Genome Evolution:
a study of MHC paralogous genes in the human genome

Vikki Rand

This dissertation is submitted for the degree of Doctor of Philosophy

September 2003

Gonville and Caius College,
University of Cambridge

Wellcome Trust Sanger Institute
Wellcome Trust Genome Campus,
Hinxton, Cambridge

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. This dissertation does not exceed the word limit set by the Biology Degree Committee.

Abstract

One of the interesting findings of the Human Genome Project was that approximately 10% of the genome has arisen by duplication. This is exemplified by the clusters of genes, on chromosomes 1q21-q25, 9q32-q34.3 and 19p13, paralogous to genes located within the Major Histocompatibility Complex (MHC) region, on 6p22.2-p21.3. By definition, paralogues are genes within the same species that have originated through duplication of an ancestral gene. The survey of the human genome identified 82 MHC paralogues based on sequence similarity and conserved gene structure. Analysis of the distribution of the paralogues identified clusters on chromosomes 1q21-q25, 9q32-q34.3 and 19p13 (38/82), and revealed paralogues located elsewhere in the genome (44/82). In total, 44% of the paralogues identified are novel discoveries, of which 89% are located outside the previously known clusters.

Evidence from my phylogenetic analyses indicates that the MHC paralogues located within the regions on 1, 9 and 19 arose by two ancient duplication events, either by duplication of the whole genome or of chromosomal segments, prior to vertebrate emergence. Expansion of paralogous gene families has occurred by additional duplications involving individual loci or chromosomal regions resulting in paralogues outside the clusters. In-depth analysis of the chromosomal region 9q32-q34.3 revealed that the order of paralogues is not conserved and that they are interspersed by other genes, indicating the region has been subjected to genomic rearrangements.

Comparison of the expression profiles of a selected set of MHC paralogues revealed that some have functionally diverged since duplication; with members of the same paralogous gene family being ubiquitously expressed, and others, having an

expression profile restricted to only a few tissues. Evidence of co-expression of paralogues in some tissues suggests a similar function and involvement in the same pathways. This thesis highlights the importance of understanding paralogy, particularly for future investigations of phenotypes associated with paralogous genes.

Acknowledgements

I would like to thank my supervisor Stephan Beck for his tireless enthusiasm, encouragement and support throughout this project. Particular thanks go to my second supervisor at Cambridge University, John Trowsdale. I am also grateful to the past and present members of the Immunogenomics group (a.k.a Team 50) who have helped me throughout my PhD: Karen Novik, Ruth Younger, Roger Horton, Melanie Stammers, Karen Crum, Karen Halls, Jennifer Sambrook, Penny Coggill, Marcos Miretti and Vardhman Rakyan. Thankyou for all your advice, encouragement and help. I am also very grateful for your tolerance – particularly of my moaning over the last few months – and for providing regular distractions in the form of tea breaks, coffee breaks, lunch, pub..... this also goes for the members of team 30, Gavin Wright, Mark Bushall and Nick Bockett, who foolishly share our lab and office!

I am indebted to everyone who has contributed to the mapping and sequencing of the human genome - without this data the work in this thesis would not have been possible. I would also like to thank everyone who has helped me with the different aspects of the project. Special thanks goes to Sean Humphray and the members of the Chromosome 9 Sequencing and Mapping group who helped me with the work presented in chapter 3. Particular thanks go to Andrew Knight for his help with the sub-cloning, Keith Porter for the fingerprinting and the Sanger Institute Cytogenetics laboratory, namely Sheila Clegg and Pawan Dhami, who helped me with the fluorescent *in-situ* hybridisation (FISH) experiments. I am also very grateful to the members of team 41 and, the now extinct, team 49 who helped me with the sequencing and to Gavin Laird and Adrienne Hunt for their 'finishing' tutorials.

Thanks also go to Rosemary Ward for all her assistance with the tissue culture, Alison Coffey for her help with the blots and to Dave Vetrie, Cordelia Langford and the Sanger Institute Microarray Facility for all their help and guidance with the microarray experiments. Thanks also go to Ewan Birney and Michele Clamp for their help with ENSEMBL and to Kevin Howe for maintaining the FINEX database and writing some useful perl scripts. I realise I have been privileged to have worked in an environment where help has been offered so freely.

On a personal note, thanks must go to all my family and friends who have put up with me over the last few months – and years! Particular thanks go to my mam, dad and brother Simon for their pillar-like support and their unwavering belief and confidence in me – without this I don't think I would have made it this far! This also goes for all my friends who have offered encouragement and wine along the way!!! Cheers!

Table of contents

List of tables

List of figures

Chapter 6 Expression analysis of extended MHC paralogous gene families

Abbreviations

aa	amino acid
AIF	Allograft inflammatory factor
ATP	adenosine 5'-triphosphate
BAC	bacterial artificial chromosome
BLAST	basic local alignment search tool
bp	base pair
BRD	Bromodomain containing protein
°C	degrees Celsius
cDNA	complementary deoxyribonucleic acid
CLIC	Chloride intracellular chloride channel
CTP	cytidine 5'-triphophate
dbEST	database of expressed sequence tags
DNA	deoxyribonucleic acid
dNTP	2'-deoxyribonucleoside 5'-triphophate
DTT	dithiothreitol
EDTA	ethylenediamine tetra-acetic acid
EMBL	European Molecular Biology Laboratory
EST	expressed sequence tag
FISH	Fluorescent *in-situ* hybridisation
FPC	fingerprinting contig
GPX	Glutathione peroxidase
GTP	guanine 5'-triphosphate
HGMP	Human Genome Mapping Resource Centre
HGP	Human Genome Project
HLA	human leukocyte antigen
IHGSC	International Human Genome Sequencing Consortium
kb	kilobase pairs
l	litre
-L	-like
LB	Luria-Bertani
LINE	long interspersed nuclear element

M	molar
mA	milliamps
Mb	megabase pairs
µg	microgram
µl	microlitre
µM	micromolar
min(s)	minute(s)
MIPS	Munich Information Centre for Protein Sequences
mg	milligram
MHC	Major Histocompatibility Complex
ml	millilitre
mm	millimetre
mM	millimolar
NCBI	National Centre for Biotechnology Information
ng	nanogram
NOTCH	Neurogenic locus Notch homologue
OR	Olfactory receptor
PCR	polymerase chain reaction
PFAM	protein family database
PBX	Pre-B cell leukaemia transcription factor
RNA (mRNA, rRNA, tRNA)	ribonucleic acid (messenger-, ribosomal-, transfer-)
rpm	revolutions per minute
RT-PCR	reverse transcription polymerase chain reaction
RXR	Retinoic acid receptor
SDS	sodium dodecyl sulphate
sec(s)	second(s)
SINE	short interspersed nuclear element
STS	sequence tagged site
TEMED	N, N, N', N'-tetramethylethylenediamine
Tris	tris(hydroxymethyl)aminomethane
U	unit
UTR	untranslated region
V	volt

Chapter 1

Introduction

1.1 Genome evolution

Genomes have evolved and increased in complexity owing to a number of evolutionary processes acting upon them, such as insertions, deletions and inversions. Gene duplications are also believed to have played a major role in the evolution and development of vertebrate genomes. Susumu Ohno (1970) first suggested that the increase in organismal complexity during vertebrate evolution could only have occurred if there was a considerable increase in gene number and proposed that this happened by the duplication of entire genomes in a process termed polyploidisation.

When Ohno first proposed the theory of polyploidisation it generated a lot of excitement and outrage in the field of genetics, but, by the late 1980s, many had lost interest owing to the lack of evidence. With the expansion of genomic information generated during the 1990s, duplicated genes and chromosomal regions were identified in the human, and other genomes, and the theory became popular again although it remains controversial. Duplicated genes and regions are believed by some to represent remnants of whole-genome duplication events, whilst others have argued that they are the result of the duplication of chromosomal regions or of individual genes brought together by selective forces (reviewed by Wolfe, 2001; Lundin *et al*, 2003; Hughes and Friedman, 2003).

1.2 Homologues, paralogues and orthologues

Three definitions are commonly used to describe the relationship between genes: homologues, paralogues and orthologues (reviewed by Sharman, 1999). Homologous genes are members of the same family or superfamily and share a common ancestor at some point back in evolutionary time. Homologues can be further subdivided into two groups; orthologues, genes that have been separated by speciation, and paralogues, genes that have resulted from a duplication event. Orthologues can be traced by descent to the common ancestor of two organisms and will both encode equivalent evolutionarily conserved proteins. Paralogues, however, are genes within the same species that have originated through duplication of an ancestral gene; whether as part of a whole genome, chromosomal segment or a single gene duplication event. The evolutionary fate of paralogues and orthologues are very different. Orthologues often take over the function of the precursor gene in the species of origin and thus tend to be conserved. In contrast, young paralogues have redundant functions, which are an evolutionary unstable situation, thus, in the long run – with a few exceptions – paralogues either diverge functionally, or all but one of the versions are lost.

1.3 Paralogous genes and the evolution of the human genome

Paralogous genes have been identified throughout the human genome. Ohno (1973) identified duplicated chromosomal segments within the human genome containing two pairs of duplicate genes on chromosomes 11 and 12, which he proposed as being evidence of polyploidisation. In the 1990s, molecular mapping data was used to identify a number of chromosomal regions containing clusters of paralogues in the human and mouse genomes that were believed to be remnants of genome duplication

events (termed paralogous regions; Lundin, 1993).

Intriguingly, the number of paralogous regions and paralogous genes investigated at the time was generally four (this phenomenon was termed tetralogy), or less, suggesting that at least two rounds of large-scale block or genome duplications have occurred during the course of mammalian evolution. For example, Spring and co-workers (1994) found that vertebrates have four copies of a gene for a cell-surface protein called syndecan, whereas the fruit fly *Drosophila* has only one. More than fifty examples of this so-called 1-to-4 gene rule have now been identified (Spring, 1997). Independently, Sidow (1996) observed the 1-to-4 gene rule during phylogenetic and sequence surveys of developmental regulator families, in which he concluded that two large-scale gene duplication events, most likely of entire genomes, occurred in an ancestor of vertebrates (Sidow, 1996).

Ohno (1970) originally suggested that there were large-scale gene duplication events, possibly involving the whole genome, in early chordates; specifically on the lineage leading to both cephalochordates (including amphioxus) and vertebrates (including hagfish, lampreys and jawed vertebrates). He also suggested a second, and maybe a third, large-scale duplication event at the time of fish or amphibian divergence. The number of duplications and the mechanisms involved have been heavily debated, and many modifications of Ohno's model have been proposed. For example, Holland and co-workers (1994) proposed that there were two phases of duplication on the vertebrate lineage, but suggested that the first duplication occurred on the vertebrate lineage after divergence of the amphioxus lineage, and the second on the jawed vertebrate lineage after the divergence of jawless fish. Kasahara and colleagues (1996) proposed that two polyploidisation events occurred later in the vertebrate

lineage, after the divergence of lampreys. The most popular version of events has been termed the 2R hypothesis as it involved two rounds of polyploidisations; one prior to the divergence of agnatha (jawless fish, exemplified by lampreys and hagfish) and gnathostomata (jawed vertebrates), while the second occurred after the divergence of agnatha but before the divergence of chondryichthyes (cartilaginous fish) (Sidow, 1996). This is simplified in figure 1.1.

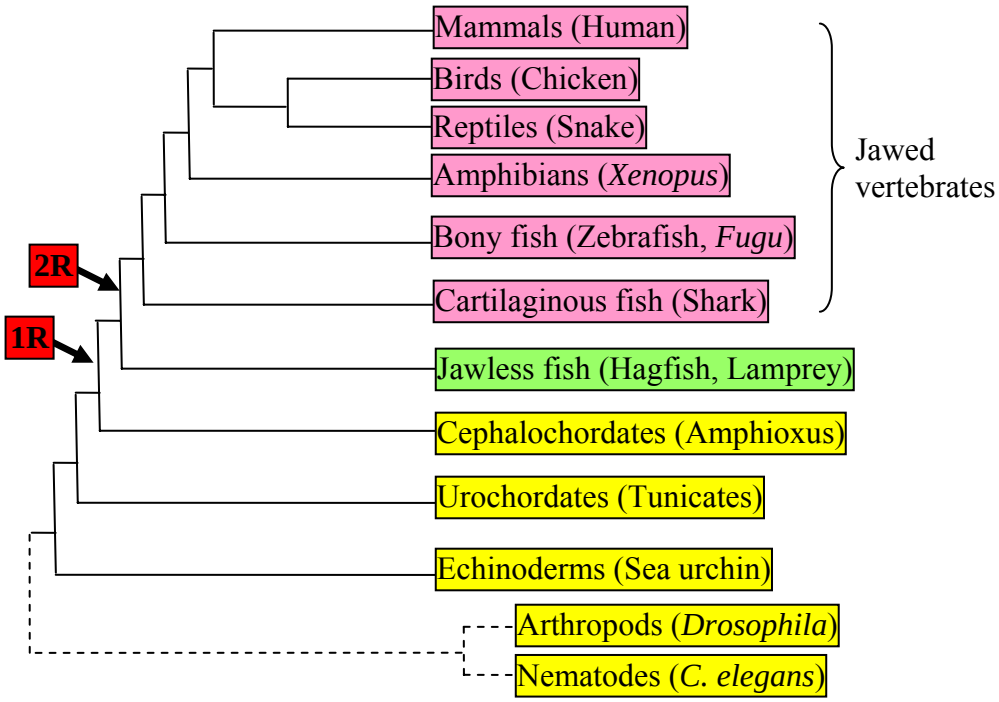


Figure 1.1 The 2R hypothesis. The two rounds of duplication are indicated by arrows. 1R corresponds to the first round of whole-genome duplication, after the emergence of amphioxus, and 2R corresponds to the second round of whole-genome duplication prior to the emergence of jawed vertebrates, more specifically cartilaginous fish.

The four Hox gene clusters in the human genome exemplify the 2R hypothesis (figure 1.2). The homoeotic complex (HOM-C) occurs as a single cluster in invertebrates such as *Drosophila, Caenorhabditis elegans* and amphioxus, but is found as four paralogous Hox gene clusters in vertebrates like mice and humans (Schughart *et al*,

1988). Interestingly, not only was the order of genes in the mammalian Hox clusters found to be conserved between human and mouse, but it was also conserved among the four mammalian clusters. The quadruplication of the Hox genes and the discovery of other paralogous genes linked to the Hox clusters provide evidence to support the involvement of large-scale chromosomal or whole-genome duplications in the evolution of vertebrate genomes (Larhammar *et al*, 2002).

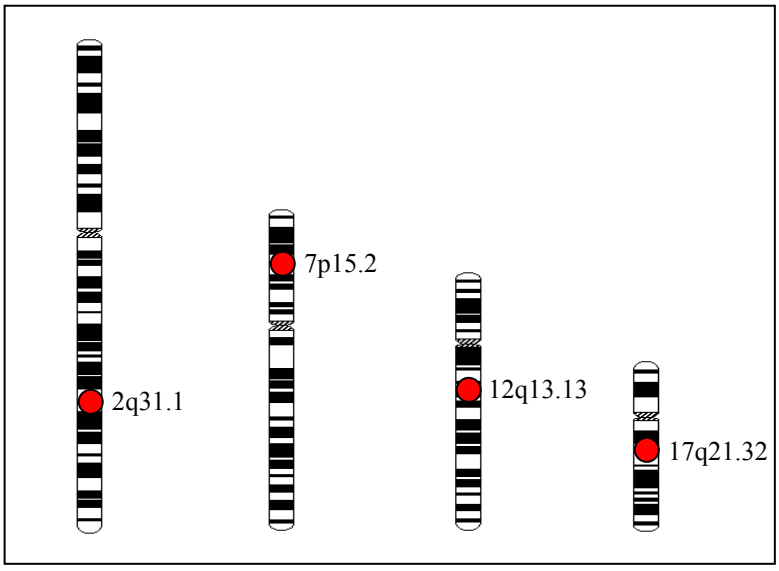


Figure 1.2 Distribution of Hox gene clusters in the human genome (represented by the red circles).

The 2R hypothesis is controversial and continues to be heavily discussed in the literature (reviewed by Wolfe, 2001). It was widely believed that the debating over the evolution of the human genome would be resolved once the entire human genome sequence was available. However, the initial analysis of the draft human genome sequence did not reveal overwhelming evidence for tetralogy and the 2R hypothesis remains controversial (International Human Genome Sequencing Consortium (IHGSC), 2001; Venter *et al*, 2001).

1.4 Genome sequencing projects

Between 1977 and 1982 the genomes of the bacterial virus ΦX174 (Sanger *et al*, 1977a, 1978), bacteriophage lambda (Sanger *et al*, 1982), animal virus SV40 (Fiers *et al*, 1978) and the human mitochondrion (Anderson *et al*, 1981) were successfully sequenced and assembled. During the early 1990s, the genomes of the yeast *Saccharomyces cerevisiae* (Oliver *et al*, 1992) and the nematode worm *Caenorhabditis elegans* (Wilson *et al*, 1994) were sequenced, thus demonstrating the feasibility of large-scale genome sequencing. By September 2003, the sequencing of 160 genomes had been completed, with 393 prokaryotic and 242 eukaryotic genome-sequencing projects still ongoing (http://igweb.integratedgenomics.com/GOLD/). The time-line of a number of genome sequencing projects is shown in figure 1.3.

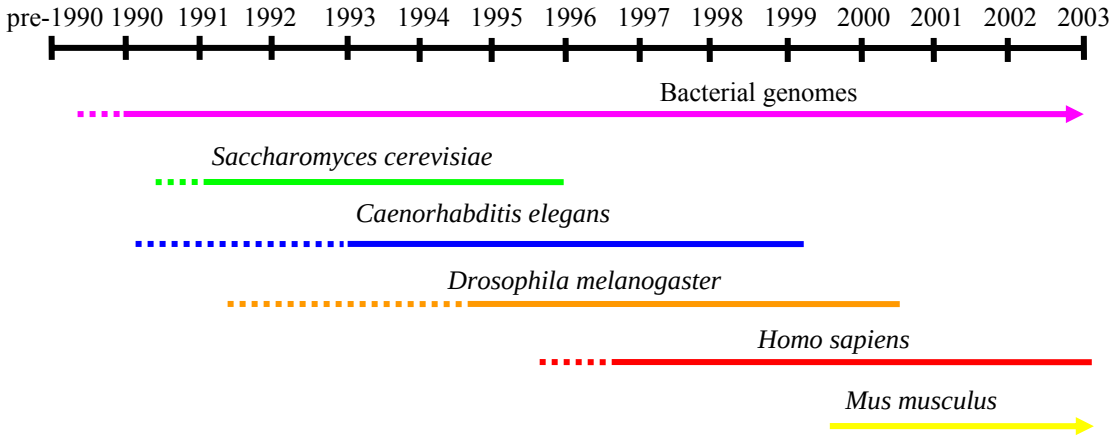


Figure 1.3 Time-line of a range of genome sequencing projects. The arrows signify ongoing sequencing projects.

1.5 The Human Genome Project

The Human Genome Project (HGP) was established in 1990 with the aim of

sequencing the entire human genome by 2005. In 1999, the year I started this project, the HGP effort moved into full-scale production, and the overall sequencing output increased significantly (figure 1.4). By 2000, the 'draft' human sequence was completed consisting of mainly 'unfinished' sequence covering approximately 90% of the human genome. Two 'draft' sequences were published by separate organisations (IHGSC, 2001; Venter *et al*, 2001) offering the chance to compare the genomic data produced. The data generated by the International Human Genome Sequencing Consortium (IHGSC) was a collaborative effort involving 20 groups from around the world. Venter and colleagues were part of the biotechnology company Celera Genomics which was formed in 1998. The completion of the HGP was announced by the IHGSC in 2003, two years ahead of schedule.

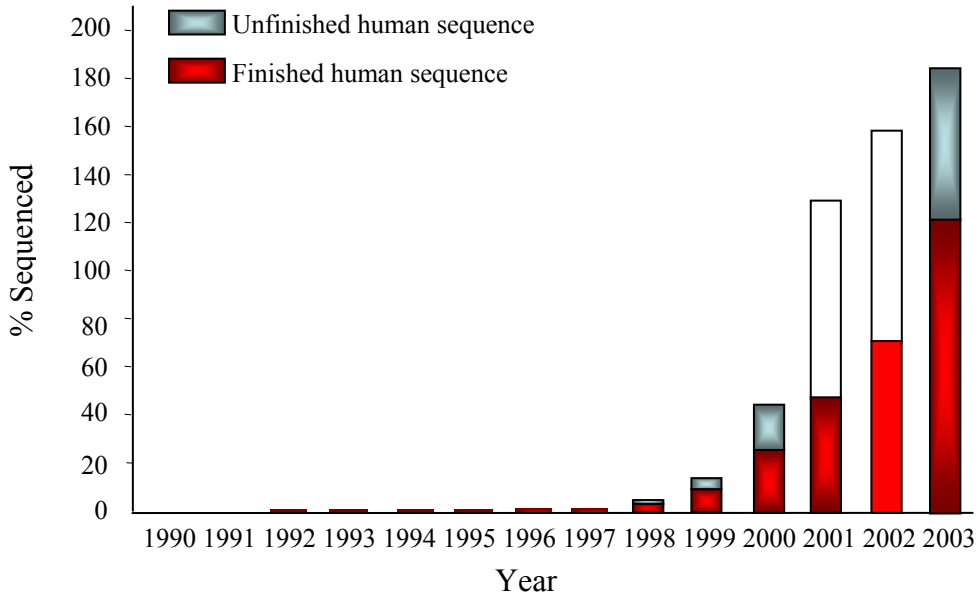


Figure 1.4 Progress of the Human Genome Project from the launch in 1990 to its completion in 2003. The % of finished (red) and unfinished (grey) sequence was calculated for January of each year using the Genome Monitoring Table (http://www2.ebi.ac.uk/genomes/mot/). Finished sequence is the final stage of the sequencing project when the sequence is contiguous with reads covering depths of greater than 8 times redundant sequence with 99.99% accuracy. Unfinished sequence is a working draft covering depths of 2-4 times redundant sequence and contains gaps.

1.6 Analysis of the human genome sequence

The sequence of the entire human genome has enabled a number of key aspects of the genome to be investigated in order to test the theory of polyploidisation and the 2R hypothesis. These are discussed below.

1.6.1 Gene numbers

One of the most important pieces of information revealed by sequencing projects is the number of genes. Ohno (1970) first observed that the gene number and genome sizes increased when looking at more complex organisms. This observation has been confirmed by various sequencing projects (table 1.1).

Table 1.1 Gene number and genome size for a range of organisms.

Organism	***Genome size (Mb)***	***Gene Number***	***Reference***
Homo sapiens	3000	~30,000	IHGSC, 2001 and Venter *et al*, 2001
Mus musculus	3000	30,000	Marshall, 2001
Fugu rubripes	365	31,059	Aparacio *et al*, 2002
Drosophila melanogaster	135.6	13,061	Adams *et al*, 2000
Caenorhabditis elegans	97	19,099	*C.elegans* Sequencing Consortium, 1998
Saccharomyces cerevisiae	12.1	6,034	Mewes *et al*, 1997
Escherichia coli	4.67	3,237	Blattner *et al*, 1997

The human genome was originally estimated to have over 80,000 genes while invertebrates have less than 20,000. The fourfold increase between human and invertebrate gene numbers was previously used as evidence in support of the 2R hypothesis (Makalowski, 2001). One of the most interesting discoveries in the human

sequencing projects has been the identification of only approximately 30,000 protein-coding genes in the human genome (IHGSC, 2001; Venter *et al*, 2001). On average, this would give only two paralogues in humans for every invertebrate gene and would support only one round of genome duplication. However, it could be argued that extensive gene loss may follow genome duplication i.e. if two rounds of duplication occurred then a significant proportion of duplicate genes were lost after each duplication event leaving no obvious trace of two genome duplication events.

1.6.2 1-to-4 gene rule

Initial analysis of the protein coding genes in the draft human genome does not support a strict 1-to-4 gene rule (IHGSC, 2001; Venter *et al*, 2001). The International Human Genome Sequencing Consortium employed an all-against-all sequence comparison to identify orthologous groups in human, *C. elegans* and *Drosophila* genomes. A total of 1308 groups were identified with a mean of 2.4 genes per human orthologue group and 1.1 genes per group in *C. elegans* or *Drosophila*. On closer analysis, almost half of the identified orthologue groups had just a single gene in the human genome, and the remainder had two, three, four or more genes. When the ratio of the number of orthologue groups with a single gene in *C. elegans* and *Drosophila* and the number of genes in human were plotted for each analysis, the peak of this distribution was found over the 1:1 ratio and not the 1:4 ratio needed to support a strict 1-to-4 gene rule. In both cases, there are a significant number of gene families (greater than 50%) with two or more members implying that gene families have expanded via duplication (IHGSC, 2001; Venter *et al*, 2001). These gene family expansions could have been generated through whole-genome duplication events.

1.6.3 Paralogy and the human genome

With the advent of the 'draft' human genome sequence, a number of analyses have now been performed to identify all the paralogous regions. Prior to the release of the draft sequence several lists of paralogous regions had been published and were believed to represent only a small percentage of the total (Lundin, 1993; Lundin and Larhammar, 1998; Skrabanek and Wolfe, 1998; Pollard and Holland, 2000).

The International Human Genome Sequencing Consortium (2001) concluded that approximately 5% of the human genome consists of paralogous regions. The duplicated regions tend to be large, greater than 10 kb, and highly homologous. Evidence of ancient duplications, characterised by high sequence similarity between coding regions, were identified along with evidence of more recent segmental duplications. The latter duplicated regions share high sequence identity between both exons and introns, with many showing less than 6% nucleotide divergence between paralogous regions. Such duplications seem to have emerged very recently in evolution as they are absent from closely related species.

Analysis of the draft human genome sequence by Venter and co-workers (2001) using a multiple alignment algorithm, identified 1077 blocks of paralogy spread throughout the genome. Out of the 1077 blocks, 159 contained only three genes, 137 contained four genes and 781 contained five or more genes thus illustrating the extent of duplications in the human genome. McLysaght and colleagues (2002) conducted one of the most thorough investigations into duplicate genes in the human genome. Of the 24,046 genes used in the analysis, 6,120 (almost a quarter) were identified located in 1642 paralogous regions containing two or more linked duplicated genes. The Hox gene clusters were amongst the largest paralogous regions identified; they found 28

paralogous genes on chromosomes 7p and 17q, and 26 genes on chromosomes 2q and 12q. Owing to the number and sizes of the paralogous regions identified in all three analyses the most likely explanation is that they arose by whole-genome or large-scale block duplication events rather than through duplication of individual genes.

1.6.4 Evolutionary analysis of paralogous gene families

A number of phylogenetic studies have been conducted in order to understand the evolutionary histories of the paralogous gene families. The 2R hypothesis proposes that one round of duplication occurred after the divergence of cephalochordates (exemplified by amphioxus) and the second after the divergence of jawless fish (including hagfish and lamprey). Therefore, the phylogenetic trees of the gene families should show similar histories.

The phylogenetic analyses of gene families supporting the 1-to-4 (or less) gene rule revealed that the evolution of the human genome is complicated. Wang and Gu (2000) analysed 49 vertebrate gene families, each consisting of three or four gene members, generated in the early stages of vertebrates, and/or shortly before the origin of vertebrates, including the early growth response protein, EGR, and the glycine receptor, GLR. Of the 49 gene families studied, they determined that 26 families with three members were consistent with the 2R hypothesis but the evolution of the remaining 23 gene families with four members was more complicated. Of these 23, only five were consistent with the 2R hypothesis, with 11 families supporting a third round of genome duplication and the remaining seven families suggesting at least one round of duplication prior to the divergence between *Drosophila* and vertebrates.

In contrast, Friedman and Hughes (2001) found that of a total of 134 families with four members 70% were not consistent with the 2R hypothesis. Similar results were also reported for a smaller number of gene families by the International Human Genome Sequencing Consortium (2001). However, it is considered by some that organisms, such as amphioxus, hagfish and lamprey, are more appropriate to study vertebrate evolution than *Drosophila* as they are actually on the vertebrate lineage (Holland, 2003). Escriva and colleagues (2002) investigated 33 gene families, where the sequence was available for both lamprey and hagfish. According to their phylogenetic analyses, all 33 families were found to support the 2R hypothesis.

1.7 Polyploidy

Humans and other species are generally diploid and have two copies of each gene; one from each parent. As stated earlier, it has been suggested that the vertebrate genome evolved via whole-genome duplication events, in which the chromosome complement doubled at some point in time. Therefore, the vertebrate genome underwent a stage when it was polyploid, then, through processes such as gene silencing and mutation, reverted to a diploid-like state. Several polyploid species have been identified in both the animal and plant kingdoms. One example is the amphibian, *Xenopus laevis,* which is tetraploid and has double the number of chromosomes than its cousin, *Xenopus tropicalis*. In 1999, the first polyploid mammal, the red viscacha rat (*Tympanoctomys barrerae*) was discovered (Gallardo *et al*, 1999). The rodent is unaffected by having double the number of chromosomes showing that the vertebrate genome can duplicate and that organisms can survive with multiple copies of a genome.

In addition to the two rounds of genome duplication in the vertebrate lineage Ohno (1970) proposed a round of genome duplication in fish; after the divergence of lobed-fin fish that led to land-based organisms. Evidence in support of the third duplication was detected in zebrafish (Postlethwait *et al,* 1998) and *Medaka* (Wittbrodt *et al,* 1998) based on the observation that they generally have larger multigene families than mammals. In particular, Amores and co-workers (1998) observed that zebrafish had seven Hox gene clusters in comparison to the four present in mammals and one in amphioxus (Garcia-Fernandez and Holland, 1994). Further mapping and sequence data has shown that for any four paralogous (or tetralogous) genes or regions in mammals there are probably an additional three or four in teleost fish.

1.8 Mechanisms of gen(om)e duplication

Gene duplication has played a major role in the evolution of the human genome. Duplication may involve part of a gene, a single gene, part of a chromosome, an entire chromosome, or the whole genome. The duplication of part of, or a whole, gene is also referred to as a tandem duplication event. Chromosomal regions are duplicated as part of either a block or segmental duplication event. Segmental duplications are defined as involving the transfer of genomic sequence to one or more locations in the human genome and, because of the strong sequence identity between both exons and introns, are relatively recent events (IHGSC, 2001). More ancient duplication events are characterised by similarities only in the coding regions and, in this thesis, are referred to as block duplication events.

The duplication of an entire chromosome is also known as aneuploidy or polysomy. There are several examples of trisomy of human chromosomes that are linked to a

number of conditions. A well known example of this is the trisomy of chromosome 21, which causes Down's syndrome. As discussed previously in this thesis the duplication of an entire genome is referred to as whole-genome duplication or polyploidisation.

There are several mechanisms by which duplication can occur; they are unequal crossing-over, unequal sister chromatid exchange, duplicative transposition, replication slippage and polyploidisation. Unequal crossing-over is a recombination event initiated by similar nucleotide sequences that are not at identical places in a pair of chromosomes. Unequal sister chromatid exchange is essentially the same as unequal crossing-over except that it involves chromatids from a pair of homologous chromosomes. The result can be duplication of a segment of DNA in one of the recombination products. This mechanism can create both small families, such as the five related genes of the β-globin cluster on chromosome 11, and large ones, such as the olfactory receptor gene clusters, which together contain nearly 1,000 genes and pseudogenes.

Transposition is defined as the movement of genetic material from one chromosomal location to another. During the process termed duplicative transposition, the transposable element is copied, therefore if this element contains a gene, the original copy is retained at the original site while a new copy is inserted elsewhere in the genome. The process, replication slippage is more commonly associated with the duplication of very short sequences, such as repeat units in microsatellites, but can also result in gene duplication if the genes are relatively short. In either case, the recombination occurs between two different copies of a short repeat sequence leading to duplication of the sequence between the repeats.

Whole-genome duplication occurs as a consequence of the lack of separation between all daughter chromosomes following DNA replication. Since it immediately doubles the size of a genome it is considered as the most effective mechanism for increasing genome size. Whole-genome duplication can occur via two mechanisms; allotetraploidy and autotetraploidy (figure 1.5). Autotetraploidy occurs within a single species and allotetraploidy occurs between genomes from different individuals (Wendel, 2000).

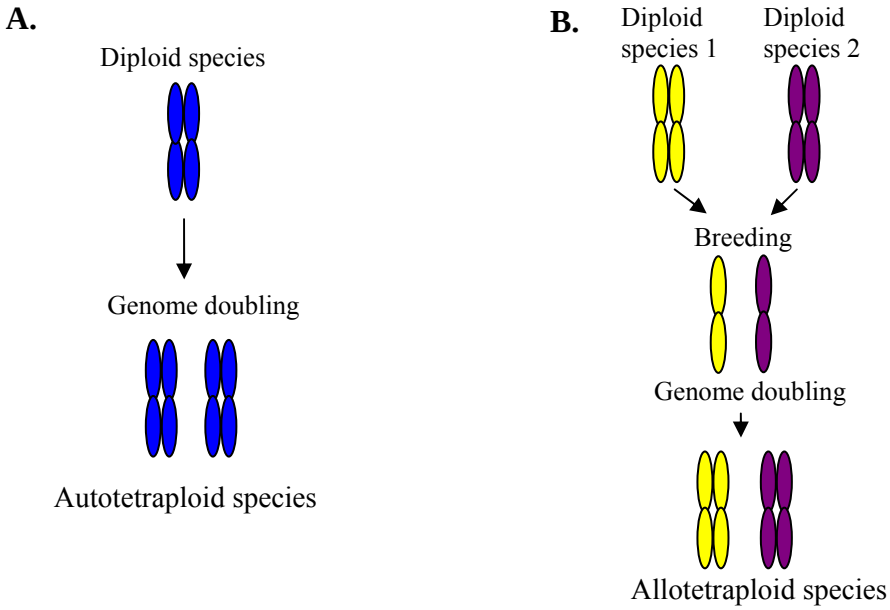


Figure 1.5 Models of genome duplication by (A) autotetraploidisation and (B) allotetraploidisation.

In plants, polyploidy is widespread and numerous studies have been conducted to understand the prevalence and consequence of polyploidy. Artificially produced autopolyploids are generally inferior to their diploid progenitors, and have lower fertility and, often, lowered ability to compete with diploid species owing to physiological effects such as, genetic imbalances and irregularities in chromosomal segregation (reviewed by Li, 1997).

Polyploidy is extremely rare in bisexually reproducing animals. Muller (1925) proposed that this is because in bisexual animals the two sexes are differentiated by means of a process involving the diploid mechanism of segregation and combination, and polyploidy invariably disturbs this process. In amphibians and fish, where there is evidence of successful polyploidy, the chromosomal determiners of the opposite sexes are still in a rather initial state of differentiation, and the X and the Y or Z and the W chromosomes can substitute for each other (Ohno, 1970). In these animals, genome duplication would not result in sexual imbalance, and many tetraploid species have been found (Ohno, 1970; Bogart, 1980; Schultz, 1980). It is interesting to see that, the only example of a tetraploid mammal identified to date is tetraploid for all autosomal chromosomes but is diploid for the sex chromosomes (figure 1.6; Gallardo *et al*, 1999).

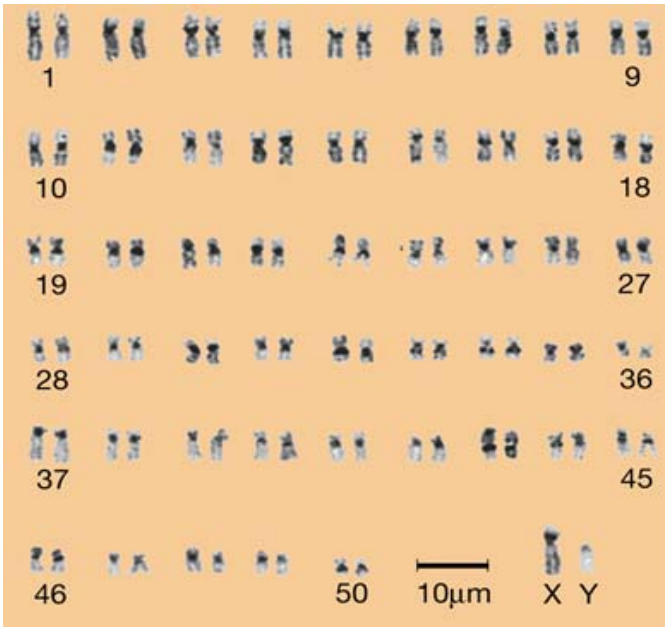


Figure 1.6 Karyotype of a male tetraploid *Tympanoctomys barrerae* from Mendoza, Argentina taken from Gallardo *et al* (1999). The karyotype contains 36 pairs of metacentric to submetacentric chromosomes and 14 pairs of subtelomeric autosomes. The X chromosome is the largest element (present in two copies in females) and the Y chromosome is the only acrocentric element of the karyotype.

Ohno (1970) argued that genome duplication has been more important than tandem duplication because the latter may duplicate only parts of the genetic system of structural genes and regulatory genes. This may disrupt the function of the duplicate genes, whereas polyploidisation duplicates the entire genetic system. However, evidence from the human genome has shown that most genes do not exist as a single copy in the genome but rather as clusters. Thus, showing that tandem duplication has played an important role in moulding the present-day structure of the human genome. It has also been seen that tandem duplication is both important for increasing the number of genes with the same function, exemplified by the HLA class I genes, and for generating genes with new functions, such as the human β-globin genes.

1.9 What happens after gen(om)e duplication?

Gene duplication is an important mechanism for the creation of new gene function (Ohno, 1970; Lynch and Conery, 2000; Wagner, 2001). After gene duplication the two resulting paralogues can evolve in different ways. The classical model of functional diversification after duplication indicates that one copy of a gene maintains the original function of the ancestral gene whereas the other gene is redundant and will either diverge functionally or be lost from the genome altogether by the accumulation of random mutations (Ohno, 1970). More recently an alternative model has been proposed, in which the two gene copies acquire complementary loss-of-function mutations and develop independent sub-functions, such that both genes are required to produce the full complement of functions of the ancestral gene (Force *et al*, 1999). The process is known as both the sub-functionalisation model and the duplication-degeneration-complementation (DDC) model.

1.10 The extended Major Histocompatibility Complex

The human Major Histocompatibility Complex (MHC) is located on the short arm of chromosome 6 (6p22.2-p21.3). The region was identified in humans over 50 years ago because of its role in tissue transplant rejection (Dausset, 1958) and is now one of the best characterised and studied regions in the human genome. It contains a high density of immune-related genes responsible for recognising foreign antigens and eliciting an adaptive immune response. The region has been linked with more diseases than any other region in the human genome (Price *et al*, 1999). It is of particular biological importance due to its association with a number of autoimmune diseases, including insulin dependent diabetes mellitus, multiple sclerosis, systemic lupus erythmatosus and rheumatoid arthritis (Thomson, 1995). In addition it has been linked with a range of aetiologies from cancer to sleeping disorders (The MHC Sequencing Consortium, 1999).

The MHC has traditionally been divided into three regions: the class I (most telomeric), class III and class II (most centromeric). The complete 3.6 Mb contiguous sequence of the three MHC regions was published in 1999 by the MHC Sequencing Consortium prior to the release of the 'draft' human genome sequence. It was estimated that 40% of the 224 genetic loci (of which 128 are expressed) have an immune function, although many still have an unknown function. Work on these three regions revealed that sequence conservation and possibly linkage disequilibrium extended further; the immediate flanking regions were termed the extended class I region and extended class II regions of the MHC (Stephens *et al*, 1999). The region of the human genome encompassing all five regions is now termed the 'extended Major Histocompatibility Complex' and spans almost 8 Mb and contains over 390 genetic

loci. The extended MHC represents a well characterised region of the human genome and is one of the best examples for the involvement of gene duplication events during its evolution.

1.10.1 The extended class I region

The extended MHC class I region (figure 1.7) has been defined as the region between the hereditary haemochromatosis locus (HFE) and the MOG locus, spanning almost 4 Mb of genomic sequence (Stephens *et al*, 1999). The region is characterised by a number of gene clusters suggesting that this region has evolved via numerous local duplication events, or through the recruitment of similar genes into the region.

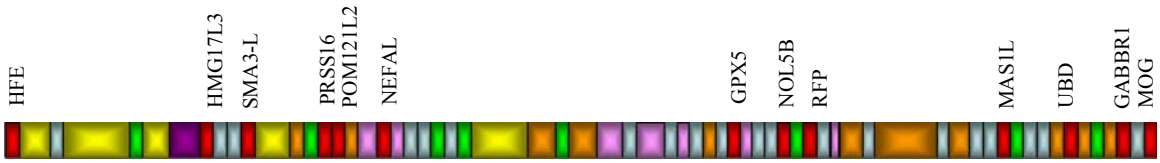


Figure1.7 Schematic representation of the extended MHC class I region. Gene clusters and individual genes are coloured according to family: histones (yellow), ribosomal proteins (green), butyrophilin receptors (purple), zinc finger proteins (pink) and olfactory receptor genes (orange). The expressed genes (red) that do not belong to a gene family cluster are labelled accordingly. Pseudogenes that do not belong to a gene family cluster are coloured grey.

There are 55 histone genes within this region, which is the largest cluster of histone genes in the human genome (Marzluff *et al*, 2002). There are also over 160 small single exon (50-100 bp in length) tRNA genes which produce 18 out of the 20 commonly used amino acids and represents approximately 25% of the human tRNA repertoire (not shown on figure 1.7). Other clusters located within the extended MHC class I region include; 20 zinc-finger proteins, 10 ribosomal proteins, two clusters of olfactory receptor genes and seven butyrophilin genes. There is further evidence of

local duplication events involving the GPX5 gene and POM121L2 gene, which both have a pseudogene in close proximity. In addition, there are also a number of expressed single copy genes, as well as pseudogenes in the extended class I region.

1.10.2 The class I region

The MHC class I region (figure 1.8) contains the three functional, classical class I genes, HLA-A, HLA-B and HLA-C, which are highly polymorphic and are expressed by most nucleated cells. In addition, there are several other functional class I loci, including the non-classical class I genes, HLA-E, HLA-F and HLA-G, which are less poylmorphic and have restricted expression. These genes are termed HLA class I genes in this thesis. The HLA-H, HLA-J and HLA-K gene fragments are thought to be pseudogenes.

The HLA class I genes encode the heavy (α) chain of the cell-surface class I molecule which, along with the β chain encoded by the β2-microglobulin locus on chromosome 15, is responsible for presenting antigens (short, specific processed peptides) to T-cells. The peptides loaded onto the class I molecules are generally derived from an endogenous (intracellular) source by the proteasome, of which PSMB8 and PSMB9 (both found within the MHC class II region; Driscoll *et al*, 1993) are subunits. The peptides are then transported to the endoplasmic reticulum by the TAP1/TAP2 molecules (also encoded by genes within the MHC class II region; Ortmann *et al*, 1994), where they are loaded onto MHC class I molecules and proceed as a complex to the cell surface via the Golgi apparatus. At the cell surface, the MHC class I molecule-peptide complex is accessible to $CD8^+$ cytotoxic T lymphocytes that elicit an immune response, which results in the lysis of the cell presenting the antigen (for a

review of class I antigen presentation see Monaco, 1992).

Of the 50 or more non-HLA related genes within the class I region, there are genes that are distantly related to the conventional class I sequence, namely the MIC genes. There are also a large number of pseudogenes (almost half of the genes) and multigene families, such as the P5 and HCG families; suggesting that duplication events contributed to the evolution of the class I region (Shiina *et al*, 1999).

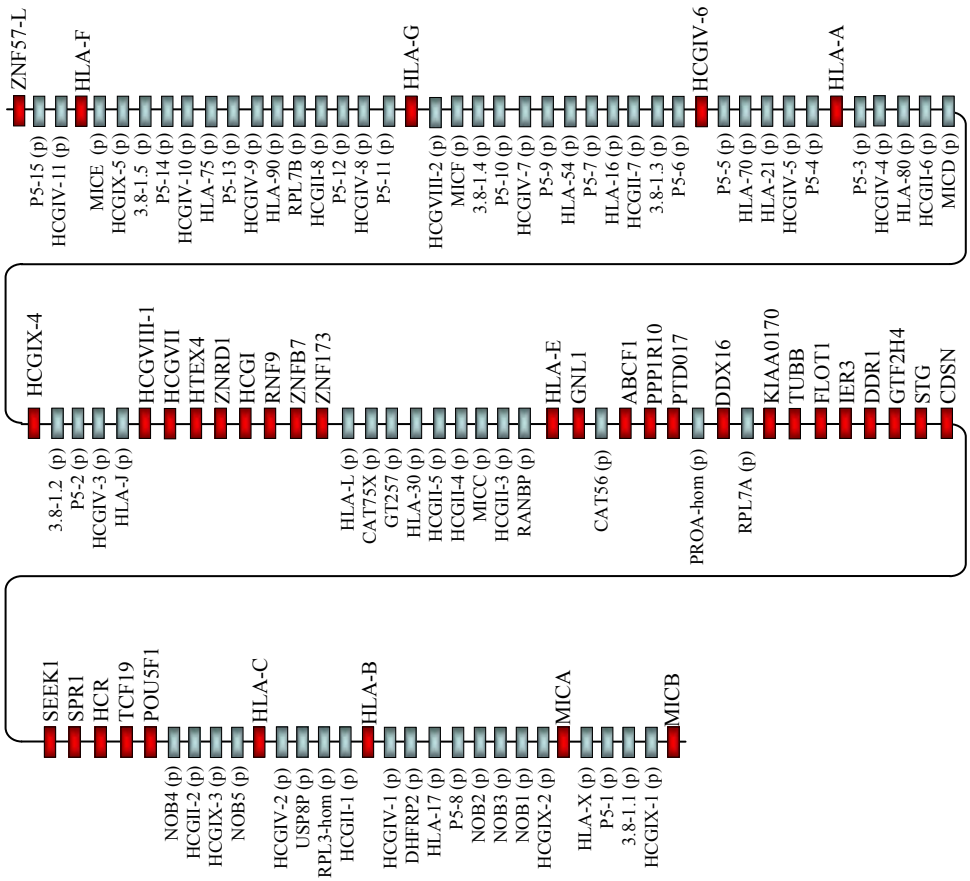


Figure 1.8 The MHC class I region. The expressed genes (red) are labelled above the gene track and the pseudogenes (grey) are labelled with a 'p' below the gene track.

1.10.3 The class III region

The MHC class III region (figure 1.9) spans approximately 0.7 Mb and is extremely

gene dense with 58 genes (this corresponds to 1 gene every 12 kb of DNA). The extent of gene density is demonstrated by the overlapping genes AGPAT1 and C6orf8, which are transcribed in different directions but overlap by 87 bp at their 3-prime ends. Furthermore, the TNXB and CYP21A2 genes overlap in the 3-prime untranslated regions. The high gene content of the class III region is complemented by a corresponding high GC content (53%). This produces a distinct boundary between the class III region and the rest of the MHC (reviewed by Beck and Trowsdale, 2000). The genes encoded in the class III region have a variety of functions and are associated with diseases, such as congenital adrenal hyperplasia and C2 deficiency (reviewed by Gruen and Weissman, 2001). There are a number of genes with an immune-related function, including members of the complement cascade (C2, C4 and BF) and the tumour necrosis factor family (TNF, LTA and LTB). Genes that are expressed in specialised cells of the immune system, such as LST1 and 1C7, are located next to each other (Holzinger *et al*, 1995).

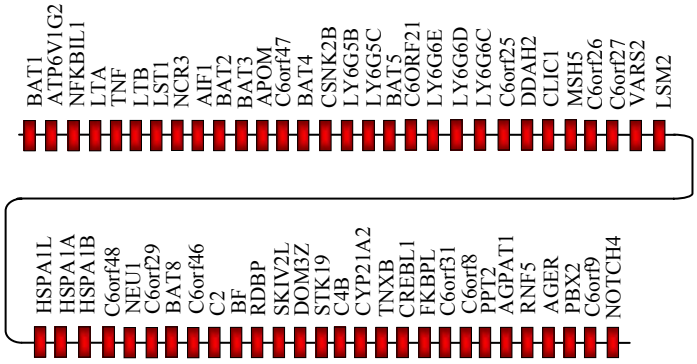


Figure 1.9 The MHC class III region. Expressed genes labelled in red are plotted from the most telomeric (BAT1) to centromeric (NOTCH4) end of the chromosome.

Gene duplication has occurred to a lesser extent in the class III region as compared with other regions of the MHC. Three genes, C4/CYP21A/TNX, have undergone

tandem duplications yielding a complex comprising overlapping genes and genes within genes (Bristow *et al*, 1993). There are also small clusters of gene families, including the three heat shock proteins located next to each other and members of the Ly6 superfamily. The class III region is unique compared with the rest of the MHC region as it does not contain any pseudogenes (with the exception of the C4 duplicate in certain haplotypes), has few duplicated genes and the genes have diverse functions, suggesting a distinct origin of the class III region.

1.10.4 The class II region

The MHC class II region (figure 1.10) takes its name from the classical and non-classical HLA class II genes, termed HLA class II genes in this thesis. The classical HLA class II genes (HLA-DP, HLA-DQ, HLA-DR) either encode proteins with α chains (HLA-DPA, HLA-DQA, HLA-DRA) or proteins with β chains (HLA-DPB, HLA-DQB, HLA-DRB). The α and β chains combine to form class II MHC molecules. The class II molecules are polymorphic and are expressed on specialised antigen-presenting cells (e.g. dendritic cells, B lymphocytes, macrophages) and present peptides mainly derived from extracellular proteins to $CD4^+$ T cells.

MHC class II molecules differ from MHC class I molecules in that the groove of the peptide-binding region (PBR) is open-ended, thus allowing longer peptides to be bound. Prior to a peptide binding, the class II molecules are assembled in the endoplasmic reticulum (ER) with a membrane-bound chaperone protein (known as the MHC class II associated invariant chain or γ chain) acting to stabilise the complex. This γ chain is degraded by proteases in the trans-Golgi reticulum with the exception of a small fragment that is buried in the PBR. The removal of this small fragment

(prior to peptide binding) is catalysed by gene products of the HLA-DM gene – a non-classical class II gene. After binding, the MHC class II molecule-peptide complex is transported to the cell surface where it is recognised by $CD4^+$ helper T lymphocytes (for a review of class II antigen presentation see Neefjes and Ploegh, 1992 and Pieters, 1997).

Within the class II region there are also a number of other genes that have an immune related function. The PSMB8, PSMB9, TAP1 and TAP2 genes are involved with antigen processing of MHC class I molecules as described in section 1.10.2. There are also a number of genes with quite diverse functions, such as the butyrophilin-like gene BTNL2, the testis-specific basic protein TSBP and the bromodomain-containing protein BRD2. Furthermore, a number of pseudogenes are located within this region, including a ribosomal protein pseudogene and the pseudogene of the extended class II gene COL11A2.

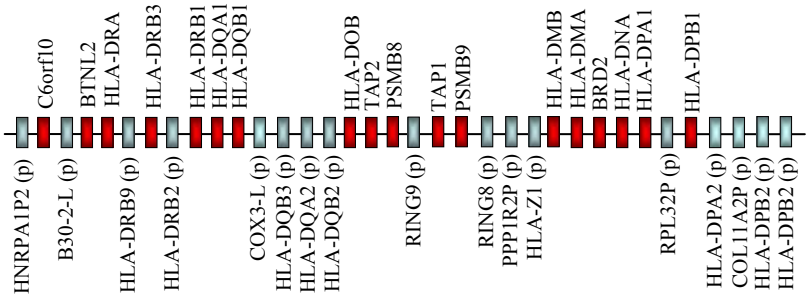


Figure 1.10 The MHC class II region. The expressed genes are shown in red and are labelled above the gene track and the pseudogenes (labelled with a 'p') are in grey and are labelled below.

1.10.5 The extended class II region

The identification of the tapasin gene, required for antigen presentation by MHC class

I molecules, in the region flanking the MHC class II region suggested that the MHC extended further than previously thought (Herberg *et al*, 1998a; 1998b). Detailed analysis of the region centromeric to the MHC class II region, now termed the extended class II region (figure 1.11), revealed several other genes, including collagen gene type 11A2 (COL11A2), a ribosomal protein RPS18 and, the most centromeric gene in the extended MHC, KNSL2.

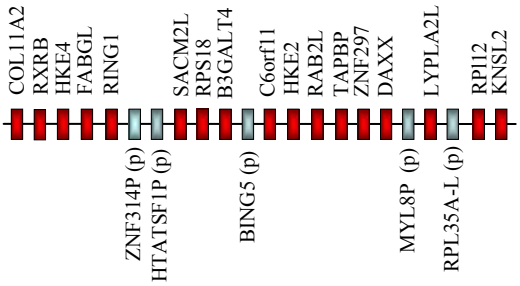


Figure 1.11 The extended MHC class II region. The expressed genes are shown in red and are labelled above the gene track and the pseudogenes (labelled with a 'p') are in grey and labelled below.

1.11 Origin of the extended MHC

There is conservation of some of the genes within the extended MHC regions between species suggesting that there is an evolutionary advantage in conserving the MHC as a unit. This MHC 'unit' can be observed in species evolving after the divergence of the jawless vertebrates. In particular, the MHC class I and class II region genes have been identified in all jawed vertebrates studied to date, but have not been identified in the jawless vertebrates, hagfish and amphioxus (Kasahara *et al*, 1996b; Flajnik *et al*, 1999). Jawless vertebrates also lack other molecules of the adaptive immune system, such as RAG1 and RAG2, as well as the lymphoid organs thymus and spleen. Thus, the adaptive immune system has arisen in a very short period of geological time since

the emergence of jawed vertebrates (Bernstein *et al*, 1996). Several MHC genes (including NOTCH4, RXRB and PBX2) are syntenic in invertebrate genomes, such as *Drosophila* and *C. elegans* indicating that the origin of the MHC locus predates the emergence of the adaptive immune system

The three classical regions of the human MHC (class I, class III and class II) appear to have been subject to different evolutionary mechanisms: whilst MHC class II and class III genes often appear to have direct orthologues, the MHC class I genes appear to have expanded and contracted in different species. The class III region is considered to be the oldest region of the MHC (reviewed by Beck and Trowsdale, 2000). It is evident that both the class I and class II regions have evolved via a series of duplications, but it is not known which region came first. One hypothesis claims the class II region evolved first (Hughes and Nei, 1993), whereas another hypothesis holds that the class I region originated first as a result of a recombination between an immunoglobulin-like C-domain and the peptide-binding domain of an HSP70 heat shock protein (Flajnik *et al*, 1991). Phylogenetic analysis supported the prior hypothesis, albeit with low statistical support (reviewed by Hughes and Yeager, 1997; Klein and Sato, 1998).

1.12 MHC Paralogy

MHC paralogous genes were observed during the study of MHC class III genes (Sugaya *et al*, 1994; Katsanis *et al*, 1996) and the class II proteasome genes (Kasahara *et al*, 1996a). It was concluded that the region 9q33-q34 was paralogous to the MHC. Furthermore, Katsanis and colleagues (1996) also noted two additional regions in the human genome, 1q21-q25/1p11-p32 and 19p13.1-p13.3, which contained MHC

paralogues (Figure 1.12). Initially, only a few genes were reported to have paralogues on chromosomes 1, 9 and/or 19 but the number has increased to 40, approximately one third of the expressed MHC genes (reviewed by Kasahara, 1999b).

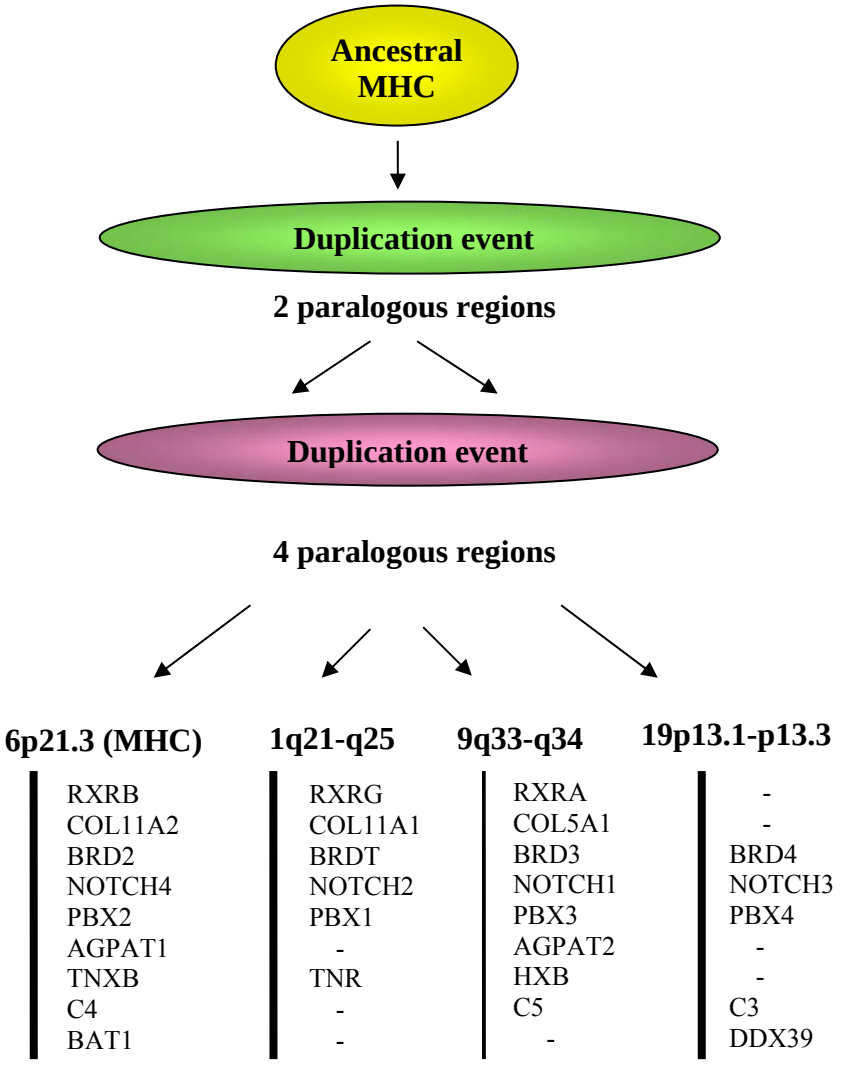


Figure 1.12 Summary of the MHC paralogous regions in the human genome.

The MHC genes with paralogues reside in both the classical and extended regions of the MHC region and constitute a diverse group of genes in terms of structure, function and gene size. Some families, such as NOTCH and PBX, have copies in all four

regions, but most only have two or three copies. MHC paralogues may not be identified in all regions, for each gene, as duplicated genes are likely to be silenced or lost from the genome altogether (Nadeau and Sankoff, 1997). Interestingly, there are also a number of other gene families that have copies in the 1, 9 and 19 paralogous regions but not in the MHC (reviewed by Kasahara *et al*, 2000).

1.12.1 Origin of the extended MHC paralogous regions

The origin of the MHC and the three paralogous regions is controversial. Currently there are two main hypotheses. The first is that they descended from a common ancestral region and emerged as a result of large-scale block duplications (Kasahara *et al*, 1996a; Kasahara, 1999a; Abi-Rached *et al*, 1999; Flajnik and Kasahara, 2001). The second is that they represent assemblies of independently duplicated genes and are grouped together by selective forces (Hughes, 1998). In general, the block duplication mechanism is preferred, as it can best explain why assortments of functionally and structurally varied genes are clustered on four specific regions of the human genome.

It is possible that the MHC and the paralogous regions on chromosomes 1, 9 and 19 arose from two rounds of whole-genome duplications (the 2R hypothesis) based on the estimated timings and numbers of duplications that appear to have occurred (reviewed by Flajnik and Kasahara, 2001). It is believed that the first round of duplication occurred close to the origin of jawed vertebrates. This is supported by the identification of a single orthologue of the MHC paralogues, exemplified by the complement genes C3, C4 and C5. All have been identified in cartilaginous fish, however, jawless fish lack C4 and C5 but do have a C3-like gene that shares features

with the common ancestor of C3 and C4 (Kasahara, 1999a).

In order to understand the evolution of the MHC and associated paralogous regions, numerous genes have been analysed in a number of organisms, including *Drosophila* and amphioxus which are thought to predate the whole genome duplication events proposed by the 2R hypothesis. The identification of 19 MHC paralogous genes in *Drosophila* (Danchin *et al*, 2003) and nine MHC paralogous genes in amphioxus (Abi-Rached *et al*, 2002) residing in close proximity to other genes found on human 1q21-q25/1p11-p32, 9q33-q34 and 19p13.1-p13.3 provides evidence for a block duplication event in vertebrates. Phylogenetic analysis has demonstrated that the duplications occurred prior to vertebrate emergence but after the divergence of amphioxus from the vertebrate lineage supporting the 2R hypothesis.

If the paralogous regions have a common origin there should be evidence of conserved synteny between them. Preliminary analysis of the gene order within the paralogous regions indicated that the order is poorly conserved (Endo *et al*, 1997). This is no surprise considering that more than 500 million years have passed since the last duplication event occurred and each region has undergone major structural rearrangements, including inversions and translocations. Rearrangements are particularly dramatic on chromosome 1, as the MHC paralogues are located on both arms of the chromosome. There is compelling evidence that chromosome 1 underwent a pericentromeric inversion after the divergence of the human and chimpanzee lineages and this is probably responsible for the occurrence of the paralogous genes on both arms (Maresco *et al*, 1996).

1.13 Thesis aims

With the discovery that approximately 10% of the human genome arose by duplication it is evident that this process has played a major role in the evolution of our genome. Whether the duplication events involved the entire genome (as proposed by the 2R hypothesis), chromosomal segments or individual genes is unclear. Therefore, the aim of this thesis was to investigate the mechanism(s) that gave rise to the present-day organisation of the human genome.

Using the MHC region as a model a number of aspects of paralogy and the genome were investigated. Firstly, in order to study the genomic regions containing MHC paralogous genes, the chromosomal region 9q32-q34.3 was mapped, sequenced and analysed. Comparison of 9q32-q34.3 with the MHC region on 6p22.2-p21.3 will reveal the level of synteny between these two regions and determine whether they have a common origin. Secondly, a survey of the entire human genome sequence was conducted to identify the MHC paralogues and determine their distribution. Thirdly, phylogenetic trees were used to reconstruct the evolutionary histories of the MHC paralogues. Analysis of the topology of the trees and the arrangement of the paralogues and orthologues will determine the mechanism(s) of evolution. Finally, the expression profiles were generated to understand how the MHC paralogues have evolved since their emergence. In summary, the data presented in this thesis aims to provide a unique insight into the evolution of the MHC paralogous genes and the human genome.

Chapter 2

Materials and Methods

2.1 Materials

The majority of chemical reagents were bought from Sigma Chemical Co., BDH Chemical Ltd., and Difco Laboratories unless stated in the text. Similarly restriction enzymes were bought from New England Biolabs unless stated differently. The sources of the commercial kits used in this thesis purchased from various companies are stated in the text. All primers were synthesised in-house by the Sanger Institute Oligo Preparation laboratory. PCR was generally performed using Amplitaq® DNA Polymerase from Perkin Elmer.

2.1.1 Solutions, buffers and media

Solutions used in this thesis are listed according to the methods section they were used. All solutions were made up in double-distilled water (ddH_20), unless stated otherwise.

used in section 2.4

2 x TY: 15 mg/ml bacto-tryptone, 10 mg/ml yeast extract, 5 mg/ml NaCl (pH 7.4) ddH_20 up to 1 litre and autoclave to sterilise.

Chloramphenicol: 30 mg/ml stock made up in 100% ethanol, filtered to sterilise and

stored at -20°C. Used at 30µg/ml final concentration.

GTE: 50 mM Glucose, 25 mM Tris (pH7.5), 10 mM EDTA

NaOH/SDS: 0.2M NaOH, 1% (w/v) SDS

3 M KOAc (pH5.5): 300mM potassium acetate (pH4.8), 11.5ml glacial acetic acid, 28.5ml H_20

Boehringer buffer B: 50mM NaCl, 10mM Tris-HCl, 10mM $MgCl_2$, 1mM DTE, pH7.5.

6 x Buffer II: 0.25% bromophenol blue, 0.25% xylene cyanol, 15% ficoll

Vistra green solution: mix 5ml 1 M Tris HCl, 0.5ml 0.1 M EDTA, 50µl Vistra Green (Amersham Life Sciences) made up to 500 ml with ddH_20.

used in section 2.5

10x nick translation buffer: 0.5M Tris-HCl (pH7.5), 0.1M $MgSO_4$, 1mM dithiothreitol, 500µg/ml bovine serum albumin

DNase I (1µg/ml): dilute 10 mg/ml Deoxyribonuclease I (Sigma) stock to 1µg/ml working solution with enzyme diluent.

Enzyme diluent: 500µl glycerol, 100µl nick translation buffer, 400µl ddH_20

Fixative: 3:1 methanol/glacial acid

Formaldehyde fixative: 1% v/v formaldehyde (from 40% stock), 50 mM $MgCl_2$ in PBS. For 50 ml add 1.25 ml formaldehyde, 2.5 ml 1M $MgCl_2$ and make-up to 50mls

with 2xSSC.

FISH Hybridisation buffer: 50% deionised formamide, 2x SSC (pH7.0), 10% dextran sulphate, 0.1% SDS, 1x Denhardt's solution, 40mM sodium phosphate pH7.0.

4 x TNFM: 4 x SSC, 0.05% Tween 20, 5 % non-fat milk powder, filtered through several layers of Whatman No.4 filter paper

used in section 2.6

Mung bean nuclease buffer: 100µl 3 M sodium acetate, 250 µl 2 M sodium chloride, 10 µl 1 M zinc chloride, 140 µl water, 500 µl mung bean nuclease (Pharmacia), 500 µl glycerol

Buffered phenol: 1 ml phenol, 200 µl 1 M Tris-hydrogen chloride (shaken and placed on ice for 5 minutes, centrifuged, top layer removed and discarded, 200 µl TE added, mixed, shaken and centrifuged. Kept on ice until needed.

SOC: SOB + 200 µl 20% glucose

SOB: 20 g tryptone, 5 g yeast extract, 10 ml 1M sodium chloride, 0.5 g potassium chloride, water added up to 1 litre

TYE agar: 8 g tryprone, 5 g yeast extract,8 g sodium chloride, 12 g agar, water upto 1 litre

TYE/Amp plates: 2 ml of 25 mg/ml ampicillin was added to 1 ml TYE autoclaved solution which was allowed to cool to 48°C before addition.

IPTG: 40 mg/ml in DMSO. Sterilised by filtration and stored at -20°C.

Xgal: 50 mg/ml in ddH_20. Sterilised by filtration and stored at -20°C.

0.1% DMSO: Dimethyl sulfoxide diluted in ddH_20 and autoclaved.

used in section 2.7

2 x LB: 10 mg/ml bacto-tryptone, 5 mg/ml yeast extract, 10 mg/ml NaCl (pH 7.4) ddH_20 up to 1 litre and autoclaved to sterilise.

Ampicillin: 25 mg/ml Ampicillin stock made up in ddH_20, filtered to sterilise and stored at -20°C.

