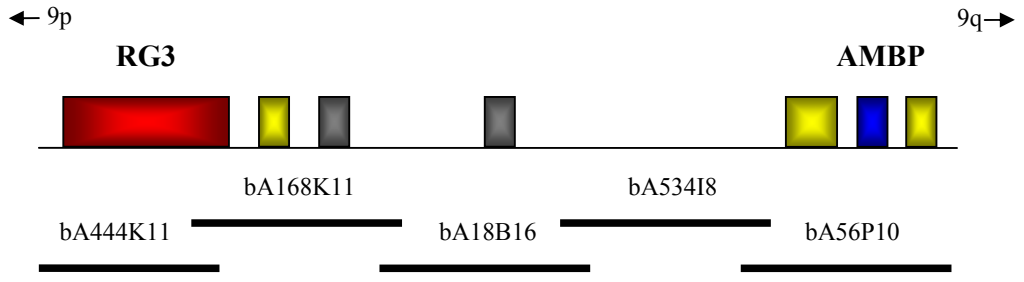

Figure 3.5 Overview of the gene content of region analysed to identify a putative GPX5 paralogue. The two known genes are labeled RG3 (red) and AMBP (blue) and novel genes are shown in grey and genes with no assigned name are shown in yellow (the latter refer to database entries NM_152575, Q8TF49 and NM_18424 from left to right). No GPX5 paralogue was identified.

3.2.5 Orientation of contigs containing putative paralogues

Conservation of gene order provides an insight into the evolution of the identified paralogous genes. If the genes had evolved by whole-genome duplication, or as part of a block duplication event, the overall gene order may still be visibly conserved. The determination of gene order on chromosome 9q32-q34.3 was hindered by the number of gaps in the physical map. For example, two putative paralogues, BRD3 and RALGDS, were identified on separate contigs on 9q34.2 but the orientation and order of the contigs in this region had not been confirmed. It was therefore necessary to determine the order of the contigs to ultimately determine the order of these two paralogues. This was achieved using interphase and fibre fluorescent *in-situ* hybridisation.

During interphase chromosomes are at their most unpacked allowing higher resolution mapping of clones to be achieved compared with metaphase chromosomes. In order to orientate the contigs containing the two paralogues of interest three clones were selected: one from the contig containing BRD3 (bA317B10), one from the contig containing RALGDS (bA244N20) and another neighbouring contig (bB97D14). Each clone was fluorescently labelled using two different dyes, Texas Red (red) and FITC (green), and hybridised in different combinations to interphase chromosomes. The resulting combinations of clones labelled in the different dyes enabled the exact order of the clones to be determined (figure 3.6.A and B).

In order to clarify the order of the clones and determine the distance between them, the labelled clones were hybridised against extended DNA fibres (figure 3.6.C). The precise order and distance between the clones and contigs was determined and the precise location of the clones clarified. The resulting order of the genes was, from

centromere to telomere, RALGDS-BRD3 separated by one contig and two 100 kb gaps (figure 3.6.D).