GET: 30 mM Glucose, 15 mM Tris-HCl (pH8.0), 30 mM Na_2EDTA, 60 µg/ml RNase A

Bind solution: (6.1 M Potassium Iodide) 40 g potassium iodide in 28 ml ddH_20. Stored in the dark at room temperature.

Precipitation mix: 100 ml 96% ethanol, 2 ml 3 M sodium acetate, 4 ml 0.1 mM EDTA

Sequencer Loading dye: 25 mM EDTA (pH 8.0), 50 mg/ml Blue dextran, deionised formamide (5:1 formamide: EDTA/Blue dextran).

used in sections 2.8-2.13

10 x PCR buffer: 500 mM KCl, 50 mM Tris (pH8.5), 25 mM $MgCl_2$.

PBS: 10 g sodium chloride, 0.25 g potassium chloride, 1.44 g sodium hydrogen

phosphate (dibasic), 0.25 g potassium dihydrogen phosphate, made up to 1 litre with water and made to pH with sodium hydroxide. Stored at 4°C.

4 x Spotting buffer: 1 M sodium phosphate buffer pH 8.5, 0.001% sarkosyl

Bacterial mRNA "cocktail": pool of cDNA bacterial clones for B.subtilis *trp* gene (30 ng/µl), *lysA* gene (0.3 ng/µl), *thrB* gene (3 ng/µl), *dapB* gene (15 ng/µl), *pheB* gene (1.5 ng/µl) all purchased from American Type Culture Collection (ATCC catalogue numbers 87482 to 87486)

Microarray hybridisation buffer: 5 x SSC, 6 x Denhardts solution, 60mM Tris HCl (pH7.6), 0.12% sarkosyl, 48% formamide filter sterilised

100 x Denhardt's solution: 20 mg/ml Ficoll 400-DL, 20mg/ml polyvinylpyrrolidine 40, 20mg/ml BSA (pentax fraction V)

Microarray wash solution 1: 2 x SSC, filter sterilised

Microarray wash solution 2: 0.1 x SSC, 0.1 % SDS, filter sterilised

Microarray wash solution 3: 0.1 x SSC, filter sterilised

used in section 2.14

Spermidine stock: 25.46 mg/ml spermidine (Sigma) in 10 mM Tris (pH 7.4)

1 x denaturation solution: 0.5 M NaOH, 1,5 M NaCl

1 x neutralisation solution: 1.5 M NaCl, 1 M Tris-HCl (pH7.4)

MTN/Southern Wash solution I: 2 x SSC, 0.05% SDS, filter sterilised

MTE Wash solution I: 2 x SSC, 1% SDS, filter sterilised

MTN/Southern Wash solution II: 0.1 x SSC, 0.1% SDS, filter sterilised

MTE Wash solution II: 0.1 x SSC, 0.5% SDS, filter sterilised

General solutions, buffers and media used in this thesis

10 x TBE: 890 mM Tris base, 890 mM Borate, 20 mM EDTA (pH8.0)

50 x TAE: 2 M Tris base, 5.7% v/v glacial acetic acid, 50 mM EDTA

20 x SSC: 175.3 g sodium chloride, 88.2 g sodium citrate, made up to 1 litre with ddH_20.

1 x $T_{0.1}E$: 10 mM Tris-HCl (pH8.0), 0.1 mM EDTA

1 x TE: 2 ml Tris (pH7.4), 200 µl 0.1 M EDTA, ddH_20 to 200 ml.

10 mM dNTPs mix: 1 ml of each dNTP (100 mM) in 6 ml of ddH_20. Stored at -20 °C.

1µl 10 mM dA,T,GTP/5 mM dCTP mix: 25 µl dATP, 25 µl dTTP, 25 µl dGTP, 10 µl dCTP and 15 µl ddH_20.

2.1.2 Loading dyes

Loading dye: 5 mg bromophenol blue, 0.5 g Ficoll, 0.5 ml 10 x TBE, 4.5 ml ddH_20

Loading buffer: 10 µl 10 x TBE, 20 µl loading dye, 50 µl water

Sequencer loading dye: 25 mM EDTA (pH 8.0), 50 mg/ml Blue dextran, deionised formamide (5:1 formamide: EDTA/Blue dextran).

2.1.3 Nucleotides

Amersham Biosciences	Redivue™ deoxycytidine 5'-[α-^{32}P]-dCTP-triphosphate, triethylammonium salt (AA0075)
Invitrogen	Renaissance R Cyanine 3-dCTP (NEL576) Renaissance R Cyanine 5-dCTP (NEL575)
Pharmacia	100mM dATP, dCTP, dGTP, dTTP (27-2035-01)
Boehringer	Biotin-16-dUTP (1 mM) (1093-070)

2.1.4 Size markers and ladders

1kb DNA ladder (1 µg/µl) (Gibco BRL Life Technologies): 5 µl 1 Kb DNA ladder mixed with 1 µl 50 x TAE, 10 µl Ficoll dye and 34 µl ddH_20. The 1 Kb DNA ladder contains 1 to 12 repeats of a 1018 bp fragment and vector fragments from 75 to 1636 bp to produce the following sized fragments in bp: 75, 142, 154, 200, 220, 298, 344, 394, 516/506, 1018, 1635, 2036, 3054, 4072, 5090, 6108, 7125, 8144, 9162, 10180, 11198, 12216.

100 bp DNA ladder (1µg/µl) (Gibco BRL Life Technologies): 50µl (1µg/µl) 100 bp

DNA ladder, 60 µl loading buffer and 390 µl ddH_20. The 100 bp DNA ladder consists of 15 blunt-ended fragments between 100 and 1500 bp in multiples of 100 bp and an additional fragment at 2072 bp. The 600 bp fragment is approximately 2 to 3 times brighter that the other ladder bands to provide orientation.

Fingerprinting gel marker: 19.2µl $T_{0.1}E$, 1.5µl Analytical Marker DNA wide range (Promega), 0.2µl Molecular Weight Marker V (Boehringer-Mannheim) and 4.2µl 6x loading dye were added to a 1.5ml microfuge tube and stored at -20°C). The Analytical Marker DNA wide range provides an evenly spaced distribution of DNA fragments from 0.702 kb to 29.95 kb.

Lambda DNA-Hind III marker (Gibco BRL Life Technologies): 8 µl lambda DNA-*Hind* III digest, 60 µl TBE buffer, 252 µl ddH_20 was incubated at 65°C for 5 minutes then snap chilled on ice before 80 µl loading dye was added. The *Hind* III digest of lambda DNA yields 8 fragments suitable for use as molecular weight standards for agarose electrophoresis of the following sizes; 125, 564, 2027, 2322, 4361, 6557, 9416, 23130.

Lambda Hind III/pBR322 marker: 8 µl lambda *Hind* III (NEB), 60 µl 10 x TAE, 252 µl ddH_20 were heated at 65°C for 5 minutes then snap chilled on ice then 6 µl pBR322 BstNI (NEB) and 80 µl loading dye were added.

2.1.5 Sources of DNA and RNA

Human genomic DNA was purchased from Clontech (catalogue number 6550-1). Total RNA was extracted from Raji, Jurkat, 293T and U937 cell-lines (a kind gift from John Trowsdale, Division of Immunology, Department of Pathology, University of Cambridge) and a THP1 cell-line (kindly provided by Paul Lehner, CIMR, Cambridge). Total human Adrenal Gland, Brain, Skeletal Muscle, Spleen and Testis RNA were purchased from Ambion (catalogue numbers 7994, 7962, 7982, 7970 and 7972, respectively). Universal Human Reference RNA was purchased from Stratagene (catalogue number 740000). The human multiple tissue expression (MTE™) array and the multiple tissue northern (MTN™) were purchased from Clontech (catalogue numbers 7776-1 and 7760-1, respectively). Information regarding the sources of RNA can be found on the web-site http://www.clontech.com.

Methods

2.2 Agarose gel preparation and electrophoresis

Unless stated otherwise in the text: agarose gels were prepared in either 1x TBE or 1 x TAE buffer containing 250ng/µl ethidium bromide and the appropriate percentage of agarose was used according to the size of fragments being separated; a 2.5% agarose gel was used for electrophoresis of fragments below 1kb, and a 0.8-1.0% agarose gel for analysis of larger fragments. Electrophoresis was performed at 50-100 V for 15-45 minutes depending on separation required. The sizes of the DNA fragments were estimated by running either the 1 kb or 100 bp ladder size standards.

2.3 Sequencing gel

The denaturing acrylamide gel (6%) was made up using 30 g urea in 9 ml acrylamide/bisacrylamide solution, 4 ml 10 x TBE and 37 ml ddH_20. The urea was dissolved by heating to 60°C and stirring. The solution was made up to 60 ml with water and placed in a dessicator for 4 minutes. Just prior to pouring the gel, 138 µl of 25% ammonium persulphate and 138 µl TEMED were added. The gel was then carefully syringed between the glass plates whilst tapping the glass gently to get rid of air bubbles. The appropriate comb was inserted and the gel was left to set for at least 90 minutes prior to use.

Mapping and sequencing

2.4 Restriction Digest Fingerprinting

The BAC genomic clone, bA465F21 (AC006313) was fingerprinted using the *HindIII* digest fingerprinting method essentially as described by Olson *et al*, 1986.

2.4.1 Filterprep isolation of BAC DNA

1. 500 µl of 2 x TY containing 30 µg/ml of chloramphenicol were added to a 96-well 1 ml Beckman box.
2. Each well was inoculated from a glycerol stock with either a 96-well inoculating tool, or a sterile cocktail stick. A plate sealer was placed on top of a plate to seal the wells and the cultures grown for 16-18 hours at 37°C with gentle shaking (300 rpm).
3. For each well, 250 µl of the overnight growth were transferred to a clean round-bottomed Corning microtitre plate using a 50- to 250-multichannel pipette (Finnpipette). The cells were pelleted by centrifugation at 2500 rpm at 20°C from 4 minutes.
4. For each well, the supernatant was discarded and the pellet re-suspended in 25 µl of GTE and mixed gently by vortexing. 25 µl of freshly prepared NaOH/SDS solution was added and mixed by tapping the plate gently and left to stand at room temperature for 5 minutes.
5. 25 µl of chilled 3 M KOAc (pH5.5) solution were added, mixed and left at room temperature for 5 minutes.
6. A microtitre plate containing 100 µl of isopropanol was taped to the bottom of

2 µm filter-bottomed plate (Millipore). The total well volume of the sample was transferred to the filter-bottomed plate.

7. These 2 plates were then spun at 2500 rpm, 20°C for 2 minutes to ensure all liquid had been transferred from the filter plate to the lower plate; the filter plate was then discarded.
8. After separation from the filter plate, the lower (Corning) plate was left at room temperature for 30 minutes before being centrifuged at 3200 rpm, 20°C for 20 minutes.
9. The supernatant was discarded from the plate and the DNA pellet was briefly dried by inverting the plate and placing on clean tissue paper.
10. 100 µl of 70% ethanol were added to the dried DNA to wash the pellet, mixed gently, and the DNA precipitated by centrifuging at 3200 rpm 20°C for 10 minutes. This step was repeated.
11. Finally, the supernatant was discarded and the DNA pellet was dried before being resuspended in 5 µl of fresh $T_{0.1}E$ with RNase (1µg/ml).
12. Samples were stored at -20°C.

2.4.2 Restriction digest fingerprinting (*Hind* III) of BAC DNA

1. For one 96-well microtitre plate of sample DNA, a premix containing 286 µl ddH_20, 99 µl Boehringer buffer B, 55 µl *Hind* III was prepared in a 1.5 ml microfuge tube, and mixed by vortexing.
2. 4 µl of the premix was added to each well of a 96 well-microtitre plate containing previously prepared DNA (see section 2.4.1) and mixed gently by vortexing at 37°C for 2 hours.

3. The reaction was terminated by adding 2 µl of 6x Buffer II and either stored at 4°C or loaded immediately.
4. 0.8 µl of the fingerprinting marker was added to the first well and then every sixth well of a freshly prepared 1% agarose/1 x TAE gel. 1 µl of each sample was loaded (i.e. wells 2-5, 7-10 *etc)* between the marker lanes. Fragments were resolved by electrophoresis through the gel at 4°C in a cold room for 15 hours at 90 volts.
5. Following electrophoresis, the gel was cut down so the length was 19-20 cm and stained with Vistra Green solution for 30-45 minutes on a shaker. The gel was washed with ddH_20 to remove excessive stain.
6. The gels were scanned on a FluorImager SI. The parameters were set to 530 nm for emission filter, the pixel size was 100 microns, detection sensitivity was normal, digital resolution was at 16 bits, dye was single label, excitation filter was 488 nm, Em filter 1530 nm and PMT voltage was 800.
7. The gel image was transferred to a UNIX workstation and entered into the fingerprint 'IMAGE' analysis system (Sanger Institute in house software).The band pattern was extracted using 'IMAGE' and the data entered into another program, fingerprinted contigs, FPC (Soderlund *et al*, 1997), where the fingerprint patterns were compared to those already in the database and the position of the clone within a contig determined.

2.5 Fluorescent *in-situ* hybridisation (FISH) mapping

Cytogenetic mapping using FISH techniques were performed using chromosome 9 clones. The BAC clone, bA465F21, was fluorescently labelled and hybridised to

metaphase chromosomes to determine which chromosome it maps to (Pinkel *et al*, 1986). In addition, the orientation and order of 3 contigs were resolved by interphase FISH (Wilke *et al*, 1994) and the sizes of gaps between 5 contigs determined using extended DNA fibres using Fibre FISH (Heiskanen *et al*, 1994).

2.5.1 Labelling of FISH probe using Nick translation

1. 1 µg of clone DNA was labelled with 1mM biotin-16-dUTP (Boehringer) in a 25 µl reaction containing; 2.5 µl nick translation buffer, 1.9µl 0.5 mM dATP, dCTP and dGTP mix, 0.7 µl 1mM biotin-16-dUTP, 1µl DNase I* (Sigma), 0.5 µl DNA polymerase I (10U/µl Sigma) made-up to 25 µl with H_20.

*In order to determine the concentration of DNase I and incubation time a series of dilutions were carried out using different amounts of DNase I in 50 µl reaction volumes containing; 2 µg test DNA, 5 µl nick translation buffer, 1-2 µl DNase I (1µg/ml working stock in enzyme diluent). The reactions were incubated at 14°C for 60 minutes. A 10 µl aliquot was removed after 20 minutes with further 10 µl aliquots removed at 10 minute intervals. All samples were run on a 1% agarose gel and the DNase I concentration and incubation time which gave fragment smears with a size range of 200-700 bp used.

2. The 25 µl reaction was incubated at 14°C for 60 minutes and the labelling reaction terminated by adding 2.5 µl of 0.5 M EDTA (pH8.0)
3. 2.5 µl 3M sodium acetate (pH7.0) and 60 µl of 100% ethanol were added to the reaction and the probe precipitated at -70°C for 30 minutes.
4. The mixture was centrifuged at 13,000rpm for 10 minutes and the pellet washed twice with the addition of 500 µl 70% ethanol and centrifuged at 13,000 rpm for a further 2 minutes. The pellet was air-dried at 37°C.

5. The pellet was resuspended in 10 µl $T_{0.1}E$ and 2 µl of sample was run on a 1% agarose gel to check efficacy of the reaction.

2.5.2 Preparation of microscope slides

The slides containing the extended DNA fibres and the metaphase and interphase cell-suspensions were kindly provided by the Sanger Institute Molecular Cytogenetics group.

1. Microscope slides were washed in 2% Decon and sonicated in a sonicator bath then rinsed under cold running water for 60 minutes. Stored in 96% ethanol.
2. The slides were removed from the ethanol and polished with a dry, lint-free tissue.
3. The metaphase or interphase cell-suspension was mixed by gentle flicking of the tube and a single drop was dropped onto the slide using a Pasteur pipette.
4. A drop of fixative was added whilst the first drop was still spreading and the slide air-dried.
5. The slides were fixed in a coplin jar of fixative at room temperature for 30-60 minutes, air-dried and stored in a sealed box at room temperature until needed.
6. Prior to use the slides were incubated in 2 x SSC at 37°C for 5 minutes, followed by 5 minute incubation at 37°C in 0.01 M HCl and 10 µl of 25% pepsin (in ddH_20).
7. The slides were rinsed 3 times in 2 x SSC for 2 minutes each at room temperature and then fixed in formaldehyde fixative for 10 minutes also at room temperature.
8. The slides were rinsed again 3 times in 2 x SSC for 2 minutes at room

temperature then dehydrated through exposure to 3 concentrations of ethanol: slides were incubated at room temperature with 70%, 70%, 90%, 90%, 100% ethanol for 1 minute each.

9. The slides were air dried to evaporate the remaining ethanol.

2.5.3 Hybridisation of FISH probes

1. 1 µl Cot1 DNA (1µg) and 11.5 µl FISH hybridisation buffer was added to 0.5 µl labelled DNA (30-50 ng), mixed thoroughly and denatured at 65°C for 10 minutes.
2. The denatured probe was then incubated at 37°C for 1 hour.
3. Prior to hybridisation the slides were denatured in 70% formamide (in 2 x SSC) at 65°C for 2 minutes then quenched in ice cold 70% ethanol then dehydrated through an ethanol series (70%, 70%, 90%, 90%, 100% for 1 minute each) and air-dried.
4. The hybridisation mix was placed onto the denatured slides and covered with 22 x 22 mm cover slip. Rubber cement was used to seal the cover slips and the slides were incubated in a moist chamber at 37°C for 24 hours.
5. After hybridisation the cover slip was removed and the slides rinsed in 2 x SSC for 5-20 minutes.
6. The slides were washed twice at 42°C in 50% formamide (in 2 x SSC) for 5 minutes.
7. Further washing was performed at 42°C in 2 x SSC for 5 minutes.
8. For detection of biotinylated probes 100 µl of 4 µg/ml avidin Texas Red DCS (Vector) was added and the slide covered with Nescofilm and incubated at

37°C for 20-60 minutes in a humid chamber.

9. The slides were washed in 4 x TNFM at 42°C for 5 minutes.
10. The slides were drained and 100 µl of 4 µg/ml biotinylated anti-avidin D plus 1:500 dilution of mouse anti-digoxin (Sigma) was added. The slide covered in Nescofilm and incubated at 37°C for 20-60 minutes in a humid chamber.
11. The slides were washed in 4 x TNFM at 42°C for 5 minutes.
12. The slide covered in Nescofilm and incubated at 37°C for 20-60 minutes in a humid chamber.
13. The slides were drained and 100 µl of 4 µg/ml avidin Texas Red DCS plus 10 µg/ml goat anti-mouse FITC conjugate (Sigma) was added. The slide covered in Nescofilm and incubated at 37°C for 20-60 minutes in a humid chamber.
14. The slides were washed in 4 x TNFM at 42°C for 5 minutes.
15. The slides were drained and 100 µl of 4 µg/ml avidin Texas Red DCS (Vector) was added and the slide covered with Nescofilm and incubated at 37°C for 20-60 minutes in a humid chamber.
16. The slides were washed in 4 x SSC, 0.05% Tween 20 at room temperature and counterstained in 0.08 µg/ml DAPI (4',6'-diamidino-2-phenylindole hydrochloride) in 2 x SSC for 2-3 minutes.
17. The slides were rinsed in 2 x SSC then dehydrated through the ethanol series (70%, 70%, 90%, 90%, 100% for 1 minute each) and air-dried.
18. 20 µl of antifade solution (Citifluor AF1) was added to a clean 22 x 32 mm cover slip and overlayed on the slide. The cover slip was sealed using nail varnish.
19. The slides were analysed using a Zeiss Axioscop fluorescence microscope equipped with a CCD camera. Separate images of the DAPI staining of the

chromosomes and the biotinylated probes were merged using SmartCapture software (Digital Scientific Ltd).

2.6 Production of shotgun libraries for shotgun sequencing (essentially as described by Bankier *et al*, 1987)

The minimum set of clones to cover chromosome 9 were selected for sequencing using the large-scale maps produced by FPC fingerprinting methods. Each clone is divided into fragments by sonication which are then assembled so overlapping fragments of sequence provide the complete sequence across the clone. The random nature of sonication produces fragments that will be sequenced on average 6-8 times before a project is considered complete; this redundancy is necessary to ensure that sequencing errors are resolved. The chromosome 9 BAC clone DNA for bA18B16 was provided by the Sanger Institute Sub-cloning laboratory and sub-cloned by me.

2.6.1 Sonication and subfragment end repair of plasmid DNA

1. In order to estimate the concentration of DNA of the BAC clone, a 0.5 % agarose, 1 x TBE gel was run on a 10 x dilution of the BAC. The DNA sample was diluted 1/10 in $T_{0.1}E$ and 1 µl was run alongside lambda *Hind* III/pBR322 marker. Samples were visualised by soaking the gel in 500 ml of 1 x TBE containing 25 µl ethidium bromide (10 mg/ml) for 5 minutes the de-stained in ddH_20 for 10 minutes.
2. From the gel image, the amount of DNA required to obtain 10 µg was taken for sonication.
3. To the 10 µg DNA ddH_20 was added to a final volume of 54 µl. 6 µl of mung

bean buffer was added, mixed and collected by centrifugation.

4. The sample tube was placed in the 'cup-horn' of the sonicator containing ice cold water 1 mm from the face of the probe.
5. An output of approximately 12% on the 400 watt Virsonic 300 sonicator was used for 10 seconds in order to produce fragments of the required length.
6. 1 µl of sonicated DNA was mixed with 4 µl of loading buffer and the sample was run alongside lambda *Hind* III/pBR322 markers on a 0.8% agarose gel with 1 x TBE.
7. If sonication was successful the DNA was visible as a smear with no sign of a band of high molecular weight DNA. If a band was visible the samples were sonicated for a further 5 seconds and checked again on a 0.8% / 1 x TBE agarose gel.
8. The ends of the sonicated DNA fragments were repaired by adding 0.3 µl of mung bean nuclease buffer to the DNA. This mixture was placed in a 30°C water bath for 10 minutes.
9. The volume in the tube was made up to 200 µl with H_20, 20 µl of 1 M sodium chloride, 550 µl of ice cold 100% ethanol and 1 µl of pellet paint (Novagen) were added to the DNA.
10. In order to precipitate the DNA, it was left for 2-18 hours at -20°C and then centrifuged for 30 minutes at 4°C at 13,000 rpm.
11. The supernatant was removed from the tube and the DNA pellet was washed in 1 ml 100% ethanol by centrifugation for 10 minutes at 4°C at 13,000 rpm.
12. The ethanol was removed and the pellet was dried in a vacuum dryer for 10-15 minutes.

2.6.2 Selection of suitably sized DNA fragments for subcloning

1. The pellet was thoroughly resuspended for loading in 6.25 µl $T_{0.1}E$, 0.75 µl 10 x TAE and 2 µl loading dye.
2. All 9 µl of sample was loaded on a 0.8% agarose/1 x TAE gel with a lambda *Hind* III/pBR322 marker for 2 hours at 35 mA, 50-60 v.
3. The bands were visualised on an ultra violet transilluminator (312 nm) and the bands corresponding to 1.4-2 Kb (ideal) size were cut out. Additional bands of 1-1.4 Kb and 2-4 Kb were also cut from the gel and stored at 4°C. The pieces of gel were weighed to estimate the gel volume.
4. The 1.4-2 Kb gel fragment was placed in a tube and incubated at 65°C for 5-10 minutes.
5. 4 µl of AgarACE (Promega) was added to the tube in a 42°C waterbath. The molten gel was incubated at 42°C for 15 minutes.
6. 15 µl of sodium chloride, 150 µl of buffered phenol and the appropriate volume of $T_{0.1}E$ buffer corresponding to the weight of the gel piece was added to the tube to a final volume of 135 µl.
7. The tube was mixed by vortexing and centrifuged at 13,000 rpm.
8. The upper (aqueous) phase (approximately 230 µl) was extracted and added to the tube containing the 1.4-2 Kb gel fragment. 30 µl of $T_{0.1}E$ was added to the bottom layer, vortexed and centrifuged at 13,000 rpm for 3 minutes.
9. The upper (aqueous) phase was removed and pooled with the first layer removed.
10. 1 µl of pellet paint (Novagen) and 350 µl 100% ethanol were added to the tube which was placed at -20°C overnight.
11. The tube was centrifuged at 4°C at 13,000 rpm for 30 minutes and the ethanol

was discarded.

12. The pellet was resuspended in 1 ml of ethanol and spun at 4°C, 13,000 rpm for 10 minutes.
13. Ethanol was removed from the pellet which was vacuum dried for 5-10 minutes before resuspension in 5 µl of $T_{0.1}E$.
14. To check for successful elution, 0.5 µl DNA with 4.5 µl loading buffer was run on a 0.8%/ 1 x TBE agarose gel with lambda *Hind* III/pBR322 markers.

2.6.3 Ligation into pUC18 vector

1. A premix of pUC18 (SmaI/CIP, Amersham) and buffer (provided with vector), consisting of 0.05 µl of pUC18 per reaction and 0.1 µl of buffer (supplied with the pUC18) was prepared by vortexing and placing the tube on ice.
2. 0.15 µl of the pUC18-buffer mix was dispensed into the 600 µl Sarstedt tube set-up for each reaction.
3. 0.7 µl of DNA was added to each tube. In addition 3 control tubes were set-up with the following: (a) 0.7 µl ddH_20 (b) 0.7 µl ddH_20 and (c) 0.7 µl Φx174/HaeIII (1.4 ng).
4. 5 µl of mineral oil was added to each tube.
5. With the exception of tube (b), 0.15 µl T4 DNA ligase (Pharmacia) was dispensed to each tube, aiming for the 'bubble' under the oil, and the tubes were mixed and centrifuged for a few seconds.
6. Tubes were transferred to a 16°C incubator and left overnight to allow ligation to occur.

7. Tubes were heated to 65°C for 7 minutes before being left at room temperature for 5 minutes and centrifuged briefly.
8. 49 µl of ddH_20 was added to each reaction and tubes were stored at -20°C until transformations were performed.

2.6.4 Transformation of pUC18 vector

1. 1 µl of ligated DNA was aliquoted into 15 ml glass test-tubes and 500 µl of SOC was added to each 1 ml tube.
2. TG-1 cells (Invitrogen, maintained in 10% glycerol and stored at -70°C) were removed from the freezer and 150 µl 10% glycerol was added to each tube of cells which were left on ice.
3. Cells and glycerol were mixed using a P200 Gilson pipette and 40 µl of this mixture was added to the ligated DNA.
4. The cells, glycerol and DNA were aliquoted into a cuvette placed on ice, then electroporated using a Bio Rad Micropulser at 3.1 ms and 1.9 kv.
5. The cuvette was removed from the Micropulser and 400 µl SOC (pre-warmed to 20-30°C) was added to the cuvette: the mixture of SOC, cells and DNA was taken up and ejected into a test-tube.
6. The test-tubes were incubated in a shaker at 30°C for 1 hour with agitation.
7. TYE/Amp plates (90 mm) were placed at room temperature.
8. The test-tubes were removed from the shaker and 50 µl IPTG (40 mg/ml) 50 µl Xgal (50 mg/ml) were added to each tube.
9. 125 µl of the solution was dispensed into one TYE/Amp plate and 250 µl was dispensed onto a second plate.
10. A sterile spreader was used to make the solution cover the plate in an even

manner.

11. Plates were placed in a 37°C incubator overnight and the number of recombinant (white) and non-recombinant (blue) colonies counted.
12. Successful ligations were stored at -20°C.

2.7 Shotgun sequencing

The chromosome 9 clones, bA18B16 and bA544A12, were sequenced using a modified version of the method described by Sanger *et al* (1977b). Essentially, the DNA was sequenced using the dideoxy termination system in which DNA polymerase uses directed primers to extend a DNA strand from a single stranded template. Extension occurs with the addition of deoxynucleotides complementary to the template strand until the dideoxynucleotide that inhibits further extension is incorporated. The latter are labelled with fluorescent dyes and visualised when separated by gel electrophoresis. The biochemistry will produce populations of products specifically terminating at either A, G, C or T.

2.7.1 Vacuum preparation of template DNA in pUC18 vector

1. 1 ml of 2 x LB containing ampicillin was aliquoted into each well of a 96 well Beckman box, and separate (white recombinant) colonies were picked into each of these wells.
2. Boxes were sealed and the lids were pierced before boxes were placed in a 37°C incubator at 320 rpm and left to grow for 20-24 hours.
3. After growth, 100 µl of the cells were removed from each well and added to a 96 well plate (Corning) containing 50 µl 100% glycerol. The plates were

sealed and stored at -70°C.

4. Boxes were spun for 2 minutes at 4000 rpm, the supernatant was discarded and boxes were inverted on several layers of towels for 20 minutes to remove residual culture supernatant.
5. The pellets were resuspended in 120 µl GET using a plate shaker (Luckham V400 Vortexer) completely resuspended.
6. 120 µl NaOH/SDS solution was added, mixed thoroughly then incubated at room temperature for 2-5 minutes.
7. 120 µl 3 M KOAc (pH 5.5) was added and mixed gently.
8. A filter-bottomed plate (FB; Millipore catalogue number MAFBNOB50) was placed in the bottom of the vacuum manifold (Eppendorf). The lysate was removed from the Beckman box and dispensed into a Multiscreen-NA lysate clearing plate (NA; Millipore catalogue number MANANLY50) which was then placed on top of the manifold.
9. The lysate was drawn through the NA plate into the FB plate inside the manifold by applying the vacuum for 3 minutes not exceeding 8 Hg vacuum setting.
10. The NA plate was discarded and 150 µl of Bind Solution added to the FB plate and mixed.
11. The FB plate was placed on the empty manifold and full vacuum (30 Hg) was applied for 1 minute.
12. The plasmid DNA, bound to the FB plate, was washed with ice cold 80 % ethanol and vacuum filtered at full vacuum for 1 minute.
13. The plasmid DNA was washed again with ice cold 80% ethanol and vacuum filtered at full vacuum for 3 minutes.

14. The FB plate was removed from the vacuum manifold and dried thoroughly at 90°C for 10 minutes or 2 hours at room temperature.
15. 50 µl ddH_20 was added to the centre of each well and left to stand for 5 minutes at room temperature.
16. The plasmid DNA was eluted by placing the FB plate on top of a new microtiter plate (AB gene Thermo-fast® 96 well skirted plate; catalogue number AB-0800) and centrifuging for 2-5 minutes at 4000 rpm.
17. The plasmid DNA was checked on a 0.8 % agarose gel made up in 1 x TBE.

2.7.2 The sequencing reaction

1. 2 µl of DNA was added to 8 µl of a mix made up of 1 µl of forward primer (M13F-21F 5'-TGTAAAACGACGGCCAGT-3'; 6 pM/µl) or reverse primer (pUC18R 5'-GCGGATAACAATTTCACACAGGA-3'; 6 pM/µl), 4 µl BigDye™ Terminator Ready Reaction mix (supplied by PE Applied Biosystems) and 3 µl water.
2. The mixture was centrifuged and placed on a PTC-225 Peltier Thermocycler (MJ Research) with the following program: (i) 96°C for 30 seconds (ii) 50°C for 15 seconds (iii) 60°C for 2 minutes 30 seconds, (iv) repeat (i) – (iii) for 25 cycles (v) 4°C until stopped.
3. To each reaction, 10 µl ddH_20 and 50 µl precipitation mix was added.
4. The plate was centrifuged at 4°C, 4000 rpm for 25 minutes, and the ethanol was decanted.
5. 100 µl of ice-cold 70% ethanol was added, and the plate was centrifuged for 2-3 minutes at 4°C, 4000 rpm. This step was repeated.

6. The ethanol was removed and the plate inverted on a tissue and centrifuged at 250 rpm to remove all traces of ethanol. The plate was dried at 90°C for 10 minutes in the dark. Plates were stored at -20°C until loaded onto the sequencer.

2.7.3 Sequencing instrumentation

DNA sequenced by me was loaded on either an ABI PRISM® 373 DNA sequencer or an ABI PRISM® 377 DNA sequencer and generated by the Sanger Sequencing Centre on an ABI PRISM® 3100 DNA analyser (Applied Biosystems).

2.7.3.1 ABI PRISM® 373 DNA sequencer set-up:

1. The sequencing gel plate (see section 2.3 for preparation) was inserted into the ABI cassette of the ABI PRISM® 373 DNA sequencer and secured using clips; ensuring that the gel plates were flat in the cassette.
2. The plates were cleaned using a lint free tissue and the plates-checked by scanning the glass plates. If 4 flat coloured lines appeared in the scan window the upper buffer chamber was put in place and both upper and lower chambers were filled with 1 x TBE buffer before pre-running the machine for 30 minutes. If there were peaks in the trace the plate was removed from the cassette and cleaned before repeating the plate-checking process.
3. 3 µl of sequencer loading dye was added to each sequencing reaction, briefly centrifuged then denatured by heating at 80°C for 10 minutes before loading.
4. The comb was removed from the gel and wells were rinsed using 1x TBE to

ensure that there were no air bubbles before 3 µl of sample was loaded to each well (36 maximum) using a Gilson pipette.

5. Data was collected over a run-time of 8 hours.

2.7.3.2 ABI PRISM® 377 DNA sequencer set-up:

1. The sequencing gel plate (see section 2.3 for preparation) was inserted into the ABI cassette of the ABI PRISM® 377 DNA sequencer and secured using clips; ensuring that the gel plates were flat in the cassette.
2. The plates were cleaned using a lint free tissue and the plates-checked by scanning the glass plates. If 4 flat coloured lines appeared in the scan window the upper buffer chamber and heat plate that clipped in front of the gel plate were put in place. If there were peaks in the trace the plate was removed from the cassette and cleaned before repeating the plate-checking process.
3. The upper and lower buffer chambers were filled with 1 x TBE buffer before pre-running the machine for 30 minutes.
4. 2 µl of loading dye was added to each dried sequencing reaction and the samples were then briefly centrifuged.
5. The comb was removed from the gel and wells were rinsed using 1x TBE to ensure that there were no air bubbles before 2 µl of sample was loaded to each well (48-60) using a Gilson pipette.
6. Data was collected over a run-time of 4 hours.

2.7.4 Data analysis of shotgun sequencing reactions and clone assembly

The data produced from the ABI sequencers was transferred to the UNIX system

where a number of Sanger Institute in house software programs have been developed for the analysis of this data. The first procedure involved in analysing an ABI-PRISM® 373 or ABI-PRISM® 377 sequencing gel is to establish the position of each sample on a gel. This lane tracking is automatically performed by the program 'Gelminder' (Platt and Mullikin, unpublished) but manual checking and in some cases, repositioning is required. After manual checking of the lane tracking, the individual bases are called by 'Gelminder' using the program 'Phred'. The sequencing data produced by the capillary sequencer ABI-PRISM® 3100 is automatically uploaded into 'Capminder' and the bases are identified using the program 'Phred'.

Data from each sequencing reaction is then passed into the 'Automated Sequence Preprocessor (ASP)' program (Hodgson, unpublished) which cuts off sequence according to whether it is cloning or sequencing vector, *E.coli* contamination and sequence of an unacceptably poor quality. Clipped good quality sequences are then passed into the 'Phrap2Gap' program (Mott and Dear, unpublished), a modification of 'Phred' (a base-calling program) and 'Phrap' (a sequence assembly program; Gordon *et al*, 1998). 'Phrap2Gap' allows phrap-assembled reads to be transferred into the 'GAP' editing package. The 'GAP' sequence assembly program was developed as part of the Staden package (Bonfield *et al*, 1995; Staden, 1980; Staden *et al*, 2000); over the years versions have been updated from 'xGAP' to 'GAP' to 'GAP4' to 'GAP4.new'. Clones assembled as part of this project were largely assembled using 'GAP4.new' packages.

2.7.5 Contiguation or 'finishing' of a clone

Generally, a clone is not a contiguous piece of DNA sequence upon transfer into a

'GAP' package. The clones, bA544A12 and bA18B16, were not contiguous and a number of steps were required in order to produce a 'finished' clone (defined as a contiguous piece of sequence with both cloning vector arms present). A 'finished' clone also required that all the sequence was 'double stranded', which refers to the idea that the entire clone should be covered by at least two individual reads. Assembling a clone, therefore, required the use of a number of pieces of software, resequencing certain subclones and generating specific segments of DNA using the PCR reaction with the addition of reaction additives (section 2.7.5.1).

After a clone was contiguous and double stranded, the virtual restriction digest of the clone was checked against fragments generated by 3 actual restriction digests. This involved generating the real digests (described in section 2.4) and generating the virtual digests. Virtual digests were generated by the program 'Confirm' (Production Software Group, Sanger Institute, unpublished) which also has a graphical display showing the real and virtual digests alongside each other.

2.7.5.1 'Finishing' PCR reaction

The additives A, E and F (Invitrogen) are used to sequence 'difficult' regions. Additive A is the 'universal additive' and is designed to generally aid the sequencing reactions on problematic areas. Additive E is used to sequence regions with high GA composition and additive F for high AT composition. The dGTP BigDye™ terminator mix (Applied Biosciences) is used for regions of high GC content.

1. 2 µl (40 nM) of forward primer and 2 µl (40 nM) reverse primer were dispensed into a 96 well plate (Costar), spun briefly and dried down in a 90°C

oven for 10 minutes.

2. To 3 µl DNA template, 4µl BigDye™ Terminator mix (Applied Biosystems) or dGTP BigDye™ Terminator Ready Reaction mix (Applied Biosystems), 2 µl additive A, E or F (Invitrogen) and 4 µl ddH_20 was added.
3. Samples were placed on the PTD-225 Peltier Thermocycler (MJ Research) with the following program: (i) 95°C for 15 seconds, (ii) 45°C for 5 seconds, (iii) 60°C for 2 minutes, (iv) steps (i)-(iii) repeated 25 times, (v) 4°C until program stopped.
4. Sequencing protocols were performed as described in section 2.7.2, using the appropriate primer(s).

Expression profile analysis

2.8 Design of paralogue specific primers

Primers were designed manually and using the Primer3 program (Rozen and Skaletsky, 2000) for each paralogue[1] to ensure that the primers and product were 'paralogue specific'[2]. Primer sequences, 18-25 bp in length with an average GC-content of 40-60% and melting temperature of 55°C-65°C, were designed, in most cases, in the 3' UTR of the paralogue mRNA to generate a PCR product between 250-500 bp. The primer sequences used are given in Appendix 3 and 4.

Sequences were chosen:

(i) to avoid areas of simple sequence showing non-representative use of the bases and obvious repetitive sequence i.e. runs of single nucleotides (e.g. TTTT) or double nucleotides (e.g. CGCGCG) motifs.

(ii) to avoid complementarity between primer pairs as this would result in primers annealing to each other and forming primer dimers.

(iii) to exclude palindromes which will form inhibitory secondary structure (e.g. GACGTC)

Each primer was also designed with the universal 5' adaptor sequence '5'-TGACCATG-3'' necessary for attaching the paralogue specific amplicon to the

[1] Paralogues (or paralogous genes) are genes found within the same species which have arisen by duplication of a common ancestral gene.

[2] 'Paralogue specific' indicates that the PCR primers and product have been designed to be specific to a particular paralogue and not to cross-hybridise with other members of the same paralogous gene family which might share high sequence homology.

surface of the microarray (as per section 2.13.2).

The specificity of each primer and product was determined by BLAST searching the sequence against the ENSEMBL human genome build (UCSC (ENSEMBL 1.1.0). Each PCR product was also verified by sequencing (section 2.7 using appropriate primer).

2.9 PCR amplification of paralogue specific PCR products

The primers used to amplify the paralogue specific PCR products for the Southern, Northern and dot-blot experiments are summarised in Appendix 3 and the primers used in the RT-PCR and microarray experiments are summarised in Appendix 4.

1. To 5 µl human genomic DNA (1µg/µl), 2 µl 10x PCR Buffer, 10mM dNTPs, 0.5 µl of primer 1 (200ng/µl), 0.5 µl primer 2 (200ng/µl), 0.125 µl Taq Polymerase and 10.875 µl of ddH_20 was added.
2. The paralogue-specific PCR products were amplified using a PTD-225 Peltier Thermocycler (MJ Research) with the following program: (i) 95°C for 5 minutes, (ii) 95°C for 1 minute (iii) [annealing temperature]°C for 1 minute 30 seconds (iv) 72°C for 1 minute 30 seconds (v) repeat (ii)-(iv) 35 times (vi) 72°C for 5 minutes.
3. 5 µl of PCR product with 10 µl loading buffer were separated on a 2.5% agarose gel made up with 1x TBE and visualised with ethidium bromide.

2.10 Total RNA extraction from mammalian cell-lines

The 5 cell-lines growth medium (with serum and antibiotics): RPMI 1640 Medium

(GIBCO) supplemented with 10% fetal calf serum (FCS, GIBCO) and 5 ml penicillin/streptomycin (10,000U/ml; GIBCO BRL). Stored at 4°C.

1. The cell-line liquid nitrogen stocks were first thawed then washed by adding 25 ml of the tissue-culture medium and gently mixed.
2. The cell pellet was collected by centrifugation at 1500 rpm for 5 minutes and the supernatant discarded.
3. The cell pellet was resuspended in 15 ml of growth medium and grown in suspension at 37°C in 5% $C0_2$ / 95% air in 75 cm^2 filter capped flasks.
4. A cell culture with 70-80% confluence (~10^7 cells) was taken and centrifuged to pellet the cells at 1500 rpm
5. The medium was removed with a pasteur pipette and the cells washed twice with 50 ml PBS and the cell pellet collected by centrifugation at 1000rpm for 5 minutes.
6. 1ml TRIZOL reagent (GIBCO BRL) was added to each pellet and mixed well by pipetting until the pellet was completely resuspended (TRIZOL is a clear red liquid which will become cloudy once pellet is resuspended). An additional 1 ml TRIZOL was added if pellet did not resuspend completely.
7. The samples were dispensed into 1 ml aliquots in 2 ml round-bottom tubes and incubated at 60°C (heating block) for 10mins to fully resuspend the pellet.
8. 200µl chloroform was added to each 1 ml aliquot and mixed vigorously by shaking for ~15 seconds then incubated at room temperature for 2-3mins.
9. The samples were centrifuged at 14,000rpm for 15mins at 4°C.
10. The aqueous upper phase (clear, colourless) was dispensed into a new 2ml tube without disturbing the other layers and the remaining layers were discarded.

11. 0.5ml isopropanol was added to the aqueous layer and mixed by inversion then incubated at room temperature for 10mins.
12. The RNA pellet was collected by centrifugation at 14,000rpm for 15mins at 4°C (the RNA was visible as a white pellet at the bottom of the tube).
13. The supernatant was removed and discarded and the pellet washed once with 1 ml 75% ethanol. Vortexed to mix and centrifuged at 7,500rpm for 5mins at room temperature
14. The supernatant was removed and the pellet was centrifuged at 7,500 rpm for 2 minutes and the remaining supernatant carefully removed.
15. The pellet was air dried for ~30mins.
16. The pellet was re-suspended in 100µl DEPC and incubated at 60°C (heating block) to ensure the pellet was completely resuspended.
17. The total RNA was quantitated using a spectrophotometer and qualitated by assessing 2µg RNA by electrophoresis on a 1% agarose gel made with DEPC/1 x TBE (not exceeding 80mA as RNA will smear).
18. 3x volume 75% ethanol was added to the RNA sample. Stored at -70°C.

2.11 DNase treatment of RNA

Prior to use the total cell-line RNA was treated with DNase to remove any DNA contamination using the Ambion DNA-free™ kit.

1. 1 µg of RNA was incubated with 0.1 volumes 10 x DNase I Buffer (provided with kit) and 1 µl DNase I (2 units) at 37°C for 30 minutes.
2. 5 µl of DNase Inactivation Reagent (provided with kit) was added to the reaction mix and incubated at room temperature for 2 minutes.

3. The DNase Inactivation Reagent pellet was collected by centrifugation at 13,000rpm for 1 minute and the supernatant containing the DNA free RNA removed into a fresh tube.

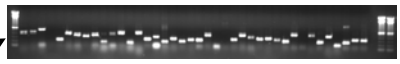
2.12 First strand cDNA synthesis and amplification of target cDNA using paralogue specific primers

The cDNA was synthesised using Superscript™ First-Strand Synthesis System for RT-PCR (Invitrogen). All reagents were provided with the kit.

1. To 1 µg DNA free total RNA 10mM dNTP mix, 1 µl Oligo$(dT)_{12\text{-}18}$ (0.5 µg/µl) was added and made-up to 10 µl with DEPC-treated water.
2. The reaction mix was incubated at 65°C for 5 minutes, then snap chilled on ice for 1 minute.
3. To the reaction mix, 2 µl 10x RT buffer, 25 mM $MgCl_2$, 0.1 M DTT and 1 µl RNaseOUT™ Recombinant RNase Inhibitor was added then incubated at 42°C for 2 minutes.
4. 1 µl of SuperScript™ II RT (50 units) was added to the reaction, mixed and incubated at 42°C for 50 minutes.
5. The reaction was terminated by incubating at 70°C for 15 minutes then chilled on ice.
6. The reaction was collected by centrifugation and 1 µl RNase H was added. The reaction was incubated at 37°C for 20 minutes.
7. Amplification of target cDNA was carried out according to section 2.9 substituting 5 µl genomic DNA with 2µl first-strand synthesised cDNA and increasing ddH_20 from 10.375µl to 13.375µl.

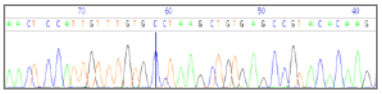
2.13 Overview of microarray experiments

Design of paralogue specific primers



PCR amplification of paralogue specific PCR products

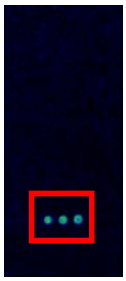
Sequence verified

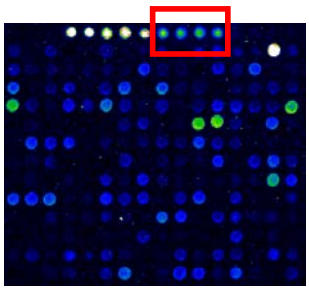


Generation of paralogue specific PCR products with a universal adaptor for use on Microarrays

Manufacture of Microarrays by Sanger Institute Microarray Facility

Paralogue Microarray	10K/Paralogue Microarray
Cross-hybridisation control experiments	**Expression experiments**
Fluorescent labelling of PCR product (2.13.3.1)	Fluorescent labelling of RNA (2.13.3.2)
Hybridisation and scanning	Hybridisation and scanning
Spot identification and analysis	Spot identification and analysis





2.13.1 Description of microarrays used

Two different microarrays were produced by the Sanger Institute Microarray Facility;

(i) The 'Paralogue Microarray' has 40 paralogue-specific PCR products arrayed in triplicate. This microarray was used to ensure that the amplified PCR products do not cross-hybridise with any of the other paralogue-specific PCR products.

(ii) The '10K/Paralogue Microarray' is a modification of the Sanger Institute human 10K microarray (Hver1.2.1). Further information can be found at http://www.sanger.ac.uk/Projects/Microarrays. The 10K array consists of 12 x 4 super-arrays corresponding to 48 sub-arrays each containing 224 DNA elements arranged as 14 rows and 16 columns (figure 2.1). There are currently 9932 DNA elements corresponding to human genes on the 10K microarray. The first row of each sub-array is generally reserved for duplicate sets of positive and negative controls consisting of. Cy3 positive control (a spot of Cy3), a negative control (empty spot) and 5 bacterial controls representing *Bacillus subtilis trp*, *lysA*, *thrB*, *dapB* and *pheB* genes. The duplicate set of controls has been substituted on the 10K/Paralogue Microarray by 2 of the 40 paralogue-specific PCR products arrayed in quadruplicate. The quadruplicate sets of the 40 paralogues are represented in 2 different locations on the microarray; once in super-array rows 1-6 and again in super-array rows 7-12. Therefore, each paralogue specific PCR product appears 8 times on the microarray. The 10K/Paralogue Microarrays were used to determine the expression profiles of the paralogues and the standard 10K DNA elements in

10 human RNAs.

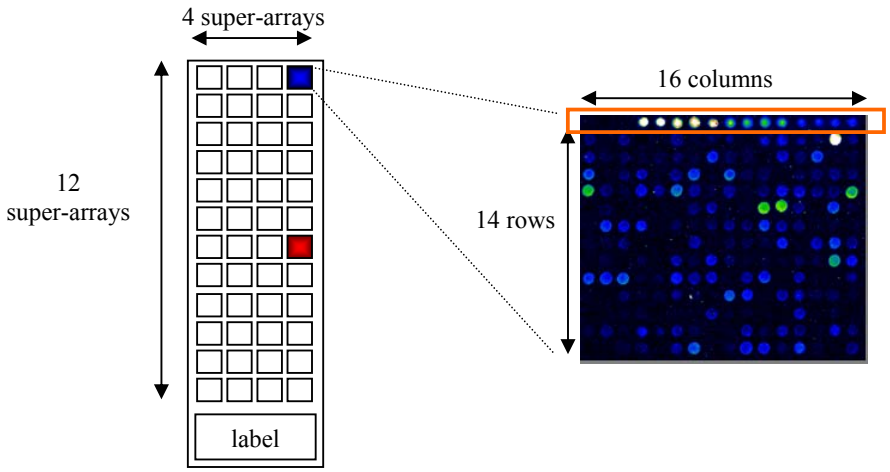


Figure 2.1 The 10K/Paralogue Microarray. The layout of the 48 sub-arrays in 12 x 4 super-arrays is shown and the sub-array coloured blue is expanded. The first row of the sub-array is boxed in orange. Columns 1 to 8 of row 1 contain the controls described in the text. The paralogue specific PCR products of one paralogue are arrayed in rows 9 to 12 (shown as 4 green spots) and a second paralogue in rows 13 to 16 (shown as 4 blue spots). The paralogue specific PCR products are represented in two different locations on the microarray and the super-array containing the same paralogue specific PCR products is coloured red.

2.13.2 Generation of paralogue specific PCR products with a Universal Adaptor for use on microarrays

The paralogue specific PCR products generated as described in section 2.9 were subjected to a second round of PCR in which a universal primer (5' GCTGAACAGCTATGACCATG-3') was used to attach an aminolinker to the 5' of the PCR product. The aminolinker enables the attachment of the PCR product to the microarray surface.

1. The paralogue specific PCR products were amplified according to section 2.9. The bands corresponding to the required PCR products were excised from the gel and transferred into 1 ml $T_{0.1}E$ and placed at 4°C for 18 hours, then stored

at -20°C.

2. To 15µl of paralogue specific PCR product in $T_{0.1}E$, 6µl 10x PCR buffer, 3µl 10mM dNTPs, 1.5µl universal primer, 1.5µl paralogue specific reverse primer, 0.375µl Taq polymerase and 32.625µl ddH_20 was added.
3. The thermocycler program was as follows (i) 95°C for 5 minutes, (ii) 95°C for 1 minute (iii) [Annealing temperature]°C for 1 minute 30 seconds (iv) 72°C for 1 minute 30 seconds (v) repeat (ii)-(iv) 35 times (vi) 72°C for 5 minutes.
4. 2µl PCR was analysed by electrophoresis on a 2.5% agarose/1 x TBE gel.
5. 15µl spotting buffer was added to each PCR product and this was arrayed onto the Microarrays.

2.13.3 Generation of fluorescently labelled DNA

2.13.3.1 Generation of fluorescently labelled paralogue-specific PCR products using the Cyanine 3-dCTP dye for hybridisation onto the 'Paralogue Microarray'

In order to ensure the specific PCR products do not cross-hybridise to the other paralogues the paralogue-specific PCR products generated in section 2.9 were labelled with a fluorescent dye and hybridised to the Paralogue Microarray.

1. To 5µl $T_{0.1}E$ DNA stocks of the paralogue specific PCR products 2µl 10x PCR buffer, 1µl 10 mM dA,T,GTP/5 mM dCTP mix, 0.5µl primer 1, 0.5µl primer 2, 0.125µl Taq polymerase, 2µl dCTP-Cy3 and 8.875µl ddH_20.
2. The thermocycler program was as follows (i) 95°C for 5 minutes, (ii) 95°C for

1 minute (iii) [Annealing temperature]°C for 1 minute 30 seconds (iv) 72°C for 1 minute 30 seconds (v) repeat (ii)-(iv) 35 times (vi) 72°C for 5 minutes.

3. Excess nucleotides were removed from the PCR reaction using QIAquick Nucleotide Removal Kit (Qiagen) according to the manufacturers' instructions and eluted with 40µl ddH_20.
4. 5µl of product with 5µl loading buffer were analysed on a 1% agarose 0.5 x TBE gel.
5. Depending on how successful the labelling reactions was, between 3-10µl of fluorescently labelled paralogue specific PCR product was denatured at 100°C for 5 minutes then snap chilled on ice and mixed with 38µl hybridisation buffer.

2.13.3.2 Generation of fluorescently labelled single-stranded cDNA target using direct incorporation of Cyanine dyes for hybridisation onto the '10K/Paralogue Microarray'

The Bacterial mRNA "cocktail" was provided by Sanger Institute Microarray Facility.

1. 1µl of bacterial "cocktail" (1 x stock in 75% ethanol) was added to 40µg of total RNA (in 75% ethanol) and precipitated by adding 1/40th volume of 3M sodium acetate at -70°C for 30 minutes.
2. The RNA pellet was collected by centrifugation at 13,000rpm and washed briefly in 100µl 70% ethanol and air-dried for 30 minutes.
3. The RNA pellet was resuspended in 12.9µl DEPC and 2.5µl anchored oligo-dT (2µg/µl final concentration; mixture of $T_{17}A$, $T_{17}G$ and $T_{17}C$ primers).
4. The RNA/oligo mixture was heated to 70°C for 10 minutes and then snap

chilled on ice.

5. To 15.4µl RNA/oligo mixture, 6µl 5x first strand buffer (Invitrogen), 3µl 0.1M DTT (Invitrogen), 0.6µl 10 mM dA,T,GTP/5 mM dCTP mix dNTPs, 3µl dCTP-Cy3 or dCTP-Cy5 and 2µl Superscript II (Invitrogen) was added.
6. The reaction was incubated at 42°C for 2 hours.
7. 1.5µl 1M NaOH was added to the reaction and incubated at 70°C for 20 minutes to hydrolyse the RNA.
8. 1.5µl 1M HCl was added to neutralise the reaction.
9. The nucleotides and short oligomers were removed using the Autoseq G-50 columns (Amersham Biosciences) according to the manufacturers' instructions resulting in ~33µl of labelled cDNA sample.
10. 33µl of test cDNA sample was combined with 33µl of control cDNA and 4µl polyA DNA (Sigma), 8µl C_ot1 DNA (Gibco BRL) and precipitated with 7.8µl 3M sodium acetate pH5.2 and 260µl 100% ethanol at -70°C for 25 minutes.
11. The pellet was collected by centrifugation at 13,000 rpm and washed briefly in 70% ethanol. All traces of ethanol were carefully removed and the pellet air-dried.
12. The pellet was resuspended in 40µl microarray hybridisation buffer and 8µl dd$H_2$0.

2.13.4 Hybridisation, washing and scanning of microarrays

1. 46µl of hybridisation mixture was spotted onto the microarray cover slip (25 x 60 mm) and the microarray was inverted and lowered onto it.
2. The microarray was placed in a humid chamber (2 cm x 7 cm 3MM paper

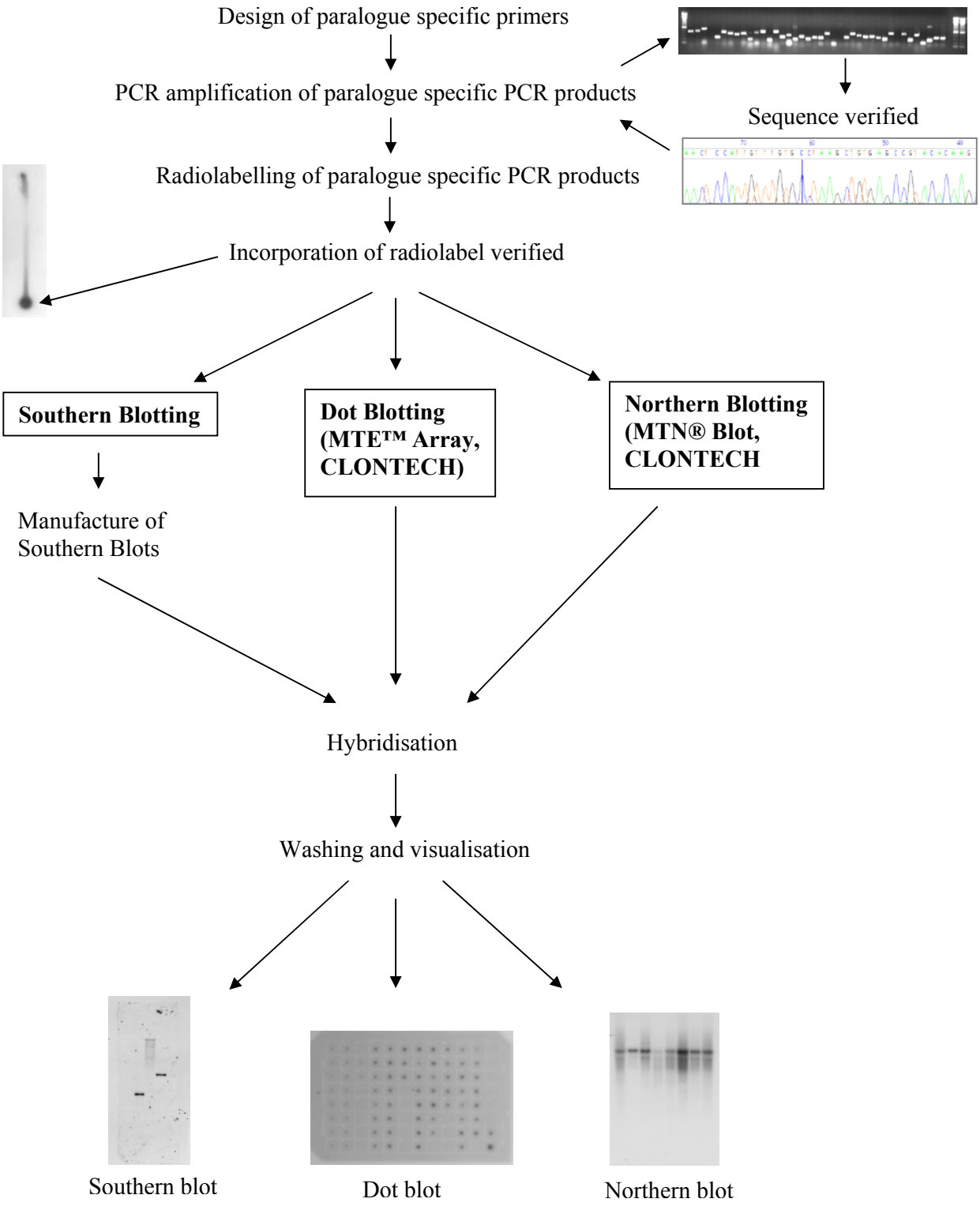
moistened with 2 ml 40% formamide, 2 x SSC in a Petri dish) and incubated for 12-24 hours at 47°C.

3. The cover slip was carefully removed from the microarray by rinsing in microarray wash solution 1 for 10-15 seconds.
4. The microarray was first washed in microarray wash solution 1 for 5 minutes at room temperature with gentle shaking. Followed by 2 washes in microarray wash solution 2 for 30 minutes at room temperature with vigorous shaking and, finally, in microarray wash solution 3 for 5 minutes with vigorous shaking at room temperature.
5. The microarray was dried by centrifugation at 1000 rpm for 1-2 minutes.
6. Using a laser-based scanner (GSI Lumonics ScanArray® 5000) the microarray was scanned at the two wavelengths compatible with efficient excitation for Cy3 and Cy5 (550nm and 650nm respectively) at 10 μm scanning resolution.

2.13.5 Analysis of microarrays

GSI Lumonics Quantarray® microarray analysis application software was used to determine the fluorescence intensity of spots in microarray images produced by ScanArray®. A three stage protocol was observed: (i) spot finding, (ii) spot quantitation; (iii) data export and visualisation. Once the spots have been identified and quantitated the standard deviation between the spot intensity and background intensity was calculated. In most cases, if a spot was present the standard deviation was greater than 2. To verify these results each spot was also assessed by-eye for each experiment using Quantarray®. The microarray data was clustered using the program EPCLUST at EMBL-EBI as described in section 2.18.

2.14 Overview of blot expression analysis



2.14.1 Radioactive labelling of DNA

2.14.1.1 Radioactive labelling of paralogue-specific PCR products

1. In a 0.5µl microcentrifuge tube, 2µl 10x PCR buffer, 1µl 10mM dNTPs mix, 0.5µl primer 1, 0.5µl primer 2, 0.125µl Taq polymerase, 4µl [α-^{32} P]-dCTP and 6.875µl dd$H_2$0 was added to 5µl of the $T_{0.1}$E DNA stocks of the paralogue specific PCR products generated as described in section 2.9.
2. The reaction was overlaid with mineral oil to prevent evaporation and subjected to PCR in a DNA thermal cycler (Perkin Elmer, USA). PCR cycling conditions were as follows (i) 95°C for 5 minutes, (ii) 95°C for 1 minute (iii) [Annealing temperature]°C for 1 minute 30 seconds (iv) 72°C for 1 minute 30 seconds (v) repeat (ii)-(iv) 35 times (vi) 72°C for 5 minutes.
3. Excess nucleotides were removed using QIAquick Nucleotide Removal Kit (QIAGEN) according to manufacturers' instructions and the labelled PCR product was eluted in 50µl dd$H_2$0.

2.14.1.2 Radioactive labelling of DNA using MegaPrime™ DNA labelling system (Amersham)

1. To 25 ng DNA template, 5 µl of primers was added and the final volume made up to 50 µl with dd$H_2$0.
2. The reaction mix was denatured at 95 °C for 5 minutes.
3. The reaction was collected by centrifugation at 13,000 rpm.
4. 10 µl labelling buffer, 5 µl [α-^{32} P]-dCTP and 2 µl enzyme was added to the reaction mix. Mixed then centrifuged.

5. The reaction was incubated at 37°C for 1 hour and the reaction stopped by the addition of 5 µl 0.2 M EDTA.

2.14.2 Probe verification

2.14.2.1 Assessment of radiolabel incorporation using thin-layer chromatography

1. 1µl of PCR product was spotted onto a 5 x 10 cm Polygram CEL 300 PEI/UV thin-layer chromatography sheet (Macherey-Nagel, GmbH & Co) approximately 1.5 cm from the bottom edge.
2. This was placed in a beaker containing 0.75 M KH_2PO_4 (pH3.5) 1cm in depth and left for 30 minutes.
3. The chromatogram was subjected to autoradiography for 30 minutes.
4. Incorporated isotope remains at the spotting position, whereas unincorporated migrates with the buffer front.

2.14.2.2 Measurement of radioactively labelled PCR product concentration

The optimal concentration of radioactively labelled PCR product (or probe) is 1-2 x 10^7 cpm/ml. This was calculated using either a mini Geiger counter or a Scintillation counter (Easicount 4000, Scotlab, UK).

2.14.3 Manufacture of Southern Blots

2.14.3.1 Restriction digest of human genomic DNA

1. To 10µg human genomic DNA, 1mM Spermidine, 1mM DTT, 1.5µl restriction enzyme (either *Pst*I, *EcoR*I and *Hind*III), 5µl appropriate NEBuffer were added and the total reaction volume made up to 50µl with $T_{0.1}E$.
2. The reaction was incubated at 37°C for several hours (time optimised for each enzyme). After 1 hour another 1µl enzyme was added.
3. 3µl of digest was analysed on a 0.8% agarose 0.1 x TAE gel with a 100bp ladder. If digestion was not occurring after 24 hours more enzyme was added and the reaction incubated at 37°C until genomic DNA was completely digested and a further 3µl analysed by electrophoresis.
4. Once the genomic DNA had completely digested all 3 digests were loaded on a 0.8% agarose, 1 x TAE gel (2.4 g agarose, 300 ml 1 x TAE, 10 µl Ethidium Bromide) with a lambda *Hind* III ladder (100ng) and run at 50v for ~16 hours.

2.14.3.2 Transfer of digested genomic DNA onto filter

1. Excess gel was cut away and the gel was denatured in 1 x denaturation solution for 30 minutes with gentle shaking.
2. The gel was rinsed twice in ddH_20 then washed twice in 1 x neutralisation solution for 30 minutes each with gentle shaking.
3. Gels were blotted for 24 hours in 10 x SSC onto hybridisation transfer membrane (Hybond™-N ; Amersham) with frequent changing of towels.
4. The membranes were rinsed in 2 x SSC, dries on Whatman paper and the

DNA cross-linked on a UV transilluminator (320nm) for 2.5 minutes.

2.14.4 Hybridisation of radiolabelled PCR product to blots

1. The blots were prehybridised in ExpressHyb™ Hybridisation Solution (Clontech) containing 1.5 mg sheared salmon testes DNA (Stratagene) at 65°C for 1-4 hours.
2. In the case of the Human Multiple Tissue Northern (MTN®) blots (Clontech) and the Southern blots, the radiolabelled DNA were denatured at 95-100°C for 10 minutes then snap chilled on ice for 10 minutes before being added to appropriate volume fresh ExpressHyb™ solution.
3. In the case of the Human Multiple Tissue Expression (MTE™) Arrays (Clontech) 30µg of C_ot-1 DNA, 150µg of sheared salmon testes DNA (Stratagene) and 50µl 20 x SSC were added. The reaction volume was made up to 200µl with dd$H_2$0 then heated to 95-100°C for 5 minutes then incubated at 68°C for 30 minutes.
4. The pre-hybridising solution was discarded and replaced with fresh ExpressHyb™ solution containing 1.5 mg sheared salmon testes DNA and the denatured radiolabelled PCR product. All the blots were hybridised at 65°C for 16-18 hours.

2.14.5 Washing

1. The hybridisation solution was discarded and replaced with the appropriate wash solution I. The blots were rinsed 5 times with wash solution I before being washed in fresh wash solution I for 30-40 minutes (the wash solution

was replaced several times) at room temperature for MTN and Southern blots and 65°C for MTE Arrays.

2. Wash solution I was discarded and the MTN and Southern blots washed at room temperature in wash solution II and MTE array washed at 65°C until the background signal was significantly reduced and activity detected with a Geiger counter more specific (~5 cpm).
3. Excess liquid was removed and from the filters by laying them briefly onto Whatman 3MM paper. The filters were then subjected to autoradiography using pre-flashed film and intensifier screens at -70°C for 24 hours, 3 days in all cases, and longer if necessary.

2.15 Computational analysis

A multitude of bioinformatics programs were used in this thesis in order to identify and characterise genes; both in the annotation of genomic clones and the identification and characterisation of MHC paralogues. The individual tools are discussed in section 2.15.1 and the methods in which they were used for a particular analysis are discussed in later sections.

2.15.1 General programs used in this thesis

'BLAST' is an acronym for the basic alignment search tool (Altschul *et al*, 1990). The program has become widely used in DNA and protein database searches. It is based on measuring local similarity between sequences, calculated by the maximal segment pair (MSP) score. There are several types of BLAST searches available for both nucleotide and protein sequences. In general, 'tblastn' was used to search the protein sequence against the selected nucleotide database translated in all six reading frames. In addition, PSI-BLAST (Position Specific Iterated BLAST) was used in the identification of paralogues. This uses an iterative search in which sequences found in one round of searching are used to build a score model for the next round of searching thus identifying paralogues sharing weak sequence homology. The databases searched using BLAST were either stored at EMBL-EBI or the NCBI.

'NIX' is a tool at the HGMP used to view the results of running many DNA analysis programs on a DNA sequence. In the initial step the sequence is masked for repeats using 'RepeatMasker' (Smit and Green, unpublished). This program screens against a library of interspersed repeats and low complexity DNA sequence called 'Repbase'

(Jurka, 2000). BLAST searches are started using the masked sequence against a number of databases, including Swissprot, TrEMBL, EMBL, EST, HTG, Unigene, Ecoli and Vector. The DNA sequence is also run through a number of gene finding programs, including 'Grail', 'Genefinder', 'Hexon' and 'Fgene'. The results from so many different programs are presented in a graphical interface and the features in the DNA sequence identified. By viewing all the results side-by-side it makes it easier to see when many programs have a consensus about a feature.

'Electronic PCR' (e-PCR) is a tool used to identify molecular markers, such as STSs, in a query sequence. In order to determine the true locations of known genes on chromosome 9, e-PCR was performed using the cDNA sequences of the genes as the query sequences. The STSs matching the cDNA sequence of the gene were identified and used to determine which genomic clone the genes were located within the chromosome 9 database, '9ace'.

The 'ENSEMBL' genome browser was used extensively throughout this project. The 'ENSEMBL' genome browser is a joint project between EMBL-EBI and the Sanger Institute and provides a bioinformatics framework to organise biology around the sequences of large genomes (Hubbard *et al*, 2002). ENSEMBL provides a fully annotated human genome incorporating the data from existing biological databases and as *ab initio* gene predictions. Other genome browsers were also used to view the genomic sequence and annotation; the UCSC genome browser based at UC Santa Cruz (Kent and Haussler, 2001; Kent *et al*, unpublished) and the NCBI Map Viewer. All three genome browsers are now built using the same reference sequence constructed directly from the physical maps representing the minimum clone tiling path being finished by the genome centres. All analyses for this thesis were performed

using NCBI31 genome data freeze (November 2002).

EMBOSS (Rice *et al*, 2000) is the 'European Molecular Biology Open Software Suite' which provides a comprehensive suite of sequence analysis programs (~100). Programs such as 'water' and 'needle' were used to generate local and global sequence alignments, respectively. Several other programs were used to manipulate both DNA and protein sequences.

2.16 Identification of extended MHC paralogous genes in the human genome

Chapter 4 summarises the results of a search to identify the extended MHC genes in the human genome. Several methods and programs have been developed to facilitate the identification of the paralogues but only the final method used to generate the results presented in chapter 4 is described here. The paralogues were identified with increasing levels of confidence using a number of criteria; the method used is described in sections 2.16.1 and 2.16.2.

2.16.1 Identification of extended MHC paralogues based on protein sequence homology

The extended MHC protein sequences were BLAST searched against the ENSEMBL human genome build NCBI31 using the 'tblastn' executable. The BLAST program used in this analysis was Washington University BLAST version 2.0 (WU-BLAST2.0) which is capable of detecting relationships between proteins with low

sequence identities (Brenner *et al*, 1998). The BLAST search parameters have been optimised to identify the paralogues based on sequence homology and to eliminate false-positives and reduce background noise as much as possible without losing the sensitivity of the analysis. The 2 critical parameters optimised were; (i) the substitution matrix and (ii) the Expected (E) value.

i. The substitution matrix is a key element in evaluating the quality of an alignment and assigns a score for aligning any possible pair of residues. The BLOSUM62 (Henikoff and Henikoff, 1992) matrix was used in this analysis as it is one of the best for detecting weak sequence similarities of query length greater than 85 amino acids/nucleotides.
ii. The Expect value (E) is a parameter that describes the number of hits one can "expect" to see just by chance when searching a database of a particular size (Karlin and Altschul, 1990). Essentially, the E value describes the random background noise that exists for matches between sequences and the E value is used as a convenient way to create a significance threshold for reporting results. The E-value used in this analysis was 10. Using a set of protein sequences I found that by increasing the E value from 10 a larger list with more low-scoring hits was reported and no new paralogues were identified. By decreasing the E-value the number of hits was reduced and the low-scoring hits were practically eliminated but known paralogues were also not identified.

The resulting BLAST hits were then filtered according to the P-value and results with a P-value $\geq E^{-5}$ removed from the analysis. The P value is the probability of finding such a hit by chance. In short, small P values are considered to be good and very unlikely to be random, therefore meaningful, but P values become unreliable above

10^{-5} (Lesk, 2002). This filtering value was selected in order to maintain both sensitivity and specificity of the experiment.

2.16.2 Identification of extended MHC paralogues with increasing levels of confidence

The initial BLAST search identified 1000's of BLAST hits. Using knowledge of the protein sequence and the gene structure these BLAST hits were filtered to identify the extended MHC paralogues with the highest level of confidence.

2.16.2.1 Filter 1: Domain-masking

The protein domains were identified by searching the protein sequence against the 'PFAM' database of protein domain families using the perl script 'pfam_scan.pl' (written by the PFAM Software Group and kindly provided by K.Howe). The domains were masked using another perl script 'x_out_domains.pl' (written by K.Howe). The domain-masked protein sequences were then BLAST searched against the human genome as described in section 2.16.1 and the results sorted according to the P value. By masking the protein domains a large number of BLAST hits were identified that were just to a particular protein domain. More significantly, it also identified the extended MHC paralogues which still share good sequence homology outside the domains.

2.16.2.2 Filter 2: FINEX

Putative paralogues were initially identified by sequence homology using similarity searching to find relationships. However, genomic sequence data provides gene architecture information not used by conventional search methods. In particular, intron positions and phases are expected to be relatively conserved features, because mis-splicing and reading frame shifts should be selected against. 'FINEX' (Fingerprinting of INtron EXon boundaries) is an alignment technique which exploits the gene structure information provided by a genomic sequence (Brown *et al*, 1995). A single exon fingerprint can be compared rapidly against all the entries in a library of fingerprints which is generated using the CDS (coding sequence information) features in the annotated EMBL entry (EMBL release 73). The phases of the exon fingerprints are classified according to their position relative to the reading frame of the gene: introns lying between two codons (phase 0); introns interrupting a codon between the first and second base (phase 1); and, introns interrupting a codon between the second and third base (phase 2). These intron positions and phases are expected to be relatively conserved features, because mis-splicing and reading frame shifts should be selected against.

The FINEX database relies on coding sequence (CDS feature) information available in annotated EMBL entries for genomic clones. Only a small percentage of genomic clones are annotated therefore the FINEX database does not contain the fingerprint for every gene in the genome. As the MHC region is one of the best characterised regions and the majority of clones covering the region are annotated all the MHC gene fingerprints are present in the FINEX database compiled using EMBL release version 73. Therefore, the FINEX fingerprint was generated for all putative paralogues and

used to search the FINEX fingerprint database using the optimised parameters: weight = 0.5, power = 4.0 and gap penalty = 0.5.

A number of scores were generated for each alignment to add statistical significance. The important scores to consider when determining a cut-off threshold are the D_{avg} score, which is the alignment length normalised score used to rank the alignments, the D_{mat} score, which is the global alignment score that measures the quality of the alignment and the z-score, which is the significance of the D_{mat} score for a given query/hit. The best alignment attainable is with self and, by definition, the dissimilarity scores D_{mat} and D_{avg} are zero and has the highest z-score (Brown *et al*, 1995). With this in mind the putative paralogues were used to search the FINEX fingerprint database and it was generally observed that the highest z-score obtained corresponded to the MHC gene it is paralogous to, or to another paralogue. Analysis of the paralogues identified ambiguous matches with a z-score less than 3.00 therefore a z-score greater than 3.00 was taken as significant. This is in agreement with significant z-score values for conventional sequence comparison (Dayhoff *et al*, 1978; Feng *et al*, 1984).

2.17 *In-silico* expression analysis

The 'UNIGENE' database at the NCBI automatically clusters the 'GenBank' sequences into a non-redundant set of gene-oriented clusters. Each UniGene cluster contains sequences that represent a unique gene, as well as related information including the tissue types in which the gene has been expressed. This information, in the form of expressed sequence tags (EST) was downloaded into a text file <gene_name>_unigene.txt for each gene. In addition to ensure no ESTs had been

omitted the gene mRNA or cDNA was used to BLAST search the database 'dbEST'; a division of GenBank that contains sequence data and other information on 'single-pass' cDNA sequences.

The 'BLAST' results were then filtered to remove insignificant hits (generally <90% sequence identity) and the results saved in a text file <gene_name>_dbest.txt. The 2 text files were then parsed to produce a list of EST accession numbers using 'parse.pl' and the 2 lists were compared to identify 'unique' and 'not-unique' ESTs for a particular gene using the perl script 'check_unique.pl'. The 'not-unique' ESTs were removed from the list and the non-redundant EST lists for each paralogue were compared against each other and 'not-unique' ESTs removed resulting in a paralogue specific list of ESTs.

The DNA sequence for each paralogue specific EST was extracted using 'sequence_retrieval.pl' and aligned against the gene cDNA using the program 'b2b' (written by R.Horton) and any false positive ESTs removed from the analysis. A database containing the full database entry for each paralogue specific EST list was produced using 'db_extract.pl' and the tissue information extracted using 'tissue.pl'. As the annotation within EST database entries is not consistent several versions of 'tissue.pl' exist to extract the tissue information. Finally, the EST data was inputted into Excel and sorted by tissue and system. All perl scripts used in this section were written by K.Crum and modified by myself, unless stated otherwise.

2.18 Clustering methods

The data presented in this thesis was clustered using the unsupervised clustering methods, hierarchical clustering (clustering methods are reviewed by Brazma and Vilo, 2001) using using EPCLUST (Expression Profile Data CLUSTering and Analysis) at the EBI. Hierarchical clustering arranges the data in a tree-like structure (similar to phylogenetic trees), where genes with similar expression patterns occupy neighbouring 'leaves' of the tree. The algorithms used for hierarchical clustering are largely the same as used for distance-based phylogenetic reconstruction from sequence data, but restricted to those methods that are fast enough to deal with large numbers of nodes. Hierarchical clustering works by iteratively partitioning clusters starting with the complete set. After the joining of two clusters, the distances between all other clusters and a new joined cluster are recalculated. The complete linkage method used in this analysis uses the maximum distances between the members of two clusters to cluster the data.

Expression data was also clustered using the perl script ‘exprofile’ (written by R.Younger), or modifications of this script. The input file consists of a tab delimited table containing the expression data. The cells contain values 0 and 100 which correspond to whether the gene is expressed in the corresponding tissue or not, respectively. Alternatively, the cells may contain a value between 0 and 100, which is the percentage of genes expressed in a particular tissue. The output of the program is a postscript file containing the image corresponding to the input data, i.e. where a gene is expressed there is a black bar and when there is no expression it is white. The thickness of the black bar corresponds to the percentage of genes expressed in a particular tissue. This is described in more detail in chapter 6.

2.19 Phylogenetic analysis

The consensus phylogenetic trees presented in Chapter 5 were produced by merging the trees generated by 3 different software packages; PHYLIP, TREE-PUZZLE and MEGA2. The PHYLIP (Phylogeny Interface Program version 3.6; Felsenstein, 1989) package is a public domain package that provides a wide range of programs for constructing phylogenetic trees from molecular and other types of data. TREE-PUZZLE version 5.0 (Schmidt *et al*, 2002) is a computer program to reconstruct phylogenetic trees from molecular sequence data by maximum likelihood using the quartet-puzzling algorithm and MEGA2 (Molecular Evolutionary Genetics Analysis) software (Kumar *et al*, 1994; Kumar *et al*, 2001) is a software package for exploring and analysing aligned DNA or protein sequences from an evolutionary prospective and offers useful and easy-to-use methods of comparative sequence analysis.

2.19.1 Protein sequence alignments

Protein sequences were aligned using the 'ClustalW' program (Thompson *et al*, 1994). This is a progressive multiple sequence alignment method which, firstly, assigns individual weights to each sequence in a partial alignment in order to down-weight near-duplicate sequences and up-weight the most divergent ones. Secondly, it varies amino acid substitution matrices at different alignment stages according to the divergence of the sequences to be aligned. Thirdly, residue-specific gap penalties and locally reduced gap penalties in hydrophilic regions encourage new gaps in potential loop regions rather than regular secondary structure. After a gap has been opened, locally reduced gap penalties are applied to positions around this gap. The alignments

produces by 'ClustalW' were viewed in 'belvu' (Sonnhammer, unpublished) and edited using 'Jalview' (Clamp, unpublished).

2.19.2 Estimation of the gamma distribution

Evolutionary analysis of DNA and protein sequences is typically performed by either assuming that all evolutionary lineages evolve at the same rate or by avoiding any attempt to directly consider the fact that the rate of evolution changes over time. The default parameters for the 3 programs used assume that the rate of evolution is constant. However, there are several factors that affect the rate of molecular evolution (e.g., mutation, population size, selection) and therefore the rate of molecular evolution is extremely unlikely to be identical for different evolutionary lineages or individual amino acids or nucleotides. This was taken into account in this analysis and the rate of variation (or rate of heterogeneity) between sites was calculated using the gamma distribution. The shape of this distribution is determined by the value of a parameter known as the gamma distribution parameter alpha and was calculated using TREE-PUZZLE.

2.19.3 Bootstrapping and tree-puzzling steps

The aligned sequences were bootstrapped using the program SEQBOOT in PHYLIP and by selecting the bootstrapping option in MEGA2. Bootstrapping (Felsenstein, 1985) involves taking each site within a protein and rearranging sites to create a number of 'pseudoalignments'. These 'pseudoalignments' are then used to recreate a number of trees which are compared to the original tree. Groupings obtained in the

original tree are then given a percentage expressing how many times they are recreated in the 'pseudoalignment' trees.

The 'puzzling-step' parameter was selected in TREE-PUZZLE which is similar to bootstrapping and trees are composed into so-called intermediate trees. This step results in many intermediate trees (default 1000) and from these a majority rule consensus tree is built and the number of intermediate trees lending support for the consensus topology is displayed at each node. Bootstrap or puzzling-step values of over 50 % were considered to represent reliable groupings those below were considered to show little or no support. However, low values at branches are not considered worthless as every phylogenetic tree is the best tree obtainable using a specific method and sequences. Computer simulations have shown that the branching patterns of an inferred tree may be correct even if they are not supported by high bootstrap values (Nei and Kumar, 2000).

2.19.4 Phylogenetic analysis using distance methods

Trees were generated using the Neighbour-Joining method (Saitou and Nei, 1987). This method uses an algorithm to convert pairwise distances between sequences into a matrix, from which branching order and branch lengths are computed. The Jones, Taylor and Thornton, or JTT, (Jones *et al*, 1992) model of amino acid change was used. This model is very similar to another model, the PAM Dayhoff (Dayhoff *et al*, 1978) model, which provides a measure of probability calculating how likely the amino acid in one sequence is likely to change the amino acid in the other sequence. These probabilities were based on a subset of closely related proteins that were organised into a phylogenetic tree and the frequency of change from each amino acid

to another was determined by adding up the changes at each evolutionary step. The JTT model is based on a recounting of the number of observed changes in amino acids of a much larger set of proteins therefore this model is to be preferred over the original Dayhoff PAM model. Using this model the Neighbour-Joining method constructed trees from the matrices of the multiple data sets from bootstrapping by the successive clustering of lineages and the setting of branch lengths as the lineages join.

2.19.4.1 PHYLIP

The output of the SEQBOOT program in PHYLIP was used as the input into the distance program PROTDIST. The program corrects distances for unequal rates of change at different amino acid positions using the coefficient of variation (CV) which was calculated using the gamma distribution alpha parameter from TREE-PUZZLE. The square of the CV is the value of the alpha parameter. The PROTDIST output was used as an input file to the program NEIGHBOR. The consensus tree is produced by using the output of the NEIGHBOR program as the input of the CONSENSE program.

2.19.4.2 MEGA2

MEGA2 is an easy to use software package and phylogenetic trees are generated quickly and in one simple step. First, the protein alignments were converted into MEGA2 format (.meg file) within the software package then the trees were generated using the neighbour-joining method (Saitou and Nei, 1987) under the JTT (Jones *et al*, 1992) model with 1000 bootstraps. Two additional parameters were selected; the

pairwise deletion comparison option and the Gamma distance option. The former removes sites containing missing data or alignment gaps from the analysis as they arise. This is in contrast to the complete-deletion option which removes all such sites prior to analysis. Both options were initially used but no significant difference was observed. The Gamma distance was used to take care of the inequality of the substitution rates among sites and the gamma shape parameter, or alpha parameter, calculated using TREE-PUZZLE was used in this analysis.

2.19.5 Phylogenetic analysis using the maximum likelihood method

The maximum likelihood (ML) method allows the inference of evolutionary trees from nucleotide or amino acid sequences under a probabilistic model of nucleotide/amino acid evolution (Felsenstein, 1981). The ML method looks for all possible tree topologies between the sequences by initially constructing an unrooted tree using three sequences then the 4^{th} is added to the tree and the 'best' tree topology for the four sequences chosen under likelihood criterion. This is repeated for the 5^{th}, 6^{th}, 7^{th} etc sequences until the final tree is produced. The log likelihood value is calculated and the best tree is the one with the most positive log likelihood value.

2.19.5.1 PHYLIP

The PHYLIP program PROML implements the maximum likelihood method for protein amino acid sequences using the JTT model of changes between amino acids. The model assumes that each position and each lineage have evolved independently and the different rates of evolution were determined using the Gamma distribution. As

previously the gamma distribution alpha parameter was calculated using TREE-PUZZLE and the Coefficient of Variation used as input into PROML. PROML is CPU intensive and, for this reason, data analysed was not bootstrapped.

2.19.5.2 TREE-PUZZLE

TREE-PUZZLE (TREE-PUZZLE version 5.0) constructs phylogenetic trees using maximum likelihood by implementing the fast tree search algorithm, quartet-puzzling. The protein alignment was used as input into the program and the 1000 'puzzling-step' option selected. Trees were generated using the JTT model and the rate of heterogeneity was set as Gamma distance. The 'outfile' generated contained information regarding the calculation of the gamma distribution alpha parameter and the consensus tree.

2.20 Useful web-sites

BLAST	http://www.ncbi.nlm.nih.gov/BLAST/
Chromosome 6	http://www.sanger.ac.uk/HGP/Chr6/
Chromosome 9	http://www.sanger.ac.uk/HGP/Chr9/
ClustalW	http://www.ebi.ac.uk/clustalw/index.html
Electronic PCR	http://www.ncbi.nlm.nih.gov/genome/sts/epcr.cgi
EMBL-EBI	http://www.ebi.ac.uk
EMBOSS	http://www.hgmp.mrc.ac.uk/Software/EMBOSS/overview.html
ENSEMBL	http://www.ensembl.org/Homo_sapiens
EPCLUST	http://www.ebi.ac.uk/microarray/ExpressionProfiler/ep.html

FINEX	http://www.sanger.ac.uk/cgi-bin/finex/finex_search.pl
GeneMap99	http://www.ncbi.nlm.nih.gov/genemap99/
HGMP	http://www.hgmp.mrc.ac.uk
HUGO	http://www.gene.ucl.ac.uk/hugo/
LocusLink	http://www.ncbi.nlm.nih.gov/LocusLink
MEGA2	http://www.megasoftware.net
MIPS	http://mips.gsf.de
NCBI	http://www.ncbi.nlm.nih.gov
NIX	http://www.hgmp.mrc.ac.uk/Registered/Webapp/nix/
PFAM	http://www.sanger.ac.uk/Software/pfamservice.shtml
Primer3	http://www.genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi
Repbase	http://www.geospiza.com/products/tools/repbase.htm
RepeatMasker	http://ftp.genome.washington.edu/cgi-bin/RepeatMasker
PHYLIP	http://evolution.genetics.washington.edu/phylip.html
Sanger Institute	http://www.sanger.ac.uk
TREE-PUZZLE	http://www.tree-puzzle.de
UniGene	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene

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.

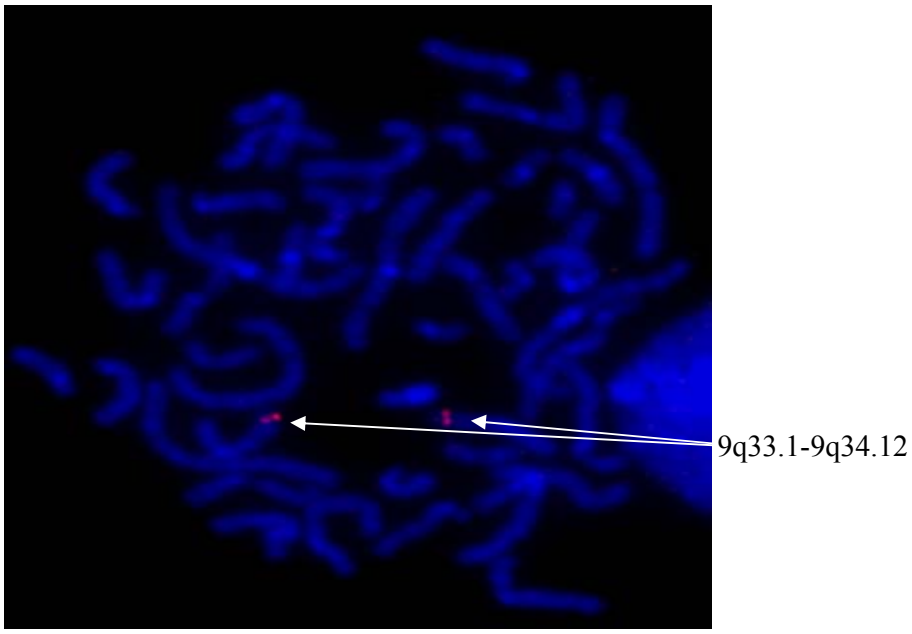


Figure 3.1 FISH analysis of bA465F21. Arrows indicate the location of the clone on the two copies of chromosome 9.

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).

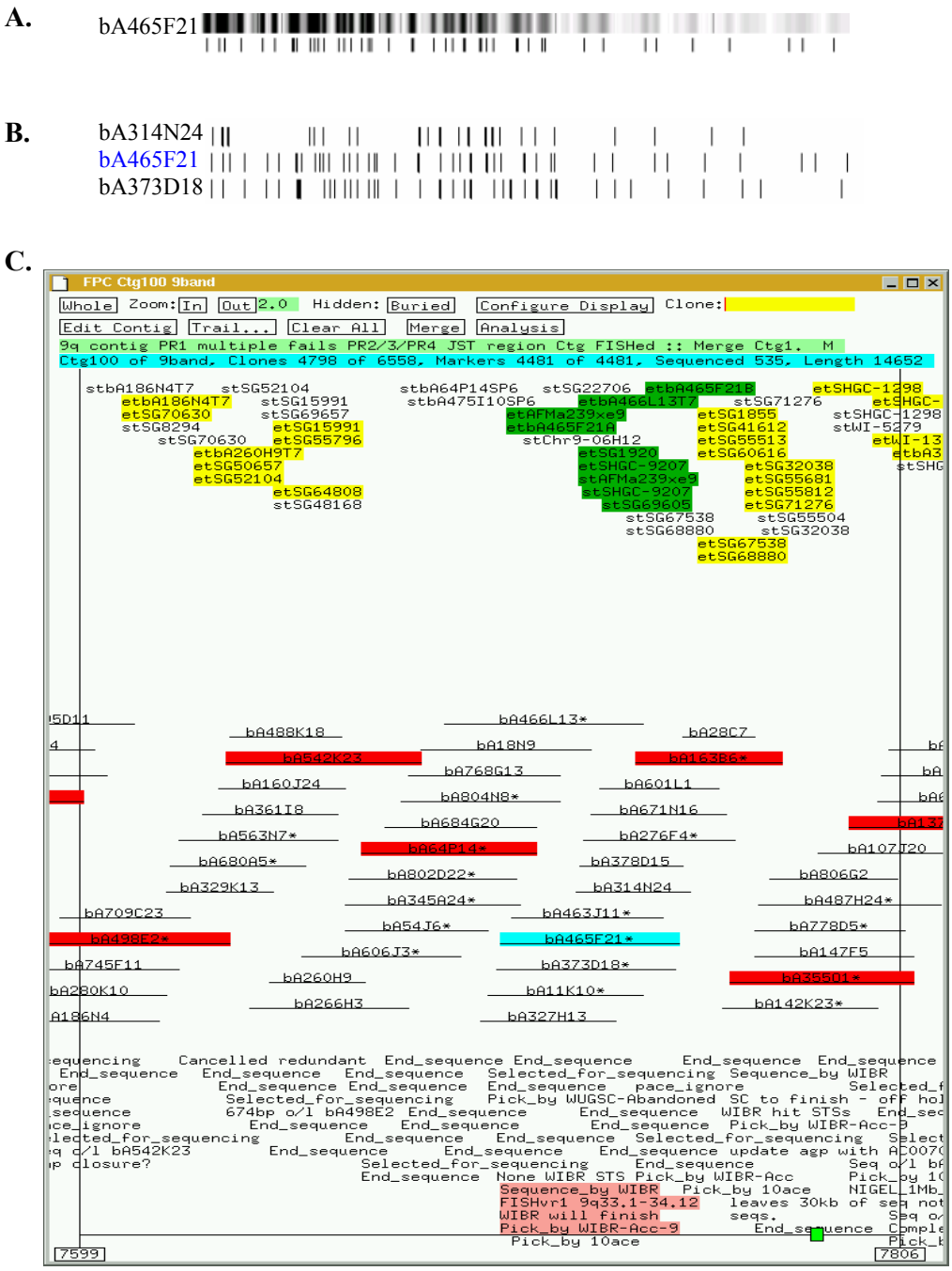


Figure 3.2 Localisation of the clone bA465F21 to the chromosome 9 tiling path. The gel image and the computationally determined fingerprint generated by *Hind* III restriction digest (A) was compared with the fingerprints of other chromosome 9 clones (B). Overlapping bands with clones bA314N24 and bA373D18 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).

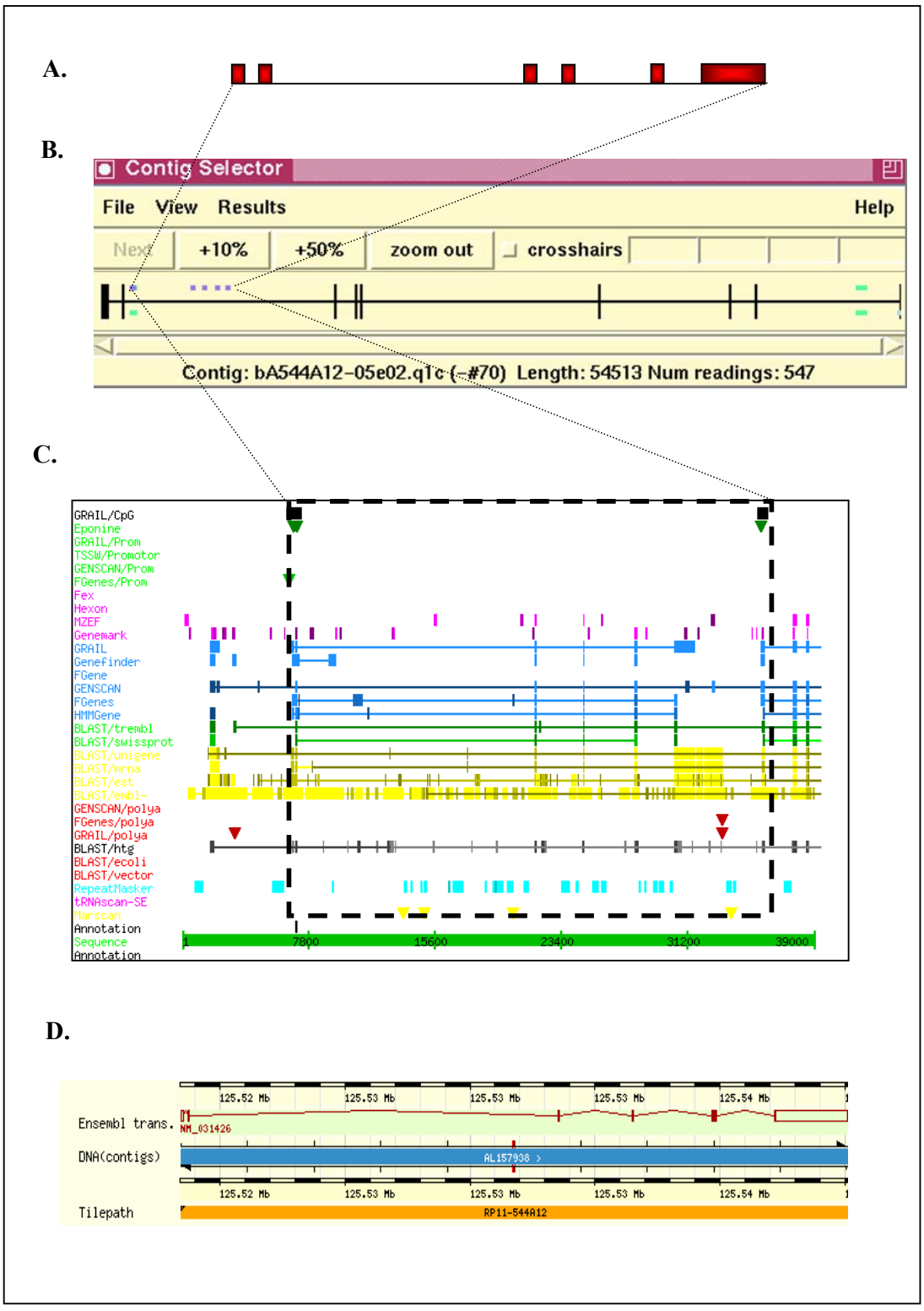


Figure 3.3 Computational identification of the AIF1 paralogue. The six exons of the AIF1 paralogue on chromosome 9, as shown in (A) were determined during sequence assembly (B) and NIX analysis (C). The gene structure has since been confirmed in ENSEMBL (D).

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).

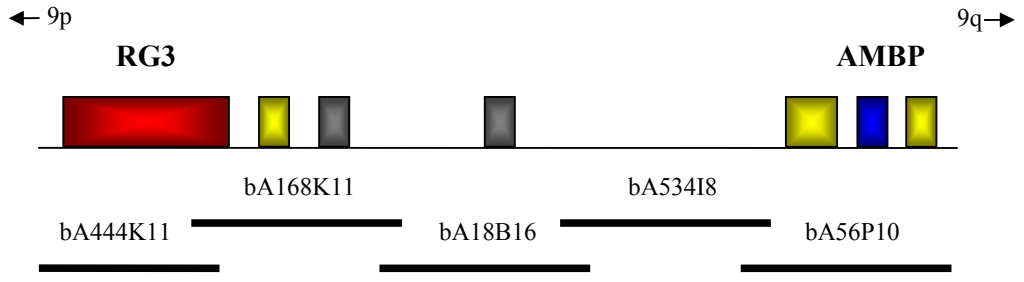


Figure 3.5 Overview of the gene content of region analysed to identify a putative GPX5 paralogue. The two known genes are labeled RG3 (red) and AMBP (blue) and novel genes are shown in grey and genes with no assigned name are shown in yellow (the latter refer to database entries NM_152575, Q8TF49 and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

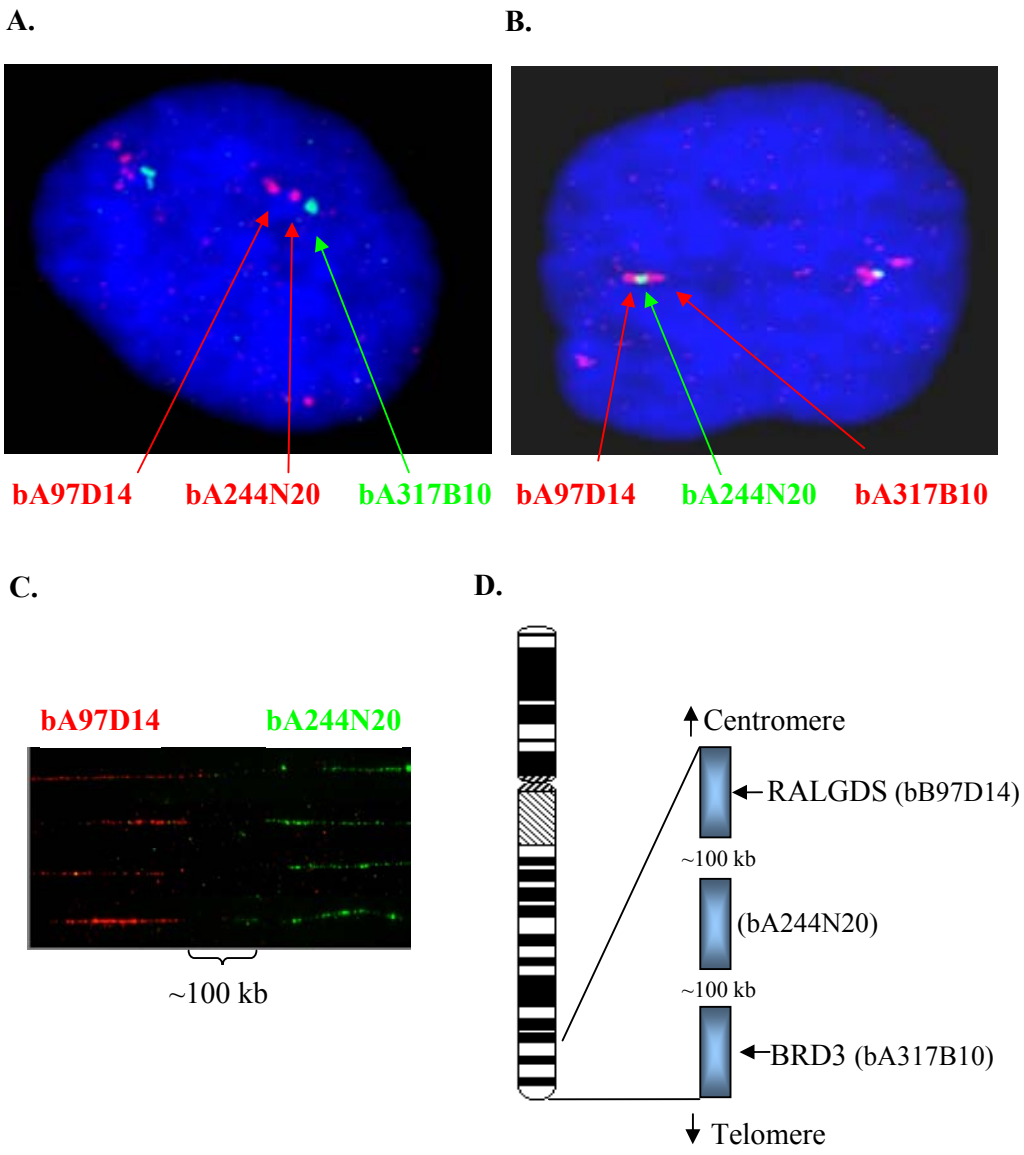


Figure 3.6 Overview of methods used to order and orientate the contigs containing RALGDS and BRD3 putative paralogues. Three clones representing three different contigs were labelled using fluorescent dyes then hybridised in different combinations to interphase chromosomes (A and B). By analysing the 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.

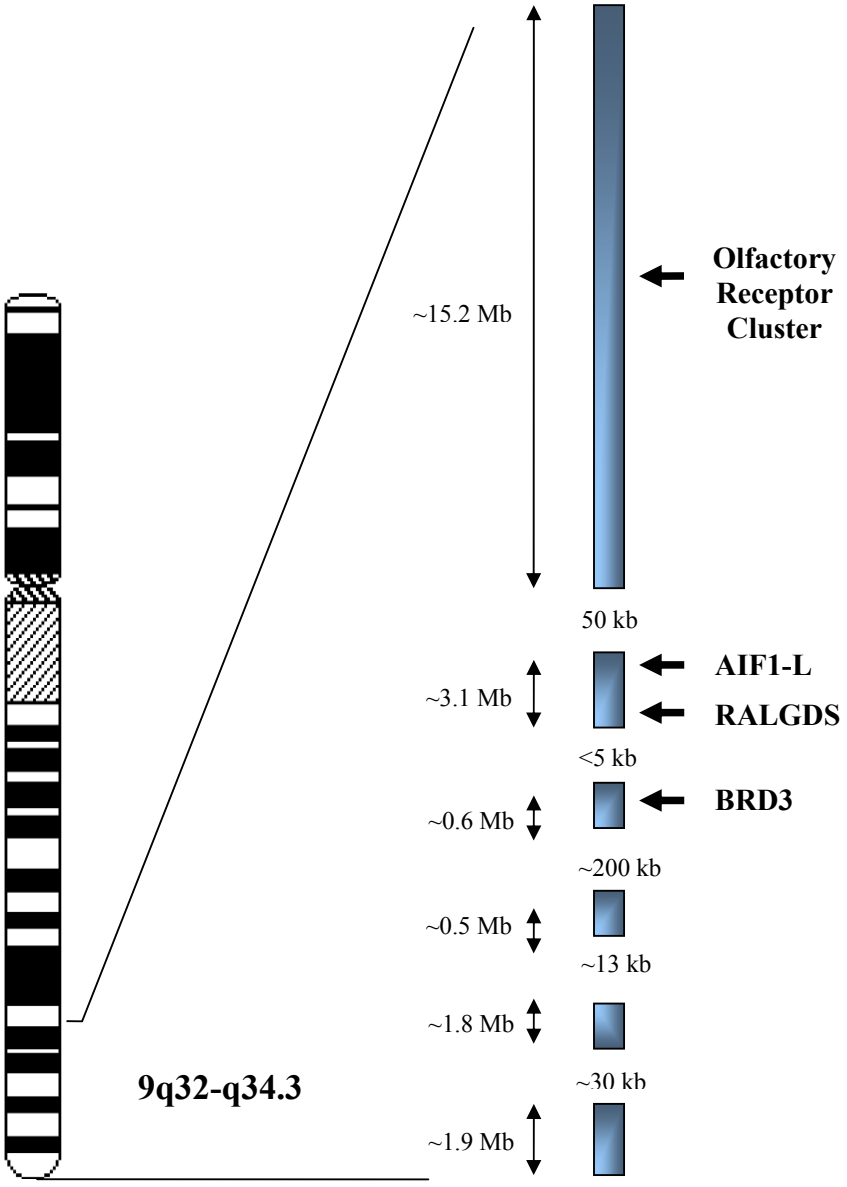


Figure 3.7 Schematic representation of the status (August 2003) of the MHC paralogous region on 9q32-q34.3.

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).

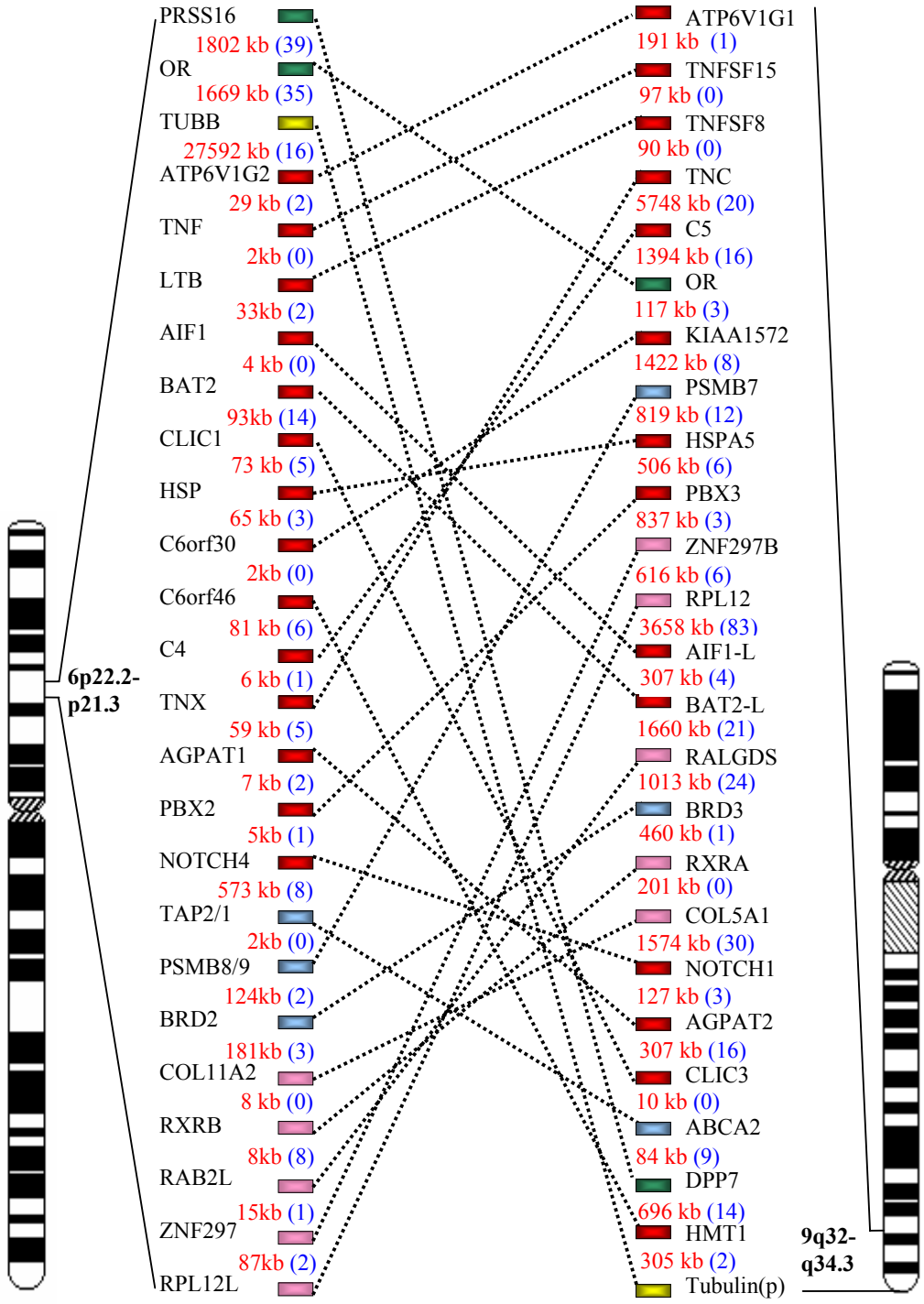


Figure 3.8 Comparison of the order of paralogues between the MHC region on 6p22.2-p21.3 and the paralogous region on 9q32-q34.3. The numbers in red indicate the distances between paralogues and the number in blue shown in parentheses corresponds to the number of coding genes interspersed between the 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.

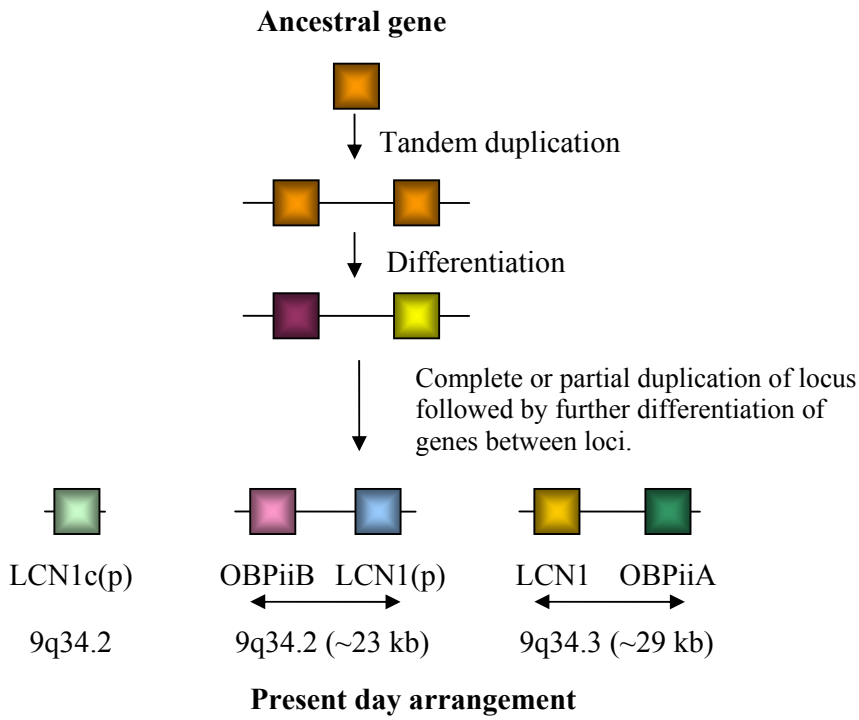


Figure 3.9 Evolution of the lipocalin paralogous gene family on 9q34. Gene differentiation is demonstrated by the change in colour of the boxes (genes) during evolution.

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.

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).

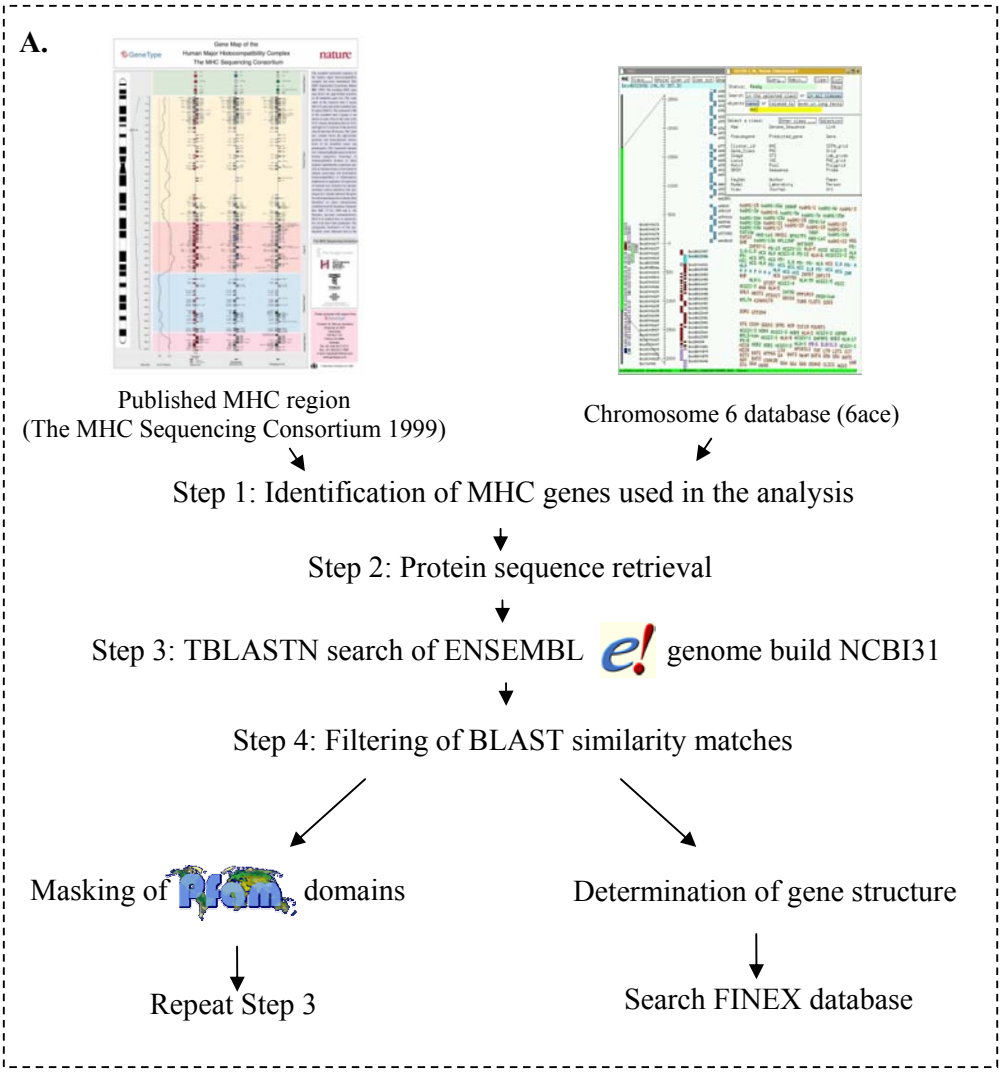


Figure 4.1 (A) Overview of the strategy used to identify MHC paralogues with increasing levels (L0 to L3) of confidence and definitions (B).

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).

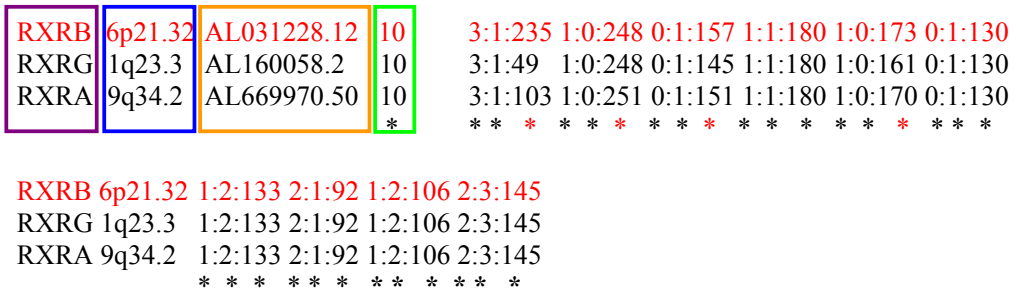


Figure 4.2 Alignment of the exon fingerprints of the extended MHC class I gene RXRB (in red) and its paralogues, RXRA and RXRG, identified in the genome survey (discussed in section 4.4.1). The gene names corresponding to the exon fingerprints are boxed in purple and the genomic location in blue. The 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.

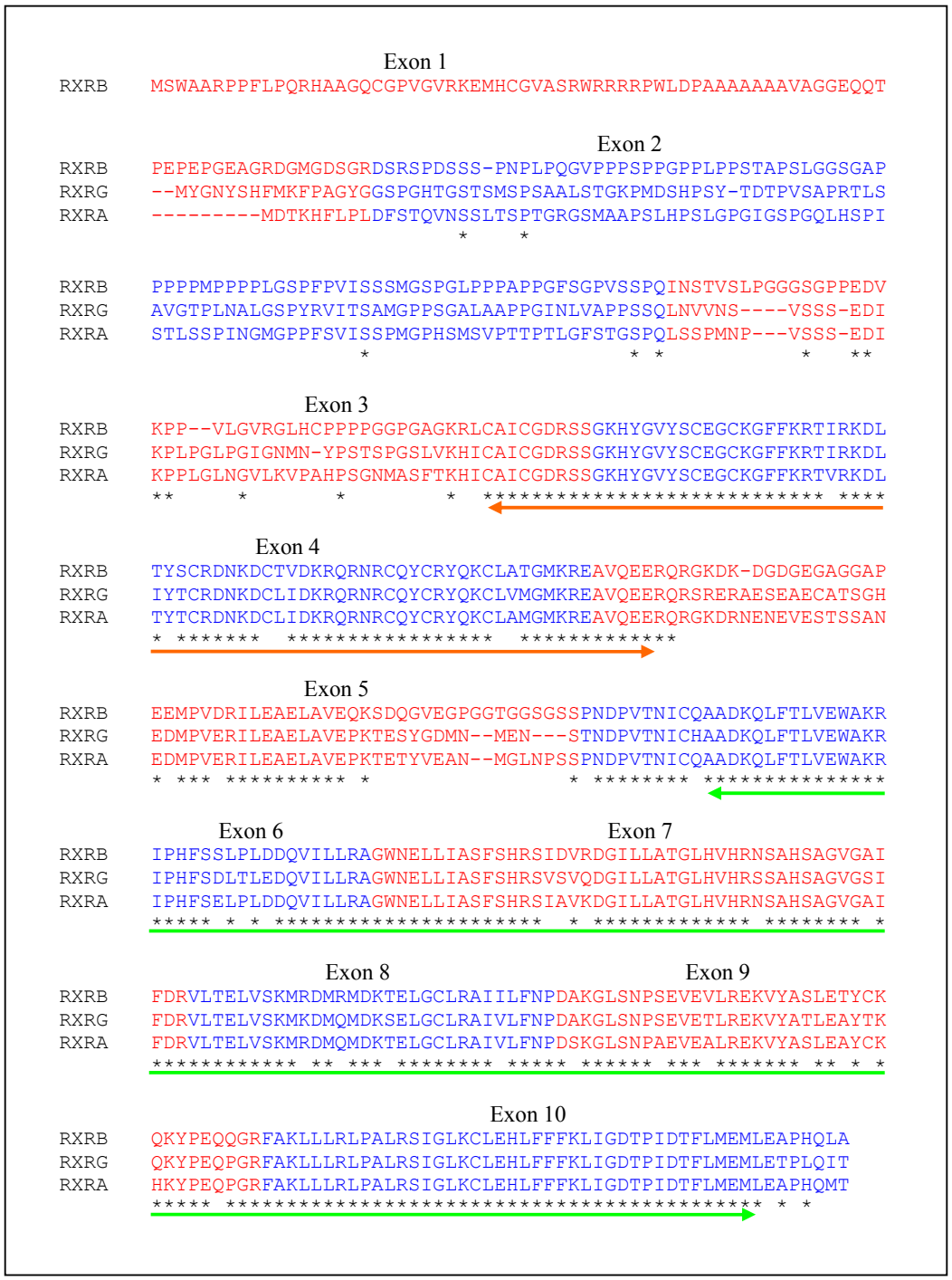


Figure 4.3 Protein sequence alignment of the extended MHC class II encoded protein, RXRB, and its two paralogues, RXRA and RXRG. The amino acid residues encoded by the 10 exons are shown in red and blue alternatively. The regions indicated by the orange and green arrows correspond to the zinc-finger 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).

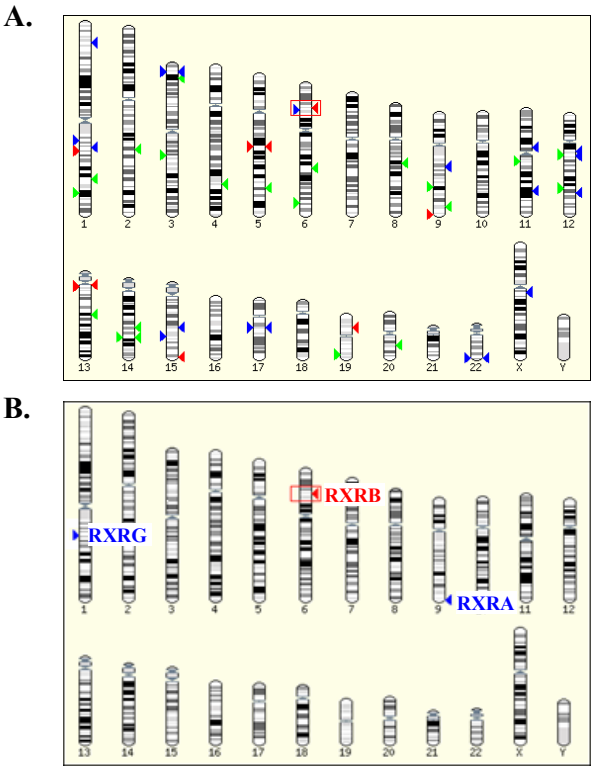


Figure 4.4 Summary of the results of the initial (A) and domain-masked (B) TBLASTN search of the human genome using the RXRB protein sequence. The coloured arrows correspond to the ENSEMBL BLAST score which roughly corresponds to the P-value. In short, a green arrow implies low score and high P-value, 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**

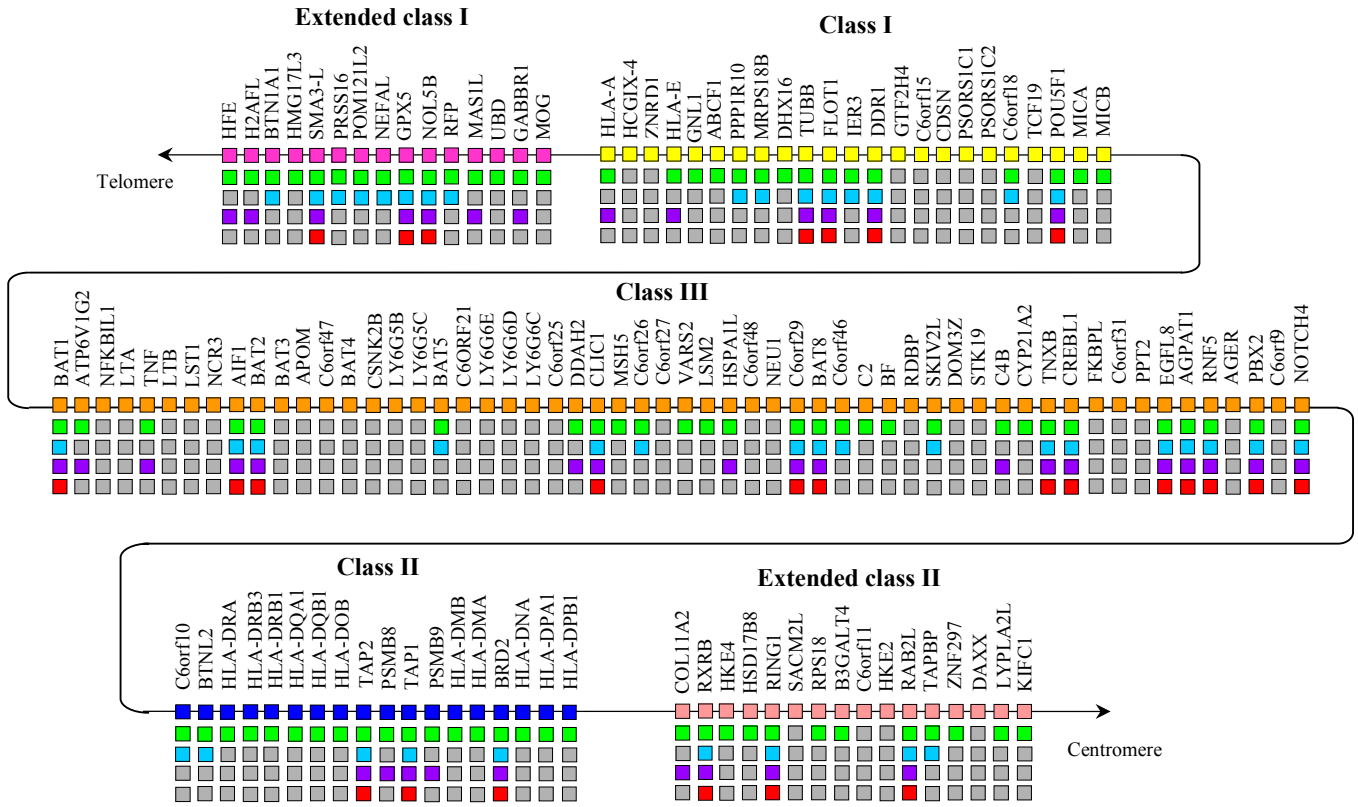


Figure 4.6 Summary of the results of the whole-genome survey using 128 MHC genes. The MHC region is divided into five classes and the genes within each class are represented by coloured boxes; extended class I are cerise, class I yellow, class III orange, class II blue and extended class II are pink. 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).

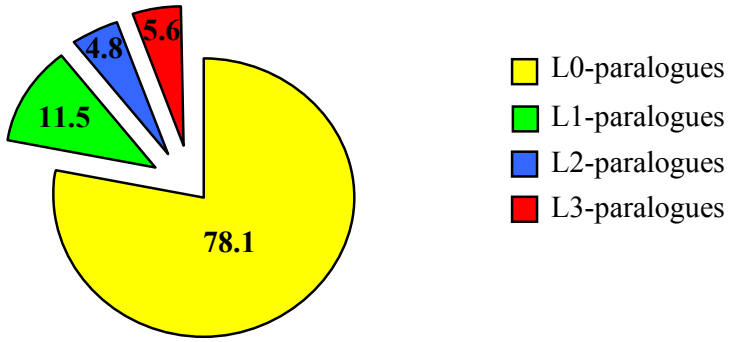


Figure 4.7 Summary of the proportion (%) of BLAST hits corresponding to the paralogues with different levels of confidence.

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.

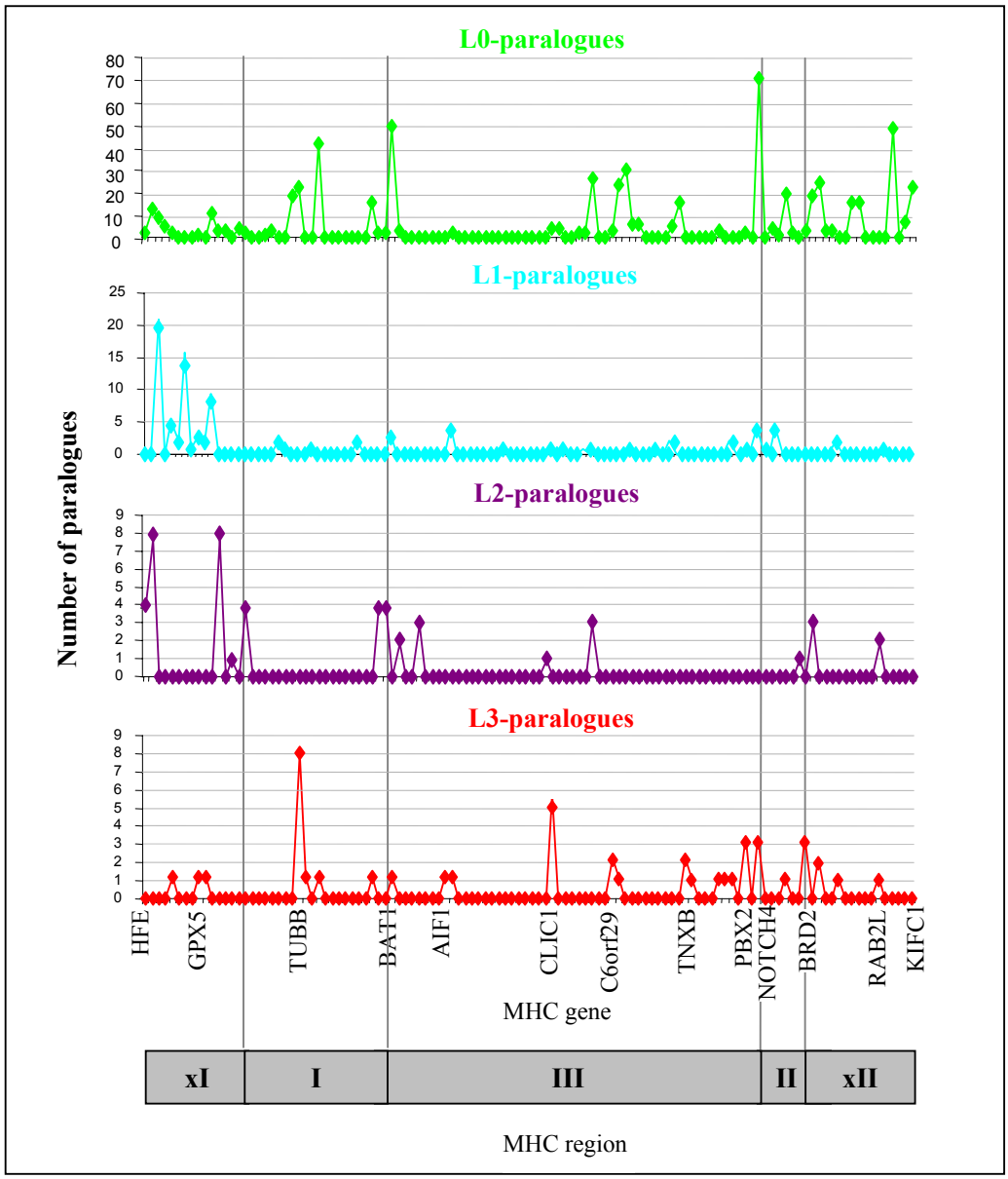


Figure 4.8 Summary of the MHC genes with L0- to L3-paralogues. The L0-paralogues are shown in green, the L1-paralogues are in turquoise, the L2-paralogues are shown in purple and the L3-paralogues are in red. The y-axis on each graph represent the number of paralogues identified with each level of confidence (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.