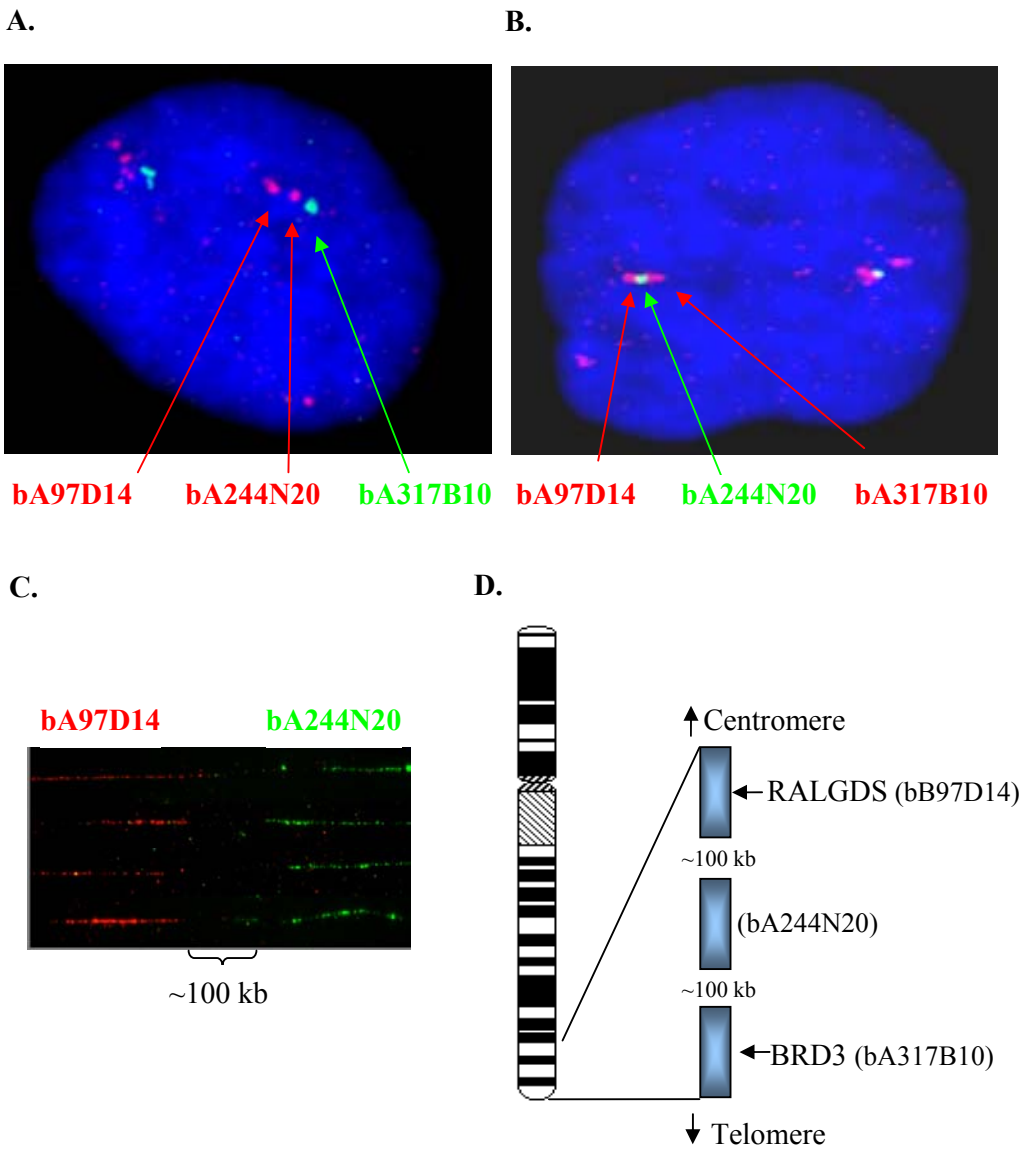

Figure 3.6 Overview of methods used to order and orientate the contigs containing RALGDS and BRD3 putative paralogues. Three clones representing three different contigs were labelled using fluorescent dyes then hybridised in different combinations to interphase chromosomes (A and B). By analysing the order of the labelled clones in the different combinations the contig order could be determined. The clones were then used to hybridise against chromosome fibres which revealed the distance between the clones (C). The clone bB97D14 had previously been anchored to a contig centromeric of this region and using this information the precise order, orientation and gap sizes were determined (D).

3.2.6 Current status of 9q32-q34.3

The putative MHC paralogous region in August 2003 is in six contigs with gap sizes ranging from less than 5 kb to 200 kb (the latter is approximately the size of a BAC clone insert). The current status is summarised in figure 3.7.