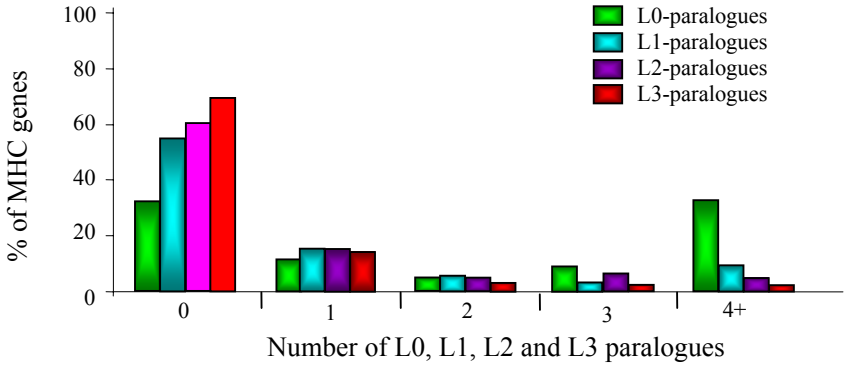


Figure 4.9 Summary of the percentage (%) of MHC genes with no, 1, 2, 3, 4 or more L0, L1, L2 and L3-paralogues in the human genome

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**

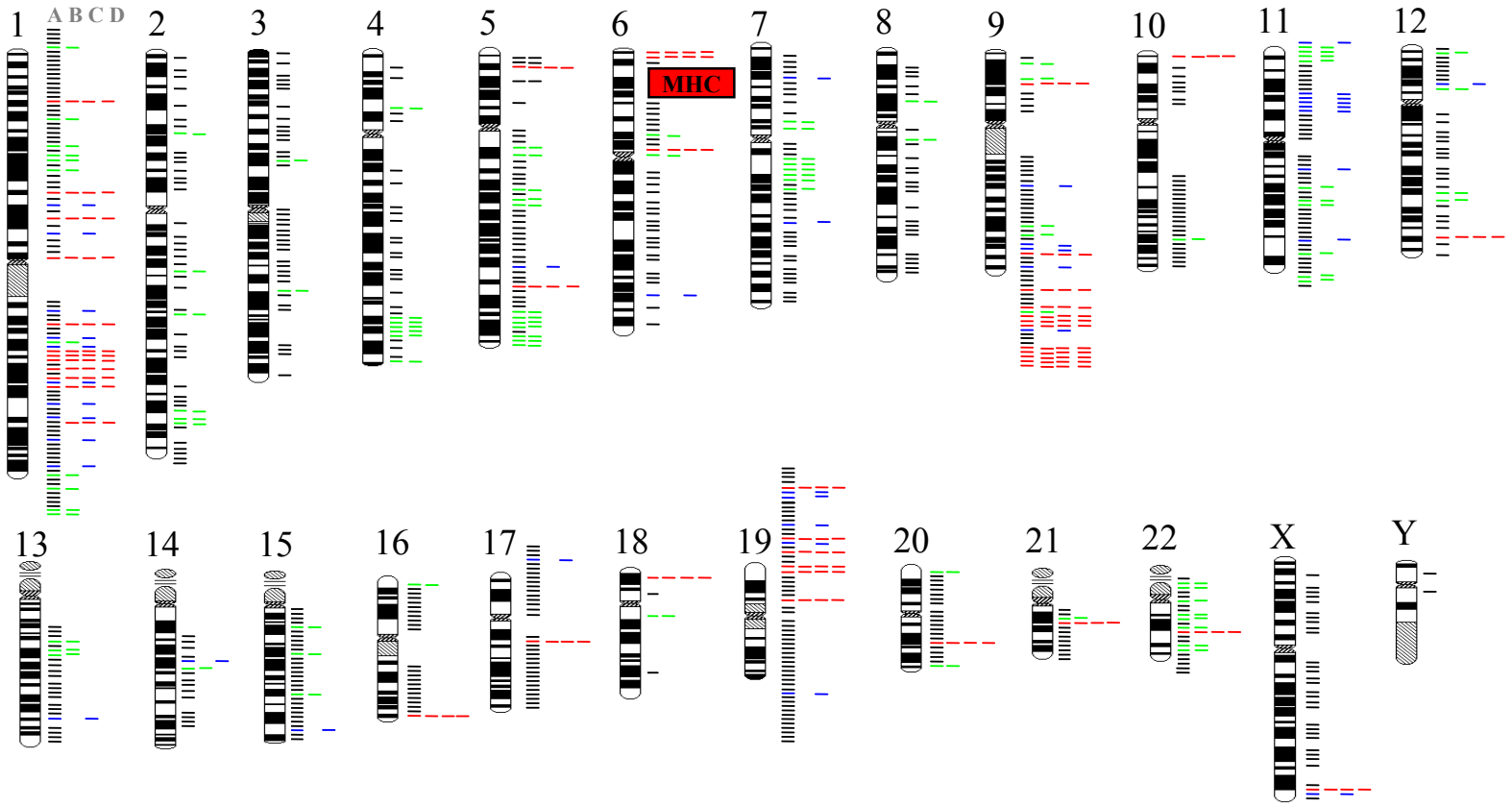


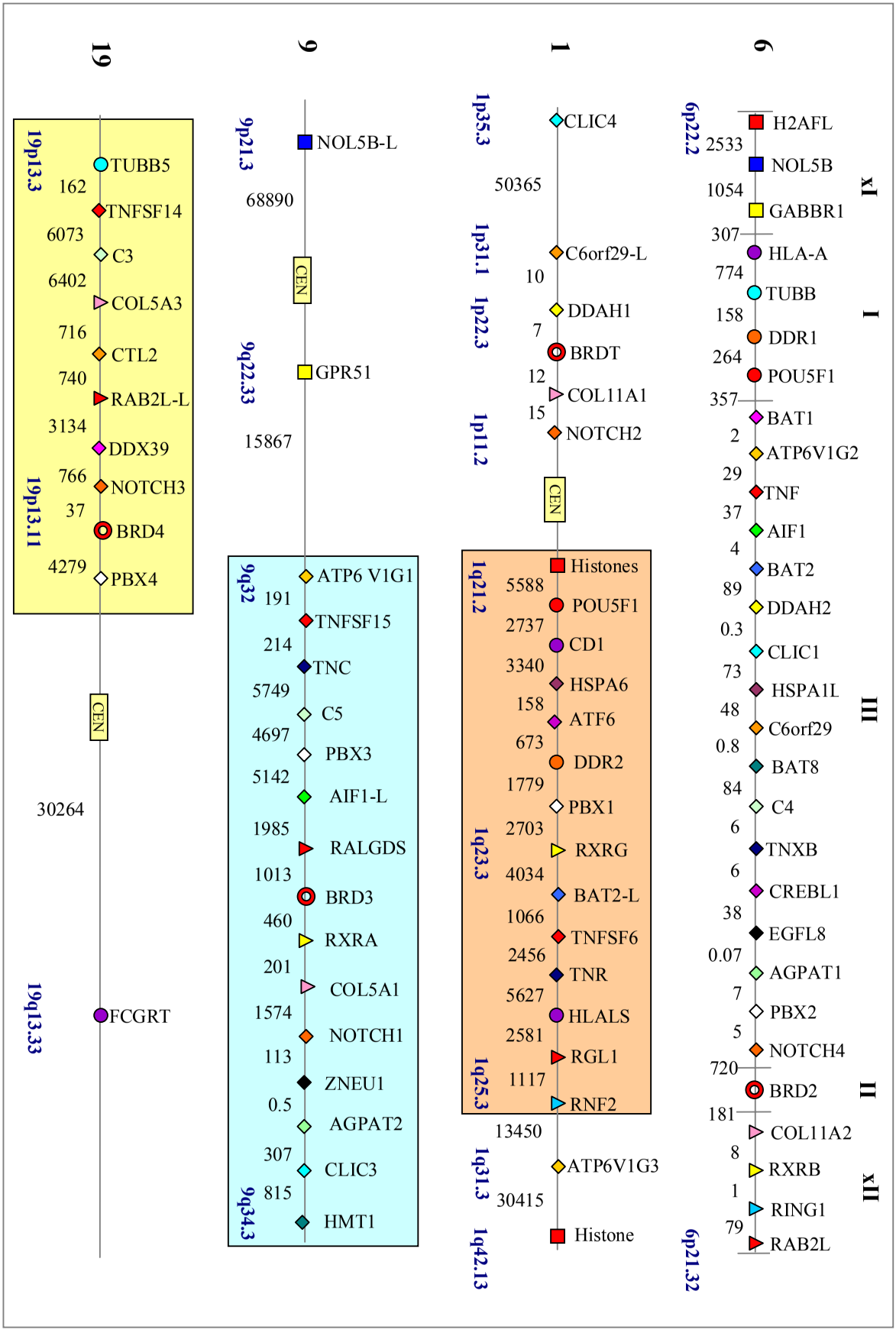
Figure 4.10 Distribution of MHC paralogues in the human genome. Column A represents all BLAST similarity matches with a P-value less than 10^{-5}. Column B represents the BLAST matches still detected after the domain-masking filtering step, column C represents BLAST matches still detected after the FINEX 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.


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

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).

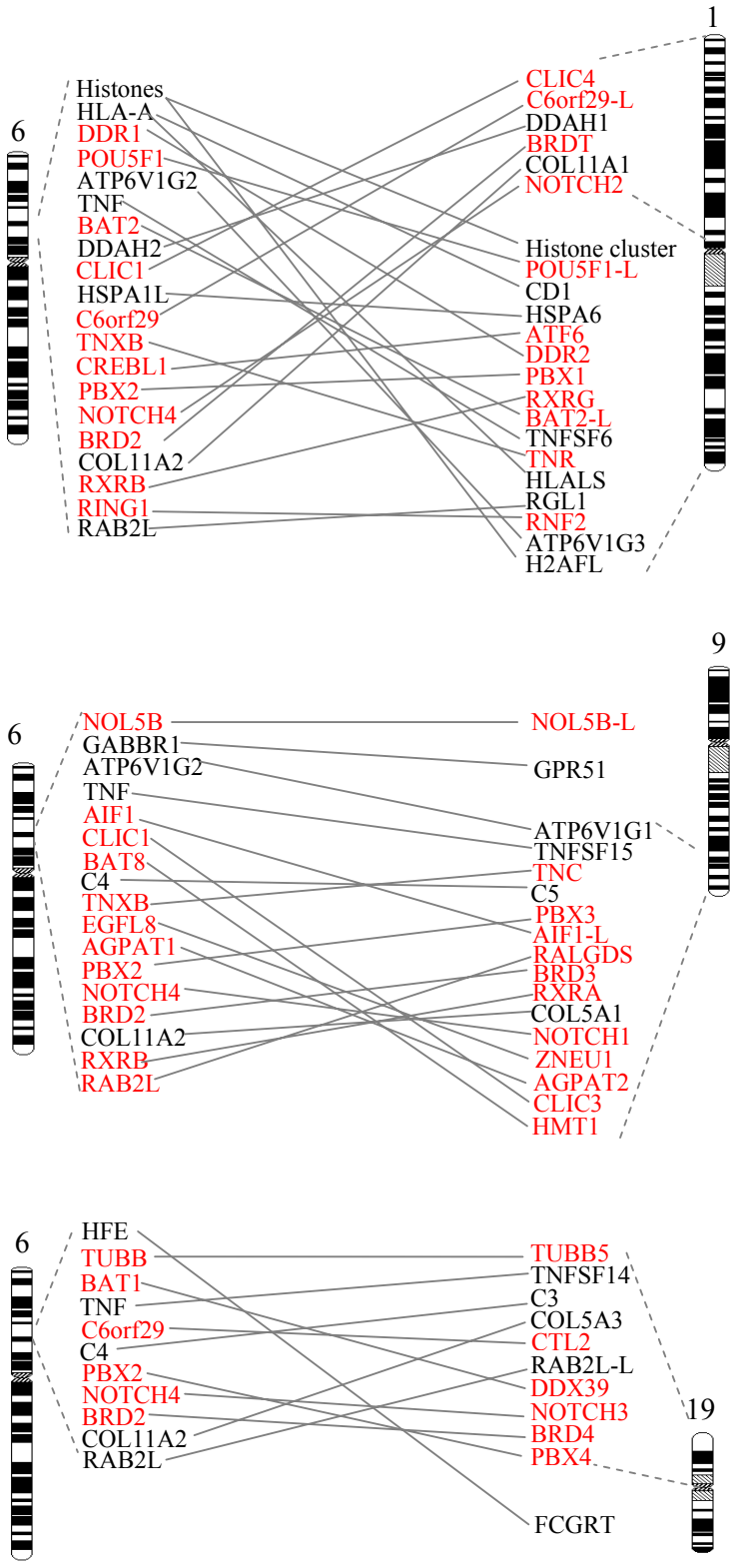


Figure 4.12 Comparison of the order of L2- and L3-paralogues on chromosomes 1, 9 and 19. The gene names in red represent L3-paralogues and the L2-paralogues are shown in black. Continued on next page.

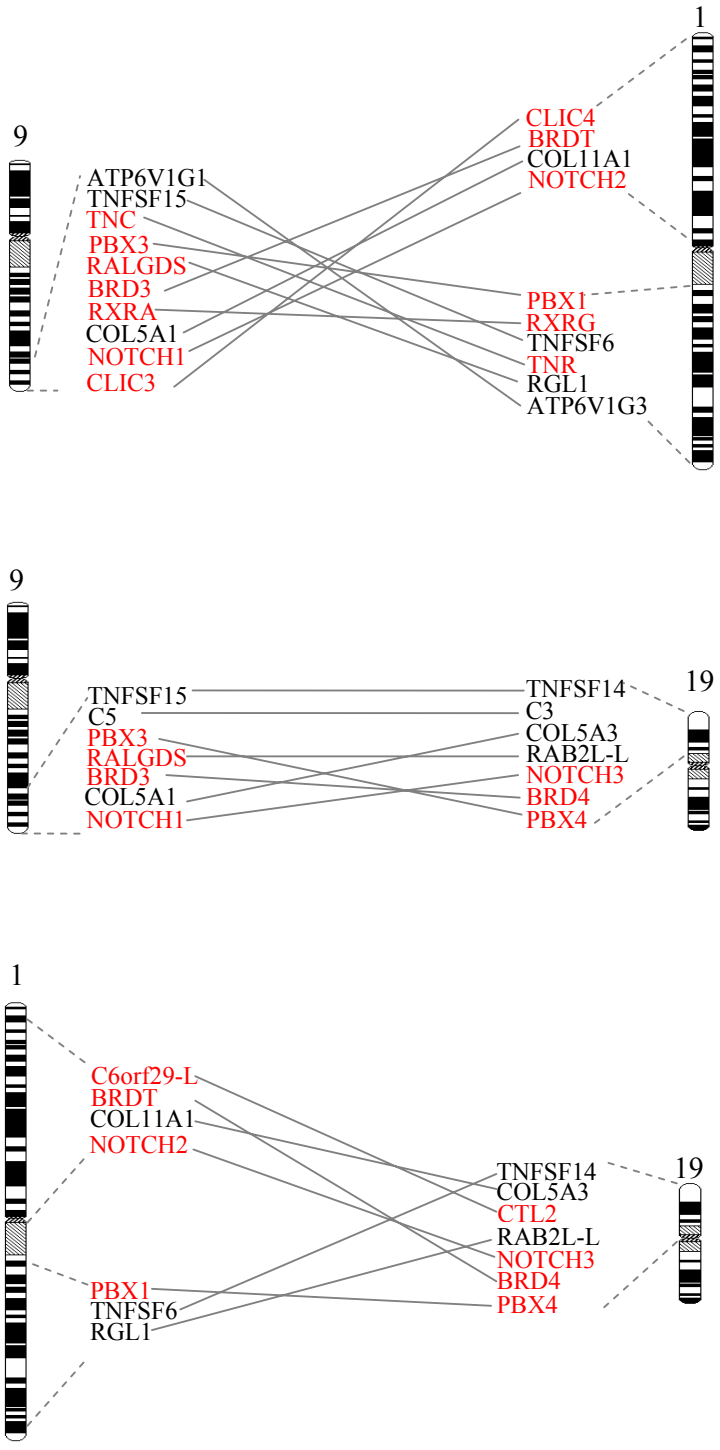


Figure 4.12 Continued. See previous page for legend.

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.

Chapter 5

Phylogenetic analysis of extended MHC paralogous gene families

5.1 Introduction

The genome-wide survey presented in chapter 4 identified over 700 MHC paralogues with varying levels of confidence. Analysis of the distribution of the 82 L2- and L3-paralogues confirmed that there were paralogous regions on chromosomes 1q21.2-q25.3, 9q32-q34.3 and 19p13.3-p13.11. One of the most interesting and novel findings was that there are also paralogues scattered throughout the genome. However, the origin of these paralogues is not known. By definition paralogues have arisen by duplication of an ancestral gene, which can involve a chromosomal segment containing one or more genes (block duplication), an entire chromosome or the whole genome. One of the most useful approaches to study the history of paralogues is to reconstruct the evolutionary relationships using orthologous sequences.

These relationships are commonly represented by means of a phylogenetic tree using sequence data from a range of evolutionary distant organisms. A phylogenetic tree is simply a branching diagram in which each terminal element (e.g. a protein sequence) is linked only once to one or more other protein sequences, thus specifying a hierarchy. Trees can be rooted using a distantly related sequence, such as the *Drosophila* or amphioxus orthologue, and corresponds to a point at the base of a tree indicating the evolutionary direction. Internal branch points, or 'nodes', represent putative ancestors and are connected by 'branches'. Two sequences that are very

much alike will be located as neighbouring outside branches and will be joined to a common branch beneath them. Evolutionary trees can be constructed such that the length of a branch connecting two proteins is proportional to the number of residue differences in the sequences. Thus, the object of phylogenetic analysis is to discover all of the branching relationships in the tree and the branch lengths.

The paralogues located in the paralogous regions on chromosomes 1, 9 and 19 are believed to be remnants of two rounds of large-scale duplication events involving the whole genome early in vertebrate history. This phenomenon is referred to as the 2R hypothesis (Sidow, 1996). The first whole-genome duplication event occurred after an 'amphioxus stage' prior to the divergence of Agnatha (jawless vertebrates, represented by lamprey and hagfish) and Gnathostomata (jawed vertebrates), while a second occurred after the divergence of Agnatha but before the divergence of cartilagenous fish (represented by sharks) (summarised in figure 5.1).

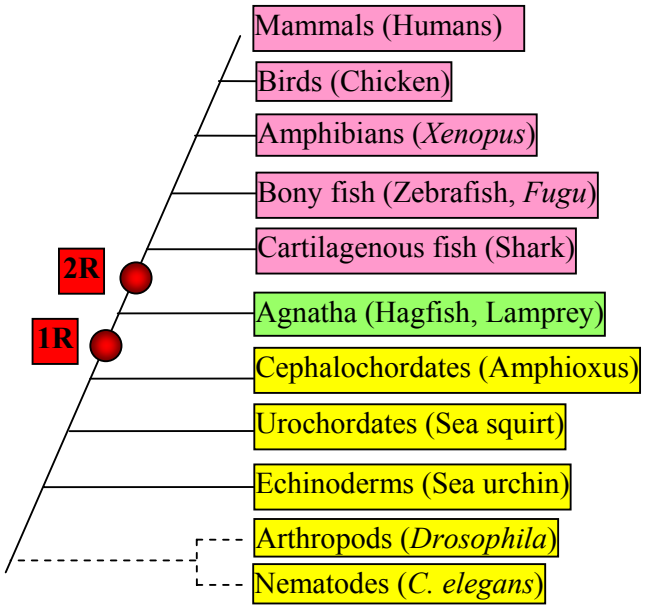


Figure 5.1 Summary of the 2R hypothesis. The first round of genome duplication (1R) occurred after the emergence of amphioxus (yellow), prior to agnatha divergence (green) and the second occurred after the divergence of agnatha but before the divergence of cartilagenous fish (pink). The two duplication events are represented by red circles.

If two or more genes have been duplicated simultaneously as the result of a block duplication event, this should be revealed by phylogenetic analysis. If the 2R hypothesis is correct (assuming no genes have been lost since duplication) four paralogous genes should be found in humans and other jawed vertebrates, such as mice and chickens. Jawless fish, such as hagfish, should only have two paralogous genes, which are considered orthologous to the four paralogues in jawed vertebrates and the cephalochordate will have only one; corresponding to the closest relative of the 'ancestral gene' (figure 5.2). The branching pattern of the phylogenetic tree should be representative of the duplication events showing the double-forked tree topology, or the so-called 2+2 or (A,B)(C,D) topology. The age of the split of AB and CD is the same thus showing the history of successive rounds of duplication (summarised in figure 5.2).

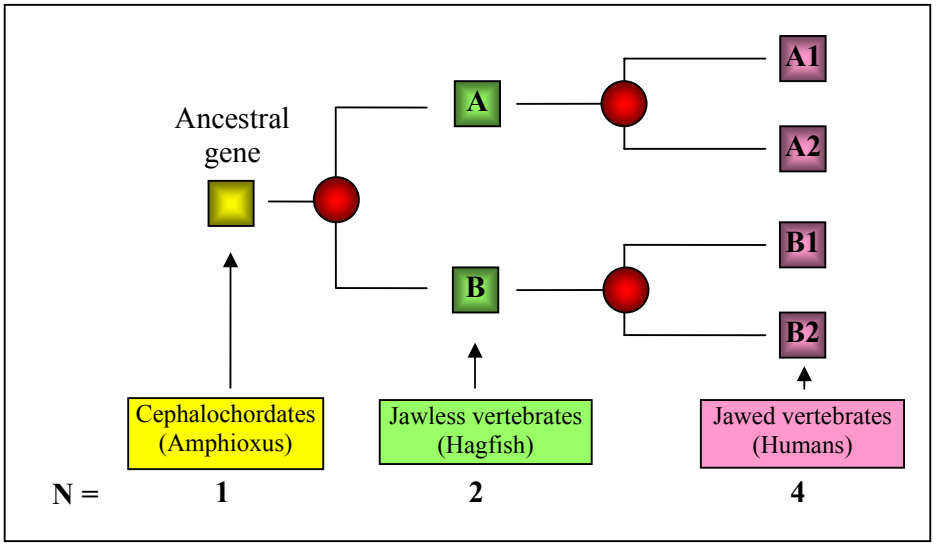


Figure 5.2 Schematic representation of the effects of two rounds of gene, or genome, duplication on the topology of the phylogenetic tree (A1,A2)(B1,B2) and the resulting number of (N=) paralogues in 'key' species (1:2:4 ratio between amphioxus:hagfish:humans).

In this chapter, I present the phylogenetic analyses of ten paralogous gene families in order to determine the mechanism(s) by which they arose. Figure 5.3 summarises the topology of the phylogenetic tree expected if the paralogues arose from a common ancestor via two rounds of genome duplication (i.e. support the 2R hypothesis).

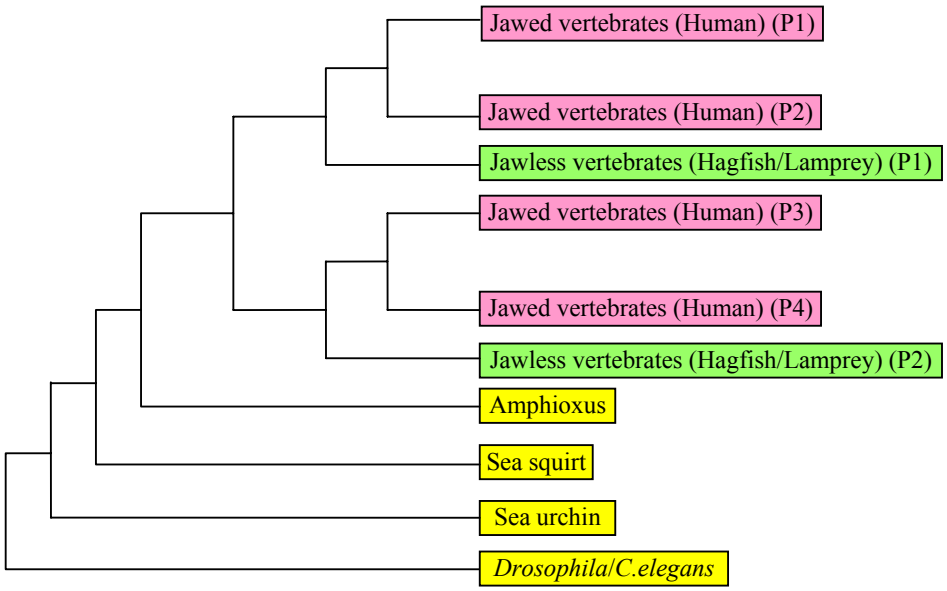


Figure 5.3 Schematic representation of the 'ideal' phylogenetic tree in support of the 2R hypothesis. The species are colour coded according to the number of expected paralogues; species with one copy are highlighted in yellow, two copies (P1-2) in green and four paralogues (P1-4) in pink. It is important to note that in the phylogenetic trees presented in this chapter the species zebrafish, *Fugu* and *Xenopus* are highlighted pink, since they are jawed vertebrates, but they are expected to have more than four paralogues as an additional genome duplication event has occurred in their lineage.

5.2 MHC paralogous gene families used in phylogenetic analysis

In order to understand the evolutionary history of the MHC paralogues, 10 MHC genes and their paralogues (termed paralogous gene families) were selected for further analysis (summarised in figure 5.4).

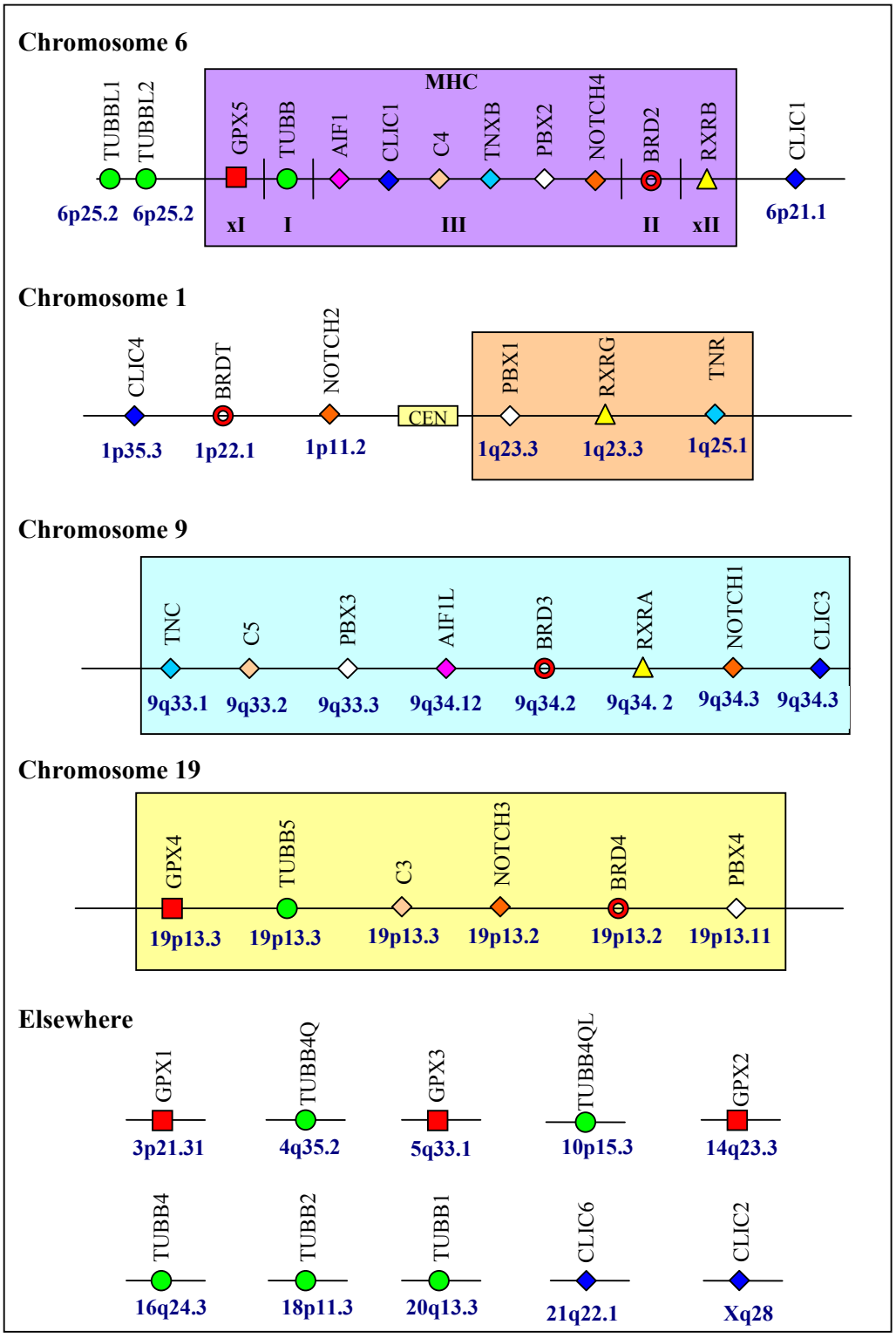


Figure 5.4 Summary of the MHC genes and paralogues selected for further investigation. The MHC genes and corresponding paralogues are represented by a shaded symbol. The cytogenetic locus for each gene on chromosomes 1, 9 and 19 is shown in blue text. 'CEN' corresponds to the centromere. The shaded areas correspond to the paralogous regions as defined in chapter 4.

The 10 paralogous gene families presented in this chapter were selected in order to satisfy a number of criteria. Firstly, the families were chosen to ensure that each of the five classes of the MHC region were represented by at least one paralogous gene family. Secondly, there were families with L2- and L3-paralogues located within the gene clusters on 1q21-q25, 9q32-q34.3 and 19p13.3-p31.3 only and, finally, there were also families with paralogues located elsewhere in the human genome.

5.3 Results

The protein sequences corresponding to the orthologues and paralogues of the 10 MHC genes were identified by searching the annotated protein databases and literature. The protein sequences were aligned with the ClustalW program using default parameters, and edited in Jalview. The sequence similarity between the MHC paralogues showed varying levels of divergence and, it was found that, the sequence alignments were often only reliable for conserved regions of the proteins. Therefore, in most circumstances, only these conserved regions were used to generate the trees. However, in cases, such as the TUBB family, where the sequence identity is very high (between 72.9 and 99.6%), the full length protein sequences were used. The number of sequences and protein regions used to produce the trees is summarised in table 5.1.

Table 5.1 Summary of the MHC paralogous gene families used to generate phylogenetic trees. The first three columns show the MHC gene locus, the location within the MHC region and the location of their paralogues, respectively. The remaining four columns, from left to right, show the number of sequences, the gamma-distribution alpha-parameter (α), number of amino acid (aa) residues and a description of the protein region used to generate the trees. The alpha-parameter is a measure of the rate of heterogeneity or change between amino acid sites (as described in 2.19.2). PR stands for paralogous region.

MHC Locus	***MHC class***	***Location of paralogues***	***No. of sequences used***	***α***	***aa residues used/length***	***Description of protein region used***
GPX5	xI	Outside PRs	22	1.09	221/221	Complete sequence
TUBB	I	Inside and outside PRs	27	0.29	444/444	Complete sequence
AIF1	III	In PR	11	1.21	92/147	Includes EF-hand domain
CLIC1	III	Inside and outside PRs	13	1.82	232/241	Most of sequence
C4	III	In PRs	27	1.42	1275/1744	Includes anaphylatoxin and macroglobulin domains
TNXB	III	In PRs	14	1.28	309/4289	Includes a fibronectin III domain and the fibrinogen c-terminal.
PBX2	III	In PRs	13	0.33	180/430	Includes homebox
NOTCH4	III	In PRs	17	0.97	385/2003	Includes 11 EGF-like domains
BRD2	II	In PRs	14	0.74	113/801	Includes a bromodomain
RXRB	xII	In PRs	20	0.23	313/533	Includes the DNA binding domain.

The phylogenetic trees presented in this chapter, unless otherwise stated, are a consensus of four trees generated using the three software packages: PHYLIP, MEGA2 and PUZZLE (as described in section 2.19). In each tree, the number on the branches of the tree correspond to the average percentage bootstrap or puzzling-step confidences from the three software packages. It should be noted that the protein names for all species, apart from human, are given in lower case.

5.3.1 Phylogenetic analysis of the BRD paralogous gene family

The BRD2, or the bromodomain containing 2, gene is located in the MHC class II region (Beck *et al*, 1992b). Denis and Green (1996) discovered that the RING3 product is a mitogen-activated nuclear kinase involved in signal transduction and that it is upregulated in certain types of leukeamia. In total, three paralogues of the BRD2 gene have been identified in the human genome with the highest level of confidence; these are BRDT on 1p22.3, BRD3 on 9q34.2 and BRD4 on 19p13.12. They all belong to the BET subgroup of bromodomain proteins and contain two bromodomains and an ET (or extraterminal) motif, which is a protein-protein interactive surface. The precise function of the bromodomains is unclear but it may be involved in protein-protein interactions and may play a role in assembly or activity of multi-component complexes involved in transcriptional activation (Tamkun, 1995).

The topology of the phylogenetic tree of the BRD paralogous gene family is (BRD2,BRD3)(BRD4,BRDT), thus supporting the 2R hypothesis (figure 5.5). Phylogenetic analysis shows that the timings of the two duplication events occurred after the divergence of cephalochordates and prior to the emergence of jawless fish. This is indicated by the positions of the amphioxus and hagfish orthologues on the

tree. Overall, the phylogenetic analysis of the BRD2 paralogues and othologues shows that the BRD paralogous gene family arose by two rounds of duplication, but it should be noted that some branches show low levels of support.

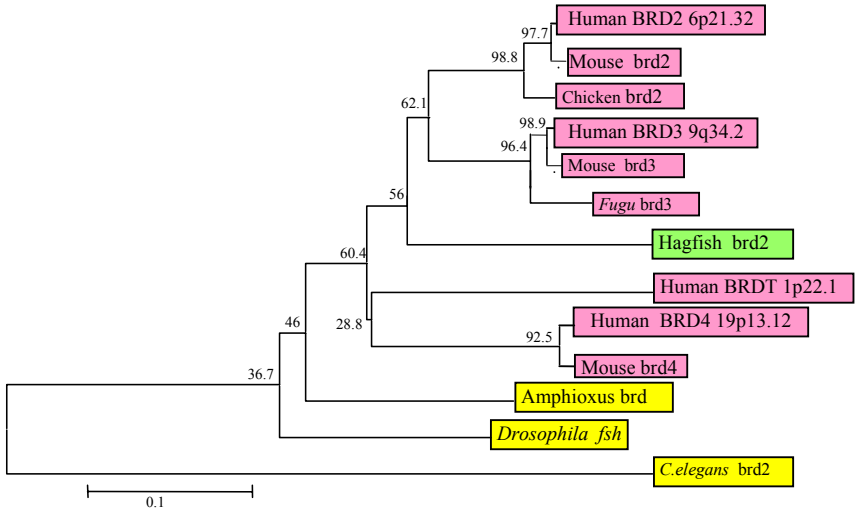


Figure 5.5 Phylogenetic tree of the BRD paralogous and orthologous family. The accession numbers are: P25440 (human BRD2), Q15059 (human BRD3), O14789 (human BRDT), O60885 (human BRD4), O54795 (mouse brd2), Q8K2F0 (mouse brd3), Q9ESU6 (mouse brd4), Q90971 (chicken brd2), Q8QFT7 (*Fugu* brd3), Q8T775 (amphioxus brd), P13709 (*Drosophila fsh*) and Q20948 (*C.elegans* brd2).

5.3.2 Phylogenetic analysis of the PBX paralogous gene family

The PBX2 (pre-B-cell leukaemia 2) gene encodes a homeodomain-containing protein. It was first identified on the basis of the extensive homology to the PBX1 gene involved in t(1;19)(q23;p13.3) translocation in acute pre-B-cell leukaemias (Monica *et al*, 1991). The genome survey identified three paralogues located within the paralogous regions on 1q23.3, 9q33.3 and 19p13.11, named PBX1, PBX3 and PBX4, respectively. Phylogenetic analysis shows that the PBX paralogous gene family arose by two rounds of duplication (figure 5.6). The topology of the tree is (PBX2,

PBX4)(PBX1, PBX3), which supports the 2R hypothesis. The timings of the two duplication events can be determined as occurring after the divergence of cephalochordates and prior to the emergence of jawed fish, indicated by the positions of the amphioxus and zebrafish orthologues on the tree.

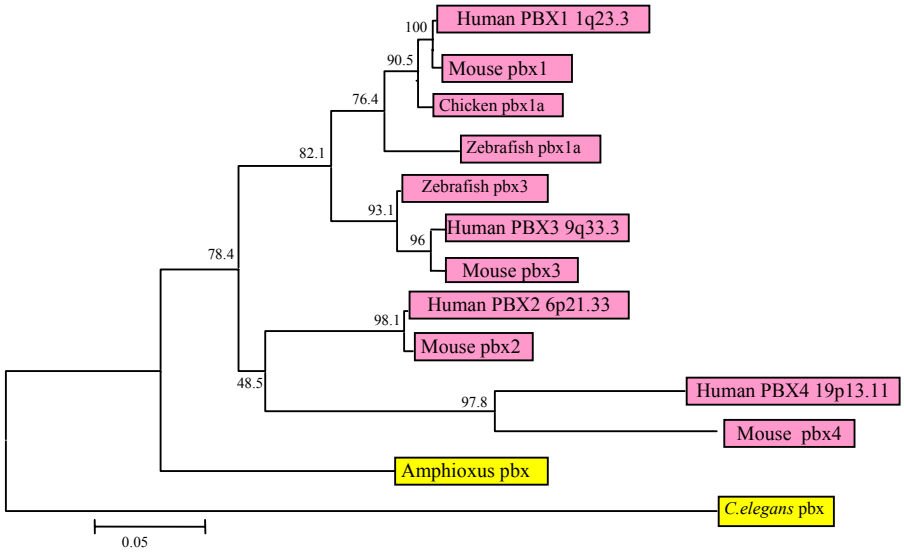


Figure 5.6 Phylogenetic analysis of the PBX paralogous gene family. The accession numbers of the vertebrate protein sequences used are: P40425 (human PBX2), P40424 (human PBX1), P40426 (human PBX3), Q9BYU1 (human PBX4), O35984 (mouse pbx2), P41778 (mouse pbx1), O35317 (mouse pbx3), Q99NE9 (mouse pbx4), Q9IB15 (chicken pbx1a), Q9I9B7 (zebrafish pbx1a), Q9I9B5 (zebrafish pbx3), AF39192_1 (amphioxus pbx) and P41779 (*C.elegans* pbx).

5.3.3 Phylogenetic analysis of the NOTCH paralogous gene family

The Notch gene was first identified in *Drosophila* as a regulator of cell fate determination and has been implicated in a large number of developmental processes in *Drosophila* and vertebrate systems (reviewed by Bray, 1998; Lewis, 1998). The phylogenetic tree using 14 vertebrate protein sequences and three invertebrate protein sequences is presented in figure 5.7.

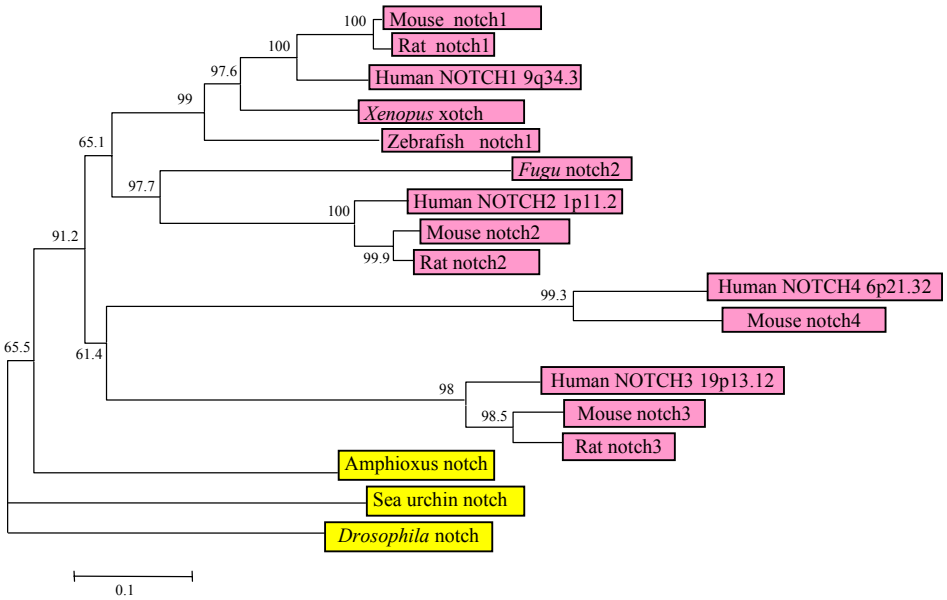


Figure 5.7 Phylogenetic analysis of the NOTCH paralogous gene family. The protein sequences, with the accession numbers given in parentheses, are: human NOTCH4 (O00306), human NOTCH1 (P46531), human NOTCH2 (Q04721), human NOTCH3 (Q9UM47), mouse notch4 (P31695), mouse notch1 (Q01705), mouse notch2 (O35516), mouse notch3 (Q61982), rat notch1 (O07008), rat notch2 (Q9QW30), rat notch3 (Q9R172), *Xenopus* xotch (P21783), zebrafish notch1 (P46530), *Fugu* notch2 (O13149), amphioxus notch (Q9GPA5), sea urchin notch (O16004) and *Drosophila* notch (P07207).

In *Drosophila* and lower deuterostomes (such as sea urchins) there is a single Notch gene, while in vertebrates there are multiple Notch genes (four in humans and mouse). Phylogenetic analysis of the NOTCH4 paralogues and orthologues supports the hypothesis that NOTCH1-4 arose from a common ancestor via two duplication events. The single amphioxus notch protein branches at the base of the four vertebrate Notch proteins. Together with the presence of single Notch gene in the sea urchin it suggests that Notch duplicated within the vertebrate lineage. Thus, both duplications occurred after the divergence of amphioxus and prior to the divergence of bony fish and tetrapods.

5.3.4 Phylogenetic analysis of the complement paralogous gene family

The C4 gene is located in the MHC class III region and encodes the complement factor 4 protein. C4 plays a central role in the activation of the classical pathway of the complement system. The complement system is the principle effector mechanism of humoral immunity and consists of at least 24 serum proteins and 11 membrane-bound proteins. The interaction of these proteins leads to a complement cascade and results in a number of responses, including cell lysis, opsonisation of targets for phagocytosis by macrophages, regulation of B cell responses and the generation of potent anaphylatoxins (for review see Reid and Porter, 1981). Two paralogues of the C4 gene have been identified in the human genome; these are C5 located on 9q33.2 and C3 located on 19p13.3.

Phylogenetic analysis of the full length C3, C4 and C5 protein sequences (figure 5.8) supports the view that C5 diverged first with C3 and C4 subsequently diverging before the separation of jawed and jawless fishes (Hughes, 1994). The presence of C3 in jawless deuterostomes, such as sea urchin (Smith *et al*, 1999), hagfish (Ishiguro *et al*, 1992) and lamprey (Nonaka and Takashii, 1992) enables the divergence times of the complement genes to be determined and establishes the ancient origin of the complement system. The clustering of the lamprey and hagfish C3 with the other vertebrate C3 proteins clearly indicates that the duplication of the C3 and C4 genes from the ancestral gene occurred after the divergence of jawless vertebrates and prior to the divergence of jawed vertebrates. One would expect the orthologue of C4 to be revealed upon full sequencing of the hagfish and lamprey genomes. The divergence of C5 occurred after the cephalocordate split prior to the divergence of jawless fish. Phylogenetic analysis shows a ratio of 1:2:2:3 Amphioxus:Hagfish:Lamprey:Human.

Thus, supporting the 2R hypothesis, accompanied with the loss of one of the C4 duplicate giving the topology (C5(C4, C3)) rather than the predicted (A,B)(C,D).

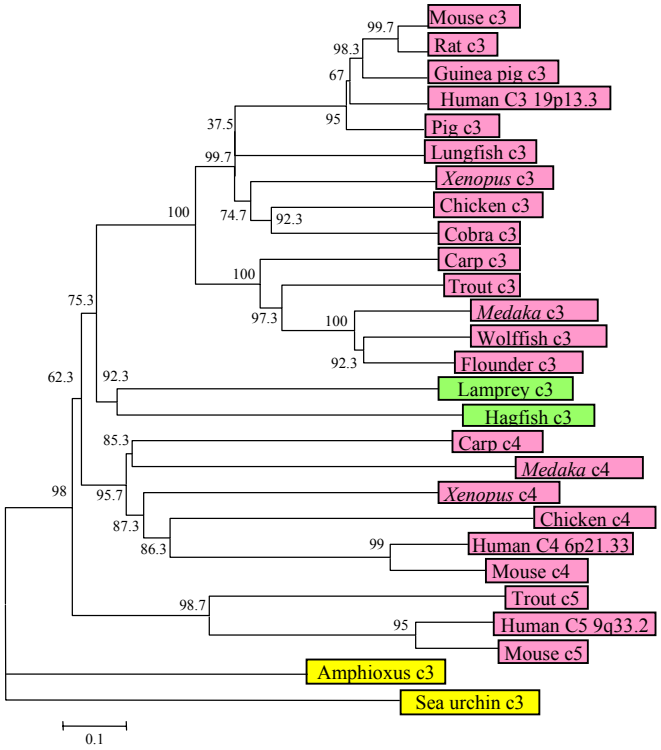


Figure 5.8 Phylogenetic analyses showing the relationship of the C4 paralogues and orthologues. The accession numbers of the proteins used to generate the trees are: human C4 (P01028), mouse c4 (P01029), chicken c4 (O73905), *Xenopus* c4 (Q91741), *Medaka* c4 (Q9IBG9), carp c4 (Q9I933), human C5 (P01031), mouse c5 (P06684), trout c5 (Q90XS7), human C3 (P01024), mouse c3 (P01027), chicken c3 (Q90633), *Xenopus* c3 (Q91588), rat c3 (P01026), guinea pig c3 (P12387), pig c3 (Q9GKP1), cobra c3 (Q01833), lungfish c3 (Q9W6G1), carp c3 (Q9YIB0), trout c3 (P98093), *Medaka* c3 (Q9IBH1), wolffish c3 (Q98TS6), flounder c3 (Q9PTY1), lamprey c3 (Q00685), hagfish c3 (P98094), amphioxus c3 (Q969A4) and sea urchin c3 (O44344).

5.3.5 Phylogenetic analysis of the RXR paralogous gene family

The retinoid X receptor beta, or RXRB, protein is a retinoid receptor and belongs to the steroid/thyroid hormone receptor superfamily of transcriptional regulators (Mangelsdorf *et al*, 1992). Retinoid receptors are soluble nuclear proteins that fall into

two classes: retinoic acid receptors (RAR) and retinoid X receptors (RXR). The RXR subfamily consists of three polypeptide chains, namely alpha, beta and gamma, encoded by separate loci. The three loci encoding the alpha (RXRA), beta (RXRB) and gamma (RXRG) proteins are located on chromosomes 9, 6 and 1, respectively.

The RXR phylogenetic tree was rooted with the *Drosophila* orthologue, usp (figure 5.9).

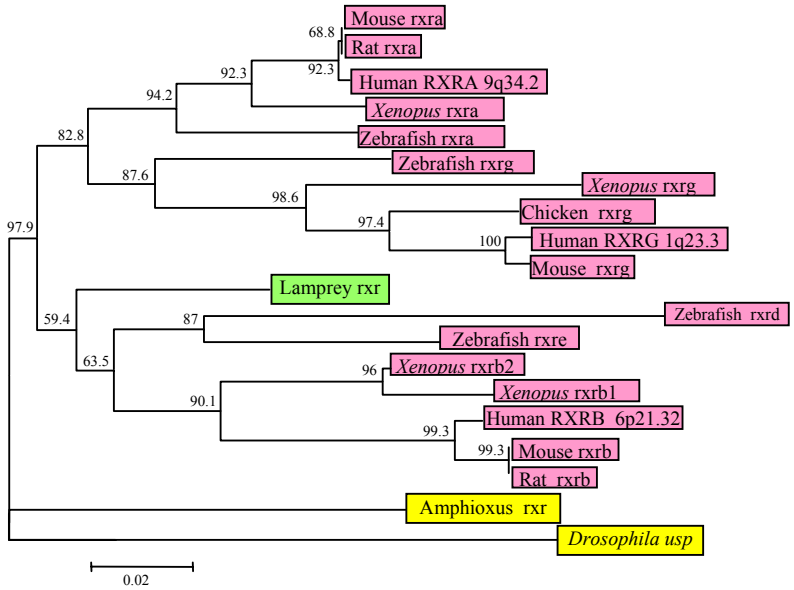


Figure 5.9 Phylogenetic tree showing the evolutionary relationship between the RXRB paralogues and orthologues. The accession numbers are: X52773 (human RXRA), X63522 (human RXRB), U38480 (human RXRG), M84817 (mouse rxra), M84818 (mouse rxrb), M84819 (mouse rxrg), L06482 (rat rxra), M81766 (rat rxrb), X58997 (chicken rxrb), L11446 (*Xenopus* rxra), X87366 (*Xenopus* rxrb2), S73269 (*Xenopus* rxrb1), L11443 (*Xenopus* rxrg), U29940 (zebrafish rxra), U29894 (zebrafish rxrg), U29941 (zebrafish rxrd), U29942 (zebrafish rxre), AF316878 (lamprey rxr), AF391296/5 (amphioxus rxr) and P20153 (*Drosophila* usp).

The results show that the human paralogues cluster, as expected, with equivalent orthologues. The RXR orthologues of the invertebrate species, *Drosophila* and

amphioxus, both fall outside all of the vertebrate genes. However, the RXR orthologue of the invertebrate lamprey clusters with vertebrate RXRB. This indicates that RXRB diverged first, after the cephalochordate split prior to the divergence of jawless fish. This was followed by a duplication event between RXRA and RXRG after the divergence of jawless fish. The zebrafish RXRD and RXRE genes resulted from a duplication occurring around the time of teleost/mammalian divergence. The topology of the tree, (RXRB(RXRA, RXRG)), clearly supports at least one round of large-scale duplication but it is possible that the RXR family arose by two-rounds of large-scale duplication events and one paralogue has been lost over time. Thus, the present day topology is (RXRB (RXRA, RXRG)).

5.3.6 Phylogenetic analysis of the tenascin paralogous gene family

The tenascin proteins are a family of extracellular matrix proteins (ECM) (for a review see Erickson, 1993). The Tenascin X (TNX) gene is located within the MHC class III region overlapping the CYP21A2 and C4 genes. Two paralogues, tenascin C (TNC, cytoactin, hexabrachion) and tenascin R (TNR, restrictin) have been identified in the paralogous regions on chromosomes 9q33.1 and 1q24.1, respectively. Tenascin orthologues have been identified in a range of vertebrates but only one invertebrate (summarised in the legend of figure 5.10). The Drosophila protein, ten^{m}, contains the EGF-like and FN-III domains and is believed to be the closest relative of the vertebrate tenascins (Baumgartner *et al*, 1994). This has been used as the outgroup to root the phylogenetic tree presented in figure 5.10.

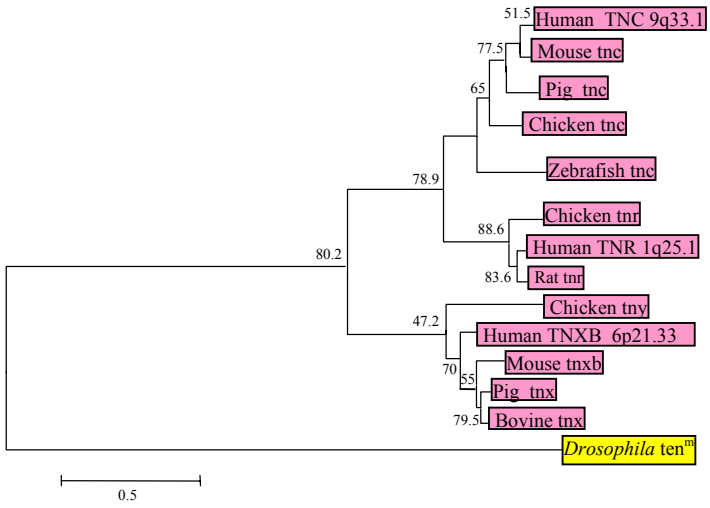


Figure 5.10 Phylogenetic analyses of the TNXB paralogues and orthologues. The accession numbers of the protein sequences used to generate this tree are as follows; P22105 (human TNXB), P24821 (human TNC), Q15568 (human TNR), O35452 (mouse tnxb), Q64706 (mouse tnc), Q05546 (rat tnr), P10039 (chicken tnc), Q00546 (chicken tnr), Q91008 (chicken tny), Q29038 (pig tnxb), Q29116 (pig tnc), O18977 (bovine tnxb) and Q24551 (*Drosophila* ten^{m}).

The topology of the tree strongly supports that TNXB diverged prior to the divergence of TNR and TNC as suggested by Katsanis and co-workers (1996) and Hughes (1998). Phylogenetic analysis shows that the TNC and TNR paralogues are most closely related and have arisen from a common ancestor. The clustering of the zebrafish TNC orthologue with the other TNC orthologous sequences indicates that the duplication which gave rise to the TNC and TNR paralogues occurred prior to the divergence of bony fish and tetrapods, approximately 450 million years ago. However, without the orthologous protein sequences of the key species (amphioxus, hagfish and lamprey) it cannot be determine whether the tenascin X gene supports the 2R hypothesis. Compelling evidence from the five other MHC paralogous gene families presented in sections 5.3.1-5.3.6 implies that these genes may have arisen via the same mechanism.

5.3.7 Phylogenetic analysis of the AIF paralogous gene family

The Allograft inflammatory factor 1 (AIF-1) gene was first isolated from activated macrophages in rat atherosclerotic allogenic heart grafts undergoing chronic transplant rejection (Utans *et al*, 1995). In humans, the full-length clone has been isolated and characterised (Autieri, 1996). Only one AIF-1 paralogue (AIF1-L) has been identified in the human genome (discussed in chapters 3 and 4) and is located in the chromosome 9 paralogous region.

The AIF1 encoded protein is evolutionarily well conserved within vertebrate species (Utans *et al*, 1996). To-date, it has been identified in seven vertebrates: humans, pig, rat, macaque, mouse, bovine, red sea bream and carp. It has only been identified in two invertebrates, the sea sponge and amphioxus. Phylogenetic analysis of the AIF1 paralogues and orthologues was carried out using the distantly related amino acid sequence from sea sponge as the outgroup (figure 5.11).

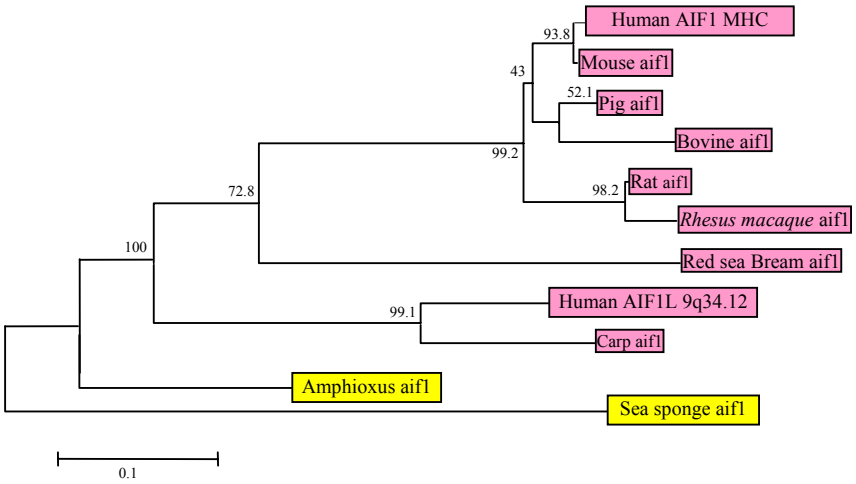


Figure 5.11 Phylogenetic tree of the AIF1 paralogues and orthologues. The species, and corresponding accession numbers given in parentheses, used to generate the tree are as follows: human AIF1 (P55008), human AIF1-L (Q9BQI0), pig aif1 (P81076), rat (P55009), *Rhesus macaque* aif1 (Q9GMH2), mouse aif1 (070200), bovine aif1 (Q9BDK2), red sea bream aif1 (Q9YI94), carp aif1 (O93246), sea sponge aif1 (Q966Y8) and aif1 amphioxus (translated from EMBL entry AU234552).

Evidence provided by the phylogenetic tree indicates that the duplication event involving the ancestral gene of the two AIF1 paralogues occurred prior to divergence of bony fish and post-dates the divergence of Amphioxus. Thus, this analysis supports at least one round of duplication prior to vertebrate emergence, or the 1R hypothesis. Interestingly, the AIF1 protein in carp clusters with AIF1-L in the phylogenetic tree suggesting that the carp AIF1 protein may actually be the orthologue of the human AIF1-L gene on chromosome 9. A sequence similarity search using the AIF1-L protein did not identify any orthologous sequences previously not identified for AIF1. To-date, a second AIF1 orthologue has not been identified in the carp genome to confirm that this is the true AIF1L orthologue. In summary, the AIF1 paralogous gene family have occurred via a large-scale duplication after the divergence of the cephalochordate lineage prior to the emergence of bony fish. Thus, supporting one round of large-scale duplication, or the 1R hypothesis.

5.3.8 Phylogenetic analysis of the β-tubulin paralogous gene family

The β-tubulins form the basic building blocks of the microtubulins when they form heterodimers with α-tubulins (reviewed by McKean *et al*, 2001). Microtubulins constitute a major component of the cytoskeleton in eukaryotic cells and are involved in essential processes, including cell division and intracellular transport. The survey of the human genome revealed seven paralogues of the TUBB gene scattered throughout the genome. The paralogues share very high sequence similarity, ranging from 72.9% to 99.6% at the protein level. The high level of similarity has resulted in the mis-annotation of these genes, i.e. the same SWISSPROT or SPTREMBL accession number has been given as the encoded protein sequence for multiple genes. In order to

prevent confusion, the corresponding ENSEMBL accession numbers is given in table 5.2 that was identified in the genome survey.

Table 5.2 Summary of the TUBB paralogues in the human genome

Gene	***Locus***	***Genomic clone accession number***	***ENSEMBL gene ID***	***No. of amino acids***
TUBB	6p21.3	AB023051	ENSG00000137379	444
TUBBL1	6p25.2	AL031963	ENSG00000137267	445
TUBBL2	6p25.2	AL445309	ENSG00000137285	445
TUBB4QL	10p15.3	AL713922	ENSG00000173876	444
TUBB4	16q24.3	AC0092143	ENSG00000141037	442
TUBBL	18p11.3	AP001005	ENSG00000173213	433
TUBB5	19p13.3	AC010503	ENSG00000104833	444
TUBB1	20q13.3	AC109840	ENSG00000101162	451

The β-tubulin genes are extensively conserved evolutionarily, but the number of encoding genes varies dramatically among species (Lewis and Cowan, 1990). A search of the protein databases, SWISSPROT and SPTREMBL, revealed several vertebrate β-tubulin proteins; one chimpanzee, one squirrel monkey, one rhesus macaque, one baboon, two mouse, one rat, two chicken and three *Xenopus* β-tubulin proteins. In addition, seven invertebrate β-tubulin proteins were extracted from the database; two sea squirt, one sea urchin, two *Drosophila*, one *C.elegans* and one *C.briggsae*.

Phylogenetic analysis using the protein sequences encoded by the TUBB paralogues reveals evidence in support of a number of duplication events (figures 5.12 and 5.13). The phylogenetic tree presented in figure 5.12 reveals two main gene clusters; one including only new world monkey and human sequences and the other, also containing human sequences, including a number of more ‘ancient’ species, namely the sea squirt and sea urchin clustered with vertebrates.

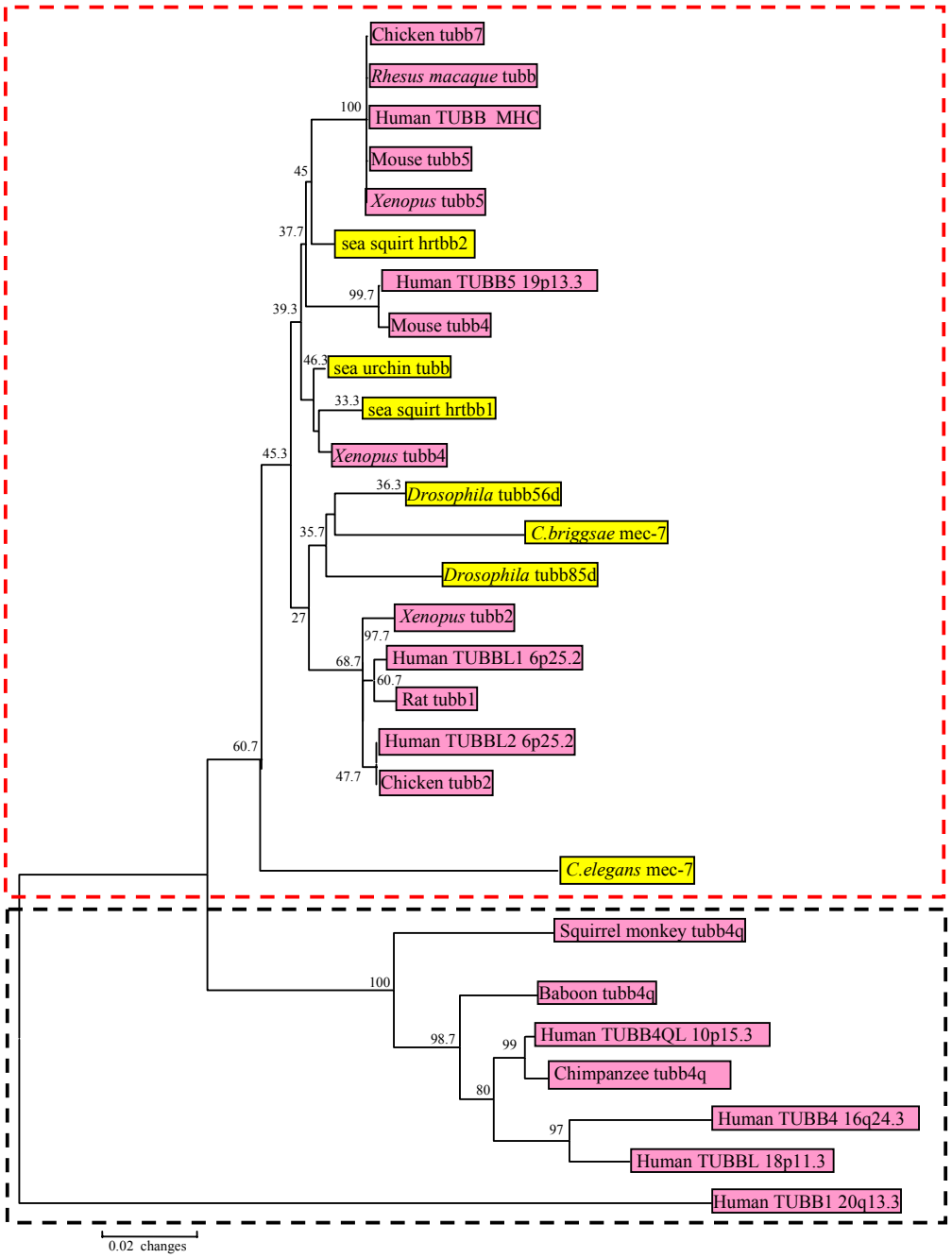


Figure 5.12 Phylogenetic analysis of the β-tubulin paralogues and orthologues. The two major species clusters corresponding to ancient duplication events (surrounded by a dashed red line) and more recent duplications (black dashed line). The accession numbers of the protein sequences used to generate this tree are: Q8WP14 (tubb4q, chimpanzee), Q8WP12 (tubb4q, squirrel monkey), AAD33992 (tubb4q, *Rhesus macaque*), Q8WP13 (tubb4q, baboon), Q9D6F9 (tubb4, mouse), P05218 (tubb5, mouse), P04691 (tubb1, rat), P32882 (tubb2, chicken) P09244 (tubb7, chicken), P13602 (tubb2, *Xenopus*), P30883 (tubb4, *Xenopus*), Q91575 (tubb5, *Xenopus*), O18343 (hrtbb2, sea squirt), O18342 (hrtbb1, sea squirt), P11833 (tubb, sea urchin), Q24560 (tubb56d, *Drosophila*), P08840 (tubb85d, *Drosophila*), P12456 (mec-7, *C.elegans*) and Q17299 (mec-7, *C.briggsae*).

The cluster containing the new world monkey β-tubulin proteins indicates that some of the tubulin paralogues in the human genome are the result of recent duplication events. This is supported by the analysis of the TUBB4Q pseudogene on 4q35.2 and related paralogues and orthologues by van Geel and co-workers (2002). Analysis of the human chromosomal segment, 4q35, containing the TUBB4Q pseudogene has indicated a substantial amount of duplication throughout the genome (Grewal *et al*, 1999; van Geel *et al*, 1999). Van Geel and colleagues (2002) revealed that this segment has undergone a number of duplications at different time points within the last 25 million years of catarrhine (New World Monkeys and humans) evolution.

The phylogenetic tree presented in figure 5.13 reveals evidence of ancient duplication events which occurred earlier than those proposed by the 2R hypothesis. The timings of the two rounds of duplication are proposed as follows; the first round of duplication occurred prior to the divergence of sea squirt and sea urchin and the second, after their divergence, prior to vertebrate emergence. The β-tubulin paralogues have been involved in a much earlier round of duplication followed by further duplications; this is supported by the clustering of the sea squirt and sea urchin orthologues with the mammalian counterparts.

There is evidence of a more recent duplication event telomeric to the MHC region on chromosome 6. The two newly identified paralogues, termed TUBBL1 and TUBBL2, on 6p25.2 share identical exon fingerprints and have 99.6% protein sequence similarity. The two paralogues are located within 70 kb of each other and span approximately 3.4 kb. They appear to have arisen by a tandem duplication event after amphibian divergence and prior to the emergence of rodents (the orthologue of TUBBL2 in rat may not be functional or the dataset used to generate the tree may not be complete).

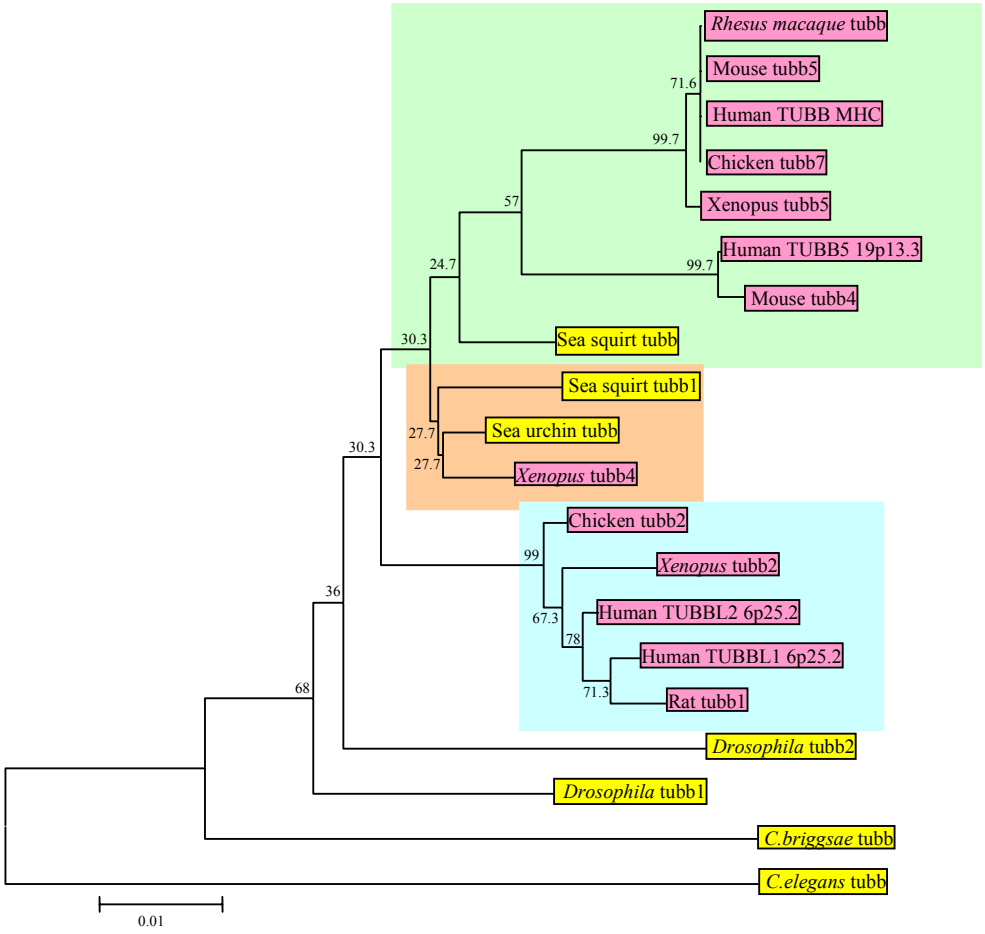


Figure 5.13 Phylogenetic tree showing the ancient duplication events that have shaped the present day β-tubulin paralogues and orthologues. The three main groups are highlighted in different colours. The protein accession numbers are as described in figure 5.12.

Analysis of the distribution of β-tubulin paralogues in the human genome (Chapter 4) reveals a strong positional bias towards the pericentromeric and subtelomeric regions of the genome. It is known that frequent exchange of sequences occurs between these dynamic chromosome regions (Eichler *et al*, 1996; Pryde *et al*, 1997; Eichler, 1998; IHGSC, 2001), which can result in the acquisition of new genes as well as genetic diversity. There is evidence to suggest that the TUBB4Q pseudogene on 4q35.2 was

once a functional gene and, because of its proclivity to duplicate to subtelomeric locations, a novel tubulin member was transposed to 10p15.3 (TUBB4QL) approximately 7.3 MYA (van Geel *et al*, 2002). It has also been suggested that GC-rich repeat elements play a direct role in the pericentromeric localisation of intra- and interchromosomal duplication events (Eichler *et al*, 1999). It would be interesting to investigate whether there are GC-rich repeat elements bordering the duplicated segments containing the TUBB paralogues but this is beyond the scope of this thesis.

5.3.9 Phylogenetic analysis of the GPX paralogous gene family

The glutathione peroxidase proteins (GPX) are enzymes involved in the protection of the cell against oxidative damage. Using glutathione as the reducing agent they metabolise hydroperoxides generated by normal oxidative metabolism which otherwise would have deleterious effects, mainly on cell-wall integrity (Dufaure *et al*, 1996) phylogenetic relationship of the GPX5 paralogues and orthologues is shown in figure 5.14.

The human GPX protein sequences were aligned with the protein sequences obtained from a range of vertebrates and a single invertebrate species, *Suberites domuncula* also known as the sea sponge (see legend of figure 5.14). In total, one true paralogue of GPX5, named GPX3, was identified in the genome survey on 5q33.1. Two putative paralogues were also identified on chromosomes 3p21.31, GPX1, and GPX3, 5q33.1 but have a very different exon structure compared with GPX5. In addition, another member of the glutathione peroxidase family, GPX4, has been identified in the paralogous region on 19p13.3 but was not identified as a paralogue in this analysis.

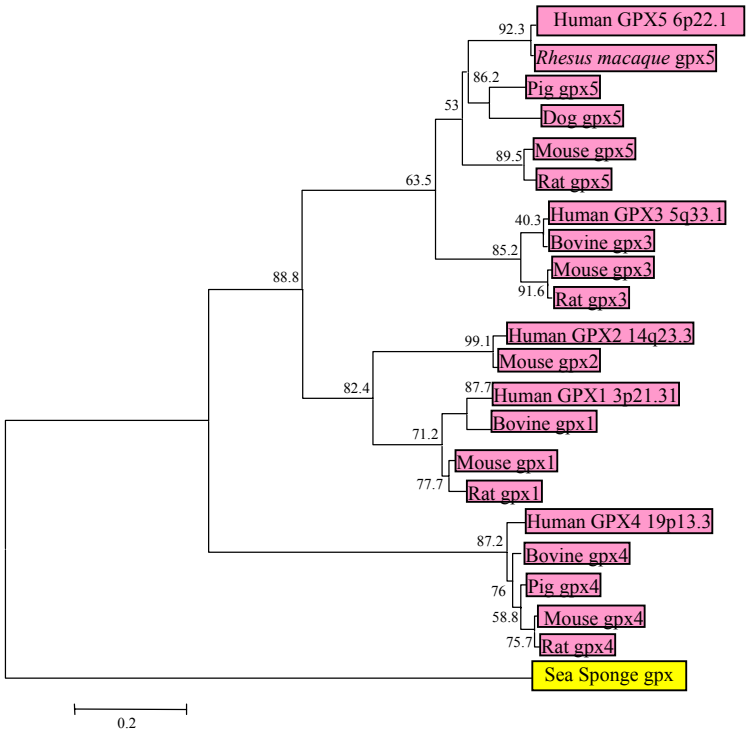


Figure 5.14 Phylogenetic analysis of the GPX family. The protein sequences used, with the corresponding accession numbers given in parentheses, were as follows; human GPX5 (O75715), human GPX3 (P22352), human GPX1 (P18283), human GPX2 (P18283), human GPX4 (P36969), mouse gpx5 (P21765), mouse gpx3 (P46412), mouse gpx1 (P11352), mouse gpx2 (Q9JHC0), mouse gpx4 (O70325), pig gpx5 (O18994), pig gpx4 (P36968), rat gpx5 (P30710), rat gpx3 (P23704), rat gpx1 (P04041), rat gpx4 (P36970), bovine gpx3 (P37141), bovine gpx1 (P00435), bovine gpx4 (Q9N2N2), dog gpx5 (O46607), *Rhesus macaque* gpx5 (P28714) and sea sponge gpx (Q966Y9).

Phylogenetic analysis shows that GPX4 is the most distantly related member. This is to be expected as it has a very different exon structure than the other GPX family members and shares less than 30% sequence identity. The GPX genes with identical exon fingerprints and highest protein sequence identity, GPX5-GPX3 and GPX1-GPX2, cluster together indicating that the genes have descended from a common ancestor. The duplications occurred after sea sponge divergence but prior to rodent divergence thus could have resulted from two rounds (or more) of whole-genome

duplication. More data is needed to determine the precise times of the duplication events.

5.3.10 Phylogenetic analysis of the CLIC paralogous gene family

The chloride intracellular channel (CLIC) paralogous gene family encode for chloride channels, which are involved in chloride ion transport within various subcellular compartments (reviewed by Jentsch *et al*, 2002). Phylogenetic analysis of the CLIC family suggests a series of successive duplication events including at least 2 rounds of whole-genome duplications (figure 5.15).

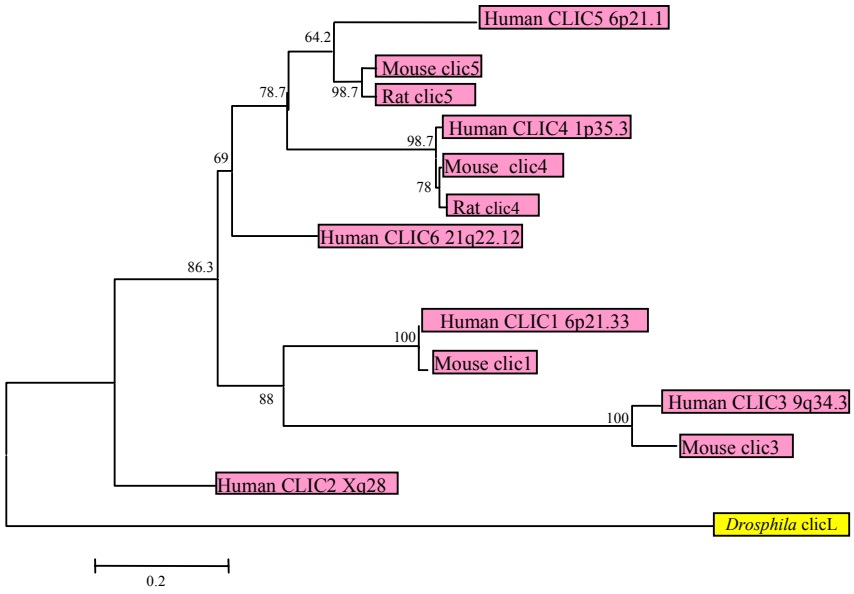


Figure 5.15 Phylogenetic analysis of the CLIC family. The protein sequences used, with the corresponding accession numbers given in parentheses, were as follows; human CLIC1 (O00299), human CLIC2 (O15247), human CLIC3 (O95833), human CLIC4 (Q9Y696), human CLIC5 (Q9NZA1), human CLIC6 (Q9NY7), mouse clic1 (Q9Z1Q5), mouse clic3 (Q9D7P7), mouse clic4 (Q9QYB1), mouse clic5 (Q9CYD1), rat clic4 (Q9Z0W7), rat clic5 (Q9EPT8), *Drosophila* clicL (NM_132700).

The topology of the tree indicates that the Xq28 CLIC2 gene diverged first. Following the divergence of CLIC2 there were two rounds of whole-genome duplication. The first duplication event resulted in the CLIC (MHC,9q34.3) and the CLIC (21q22.12(1p35.3,6p21.1)) gene precursors. This was followed by a second round of whole-genome duplication resulting in CLIC1 (MHC), CLIC3 (9q34.3), CLIC6 (21q22.12) and the CLIC (1p35.3, 6p21.1) precursor (possibly located on 19p13). A further segmental duplication event occurred resulting in the present day location of the CLIC4 and CLIC5 genes on 1p35.3 and 6p21.1, respectively. The latter segmental duplication is supported by the observation that there has been a large-scale triplication involving the chromosomal regions 1p35, 6p21.1 and 21q22.12 (Strippoli *et al*, 2002). Strippoli and colleagues (2002) identified a large (approximately 500 kb) segment on human chromosome 21q22 that is triplicated on chromosomes 1p35 and 6p12-p21. The region on chromosome 21 contains the CLIC6 gene, along with two other genes, DSCR1 and AML1, which have functional copies in the other regions. The gene order within these regions, termed the ACD clusters, is identical and it was suggested that the triplication occurred by segmental duplication as part of the genome-duplication events before the divergence of tetrapods and teleosts. However, more sequence data is needed to confirm this prediction and to fully understand the complex history of this paralogous gene family.

5.4 Discussion

The evolutionary histories of 10 MHC paralogous gene families have been reconstructed using phylogenetic trees. Analysis of the topologies of the trees and the arrangement of the paralogues and orthologues has revealed that the evolution of the MHC paralogues is complex. What is evident is that gene duplication has played a major role in the evolution of these gene families. In particular, there is evidence in support of the 2R hypothesis. The 2R hypothesis proposes that the genome evolved via two rounds of whole-genome duplication events early in the vertebrate lineage; one occurring after amphioxus divergence, prior to the emergence of hagfish and lamprey and the second just after. In order to support the 2R hypothesis, phylogenetic analyses of gene families should meet the following criteria: (a) the vertebrate members of the gene family can be shown to have duplicated within the vertebrate lineage and (b) the gene family phylogenies show the (A,B)(C,D) topology.

The 2R hypothesis would give rise to four copies of an ancestral gene therefore this is best exemplified by the paralogous gene families with four members in the human genome. The BRD, NOTCH and PBX paralogous gene families all have four paralogues, including the MHC locus, in the human genome. The topology of the three phylogenetic trees support the 2R hypothesis, showing the (A,B)(C,D) topology. Furthermore, some of the orthologous genes have been identified in the three key organisms, amphioxus, lamprey and hagfish, and the positions of these organisms in the phylogenetic trees are in support of the timings of the duplication events proposed by the 2R hypothesis. Thus, if the sequences are available, a single amphioxus orthologue is positioned at the base of each tree and at least one hagfish or lamprey orthologue clusters with the mammalian counterparts.

The paralogous gene families with three members also support the 2R hypothesis, albeit accompanied by gene loss. This appears particularly likely as extensive gene loss has been shown to take place after gene duplication events (Gu and Huang, 2002). Furthermore, the paralogous gene family with only two members is also in support of at least one round of genome duplication in the vertebrate history (the 1R hypothesis). Alternatively, it also supports the 2R hypothesis accompanied with the loss of two genes. The timings of the duplication events as suggested by the 2R hypothesis are also supported by the clustering of the 'key' organisms in these phylogenetic trees.

Ideally, if the paralogues emerged simultaneously by block or whole-genome duplication, the genes from the same chromosomal regions should cluster together on the tree. For example, previously published phylogenetic analyses of three paralogous gene families indicated that the paralogous regions on 1q21-q25 and 9q33-q34 were most related (Katsanis *et al*, 1996; Kasahara, 1997; Hughes, 1998). It would therefore be expected that the paralogues on chromosomes 1 and 9 will cluster and 6 and 19 will also cluster. However, this is not the case. The NOTCH and PBX paralogous gene families support this clustering however the BRD paralogues do not; with the BRD paralogues on chromosomes 6 and 9, and, 1 and 19 clustering. Since the construction of phylogenetic trees utilises sequence information the different rates by which the sequences of the paralogues have evolved since duplication, dictated by the evolutionary pressures acting upon them, may explain why different sets of paralogues cluster.

Phylogenetic analysis of the MHC paralogous gene families with five or more members revealed that the evolution of the MHC paralogues involved more than just

the two rounds of large-scale duplication events proposed by the 2R hypothesis. This is exemplified by the β-tubulin family, which shows evidence of both ancient duplication events, dated prior to the divergence of sea squirt and sea urchin prior to the emergence of amphioxus, as well as much more recent duplications. This is in concordance with previously published phylogenetic studies of the MHC paralogous gene families (Endo *et al*, 1997; Hughes; 1998). These studies revealed that duplication events of multigene families, such as the proteasome component (PSMB) genes, occurred much earlier than those proposed by the 2R hypothesis.

In conclusion, there is strong evidence that some MHC paralogues evolved via the mechanism proposed by the 2R hypothesis but that others have emerged by independent means. Therefore, there could still be a selective advantage, potentially related to function, for these genes to have been brought together and remained clustered.

Chapter 6

Expression analysis of extended MHC paralogous gene families

6.1 Introduction

The MHC paralogous genes identified in the human genome (presented in chapters 3 and 4) have arisen by duplication (discussed in chapter 5). Duplication results in new genes and, if they are duplicated in their entirety (including the regulatory elements) there will be some inter-gene redundancy, with the two paralogues being able to fulfil the same function. In principle, the genetic redundancy created by duplication will allow evolutionary experimentation; since only one copy is required to maintain the function provided by the single, ancestral gene the other copy is free to diverge. Thus, one of the duplicate genes is left under purifying selection (selection against deleterious alleles) and therefore maintains the original function of the ancestral gene and the other duplicate gene is freed from all functional constraints to diverge.

The classical model, originally proposed by Ohno in 1970, predicts two potential fates for the 'other' duplicate gene. The most likely fate is that it will degenerate into a pseudogene or will be lost from the genome altogether, due to locus deletions or point mutations, by a process called non-functionalisation (figure 6.1.A). The less frequently expected outcome is that the duplicated gene acquires mutations that modify either the expression pattern of the gene or the function of the encoded protein in an advantageous way. The novel allele could then become fixed in the population, exposing the formerly redundant gene to new and distinct selective constraints in a

process known as neo-functionalisation (figure 6.1.B).

It is believed that neo-functionalisation is rare and that few duplicates will be retained in the genome (reviewed by Prince and Pickett, 2002). However, analysis of the human genome has revealed that at least 15% of human genes are duplicates (Li *et al*, 2001) and that segmental duplications cover approximately 10% of the genome (IHGSC, 2001; Bailey *et al*, 2002). In order to explain the preservation of duplicate genes in the genome, the sub-functionalisation model has been proposed (figure 6.1.C; Force *et al*, 1999; Lynch and Force, 2000). Sub-functionalisation proposes that, after duplication, the two duplicate gene copies acquire complementary loss-of function mutations. The two genes therefore develop independent functions, and are both required to produce the full complement of functions of the ancestral gene.

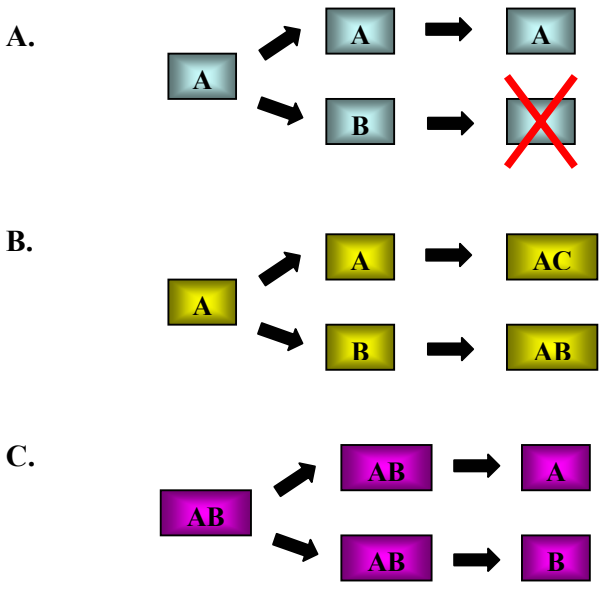


Figure 6.1 Fates of duplicated genes (adapted from Mazet and Shimeld, 2002). (A) non-functionalisation, in which one copy degenerates after duplication, (B) neo-functionalisation, when initially identical duplicates with function *A* diverge by acquiring new functions *B* and *C* and (C) sub-functionalisation, in which duplicate genes with multiple functions *A* and *B* diverge by reciprocal loss.

The process or mechanism by which the MHC paralogous gene families have evolved since duplication is not known. However, what is clear is that their emergence by gene duplication created genetic redundancy. It is therefore interesting to determine the present-day function(s) of the paralogues in order to gain some understanding of the mechanism(s) by which they have evolved. The first step to understanding the function and phenotype of the genes and the corresponding proteins is to generate the expression profile of the human paralogues in a range of normal human tissues.

Each tissue in the human body is different from another because of the synthesis of a distinct set of RNA molecules. The proportion of the genes expressed as mature messenger RNA (mRNA), collectively known as the transcriptome, represent only a small part of the human genome. Messenger RNA (mRNA) represents approximately 2.5% of the RNA in a cell, with ribosomal RNA (rRNA) and transfer RNA (tRNA) making up 75% and 10% respectively (Jackson *et al*, 2000). The remainder is made up of RNA molecules such as small nuclear RNAs (snRNAs) and small nucleolar RNAs (snoRNAs). The analysis of the transcriptome can provide many clues to the functional significance of a particular gene. For example, the presence of an RNA transcript in one specific tissue and absence in all others would suggest a specialised function of the gene in that tissue. Therefore, by generating a comprehensive expression profile in a range of human tissues we can discover whether the paralogues have similar or divergent functions to ultimately understand how the paralogues have evolved since their emergence by duplication.

This chapter focuses on the characterisation of 40 MHC paralogues, corresponding to the 10 MHC paralogous gene families discussed in chapter 5 (see section 5.2 for more detail), in a range of normal human tissues and cell-lines using different approaches.

6.2 Terminology

In total, five different methods were used in this project to obtain a comprehensive profile of the expression of 40 paralogues in a range of human tissues and cell-lines; these were *In-silico*, Northern blot, Dot-blot, RT-PCR and microarray analysis. Each method will be discussed individually within the results section. One point to note is the use of the terms 'probe' and 'target' when referring to the hybridisation methods. In this chapter the terms 'probe' and 'target' have been used to describe elements in both the blotting (namely Southern, Northern and Dot blots) and microarray experiments. In the case of the blotting methods the 'target' is referred to as either the DNA or RNA attached to the membrane and the 'probe' is the free nucleic acid which is labelled and used to hybridise to the blot.

The microarray experiments were divided into two phases; I and II. Phase I corresponds to the cross-hybridisation (control) experiments using the 'Paralogue Microarray' (as described in section 2.13.1) and phase II refers to the expression profiling experiments using the '10K/Paralogue Microarray' (as described in section 2.13.1). In the phase I microarray experiments, the paralogue specific PCR products represent both the 'target' and the 'probe', as they are attached to the surface of the 'Paralogue Microarray' and used to hybridise with the array. In the case of the phase II experiments using the '10K/Paralogue Microarray', the 'probe' is the free labelled nucleic acid used to hybridise with the array, i.e. the complementary DNA of the RNA either extracted from the cell-line (as described in section 2.10) or purchased from Ambion. The 'target' is the DNA attached to the '10K/Paralogue Microarray', and corresponds to the paralogue specific PCR products generated for each of the 40 genes and the DNA elements already on the standard Sanger Institute 10K microarray.

6.3 Results

6.3.1 Cross-hybridisation (control) experiments

The potential for cross-hybridisation needs to be considered when working with paralogous genes and proteins. It has been shown that gene targets with 77-100% sequence identity cross-hybridise in hybridisation experiments using nylon membranes (Vernier *et al*, 1996) and over 80% in glass cDNA microarray experiments (Evertsz *et al*, 2001). Although the primers for each paralogue were designed to amplify a paralogue specific PCR product it was still essential to ensure that they did not demonstrate any cross-reactivity. Two methods were used to verify that they were paralogue specific.

First, the paralogue specific PCR products were arrayed, or printed, onto the 'Paralogue Microarray' in triplicate (the primers used to amplify the PCR products are summarised in Appendix 4). The same PCR products were also fluorescently labelled (as described in section 2.13.3) and hybridised to the array (as described in section 2.13.4). If the probes were specific to the particular paralogue they did not cross-hybridise with other members of the same paralogous gene family. This was detected upon scanning of the array after hybridisation. The probes corresponding to the 10 extended MHC genes were labelled and individually hybridised to the 'Paralogue Microarray'. In addition, the 10 probes were pooled and used to hybridise to the array (this is presented in figure 6.2.A and 6.2.B).

Secondly, the paralogue specific PCR products were hybridised to Southern blots to ensure that there was only a single copy in the genome (the primers used to amplify the PCR products are summarised in Appendix 3). Southern blots were made (as

described in section 2.14.3) by digesting human genomic DNA with three different restriction endonucleases, *HIND*III, *Pst*I and *BamH*I. Restriction endonucleases are enzymes that bind to a DNA molecule at a specific sequence and make a double-stranded cut at or near to that sequence, resulting in restriction fragments of genomic DNA. After treatment with the restriction endonucleases, the resulting fragments were examined by agarose electrophoresis to determine their size. When the digested genomic DNA was run on a gel it appeared as a smear because there were DNA fragments of every possible length merged together (data not shown). The restriction fragments from the agarose gel were then transferred from the agarose gel to a nylon membrane and fixed by UV irradiation. This process resulted in the DNA bands becoming immobilised in the same relative positions on the surface of the membrane, and is referred to as the target.

The hybridisation probe was prepared by radioactively labelling the paralogue specific PCR product as described in section 2.14.1, which was then verified as described in section 2.14.2. The Southern blots were probed using the radioactively labelled paralogue specific PCR products as described in section 2.14.4. The sequence of the labelled DNA molecule was complementary to the target DNA, therefore they hybridised. The position of the hybridised probe on the membrane was identified by detecting the signal given out by the label attached to the probe. The signal was detected by autoradiography. If the probes were specific to a particular paralogue only a single band was seen on the autoradiograph corresponding to the restriction fragment that hybridises to the probe and which contains the paralogue of interest (figure 6.2.C).

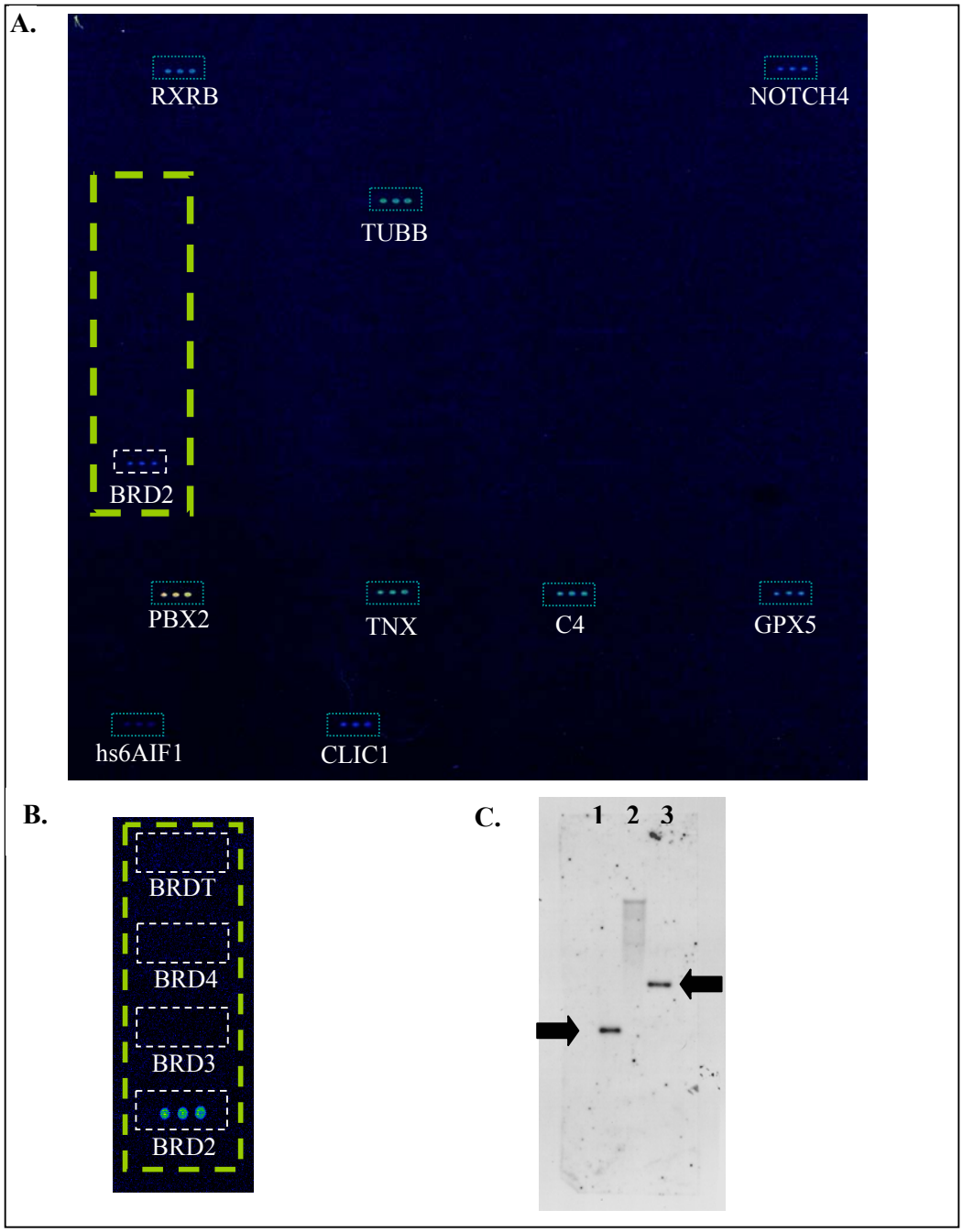


Figure 6.2 Verification of probe specificity. (A) represents the results of the 'Paralogue Microarray' hybridisation using the 10 pooled probes. The corresponding probes and targets have hybridised and the spots (in triplicate) are visible (they are boxed and the name of the extended MHC gene given). The area highlighted by the yellow dashed lines is shown in more detail in (B). Within this region there are spots corresponding to the four members of the BRD paralogous gene family. The BRD2 probe was fluorescently labelled and hybridised to the array. The probe only hybridises to the BRD2 target and not to the three paralogues, indicating that the probe is specific to the BRD2 gene. (C) Southern Blot analysis using the BRD2 probe confirms that it is specific to the BRD2 gene. There is only a single band in lanes 1 and 3 (indicated with arrows), which contain genomic DNA digested with the restriction enzymes *HIND*III and *BamH*I, respectively. The smear in lane 2 indicates that the digest with the restriction enzyme *Pst*I was not successful.

6.3.2 Expression profiling

There are many ways in which to study the expression pattern of a gene. Classical techniques, such as Northern blotting, can be used to discover the expression profile on a low-throughput scale, while microarrays can be used to give a high-throughput analysis of gene expression. In total, five different methods were used to study the expression profile; *In-silico*, Dot-blotting, Northern Blotting, RT-PCR and Microarrays.

6.3.2.1 *In-silico* analysis

The aim of the human genome project was to produce a complete and accurate sequence of the entire genetic material. It was realised that the transcriptome was the information of most interest to scientists and this was addressed in part by the EST sequencing project. ESTs are Expressed Sequence Tags, which are short single-pass DNA sequences obtained from either end of complementary DNA (cDNA) clones. These ESTs are derived from a vast number of cDNA libraries obtained from different tissues, and species. Complementary DNA is prepared by converting an mRNA preparation into double-stranded DNA. Because the mRNA in a cell is derived from protein-coding genes, cDNAs and the ESTs obtained from them represent the genes that were being expressed in the cell from which the mRNA was prepared. ESTs are looked upon as a rapid means of gaining access to the sequences of important genes, and they are valuable even if their sequences are incomplete. They are also very useful for the preliminary analysis of gene expression in different tissues or pathological states. As this analysis is performed solely using computational techniques it has been termed '*in-silico*'.

In silico analysis of EST data was performed as described in section 2.17. In summary, ESTs were retrieved from the UNIGENE cluster and by BLAST searching the EST database (dbEST) using the protein sequences. UNIGENE is an experimental system which automatically partitions the GENBANK sequences into non-redundant sets of gene-specific clusters. Each cluster contains sequences that represent a unique gene, as well as related information including the EST data. The EST data in the UNIGENE clusters is compiled using the EST database. In order to ensure I had the most comprehensive list of ESTs for each gene, the EST database was independently searched. However, in all cases no additional ESTs were identified. The ESTs were filtered in order to produce a non-redundant, unique set of ESTs for 36 MHC paralogues (summarised in Appendix 5).

Figure 6.3 summarises the results of the *in-silico* analysis of the BRD2 gene and the three paralogues, BRDT, BRD3 and BRD4. The expression profile of the BRD paralogous gene family was achieved in 59 different tissues corresponding to eight systems of the human body. In addition, the genes were all identified in pools of tissues that were categorised as mixed and in tissues of unknown sources, termed unknown. The transcript patterns of BRD2 and BRD3 have previously been determined in 43 human adult tissues and were found to be ubiquitously expressed (Thorpe *et al*, 1997). The expression profile using the EST data indicates that they are not ubiquitously expressed and have a more specialised transcript pattern.

One of the main advantages of EST data is that information is freely available for the majority of genes in the human genome in an array of tissues and cell-lines. Therefore, an extensive profile for a particular gene can be obtained relatively quickly. However, EST data has its limitations, including the types and sizes of the libraries available.

Figure 6.3 Summary of the results of the *in-silico* expression analysis of the BRD2 gene and its three paralogues. The tissues are divided into eight systems of the human body; nervous (red), cardiovascular (yellow), digestive (orange), genitourinary (blue), immune (purple), respiratory (green), secretory (pink) and muscle (grey). A black bar indicates that the gene is expressed in the tissue (i.e. there was one or more EST hits for the gene in the tissue). A white bar indicates that no EST was identified for a particular gene in the corresponding tissue; therefore there is no evidence of expression. Each tissue is separated by a horizontal grey line.

There is also a vast amount of redundancy within the EST libraries, which has been associated with the different rates and levels at which genes are expressed within various tissues, for example, in the UNIGENE dataset for the GPX3 gene there are 1335 ESTs compared with just 3 for the GPX5 gene. This implies that GPX3 is more highly expressed than GPX5, which is correct as GPX5 has a restricted expression Perry *et al*, 1992; Hall *et al*, 1998) whereas GPX3 is expressed in a range of tissues (Chu *et al*, 1992). In this thesis the EST data was used as a preliminary screen in order to determine in which tissues the genes were expressed and all findings were experimentally verified using a number of techniques.

6.3.2.2 Dot-blot analysis

The dot-blot, or Multiple Tissue Expression Array (MTE™ Array), enabled the accurate profile of gene expression over a range of human tissues and cancer cell-lines in one experiment. In total, 76 tissue-specific poly A^+ RNAs were spotted onto the nylon membrane, including 17 areas of the brain, seven regions of the heart and RNA from other major organs of the body (summarised in figure 6.4.D). The paralogue specific probes were amplified using the primers summarised in Appendix 3 and radioactively labelled as described in section 2.14.1. The activity and the amount of incorporation of radioactivity were verified as described in section 2.14.2. The labelled probes were then hybridised to the dot-blot for 16 hours, washed and exposed for up to 8 days (as described in sections 2.14.4 and 2.14.5). The results for the 37 MHC paralogues analysed are summarised in Appendix 6.

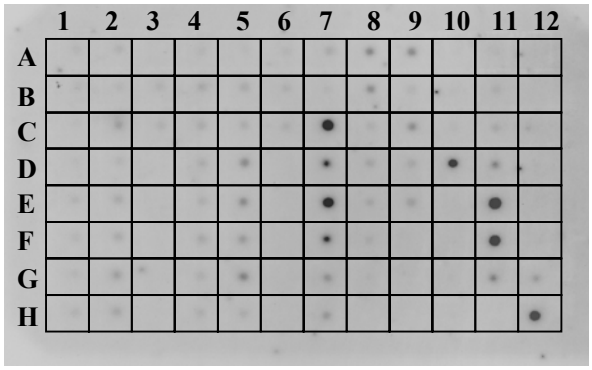
Figure 6.4 summarises the results of the dot-blot analysis of the Allograft inflammatory factor 1 gene (AIF1; figure 6.4.A) and the paralogue on 9q34.12 (AIF1-

L; figure 6.4.B). Both genes show expression in most tissues spotted on the blot, albeit with varying levels of expression. The AIF1 gene was first isolated from activated macrophages in rat atherosclerotic allogenic heart grafts undergoing chronic transplant rejection (Utans *et al*, 1995). Autieri (1996) showed that AIF1 was a cytokine-inducible, tissue-specific, and highly conserved transcript transiently expressed in response to vascular trauma. AIF1 is also known to be expressed in a variety of human tissues, with highest expression in tissues of lymphoid origin, in particular, spleen and thymus. This has been confirmed by my dot-blot analysis of the gene. Both paralogues, AIF1 and AIF1-L, are highly expressed in adult and foetal spleen suggesting an overlap (or redundancy) in function. However, only AIF1 is expressed in adult and foetal thymus showing functional divergence.

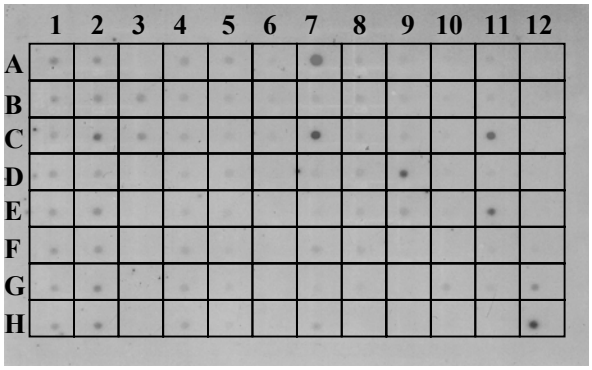
There are other examples of co-expression and divergence of expression. In particular, both are expressed in areas of the brain. AIF1-L is expressed in whole brain as well as the 17 different sections of the brain individually spotted on the blot, whereas AIF1 is more selective and is expressed in 10 areas of the brain but is not detected in whole brain. The overlap in expression suggests that the two paralogues maybe involved in the same pathway and, owing to redundancy, can perform the same function in certain parts or stages of that pathway. One of the other interesting findings of the dot-blot analysis is that AIF1-L is highly expressed in kidney whereas the expression of the AIF1 gene is very weak. This indicates that these paralogues also have divergent functions.

Figure 6.4 Transcription pattern of the AIF1 (A), AIF1L (B) and β-actin control (C) genes after hybridisation with paralogue-specific probes to the dot blot with RNA from different tissues. (D) Tissue key of the RNA dot-blot as supplied by the manufacturer (Clontech). The tissues shaded red indicate that both AIF1 and AIF1L were expressed in that tissue, blue shows only AIF1 was expressed, yellow indicates only AIF1L was expressed and white shows that neither gene were expressed in that tissue. Blots A and B were exposed for 3 days and blot C for 2 days.

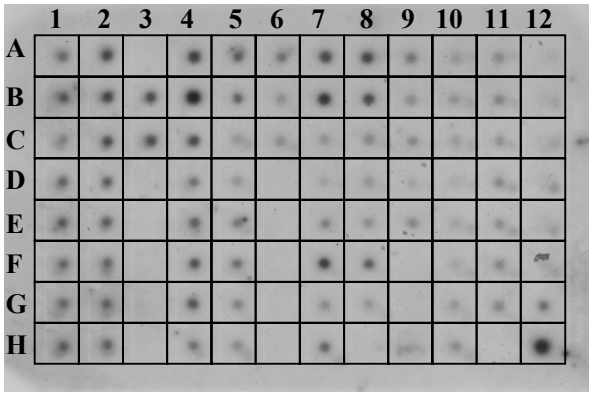
A. AIF1 (6p21.33)



B. AIF1L (9q34.12)



C. Control



D. Template

	1	2	3	4	5	6	7	8	9	10	11	12
A	Whole brain	Cerebellum, left		Heart	Oesophagus	Colon, transverse	kidney	lung	liver	Leukaemia, HL-60	Foetal brain	Yeast Total RNA
B	Cerebral cortex	Cerebellum, right	Accumbens	Aorta	Stomach	Colon, descending	Skeletal muscle	Placenta	Pancreas	HeLa S3	Foetal heart	Yeast tRNA
C	Frontal lobe	Corpus collosum	thalamus	Atrium, left	duodenum	rectum	Spleen	Bladder	Adrenal gland	Leukeamia, K562	Foetal kidney	E.coli rRNA
D	Parietal lobe	Amygdala		Atrium. Right	Jejunum		Thymus	Uterus	Thyroid gland	Leukaemia, MOLT4	Foetal liver	E.coli DNA
E	Occipital lobe	Caudate nucleus		Ventricle, left	Ileum		Peripheral blood leukocyte	Prostate	Salivary gland	Burkitt's lymphoma, Raji	Foetal spleen	Poly-(A)
F	Temporal lobe	Hippo-campus		Ventricle, right	Ilocecum		Lymph node	Testis		Burkitt's lymphoma, Daudi	Foetal thymus	Human Cot-1 DNA
G	* of cerebral cortex	Medulla oblongata		Intre-venricular septum	Appendix		Bone marrow	ovary		Colorectal adeno-carcinoma SW480	Foetal lung	Human DNA 100ng
H	Pons	Putamen		Apex of the heart	Colon, ascending		Trachea			Lung carcinoma A549		Human DNA 500ng

* paracentral gyrus

6.3.2.3 Northern blot analysis

Northern blotting can be used to determine the expression profile of a gene as well as identify the number of alternative splice variants. Alternative splicing is a widely occurring and important mechanism for controlling differential expression of cellular genes. The process changes the effect of a gene in different tissues and developmental states by generating distinct mRNA isoforms composed of different selections of exons, which produce variant proteins. This phenomenon is widespread in the human genome and it has been predicted that between 40-60% of human genes are alternatively spliced (Modrek and Lee, 2002).

Whilst studying the expression profile of the MHC paralogous genes it was important to understand how many splice variants that utilise the region amplified by the paralogue specific primers were expressed in a particular tissue. This was achieved using Multiple Tissue Northern (MTN™) blots purchased from Clontech, referred to as Northern blots throughout this chapter. The Northern blots used in this thesis enabled the assessment of the alternative splice forms, sizes (ranging from 0.5 to 10 kb) and relative abundance of the transcript in eight different normal human tissues (figure 6.5). The Northern blots were made using poly A^+ RNA extracted from the eight different normal human tissues.

The paralogue specific probes were used to hybridise to the Northern blots in order to assess how many splice variants were present in each tissue for 37 MHC paralogues (the tenascin paralogous gene family has been removed from the analysis, see section 6.3.3.1). The primers used and the results for the paralogues analysed are summarised in Appendix 3 and 7 respectively. The results of the Northern blot analyses for the four members of the BRD paralogous gene family is shown in figure 6.5.

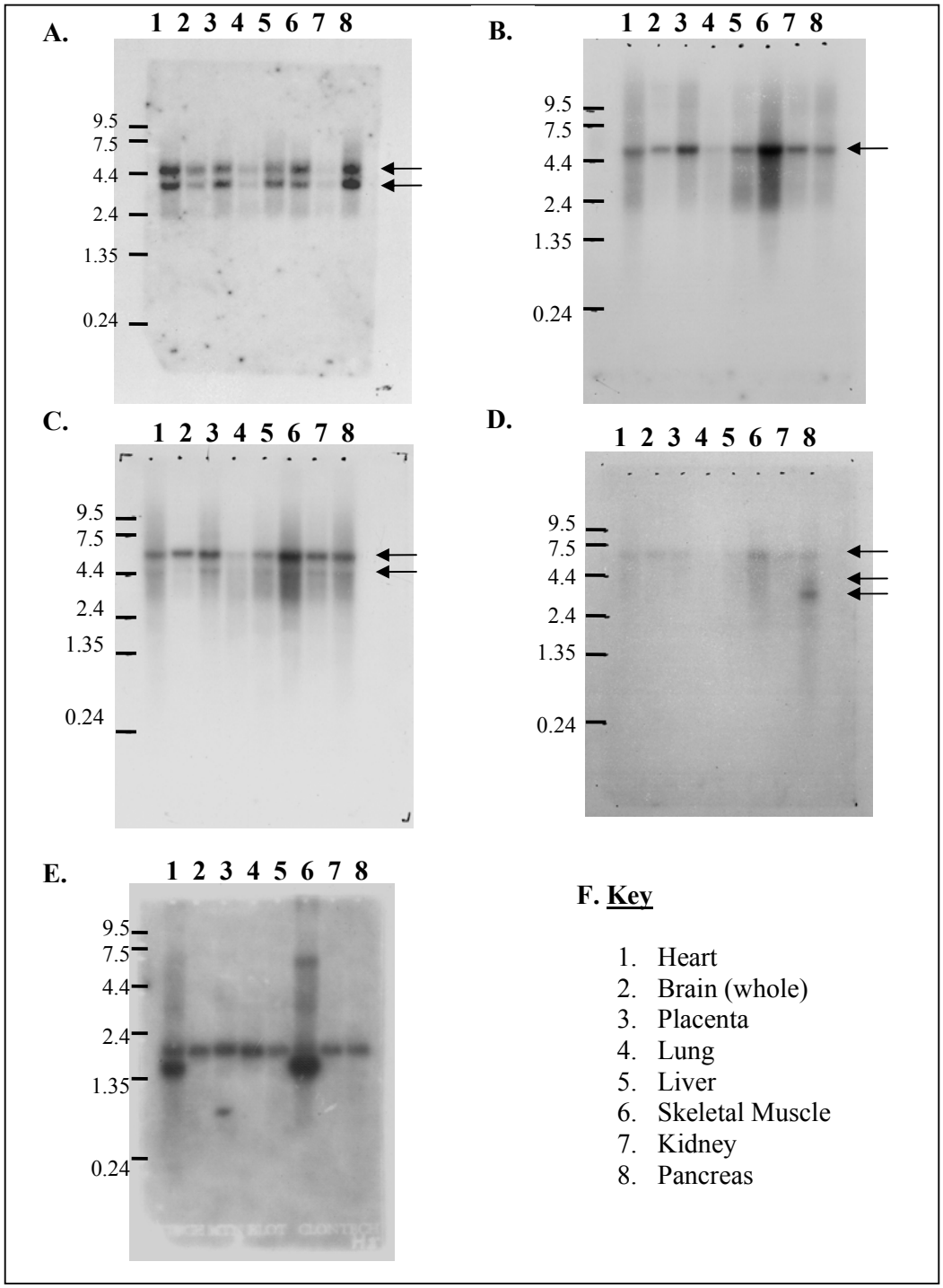


Figure 6.5 Transcription pattern and splice variants of the BRD2 (A), BRD3 (B) BRD4 (C), BRDT (D) and β-actin control (E) genes after hybridisation with specific probes to a Northern blot with eight different tissues. (F) Tissue/source key of the Northern blot as supplied by the manufacturer (Clontech). The splice variants are indicated with arrows for the BRD paralogous genes. Blots in A, B, C and E were exposed for 3 days whereas the blot in D was exposed for 18 days.

Northern analysis revealed that the BRD2, BRD4 and BRDT genes have multiple transcripts, whereas BRD3 only has one (indicated by arrows in figure 6.5). Two alternative splice variants of 4.6 kb and 3.8 kb were detected in all eight tissues probed using the BRD2 gene (figure 6.5.A). The weakest transcripts were identified in brain, lung and kidney. The strongest signals were for the heart and pancreas. A single transcript of approximately 6.5 kb was observed for BRD3 in the eight tissues on the Northern blot, with the strongest signal in skeletal muscle and the weakest in lung (figure 6.5.B). Two splice variants of the BRD4 gene were identified in all eight tissues of approximately 6.0 kb and 4.4 kb corresponding to the long and short isoforms of the BRD4 gene (French *et al*, 2003). The weakest expression was in lung and the strongest in skeletal muscle, which is similar to the BRD3 gene.

Weak expression of a single BRDT transcript of approximately 7 kb was detected in heart, brain, placenta, liver, skeletal muscle and kidney. In addition three transcripts were identified in pancreas, corresponding to the 7 kb transcript and two pancreas specific splice variants of approximately 3.5 kb and 4.0 kb. The strongest signal corresponds to the 3.5 kb variant in pancreas and the 7 kb transcript in skeletal muscle. These findings are interesting as the BRDT gene was identified using an EST from a testis-specific library (Diatchenko *et al*, 1996). Further expression analysis using 16 normal human tissues and eight cancer cell-lines indicated that there were only two BRDT transcripts of 3.5 kb and 4.0 kb which were both specific to testis (hence the gene being named bromodomain, testis-specific or BRDT; Jones *et al*, 1997). I have shown that this transcript is expressed at very low levels in a number of tissues by using the BRDT paralogue specific probe.

To summarise, Northern analysis shows that the BRD paralogous genes are co-

expressed in most of the tissues tested. In particular, the BRD3 and BRD4 genes both demonstrate elevated expression in skeletal muscle suggesting an important role in this tissue.

6.3.2.4 Microarray analysis

Microarrays can be used to simultaneously determine the expression profile of thousands of genes in a particular tissue or cell-line. Each experiment provides static information about gene expression (i.e. in which tissue(s) the gene is expressed) and dynamic information (i.e. how the expression pattern of one gene relates to those of others). In the expression microarray experiments a modified standard Sanger Institute 10K microarray was used to establish the expression profile of all the DNA elements on the array in ten different RNAs (described in more detail later in this section).

First, the RNAs were extracted from five cell-lines (as described in section 2.10) and five were purchased from Ambion. As a control, a standard RNA was purchased from Stratagene that is routinely used by the Sanger Institute Microarray Facility for the quality control of the microarrays they manufacture. The quality of the RNA was determined by electrophoresis of 2 μg of each of the RNAs on a 1% agarose gel (figure 6.6.A). The quality of the RNA is indicated by two bands corresponding to 28S and 18S ribosomal RNA. Sharp and distinct bands indicate good quality RNA but diffused and smeared are indicative of degradation. The RNA was also checked for DNA contamination using 'no RT-PCR'. The RNA was used as a template in a standard PCR reaction and the paralogue specific primers (summarised in Appendix 4) for the BRD2 gene were used, as described in section 2.9. If the RNA was contaminated a faint band was visible after 35 PCR cycles and the RNA was DNase

treated to remove the DNA in the sample and the quality of the RNA re-checked (as described in section 2.11).

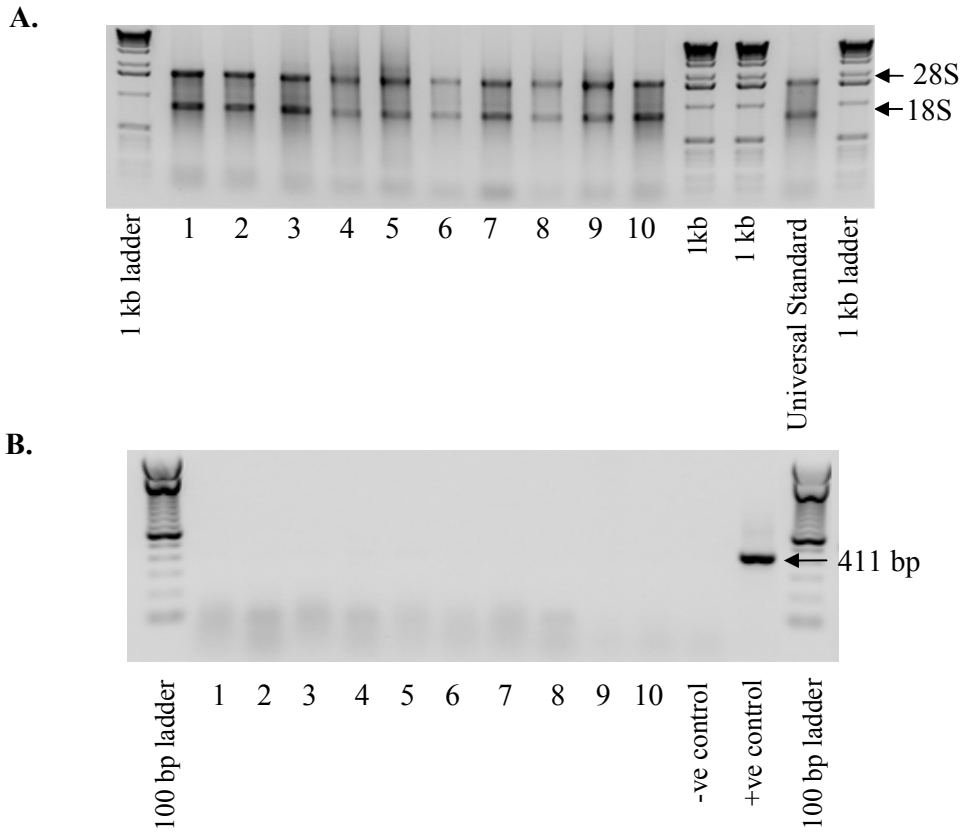


Figure 6.6 (A) Assessment of the quality of the eleven RNAs used in the expression microarray experiments. The RNA is of good quality and has the two distinct bands corresponding to 28S and 18S ribosomal RNA, indicated by arrows. (B) RNA was checked for DNA contamination using 'no RT-PCR'. The primers used were specific for the BRD2 gene which amplifies a 411 bp product in the positive control, where genomic DNA was used as the template, only. (C) is the key to the ten RNAs used in the analysis. RNAs 1 to 5 were purchased from Ambion and RNAs 6 to 10 were extracted from cell-lines as described in section 2.10. Water was used as the template in the negative control.

In the expression microarray experiments presented in this thesis the standard Sanger Institute 10,000 gene (or 10K) microarray was modified to accommodate the 40

paralogous genes and is termed the '10K/Paralogue Microarray' in this thesis (as described in section 2.13.1). In short, the 40 paralogue specific targets were amplified using the PCR specific primers (summarised in Appendix 4) and arrayed in quadruplicate onto the matrix, in this case a glass microscope slide. The mRNA from the tissue or cell-line was reverse-transcribed into cDNA, labelled with a fluorescent dye and hybridised to the '10K/Paralogue Microarray'. After hybridisation, a laser scanner measured the amount of fluorescence at each spot. The results of a hybridisation using the standard RNA purchased from Stratagene is shown in figure 6.7.

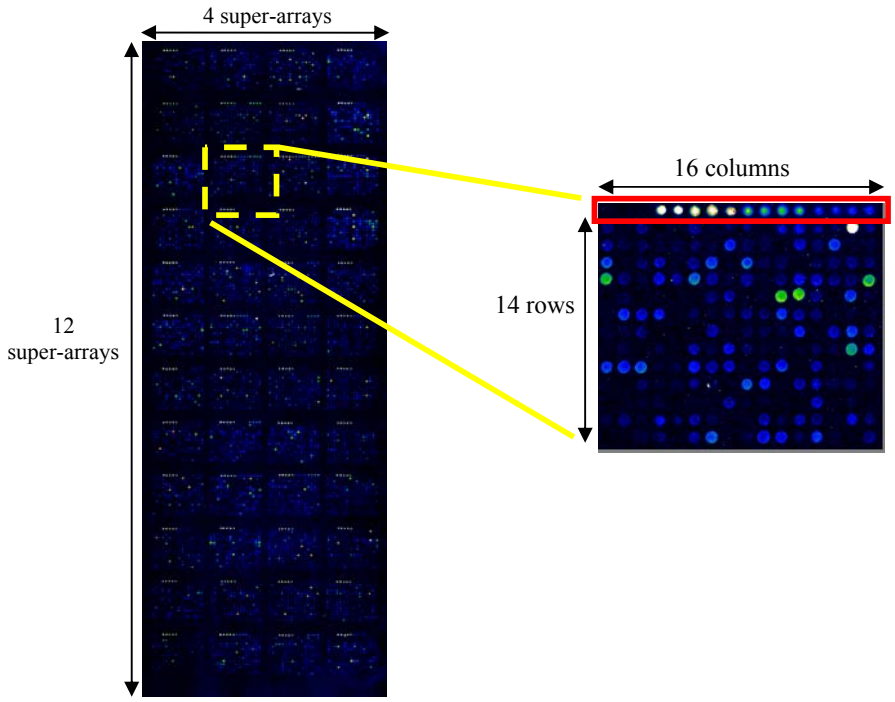


Figure 6.7 Results of a hybridisation with the standard Stratagene RNA to the '10K/Paralogue Microarray'. The layout of the 48 sub-arrays in 12 x 4 super-arrays is visible after hybridisation and the sub-array boxed in yellow is expanded. The first row of the sub-array is boxed in red. Columns 1 to 8 of row 1 contain the controls described in section 2.13.1. The paralogue specific PCR products of one paralogue are arrayed in rows 9 to 12 (shown as 4 green spots) and of a second paralogue in rows 13 to 16 (shown as 4 blue spots). The level of expression is indicated by the intensity of the spot, which is, in turn, is indicated by the colour of the spot. The colour intensities are, from highest to lowest, white > red > yellow > green > blue > black (i.e. no spot).

In short, when a spot is visible after hybridisation with the labelled cDNA the corresponding DNA element on the array is expressed in that tissue but when the particular transcript is not expressed no spot is visible. This is summarised in figure 6.8.

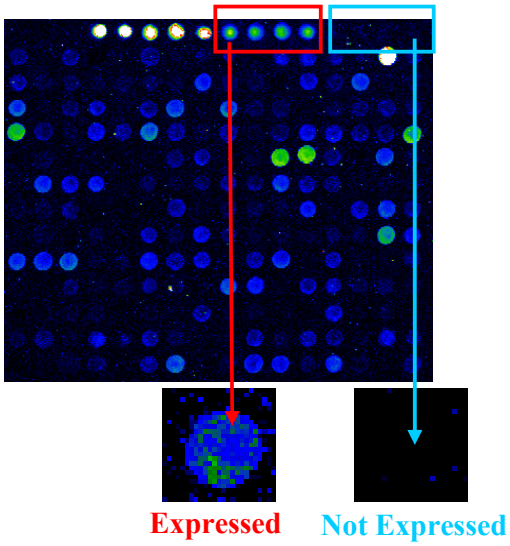


Figure 6.8 One of the 48 sub-arrays of the '10K/Paralogue Microarray' after hybridisation using the Stratagene standard RNA. The four spots boxed in red correspond to one paralogue which is expressed in the standard RNA. The area boxed in blue contains four DNA elements corresponding to one paralogue which is not expressed in the standard RNA, indicated by the absence of spots. The level of expression is indicated by the intensity of the spot, which is, in turn, indicated by the colour of the spot. The colour intensities are, from highest to lowest, white > red > yellow > green > blue > black (i.e. no spot).

The intensity of the spot and the background were determined using the fixed circle method in the Quantarray® software package. The highly regular arrangement of the spots in rows and columns resulting from the robotic printing rendered the image data amenable to extraction by highly developed, digital image processing procedures. In order to detect the spots a grid was overlaid on the scanned array image. Firstly, the array pattern was established and the initial definition of the area of the spot (i.e. the spot diameter and the row and column information) determined. In an ideal situation

the spots are perfectly circular, homogenous (i.e. the intensity is the same at each pixel in the spot) and the background signal is well defined. However, the spots on the slide may be somewhat irregular in nature and are not perfectly placed on the slide. Therefore, to counteract this, the precise location of each spot was identified by editing the array pattern and the reading was taken within the region defined as containing the spot. A spot typically consists of a number of pixels and the image analysis algorithms either assign pixels to a spot or not and produces a summary of the intensity of the fluorescence at each spot and the surrounding unspotted area i.e. the background.

In total, three experiments were performed for each of the 10 test RNAs and the Stratagene standard RNA. The outputs of the analysis of each of the 30 arrays were independently analysed. In order to determine whether a spot was present or not, i.e. the gene is expressed in that tissue or not, the standard deviation between the spot intensity and the background intensity was calculated. It was determined that a standard deviation greater than two indicated that a spot was present and the gene therefore expressed. For example, the standard deviation of the BRD2 gene in brain tissue for one experiment across all four spots in the upper section of the array were 7.09, 7.09, 8.08 and 8.02, thus clearly indicating that the BRD2 gene is expressed. This was also confirmed by the *in-silico*, Dot-blot and Northern blot analyses. In the case of the 40 paralogues each of the corresponding spots were manually analysed.

As a control, the 10 test RNAs were reverse transcribed into complementary DNA and were used as the templates in PCR reactions to amplify the paralogue specific PCR products for the AIF1 gene and its paralogue, AIF1-L. The results of the microarray and RT-PCR experiment performed these two genes shows that the same

expression profile was achieved using both techniques (summarised in figure 6.9A).

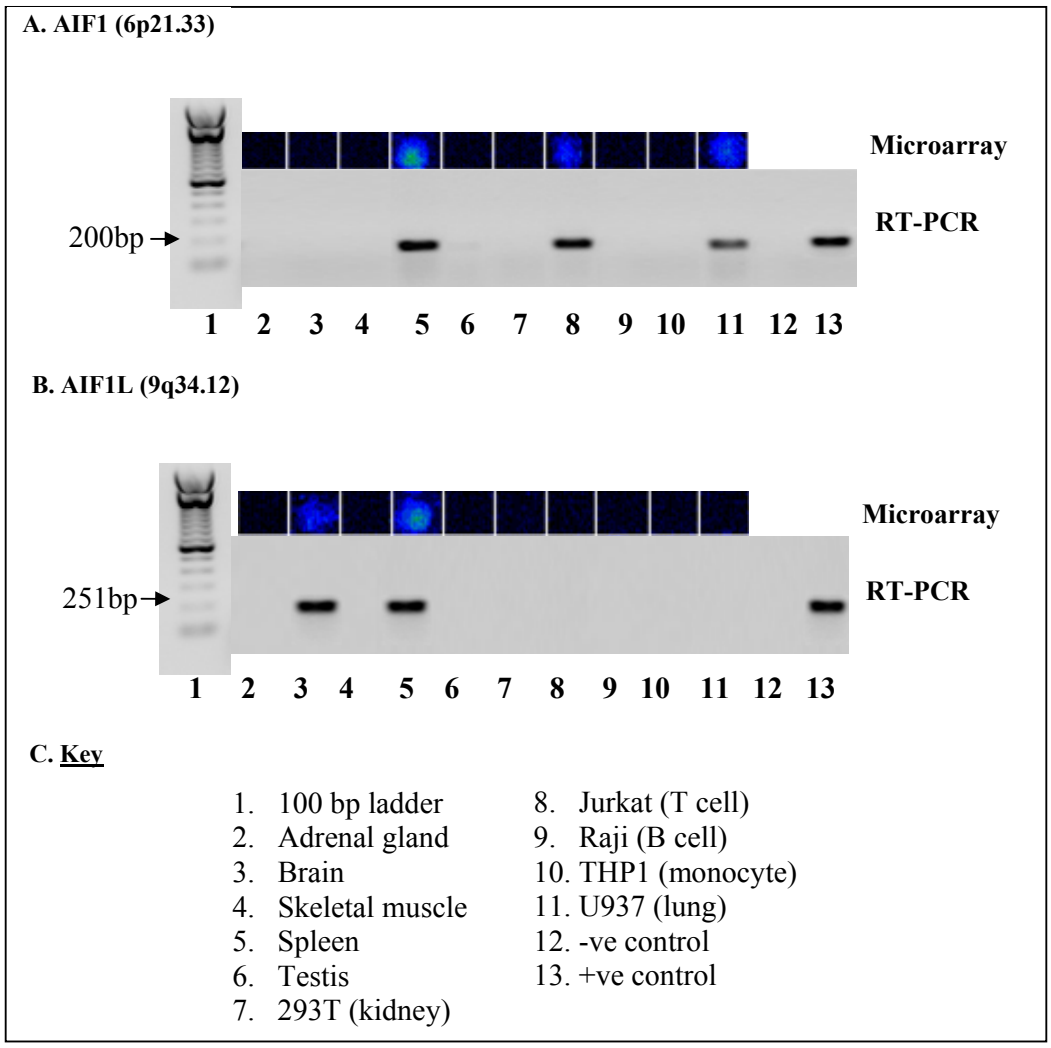


Figure 6.9 Microarray results confirmed by RT-PCR. (A) RT-PCR (using the primers AIF1.F1 and AIF1.R1 summarised in appendix 4) and (B) RT-PCR results (using AIF1-L.F and AIF1-L.R see Appendix 4) and microarray results of AIF1-L. The negative (-ve) control used water as a template and the positive (+ve) control used genomic DNA as the template and only apply to RT-PCR. (C) summarises the marker and tissue key.

6.3.2.5 Importance of designing specific microarray targets

In addition to the 40 paralogues selected for further analysis there are 9464 other DNA elements spotted onto the standard Sanger Institute 10K array. These DNA

elements correspond to cDNAs derived from direct sequencing of I.M.A.G.E (or Integrated Molecular Analysis of Genomes and their Expression) clones, which are generated as part of the EST project, and 468 chromosome 22 gene-specific PCR products. In total, 15 of the 40 paralogues selected for further analysis are already represented on the standard Sanger Institute 10K microarray and are part of the gene repertoire. The 15 genes correspond to eight of the 10 paralogous gene families. It is, therefore, of interest to compare the expression profiles of the paralogue specific PCR products designed in this thesis and those already on the microarray. Examples of such a comparison are shown in figure 6.10.

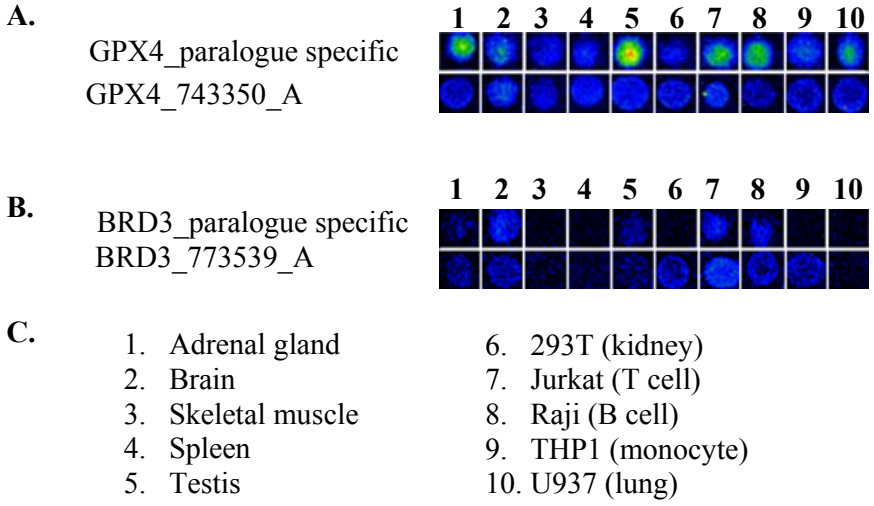


Figure 6.10 Comparison of the expression profiles of the paralogue specific PCR products designed in this thesis and those already on the standard Sanger Institute 10K microarray corresponding to (A) GPX4 and (B) BRD3 genes. (C) is the key to the tissues and cell-lines used.

The expression profile of the paralogue specific PCR product designed for GPX4 in this thesis is identical to that of the DNA element (GPX4_743350_A) already spotted onto the standard Sanger Institute 10K array (figure 6.10.A). However, the DNA element BRD3_773539_A is expressed in all of the same tissues as the paralogue

specific PCR product I designed for the BRD3 gene but it is also expressed in two additional tissues (figure 6.10.B). The difference in expression may be due to the DNA element cross-hybridising with another paralogue or may correspond to different splice variants that are not represented by my paralogue specific PCR product. This emphasises the importance of understanding, not only, which gene, but the splice variant a DNA element on a microarray corresponds to when interpreting the results of a hybridisation experiment. It also shows the value of designing a paralogue and splice variant-specific microarray.

6.3.3 Interpretation of expression data

In order to interpret the vast amounts of expression data generated in this thesis and determine the relationships between the MHC paralogous genes the data was clustered. Clustering is a technique used in exploratory data analysis and pattern discovery to extract underlying cluster structures. The data presented in this chapter was clustered using the unsupervised clustering methods, hierarchical clustering (clustering methods are reviewed by Brazma and Vilo, 2001) using EPCLUST (Expression Profile Data CLUSTering and Analysis) available from the EBI, unless stated otherwise.

6.3.3.1 Tenascin paralogous gene family

The tenascin paralogous gene family was removed from the analysis when it became apparent that I had designed paralogue specific primers to the wrong transcript of Tenascin X. The tenascin proteins are a family of extracellular matrix proteins (ECM) (for a review see Erickson, 1993). The Tenascin X gene was partially duplicated

during the duplication of the region on 6p, which gave rise to two isoforms, the 65 kb TNXB and 4.5 kb TNXA. A truncated version of the TNXB gene, termed TNXB-S (or TNXB-short) has been identified as an adrenal gland specific transcript (Tee *et al*, 1995). Evidence from the expression profiling experiments indicated that I had designed TNXB specific primers to the TNXB-S transcript as microarray experiments determined that it was only expressed in adrenal gland (figure 6.11). This was confirmed by Dot-blot analysis.

1 2 3 4 5 6 7 8 9 10

Key

1. Adrenal gland
2. Brain
3. Skeletal muscle
4. Spleen
5. Testis
6. 293T (kidney)
7. Jurkat (T cell)
8. Raji (B cell)
9. THP1 (monocyte)
10. U937 (lung)

Figure 6.11 Expression profile of the TNXB gene indicates that it is adrenal gland specific. This suggests that the paralogue specific primers were designed for the truncated TNX transcript rather than the full length gene.

6.3.3.2 Microarray expression data

The results of the microarray expression experiments and the resulting clustering are summarised in figure 6.12. Each spot is representative of one of the 24 spots corresponding to an individual paralogue[1]. In order to cluster the data it was necessary to assign numerical values to the expression profiles. When the gene was expressed in a particular tissue, indicated by the presence of a spot, it was assigned the number 1. When no spot was present, 0 was assigned. As the experiments were performed in triplicate, in the case of uncertainty, the majority rule was applied. In other words, if

[1] each paralogue specific PCR product was spotted in quadruplicate in two separate locations on the microarray and the experiments were repeated three times

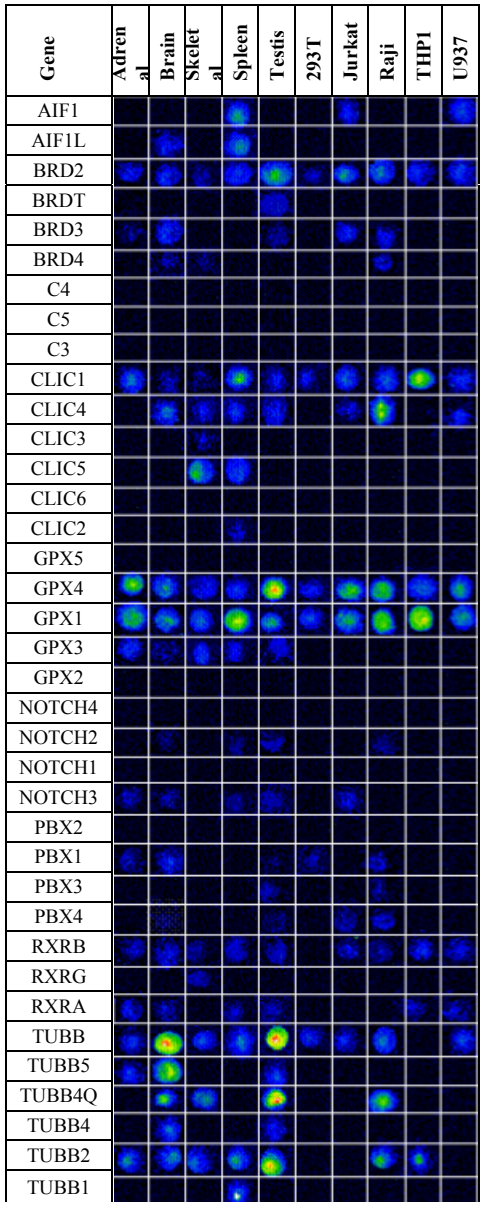
two out of the three experiments showed expression the gene in that tissue was assigned the value of 1.