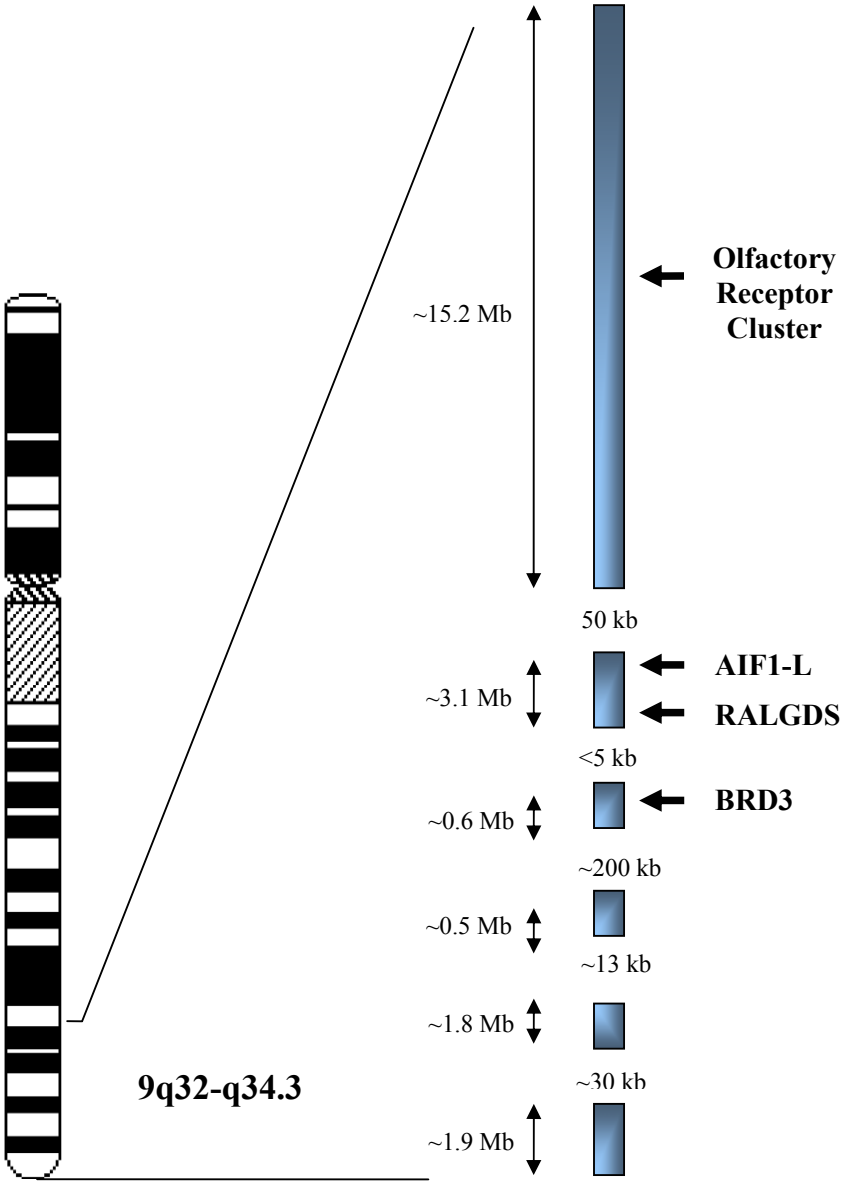

Figure 3.7 Schematic representation of the status (August 2003) of the MHC paralogous region on 9q32-q34.3.

3.2.7 Comparison of the MHC paralogous region on 9q32-q34.3 and the MHC region on 6p22.2-p21.3

3.2.7.1 Gene and paralogue content

The region 9q32-q34.3 contains 198 fully sequenced clones that have all been individually analysed using NIX and ENSEMBL to identify known genes and paralogues (summarised in Appendix 1). Now that the human genome sequence has been assembled using the official minimum tile-path clones (previous human genome assemblies did not use the tile-path clones being sequenced and finished by the various genome centres), the locations of genes identified in NIX and early ENSEMBL freezes, based on sequence similarity and gene prediction programs, could finally be confirmed and integrated into the 9q32-q34.3 gene list. In total, 322 genes have been identified, of which 178 are known genes, 44 are 'Novel' genes characterised by the ENSEMBL genome browser, 24 have no assigned name, 36 are hypothetical proteins and 40 are putative MHC paralogues (see Flajnik and Kasahara, 2001 for most recent paralogue list).

The proposed paralogous region on chromosome 9 is gene rich (one gene per 73 kb) compared with the rest of the chromosome (one gene per 129 kb) (summarised in table 3.3). The gene dense nature of 9q32-q34.3 is comparable to the MHC region on 6p22.2-p21.3 that has approximately one gene per 33 kb, which is high when compared to the chromosome 6 average of one gene per 132 kb. Overall, 9q32-q34.3 is a less gene dense region as opposed to the MHC region on 6p22.2-6p21.3; however, the gene density is still greater than the genome average of approximately one gene per 100 kb (table 3.3).

Table 3.3 Summary of the gene content and sizes of chromosomes 6 and 9 and the paralogous regions compared with genome average.

Chromosome or region	**Number of genes**	**Size (Mb)**	**Approximate gene density**
Chromosome 6	1296	170.67	1 gene per 132 kb
6p22.2-p21.3	222	7.22	1 gene per 32 kb
Chromosome 9	1031	132.88	1 gene per 129 kb
9q31.2-q34.3	322	23.78	1 gene per 73 kb
Human genome	~30,000	3,000	~1 gene per 100 kb

The distribution of the paralogues, including distances between proposed paralogues and the number of interspersed genes, has been summarised in figure 3.8. Identification of the MHC paralogous genes has determined that the proposed paralogous region spans approximately 24 Mb from 9q32 through to 9q34.3. Within this region, paralogues represent 12.4% of the total gene repertoire (39/322). In comparison, the MHC region spans approximately 7.2 Mb of which the cited paralogues represent almost 18% of the total gene repertoire (40/222).

It has been noted that the order of genes within some paralogous regions has been conserved, namely in the case of the Hox gene clusters (Garcia-Fernandez and Holland, 1994). Initial analysis of the proposed MHC paralogous regions indicated that the gene order of the paralogues located using mapping data was not conserved (Katsanis *et al*, 1996). Now that the precise locations of the putative paralogues have been determined, a full comparison of the gene order is possible (figure 3.8).

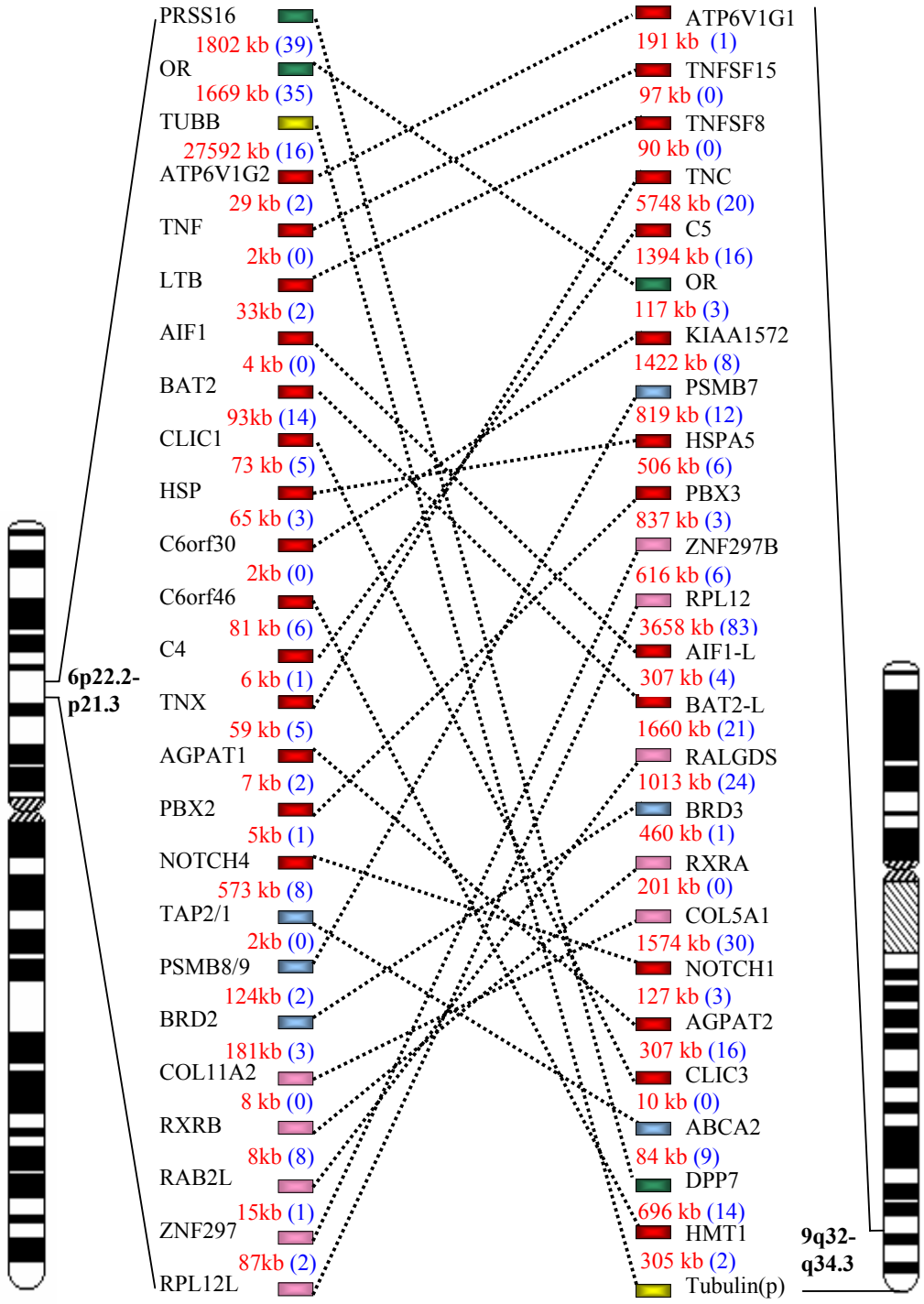

Figure 3.8 Comparison of the order of paralogues between the MHC region on 6p22.2-p21.3 and the paralogous region on 9q32-q34.3. The numbers in red indicate the distances between paralogues and the number in blue shown in parentheses corresponds to the number of coding genes interspersed between the paralogues. The extended class I genes and corresponding chromosome 9 paralogues are shaded green, class I in yellow, class III in red, the class II in blue and the extended class II in pink. The tubulin paralogue is a pseudogene, indicated by the letter 'p' in parentheses.