Analysis of the microarray results presented in figure 6.12.A shows that the members of only one of the ten MHC paralogous gene families have identical expression patterns. This family is the complement paralogous gene family which has three members, C4, C3 and C5. These genes are not expressed in any of the ten tissues used in this analysis. Interestingly, upon clustering the complement genes do cluster together, along with the six other genes also not expressed in any of the tissues tested (figure 6.12.B).

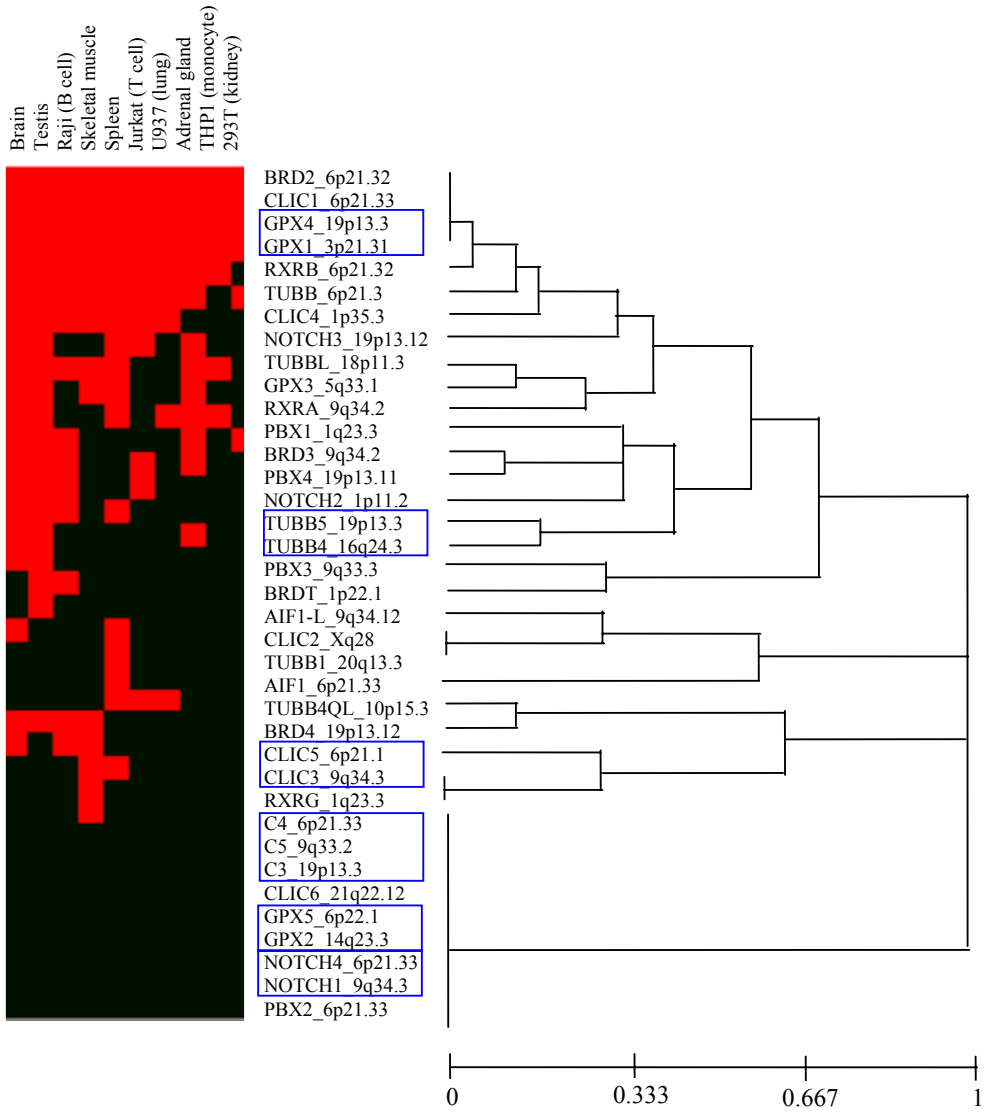
In addition to highlighting the relationships between members of the same paralogous gene families, clustering the data reveals the relationships between all the paralogues used in the expression analysis (summarised in figure 6.12.B). It is interesting to note that the four paralogues that are expressed in all ten tissues are clustered, of which two are members of the GPX paralogous gene family, GPX1 and GPX4, and the other two, BRD2 and CLIC1, are members of different paralogous gene families. As stated earlier, the nine genes not expressed in any of the ten tissues are also clustered. Although it was apparent prior to clustering the data that the members of the same paralogous gene families are differentially expressed there are some members of the β-tubulin paralogous gene family which do cluster; they are TUBB5 and TUBB4. However, the other four members investigated cluster with members of other paralogous gene families, including the GPX and the CLIC paralogous gene families, which have different functions in the human body.

Figure 6.12 (A) Summary of the microarray expression data and (B) the result of applying Hierarchical clustering methods. Red indicates that the gene is expressed in the corresponding tissue, whereas black shows lack of expression. Members of the same paralogous gene families that cluster together are highlighted by blue box in (B). The raw data is summarised in Appendix 8.

A.


Gene
Adrenal
Brain
Skeletal
Spleen
Testis
293T
Jurkat
Raji
THP1
U937
AIF1
AIF1L
BRD2
BRDT
BRD3
BRD4
C4
C5
C3
CLIC1
CLIC4
CLIC3
CLIC5
CLIC6
CLIC2
GPX5
GPX4
GPX1
GPX3
GPX2
NOTCH4
NOTCH2
NOTCH1
NOTCH3
PBX2
PBX1
PBX3
PBX4
RXRB
RXRG
RXRA
TUBB
TUBB5
TUBB4Q
TUBB4
TUBB2
TUBB1

B.

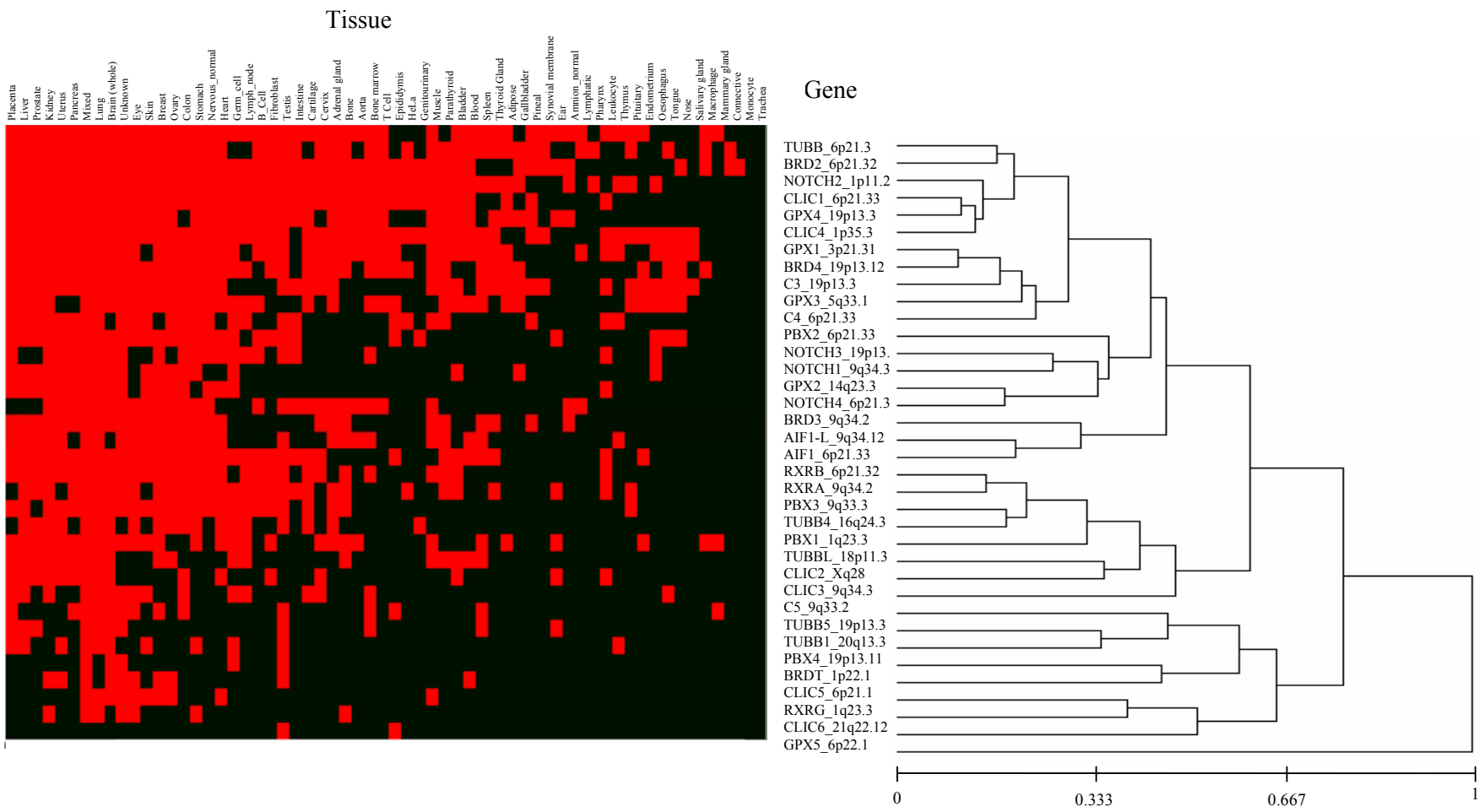

Brain
Testis
Raji (B cell)
Skeletal muscle
Spleen
Jurkat (T cell)
U937 (lung)
Adrenal gland
THP1 (monocyte)
293T (kidney)
BRD2_6p21.32
CLIC1_6p21.33
GPX4_19p13.3
GPX1_3p21.31
RXRB_6p21.32
TUBB_6p21.3
CLIC4_1p35.3
NOTCH3_19p13.12
TUBBL_18p11.3
GPX3_5q33.1
RXRA_9q34.2
PBX1_1q23.3
BRD3_9q34.2
PBX4_19p13.11
NOTCH2_1p11.2
TUBB5_19p13.3
TUBB4_16q24.3
PBX3_9q33.3
BRDT_1p22.1
AIF1-L_9q34.12
CLIC2_Xq28
TUBB1_20q13.3
AIF1_6p21.33
TUBB4QL_10p15.3
BRD4_19p13.12
CLIC5_6p21.1
CLIC3_9q34.3
RXRG_1q23.3
C4_6p21.33
C5_9q33.2
C3_19p13.3
CLIC6_21q22.12
GPX5_6p22.1
GPX2_14q23.3
NOTCH4_6p21.33
NOTCH1_9q34.3
PBX2_6p21.33
0
0.333
0.667
1

6.3.3.3 *In-silico* expression data

The *in-silico* data compiled for nine MHC genes and 27 paralogues was clustered for 61 different tissues and cell-lines (figure 6.13). It is important to note that the TUBB4QL gene located on 10p15.3 did not have a UNIGENE cluster and no unique ESTS were identified, therefore, there is no *in-silico* data for this gene. However, there is a vast amount of information regarding the remaining 36 paralogues. Clustering has enabled relationships to be discovered between the paralogues which were not apparent upon initial analysis of the raw data (summarised in Appendix 5). It is interesting to see that some members of the same paralogous gene families are clustered. For example, both members of the AIF paralogous gene family are clustered together, which is interesting as they demonstrate both co-expression as well as divergence in their expression patterns. As in the microarray experiments two members of the β-tubulin paralogous gene family cluster, albeit they are TUBB5 and TUBB1 rather than TUBB5 and TUBB4.

Two members of the RXR family, RXRA and RXRB, are also clustered and demonstrate overlapping expression profiles in a number of tissues. The third member of this group, RXRG, does not cluster with them as it has more specialised expression profile. In addition, both NOTCH1 and NOTCH3 genes cluster and it has previously been proposed that these genes may have an overlapping function (Lardelli *et al*, 1994). It is interesting to note that all 36 genes were represented in the EST libraries, with the GPX5 gene only represented in testis and epididymis, as previously described by Perry and co-workers (1992) and Hall and colleagues (1998).

Figure 6.13 Clustering of the *in-silico* expression profile results. Red indicates the gene is expressed in the corresponding tissue and black shows that there is no evidence of expression. The raw data is summarised in Appendix 5.


Tissue
Placenta
Liver
Prostate
Kidney
Uterus
Pancreas
Mixed
Lung
Brain (whole)
Unknown
Eye
Skin
Breast
Ovary
Colon
Stomach
Nervous_normal
Heart
Germ_cell
Lymph_node
B_Cell
Fibroblast
Testis
Intestine
Cartilage
Cervix
Adrenal gland
Bone
Aorta
Bone marrow
T Cell
Epididymis
HeLa
Genitourinary
Muscle
Parathyroid
Bladder
Blood
Spleen
Thyroid Gland
Adipose
Gallbladder
Pineal
Synovial membrane
Ear
Amnion_normal
Lymphatic
Pharynx
Leukocyte
Thymus
Pituitary
Endometrium
Oesophagus
Tongue
Nose
Salivary gland
Macrophage
Mammary gland
Connective
Monocyte
Trachea
Gene
TUBB_6p21.3
BRD2_6p21.32
NOTCH2_1p11.2
CLIC1_6p21.33
GPX4_19p13.3
CLIC4_1p35.3
GPX1_3p21.31
BRD4_19p13.12
C3_19p13.3
GPX3_5q33.1
C4_6p21.33
PBX2_6p21.33
NOTCH3_19p13.
NOTCH1_9q34.3
GPX2_14q23.3
NOTCH4_6p21.3
BRD3_9q34.2
AIF1-L_9q34.12
AIF1_6p21.33
RXRB_6p21.32
RXRA_9q34.2
PBX3_9q33.3
TUBB4_16q24.3
PBX1_1q23.3
TUBBL_18p11.3
CLIC2_Xq28
CLIC3_9q34.3
C5_9q33.2
TUBB5_19p13.3
TUBB1_20q13.3
PBX4_19p13.11
BRDT_1p22.1
CLIC5_6p21.1
RXRG_1q23.3
CLIC6_21q22.12
GPX5_6p22.1
0
0.333
0.667
1

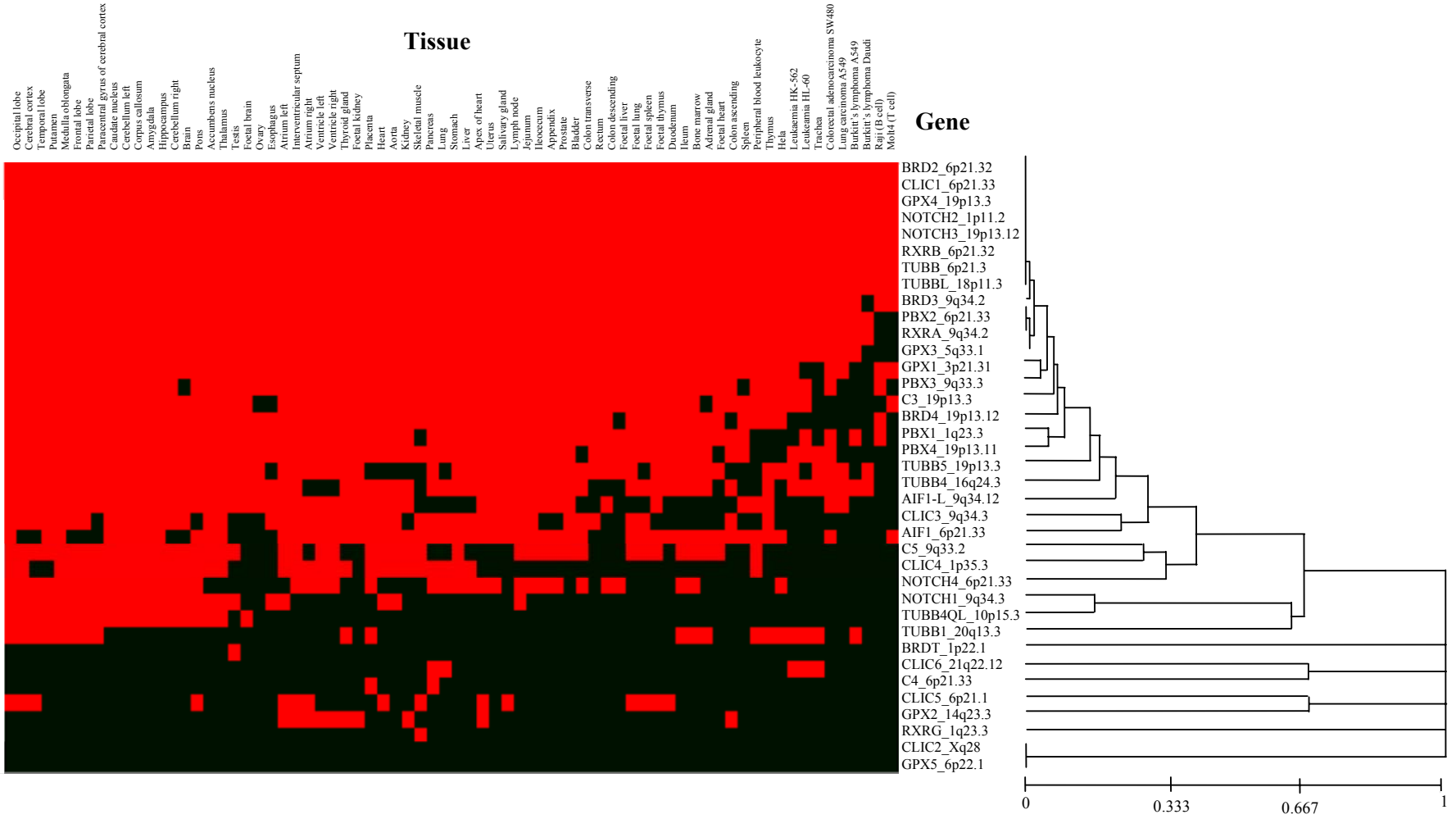
6.3.3.4 Dot-blot expression data

Clustering of the dot-blot expression data reveals a number of relationships between the expression profiles of the 37 paralogues investigated (figure 6.14). One of the most interesting findings is that eight of the 37 paralogous genes are ubiquitously expressed and are clustered as one group. Some members within this group, namely BRD2, CLIC1 and GPX4 also demonstrate similar expression profiles in the microarray and *in-silico* clustering figures. In addition, two members of the NOTCH paralogous gene family cluster, albeit they are NOTCH2 and NOTCH3 rather than NOTCH1 and NOTCH2 which cluster together in the *in-silico* analyses. The clustering of two members of the PBX paralogous gene family, PBX1 and PBX4, is unique to the dot-blot analysis.

There are two genes that are not expressed in any of the tissues, CLIC2 and GPX5, which are clustered. In addition, both the in-silico and microarray analyses showed a restricted expression pattern for the GPX5 gene in the tissues analysed. The expression profile of the CLIC2 gene generated by the microarray expression experiments showed limited expression whereas *in-silico* analysis showed expression in a wider range of tissues. This highlights the advantage of using multiple techniques, and RNAs from a range of tissues and sources, to generate expression profiles.

One point to note regarding the clustering of the expression data is that, although the clustering presented in this section reveals relationships between paralogues regarding their expression in various tissues, it should be viewed with caution. Further investigation is required to determine true relationships.

Figure 6.14 Clustering of the dot-blot expression profile results. Red indicates the gene is expressed in the corresponding tissue and black shows that there is no expression. The raw data is summaries in Appendix 6.


Tissue
Occipital lobe
Cerebral cortex
Temporal lobe
Putamen
Medulla oblongata
Frontal lobe
Parietal lobe
Paracentral gyrus of cerebral cortex
Caudate nucleus
Cerebellum left
Corpus callosum
Amygdala
Hippocampus
Cerebellum right
Brain
Pons
Accumbens nucleus
Thalamus
Testis
Foetal brain
Ovary
Esophagus
Atrium left
Interventricular septum
Atrium right
Ventricle left
Ventricle right
Thyroid gland
Foetal kidney
Placenta
Heart
Aorta
Kidney
Skeletal muscle
Pancreas
Lung
Stomach
Liver
Apex of heart
Uterus
Salivary gland
Lymph node
Jejunum
Heloeceum
Appendix
Prostate
Bladder
Colon transverse
Rectum
Colon descending
Foetal liver
Foetal lung
Foetal spleen
Foetal thymus
Duodenum
Ileum
Bone marrow
Adrenal gland
Foetal heart
Colon ascending
Spleen
Peripheral blood leukocyte
Thymus
Hela
Leukaemia HK-562
Leukaemia HL-60
Trachea
Colorectal adenocarcinoma SW480
Lung carcinoma A549
Burkitt's lymphoma A549
Burkitt's lymphoma Daudi
Raji (B cell)
Molt4 (T cell)
Gene
BRD2_6p21.32
CLIC1_6p21.33
GPX4_19p13.3
NOTCH2_1p11.2
NOTCH3_19p13.12
RXRB_6p21.32
TUBB_6p21.3
TUBBL_18p11.3
BRD3_9q34.2
PBX2_6p21.33
RXRA_9q34.2
GPX3_5q33.1
GPX1_3p21.31
PBX3_9q33.3
C3_19p13.3
BRD4_19p13.12
PBX1_1q23.3
PBX4_19p13.11
TUBB5_19p13.3
TUBB4_16q24.3
AIF1-L_9q34.12
CLIC3_9q34.3
AIF1_6p21.33
C5_9q33.2
CLIC4_1p35.3
NOTCH4_6p21.33
NOTCH1_9q34.3
TUBB4QL_10p15.3
TUBB1_20q13.3
BRDT_1p22.1
CLIC6_21q22.12
C4_6p21.33
CLIC5_6p21.1
GPX2_14q23.3
RXRG_1q23.3
CLIC2_Xq28
GPX5_6p22.1
0
0.333
0.667
1

6.3.3.5 Comparison of the expression profiles of the MHC paralogues located in the paralogous regions on chromosomes 1, 9 and 19

In order to understand the relationship between the expression profiles and the location of the paralogous genes the data was clustered as described in section 2.18 (presented in figure 6.15).

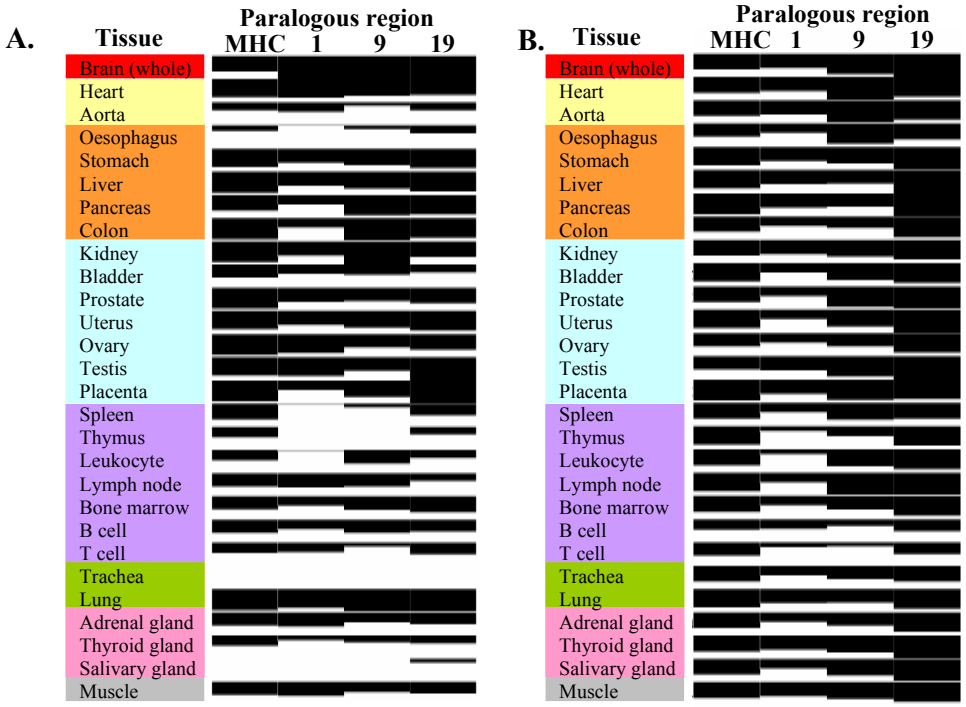


Figure 6.15 Comparison of the expression profiles of the paralogues located within the paralogous regions on chromosomes 1, 9 and 19 with the MHC genes using (A) *in-silico* and (B) dot blot analysis in 28 normal human tissues. The tissues are divided into eight systems of the human body; nervous (red), cardiovascular (yellow), digestive (orange), genitourinary (blue), immune (purple), respiratory (green), secretory (pink) and muscle (grey). A black bar indicates that genes within that region are expressed; the thickness of the bar is indicative of the percentage of genes expressed (i.e. the thicker the bar the more genes are expressed). Each tissue is separated by a grey horizontal line.

A total of 28 different normal human tissues corresponding to eight different systems of the body were common to both the *in-silico* and the dot-blot analyses. The results

presented in figure 6.15 correspond to 27 genes; of which nine are located within the MHC region, five are on chromosome 1, seven on chromosome 9 and six on chromosome 19. Overall, the genes located within the MHC region are expressed in most systems of the body and the profile is most similar to that of chromosome 9. It is apparent from figure 6.15 that the genes located within the region on chromosome 1 have a more specialised expression pattern whereas the chromosome 19 genes are more highly expressed throughout the different systems of the body.

6.3.3.6 Comparison of the methods used to generate expression profiles

Nine of the tissues used in the microarray, dot-blot and *in-silico* analyses are common to all three methods. It is, therefore, of interest to compare the expression profiles of the paralogous genes in these tissues in order to see how expression differs between the techniques used in this thesis (summarised in table 6.1 and Appendix 9). The number of differences between each method was determined and the percentage differences calculated. For example, when comparing the expression profiles of the 36 genes in testis there were four differences between the microarray results, four differences between the dot-blot results and five differences between the *in-silico* data and the other two methods. These corresponded to 11%, 11% and 14% differences respectively.

Table 6.1 Comparison of three methods used to generate the expression profiles for nine MHC paralogous gene families. M refers to microarray data, D is dot-blot data and S refers to the *in-silico* data. Full table can be found in Appendix 9.

	Adrenal Gland			***Brain***			***Skeletal muscle***			***Spleen***			***Testis***			***Kidney***			***T cell***			***B cell***			***Lung***		
	M	**D**	**S**	**M**	**D**	**S**	**M**	**D**	**S**	**M**	**D**	**S**	**M**	**D**	**S**	**M**	**D**	**S**	**M**	**D**	**S**	**M**	**D**	**S**	**M**	**D**	**S**
% differences between methods	14	8	25	11	3	22	14	11	14	8	14	19	11	11	14	44	3	22	8	6	14	6	6	14	42	0	28

There are some very large differences between the expression profiles of the genes within certain tissues. For example, 44% of the results obtained in the microarray analysis for the expression profiles of the genes in kidney are different to those determined by dot-blot and *in-silico* analyses. This is probably due to the sources of the RNA. In the case of the microarray experiments, the RNA was extracted from the human kidney cell-line, 293T, whereas the RNA on the dot-blot is pooled from 14 different individuals and the kidney EST libraries have been generated from a range of different kidney tissues. The age and gender of the individual from which the RNA was extracted may also affect the expression profiles of the genes in a particular tissue. This emphasises the importance of understanding the source of RNA when studying the expression of transcripts in a particular tissue.

6.4 Discussion

In order to understand the mechanism(s) by which the MHC paralogues have evolved since their emergence by duplication, the expression profiles of nine MHC paralogous gene families were generated in a range of tissues corresponding to eight different systems of the human body. A lot of information regarding the potential functions of the MHC paralogues can be deduced from the knowledge in which tissues they are expressed, but it is difficult to determine the precise mechanism by which the MHC paralogues have evolved without prior knowledge of the function, or functions, of the ancestral gene.

The function(s) of the ancestral gene can be deduced by analysing the expression profiles of the orthologues in organisms near the origin of the vertebrate lineage. For example, comparison of the expression of the four vertebrate NOTCH genes with that of the single Notch gene in amphioxus during embryogenesis indicates that they have similar roles (Holland *et al*, 2001). This organism is in a unique phylogenetic position as it is located at the base of the vertebrate lineage and can therefore be used as a 'stand-in' for the ancestral species. However, further investigation of the individual functions of the vertebrate NOTCH genes is necessary to determine the process by which paralogues have evolved. For example, functional comparisons will determine whether just one gene has maintained the functions of the ancestral gene, indicating that the paralogues have evolved via the process of neo-functionalisation, or whether the functions of the ancestral gene have been 'shared' between the four paralogues, thus they have evolved by the process of sub-functionalisation.

The data generated in this chapter has enabled the comparison of members of the same paralogous gene family as well as with members of other families in the same

tissues. In general, the MHC paralogues have distinct overall expression profiles to one another; but, in many cases there is still some level of overlap, or co-expression, in a number of tissues. For example, the expression profiles generated in this thesis for the NOTCH paralogues demonstrate some overlap in expression. One reason to explain why a certain amount of redundancy has been maintained over several hundred million years of evolution is that the paralogues concerned may perform the same functions.

It has been shown that experimental disruption of many individual genes does not exert lethal effects on an organism or even visible changes in phenotype of the organism with the knockout. In the unicellular eukaryote *Saccharomyces cerevisiae* there are approximately 5350 protein coding genes (Mackiewicz *et al*, 1999; 2002), of which only 924 are essential and probably unique, since their elimination from the genome has a lethal effect, while for about half of the other genes no changes in phenotype after disruption has been found (MIPS 2002 database (http://mips.gsf.de); Cebrat and Stauffer, 2002). If paralogues can perform the same function they may act as a 'back-up' system, thus if one paralogue is knocked-out, or rendered non-functional, another paralogue may be able to act as a substitute to prevent changes in phenotype, especially harmful ones.

To-date, there is little evidence to prove or refute this hypothesis. Experiments involving mice deficient for a member of the BRD paralogous gene family, the BRD4 gene, showed phenotypic changes in heterozygotes and fatality in homozygous mice (Houzelstein *et al*, 2002). The mice heterozygous for the BRD4 allele displayed pre and postnatal growth defects and exhibited a variety of malformations, including head malformations, absence of subcutaneous fat, cataracts and abnormal liver cells. The

BRD4 homozygous mice died shortly after implantation and were compromised in their ability to maintain an inner cell mass *in-vitro*. These experiments suggest that BRD4 plays an important role but it can not be substituted by the other three paralogues, or any other gene in the genome.

Comparison of the expression profiles of the BRD4 gene and the three paralogues generated in this thesis shows overlapping expression in a number of tissues. BRD4, BRD2 and BRD3 are expressed in most of the tissues used in the analysis and are expressed in all systems of the body, whereas the BRDT gene demonstrates a more restricted tissue distribution. Co-expression of the BRD genes in the same tissues would indicate that they have a similar function and could act as substitutes for each other if necessary. However, the experiments in mouse discussed above indicate otherwise.

Evidence from other paralogous genes in the mammalian genome has shown that paralogues with similar expression profiles have complementary functions in certain tissues and demonstrate partial redundancy rendering them functionally interchangeable. An example of this is demonstrated by two members of the Hox paralogous gene family, Hoxa3 and Hoxd3 (reviewed by Prince and Pickett, 2002). These genes have virtually identical expression patterns (Greer *et al*, 2000 and references therein). Intriguingly, mice lacking either a functional Hoxa3 or Hoxd3 gene show no obvious overlap in phenotype thus indicating that they have independent functions. However, the analysis of mice carrying different combinations of the mutant alleles of the Hox3 paralogues suggests that there is also a functional overlap between these genes.

It is apparent that a lot more functional information is needed before we will fully

understand the role of paralogous genes in the human body and just how much genetic redundancy is maintained or lost during evolution. To-date, there is evidence of functional redundancy as well as complementation between members of paralogous gene families. Expression analysis is only the first step in exploring the function of a gene and we are still a long way from having a detailed gene expression profile of every gene and paralogue in the human genome. The transcript patterns of nine MHC paralogous gene families presented in this chapter will therefore act as the basis for future research. The results in this chapter show that, like genome sequencing, systematic gene expression profiling is valuable but is not the end in itself. In order to understand the mechanism(s) by which the MHC paralogues have evolved we need to have a complete understanding of the functions of these genes. Therefore, further genetic analysis combined with biochemical studies are necessary to shed light on the functional evolution of the paralogues.

Chapter 7

Conclusions and future work

7.1 Conclusions

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

The study of the distribution of the MHC paralogous genes has confirmed that there are clusters of MHC paralogues located in the previously proposed regions on human chromosomes 1, 9 and 19. Almost 50% of the L2- and L3-paralogues are located within the regions 1q21.2-q25.3, 9q32-q34.3 and 19p13.3-p13.11. No further clusters of MHC paralogous genes were identified in the human genome, as postulated by Hughes and Pontarotti (2000). However, one of the most interesting, and novel,

findings of this thesis is that the MHC paralogous genes are not confined just to these regions but there are paralogues scattered throughout the human genome.

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

The existence of chromosomal regions containing clusters of duplicated genes is indicative of a common origin by large-scale duplication of either the whole-genome or of a block. In this thesis, I have identified three regions containing clusters of genes paralogous to those found within the MHC region, which is indicative of at least two rounds of large-scale duplication events. This is in support of the 2R hypothesis

which, in its simplest form, assumes two rounds of whole-genome duplication early in the vertebrate lineage; the first in the common ancestor of all vertebrates and the second in a common ancestor of jawed vertebrates after its separation from jawless fish (reviewed by Wolfe, 2001). It is also in support of two rounds of duplication of a chromosomal segment, or block duplication. Either way, if they did have a common origin, it is expected that the regions are syntenic, which is not strictly obeyed.

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

Prior to my genome survey, the proposed genomic distribution of the MHC paralogous genes in the regions on 1, 9 and 19 was considered as evidence of past large-scale duplication (Kasahara, 1997; 1999a; 1999b). It was presumed that the regions emerged as part of two rounds of whole-genome duplication believed to have occurred early in the history of vertebrates, approximately 500 million years ago. The

results of the phylogenetic studies presented in this thesis indicate that some of the MHC paralogues did emerge via large-scale duplication events early in the vertebrate lineage. This is consistent with the extensive evidence emerging in the literature to show that there was a burst of gene duplication during early chordate evolution (Pépusque *et al*, 1998; Wang and Gu, 2000; Miyata and Suga, 2001; Escriva *et al*, 2002; McLysaght *et al*, 2002; Panopoulou *et al*, 2003).

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

The only two member family (AIF) studied also shows that there was at least one round of duplication in the vertebrate lineage after the emergence of amphioxus. The existence of only two paralogues indicates that this family was only involved in one round of genome duplication (supporting a 1R hypothesis) or two paralogues have been lost after two genome duplication events (supporting the 2R hypothesis). It has

been calculated that the average time before silencing of one member of a duplicate gene pair is approximately four million years in animals (Lynch and Conery, 2000), therefore the likelihood of these paralogues being lost since their emergence, approximately 500 million years ago, is high.

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

Gene and, potentially, genome duplication have played a central role in the evolution of the MHC paralogues. It has been shown that gene duplications are frequent events in the mammalian genome with an average duplication rate of approximately 1% per gene per million years (Lynch and Conery, 2000). But what happens after gene or genome duplication? The classical model predicts that one copy will be maintained under purifying selection whereas the other will accumulate mutations which will generally lead to the loss of function of that gene copy. In rare cases, new functions will be created and both duplicate genes will be conserved. In contrast, under the sub-functionalisation model both duplicates are preserved due to the partition of different functions between duplicates: the development of new functions is also possible. In

order to understand how the paralogues have evolved since duplication the first step was to determine the function(s) of these genes in humans. This was addressed in this thesis by generating the expression profiles of the members of nine MHC paralogous gene families. The profiles were generated in a range of tissues corresponding to the major systems of the human body using several different approaches.

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

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

ancestral genes as well as the functions of the human genes.

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

7.2 Future work

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

1. Comparative studies

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

2. Further analysis of the human genome

Initial analysis of the human genome identified significantly less genes than expected and it has been proposed that differential splicing of genes and the different encoded proteins play a crucial role in humans. The analysis of the splice variants identified in this thesis will help understand the role paralogues play and it will also be of interest to determine whether the same splice variants are maintained and used by the different paralogues. It will also be of interest to

determine whether the functions of the ancestral genes have been split between the different splice variants of the paralogues and whether they have specific functions. The study of the regulatory features of the paralogues will also provide insight to both the evolution and control of the paralogues.

3. Improved strategy to identify paralogues

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

4. Functional studies

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

5. Paralogue-specific microarrays

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

Final conclusion

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

Bibliography

Abi Rached L., McDermott M. F., and Pontarotti P. (1999). The MHC big bang. *Immunol Rev* 167: 33-44.

Abi-Rached L., Gilles A., Shiina T., Pontarotti P., and Inoko H. (2002). Evidence of en bloc duplication in vertebrate genomes. *Nat Genet* 31: 100-5.

Achermann J. C., Ito M., Hindmarsh P. C., and Jameson J. L. (1999). A mutation in the gene encoding steroidogenic factor-1 causes XY sex reversal and adrenal failure in humans. *Nat Genet* 22: 125-6.

Adams M. D., Celniker S. E., Holt R. A., Evans C. A., Gocayne J. D., Amanatides P. G., Scherer S. E., Li P. W., Hoskins R. A., Galle R. F., George R. A., Lewis S. E., Richards S., Ashburner M., Henderson S. N., Sutton G. G., Wortman J. R., Yandell M. D., Zhang Q., Chen L. X., Brandon R. C., Rogers Y. H., Blazej R. G., Champe M., Pfeiffer B. D., Wan K. H., Doyle C., Baxter E. G., Helt G., Nelson C. R., Gabor G. L., Abril J. F., Agbayani A., An H. J., Andrews-Pfannkoch C., Baldwin D., Ballew R. M., Basu A., Baxendale J., Bayraktaroglu L., Beasley E. M., Beeson K. Y., Benos P. V., Berman B. P., Bhandari D., Bolshakov S., Borkova D., Botchan M. R., Bouck J., Brokstein P., Brottier P., Burtis K. C., Busam D. A., Butler H., Cadieu E., Center A., Chandra I., Cherry J. M., Cawley S., Dahlke C., Davenport L. B., Davies P., de Pablos B., Delcher A., Deng Z., Mays A. D., Dew I., Dietz S. M., Dodson K., Doup L. E., Downes M., Dugan-Rocha S., Dunkov B. C., Dunn P., Durbin K. J., Evangelista C. C., Ferraz C., Ferriera S., Fleischmann W., Fosler C., Gabrielian A. E., Garg N. S., Gelbart W. M., Glasser K., Glodek A., Gong F., Gorrell J. H., Gu Z., Guan P., Harris M., Harris N. L., Harvey D., Heiman T. J., Hernandez J. R., Houck J., Hostin D., Houston K. A., Howland T. J., Wei M. H., Ibegwam C., et al. (2000). The genome sequence of *Drosophila melanogaster*. *Science* 287: 2185-95.

Altschul S. F., Gish W., Miller W., Myers E. W., and Lipman D. J. (1990). Basic local alignment search tool. *J Mol Biol* 215: 403-10.

Amores A., Force A., Yan Y. L., Joly L., Amemiya C., Fritz A., Ho R. K., Langeland J., Prince V., Wang Y. L., Westerfield M., Ekker M., and Postlethwait J. H. (1998). Zebrafish hox clusters and vertebrate genome evolution. *Science* 282: 1711-4.

Anderson S., Bankier A. T., Barrell B. G., de Bruijn M. H., Coulson A. R., Drouin J., Eperon I. C., Nierlich D. P., Roe B. A., Sanger F., Schreier P. H., Smith A. J., Staden R., and Young I. G. (1981). Sequence and organization of the human mitochondrial genome. *Nature* 290: 457-65.

Aparicio S., Chapman J., Stupka E., Putnam N., Chia J. M., Dehal P., Christoffels A., Rash S., Hoon S., Smit A., Gelpke M. D., Roach J., Oh T., Ho I. Y., Wong M., Detter C., Verhoef F., Predki P., Tay A., Lucas S., Richardson P., Smith S. F., Clark M. S., Edwards Y. J., Doggett N., Zharkikh A., Tavtigian S. V., Pruss D., Barnstead M., Evans C., Baden H., Powell J., Glusman G., Rowen L., Hood L., Tan Y. H., Elgar G.,

Hawkins T., Venkatesh B., Rokhsar D., and Brenner S. (2002). Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* 297: 1301-10.

Autieri M. V. (1996). cDNA cloning of human allograft inflammatory factor-1: tissue distribution, cytokine induction, and mRNA expression in injured rat carotid arteries. *Biochem Biophys Res Commun* 228: 29-37.

Bailey J. A., Gu Z., Clark R. A., Reinert K., Samonte R. V., Schwartz S., Adams M. D., Myers E. W., Li P. W., and Eichler E. E. (2002). Recent segmental duplications in the human genome. *Science* 297: 1003-7.

Bairoch A., and Apweiler R. (1997). The SWISS-PROT protein sequence data bank and its supplement TrEMBL. *Nucleic Acids Res* 25: 31-6.

Bankier A. T., Weston K. M., and Barrell B. G. (1987). Random cloning and sequencing by the M13/dideoxynucleotide chain termination method. *Methods Enzymol* 155: 51-93.

Baumgartner S., Martin D., Hagios C., and Chiquet-Ehrismann R. (1994). Tenm, a Drosophila gene related to tenascin, is a new pair-rule gene. *Embo J* 13: 3728-40.

Beck S., Kelly A., Radley E., Khurshid F., Alderton R. P., and Trowsdale J. (1992a). DNA sequence analysis of 66 kb of the human MHC class II region encoding a cluster of genes for antigen processing. *J Mol Biol* 228: 433-41.

Beck S., Hanson I., Kelly A., Pappin D. J., and Trowsdale J. (1992b). A homologue of the Drosophila female sterile homeotic (fsh) gene in the class II region of the human MHC. *DNA Seq* 2: 203-10.

Beck S., and Trowsdale J. (2000). The human major histocompatability complex: lessons from the DNA sequence. *Annu Rev Genomics Hum Genet* 1: 117-37.

Bernstein R. M., Schluter S. F., Bernstein H., and Marchalonis J. J. (1996). Primordial emergence of the recombination activating gene 1 (RAG1): sequence of the complete shark gene indicates homology to microbial integrases. *Proc Natl Acad Sci U S A* 93: 9454-9.

Bichara M., Schumacher S., and Fuchs R. P. (1995). Genetic instability within monotonous runs of CpG sequences in Escherichia coli. *Genetics* 140: 897-907.

Bichara M., Pinet I., Schumacher S., and Fuchs R. P. (2000). Mechanisms of dinucleotide repeat instability in Escherichia coli. *Genetics* 154: 533-42.

Blattner F. R., Plunkett G., 3rd, Bloch C. A., Perna N. T., Burland V., Riley M., Collado-Vides J., Glasner J. D., Rode C. K., Mayhew G. F., Gregor J., Davis N. W., Kirkpatrick H. A., Goeden M. A., Rose D. J., Mau B., and Shao Y. (1997). The complete genome sequence of Escherichia coli K-12. *Science* 277: 1453-74.

Bogart J. P. (1980). Evolutionary implications of polyploidy in amphibians and reptiles. *In* "Polyploidy: Biological Relevance" (W. H. Lewis, Ed.), Plenum, New York.

Bonfield J. K., Smith K., and Staden R. (1995). A new DNA sequence assembly program. *Nucleic Acids Res* 23: 4992-9.

Bouchireb N., Grutzner F., Haaf T., Stephens R. J., Elgar G., Green A. J., and Clark M. S. (2001). Comparative mapping of the human 9q34 region in Fugu rubripes. *Cytogenet Cell Genet* 94: 173-9.

Bray S. (1998). Notch signalling in Drosophila: three ways to use a pathway. *Semin Cell Dev Biol* 9: 591-7.

Brazma A., and Vilo J. (2001). Gene expression data analysis. *Microbes Infect* 3: 823-9.

Brenner S. E., Chothia C., and Hubbard T. J. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc Natl Acad Sci U S A* 95: 6073-8.

Bristow J., Tee M. K., Gitelman S. E., Mellon S. H., and Miller W. L. (1993). Tenascin-X: a novel extracellular matrix protein encoded by the human XB gene overlapping P450c21B. *J Cell Biol* 122: 265-78.

Brown N. P., Whittaker A. J., Newell W. R., Rawlings C. J., and Beck S. (1995). Identification and analysis of multigene families by comparison of exon fingerprints. *J Mol Biol* 249: 342-59.

C. elegans Sequencing Consortium (1998). Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* 282: 2012-8.

Cebrat S., and Stauffer D. (2002). Monte Carlo simulation of genome viability with paralog replacement. *J Appl Genet* 43: 391-5.

Chenna R., Sugawara H., Koike T., Lopez R., Gibson T. J., Higgins D. G., and Thompson J. D. (2003). Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res* 31: 3497-500.

Chu F. F., Esworthy R. S., Doroshow J. H., Doan K., and Liu X. F. (1992). Expression of plasma glutathione peroxidase in human liver in addition to kidney, heart, lung, and breast in humans and rodents. *Blood* 79: 3233-8.

Colombo P., Yon J., Garson K., and Fried M. (1992). Conservation of the organization of five tightly clustered genes over 600 million years of divergent evolution. *Proc Natl Acad Sci U S A* 89: 6358-62.

Danchin E. G., Abi-Rached L., Gilles A., and Pontarotti P. (2003). Conservation of the MHC-like region throughout evolution. *Immunogenetics*.

Dausset J. (1958). Iso-leuco-anticorps. *Acta Haem* 20: 156-166.

Dayhoff M. O., Schwartz R. M., and Orcutt B. C. (1978). "A model of evolutionary change in proteins," National Biomedical Research Foundation, Washington D.C.

Deloukas P., Schuler G. D., Gyapay G., Beasley E. M., Soderlund C., Rodriguez-Tome P., Hui L., Matise T. C., McKusick K. B., Beckmann J. S., Bentolila S., Bihoreau M., Birren B. B., Browne J., Butler A., Castle A. B., Chiannilkulchai N., Clee C., Day P. J., Dehejia A., Dibling T., Drouot N., Duprat S., Fizames C., Bentley D. R., and et al. (1998). A physical map of 30,000 human genes. *Science* 282: 744-6.

Denis G. V., and Green M. R. (1996). A novel, mitogen-activated nuclear kinase is related to a Drosophila developmental regulator. *Genes Dev* 10: 261-71.

Diatchenko L., Lau Y. F., Campbell A. P., Chenchik A., Moqadam F., Huang B., Lukyanov S., Lukyanov K., Gurskaya N., Sverdlov E. D., and Siebert P. D. (1996). Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proc Natl Acad Sci U S A* 93: 6025-30.

Doolittle R. F., Feng D. F., Johnson M. S., and McClure M. A. (1986). Relationships of human protein sequences to those of other organisms. *Cold Spring Harb Symp Quant Biol* 51 Pt 1: 447-55.

Driscoll J., Brown M. G., Finley D., and Monaco J. J. (1993). MHC-linked LMP gene products specifically alter peptidase activities of the proteasome. *Nature* 365: 262-4.

Dufaure J. P., Lareyre J. J., Schwaab V., Mattei M. G., and Drevet J. R. (1996). Structural organization, chromosomal localization, expression and phylogenetic evaluation of mouse glutathione peroxidase encoding genes. *C R Acad Sci III* 319: 559-68.

Eichler E. E. (1998). Masquerading repeats: paralogous pitfalls of the human genome. *Genome Res* 8: 758-62.

Eichler E. E., Archidiacono N., and Rocchi M. (1999). CAGGG repeats and the pericentromeric duplication of the hominoid genome. *Genome Res* 9: 1048-58.

Eichler E. E., Lu F., Shen Y., Antonacci R., Jurecic V., Doggett N. A., Moyzis R. K., Baldini A., Gibbs R. A., and Nelson D. L. (1996). Duplication of a gene-rich cluster between 16p11.1 and Xq28: a novel pericentromeric-directed mechanism for paralogous genome evolution. *Hum Mol Genet* 5: 899-912.

Ellisen L. W., Bird J., West D. C., Soreng A. L., Reynolds T. C., Smith S. D., and Sklar J. (1991). TAN-1, the human homolog of the Drosophila notch gene, is broken by chromosomal translocations in T lymphoblastic neoplasms. *Cell* 66: 649-61.

Endo T., Imanishi T., Gojobori T., and Inoko H. (1997). Evolutionary significance of intra-genome duplications on human chromosomes. *Gene* 205: 19-27.

Erickson H. P. (1993). Tenascin-C, tenascin-R and tenascin-X: a family of talented proteins in search of functions. *Curr Opin Cell Biol* 5: 869-76.

Escriva H., Manzon L., Youson J., and Laudet V. (2002). Analysis of lamprey and hagfish genes reveals a complex history of gene duplications during early vertebrate evolution. *Mol Biol Evol* 19: 1440-50.

Evertsz E. M., Au-Young J., Ruvolo M. V., Lim A. C., and Reynolds M. A. (2001). Hybridization cross-reactivity within homologous gene families on glass cDNA microarrays. *Biotechniques* 31: 1182-1192.

Felsenstein J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* 17: 368-76.

Felsenstein J. (1985). Confidence-Limits On Phylogenies - an Approach Using the Bootstrap. *Evolution* 39: 783-791.

Felsenstein J. (1989). PHYLIP -- Phylogeny Inference Package (Version 3.2). *Cladistics* 5: 164-166.

Feng D. F., Johnson M. S., and Doolittle R. F. (1984). Aligning amino acid sequences: comparison of commonly used methods. *J Mol Evol* 21: 112-25.

Fiers W., Contreras R., Haegemann G., Rogiers R., Van de Voorde A., Van Heuverswyn H., Van Herreweghe J., Volckaert G., and Ysebaert M. (1978). Complete nucleotide sequence of SV40 DNA. *Nature* 273: 113-20.

Flajnik M. F., Canel C., Kramer J., and Kasahara M. (1991). Which came first, MHC class I or class II? *Immunogenetics* 33: 295-300.

Flajnik M. F., and Kasahara M. (2001). Comparative genomics of the MHC: glimpses into the evolution of the adaptive immune system. *Immunity* 15: 351-62.

Flajnik M. F., Ohta Y., Namikawa-Yamada C., and Nonaka M. (1999). Insight into the primordial MHC from studies in ectothermic vertebrates. *Immunol Rev* 167: 59-67.

Force A., Lynch M., Pickett F. B., Amores A., Yan Y. L., and Postlethwait J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151: 1531-45.

French C. A., Miyoshi I., Kubonishi I., Grier H. E., Perez-Atayde A. R., and Fletcher J. A. (2003). BRD4-NUT fusion oncogene: a novel mechanism in aggressive carcinoma. *Cancer Res* 63: 304-7.

Friedman R., and Hughes A. L. (2001). Pattern and timing of gene duplication in animal genomes. *Genome Res* 11: 1842-7.

Gallardo M. H., Bickham J. W., Honeycutt R. L., Ojeda R. A., and Kohler N. (1999). Discovery of tetraploidy in a mammal. *Nature* 401: 341.

Garcia-Fernandez J., and Holland P. (1994). Archetypal organisation of the amphioxus Hox gene cluster. *Nature* 370: 563-6.

Gaston K., and Fried M. (1994). YY1 is involved in the regulation of the bi-directional promoter of the Surf-1 and Surf-2 genes. *FEBS Lett* 347: 289-94.

Giglio S., Broman K. W., Matsumoto N., Calvari V., Gimelli G., Neumann T., Ohashi H., Voullaire L., Larizza D., Giorda R., Weber J. L., Ledbetter D. H., and Zuffardi O. (2001). Olfactory receptor-gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. *Am J Hum Genet* 68: 874-83.

Giglio S., Calvari V., Gregato G., Gimelli G., Camanini S., Giorda R., Ragusa A., Guerneri S., Selicorni A., Stumm M., Tonnies H., Ventura M., Zollino M., Neri G., Barber J., Wieczorek D., Rocchi M., and Zuffardi O. (2002). Heterozygous submicroscopic inversions involving olfactory receptor-gene clusters mediate the recurrent t(4;8)(p16;p23) translocation. *Am J Hum Genet* 71: 276-85.

Gordon, D., Abajian, C. and Green, P. (1998) Consed: a graphical tool for sequence finishing *Genome Res* 8 (3): 195-202

Greer J. M., Puetz J., Thomas K. R., and Capecchi M. R. (2000). Maintenance of functional equivalence during paralogous Hox gene evolution. *Nature* 403: 661-5.

Grewal P. K., van Geel M., Frants R. R., de Jong P., and Hewitt J. E. (1999). Recent amplification of the human FRG1 gene during primate evolution. *Gene* 227: 79-88.

Gruen J. R., and Weissman S. M. (2001). Human MHC class III and IV genes and disease associations. *Front Biosci* 6: D960-72.

Gu X., and Huang W. (2002). Testing the parsimony test of genome duplications: a counterexample. *Genome Res* 12: 1-2.

Habuchi T., Luscombe M., Elder P. A., and Knowles M. A. (1998). Structure and methylation-based silencing of a gene (DBCCR1) within a candidate bladder cancer tumor suppressor region at 9q32-q33. *Genomics* 48: 277-88.

Hall L., Williams K., Perry A. C., Frayne J., and Jury J. A. (1998). The majority of human glutathione peroxidase type 5 (GPX5) transcripts are incorrectly spliced: implications for the role of GPX5 in the male reproductive tract. *Biochem J* 333 (Pt 1): 5-9.

Heiskanen M., Karhu R., Hellsten E., Peltonen L., Kallioniemi O. P., and Palotie A. (1994). High resolution mapping using fluorescence in situ hybridization to extended DNA fibers prepared from agarose-embedded cells. *Biotechniques* 17: 928-9, 932-3.

Henikoff S., and Henikoff J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* 89: 10915-9.

Herberg J. A., Beck S., and Trowsdale J. (1998a). TAPASIN, DAXX, RGL2, HKE2 and four new genes (BING 1, 3 to 5) form a dense cluster at the centromeric end of the MHC. *J Mol Biol* 277: 839-57.

Herberg J. A., Sgouros J., Jones T., Copeman J., Humphray S. J., Sheer D., Cresswell P., Beck S., and Trowsdale J. (1998b). Genomic analysis of the Tapasin gene, located close to the TAP loci in the MHC. *Eur J Immunol* 28: 459-67.

Holland L. Z., Rached L. A., Tamme R., Holland N. D., Kortschak D., Inoko H., Shiina T., Burgtorf C., and Lardelli M. (2001). Characterization and developmental expression of the amphioxus homolog of Notch (AmphiNotch): evolutionary conservation of multiple expression domains in amphioxus and vertebrates. *Dev Biol* 232: 493-507.

Holland P. W. (2003). More genes in vertebrates? *J Struct Funct Genomics* 3: 75-84.

Holland P. W., Garcia-Fernandez J., Williams N. A., and Sidow A. (1994). Gene duplications and the origins of vertebrate development. *Dev Suppl*: 125-33.

Holzinger I., de Baey A., Messer G., Kick G., Zwierzina H., and Weiss E. H. (1995). Cloning and genomic characterization of LST1: a new gene in the human TNF region. *Immunogenetics* 42: 315-22.

Houzelstein D., Bullock S. L., Lynch D. E., Grigorieva E. F., Wilson V. A., and Beddington R. S. (2002). Growth and early postimplantation defects in mice deficient for the bromodomain-containing protein Brd4. *Mol Cell Biol* 22: 3794-802.

Hubbard T., Barker D., Birney E., Cameron G., Chen Y., Clark L., Cox T., Cuff J., Curwen V., Down T., Durbin R., Eyras E., Gilbert J., Hammond M., Huminiecki L., Kasprzyk A., Lehvaslaiho H., Lijnzaad P., Melsopp C., Mongin E., Pettett R., Pocock M., Potter S., Rust A., Schmidt E., Searle S., Slater G., Smith J., Spooner W., Stabenau A., Stalker J., Stupka E., Ureta-Vidal A., Vastrik I., and Clamp M. (2002). The Ensembl genome database project. *Nucleic Acids Res* 30: 38-41.

Hudson T. J., Stein L. D., Gerety S. S., Ma J., Castle A. B., Silva J., Slonim D. K., Baptista R., Kruglyak L., Xu S. H., and et al. (1995). An STS-based map of the human genome. *Science* 270: 1945-54.

Hughes A. L., and Nei M. (1993). Evolutionary relationships of the classes of major histocompatibility complex genes. *Immunogenetics* 37: 337-46.

Hughes A. L. (1994). Phylogeny of the C3/C4/C5 complement-component gene family indicates that C5 diverged first. *Mol Biol Evol* 11: 417-25.

Hughes A. L., and Yeager M. (1997). Molecular evolution of the vertebrate immune system. *Bioessays* 19: 777-86.

Hughes A. L. (1998). Phylogenetic tests of the hypothesis of block duplication of homologous genes on human chromosomes 6, 9, and 1. *Mol Biol Evol* 15: 854-70.

Hughes A. L., and Pontarotti P. (2000). Gene duplication and MHC origins. *Immunogenetics* 51: 982-3.

Hughes A. L., and Friedman R. (2003). 2R or not 2R: testing hypotheses of genome duplication in early vertebrates. *J Struct Funct Genomics* 3: 85-93.

Huxley C., and Fried M. (1990). The mouse surfeit locus contains a cluster of six genes associated with four CpG-rich islands in 32 kilobases of genomic DNA. *Mol Cell Biol* 10: 605-14.

International Human Genome Sequencing Consortium (IHGSC) (2001). Initial sequencing and analysis of the human genome. *Nature* 409: 860-921.

Ishiguro H., Kobayashi K., Suzuki M., Titani K., Tomonaga S., and Kurosawa Y. (1992). Isolation of a hagfish gene that encodes a complement component. *Embo J* 11: 829-37.

Jackson D. A., Pombo A., and Iborra F. (2000). The balance sheet for transcription: an analysis of nuclear RNA metabolism in mammalian cells. *Faseb J* 14: 242-54.

Jentsch T. J., Stein V., Weinreich F., and Zdebik A. A. (2002). Molecular structure and physiological function of chloride channels. *Physiol Rev* 82: 503-68.

Jones D. T., Taylor W. R., and Thornton J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* 8: 275-82.

Jones M. H., Numata M., and Shimane M. (1997). Identification and characterisation of BRDT: A testis-specific gene related to the bromodomain genes RING3 and Drosophila fsh. *Genomics* 45: 529-34.

Jurka J. (2000). Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet* 16: 418-20.

Karlin S., and Altschul S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci U S A* 87: 2264-8.

Kasahara M., Hayashi M., Tanaka K., Inoko H., Sugaya K., Ikemura T., and Ishibashi T. (1996a). Chromosomal localization of the proteasome Z subunit gene reveals an ancient chromosomal duplication involving the major histocompatibility complex. *Proc Natl Acad Sci U S A* 93: 9096-101.

Kasahara M., Kandil E., Salter-Cid L., and Flajnik M. F. (1996b). Origin and evolution of the class I gene family: why are some of the mammalian class I genes encoded outside the major histocompatibility complex? *Res Immunol* 147: 278-84; discussion 284-5.

Kasahara M. (1997). New insights into the genomic organization and origin of the major histocompatibility complex: role of chromosomal (genome) duplication in the emergence of the adaptive immune system. *Hereditas* 127: 59-65.

Kasahara M. (1999a). The chromosomal duplication model of the major histocompatibility complex. *Immunol Rev* 167: 17-32.

Kasahara M. (1999b). Genome dynamics of the major histocompatibility complex: insights from genome paralogy. *Immunogenetics* 50: 134-45.

Kasahara M., Yawata M., and Suzuki T. (2000). "The MHC paralogous group: Listing of members and a brief overview," Springer-Verlag,, Tokyo-Berlin-Heidelberg-New York.

Katsanis N., Fitzgibbon J., and Fisher E. M. (1996). Paralogy mapping: identification of a region in the human MHC triplicated onto human chromosomes 1 and 9 allows the prediction and isolation of novel PBX and NOTCH loci. *Genomics* 35: 101-8.

Kent W. J., and Haussler D. (2001). Assembly of the working draft of the human genome with GigAssembler. *Genome Res* 11: 1541-8.

Klein J., and Sato A. (1998). Birth of the major histocompatibility complex. *Scand J Immunol* 47: 199-209.

Kobayashi K., Nakahori Y., Miyake M., Matsumura K., Kondo-Iida E., Nomura Y., Segawa M., Yoshioka M., Saito K., Osawa M., Hamano K., Sakakihara Y., Nonaka I., Nakagome Y., Kanazawa I., Nakamura Y., Tokunaga K., and Toda T. (1998). An ancient retrotransposal insertion causes Fukuyama-type congenital muscular dystrophy. *Nature* 394: 388-92.

Kumar S., Tamura K., and Nei M. (1994). MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers. *Comput Appl Biosci* 10: 189-91.

Kumar S., Tamura K., Jakobsen I. B., and Nei M. (2001). MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* 17: 1244-5.

Lacazette E., Gachon A. M., and Pitiot G. (2000). A novel human odorant-binding protein gene family resulting from genomic duplicons at 9q34: differential expression in the oral and genital spheres. *Hum Mol Genet* 9: 289-301.

Lardelli M., Dahlstrand J., and Lendahl U. (1994). The novel Notch homologue mouse Notch 3 lacks specific epidermal growth factor-repeats and is expressed in proliferating neuroepithelium. *Mech Dev* 46: 123-36.

Larhammar D., Lundin L. G., and Hallbook F. (2002). The Human Hox-bearing Chromosome Regions Did Arise by Block or Chromosome (or Even Genome) Duplications. *Genome Res* 12: 1910-20.

Lennard A., Gaston K., and Fried M. (1994). The Surf-1 and Surf-2 genes and their essential bidirectional promoter elements are conserved between mouse and human. *DNA Cell Biol* 13: 1117-26.

Lesk A. (2002). "Introduction to bioinformatics," Oxford University Press, Oxford.

Lewis J. (1998). Notch signalling and the control of cell fate choices in vertebrates. *Semin Cell Dev Biol* 9: 583-9.

Lewis S. A., and Cowan N. J. (1990). Tubulin genes: structure, expression, and regulation. *In* "Microtubule proteins" (J. Avila, Ed.), pp. 37-66, CRC Press, Inc, Boca Raton.

Li W.-H. (1997). "Molecular Evolution," Sunderland Sinauer.

Li W. H., Gu Z., Wang H., and Nekrutenko A. (2001). Evolutionary analyses of the human genome. *Nature* 409: 847-9.

Lundin L. G. (1993). Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. *Genomics* 16: 1-19.

Lundin L. G., and Larhammar D. (1998). "Paralogous genes and nervous systems," Oxford, UK.

Lundin L. G., Larhammar D., and Hallbook F. (2003). Numerous groups of chromosomal regional paralogies strongly indicate two genome doublings at the root of the vertebrates. *J Struct Funct Genomics* 3: 53-63.

Lynch M., and Conery J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science* 290: 1151-5.

Lynch M., and Force A. (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154: 459-73.

Mackiewicz P., Kowalczuk M., Gierlik A., Dudek M. R., and Cebrat S. (1999). Origin and properties of non-coding ORFs in the yeast genome. *Nucleic Acids Res* 27: 3503-9.

Mackiewicz P., Kowalczuk M., Mackiewicz D., Nowicka A., Dudkiewicz M., Laszkiewicz A., Dudek M. R., and Cebrat S. (2002). How many protein-coding genes are there in the Saccharomyces cerevisiae genome? *Yeast* 19: 619-29.

Makalowski W. (2001). Are we polyploids? A brief history of one hypothesis. *Genome Res* 11: 667-70.

Mangelsdorf D. J., Borgmeyer U., Heyman R. A., Zhou J. Y., Ong E. S., Oro A. E., Kakizuka A., and Evans R. M. (1992). Characterisation of three RXR genes that mediate the action of 9-cis retinoic acid. *Genes Dev* 6: 329-44.

Maresco D. L., Chang E., Theil K. S., Francke U., and Anderson C. L. (1996). The three genes of the human FCGR1 gene family encoding Fc gamma RI flank the centromere of chromosome 1 at 1p12 and 1q21. *Cytogenet Cell Genet* 73: 157-63.

Marshall E. (2001). Genome sequencing. Celera assembles mouse genome; public labs plan new strategy. *Science* 292: 822.

Marzluff W. F., Gongidi P., Woods K. R., Jin J., and Maltais L. J. (2002). The human and mouse replication-dependent histone genes. *Genomics* 80: 487-98.

Mazet F., and Shimeld S. M. (2002). Gene duplication and divergence in the early evolution of vertebrates. *Curr Opin Genet Dev* 12: 393-6.

McKean P. G., Vaughan S., and Gull K. (2001). The extended tubulin superfamily. *J Cell Sci* 114: 2723-33.

McLysaght A., Hokamp K., and Wolfe K. H. (2002). Extensive genomic duplication during early chordate evolution. *Nat Genet* 31: 200-4.

Mewes H. W., Albermann K., Bahr M., Frishman D., Gleissner A., Hani J., Heumann K., Kleine K., Maierl A., Oliver S. G., Pfeiffer F., and Zollner A. (1997). Overview of the yeast genome. *Nature* 387: 7-65.

Miyata T., and Suga H. (2001). Divergence pattern of animal gene families and relationship with the Cambrian explosion. *Bioessays* 23: 1018-27.

Modrek B., and Lee C. (2002). A genomic view of alternative splicing. *Nat Genet* 30: 13-9.

Monaco J. J. (1992). A molecular model of MHC class-I-restricted antigen processing. *Immunol Today* 13: 173-9.

Monica K., Galili N., Nourse J., Saltman D., and Cleary M. L. (1991). PBX2 and PBX3, new homeobox genes with extensive homology to the human proto-oncogene PBX1. *Mol Cell Biol* 11: 6149-57.

Muller H. J. (1925). Why polyploidy is rarer in animals than in plants. *American Naturalist* 59: 346-353.

Nadeau J. H., and Sankoff D. (1997). Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. *Genetics* 147: 1259-66.

Nagata T., Weiss E. H., Abe K., Kitagawa K., Ando A., Yara-Kikuti Y., Seldin M. F., Ozato K., Inoko H., and Taketo M. (1995). Physical mapping of the retinoid X receptor B gene in mouse and human. *Immunogenetics* 41: 83-90.

Neefjes J. J., and Ploegh H. L. (1992). Intracellular transport of MHC class II molecules. *Immunol Today* 13: 179-84.

Nei M., and Kumar S. (2000). "Molecular Evolution and Phylogenetics," Oxford University Press, New York.

Nicholls A. C., McCarron S., Narcisi P., and Pope F. M. (1994). Molecular abnormalities of type V collagen in Ehlers Danlos syndrome. *Am. J. Hum. Genet* 55: A233.

Nonaka M., and Takahashi M. (1992). Complete complementary DNA sequence of the third component of complement of lamprey. Implication for the evolution of thioester containing proteins. *J Immunol* 148: 3290-5.