It appears that the overall gene order is poorly conserved between the two regions but small blocks of conservation can be seen, namely between gene pairs such as AIF1/AIF1-L and BAT2/BAT2-L. There are several examples of gene pairs that are located in the reverse orientation, including RXRB/A and COL11A2/COL5A1, as well as TNXB/TNC and C4/C5. Endo and co-workers (1997) identified five paralogues with conserved gene order, namely C4/C5-PBX2/PBX3-BRD2/BRD3-COL11A2/COL1A5-RXRB/RXRA, which all appeared to have arisen at the origins of vertebrate emergence. Upon identification of the precise location of the paralogues, this gene order is essentially conserved, albeit with the reverse orientation of RXRB/A and COL11A2/COL1A5. More recently, Flajnik and Kasahara (2001) analysed the gene order of all four proposed paralogous regions and identified six paralogues which have remained in the same order on 6p22.2-p21.3, 9q32-q34.3 and 19p13.3-p13.1 but not on chromosome 1. These include only two of the genes proposed by Endo and colleagues (1997). This will be discussed in more detail in chapter 4 upon full analysis of the MHC paralogues in the human genome.

The number of genes separating the MHC paralogues on 9q32-q34.3 range from zero to over 80 and are, in general, structurally and functionally unrelated, which is reminiscent of the MHC region. A number of transcription factors, solute carriers, homeobox proteins, lipocalins and kinases have been identified in the region but no HLA class I or class II genes have been found. An HLA-DR associated protein (also known as SET) and a hypothetical protein containing an Ig_MHC domain have been anchored to clones located on 9q34, but the association of the genes within this paralogous region with the immune system is not as prevalent as genes located within the MHC region. The extended MHC region is characterised by multigene families, in particular the HLA class I and class II families, histones, olfactory receptor genes and

zinc finger genes. The region on 9q32-q34.3 does not contain such a repertoire of multigene families but a number of zinc-finger proteins (six) and ribosomal proteins (four) have been identified along with the Surfeit locus. The Surfeit locus is not associated with the MHC region, but like many gene families located in the MHC region, is of biological interest.

The Surfeit locus contains an unusually tight cluster of six housekeeping genes, designated SURF1 to SURF6, which are unrelated by sequence similarity (Yon *et al*, 1993). The cluster exhibits alternation of transcription, bi-directional promoters and produce overlapping transcripts that has led to the proposition that these genes form a locus with potential regulatory and/or functional significance (Huxley and Fried, 1990; Gaston and Fried, 1994; Lennard *et al*, 1994). Colomobo and co-workers (1992) found that this cluster along with associated CpG rich islands have remained tightly clustered over 600 million years of divergent evolution that separate birds and mammals. However, it has been shown that in the teleost fish *Fugu* the five SURF genes are located in separate locations on two different chromosomes (Bouchireb *et al*, 2001). Thus, indicating that this tightly organised functional unit does not need to be next to each other in this organism. Nevertheless, the Surfeit cluster represents a gene cluster in which the gene organisation has biological significance in mammals, which is reminiscent of gene families located within the MHC region.

3.2.7.2 Genomic landscape

Surveys of genomic landscapes have noted the non-random distribution of particular sequence features, namely GC content and repeat elements. The assessment of these features is essential when characterising a genomic landscape (table 3.4). The overall

GC content of the 7.2 Mb 6p22.2-6p21.3 and the 24 Mb 9q32-q34.3 regions was calculated using Repeatmasker (http://repeatmasker.genome.washington.edu). The GC content in both regions (44% and 47%, respectively) was higher than the genome average of 41%. High GC content is associated with high gene density (IHGSC, 2001), which is a feature of both regions.

It is estimated that repeat sequences account for approximately 45% of the human genome (IHGSC, 2001). Although repeat elements are quite recent additions to the genome compared to the ancient duplication events proposed by Ohno (1970), it is interesting to compare the overall repeat content between regions as they shed light on chromosome structure and dynamics. Over time, these repeats reshape the genome by rearranging it, thereby creating entirely new genes or modifying and reshuffling existing genes.

Most human repeat sequences are derived from transposable elements and are made up of four major classes of repetitive elements (Smit, 1999): (1) short interspersed elements (SINEs), (2) long interspersed elements (LINEs), (3) elements possessing long terminal repeats (LTR elements) and (4) DNA transposons. The repeat content of the four main classes was calculated for the paralogous regions 6p22.2-p21.3 and 9q32-q34.3 using Repeatmasker and compared against the averages in the human genome (summarised in table 3.4).