Ohno S. (1970). "Evolution by Gene Duplication," Springer-Verlag, New York.

Ohno S. (1973). Ancient linkage groups and frozen accidents. *Nature* 244: 259-62.

Oliver S. G., van der Aart Q. J., Agostoni-Carbone M. L., Aigle M., Alberghina L., Alexandraki D., Antoine G., Anwar R., Ballesta J. P., Benit P., and et al. (1992). The complete DNA sequence of yeast chromosome III. *Nature* 357: 38-46.

Olson M. V., Dutchik J. E., Graham M. Y., Brodeur G. M., Helms C., Frank M., MacCollin M., Scheinman R., and Frank T. (1986). Random-clone strategy for genomic restriction mapping in yeast. *Proc Natl Acad Sci U S A* 83: 7826-30.

Ortmann B., Androlewicz M. J., and Cresswell P. (1994). MHC class I/beta 2-microglobulin complexes associate with TAP transporters before peptide binding. *Nature* 368: 864-7.

Padgett R. A., Grabowski P. J., Konarska M. M., Seiler S., and Sharp P. A. (1986). Splicing of messenger RNA precursors. *Annu Rev Biochem* 55: 1119-50.

Panopoulou G., Hennig S., Groth D., Krause A., Poustka A. J., Herwig R., Vingron M., and Lehrach H. (2003). New evidence for genome-wide duplications at the origin of vertebrates using an amphioxus gene set and completed animal genomes. *Genome Res* 13: 1056-66.

Pébusque M. J., Coulier F., Birnbaum D., and Pontarotti P. (1998). Ancient large-scale genome duplications: phylogenetic and linkage analyses shed light on chordate genome evolution. *Mol Biol Evol* 15: 1145-59.

Perry A. C., Jones R., Niang L. S., Jackson R. M., and Hall L. (1992). Genetic evidence for an androgen-regulated epididymal secretory glutathione peroxidase whose transcript does not contain a selenocysteine codon. *Biochem J* 285 (Pt 3): 863-70.

Pieters J. (1997). MHC class II restricted antigen presentation. *Curr Opin Immunol* 9: 89-96.

Pinkel D., Gray J. W., Trask B., van den Engh G., Fuscoe J., and van Dekken H. (1986). Cytogenetic analysis by in situ hybridization with fluorescently labeled nucleic acid probes. *Cold Spring Harb Symp Quant Biol* 51 Pt 1: 151-7.

Pollard S. L., and Holland P. W. (2000). Evidence for 14 homeobox gene clusters in human genome ancestry. *Curr Biol* 10: 1059-62.

Postlethwait J. H., Yan Y. L., Gates M. A., Horne S., Amores A., Brownlie A., Donovan A., Egan E. S., Force A., Gong Z., Goutel C., Fritz A., Kelsh R., Knapik E.,

Liao E., Paw B., Ransom D., Singer A., Thomson M., Abduljabbar T. S., Yelick P., Beier D., Joly J. S., Larhammar D., Rosa F., and et al. (1998). Vertebrate genome evolution and the zebrafish gene map. *Nat Genet* 18: 345-9.

Price P., Witt C., Allcock R., Sayer D., Garlepp M., Kok C. C., French M., Mallal S., and Christiansen F. (1999). The genetic basis for the association of the 8.1 ancestral haplotype (A1, B8, DR3) with multiple immunopathological diseases. *Immunol Rev* 167: 257-74.

Prince V. E., and Pickett F. B. (2002). Splitting pairs: the diverging fates of duplicated genes. *Nat Rev Genet* 3: 827-37.

Pryde F. E., Gorham H. C., and Louis E. J. (1997). Chromosome ends: all the same under their caps. *Curr Opin Genet Dev* 7: 822-8.

Radley E., Alderton R. P., Kelly A., Trowsdale J., and Beck S. (1994). Genomic organization of HLA-DMA and HLA-DMB. Comparison of the gene organization of all six class II families in the human major histocompatibility complex. *J Biol Chem* 269: 18834-8.

Radosavljevic M., Cuillerier B., Wilson M. J., Clement O., Wicker S., Gilfillan S., Beck S., Trowsdale J., and Bahram S. (2002). A cluster of ten novel MHC class I related genes on human chromosome 6q24.2-q25.3. *Genomics* 79: 114-23.

Reid K. B., and Porter R. R. (1981). The proteolytic activation systems of complement. *Annu Rev Biochem* 50: 433-64.

Reiter L. T., Murakami T., Koeuth T., Pentao L., Muzny D. M., Gibbs R. A., and Lupski J. R. (1996). A recombination hotspot responsible for two inherited peripheral neuropathies is located near a mariner transposon-like element. *Nat Genet* 12: 288-97.

Rice P., Longden I., and Bleasby A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 16: 276-7.

Rouquier S., Taviaux S., Trask B. J., Brand-Arpon V., van den Engh G., Demaille J., and Giorgi D. (1998). Distribution of olfactory receptor genes in the human genome. *Nat Genet* 18: 243-50.

Rozen S., and Skaletsky H. (2000). Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* 132: 365-86.

Saglio G., Storlazzi C. T., Giugliano E., Surace C., Anelli L., Rege-Cambrin G., Zagaria A., Jimenez Velasco A., Heiniger A., Scaravaglio P., Torres Gomez A., Roman Gomez J., Archidiacono N., Banfi S., and Rocchi M. (2002). A 76-kb duplicon maps close to the BCR gene on chromosome 22 and the ABL gene on chromosome 9: possible involvement in the genesis of the Philadelphia chromosome translocation. *Proc Natl Acad Sci U S A* 99: 9882-7.

Saitou N., and Nei M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4: 406-25.

Sanger F., Air G. M., Barrell B. G., Brown N. L., Coulson A. R., Fiddes C. A., Hutchison C. A., Slocombe P. M., and Smith M. (1977a). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* 265: 687-95.

Sanger F., Nicklen S., and Coulson A. R. (1977b). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* 74: 5463-7.

Sanger F., Coulson A. R., Friedmann T., Air G. M., Barrell B. G., Brown N. L., Fiddes J. C., Hutchison C. A., 3rd, Slocombe P. M., and Smith M. (1978). The nucleotide sequence of bacteriophage phiX174. *J Mol Biol* 125: 225-46.

Sanger F., Coulson A. R., Hong G. F., Hill D. F., and Petersen G. B. (1982). Nucleotide sequence of bacteriophage lambda DNA. *J Mol Biol* 162: 729-73.

Schmidt H. A., Strimmer K., Vingron M., and von Haeseler A. (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18: 502-4.

Schughart K., Kappen C., and Ruddle F. H. (1988). Mammalian homeobox-containing genes: genome organization, structure, expression and evolution. *Br J Cancer Suppl* 9: 9-13.

Schultz R. J. (1980). Role of polyploidy in the evolution of fishes. *In* "Polyploidy: Biological Relevance" (W. H. Lewis, Ed.), Plenum, New York.

Sharman A. C. (1999). Some new terms for duplicated genes. *Semin Cell Dev Biol* 10: 561-3.

Shiina T., Tamiya G., Oka A., Takishima N., Yamagata T., Kikkawa E., Iwata K., Tomizawa M., Okuaki N., Kuwano Y., Watanabe K., Fukuzumi Y., Itakura S., Sugawara C., Ono A., Yamazaki M., Tashiro H., Ando A., Ikemura T., Soeda E., Kimura M., Bahram S., and Inoko H. (1999). Molecular dynamics of MHC genesis unraveled by sequence analysis of the 1,796,938-bp HLA class I region. *Proc Natl Acad Sci U S A* 96: 13282-7.

Shiina T., Ando A., Suto Y., Kasai F., Shigenari A., Takishima N., Kikkawa E., Iwata K., Kuwano Y., Kitamura Y., Matsuzawa Y., Sano K., Nogami M., Kawata H., Li S., Fukuzumi Y., Yamazaki M., Tashiro H., Tamiya G., Kohda A., Okumura K., Ikemura T., Soeda E., Mizuki N., Kimura M., Bahram S., and Inoko H. (2001). Genomic anatomy of a premier major histocompatibility complex paralogous region on chromosome 1q21-q22. *Genome Res* 11: 789-802.

Sidow A. (1996). Gen(om)e duplications in the evolution of early vertebrates. *Curr Opin Genet Dev* 6: 715-22.

Skrabanek L., and Wolfe K. H. (1998). Eukaryote genome duplication - where's the evidence? *Curr Opin Genet Dev* 8: 694-700.

Smit A. F. (1999). Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev* 9: 657-63.

Smith L. C., Azumi K., and Nonaka M. (1999). Complement systems in invertebrates. The ancient alternative and lectin pathways. *Immunopharmacology* 42: 107-20.

Soderlund C., Longden I., and Mott R. (1997). FPC: a system for building contigs from restriction fingerprinted clones. *Comput Appl Biosci* 13: 523-35.

Spring J., Goldberger O. A., Jenkins N. A., Gilbert D. J., Copeland N. G., and Bernfield M. (1994). Mapping of the syndecan genes in the mouse: linkage with members of the myc gene family. *Genomics* 21: 597-601.

Spring J. (1997). Vertebrate evolution by interspecific hybridisation--are we polyploid? *FEBS Lett* 400: 2-8.

Staden R. (1980). A new computer method for the storage and manipulation of DNA gel reading data. *Nucleic Acids Res* 8: 3673-94.

Staden R., Beal K. F., and Bonfield J. K. (2000). The Staden package, 1998. *Methods Mol Biol* 132: 115-30.

Stephens R., Horton R., Humphray S., Rowen L., Trowsdale J., and Beck S. (1999). Gene organisation, sequence variation and isochore structure at the centromeric boundary of the human MHC. *J Mol Biol* 291: 789-99.

Strippoli P., D'Addabbo P., Lenzi L., Giannone S., Canaider S., Casadei R., Vitale L., Carinci P., and Zannotti M. (2002). Segmental paralogy in the human genome: a large-scale triplication on 1p, 6p, and 21q. *Mamm Genome* 13: 456-62.

Sugaya K., Fukagawa T., Matsumoto K., Mita K., Takahashi E., Ando A., Inoko H., and Ikemura T. (1994). Three genes in the human MHC class III region near the junction with the class II: gene for receptor of advanced glycosylation end products, PBX2 homeobox gene and a notch homolog, human counterpart of mouse mammary tumor gene int-3. *Genomics* 23: 408-19.

Sugaya K., Sasanuma S., Nohata J., Kimura T., Fukagawa T., Nakamura Y., Ando A., Inoko H., Ikemura T., and Mita K. (1997). Gene organization of human NOTCH4 and (CTG)n polymorphism in this human counterpart gene of mouse proto-oncogene Int3. *Gene* 189: 235-44.

Tamkun J. W. (1995). The role of brahma and related proteins in transcription and development. *Curr Opin Genet Dev* 5: 473-7.

Tee M. K., Thomson A. A., Bristow J., and Miller W. L. (1995). Sequences promoting the transcription of the human XA gene overlapping P450c21A correctly predict the presence of a novel, adrenal-specific, truncated form of tenascin-X. *Genomics* 28: 171-8.

The MHC Sequencing Consortium (1999). Complete sequence and gene map of a human major histocompatibility complex. The MHC sequencing consortium. *Nature* 401 (6756): 921-3.

Thompson J. D., Higgins D. G., and Gibson T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22: 4673-80.

Thomson G. (1995). HLA disease associations: models for the study of complex human genetic disorders. *Crit Rev Clin Lab Sci* 32: 183-219.

Thorpe K. L., Gorman P., Thomas C., Sheer D., Trowsdale J., and Beck S. (1997). Chromosomal localization, gene structure and transcription pattern of the ORFX gene, a homologue of the MHC-linked RING3 gene. *Gene* 200: 177-83.

Utans U., Arceci R. J., Yamashita Y., and Russell M. E. (1995). Cloning and characterization of allograft inflammatory factor-1: a novel macrophage factor identified in rat cardiac allografts with chronic rejection. *J Clin Invest* 95: 2954-62.

Utans U., Quist W. C., McManus B. M., Wilson J. E., Arceci R. J., Wallace A. F., and Russell M. E. (1996). Allograft inflammatory factory-1. A cytokine-responsive macrophage molecule expressed in transplanted human hearts. *Transplantation* 61: 1387-92.

van Geel M., Heather L. J., Lyle R., Hewitt J. E., Frants R. R., and de Jong P. J. (1999). The FSHD region on human chromosome 4q35 contains potential coding regions among pseudogenes and a high density of repeat elements. *Genomics* 61: 55-65.

van Geel M., Eichler E. E., Beck A. F., Shan Z., Haaf T., van der Maarel S. M., Frants R. R., and de Jong P. J. (2002). A cascade of complex subtelomeric duplications during the evolution of the hominoid and Old World monkey genomes. *Am J Hum Genet* 70: 269-78.

van Slegtenhorst M., de Hoogt R., Hermans C., Nellist M., Janssen B., Verhoef S., Lindhout D., van den Ouweland A., Halley D., Young J., Burley M., Jeremiah S., Woodward K., Nahmias J., Fox M., Ekong R., Osborne J., Wolfe J., Povey S., Snell R. G., Cheadle J. P., Jones A. C., Tachataki M., Ravine D., Kwiatkowski D. J., and et al. (1997). Identification of the tuberous sclerosis gene TSC1 on chromosome 9q34. *Science* 277: 805-8.

Venter J. C., Adams M. D., Myers E. W., Li P. W., Mural R. J., Sutton G. G., Smith H. O., Yandell M., Evans C. A., Holt R. A., Gocayne J. D., Amanatides P., Ballew R. M., Huson D. H., Wortman J. R., Zhang Q., Kodira C. D., Zheng X. H., Chen L., Skupski M., Subramanian G., Thomas P. D., Zhang J., Gabor Miklos G. L., Nelson C., Broder S., Clark A. G., Nadeau J., McKusick V. A., Zinder N., Levine A. J., Roberts R. J., Simon M., Slayman C., Hunkapiller M., Bolanos R., Delcher A., Dew I., Fasulo D., Flanigan M., Florea L., Halpern A., Hannenhalli S., Kravitz S., Levy S., Mobarry C., Reinert K., Remington K., Abu-Threideh J., Beasley E., Biddick K.,

Bonazzi V., Brandon R., Cargill M., Chandramouliswaran I., Charlab R., Chaturvedi K., Deng Z., Di Francesco V., Dunn P., Eilbeck K., Evangelista C., Gabrielian A. E., Gan W., Ge W., Gong F., Gu Z., Guan P., Heiman T. J., Higgins M. E., Ji R. R., Ke Z., Ketchum K. A., Lai Z., Lei Y., Li Z., Li J., Liang Y., Lin X., Lu F., Merkulov G. V., Milshina N., Moore H. M., Naik A. K., Narayan V. A., Neelam B., Nusskern D., Rusch D. B., Salzberg S., Shao W., Shue B., Sun J., Wang Z., Wang A., Wang X., Wang J., Wei M., Wides R., Xiao C., Yan C., et al. (2001). The sequence of the human genome. *Science* 291: 1304-51.

Vernier P., Mastrippolito R., Helin C., Bendali M., Mallet J., and Tricoire H. (1996). Radioimager quantification of oligonucleotide hybridization with DNA immobilized on transfer membrane: application to the identification of related sequences. *Anal Biochem* 235: 11-9.

von Lindern M., van Baal S., Wiegant J., Raap A., Hagemeijer A., and Grosveld G. (1992). Can, a putative oncogene associated with myeloid leukemogenesis, may be activated by fusion of its 3-prime half to different genes: characterization of the 'set' gene. *Mol. Cell. Biol.* 12: 3346-3355.

Wagner A. (2001). Birth and death of duplicated genes in completely sequenced eukaryotes. *Trends Genet* 17: 237-9.

Walter M. A., Spillett D. J., Thomas P., Weissenbach J., and Goodfellow P. N. (1994). A method for constructing radiation hybrid maps of whole genomes. *Nat Genet* 7: 22-8.

Wang Y., and Gu X. (2000). Evolutionary patterns of gene families generated in the early stage of vertebrates. *J Mol Evol* 51: 88-96.

Wendel J. F. (2000). Genome evolution in polyploids. *Plant Mol Biol* 42: 225-49.

Wilke C. M., Guo S. W., Hall B. K., Boldog F., Gemmill R. M., Chandrasekharappa S. C., Barcroft C. L., Drabkin H. A., and Glover T. W. (1994). Multicolor FISH mapping of YAC clones in 3p14 and identification of a YAC spanning both FRA3B and the t(3;8) associated with hereditary renal cell carcinoma. *Genomics* 22: 319-26.

Wilson R., Ainscough R., Anderson K., Baynes C., Berks M., Bonfield J., Burton J., Connell M., Copsey T., Cooper J., and et al. (1994). 2.2 Mb of contiguous nucleotide sequence from chromosome III of C. elegans. *Nature* 368: 32-8.

Wittbrodt J., Meyer A., and Schartl M. (1998). More genes in fish? *Bioessays* 20: 511-515.

Wolfe K. H. (2001). Yesterday's polyploids and the mystery of diploidization. *Nat Rev Genet* 2: 333-41.

Xia Y., Brown L., Yang C. Y., Tsan J. T., Siciliano M. J., Espinosa R., III, Le Beau M. M., and Baer R. J. (1991). TAL2, a helix-loop-helix gene activated by the (7;9)(q34;q32) translocation in human T-cell leukemia. *Proc Natl Acad Sci U S A* 88: 11416-20.

Yon J., Jones T., Garson K., Sheer D., and Fried M. (1993). The organization and conservation of the human Surfeit gene cluster and its localization telomeric to the c-abl and can proto-oncogenes at chromosome band 9q34.1. *Hum Mol Genet* 2: 237-40.

Zhu Z., Yao J., Johns T., Fu K., De Bie I., Macmillan C., Cuthbert A. P., Newbold R. F., Wang J., Chevrette M., Brown G. K., Brown R. M., and Shoubridge E. A. (1998). SURF1, encoding a factor involved in the biogenesis of cytochrome c oxidase, is mutated in Leigh syndrome. *Nat Genet* 20: 337-43.

Appendix 1

Summary of the annotation of the chromosomal region 9q32 to 9q34.3. Putative paralogues are in bold text.

Clone		*Ensembl Symbol*	*SPTR, Refseq or Ensembl entry*	*Transcript Start*	*Transcript end*	*Protein length (aa)*	*Description*
AL160275	**q32**	**ATP6V1G1**	**O75348**	**109082196**	**109092823**	**118**	**Vacuolar ATP synthase subunit**
	q32	NM_153045	Q8ND43	109118880	109140873	281	Unknown
AL390240	**q32**	**TNFSF15**	**O95150**	**109283763**	**109300582**	**251**	**Tumor necrosis factor ligand**
AL133412	**q32**	**TNFSF8**	**P32971**	**109398384**	**109424753**	**234**	**Tumor necrosis factor ligand**
AL162425	**q33.1**	**TNC**	**P24821**	**109514975**	**109612609**	**2201**	**Tenascin precursor**
AL355601	q33.1	NM_017418	O17418	109636267	109897093	95	Deleted in esophageal cancer 1
AL731824	q33.1	No genes					
AL714001	q33.1	No genes					
AL691420	q33.1	No genes					
AL731897	q33.1	No genes					
AL731813	q33.1	No genes					
AL732367	q33.1	EST-YD1	Q9P2X8	110398837	110399090	84	EST-YD1 protein
AL691426	q33.1	No genes					
AL353141	q33.1	No genes					
AL137024	q33.1	PAPPA	Q13219	110677328	110808083	716	Pregnancy associated plasma protein A
AL669963	q33.1	No genes					
AL133282	q33.1	ASTN2	O75129	110836698	111823883	1321	Astrotactin 1
	q33.1	Novel	ENSG00000179990	110963536	110981401	73	Unknown
AL133284	q33.1	TRIM32	Q13049	111098800	111112220	653	Zing finger protein HT2A
AL157829	q33.1	Novel	ENSG00000136913	111196159	111196227	23	Unknown
	q33.1	FLJ20958	Q9BQ00	111205755	111205826	24	Unknown
AL392085	q33.1	No genes					
AL354981	q33.1	No genes					
AL355608	q33.1	No genes					
AL358792	q33.1	No genes					
AL445644	q33.1	No genes					
AL161630	q33.1	Novel	ENSG00000179956	112057445	112065871	72	Unknown
AL160272	q33.1	TLR4	O00206	112113140	112124614	839	TOLL-like receptor 4 precursor
AL354754	q33.1	No genes					
AL445663	q33.1	No genes					
AL158831	q33.1	No genes					
AL365195	q33.1	No genes					
AL445440	q33.1	No genes					
AL355592	q33.1	No genes					
AL589703	q33.1	No genes					

AL157780	q33.1	No genes					
AL512602	q33.1	No genes					
AL445310	q33.1	No genes					
AL353773	q33.1	DCBBR1	O14618	113575169	113778264	761	Deleted in bladder cancer chromosome region
AL138894	q33.1	No genes					
AL353630	q33.1	No genes					
AC006288	q33.1	No genes					
AL445683	q33.1	No genes					
AL354931	q33.1	No genes					
AL441989	q33.1	No genes					
AL355589	q33.1	No genes					
AL592549	q33.1	No genes					
AL353736	q33.1	No genes					
AL391870	q33.2	CDK5RAP2	Q9NV90	114797706	114988994	943	CDK5 regulatory subunit associated protein 2
AL590642	q33.2	No genes					
AL138836	q33.2	EGFL5	Q9H1U4	115009649	115068409	401	EGF like domain multiple 5 protein
	q33.2	Novel	ENSG00000176341	115098762	115122988	236	Unknown
AL161911	q33.2	FBXW2	Q9UKT8	115160814	115202253	454	F-BOX/WD-repeat protein 2
	q33.2	PSMD5	Q16401	115224889	115251748	504	26S proteasome non-ATPase regulatory subunit 5
	q33.2	Novel	ENSG00000180095	115251902	115263209	105	Unknown
AL354792	q33.2	Q9UFS9	Q9UFS9	115264535	115286029	473	Transcription factor
AC006430	q33.2	PRO1995	Q9P1F7	115299108	115300075	105	Unknown
	q33.2	TRAF1	EBI6	115311227	115337603	350	TNF receptor associated factor 1
	q33.2	**C5**	**P01031**	**115361172**	**11549110**	**1676**	**Complement C5 precursor**
AL137068	q33.2	Novel	ENSG00000171635	115499108	115516745	219	Testis specfic
	q33.2	CEP1	O07018	115521379	115586444	1800	Centrosomal protein 1
	q33.2	RAB14	P35287	115586971	115610724	215	Ras-related protein
AL513122	q33.2	Novel	ENSG00000180552	115647510	115648753	409	Unknown
	q33.2	MOST2	Q9NRJ2	115689602	115694364	209	MOST2 protein
	q33.2	GSN	O06396	115708681	115741676	782	Gelosin precursor, plasma
AL161784	q33.2	EPB72	P27105	115747913	115779060	288	Erythrocyte band 7 integral membrane protein
AL359644	q33.2	Novel	ENSG00000165196	115868813	115888025	174	Unknown
AL357936	q33.2	No genes					
AL365274	q33.2	DAP2IP	Q8TDL2	115974718	116194365	964	DOC-2/DAB2 interactive protein
AL450285	q33.2	Novel					
AL596244	q33.2	Novel					
AL445587	q33.2	Novel	ENSG00000171539	116362696	116383814	140	Unknown
AL442634	q33.2	Q8NHH0	Q8NHH0	116397745	116502441	538	Unknown
AL162423	q33.2	NDUFA8	P51970	116552893	116568579	172	NADH-Ubiquinone oxidoreductase subunit
AL162424	q33.2	LHX6	Q9UPM6	116611414	11667540	363	LIM/Homeobox protein
	q33.2	NM_033117	Q96H35	116648845	11667367	190	Unknown
	q33.2	NM_138777	Q9BU92	116679683	116732299	262	RIKEN cDNA D02
	q33.2	PTGS1	P23219	116779785	116804538	596	Prostaglandin G/H synthase 1 precursor
AL359636	**q33.2**	**OR**	**Q8NGS3**	**116885796**	**116886761**	**322**	**Olfactory receptor**
		OR	**OR1J5**	**116919637**	**116920575**	**313**	**Olfactory receptor**
		OR	**Q8NGS1**	**116927976**	**116928914**	**313**	**Olfactory receptor**
		OR	**OR1N1**	**116935199**	**116936125**	**309**	**Olfactory receptor**
		OR	**Q8NGR9**	**116960247**	**116962994**	**316**	**Olfactory receptor**
		OR	**Q8NGR8**	**116976386**	**116977312**	**309**	**Olfactory receptor**
Al162254	**q33.2**	**OR**	**Q9UDD7**	**117016871**	**117017632**	**254**	**Olfactory receptor**
		OR	**OR1Q1**	**117023573**	**117024514**	**314**	**Olfactory receptor**
AC006313	**q33.2**	**OR**	**Q8NGR6**	**117037417**	**117038367**	**295**	**Olfactory receptor**
		OR	**Q8NH94**	**117070551**	**117071480**	**317**	**Olfactory receptor**
		OR	**Q8NH93**	**117083965**	**117084936**	**324**	**Olfactory receptor**
		OR	**Q8NGR5**	**117132825**	**117133757**	**311**	**Olfactory receptor**

		OR	**Q96R80**	**117158887**	**117159534**	**216**	**Olfactory receptor**
		OR	**Q8WVK7**	**117170703**	**117171065**	**121**	**Olfactory receptor**
AL359512	**q33.2**	**OR**	**Q8NGR3**	**117208958**	**117209905**	**316**	**Olfactory receptor**
	q33.2	PDCL	Q13371	117226985	117237466	301	Phosdusin-like protein
	q33.2	MNAB	O18835	117253391	117314050	1191	Membrane associated binding protein
AL731645	q33.2	ZID	Q15916	117317457	117320910	424	Zinc finger protein
	q33.2	**BIOR**	**Q9HCK0**	**117326928**	**117340094**	**441**	**Zinc finger protein**
	q33.2	GAPCenA	Q9Y3P9	117349852	117513707	997	RAB6 GTPase activating protein
AL358946	q33.3	No genes					
AL365338	q33.2	NM_030814	Q9H2N8	117518337	117522309	167	Unknown
	q33.2	STRBP	Q96S19	117533552	117593141	658	RNA binding protein
AL365504	q33.3	PRO226	Q9P180	117674928	117675167	80	Unknown
AL445489	q33.3	FLJ38464	Q8N930	117782642	117787587	215	Unknown
	q33.3	FLJ00224	Q8TEH3	117788495	118338969	896	Unknown
AL161790	q33.3	No genes					
AL390774	q33.3	No genes					
AL158208	q33.3	No genes					
AC006450	q33.3	LHX2	P50458	118420439	118441992	406	LIM/Homeobox protein
AL158052	q33.3	No genes					
Al445284	q33.3	No genes					
AL162724	q33.3	NEK6	Q9HC98	118666435	118761271	338	Serine-threonine protein kinase
AL137846	**q33.3**	**PSMB7**	**Q99436**	**118762294**	**118824271**	**277**	**Proteasome subunit beta type 7**
	q33.3	Q8NH12	Q8NH12	118874053	118892161	984	Seven transmembrane helix receptor
AL354979	q33.3	NR5A1	Q13285	118890062	118916249	461	Steroidogenic factor 1
	q33.3	NR6A1	Q15406	118928956	119180139	476	Orphan nuclear receptor
AL669818	q33.3	No genes					
AL158075	q33.3	No genes					
AL354928	q33.3	FLJ90228	Q8NCI9	119186100	119223707	318	Unknown
	q33.3	Novel	ENSG00000136918	119263034	119266399	233	Unknown
	q33.3	RPL35	P42766	119266713	119270796	141	60S ribosomal protein L35
	q33.3	NM_030978	Q9BPX5	119278120	119286561	159	Actin related protein
	q33.3	GOLGA1	Q92805	119287196	119349928	767	Golgin 97,gap junction protein
AL451125	q33.3	FLJ40705	Q8N1I4	119360938	119552351	629	Unknown
AL445930	q33.3	PPP6C	O00743	119557957	119598620	305	Serine/threonine protein phospatase 6
	q33.3	Novel	ENSG00000173602	119603477	119604650	282	40S ribosomal protein
AL354710	q33.3	P40	O00568	119609374	119642831	372	RAB9 effector P40
	q33.3	**HSPA5**	**P11021**	**119643682**	**119650159**	**654**	**78 KDA Glucose regulated protein**
	q33.3	Novel	ENSG00000176094	119672414	119672919	158	40S ribosomal protein
AL627223	q33.3	FLJ20119	Q9NXQ1	119670661	119773833	833	Unknown
AL359632	q33.3	MAPKAP1	Q9BPZ7	119846225	120116031	486	MAP kinase interacting protein 1
AL162584	q33.3	Novel	ENSG00000178022	120004502	120005454	316	Unknown
AL358074	q33.3	SIN1		119846225	120116031	522	SAPK interacting protein 1
	q33.3	NM_016158	Q9UN39	120153187	120155380	129	Erythrocyte transmembrane protein
AL627303	q33.3	No genes					
AL445186	**q33.3**	**PBX3**	**P40426**	**120156161**	**120376205**	**434**	**Pre-B-cell leukaemia transcription factor 3**
AL589923	q33.3	No genes					
AL445664	q33.3	No genes					
AL162391	q33.3	No genes					
AC006443	q33.3	FLJ00022	Q9H7P6	120735668	120915859	344	Unknown
AL356309	q33.3	No genes					
AL161908	q33.3	No genes					
AL161731	q33.3	LMX1B	O60663	121023337	121105270	379	LIM/Homeobox protein
	q33.3	Q8N243	Q8N243	121212014	121214492	115	Unknown
	q33.3	**ZNF297B**	**O43298**	**121213845**	**121244041**	**467**	**Zinc finger 297B**
AL354944	q33.3	KIAA1993	Q8NCN2	121269484	121289729	532	Unknown

AL160169	q33.3	No genes					
AL356862	q33.3	RalGPS1A	Q8WUV7	121323592	121631982	590	Ral guanine nuceotide factor
	q33.3	ANGPTL2	Q9UKU9	121496150	121531456	493	Angiopoietin-related protein 2
AL357623	q33.3	No genes					
AL450263	q33.3	Novel	ENSG00000176889	121633429	121650133	78	Unknown
	q33.3	NM_032293	Q9BQH6	121673400	121802357	802	Unknown
AL445222	q33.3	SLC2A8	Q9NY64	121806004	121816716	477	Solute carrier family 2
	q33.3	**RPL12**	**P30050**	**128564940**	**121860226**	**164**	**60S ribosomal protein L12**
	q33.3	FLJ31641	Q96MZ7	121860323	121912317	696	Unknown
	q33.3	Novel	ENSG00000176217	121901370	121904622	109	Unknown
	q33.3	NBL_HUMAN	Q96TA1	121914157	121987798	733	Niban-like protein.
AL390116	q33.3	No genes					
AL162426	q34.11	STXBP1	Q64320	122021098	122101525	594	Syntaxin binding protein 1
	q34.11	Novel	ENSG00000160401	122108029	122124506	785	Unknown
	q34.11	FLJ00176	Q8TEL7	122124875	122140409	867	Unknown
	q34.11	TOR2A	Q96LSL7	122140333	122144087	253	Torsin family 2, member A
	q34.11	SH2D3C	Q9Y2X5	122147126	122187504	703	SH2 Domain containing protein 3
AL162586	q34.11	CDK9	P50750	122194861	122198896	372	Cell division protein kinase 9
	q34.11	Novel	ENSG00000177953	122199888	122201861	85	Unknown
	q34.11	FPGS	Q05932	122211733	122222873	587	Folypolyglutamate synthase
	q34.11	FLJ33157	Q96LW6	122225388	122225972	195	Unknown
	q34.11	ENG	P17813	122224494	122263514	658	Endoglin procursor
AL157935	q34.11	AK1	P00568	122275199	122286472	194	Adenylate kinase isoenzyme 1
	q34.11	FLJ13838	Q9H8A2	122294130	122308407	352	Beta-N-Acetylgalactosaminide
	q34.11	SIAT7D	Q9H4F1	122316695	122325831	298	Sialyltransferase
	q34.11	Novel	ENSG00000167103	122330677	122339810	471	Kinase
	q34.11	Novel	ENSG00000136908	122343908	122347297	162	Unknown
	q34.11	FLJ00179	Q8TEL4	122349392	122356974	194	Unknown
	q34.11	NM_018033	Q9NW83	122374689	122375144	152	Unknown
AL360268	q34.11	Q8WU12	Q8WU12	122472894	122475975	172	Unknown
AL590708	q34.11	KIAA1896	Q96PZ1	122500273	122518054	568	Mitochondrial solute carrier
	q34.11	PTGES2	Q9H7Z7	122529502	122537271	379	Prostaglandin E synthase 2
	q34.11	Q9N1Y9	Q9N1Y9	122537454	122538008	185	Unknown
	q34.11	LCN2	P80188	122558275	122562260	192	Lipocalin
	q34.11	C9orf16	Q9BUW7	122569160	122572735	83	Unknown
	q34.11	CIZ1	Q9ULV3	122574874	122613197	967	Zinc finger protein
	q34.11	DNM1	Q05193	122612217	122664055	864	Dynamin-1
	q34.11	GOLGA2	Q08379	122665504	122684750	1008	Golgin-95
AL590722	q34.11	Q8N2W6	Q8N2W6	122685026	122697798	209	Unknown
	q34.11	FLJ21673	Q9H6Y8	122709723	122710145	423	Unknown
AL359091	q34.11	FLJ11094	O95900	122717908	122731245	331	Unknown
	q34.11	C0Q4	Q9Y3A0	122731344	122742880	265	Coenzyme Q biosynthesis protein 4
	q34.11	SLC27A4	O95186	122749454	122770025	640	Fatty acid transport protein 4
	q34.11	NM_030914	Q9BTM9	122780169	122799542	101	Unknown
	q34.11	KIAA1502	Q9P226	122820574	122846159	560	Cerebral cell adhesion molecule
	q34.11	ODF2	O14721	122864961	122909768	638	Outer dense fibre of sperm tails 2
AL445287	q34.11	GLE1L	O75458	122913513	122951096	698	Gle-1 like RNA export mediator
AL356481	q34.11	SPTAN1	Q13813	122961411	123042401	2474	Spectrin alpha chain
	q34.11	NM_052844	Q9BV46	123042469	123065595	522	Unknown
	q34.11	SET	Q01105	123092703	123105196	290	SET (HLA-DR associated protein II)
AL359678	q34.11	No genes					
AL441992	q34.11	PKNbeta	O13355	123111331	123129414	889	Protein kinase
	q34.11	ZDHHC12	O32799	123129677	123132930	267	Zinc finger protein
	q34.11	ZYG	O00156	123138594	123180701	766	ZYG homologue
	q34.11	FLJ10743	Q9NVG8	123196140	123219240	275	Unknown

	q34.11	ENDOG	Q14249	123227276	123231484	297	Endonuclease G
	q34.11	HSPC109	Q9P041	123228459	123238629	384	Unknown
	q34.11	CCBL1	Q16773	123241750	123290846	422	Cyoplasmic cysteine
AL672142	q34.11	KIAA1437	Q9P2B1	123294854	12336844	811	Unknown
	q34.11	Q96GM4	Q96GM4	123329703	123351210	206	Unknown
	q34.11	KIAA1094	Q9UPQ8	123354339	123356427	538	Unknown
AL592211	q34.11	KIAA0169	Q14675	123357991	123415903	1739	Unknown
	q34.11	SH3GLB2	Q9BRZ5	123416600	123437108	130	SH3-containing protein
	q34.11	FLJ00199	Q9TEJ6	123445452	123480880	383	Unknown
	q34.11	Q96GF8	Q96GF8	123489937	123499243	237	Unknown
AL158151	q34.11	CRAT	P43155	123503602	123519612	626	Cartinine o-acetyltransferase
	q34.11	PPP2R4	Q15257	123519773	123557754	358	Protein phosphatase 2A
	q34.11	Novel	ENSG00000167133	123585649	123586860	290	Unknown
AL161785	q34.11	FLJ35269	Q8NAJ2	123729824	123733713	232	Unknown
AL353803	q34.11	FLJ34873	Q8NAS2	123745332	123747103	144	Unknown
AL391056	q34.11	Novel	ENSG00000179068	123897948	123913240	98	Unknown
	q34.11	FLJ35803	Q8NA65	124021035	124029584	377	Unknown
AL590369	q34.11	AD003	Q9UI28	124034981	124044744	223	Adrenal gland protein
	q34.11	ASB6	Q9NWX5	124043412	124050973	421	Ankyrin repeat containing protein
	q34.11	PMX2	Q99811	12074444	124131482	253	Paired mesoderm protein
	q34.11	PTGES	O14684	124147139	124161855	152	Prostaglandin E synthase
AL592219	q34.11	No genes					
AL158207	q34.11	TOR1B	O14657	124211961	124220092	336	Torsin B precursor
	q34.11	DYT1	Q96CA0	124221751	124232942	336	Torsin A precursor
	q34.11	HSPC220	Q9NZ63	124236100	124244083	289	Unknown
	q34.11	USP20	Q9Y2K6	124244254	12490636	914	Ubiquitin carboxyl-terminal hydrolase
	q34.11	FNBP1	Q96RU3	124295995	124451976	672	Thyroid receptor interacting protein
AL136141	q34.11	GPR107	Q96T26	124462729	124548972	416	G Protein-coupled receptor
AL392105	q34.11	No genes					
AL360004	q34.11	FREQ	P36610	124581381	124645435	190	Neuronal calcium sensor 1
	q34.11	Novel	ENSG00000178890	124674690	124718869	822	Unknown
50 kb Gap							
AL354898	q34.11	Q8NDA2	Q8NDA2	124808356	124841498	1187	Unknown
	q34.11	FLJ23816	Q8TCI8	124852198	124856039	220	Unknown
	q34.11	ASS	P00966	124866845	124923190	412	Argininosuccinate synthase
AL353695	q34.11	No genes					
AL359092	q34.11	FUBP3	Q92946	125001544	125060268	542	Fuse binding protein 3
	q34.12	PRDM12	Q9H4Q4	125086510	125104913	367	PR domain containing protein 12
	q34.12	RRP4	Q13868	125115687	125126785	293	Exosome complex exonuclease RRP4
AL161733	q34.12	ABL1	P00519	125136236	125309589	1130	Abelson murine leukaemia viral oncogene
	q34.12	FLJ14810	Q96SJ7	125324358	125360767	198	Unknown
AL583807	q34.12	LAMC3	Q9Y6N6	125431028	125516389	1575	Laminin gamma-3 chain precursor
AL355872	q34.12	No genes					
AL157938	**q34.12**	**AIF1L**	**Q9BQI0**	**125518441**	**125545061**	**150**	**Ionised cacium binding adaptor molecule 2**
	q34.13	NUP214	P35658	125547506	125656586	2140	Nuclear pore complex protein
	q34.13	Q8N2W3	Q8N2W3	125679994	125698463	191	Unknown
AL354855	q34.13	FLJ90726	Q8NBV4	125711653	12731177	271	Unknown
	q34.13	Novel	ENSG00000130710	125729227	125729298	24	Unknown
AL358781	**q34.13**	**BAT2L**	**Q9BU62**	**125852061**	**125869120**	**325**	**HLA-B associated transcript**
	q34.13	LQFBS-1	O95209	125921328	125922066	245	Unknown
	q34.13	POMT1	Q9UNT2	125924841	125945722	747	Protein-o-mannosyltransferase 1
	q34.13	UCK1_HUMAN	Q9HA47	125945717	125953181	201	Uridine cytidine kinase 1
AL160276	q34.13	GRF2	Q13905	126000707	126159454	1077	Guanine nucleotide releasing factor 2
AL160271	q34.13	CRSP8	O95401	126282028	126512112	273	Cofactor required transcriptional activation
AL603649	q34.13	No genes					

AL713892	q34.13	No genes					
AL691506	q34.13	No genes					
AL513102	q34.13	No genes					
AL353631	q34.13	No genes					
AL159997	q34.13	KIAA1857	Q96JH0	126594193	126675069	541	Netrin G2
	q34.13	KIAA0625	Q8WX33	126693686	126761056	915	Unknown
AL353701	q34.13	TTF1	Q15361	126808230	126835074	882	Transcription termination factor
AL354735	q34.13	Novel	ENSG00000178595	126930940	126975292	179	Unknown
	q34.13	BARHL1	Q9BZE3	127014851	127022519	327	BARH (Drosophila)-like 1
AL160165	q34.13	DDX31	Q96NY2	127026534	127102646	851	DEAD/H Box Helicase
	q34.13	GTF3C4	Q9UKN8	127102586	127122695	822	General transcription factor
AL445645	q34.13	FLJ32704	Q96MA6	127157823	127310564	479	Unknown
	q34.13	C9orf9	Q96E40	127310608	127322275	222	Unknown
	q34.13	TSC1	Q92574	127323595	127376866	1164	Tuberous sclerosis 1 gene
	q34.13	Novel	ENSG00000176140	127379314	127383801	47	Unknown
AL593851	q34.13	GFI1B	O95270	127418923	127426295	330	Growth factor independent 1B
AL162417	q34.2	GTF3C5	Q9H4P2	127462958	12790748	528	General transcription factor
	q34.2	CEL	P19835	127494229	127504006	756	Carboxyl ester lipase
	q34.2	NM_173692		127513557	127514144	196	Unknown
	q34.2	CELL	Q14018	127514780	127519600	59	Carboxyl ester lipase-like
	q34.2	**RALGDS**	**Q12967**	**127529965**	**127553410**	**914**	**Ral Guanine nucleotide**
	q34.2	FRS	Q9UKI5	127585198	127596144	347	Forssman synthetase
AL732364	q34.2	OBPIIB	Q9NPH6	127637537	127641486	170	Odorant binding protein 2B
AL158826	q34.2	ABO	P16442	127687850	127694413	287	ABO blood group system
	q34.2	SURF6	O75683	127754393	127759885	361	SURFEIT locus protein 6
	q34.2	SURF5	Q15528	127764596	127771813	200	SURFEIT locus protein 5
	q34.2	SURF3	P11518	127771906	127775122	265	SURFEIT locus protein 3
	q34.2	Q9H3B2	Q9H3B2	127774377	127775089	101	Unknown
	q34.2	SURF1	Q15526	127775504	127780202	300	SURFEIT locus protein 1
	q34.2	SURF2	Q15527	127780269	127784875	256	SURFEIT locus protein 2
	q34.2	SURF4	O15260	127785181	127799817	269	SURFEIT locus protein 4
	q34.2	Q8NE28	Q8NE28	127800125	127828061	651	Unknown
	q34.2	Novel	ENSG00000175977	127821374	127824619	97	Unknown
	q34.2	XPMC2H	Q9GZR2	127828027	127840010	422	Prevents mitotic catastrophe 2
	q34.2	ADAMTS13	Q96L37	127843961	127881349	1427	Von Willebrand factor-cleaving protease
AL593848	q34.2	C9orf7	Q9UGQ2	127881962	127892726	172	Unknown
	q34.2	SLC2A6	Q8NCC2	127893058	127901068	515	Solute carrier family 2
BX324209	q34.2	No genes					
AC002321	q34.2	No genes					
<5 kb Gap							
AC002101	q34.2	No genes					
AL365494	q34.2	DBH	P09172	128007020	128030001	603	Dopamine beta-monooxygenase precursor
	q34.2	SARDH	Q9UL10	128056341	128124046	832	Sarcosine dehydrogenase
	q34.2	PP3781	Q8WY83	123131956	128132329	124	Unknown
AL590710	q34.2	Novel	ENSG00000176983	128211430	128251481	432	Unknown
	q34.2	SARDH	Q9UL10	128252348	128272381	396	Unknown
AL357934	q34.2	VAV2	P52735	128297977	12826304	878	Oncogene VAV-2 protein
AL445931	q34.2	Novel	ENSG00000179483	128559910	128562122	119	Unknown
	q34.2	**BRD3**	**Q15059**	**128566862**	**128602533**	**726**	**Bromodomain containing protein3**
	q34.2	Novel	ENSG00000179457	128588311	128592826	216	Unknown
AL591386	q34.2	No genes					
200kb Gap							
AL354796	q34.2	No genes					
AL683798	q34.2	No genes					
13 kb Gap							

AL669970	**q34.2**	**RXRA**	**P19793**	**129062693**	**129101647**	**453**	**Retinoid X receptor, alpha**
AL591890	**q34.3**	**COL5A1**	**Q96HC0**	**129302868**	**129503955**	**590**	**Collagen alpha 1 (V) chain precursor**
AL603650	q34.3	FCN2	Q15485	129541874	129548582	313	Ficolin 2 precursor
AL353611	q34.3	FCN1	O00602	129570647	129579025	326	Ficolin 1 precursor
AL159992	q34.3	No genes					
AL390778	q34.3	OLFM1	Q9BWJ9	129736484	129782241	467	Olfactomedin related ER localised protein
AL353615	q34.3	NM_173520	Q8N4C0	130006614	130009927	152	Unknown
AL161452	q34.3	Novel	ENSG00000178197	130131516	130144083	96	Unknown
	q34.3	NM_014811	Q9Y4D3	130146060	130152258	1209	Unknown
	q34.3	NM_144654	Q8WU44	130158546	130165104	92	Unknown
	q34.3	MRPS2	Q9Y399	130164060	130168038	296	Mitochondrial ribosomal protein S2
	q34.3	LCN1	P31025	130184820	130189897	176	Lipocalin 1
	q34.3	OBPIIA	Q9NY56	130209504	130213321	170	Odorant-binding protein 2A
AL354761	q34.3	PAEP	P09466	130225123	130230141	157	Progestagen associated endometrial
	q34.3	Novel	ENSG00000176541	130238316	130250477	104	Unknown
AL158822	q34.3	MUPL	Q8WX39	130326687	130329116	172	Putative MUP-like lipocalin
	q34.3	Q8NEE3	Q8NEE3	130356772	130362893	348	Unknown
	q34.3	KCNT1	Q9WX41	130365551	130455523	1151	Unknown
AL353636	q34.3	NM_018627	Q9WX42	130472946	130570503	1298	Unknown
AL355574	q34.3	GPDR1	Q9BSL1	130596334	130624745	405	Glialblastoma related protein
	q34.3	NM_144653	Q96BF6	130674721	130713948	587	Unknown
AL591038	q34.3	Novel	ENSG00000180858	130715097	130738981	273	Unknown
AL138781	q34.3	Q96GU2	Q96GU2	130777953	130782228	29	Unknown
	q34.3	Q8N3G2	Q8N3G2	130869698	130887479	541	Unknown
	q34.3	LHX3	Q9UBR4	130859621	130868480	397	LIM Homeobox gene 3
30 kb Gap							
AL603784	q34.3	AGS3	Q9UFS8	130939617	130943166	530	Unknown
AL592301	q34.3	CARD9	Q9H257	130947895	130957602	536	Caspase recruitment protein
	q34.3	SNAPC4	Q9Y6P7	130959516	130982736	1469	Small nuclear RNA activating complex
	q34.3	SDCCAG3	O60525	130985862	130994412	192	Serologically defined colon cancer antigen
	q34.3	INPP5E	Q10713	130994603	131007700	525	Mitochondrial processing peptidase subunit
	q34.3	PPI5PIV	Q9NRR6	131012558	131023761	644	Phosphatidylinositol (4,5) bisphosphate 5-phophatase
	q34.3	KIAA0310	Q96HP1	131024036	131059908	1433	Unknown
	q34.3	NM_152571	Q8N9P6	131067482	131070005	203	Unknown
	q34.3	**NOTCH1**	**P46531**	**131078383**	**131129726**	**2559**	**Neurogenic locus NOTCH homologue protein**
AL590226	q34.3	Novel	ENSG00000180360	131211094	131230721	251	Unknown
	q34.3	Q9P058	Q9P058	131232664	131244366	146	Unknown
	q34.3	ZNEU1	Q9UHF1	131242795	131256617	273	ZNEU1/NEU1 protein
	q34.3	**AGPAT2**	**O15120**	**131257082**	**131271362**	**278**	**Acylglycerol-phophate-acyltransferase 2**
	q34.3	NM_152421	Q8WYU5	131296511	131307989	431	Unknown
AL355987	q34.3	NM_032887	Q96IC0	131309173	131312123	37	Unknown
	q34.3	Novel	ENSG00000169672	131313093	131332422	726	Unknown
	q34.3	FLJ33328	Q8NBE9	131338327	131341477	333	Unknown, has IG_MHC domain
	q34.3	FLJ10101	Q96BU21	131383198	131425126	307	Unknown
	q34.3	FLJ30985	Q96NE7	131387866	131392787	197	Unknown
	q34.3	Nov-01	ENSG00000054148	131433033	131434977	186	Unknown
	q34.3	Novel	ENSG00000148406	131436356	131438789	350	Unknown
	q34.3	Novel	ENSG00000179285	131439193	131440765	264	Unknown
	q34.3	Q8NCX7	Q8NCX7	131440944	131444739	236	Unknown
	q34.3	EDF1	O60869	131446058	131450225	148	Endothelial differentiation-related factor 1
AL449425	q34.3	TRAF2	TRA2_HUMAN	131482651	131510546	501	TNF receptor associated factor 2
AL807752	q34.3	NM_018998	Q969U6	131524374	131528545	566	Unknown
	q34.3	C8G	P07360	131529200	131530906	198	Complement component 8, gamma subunit
	q34.3	PTGDS	P41222	131561509	131565680	190	Prostaglandin D2 synthase
	q34.3	Novel	ENSG00000176785	131567941	131570349	137	Unknown

	q34.3	**CLIC3**	**O95833**	**131578574**	**131580807**	**207**	**Chloride intracellular channel protein 3**
	q34.3	**ABCA2**	**Q9BZC7**	**131591173**	**131612250**	**2440**	**ATP-binding cassette, subfamily A, member 2**
	q34.3	Q9BUH6	Q9BUH6	131576357	131577915	212	Unknown
	q34.3	FUT7	Q11130	131614669	131615685	339	Fucosyltransferase 7
AL929554	q34.3	Q8N224	Q8N224	131618602	131619006	135	Unknown
	q34.3	NPDC1	Q9NQX5	131623413	131630157	325	Neuronal proliferation protein 1
	q34.3	ENTPD2	Q9Y5L3	131632616	131638290	494	Ectonucleoside triphosphate diphosphydrolase 2
	q34.3	Q8TEI1	Q8TEI1	131662043	131670905	270	Unknown
	q34.3	Q8WUC7	Q8WUC7	131669250	131671114	104	Unknown
	q34.3	MAN1B1	Q9UKM7	131671245	131695076	699	Endoplasmic reticulum mannosidase
	q34.3	Novel	ENSG00000179395	131687948	131691411	989	Unknown
	q34.3	**DPP7**	**Q9UHL4**	**131696546**	**131699393**	**331**	**Dipeptidyl-peptidase**
	q34.3	GRIN1	Q05586	131724505	131753424	928	Glutamate receptor subunit zeta 1
	q34.3	NM_013366	Q9UJX6	131759296	131773049	822	Anaphase-promoting complex subunit 2
	q34.3	SSNA1	O43805	131773159	131774882	119	Sjorgen's syndrome nuclear autoantigen 1
	q34.3	FLJ90254	Q8NCH2	131776627	131784420	433	Unknown
	q34.3	NM_053045	Q969S6	131788473	131790029	136	Unknown
BX255925	q34.3	Unfinished					
BX322799	q34.3	Unfinished					
AL365502	q34.3	AD038	Q96F01	132035782	132039264	205	AD038 protein, function unknown
	q34.3	NM_152285	Q8N5I2	132039236	132048941	433	Unknown
	q34.3	MZIP	Q96E35	132064147	132072376	227	Melanin-concentrating hormone receptor 1
	q34.3	NM_138778	Q9BTV6	132075663	132096389	484	Unknown
	q34.3	MRPL41	NM_032477	132098739	132099406	137	Mitochondrial ribosomal protein L41
	q34.3	NTE-L	Q8TAY5	132153193	132190541	702	Neuropathy Target Esterase
	q34.3	FLJ14568		132192950	132200987	327	Unknown
	q34.3	Q9NTU2		13295160	132200984	130	Unknown
AL590627	q34.3	Novel	ENSG00000181090	132289873	132374235	436	Unknown
AL611925	**q34.3**	**HMT1**	**Q9H9B1**	**132395728**	**132553855**	**1247**	**Histone methyltransferase**
AL772363	q34.3	CACNA1B	Q00975	132562084	132806329	2357	Calcium channel voltage-dependent
AL591424	q34.3	IL9R	ENSG00000165830	132821281	132832099	216	Interleukin 9 receptor IL 9R
	q34.3	**Novel**	**ENSG00000159247**	**132859209**	**132861370**	**425**	**Tubulin pseudogene**
	q34.3	Novel	ENSG00000179338	132868297	132868767	157	LINE 1 Reverse Transcriptase Homologue
AL954642	q34.3	No genes					

Appendix 2

Table of results the whole-genome survey. The P-values are coloured according to the level of confidence; black are L0-paralogues, green L1-paralogues, blue L2-paralogues and red are L3-paralogues.

Class	*MHC gene*	*Clone*	*Locus*	*Start*	*End*	*BLAST Match*	*P-value*
III	NOTCH4	AL390719.31.31331.88824	1p36.33	708136	744003	AGRN	1.50E-23
xII	B3GALT4	AL162741.35.1.111409	1p26.33	919626	921479	B3GALT6	1.40E-08
III	NOTCH4	AL391244.11.1.67923	1p36.33	1073552	1087162	NM_030937	9.10E-09
III	BAT8	AL391244.11.1.67923	1p36.33	1089751	1108843	Novel	7.60E-12
III	NOTCH4	AL512413.21.1.101803	1p36.32	2844459	2886399	EGFL3	2.20E-42
xI	BTN1A1	AL662907.11.1.64693	1p35.1	3343046	33466609	Q9BVG3	1.10E-18
III	NOTCH4	AL513320.27.1.132592	1p36.32	3364015	3405344	EGFL3	2.00E-25
III	C2	AL109811.40.1.112769	1p36.22	10930224	10950939	MASP2	1.30E-05
xI	BTN1A1	AC074003.3.23107.35032	1p36.13	14652770	14656331	Novel	4.20E-14
III	C6orf46/ZNF297	AL034555.2.1.86897	1p36.13	15447174	15481304	ZNF151	2.40E-14
I	DDR1	AL451042.10.1.88098	1p36.13	15564788	15596501	EPHA2	1.10E-17
xII	KIFC1	AL663074.13.1.8581	1p36.12	20071643	20079710	Novel	2.60E-24
III	NOTCH4	AL590103.12.1.175162	1p36.12	21183949	21258020	HSPG2	1.10E-16
I	DDR1	AL035703.21.1.160705	1p36.12	21925207	21965283	EPHA8	1.40E-15
I	DDR1	AL035704.9.1.113956	1p36.12	22143606	22277501	EPHB2	3.10E-20
xII	LYPLA2L	AL031295.1.1.124001	1p36.11	23188619	23193014	LYPLA2	9.00E-75
III	CLIC1	AL662924.15.1.121762	1p35.3	24141534	24239539	CLIC4	3.80E-27
I	DDR1	AL031729.16.1.125287	1p36.11	26912974	26924738	FGR	3.20E-15
xII	COL11A2	AC114488.1.90406.184673	1p35.2	31102432	31154233	COL16A1	2.10E-11
xII	ZNF297	AL033529.25.1.147167	1p35.1	31870816	32006551	NM_144621	1.00E-13
xI	RFP	AL662907.11.1.64693	1p35.1	32565928	32602196	NM_018207	7.60E-46
III	C2	AC115285.1.63883.124348	1p35.1	33035623	33094251	Q96Q03	9.80E-08
III	BF	AC115285.1.63883.124348	1p35.1	33799053	33862505	Q9H4W4	9.50E-08
I	POU5F1	AL139158.11.1.115614	1p34.3	37518563	37519168	no gene	4.50E-44
III	HSPA1L	AL354702.7.1.107422	1p34.3	38182110	38184069	Novel	4.30E-158
xII	ZNF297	AL356379.10.1.64960	1p34.2	40000812	40017418	NM_152373	7.80E-09
xI	RFP/MOG/BTNL2	AL512353.16.1.81704	1p34.2	42286363	42314156	ERMAP	6.60E-49
I	DDR1	AC093420.1.127596.194462	1p34.2	42767145	42789215	TIE	4.80E-05
II	BTNL2	AL109659.20.1.181678	1p33	47521181	181364180	genscan	1.10E-24
xI	BTN1A1	AL109659.20.1.181678	1p33	48243776	48288207	no gene	3.40E-29
xI	MOG	AL109659.20.1.181678	1p33	48253780	48253842	genscan	4.00E-22
xI	GPX5	AL356976.30.1.64323	1p32.3	51957064	51963742	NM_015696	8.30E-09
I	TUBB	AL445183.19.1.193774	1p32.3	53049002	53152605	SCP2	8.90E-12
I	DDR1	AL445205.14.1.115936	1p31.3	63525514	63564266	EST gene	3.10E-45
I	DDR1	AC093427.2.1.131877	1p31.3	64219769	64351748	JAK1	2.50E-09
III	C6orf29	AC107627.2.1.90513	1p31.1	74605198	75011843	NM_152697	8.70E-49
III	MSH5	AL445464.9.1.103097	1p31.1	75197713	75313965	MSH4	4.30E-05
III	DDAH2	AL078459.8.1.83946	1p22.3	84926028	85072691	DDAH1	3.50E-15
xII	COL11A2	AL356059.27.1.76418	1p22.3	85351949	85791154	NM_152890	1.90E-05
II	BRD2	AC004798.1.1.42497	1p22.1	91625907	91677611	BRDT	6.50E-101
III	C9orf29	AC093429.2.1.182165	1p21.3	94491811	94566684	NM_152369	1.30E-08

xII	COL11A2	AC093150.2.1.189945	1p21.1	104172486	104410718	COL11A1	4.10E-42
III	NOTCH4	AL390252.9.1.169241	1p13.3	108785784	108869934	SORT1	6.40E-22
III	BAT1	AL445483.13.1.164008	1p13.2	112714422	112726367	DDX20	3.00E-14
xI	MOG/BTN1A1	AL391476.20.1.171595	1p13.1	118145476	118212809	NM_024626	9.30E-05
III	NOTCH4	AL359752.11.1.137955	1p11.2	119292037	119450143	NOTCH2	2.30E-61
III	NOTCH4	AL592307.24.14836.157830	1q21.1	141809266	141890630	Novel	3.20E-27
III	NOTCH4	AC018381.3.23653.71017	1q21.1	141956297	141968135	Novel	3.70E-28
xI	HIST1H2AC	AL591493.13.1.113370	1q21.2	145553544	145555199	Histones	3.30E-48
III	BAT8	AL590133.32.1.192096	1q21.3	146646159	146684482	SETDB1	1.90E-12
xII	ZNF297	AL451085.20.1.182166	1q22	150706074	150720422	ZFP67	5.40E-14
I	POU5F1	AL139410.20.1.166288	1q22	151143602	151144684	Q9BZW0	4.10E-105
III	NOTCH4	AL158169.17.1.99802	1q23.1	152575891	152593889	INSRR	6.10E-12
I	DDR1	AL158169.17.1.99802	1q23.1	152595949	152616657	NTRK1	1.70E-30
xI	HFE	AL138899.23.1.134137	1q23.1	153882029	153886978	CD1D	2.50E-05
III	NOTCH4	AL356104.6.1.96693	1q23.1	154745243	154755113	Q8TEK2	4.20E-22
III	HSPA1L	AL590385.22.1.110781	1q23.3	157226591	157228843	HSPA6	6.20E-263
III	CREBL1	AL391825.15.1.211662	1q23.3	157386942	157579736	ATF6	4.30E-25
I	DDR1	AL445197.4.1.117040	1q23.3	158253102	158401084	DDR2	1.50E-131
III	PBX2	AL357568.14.1.71359	1q23.3	160179905	160302546	PBX1	3.40E-88
I	POU5F1	AL136984.20.1.169627	1q24.2	162840972	163036179	POU2F1	5.70E-19
xII	RXRB	AL160058.8.1.155369	1q23.3	163005239	163049202	RXRG	3.00E-63
xII	RPS18	AL031733.3.1.215861	1q24.2	163211499	163211834	no gene	7.20E-14
III	BAT2	AL021579.1.1.99886	1q24.3	167083736	167191695	BAT2-ISO	1.60E-63
III	TNF	Z96050.1.1.85811	1q24.3	168257203	168265061	TNFSF6	5.90E-06
III	TNXB	Z94055.1.1.134539	1q25.1	170720587	170804587	TNR	6.00E-07
III	C6orf46	AL136170.12.1.127541	1q25.1	171377563	171488882	NM_032522	2.10E-33
I	DDR1	AL139132.16.1.157866	1q25.2	174505391	174627374	ABL2	7.90E-12
I	HLA-A/HLA-E	AL162431.17.1.139006	1q25.3	176341488	176402911	STX6	1.60E-20
xI	HLA Class I and II	AL356267.27.1.181808	1q25.3	176431452	176452972	HLALS	3.50E-31
III	BAT1	AL049557.19.1.128379	1q25.2	176601093	176723077	ABL2	3.40E-10
III	BAT8	AL138776.10.1.100549	1q25.3	177972880	177984295	RNASEL	3.30E-05
xII	RAB2L	AL590422.14.1.198210	1q25.3	179033557	179326003	RGL1	8.30E-17
xII	RING1	AL109865.36.1.201823	1q25.3	180442880	180499976	RNF2	4.20E-50
I	DHX16	AL355999.9.1.76504	1q31.1	185516454	185517347	genscan	1.10E-07
xII	B3GALT4	AL390863.9.1.122864	1q31.2	188662474	188670039	B3GALT2	2.20E-21
III	NOTCH4	AL513325.13.1.212888	1q31.3	192695849	192906025	CRB1	4.20E-52
III	ATP6V1G2	AL157402.19.1.210331	1q31.3	193950307	193968515	ATP6V1G3	2.40E-09
xII	RXRB	AC096633.2.1.178152	1q32.1	195455200	195604973	NR5A2	1.50E-15
xII	KIFC1	AL445483.13.1.164008	1q32.1	195979729	196048406	KIF14	7.50E-10
III	BAT1	AL512326.24.1.189269	1q32.1	198323462	198369154	NM_031306	4.00E-06
xII	RPS18	AL606462.5.1.112401	1q42.13	223356887	223357135	genscan	1.20E-40
xI	RFP	AL139288.15.1.151563	1q42.13	224319970	224333114	TRIM11	2.10E-64
xI	HIST1H2AC	AL139288.15.1.151563	1q42.13	224383273	224384153	H2AFL	4.80E-49
xI	BTN1A1	AL139288.15.1.151563	1q42.13	224436665	224436976	genscan	4.50E-64
II	BTNL2	AL139288.15.1.151563	1q42.13	224436716	224436991	genscan	3.70E-19
II	TAP2/1	AL121990.33.1.147913	1q42.13	225346509	225388622	ABCB10	5.10E-24
III	PBX2	AL359255.12.1.20809	1q42.13	226893037	227045628	OBSCN	5.40E-11
xI	BTN1A1	AC026657.4.97959.109520	1q42.13	227053394	227055850	TRIM11	1.20E-16
xI	MOG	AL139288.15.1.151563	1q42.13	227195475	227197055	Novel	9.50E-16
xI	RFP	AL591686.9.1.150680	1q43	237498626	237934567	NM_152666	2.40E-11
III	C6orf46/ZNF297	AL590483.25.118180.187060	1q44	239460483	239466676	ZNF238	9.20E-16
III	HSPA1L	AL390728.34.1.206255	1q44	242630342	242630950	no gene	5.80E-218
xI	RFP	AC099571.1.86529.165648	1q44	243256848	243277448	NM_015431	1.50E-65
xI	BTN1A1	AC099571.1.86529.165648	1q44	245957656	245974557	Q9Y4N9	3.50E-37
III	NOTCH4	AC105450.1.1.163782	2p25.3	1491150	1620306	TP0	2.10E-07
xII	KIFC1	AC013449.8.1.120997	2q23.3	26242053	26297524	KIF3C	3.30E-37