The Alu content of both regions is higher than the genome average, which is interesting because Alu elements are associated with gene-rich regions of the genome (Smit, 1999; IHGSC, 2001). They are also associated with some chromosomal translocation breakpoint regions that suggest that these sequences could provide hot spots for homologous recombination, and could mediate the translocation process and

elevate the likelihood of other types of chromosomal rearrangements taking place.

Table 3.4 Comparison of the repeat content of the 6p22.2-p21.3 and 9q32-q34.3

Repeat element	6p22.2-p21.3 (% of sequence)	9q32-q34.3 (% of sequence)	Genome average (% of sequence)
Alu	14.83	14.42	10.60
MIR	1.06	3.45	2.20
Total SINE	**15.89**	**17.87**	**12.80**
L1	14.29	10.02	16.89
L2	2.21	2.86	3.22
L3	0.11	0.25	0.31
Total LINE	**16.61**	**13.13**	**20.42**
Total LTR	**10.47**	**5.26**	**8.29**
Total DNA	**2.35**	**2.02**	**2.84**
Unclassified	**0.65**	**0.17**	**0.12**
Total (%)	**45.97**	**38.44**	**44.83**
%GC (%)	**44**	**47**	**41**

3.2.7.3 Evidence of gene and segmental duplication

Gene and segmental duplications have shaped the MHC region (reviewed by Beck and Trowsdale, 2000) and there is strong evidence of such duplication events on 9q32-q34.3. Recent evidence indicates that duplication played a central role in the emergence of the two regions from a common ancestral region (Abi-Rached *et al*, 2002). It had previously been proposed that the MHC and 9q32-q34.3 regions, along with the proposed MHC paralogous regions on 1q21-q25 and 19p13.3-p13.1, had emerged via a series of large-genome duplication events prior to vertebrate emergence (Kasahara, 1999a). In order to test this hypothesis, Abi-Rached and colleagues (2002)

characterised the corresponding region in the cephalochordate amphioxus by identifying nine anchor genes and sequencing both the anchor genes and the regions that flank them. Analysis of the distribution of the human and amphioxus orthologues in their respective genomes revealed that they arose from a common ancestral region by block duplication events. The phylogenetic relationships determined that the duplications occurred after the divergence of cephalochordates (i.e. amphioxus) and vertebrates but before the gnathostomata (jawed vertebrates) radiation. Thus, showing the important role duplication has played in the origins of these two chromosomal segments.

Duplications have also played a major role in moulding the present-day arrangement of the 9q32-q34.3 region. For example, Lacazette and co-workers (2000) identified a new paralogous gene family on human chromosome 9q34 which they deduced were created by genomic duplications. They detected, in addition to the known, LCN1 (tear lipocalin) gene, two LCN1 pseudogenes and two OBPii genes (odorant binding proteins) paralogous to LCN1. Phylogenetic analyses indicated that the LCN1 and OBPii genes correspond to a subfamily of lipocalin genes that have arisen from a common ancestor by duplication. Figure 3.9 summarises the mechanisms involved in the emergence of the OBPii-LCN1 family.

Evidence suggests that a tandem duplication event of a seven exon lipocalin ancestor gave rise to two lipocalin paralogous genes. Following the differentiation of these two paralogues there were three complete, or partial, duplications of this 50 kb region on human chromosome 9q34. Analysis of the present day gene structures of the LCN1 and the OBPiiA and OBPiiB implied that the OBPii genes have evolved by integrating additional surrounding intronic DNA and recruiting an additional exon.

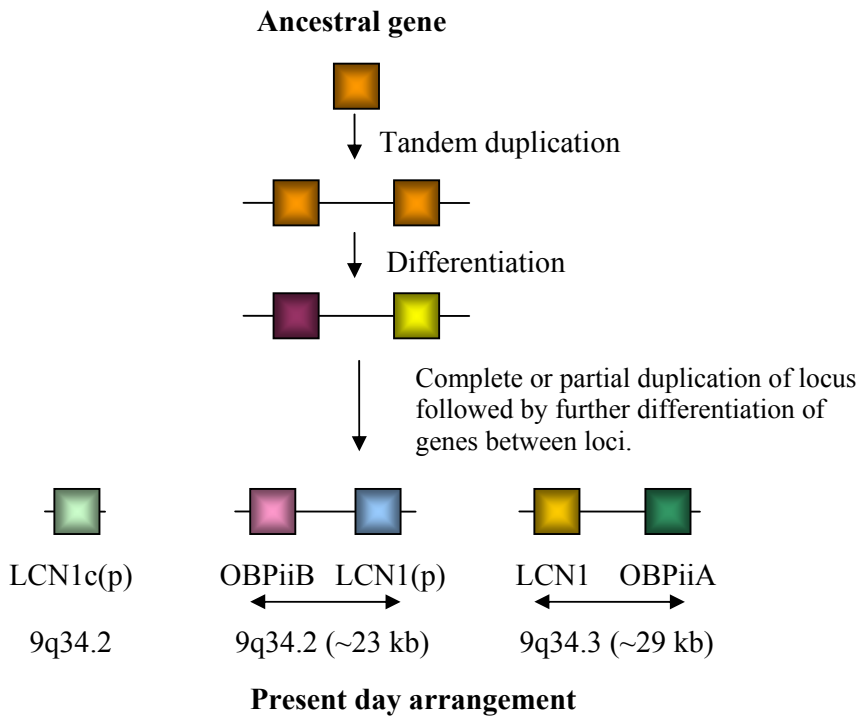

Figure 3.9 Evolution of the lipocalin paralogous gene family on 9q34. Gene differentiation is demonstrated by the change in colour of the boxes (genes) during evolution.

Recently, a 76 kb duplicon has been identified on chromosome 9q34 that is believed to mediate recombination leading to the Philadelphia chromosome (Ph) associated with leukaemia (Saglio *et al*, 2002). Segmental duplications, or duplicons, are segments of DNA with near-identical sequence. They are believed to be 'hotspots', or predisposition sites, for the occurrence of non-allelic homologous recombination or unequal crossing-over leading to genomic mutations such as inversions (Giglio *et al*, 2001), translocations (Giglio *et al*, 2002; Saglio *et al*, 2002), deletions and duplications (Reiter *et al*, 1996). The Ph chromosome is the most frequent cytogenetic abnormality present in human leukaemias and is a derivative chromosome 22 arising as a consequence of a reciprocal translocation between the long arms of chromosomes

9 and 22 (Saglio *et al*, 2002 and references therein). During the study of a patient with chronic myeloid leukeamia a large deletion on chromosome 9q32 and an unusual BCR-ABL transcript was observed. The unusual transcript was characterised by the insertion, between BCR exon 14 located on 22q11.2 and ABL exon 2 located on 9q34, of 126 bp derived from a region located on chromosome 9, 1.4 Mb 5-prime to ABL. This sequence is located in the clone bA65J3 which I have confirmed to be located approximately 1.4 Mb centromeric to the start of the ABL gene.

Fluorescence *in situ* hybridisation experiments on normal metaphase chromosomes detected two signals; a clear signal at 9q34 and a faint but distinct signal at 22q11.2. Sequence similarity search using BLAST determined that there was a large stretch of sequence similarity of 76 kb between 9q34 and a region approximately 150 kb 3-prime of the BCR gene on 22q11.2. Evolutionary studies using fluorescent *in-situ* hybridisation identified the region as a duplicon, which transposed from the region orthologous to human 9q34 to chromosome 22 after the divergence of orang-utan from the human-chimpanzee-gorilla common ancestor about 14 million years ago. The discovery of a large duplicon relatively close to the ABL and BCR genes, and the finding that the 126 bp insertion is very close to the duplicon at 9q34, opens the question of the possible involvement of the duplicon in the formation of the Philadelphia chromosome translocation as well as providing further evidence of the dynamic nature of this premier paralogous region.

3.2.7.4 Diseases associated with 9q32-q34.3

Several diseases and disorders are associated with the 322 identified genes and putative paralogues on 9q32-q34.3 (table 3.5), which is reminiscent of the MHC

region. For example, the truncation of the putative paralogue NOTCH1 is believed to play a role in human pre-T-cell acute leukeamias (T-ALL), which involves the chromosomal translocation between 7q34 and 9q34.3 (Ellisen *et al*, 1991). The association between NOTCH1 and 9q34.3 was first observed during the study of three cases of acute T-cell lymphoblastic leukaemia demonstrating the t(79)(q34;q34.3) (Ellisen *et al*, 1991). Ellisen and colleagues (1991) identified breakpoints within 100 bp of an intron in NOTCH1, resulting in the truncation of NOTCH1 transcripts. They suggested that the alteration of the NOTCH1 gene may play a role in the pathogenesis of some neoplasms. In addition, putative NOTCH1 paralogues have been identified at positions 1p13-p11 (NOTCH2) and 19p13.2-p13.1 (NOTCH3), which are also regions of neoplasia-associated translocation. The association of a variety of diseases and disorders with genes located within the paralogous region on 9q32-q34.3 is one of the similarities between this region and the MHC region.

Table 3.5 Summary of some of the disorders associated with 9q32-q34.3

Gene	**Disorder**	**Reference**
SURF1	Leigh's Disease	Zhu *et al*, 1998
TSC1	Tuberous sclerosis	van Slegtenhorst *et al*, 1997
COL5A1	Ehlers-Danlos syndrome	Nicholls *et al*, 1994
TAL2	T cell acute leukeamia	Xia *et al*, 1991
SET	Leukaemia	von Lindern *et al*, 1992
FCMD	Fukayama muscular dystrophy	Kobayashi *et al*, 1998
NR5A1	XY sex reversal	Achermann *et al*, 1999
DBCCR1	Bladder cancer	Habuchi *et al*, 1998

3.3 Discussion

This chapter presents the findings from the characterisation of one of the chromosomal regions proposed to be paralogous to the MHC. The region spans from 9q32 to 9q34.3 encompassing approximately 24 Mb of genomic sequence and represents the largest chromosomal region containing MHC paralogues to be mapped, sequenced and analysed to-date. Analysis of 9q32-q34.3 has not only provided insight into its genomic organisation but it has revealed a number of features that are shared with the MHC.

One of the main features common to both regions is that they are gene rich. Overall, the density of genes located within the MHC region is higher compared with the proposed paralogous region on 9q32-q34.3, but both contain a higher density of genes when compared with the rest of the genome. The gene-rich nature of both regions is also associated with a high GC content, which is a feature of both 6p22.2-p21.3 and 9q32-q34.3. High GC content may also explain why gaps still remain in the region 9q32-q34.3. At the time of writing (August 2003), the minimum tiling-path of 9q32-q34.3 has 198 fully sequenced clones but still contains five gaps ranging in size from approximately 5 kb to 200 kb. High GC content is believed to cause the region to be deletion-prone through frameshift mutagenesis or other unknown cellular mechanisms (Bichara *et al*, 1995; 2000) and thus making it difficult to clone and sequence.

In total, 322 genes were identified within the region 9q32-q34.3. These genes are both structurally and functionally unrelated, which is a feature of the genes located within the MHC class III region but is not mirrored by the extended MHC region as a whole. All of the 40 paralogues cited in the literature, corresponding to 25 MHC gene families, were identified within 9q32-q34.3 (Kasahara, 1999a; 1999b; Flajnik and

Kasahara, 2001). It is important to note that the paralogues discussed in this chapter are termed 'putative paralogues' as they have only been identified based on previously published data and have not been characterised within this chapter; this will be addressed in chapter 4.

One of the main differences between the extended MHC region and 9q32-q34.3 is that the prior is characterised by the human leukocyte antigen (HLA) genes located in the class I and class II regions, which are involved in antigen presentation, whereas the latter does not contain any of these genes. In contrast, it has been shown that the proposed paralogous region on 1q22 has a cluster of class I-like HLA genes, termed the CD1 gene cluster (Shiina *et al*, 2001). From this analysis it is not possible to determine whether 9q32-q34.3 once contained HLA class I-like genes and they have since been lost or whether they have never been part of the 9q32-q34.3 gene repertoire.

The linkage of the putative MHC paralogues on 9q32-q34.3 is associated with a common origin of the two regions by large-scale duplication; either as a block or the entire genome. If they did have a common origin it is expected that the regions are syntenic. However, analysis of the overall gene order of the paralogues between the MHC and 9q32-q34.3 is not strictly conserved, but there is evidence of conservation in the order of some paralogues. My findings are consistent with those of Endo and colleagues (1997) who deduced the gene order on chromosome 9 using cytogenetic and genetic maps in mouse, although two paralogues are actually in the reverse order than they proposed. The likelihood of synteny between the MHC region and 9q32-q34.3 may well be related to the time that has elapsed since their emergence. If they did emerge at the time of vertebrate emergence as proposed by Endo and co-workers

(1997), as well as others including Kasahara (1997; 1999a; 1999b), then approximately 500 million years of evolution have passed. It is also associated to the amount of rearrangement of the genomic sequence by evolutionary mechanisms, including inversions, translocations and duplications. The dynamic nature of the region 9q32-q34.3 is evident by the presence of duplicons and repetitive elements known to be involved in chromosomal rearrangements. There is also further evidence of local duplication events within both regions.

If the paralogues did not emerge simultaneously by block duplication they must have duplicated independently. Hughes (1998) proposed two hypotheses as to why these paralogues have come together; they are (1) that the cluster of paralogues is a result of chance and (2) that it is selectively advantageous for these paralogues to be together. Such a large number of independent translocations are unlikely to have occurred by chance and it has been suggested that there are selective advantages as to why the MHC paralogues are clustered on 9q32-q34.3, namely a functional reason. Analysis of the genes located within 9q32-q34.3 does not support this hypothesis as they appear to have diverse functions. However, this will be discussed in more detail in chapter 6.