III	BAT1	AL121658.4.1.162692	2p22.3	32334563	32390404	Q96NC3	6.30E-05
xI	HIST1H2AC/NOTCH4	AL133244.1.1.200368	2p22.3	33311500	33772726	LTBP1	6.50E-18
III	CYP21A2	AC009229.5.1.209156	2p22.2	38259590	38268136	CYP1B1	9.30E-15
I	DHX16	AC092833.4.1.143506	2p22.1	39171874	39237365	NM_145646	5.30E-10
I	DHX16	AC018693.8.1.164125	2p22.1	39282730	39282825	no gene	1.00E-09
xI	NEFAL	AC016722.9.1.149995	2p21	47297756	47311689	NM_139279	3.80E-32
III	MSH5	AC009600.19.1.215260	2p21	47798999	47879105	MSH2	5.00E-09
xI	UBD	AC079807.5.1.156175	2p16.3	48126446	48126718	genscan	5.80E-05
III	MSH5	AC006509.15.1.124015	2p16.3	48179027	48202837	MSH6	7.80E-08
xII	B3GALT4	AC093401.4.1.99088	2p15	62609435	62638037	B3GNT1	2.30E-13
I	DHX16	AC005041.2.1.191356	2p13.1	74958046	74966106	NM_133637	1.30E-10
III	BAT2	AC068279.6.1.135351	2p11.2	87959386	87959877	no gene	3.10E-05
III	BAT2	AC026106.12.36729.68570	2p11.2	90750748	90751398	no gene	6.50E-05
xII	ZNF297	AC092835.4.1.158404	2q-tel	94264885	94283773	ZNF2	3.70E-10
xI	RFP	AC018892.8.1.191055	2q11.2	96145683	96151491	Novel	1.60E-21
I	DDR1	AC016699.10.1.54480	2q11.2	96793061	96808883	ZAP70	3.30E-15
I	POU5F1	AC018730.7.1.154728	2q12.1	103925142	103926146	POU3F3	2.60E-44
xI	BTN1A1	AC005040.2.1.189949	2q12.3	106059831	106060295	genscan	3.40E-08
xII	ZNF297	AC013268.5.1.206457	2q13	109008786	109017957	NM_152518	4.40E-13
I	C6ORF18	AC018737.9.1.206454	2q14.3	120015449	120327116	CLASP1	9.70E-06
xII	RPS18	AC018737.9.1.206454	2q14.3	120421203	120421337	no gene	9.80E-10
I	TUBB	AC018804.9.1.195514	2q21.1	128250367	128259530	genscan	3.00E-08
I	TUBB	AC073869.5.1.195280	2q21.2	129791442	129796126	TUBA2	1.80E-10
I	MRPS18B	AC012497.8.1.212104	2q22.1	138733844	138734416	genscan	1.20E-75
xII	RXRB	AC074099.6.1.143653	2q24.1	155709873	155718141	NR4A2	1.00E-06
xII	B3GALT4	AC016723.11.1.202001	2q24.3	167216857	167269041	B3GALT1	8.10E-30
II	TAP2/1	AC069137.6.1.108836	2q24.3	168321109	168292498	ABCB11	6.90E-21
xII	COL11A2	AC066694.7.1.120381	2q32.2	187975163	188013419	COL3A1	5.40E-32
III	HSPA1L	AC013409.8.1.195478	2q34	208662170	208666396	Novel	1.00E-116
I	TUBB	AC068946.4.1.172260	2q35	218133730	218152718	TUBA4	2.00E-10
I	DDR1	AC010899.8.1.210232	2q36.1	221011996	221158367	EPHA4	9.20E-15
III	HSPA1L	AC009302.2.1.180970	2q36.1	221548101	221549027	genscan	7.00E-113
xI	BTN1A1	AC104772.3.1.106526	2q36.1	221566190	222241888	SYFB	1.40E-19
xI	RFP	AC104772.3.1.106526	2q36.1	222156619	222241888	SYFB	4.40E-44
xI	PRSS16	AC008072.3.1.206177	2q36.1	223964489	223987825	NM_024785	4.20E-10
xII	COL11A2	AC073869.5.1.195280	2q36.3	226590998	226750349	COL4A4	3.50E-07
xII	COL11A2	AC097662.4.37779.206758	2q36.3	226750355	226900581	COL4A3	8.90E-12
III	NOTCH4	AC008273.2.1.151297	2q36.3	228954450	229310970	NM_139072	3.30E-23
xII	B3GALT4	AC017104.8.1.168880	2q37.1	230992101	230995454	B3GNT7	3.50E-15
III	NOTCH4	AC005237.2.1.175179	2q37.3	240214701	240251678	PASK	2.30E-53
xII	KIFC1	AC011298.6.31675.58437	2q37.3	240584194	240668241	ATSV	3.00E-06
III	BAT8	AC034191.5.1.172215	3p26.1	4284929	4298795	SETMAR	2.00E-25
xII	RXRB	AC090947.1.1.166043	3p25.2	12270465	12415723	PPARG	5.30E-09
III	NOTCH4	AC090509.1.1.165994	3p25.1	13551690	13619799	FBLN2	2.40E-09
xII	RXRB	AC090937.1.1.160696	3p25.1	15002136	15024392	NR2C2	6.60E-09
xI	HMGN4	AC027125.4.1.173836	3p25.1	15346179	15346391	no gene	1.90E-06
III	BAT8	AC090950.1.1.199282	3p25.1	15648627	15776696	Y379	3.60E-05
III	HSPA1L	AC097635.2.1.162887	3p24.3	19380503	19387078	Novel	2.90E-102
xII	ZNF297	AC006059.3.1.185161	3p22.1	41880829	41889031	NM_145166	1.10E-11
III	C6orf46	AC099669.2.1.217035	3p21.32	43776685	43804920	Novel	4.10E-13
III	C6orf46	AC124045.1.109944.135528	3p21.32	43934102	43945288	NM_033210	5.10E-12
I	DDR1	AC104439.2.1.197279	3p21.32	45363297	45363752	genscan	7.20E-14
I	DHX16	AC026318.7.1.19068	3p21.31	47135873	47174627	DDX30	5.00E-25
xII	COL11A2	AC005903.3.1.60660	3p21.31	47884518	47915700	COL7A1	5.00E-11
III	NOTCH4	AC005923.2.1.88326	3p21.31	47956918	47983375	CELSR3	3.60E-22
xI	GPX5	AC121247.1.77964.92674	3p21.31	48542852	48544273	GPX1	2.40E-34

III	NOTCH4	AC112215.1.181144.198956	3p21.31	51781935	51811071	STAB1	1.20E-16
I	ABCF1	AC021123.4.149752.161126	3q-tel	91147420	91149286	Novel	1.50E-63
I	DDR1	AC107028.4.1.185539	3q11.2	92010160	92148744	EPHA3	1.20E-15
xII	RPS18	AC108715.2.1.176462	3q11.2	94194400	94194576	no gene	2.40E-15
xII	RPS18	AC108695.2.1.190845	3q11.2	94194185	94194358	Novel	4.30E-24
xII	COL11A2	AC069222.23.1.117000	3q12.1	96016863	96137341	COL8A1	3.50E-23
I	POU5F1	AC117460.7.1.183595	3q12.1	97141764	97141988	EST gene	2.20E-18
I	TUBB	AC046144.15.1.188840	3q13.11	100051837	100051959	genscan	2.00E-11
xII	RPS18	AC073861.18.98940.165924	3q12.3	100725089	100725481	EST gene	4.20E-62
xII	ZNF297	AC084198.24.86116.155268	3q12.3	100797978	100825652	NM_014415	3.60E-12
I	POU5F1	AC079945.13.52386.74222	3q21.3	128590475	128590648	genscan	4.70E-55
III	NOTCH4	AC080007.26.1.168551	3q21.3	129795525	129875755	WDR10	8.50E-10
III	HSPA1L	AC020632.16.1.162029	3q22.1	133021872	133186190	NM_153240	1.90E-22
I	DDR1	AC092969.6.71736.203901	3q22.2	135625566	135694383	Novel	5.10E-21
III	HSPA1L	AC117478.3.1.77155	3q22.3	138390731	138391456	genscan	4.20E-215
III	PBX2	AC018450.26.1.191474	3q24	139874190	139875716	PBXP1	4.50E-161
III	C6orf46/ZNF297	AC010184.18.1.190580	3q23	141935447	141992349	Q8NAP3	6.00E-07
III	BAT1	AC112907.4.101183.186560	3q25.2	151048380	151140293	Novel	2.00E-24
xII	B3GALT4	AC021649.18.1.209521	3q25.33	161749105	161770594	B3GALT3	9.40E-19
III	BAT1	AC092946.7.1.115500	3q27.3	183276473	183282793	EIF4A2	8.70E-15
xII	B3GALT4	AC069417.16.47453.72677	3q27.1	183826771	183845940	B3GNT5	8.40E-28
I	ABCF1	AC048331.32.147427.232441	3q27.1	184898582	184906494	NM_018358	1.60E-37
I	GNL1	AC046143.20.1.180365	3q29	197722675	197769579	NM_018385	3.90E-07
III	NOTCH4	AC021118.6.1.194612	4p15.31	20338712	20704631	SLIT2	1.00E-25
I	DHX16	AC115110.2.23128.113396	4p15.2	24612486	24669565	DDX15	4.30E-58
III	C6orf46/ZNF297	AC105287.4.52544.192873	4p14	39946115	39947420	Novel	1.40E-163
II	HLA Class II	AC097451.2.1.146808	4p13	43676798	43677424	EST gene	1.10E-22
III	LSM2	AC108054.2.101121.147995	4p12	48652844	48737332	Q9P270	1.40E-20
I	DDR1	AC098587.1.9710.175365	4q12	55065587	55126000	PDGFRA	3.00E-05
xI	RFP	AC107058.4.1.126135	4q13.1	65884515	65884994	genscan	1.00E-24
III	BAT8	AC053527.8.1.233250	4q13.3	74230262	74364018	Q9H288	8.00E-05
xI	HIST1H2AC	AC097460.3.1.164370	4q23	101179583	101181853	H2AFZ	2.70E-15
III	BAT1	AC105460.4.1.185755	4q24	104893633	104894151	Genscan	4.00E-10
III	CYP21A2	AC096564.3.1.163317	4q25	109292353	109297744	Novel	2.10E-15
xII	LYPLA2L	AC004062.1.1.154252	4q25	112148688	112149080	no gene	7.40E-66
I	TUBB	AC093663.4.1.171745	4q25	113404522	113404635	genscan	1.20E-09
I	TUBB/RNF5	AC093816.3.1.170227	4q27	123297626	123297751	genscan	4.50E-36
III	BAT8/NOTCH4	AC105421.2.1.162793	4q28.1	125865162	125873003	YB23	2.50E-16
III	NOTCH4	AC092629.2.1.148673	4q28.1	126653380	126692458	NM_024582	2.20E-30
III	HSPA1L	AC093591.3.1.158758	4q28.1	128982992	129034039	OS94	2.00E-16
I	POU5F1	AC093887.3.1.192886	4q31.22	147935518	147939053	POU4F2	5.90E-28
II	TAP1	AC017037.10.1.186106	4q32.1	159166512	159166844	no gene	8.50E-05
III	HSPA1L	AC105250.3.1.70449	4q32.3	165537822	165539805	Novel	3.40E-147
I	IER3	AC106872.5.1.174535	4q32.3	166282549	166282689	no gene	1.10E-07
xI	RFP	AC106872.5.1.174535	4q32.3	166359709	166359709	Novel	8.10E-63
xI	BTN1A1/RFP	AC106872.5.1.174535	4q32.3	166381434	166391179	NM_152620	3.00E-31
xI	BTN1A1	AC108465.3.1.48677	4q32.3	166812464	166832879	Butyrophilin	5.20E-34
xI	RFP	AC080079.5.1.112516	4q32.3	167014840	167015103	genscan	7.60E-22
xII	HSD17B8	AC021151.8.1.175081	4q32.3	170330785	170353453	NM_032783	7.10E-06
III	NOTCH4	AC079226.7.1.184032	4q35.1	184330635	184402306	NM_018104	1.30E-11
III	NOTCH4	AC110761.3.1.153458	4q35.2	188188876	188324915	FAT	2.50E-13
xI	BTN1A1/RFP	AC108073.2.107221.165848	4q35.2	189692345	189706326	NM_173553	8.20E-19
xI	POM121L2	AC093308.2.104996.134015	5p14.3	21925895	21928486	Novel	1.10E-09
xI	SMA3L	AL157879.7.1.161460	5p13.3	34638648	34642503	Novel	2.00E-45
xI	SMA3L	AC114970.1.37696.100300	5p13.3	34737381	34778708	Novel	1.40E-46

xI	POM121L2	AC114970.1.37696.100300	5p13.3	34761244	34778854	Q9H1S5	1.30E-10
I	TUBB	AC106800.1.2657.73135	5p12	43538759	43538860	genscan	8.00E-05
III	BAT1	AC016632.6.1.176784	5q11.2	55195823	55274550	DDX4	6.10E-09
xI	GPX5	AC091977.3.1.183494	5q11.2	55884790	55889596	Q8TED1	2.50E-05
I	DHX16	AC020728.4.1.201404	5q11.2	55980727	56032245	NM_019030	9.30E-13
III	SKIV2L	AC020728.4.1.201404	5q11.2	56032666	56149658	KIAA0052	9.80E-50
xI	SMA3L	AC108108.1.77200.116688	5q13.2	70698028	70703834	Novel (SMA3)	2.00E-48
III	MSH5	AC022493.12.1.153078	5q14.1	80189323	80411166	MSH3	3.90E-15
xII	RPS18	AC008799.6.1.123098	5q14.3	90941311	90941703	genscan	4.70E-49
xII	RXRB	AC106818.1.6855.31297	5q15	93388022	933399296	NRF1	6.60E-19
I	POU5F1	AC108102.2.1.161056	5q15	93545276	93545938	genscan	4.20E-48
III	BAT1	AC016567.8.1.159734	5q15	97361820	97362176	Genscan	1.30E-11
xI	POM121L2	AC114324.1.39743.91827	5q21.1	99294853	99295455	no gene	6.90E-07
xI	SMA3L	AC114324.1.39743.91827	5q21.1	99303064	99303249	genscan	8.30E-27
xI	SMA3L	AC092278.3.1.123469	5q21.1	99829117	99834848	Novel (SMA3)	1.30E-42
III	LSM2/DDR1	AC109481.3.1.20673	5q21.3	108524961	108964792	FER	9.90E-36
I	DHX16	AC093208.2.1.112115	5q22.2	113296938	113337241	Novel	7.10E-07
III	BAT8	AC010226.5.1.147140	5q22.3	115264395	115288381	FEM1C	7.20E-06
xI	HMGN4	AC109456.3.1.121848	5q23.1	115322895	115323107	no gene	5.80E-05
III	NOTCH4	AC008682.6.1.217221	5q23.2	127062260	127232628	NM_032446	1.30E-18
III	NOTCH4	AC010424.9.1.192282	5q23.2	127289102	127326544	NM_130809	9.10E-10
III	NOTCH4	AC025169.5.1.161920	5q23.2	128029494	128309063	FBN2	4.50E-35
III	HSPA1L	AC005373.1.1.112220	5q23.3	129911964	129912584	genscan	6.20E-159
xII	KIFC1	AC004237.1.1.38715	5q23.3	132479552	132517394	KIF3A	4.20E-18
III	HSPA1L	AC113410.2.1.123851	5q31.1	132834893	132887389	HSPA4	3.00E-19
III	BAT1	AC010301.7.1.155067	5q31.1	133745073	133815582	NM_014829	1.40E-09
xI	HIST1H2AC	AC026691.5.1.135062	5q31.1	135237074	135302580	H2AFY	2.10E-21
III	HSPA1L	AC011385.6.1.134599	5q31.2	138566125	138586223	HSPA9B	2.50E-72
I	POU5F1	AC011396.4.1.87692	5q32	146300493	146301825	POU4F3	5.50E-28
III	ATP6V1G2	AC008385.7.1.151712	5q33.1	150765251	150795891	Novel	1.50E-06
xI	GPX5	AC008666.5.1.99108	5q33.1	151002039	151004829	GPX3	4.20E-64
xII	KIFC1	AC008410.5.53020.91332	5q33.2	154988937	154992638	Novel	1.80E-38
III	NOTCH4	AC011369.4.1.141529	5q34	167774142	168258463	Novel	8.20E-10
III	NOTCH4	AC011365.4.1.81930	5q34	168684275	168685047	Novel	7.40E-29
xII	ZNF297	AC104117.1.114225.124230	5q35.3	179277145	179302355	Q8N9F8	5.60E-10
xI	BTN1A1/MOG	AC016572.6.1.143687	5q35.3	180455914	180507663	Butyrophilin	3.30E-54
xI	BTN1A1/MOG	AC091874.2.1.13312	5q35.3	180545631	180563128	BTNL3	4.10E-32
xI	BTN1A1	AC091874.2.13413.147570	5q35.3	180597001	180613745	Butyrophilin	9.10E-53
xI	BTN1A1	AC008443.9.1.120524	5q35.3	180699946	180712464	Q8WV44	4.00E-31
xI	BTN1A1/MOG/BTNL2	AC022413.4.1.166525	5q35.3	181312424	181364180	NM_024850	5.60E-26
II	BTNL2	AC091874.2.13413.147570	5q35.3	181402141	181419638	Butyrophilin	7.20E-23
II	BTNL2	AC091874.2.1.13312	5q35.3	181453511	151470537	NM_152547	2.70E-27
xI	RFP	AC008443.9.1.120524	5q35.3	181733455	181735630	TRIM7	3.80E-58
I	TUBB	AL031963.40.1.149546	6p25.2	3138899	3142759	TUBBL	2.40E-208
I	TUBB	AL445309.13.1.136587	6p25.2	3209729	3212968	TUBBL	2.40E-208
xII	RPS18	AL359643.27.1.166863	6p25.1	4964305	4964649	genscan	7.50E-35
III	C6orf46	AL161903.19.1.47104	6p21.32	33384295	33421772	SYNGAP1	3.30E-17
xII	ZNF297	AL161903.19.1.47104	6p21.32	33407966	33421769	NM_152735	2.90E-50
xII	RXRB	AL022721.1.1.170245	6p21.31	35306800	35392369	PPARD	5.50E-09
xII	KIFC1	AL590387.7.1.76075	6p21.2	39560349	39603922	Novel	1.20E-12
I	TUBB	AL136089.15.1.99479	6p21.2	39962644	39962751	genscan	1.00E-10
I	DDR1	AL355385.15.1.129884	6p21.1	43040485	43125871	PTK7	2.20E-17
xI	BTN1A1	AL512353.16.1.81704	6p34.2	43256510	43284362	ERMAP	2.60E-68
II	TAP2/1	AL359813.23.1.102892	6p21.1	43391706	43414579	ABCB10	2.00E-15
III	NOTCH4	AL359813.23.1.102892	6p21.1	43414506	43420784	NM_023932	5.70E-28
III	CLIC1	AL357057.19.1.58133	6p21.1	45865819	46044480	CLIC5	8.00E-53

xI	SMA3L	AL021368.1.1.188642	6p11.2	58109978	58115773	Novel(SMA3)	7.30E-41
III	NOTCH4	AL137007.9.1.105779	6q12	65571522	65587322	Q9H557	2.30E-24
xII	COL11A2	AL080275.20.1.113983	6q13	70892928	70979474	COL9A1	8.90E-08
xII	LYPLA2L	AL365267.11.1.49616	6q13	71841661	71842056	no gene	7.60E-71
III	C4B	AL590428.7.1.163577	6q13	74364147	74493108	NM_133493	4.10E-08
xII	RPS18	AL355796.11.1.152086	6q14.1	79954508	79954717	no gene	1.60E-22
I	DDR1	AL354857.13.1.199223	6q16.1	93918505	94095966	EPHA7	1.20E-14
I	POU5F1	AL022395.2.1.126882	6q16.2	99299451	99300710	POU3F2	8.00E-43
xII	HSD17B8	AL591803.10.1.90325	6q16.2	99639262	99640282	Novel	1.90E-20
xII	RXRB	AL078596.8.1.64183	6q21	108510126	108532877	NR2E1	2.00E-21
xII	ZNF297	AL109947.19.1.128960	6q21	109806583	109827304	Y441	3.50E-12
I	DDR1	Z97989.1.1.155937	6q21	112005349	112217491	FYN	2.00E-11
I	DDR1	AL357141.8.1.125184	6q22.1	116285557	115404785	FRK	9.00E-20
xII	COL11A2	AL121963.10.1.107553	6q22.1	116462985	116470164	COL10A1	4.00E-23
I	DDR1	Z98880.1.1.108260	6q22.1	117632394	117769882	ROS1	2.60E-14
xII	RPS18	AL357084.12.1.76042	6q24.1	141090774	141091166	genscan	1.60E-48
I	TUBB	AL031320.6.1.133574	6q24.2	143380844	143380963	genscan	1.10E-51
xII	RPS18	AL078581.11.1.102019	6q25.1	149763293	149791539	KATNA1	8.40E-51
III	HSPA1L	AL590413.18.1.104939	6q25.1	151621254	151668524	NM_017909	1.70E-16
xI	MAS1L	AL035691.17.1.129968	6q25.3	160201536	160202670	MAS1	3.10E-36
xII	KIFC1	AL589733.20.1.201088	6q27	168140261	168167477	KIF25	4.00E-07
III	NOTCH4	AL078605.30.1.119563	6q27	170231439	170329846	NM_032448	3.70E-41
III	CYP21A2	AC073957.7.1.196204	7p22.3	667370	673509	NM_017781	2.60E-09
III	BAT8	AC005995.3.1.80010	7p22.1	5716846	5720381	Novel	3.80E-07
III	C6orf46	AC073343.6.1.173967	7p22.1	6374466	6390938	Z325	3.10E-14
III	C4B	AC060834.8.1.113686	7p21.3	9410581	9410991	genscan	4.50E-06
III	HSPA1L	AC009945.2.1.75517	7p21.3	10135315	10136415	genscan	2.30E-231
III	NOTCH4	AC013470.10.1.170723	7p21.3	12014883	12054396	Q96SQ3	1.00E-08
II	TAP2/1	AC002486.1.1.79611	7p21.1	20327136	20342642	O14573	6.30E-24
II	TAP2/1	AC005060.3.1.120169	7p21.1	20365471	20439590	Novel	1.80E-20
I	POU5F1	AC005483.1.1.161667	7p14.1	39025428	39150485	NM_007252	1.80E-11
xI	HIST1H2AC	AC004854.3.1.98697	7p13	44512586	44533943	H2-like	7.00E-15
xI	POM121L2	AC074397.7.1.114576	7p12.1	52567227	52568144	Q8N7R1	8.00E-26
III	C6orf26	AC073057.6.1.178105	7p11.2	56890588	56890821	genscan	5.90E-05
xII	ZNF297	AC115220.1.1.115916	7q11.21	62137099	62202401	Novel	7.80E-08
xI	SMA3L	AC115220.1.1.115916	7q11.21	62213350	62213517	genscan	6.70E-11
xII	ZNF297	AC092685.2.1.183263	7q11.21	63492669	63505908	NM_152626	1.30E-10
xI	SMA3L	AC073261.8.1.93403	7q11.21	64065461	64086982	GUSB	1.00E-24
III	BAT2	AC073089.5.1.171788	7q11.21	65101611	65344287	NM_018264	3.10E-24
III	BAT2	AC091738.4.1.131928	7q11.23	70715527	70834138	Novel	2.00E-22
xI	POM121L2	AC005488.2.1.185737	7q11.23	70990215	71062258	POM121	5.30E-92
xI	POM121L2	AC073841.9.1.55588	7q11.23	71351007	71352683	EST gene	2.90E-72
xI	RFP	AC073841.9.1.55588	7q11.23	71357485	71363082	WBSCR20A	8.90E-39
xI	POM121L2	AC006014.3.1.127761	7q11.23	73578273	73650919	POM121	7.40E-71
II	TAP2/1	AC005045.2.1.123947	7q21.12	85566663	85640217	ABCB4	8.00E-22
II	TAP2/1	AC005068.2.1.98472	7q21.12	85668428	85877856	ABCB1	4.90E-25
xII	COL11A2	AC002528.1.1.141120	7q21.3	92559772	92595972	COL1A2	8.10E-24
III	HSPA1L	AC004957.1.1.160687	7q21.3	95968910	95969113	no gene	3.30E-93
xI	BTN1A1/RFP	AC011904.3.1.113879	7q22.1	98022613	98051761	TRIM4	2.30E-19
xI	HLA Class I	AC004522.2.1.100096	7q22.1	98099206	98108247	AZGP1	1.10E-24
I	DDR1	AC011895.4.1.172358	7q22.1	98934759	98959566	EPHB4	8.80E-21
I	HLA-E	AC006329.5.1.145253	7q22.1	99414126	99414392	genscan	1.10E-05
I	DDR1	AC004416.1.1.32173	7q31.2	114790113	114916094	MET	1.20E-13
xII	LYPLA2L	AC073054.2.1.154419	7q21.32	121348077	121348109	genscan	1.20E-37
III	VARS2	AC008038.1.1.202945	7q33	131061808	131063239	Novel	1.00E-11
I	TUBB	AC083874.2.1.186281	7q33	132762314	132762403	genscan	4.00E-09

I	DDR1	AC104597.3.1.161425	7q34	140878465	140894489	EPHB6	2.70E-09
I	DDR1	AC092214.3.1.72045	7q34	141413870	141431627	EPHA1	1.40E-16
III	C6orf46/ZNF297	AC073422.8.1.80743	7q36.1	147046333	147070069	Novel	2.00E-05
III	C6orf46/ZNF297	AC073314.4.1.73888	7q36.1	147374895	147404413	NM_015694	7.40E-05
II	TAP2/1	AC010973.6.1.222605	7q36.1	148971967	148989086	ABCB8	8.10E-24
III	BAT8	AC010973.6.1.222605	7q36.1	149119222	149130915	ABS10	3.60E-06
I	ABCF1	AC021097.5.1.35899	7q36.1	149151360	149170753	ABCF2	1.20E-36
III	NOTCH4	AC110288.6.1.84664	8p23.3	1266863	1277550	Novel	6.60E-11
III	BF	AC023296.6.1.189532	8p23.2	2642344	3292997	CSMD1	3.00E-06
III	C2	AC023296.6.1.189532	8p23.2	2953246	3468313	CSMD1	3.60E-05
III	BAT1	AC012119.7.1.154898	8p21.2	23396081	23396278	Genscan	3.10E-06
III	HSPA1L	AC090820.6.1.138021	8p12	30546063	30546623	genscan	7.10E-170
III	RNF5	AC069120.4.60762.91265	8p11.23	38073473	38073760	genscan	6.20E-73
III	ATP6V1G2	AC120036.3.58309.177413	8q-tel	46087951	46088247	genscan	5.50E-08
xII	RING1	AC016113.9.1.185419	8q11.23	53945099	53945155	no gene	2.20E-19
xII	LYPLA2L	AC060764.10.163950.183771	8q11.23	54898165	54953804	LYPLA1	2.10E-27
I	DDR1	AC046176.11.1.132489	8q12.1	56731370	56862134	LYN	2.10E-12
xI	HIST1H2AC	AC084251.13.1.181400	8q13.3	70956550	70956840	no gene	3.70E-18
xII	RPS18	AC022730.7.1.155468	8q13.3	71276927	71277244	genscan	1.30E-31
xII	RXRB	AC040917.6.1.158031	8q21.11	76308140	76464431	HNF4G	5.80E-28
xII	ZNF297	AC009812.17.1.155405	8q21.13	81446105	81481018	NM_023929	2.30E-10
I	TUBB	AC007992.12.1.146921	8q22.1	96147896	96147985	genscan	1.40E-09
I	POU5F1	AP002851.2.1.200610	8q22.3	103702017	103702310	no gene	1.70E-17
III	BF	AC007719.7.1.150831	8q23.3	112280606	112310596	Novel	3.10E-05
xI	HIST1H2AC	AC022360.23.1.171991	8q23.3	112695588	112695632	no gene	3.60E-05
III	C2	AC007719.7.1.150831	8q23.3	113296979	113550957	Q96PZ3	9.60E-05
I	DDR1	AC022239.14.1.171829	8q23.1	113344396	11414823	BLK	6.00E-10
xII	ZNF297	AC105210.4.125110.131923	8q24.3	145036223	145040537	Q96C28	1.20E-07
xII	HKE4	AC022505.17.153698.191078	8q24.3	145675676	145680153	SLC39A4	7.40E-08
xII	KIFC1	AC084125.8.1.197314	8q24.3	145729611	145737380	NM_145754	2.40E-17
III	NOTCH4	AC084125.4.1.26696	8q24.3	145753310	145765393	PPP1R16A	2.60E-12
III	C6orf46	AF235103.3.1.344150	8q24.3	146143706	146164020	ZNF64	1.10E-14
I	DDR1	AL161450.14.1.171146	9p24.1	5003251	5108156	JAK2	1.60E-06
xII	RING1	AL162411.23.1.59964	9p24.1	6650955	6651159	no gene	7.00E-39
III	CLIC1	AC017067.4.1.191373	9p21.3	23075247	23075813	genscan	1.30E-48
xI	NOL5B	AL445623.2.1.198637	9p21.3	23903514	23903843	genscan	1.10E-15
III	HSPA1L	AL353745.7.1.174850	9p21.1	31159147	31159635	genscan	5.50E-17
II	BRD2	AL589642.6.1.92982	9p21.1	32799398	32805026	TAF1L	2.80E-08
xII	ZNF297	AL158155.24.1.192336	9p13.2	37607553	37634838	NM_014872	2.30E-10
III	C6orf46	AL353770.18.1.130898	9p13.1	39619983	39630670	Q96M55	1.10E-14
III	CYP21A2	AL359997.8.1.169102	9q21.13	66702026	66702607	genscan	1.20E-14
xII	LYPLA2L	AL353637.16.1.133212	9q21.2	71442203	71442490	no gene	3.50E-59
III	BAT1	AL158047.9.1.201629	9q21.32	75746587	75747222	Genscan	2.30E-19
xII	KIFC1	AL354733.15.1.189579	9q21.32	78368024	78383954	Novel	5.50E-27
I	DDR1	AL445532.8.1.171629	9q21.33	79138134	79490850	NTRK2	1.80E-35
xII	ZNF297	AL136981.22.1.182280	9q22.31	87336127	87368066	Q9H559	2.80E-10
xI	GABBR1	AL445495.5.1.155837	9q22.33	92793899	93215009	GPR51	1.60E-09
xII	COL11A2	AL354923.12.1.134965	9q22.33	93449719	93576588	COL15A1	3.10E-12
III	C6orf29	AL450265.11.1.68871	9q31.1	99804995	99897178	CTL1	1.40E-12
III	C9orf29	AL450265.11.1.68871	9q31.1	99804995	99897178	NM_022109	1.40E-12
III	BF	AL158158.14.1.194835	9q31.3	103971543	103990925	Novel	1.70E-11
III	C2	AL158158.14.1.194835	9q31.3	104860730	104923764	NM_153366	2.50E-15
III	NOTCH4	AL354982.12.1.119077	9q31.3	104936930	105074331	Novel	1.20E-33
I	DDR1	AL157881.14.1.162726	9q31.3	105163309	105295448	MUSK	1.80E-41
III	C6orf46	AL159168.15.1.129010	9q31.3	106019666	106038882	Q8TF39	1.80E-14

III	BAT2	AL354877.25.1.116236	9q31.3	106181071	106289393	NM_173521	3.30E-17
xII	ZNF297	AL162588.22.1.76606	9q31.3	107536344	107551166	ZFP37	7.80E-09
III	NOTCH4	AL162425.15.1.177728	9q31.3	108581009	108678635	Novel	6.30E-28
xII	COL11A2	AL445543.20.1.140327	9q31.3	108662161	108805998	Q96JF7	2.60E-13
III	ATP6V1G2	AL160275.14.1.189709	9q32	109082196	109092823	ATP6V1G1	1.50E-12
III	TNF	AL390240.18.1.93876	9q32	109283763	109300763	TNFSF15	3.80E-10
III	TNXB	AL162425.15.1.177728	9q33.1	109514975	109612609	TNC	7.80E-25
I	TUBB	AL589703.6.1.48697	9q33.1	112631091	112631204	genscan	1.60E-10
xII	B3GALT4	AL161911.17.1.109176	9q33.2	115207621	115207998	genscan	6.40E-31
III	C4B	AC006430.22.1.194799	9q33.2	115361172	115459110	C5	5.70E-20
III	C6orf46	AC007066.4.1.190815	9q33.2	117340094	117340094	BIOR	9.80E-65
III	NOTCH4	AL445489.10.1.175869	9q33.3	117788495	118338969	NM_024820	2.20E-76
xII	RXRB	AL354979.17.1.85997	9q33.3	118928956	119180139	NR6A1	9.70E-13
III	HSPA1L	AL354710.17.1.131708	9q33.3	119643682	119650159	HSPA5	8.1E-161
III	PBX2	AL445186.4.1.156124	9q33.3	120156161	120376205	PBX3	9.10E-85
III	C6orf46/ZNF297	AL161731.20.1.182452	9q33.3	121213845	121244041	ZNF297B	7.30E-33
xII	C6orf46/ZNF297	AL354944.22.1.49144	9q33.3	121269484	121289729	Q8NCN2	9.50E-39
I	DDR1	AL161733.20.1.176466	9q34.12	125136236	125309589	ABL1	8.50E-12
III	AIF1	AL157938.22.1.197019	9q34.12	125518441	125545061	NM_031426	1.20E-56
III	BAT2	AL358781.19.1.147492	9q34.13	125852061	125869120	NM_032640	2.20E-65
xII	RAB2L	AL162417.23.1.152863	9q34.2	127529965	127553410	RALGDS	3.70E-29
II	BRD2	AL445931.29.1.175033	9q34.2	128566862	128602533	BRD3	1.90E-133
xII	RXRB	AL669970.6.1.58552	9q34.2	129062693	129101647	RXRA	6.30E-88
xII	COL11A2	AL603650.10.1.131466	9q34.3	129302868	129503955	COL5A1	2.10E-43
III	NOTCH4	AL390778.30.1.221373	9q34.3	129736484	129782241	OLFM1	4.60E-11
III	NOTCH4	AL353615.27.1.37093	9q34.3	130006614	130009927	NM_173520	6.00E-10
xII	ZNF297	AL591038.9.1.51295	9q34.3	130674721	130713948	NM_144653	7.90E-15
III	NOTCH4	AL592301.14.1.188462	9q34.3	131078383	131129726	NOTCH1	2.40E-224
III	EGFL8	AL590226.23.1.149567	9q34.3	131242795	131256617	ZNEU1	9.70E-12
III	AGPAT1	AL590226.23.1.149567	9q34.3	131257082	131271362	AGPAT2	2.50E-29
III	CLIC1	AC068451.2.53215.58850	9q34.3	131578574	131580507	CLIC3	3.30E-35
III	BAT8	AL611925.20.31668.168509	9q34.3	132395728	132553855	HMT1	7.30E-142
I	TUBB	AL713922.8.1.121218	10p15.3	33000	35178	TUBBL	2.70E-190
III	NOTCH4	AL513304.27.1.163243	10p15.3	1427949	1428792	no gene	3.20E-06
xII	KIFC1	AL161932.15.1.143423	10p11.22	3016668	32061904	KIF5B	1.00E-06
III	HSPA1L	AC069544.9.1.214866	10p13	14843884	14877306	NM_016299	1.00E-15
III	BAT8	AC069544.9.1.214866	10p13	14884428	14909880	SU92	1.40E-17
III	NOTCH4	AL133415.12.1.179912	10p13	17152297	17207242	DNMT2	1.10E-15
xII	HKE4	AL590111.14.1.41069	10p12.33	17957452	18048843	NM_152725	7.10E-08
xII	RPS18	AL513128.11.1.184685	10p12.2	22534350	22534484	no gene	1.20E-05
xII	ZNF297	AL117337.25.1.161452	10q11.21	37982417	38009185	ZNF25	1.00E-07
xII	ZNF297	AL161931.13.1.19853	10q11.21	38043229	38099906	ZNF33A	1.30E-10
xII	ZNF297	AL022345.2.1.146328	10q11.21	42553027	42602464	ZNF11B	1.30E-10
III	C6orf46	AL353801.13.1.222490	10q11.21	44964852	44969243	ZNF22	4.10E-14
III	BAT8	AL359377.18.1.172177	10q21.2	60813515	61174843	ANK3	6.70E-05
I	POU5F1	AL356741.11.1.87244	10q21.3	68768389	68953381	Q9HCH9	1.40E-75
III	BAT1	AL359844.15.1.171364	10q22.1	69527140	69554693	DDX21	1.20E-05
III	BAT1	AC016394.13.1.149726	10q22.2	73713068	73782589	Q9Y2I0	1.90E-10
xI	HIST1H2AC	AL391421.27.1.168239	10q22.3	78922958	78923248	no gene	1.00E-14
III	BAT1	AL365434.12.1.158357	10q23.31	91694792	91695643	Genscan	7.30E-17
III	BAT1	AL731553.9.1.161141	10q23.31	91759982	91760050	Genscan	7.40E-17
III	BAT1	AL158040.13.1.213648	10q23.32	92777959	92779325	Novel	1.50E-24
xII	KIFC1	AL356128.27.1.191935	10q23.33	93574641	93636806	KIF11	1.50E-11
III	CYP21A2	AL359672.19.1.143181	10q23.33	95689793	95722511	CYP2C8	3.20E-05
III	NOTCH4	AL442123.12.1.96660	10q24.1	97651046	97838934	SLIT1	6.90E-33
xII	ZNF297	AL135791.12.1.66975	10q24.1	97820294	3785029	Q9NQN2	7.80E-09

II	TAP2	AL392107.16.1.94970	10q24.2	100776166	100845227	ABCC2	7.70E-05
III	CYP21A2	AL358790.22.1.131753	10q24.32	103483494	103490378	CYP17	4.20E-12
xI	RFP	AL391121.29.1.166600	10q24.32	103638012	103651712	TRIM8	1.90E-07
I	DHX16	AL360176.22.1.155699	10q26.2	126728551	126773529	DDX32	1.10E-21
III	CYP21A2	AL161645.14.1.161644	10q26.3	134255131	134266884	CYP2E	2.00E-07
xI	MAS1L	AC108448.5.135208.198047	11p15.4	3499638	3500522	Novel	5.70E-29
xI	RFP/RING1	AC009758.8.1.141485	11p15.4	4708183	4716972	SSA1	9.40E-71
xI	RFP	AC090719.8.1.179177	11p15.4	4921932	4931481	NM_018073	6.50E-72
xI	BTN1A1	AC009758.8.1.141485	11p15.4	4972222	4981011	SSA1	6.10E-38
xI	BTN1A1	AC090719.8.1.179177	11p15.4	5185974	5195520	Novel	1.40E-35
xI	RFP	AC015691.6.1.203036	11p15.4	5919434	5967728	TRIM6	1.30E-45
xI	RFP	AC109341.7.1.202761	11p15.4	5986894	6008393	TRIM5	9.10E-44
xI	BTN1A1	AC015691.6.1.203036	11p15.4	6372568	6413722	TRIM34	1.70E-16
xI	BTN1A1	AC109341.7.1.202761	11p15.4	6459104	6478807	TRIM22	1.70E-16
III	CYP21A2	AC018795.10.1.187836	11p15.2	15932556	15932621	no gene	2.90E-18
III	CYP21A2	AC090835.6.82428.167443	11p15.2	16080713	16094757	Novel	3.10E-19
xI	MAS1L	AC090099.10.28570.173306	11p15.1	19101039	19102007	MRGX3	1.70E-39
xI	MAS1L	AC107948.7.1.156839	11p15.1	19137100	19138068	MRGX4	3.30E-38
xI	MAS1L	AC023078.9.1.163718	11p15.1	19899153	19900121	MRGX1	2.00E-40
xI	MAS1L	AC023078.9.1.163718	11p15.1	19926836	19991360	Novel	2.40E-17
xI	MAS1L	AC027026.9.1.155376	11p15.1	20020747	20021739	MRGX2	1.10E-35
xII	KIFC1	AC023206.6.1.208561	11p14.1	28817128	28904683	NM_031217	5.40E-07
III	BF	AL354921.12.1.106657	11p13	37005965	37099743	Q96JW2	2.60E-06
III	NOTCH4	AC061999.6.1.182549	11p12	37290819	37297112	RAG2	3.90E-26
xII	RXRB	AC090589.8.1.190017	11p11.2	48157255	48168103	NR1H3	1.40E-06
xII	RXRB	AC018410.19.7721.155276	11p11.2	48168666	48229300	MADD	1.10E-06
xII	HKE4	AC090559.5.26090.106816	11p11.2	48306544	48315768	NM_152264	8.60E-14
xII	RXRB	AP001453.4.1.166300	11q13.1	65754594	65765769	ESRRA	9.70E-18
III	NOTCH4	AP000769.4.1.114794	11q13.1	66974259	66987851	SCYL1	1.20E-19
III	NOTCH4	AP001362.5.1.211382	11q13.1	67025483	67041797	Novel	1.70E-22
xI	MAS1L	AP000808.4.1.176380	11q13.3	70444307	70445269	Q8TDS7	7.50E-35
xI	MAS1L	AP003071.2.1.192759	11q13.3	70468658	70477508	MRGF	4.00E-27
xI	HIST1H2AC	AP002336.3.1.112484	11q13.3	71639671	71753357	PPFIA1	4.50E-14
III	NOTCH4	AP000867.4.1.199996	11q13.4	72870983	72871905	Q8NH65	3.20E-07
I	C6ORF18	AP000719.4.1.196424	11q13.4	73253747	73331398	NUMA1	1.90E-06
xII	B3GALT4	AP000752.4.1.194140	11q13.5	78289628	78290785	NM_138706	5.70E-25
III	NOTCH4	AP002768.3.1.186084	11q14.1	79903349	79952134	Q9P2P4	1.50E-09
xI	PRSS16	AP001646.4.1.182328	11q14.1	84074231	84150294	PRCP	4.10E-06
III	BAT1	AP003390.1.1.221091	11q23.3	91694792	91695643	Genscan	2.10E-106
I	TUBB	AP002364.3.1.165702	11q14.3	92344989	92345090	genscan	1.50E-11
I	TUBB	AP002799.3.1.177564	11q14.3	94012104	94012214	genscan	2.30E-14
III	NOTCH4	AP003171.2.1.137000	11q14.3	94081540	94134035	Q8TDW7	2.90E-10
III	BAT8	AP000786.4.1.75440	11q21	95734866	95740452	NM_017704	7.20E-06
III	BAT8	AP002840.2.1.177034	11q23.2	114770675	114783056	Q98NFD2	3.30E-15
III	NOTCH4	AP002840.2.1.177034	11q23.1	114792387	114857963	DRD2	2.10E-10
III	BAT8	AP001267.4.1.194310	11q23.3	119819041	119907224	MLL	2.50E-05
xI	HIST1H2AC	AP003391.1.1.46239	11q23.3	120476378	120477968	H2AFX	2.80E-42
III	BAT1	AP000713.2.1.11316	11q23.3	120964857	120979716	DDX6	1.20E-06
I	POU5F1	AP001150.4.1.157282	11q23.3	121622699	121702405	POU2F3	6.50E-21
III	BAT1	AP001994.4.1.167376	11q23.3	121762568	121762732	No gene	1.50E-105
III	HSPA1L	AP000926.5.1.196973	11q24.1	124441468	124446116	HSPA8	9.10E-243
III	BAT1	AP000842.4.1.179369	11q24.2	127774844	147793446	DDX25	1.80E-10
III	C6orf46/ZNF297	AP001183.4.1.174526	11q24.3	131612944	131697119	NM_014155	7.80E-15
III	BAT1	AC019227.4.1.190314	11q24.3	132166105	132166485	Genscan	1.50E-13
I	PPP1R10	AP000824.4.1.186920	11q24.3	132172516	132173124	genscan	7.20E-17
I	PPP1R10	AP003486.2.1.217488	11q24.3	132258421	132299008	SNXJ	9.60E-17

III	BAT1	AP000435.5.1.124067	11q12.1	60403473	60403808	Genscan	3.00E-12
III	HSPA1L	AC007207.22.1.191877	12p13.32	4100374	4101084	genscan	1.80E-215
xII	TAPBP	AC005840.2.1.140026	12p13.31	6535709	6550143	TAPBP-R	3.30E-05
III	BF	AC006512.12.1.157115	12p13.31	7087044	7087106	no gene	3.10E-06
III	C2	AC006512.12.1.157115	12p13.31	7243215	7245179	C10	6.10E-09
III	C4B	AC006581.16.1.172931	12p13.31	8712438	8724661	Novel	2.80E-05
III	C4B	AC007436.1.1.163881	12p13.31	8928367	8976544	A2M	3.20E-09
III	C4B	AC010175.4.1.127277	12p13.31	9009493	9069023	PZP	6.20E-09
III	BAT1	AC007215.43.2235.65215	12p13.2	12695938	12712573	NM_016355	1.60E-11
xI	HIST1H2AC	AC010168.6.1.104926	12p12.3	15072102	15073039	H2AFJ	1.10E-44
xII	COL11A2	AC004801.1.1.193561	12q13.11	48379436	48410949	COL2A1	2.80E-38
III	BAT1	AC025557.4.146238.171945	12q13.13	49116936	49138712	NM_004818	1.10E-14
I	TUBB	AC011603.33.10243.45426	12q13.12	49484144	49487748	TUBA1	9.00E-11
I	TUBB	AC010173.22.67252.90665	12q13.12	49586964	49595224	TUBA6	2.20E-12
I	TUBB	AC010173.22.160578.20759	12q13.12	49655804	49660085	TUBA1	1.00E-11
xII	RXRB	AC025259.48.1.210158	12q13.13	52474503	52482549	NR4A1	1.10E-07
III	BAT1	AC055716.24.1.110819	12q13.13	53260460	53260801	Genscan	5.40E-27
III	BAT1	AC068988.19.27848.161382	12q13.13	53292388	53292489	Genscan	7.70E-27
III	BAT1	AC073573.27.1.157807	12q13.13	53706088	53706432	Genscan	4.30E-35
III	BAT8	AC073896.29.107190.140910	12q13.2	56645899	56649310	NM_173594	2.30E-05
xII	ZNF297	AC026120.33.1.171998	12q13.3	57617302	57624914	Y352	1.10E-11
III	BAT1	AC117498.1.134066.149599	12q14.1	61110934	61180767	Novel	1.50E-15
xI	NOL5B	AC027288.26.1.177080	12q12.2	80316560	80316925	genscan	6.20E-24
xI	BTN1A1	AC009771.13.122068.178104	12q23.3	107466485	107466640	no gene	2.10E-30
xI	MAS1L	AC063957.22.1.71430	12q23.3	108566149	108613888	CMKLR1	8.80E-05
III	HSPA1L	AC005805.9.96579.142875	12q24.11	111438197	111439237	Novel	1.20E-134
III	CLIC1	AC078875.25.5011.18452	12q24.31	120214564	120215259	Novel	1.60E-14
xII	B3GALT4	AC048338.22.82693.113969	12q24.31	122667939	122671768	B3GNT4	2.10E-23
II	TAP2/1	AC026362.34.74237.162900	12q24.31	123114813	123152304	ABCB9	5.50E-45
I	DHX16	AC093719.6.127047.199959	12q24.31	125177591	125220663	DDX37	8.20E-17
xII	ZNF297	AC026786.5.1.160615	12q24.33	133310141	133338979	ZNF10	6.10E-10
xII	RXRB	AL359457.12.1.129779	13q12.11	14106712	14124427	ESRRAP	3.40E-27
xII	RXRB	AL158032.32.1.172004	13q12.11	15813801	15815722	Novel	4.50E-27
I	TUBB	AL139327.18.1.149559	13q12.11	17727916	17735936	TUBA2	1.10E-09
III	BAT1	AL354828.12.1.168114	13q12.12	21259290	21260180	Genscan	1.10E-111
I	DDR1	AL591024.14.1.76721	13q12.2	22557753	22654705	FLT3	9.00E-05
xI	POM121L2	AL359741.9.1.139877	13q12.3	23332361	23332804	genscan	1.20E-19
xI	POM121L2	AL596092.8.1.153841	13q12.3	23579447	24059956	O94872	2.70E-15
III	HSPA1L	AL137142.20.1.113850	13q12.3	25697387	25722697	H105	8.20E-22
III	BAT1	AL138822.13.1.126502	13q12.3	27152990	27154150	Genscan	1.50E-50
xI	HIST1H2AC	AL159980.14.1.162044	13q13.3	31050674	31050910	genscan	5.60E-20
III	BAT1	AL138706.9.1.195032	13q13.3	35499252	35499851	Genscan	1.40E-45
III	BAT8	AL136218.26.1.159863	13q14.2	44005950	44053746	C13ORF4	3.20E-06
xII	RXRB	AL138997.18.1.172342	13q21.1	50568900	50569076	genscan	5.20E-14
III	BAT1	AL161901.18.1.150054	13q21.2	59178906	59179631	Genscan	1.90E-31
xI	RFP	AL136145.23.1.83809	13q21.32	60838247	60838498	genscan	9.90E-09
I	DHX16	AC001226.1.1.106988	13q22.3	71554239	71557564	Novel	4.40E-10
I	POU5F1	AL445209.4.1.157302	13q31.1	73168139	73172615	POU4F1	7.40E-28
II	TAP2/1	AL157818.12.1.182485	13q32.1	90059271	90340865	ABCC4	7.60E-05
xI	HIST1H2AC	AL160155.19.1.149478	13q32.3	94254802	94255185	H2A-like	1.60E-11
xII	COL11A2	AL390755.5.1.186120	13q34	105188574	105346678	COL4A1	3.30E-17
xII	COL11A2	AL159153.17.1.102319	13q34	105346805	105553028	COL4A2	2.40E-16
III	NOTCH4	AL137002.19.1.132933	13q34	108397279	108411519	F7	2.60E-18
III	NOTCH4	AL161774.49.1.162296	13q34	110946220	111086222	RASA3	9.60E-09

II	PSMB8	AL132780.5.1.191946	14q11.2	17282361	17291410	PSMB5	2.40E-67
xI	HMGN4	AL163052.4.1.181905	14q12	23108107	23108319	no gene	8.20E-06
I	TUBB	AL445383.5.1.172914	14q21.2	41065751	41065876	genscan	7.00E-10
xII	RXRB	AL161756.6.1.176257	14q23.2	58487079	58598473	ESR2	8.90E-05
III	HSPA1L	AL049869.6.1.195840	14q23.3	58801222	58803614	HSPA2	4.80E-266
xI	GPX5	AL139022.4.1.190517	14q23.3	59199531	59203136	GPX2	1.50E-27
III	BAT1	AL391262.3.1.171296	14q24.1	67056068	67056841	Genscan	1.00E-18
III	NOTCH4	AC005479.2.1.140425	14q24.3	68764077	68777511	NPC2	3.70E-22
xII	RXRB	AC008050.6.1.176975	14q24.3	70654841	70785295	ESRRB	5.10E-09
xII	RPS18	AL122020.5.1.149904	14q32.11	85048919	85049456	Novel	4.40E-53
III	NOTCH4	AL132711.4.1.184924	14q32.2	95016286	95017418	no gene	6.90E-22
III	C6orf46/ZNF297	AL590327.3.1.59297	14q32.33	99258499	99259764	Novel	3.20E-14
III	NOTCH4	AL512356.5.1.158468	14q32.33	99443586	99452823	C14orf79	1.60E-59
III	NOTCH4	AL512355.5.1.196132	14q32.33	99772045	99852517	O60342	1.90E-54
III	ATP6V1G2	AL122127.6.1.169802	14q32.33	103287219	103287536	no gene	1.00E-07
II	TAP1	AC116165.3.1.90200	15q11.2	16305068	16351483	Novel	1.70E-35
II	TAP2	AC116165.3.1.90200	15q11.2	16305068	16351483	Novel	7.20E-22
II	TAP1	AC016033.7.99902.141149	15q11.2	16392186	16403533	Novel	7.70E-36
II	TAP2	AC016033.7.99902.141149	15q11.2	16392186	16403533	Novel	7.50E-23
xI	POM121L2	AC090983.10.101166.203171	15q11.2	17724332	17724397	no gene	2.20E-11
II	TAP2	AC091304.12.1.179219	15q13.1	21513802	21524551	Novel	1.50E-44
II	TAP1	AC091304.12.1.179219	15q13.1	21513802	21524551	Novel	2.30E-40
xI	HMGN4	AC022613.13.1.188117	15q13.1	25526308	25530507	HMG17	6.40E-06
III	NOTCH4	AC020661.8.1.191655	15q15.1	34166830	34304183	Q9ULG1	1.70E-42
II	HLA Class II	AC025270.6.1.128484	15q21.1	37899544	37906166	B2M	7.10E-05
xI	BTN1A1/RFP	AC018901.8.1.199503	15q21.1	37924579	37955869	RNF36	1.20E-29
II	HLA-DPB1	AC018901.8.1.199503	15q21.1	38056731	38056793	no gene	6.90E-06
III	CYP21A2	AC020705.4.136565.149466	15q21.1	38739399	68813334	CYP1A2	1.10E-12
III	NOTCH4	AC022467.7.1.193703	15q21.1	41748164	41983082	FBN1	8.10E-27
III	BAT1	AC091700.4.1.97653	15q22.2	55722396	55722959	Genscan	4.20E-08
III	NOTCH4	AC009433.11.1.169638	15q22.31	59286605	59645098	NM_032445	4.90E-28
III	BAT8	AC067837.6.1.173919	15q23	61668952	61687004	FEM1B	1.70E-09
III	BAT8	AC021553.14.1.185596	15q23	61692853	61823052	ITGA11	2.50E-07
xII	RXRB	AC104938.2.66191.114293	15q23	65200660	65208271	NR2E3	3.40E-22
III	RNF5	AC048383.8.169960.172969	15q23	66533763	66534050	genscan	2.30E-74
II	BTNL2	AC022188.7.15746.68046	15q24.1	67107924	67122957	NM_025240	5.80E-18
III	CYP21A2	AC020705.4.92855.102206	15q24.1	68848252	68854305	CYP1A1	4.90E-15
I	DDR1	AC027243.13.89123.218680	15q24.2	69598117	69693731	ETFA	1.90E-13
xI	MOG	AC022188.7.15746.68046	15q24.1	70012753	70027796	NM_025240	9.70E-10
III	CYP21A2	AC091230.8.108454.128536	15q24.1	70970184	70976444	Novel	9.00E-12
I	DDR1	AC011966.7.1.167862	15q25.3	81649264	82028917	NTRK3	7.20E-24
xII	KIFC1	AC079075.5.54114.209978	15q26.1	83693866	83715674	ANPEP	1.10E-15
xI	HIST1H2AC	AC091544.9.1.126968	15q26.1	87091764	87110331	H2 -like	9.40E-17
xII	RXRB	AC016251.9.1.182943	15q26.2	90630767	90631006	no gene	4.20E-43
I	DDR1	AC069029.9.1.191018	15q26.3	93033248	93342019	IGF1R	3.10E-07
xI	BTN1A1/RFP	AJ003147.1.1.239566	16p13.3	3325667	3340266	MEFV	2.70E-35
II	BRD2	AC004651.1.1.42016	16p13.3	3810213	3964357	CREBBP	1.80E-05
xII	B3GALT4	AC040160.4.1.209574	16q22.1	6761048	67659124	FHOD1	3.70E-19
II	TAP2	AC025778.7.1.207614	16p13.12	15526117	15599260	ABCB6	1.30E-06
III	NOTCH4	AC106796.1.45233.67716	16p12.3	19774406	19794065	UMOD	2.00E-11
xI	HMGN4	AC093509.2.1.120576	16p12.1	25470563	25470814	Q96C64	3.90E-06
xII	KIFC1	AC023831.8.22510.115251	16p11.2	30014062	30032892	QPRT	2.00E-09
III	C6orf46	AC002310.1.1.120955	16p11.2	31053573	31058082	NM_033410	5.30E-09
III	C6orf46	AC093249.3.1.185664	16p11.2	31102367	31110170	Q96CS4	5.30E-08
xII	ZNF297	AC106886.2.20127.148471	16p11.2	31281559	31287006	Q9UEG4	6.00E-10

xI	RFP	AC009088.7.1.233305	16p11.2	31722146	31734564	Q8N4X6	1.40E-33
II	PSMB8	AC007494.7.1.206113	16q12.1	47130249	47435934	PSMB10	3.80E-43
II	TAP1	AC096996.1.1.194627	16q12.1	48249418	48329904	ABCC11	2.80E-05
xI	UBD	AC026473.7.1.170393	16q21	51486489	51486719	genscan	6.60E-05
xII	KIFC1	AC092118.2.1.148401	16q13	57876138	57920423	KIFC3	1.20E-25
III	BAT1	AC004531.1.1.191565	16q22.1	58750455	58752074	DDX28	2.00E-18
xII	B3GALT4	AC074143.4.1.152953	16q22.1	67560936	67562065	NM_033309	2.70E-19
I	DHX16	AC009087.4.1.174933	16q22.2	72851029	72870003	DDX38	2.90E-50
xII	ZNF297	AC009078.6.1.176926	16q23.1	76142879	76166439	NM_153688	7.80E-09
III	BAT1	AC093491.2.1.162178	16q24.1	76485011	76485421	Genscan	5.20E-16
III	CLIC1	AC092327.3.1.189757	16q24.1	77844902	77845267	Genscan	7.10E-08
xII	ZNF297	AC009113.5.61390.188481	16q24.3	90272052	90283297	Q96MU6	3.70E-10
I	TUBB	AC092143.3.1.183047	16q24.3	90971498	90989716	TUBBL	9.20E-195
xII	ZNF297	AC090617.7.1.169947	17p13.3	2300029	2302554	HIC1	4.70E-10
III	BAT1	AC015799.7.1.66824	17p13.3	2858331	2858387	genscan	1.70E-20
I	DDR1	AC087742.7.63895.97713	17p13.2	4468643	4469263	EST gene	4.80E-11
II	PSMB9	AC027820.9.1.56340	17p13.2	5044936	5047269	PSMB6	3.50E-26
xII	KIFC1	AC004771.1.1.91927	17q13.2	5239168	5273857	KIF1C	8.70E-05
xII	ZNF297	AC087500.12.1.136618	17p13.2	5420893	5425754	Q96JF6	6.10E-10
I	DHX16	AC004148.1.1.118276	17p13.2	5683637	5710243	DDX33	2.10E-45
I	DDR1	AC113189.3.50089.71700	17p13.1	8014004	8022650	TNK1	1.10E-16
III	BAT1	AC016876.5.1.48645	17p13.1	8215901	8221709	EIF4A1	3.20E-18
III	BAT1	AC007421.12.1.95240	17p13.1	8215901	8221709	EIF4A1	1.30E-05
xII	RPS18	AC013248.5.1.66571	17p12	15723309	15723764	Novel	5.60E-60
xII	ZNF297	AC005324.1.1.176643	17p12	16717966	16767419	ZNF386	2.20E-10
III	C6orf46	AJ009612.5.1.148978	17p11.2	17565367	17583130	ZNF287	4.10E-13
III	C6orf46	AC005822.1.1.169931	17p11.2	17634716	17638760	YD49	4.10E-14
xII	ZNF297	AC026271.6.1.171978	17p11.2	20226123	20246008	Novel	1.70E-10
III	PBX2	AC087499.8.20079.65528	17p11.2	20668666	20668839	genscan	2.80E-05
xI	UBD	AC087575.3.156902.181085	17q	24638783	24639010	UBB	1.30E-12
I	FLOT1	AC024267.9.50190.98519	17q11.2	29105687	29123905	FLOT2	1.70E-31
xII	RXRB	AC068669.4.36251.62842	17q21.1	40302169	40309811	NR1D1	6.20E-10
xII	RXRB	AC080112.4.61535.75578	17q21.2	40640652	40689179	RARA	7.90E-13
III	NOTCH4	AC006070.1.1.161987	17q21.2	41487285	41488289	KRTAP9-9	4.80E-18
III	NOTCH4	AC003958.1.1.127834	17q21.2	41624135	41630437	KRTHA3B	6.70E-10
I	DHX16	AC068675.9.124153.141665	17q21.31	43738377	43778728	DDX8	4.80E-52
xII	KIFC1	AC015936.7.29291.133312	17q21.31	45189863	45189988	no gene	1.50E-05
xII	COL11A2	AC015909.8.44136.121814	17q21.33	47864916	47882452	COL1A1	4.00E-32
xII	KIFC1	AC019315.9.1.152057	17q22	54381662	54383970	NM_032559	6.40E-24
xI	RFP	AC004584.1.1.104871	17q23.2	57450065	57472912	ZNF147	3.60E-19
I	DHX16	AC004167.1.1.124876	17q23.2	60092562	60135284	NM_024612	4.60E-25
I	DHX16	AC005702.1.1.147686	17q23.2	60503679	60528304	Novel	7.50E-15
III	BAT1	AC015651.18.1.191583	17q23.3	64290823	64323079	NM_007372	1.20E-08
III	CLIC1	AC004805.1.1.184263	17q24.1	64773117	64773701	Novel	5.70E-08
III	BAT1	AC009994.6.166827.180372	17q24.2	68976215	68982889	DDX5	2.00E-07
III	BAT1	AC087741.2.60294.77121	17q25.3	81412654	81424538	IF4N	6.80E-06
I	TUBB	AP001005.5.1.137000	18p11.32	35028	37159	TUBBL	2.80E-184
III	BAT1	AP002449.2.169334.172757	18p11.21	12998814	12998903	genscan	7.60E-17
II	C6orf10	AC006238.1.1.211945	18q11.2	23904177	23904497	genscan	2.10E-06
xII	ZNF297	AC105101.6.1.172381	18q12.1	45351393	45363238	O75453	5.40E-18
xII	ZNF297	AC006130.1.1.84984	19p13.3	2936627	2947795	NM_024967	4.90E-11
I	DDR1	AC005777.1.1.43190	19p13.3	3847245	3871088	MATX	9.30E-07
xII	ZNF297	AC016586.7.116093.145761	19p13.3	4117018	4136099	O00456	5.40E-16
III	BAT8	AC005523.1.1.41468	19p13.3	4860388	4864189	FEM1A	2.70E-09

I	TUBB	AC010503.8.1.141295	19p13.3	6562943	6570948	TUBBL	1.30E-202
III	TNF	AC008760.7.1.200167	19p13.3	6733175	673234	TNFSF14	6.20E-10
III	C4B	AC008760.7.1.200167	19p13.3	6746489	6789295	C3	3.50E-27
III	NOTCH4	AC020895.8.1.139846	19p13.3	6959105	7022006	EMR1	1.00E-08
I	DDR1	AC010311.9.1.91172	19p13.2	7254547	7432507	INSR	3.10E-12
xII	LYPLA2L	AC010336.7.1.160769	19p13.2	8042351	8049685	Novel	3.30E-96
III	NOTCH4	AC022146.6.66353.150193	19p13.2	8235200	8317297	FBN3	5.30E-34
III	EGFL8	AC022146.6.66353.150193	19p13.2	8275726	8322330	NM_032447	2.60E-08
xII	COL11A2	AC008742.8.1.194623	19p13.2	10191743	10242653	COL5A3	3.10E-67
I	DDR1	AC011557.6.1.30505	19p13.2	10684031	10714039	TYK2	2.20E-07
III	LSM2	AC011475.6.1.179953	19p13.2	10932292	10932453	no gene	5.70E-23
III	C6orf29	AC011475.6.1.179953	19p13.2	10959135	10978061	CTL2	5.20E-59
xII	RAB2L	AC024575.6.1.119638	19p13.2	11718017	11752815	Q8TEP0	1.30E-24
xII	ZNF297	AC011446.6.1.115932	19p13.2	13622357	13628643	STX10	2.20E-14
III	BAT1	AC008569.7.1.227245	19p13.13	14887089	14897635	DDX39	4.10E-98
III	NOTCH4	AC005327.1.1.37988	19p13.12	15236788	15282936	EMR2	5.70E-05
III	EGFL8	AC004663.1.1.41150	19p13.12	15649643	15690991	NOTCH3	6.00E-05
III	NOTCH4	AC004663.1.1.41150	19p13.12	15664050	15705404	NOTCH3	7.70E-227
II	BRD2	AC114486.2.1.179070	19p13.12	15741907	15784868	BRD4	2.30E-90
xII	RXRB	AC010646.5.1.41461	19p13.12	17734984	17748449	NR2F6	7.00E-41
xII	B3GALT4	AC008761.7.1.226170	19p13.12	18106912	18149110	Q9UPW8	1.10E-12
xII	B3GALT4	AC005952.1.1.39976	19p13.11	18298235	18315904	B3GNT3	2.50E-14
xII	RPS18	AC020904.7.1.148824	19p13.11	18551604	18551837	EST gene	5.30E-46
III	BAT1	AC002985.1.1.38041	19p13.11	19422473	19431417	NM_019070	1.50E-05
III	PBX2	AC011448.4.1.165122	19p13.11	20063771	20120711	PBX4	2.10E-68
xII	ZNF297	AC008751.6.1.169089	19p13.11	21436099	21452779	ZNF85	2.20E-10
III	C6orf46	AC016628.6.1.41153	19p13.11	23871643	23887148	Novel	6.70E-16
xII	ZNF297	AC020910.7.1.203201	19q13.12	35697932	35713073	Q96NL3	2.30E-11
III	BAT8	AD000671.1.1.46251	19q13.12	36657876	36678735	TRX2	1.70E-05
xII	ZNF297	AC092295.2.1.165566	19q13.12	37465837	37479151	EST gene	1.00E-10
III	C6orf46	AC008806.4.1.135173	19q13.13	38293492	38349772	NM_152484	5.30E-10
xII	ZNF297	AC022148.5.1.198751	19q13.13	38430631	38431506	Q8N3U1	1.00E-10
III	NOTCH4	AC011500.7.1.200430	19q13.2	40327364	40358466	SUPT5H	1.10E-27
III	NOTCH4	AC010412.8.1.155085	19q13.2	41494937	41527447	LTBP4	1.50E-33
III	CYP21A2	AC008537.5.1.169089	19q13.2	41988964	41996369	CYP2A6	1.60E-09
I	DDR1	AC011510.7.1.129402	19q13.2	42116547	42159395	AXL	5.40E-09
III	CYP21A2	AC008962.9.1.154169	19q13.2	42259850	42273778	CYP2F1	2.90E-08
xII	B3GALT4	AC011526.7.1.40887	19q13.2	42323217	42324407	Novel	6.40E-20
III	CYP21A2	AC011510.7.1.129402	19p13.2	42338667	42352612	CYP2S1	9.30E-06
I	POU5F1	AC024076.4.1.39443	19q13.2	42986837	43028331	POU2F2	4.30E-23
III	NOTCH4	AC011497.6.1.168586	19q13.2	43248294	43273290	EGFL4	4.00E-09
I	DHX16	AC008754.8.1.66792	19p13.32	48246893	48270700	DDX34	7.80E-27
I	DHX16	AC073548.4.1.66051	19q13.32	48322806	48366009	SLC8A2	2.20E-17
xI	HLA Class I	AC010619.7.1.179394	19q13.33	50384629	50397727	FCGRT	8.50E-10
xII	RXRB	AC008655.7.1.123149	19p13.33	51241040	51247541	NR1H2	1.10E-06
xI	MAS1L	AC005946.1.1.37392	19q13.33	52688362	52689423	FPRL2	9.40E-06
III	C6orf46	AC010320.9.1.220458	19q13.41	53262462	53283017	Q96JK0	3.60E-15
III	C6orf46	AC022150.6.1.228156	19q13.41	53461403	53462023	ZNF137	9.00E-16
xII	ZNF297	AC013256.1.1.36095	19q13.43	57406251	57442939	NM_022103	1.40E-10
xII	ZNF297	AC005498.1.1.37321	19q13.43	57504539	57522397	ZFP28	2.80E-10
xII	ZNF297	AC003682.1.1.153875	19q13.43	58536894	58544255	Q9BWM5	1.10E-11
III	C6orf46	AC003682.1.1.153875	19q13.43	58579587	58587258	ZNF134	1.60E-15
III	C6orf46	AC003006.1.1.84114	19q13.43	58734983	58745942	NM_017652	5.20E-12
III	C6orf46	AC012313.7.1.185417	19q13.43	59398153	59405560	ZNF132	1.70E-25
xII	ZNF297	AC012313.7.1.185417	19q13.43	59478868	59485159	NM_032792	5.20E-29
xI	NOL5B	AL049712.12.1.159272	20p13	2580791	2587039	NOL5A	4.00E-29

III	NOTCH4	AL035456.26.1.125952	20p12.2	10566334	10602636	JAG1	1.30E-49
xII	KIFC1	AL049794.16.1.119696	20p12.1	16200749	16502021	C20ORF23	1.30E-09
III	NOTCH4	AL049651.2.1.97912	20p11.21	22964121	22965287	SSTR4	4.00E-10
III	EGFL8	AL118508.27.1.123832	20p11.21	23048052	23055034	C1QR1	1.90E-05
III	NOTCH4	AL118508.27.1.123832	20p11.21	23054616	23054911	Q8WY72	1.90E-16
xI	BTN1A1	AL080312.14.1.94664	20p11.21	25027858	25028403	genscan	1.60E-08
I	DDR1	AL049539.21.1.111694	20q11.21	30388101	30437940	HCK	6.80E-11
xII	KIFC1	AL121897.32.1.145414	20q11.21	30613467	30669435	KIF3B	3.10E-46
I	DDR1	AL133293.28.1.68662	20q11.23	35700500	35722250	SRC	3.00E-15
I	DHX16	AL023803.3.1.155379	20q11.23	37279429	37356793	DDX35	2.10E-19
xII	RXRB	AL132772.14.1.83798	20q13.12	42718338	42747410	HNFA4	4.50E-26
xII	ZNF297	AL354745.11.1.13535	20q13.12	44818128	44830619	ZNF334	1.60E-05
III	BAT1	AL049766.14.1.110293	20q13.13	47524305	47549031	DDX27	1.70E-10
I	TUBB	AL109840.24.1.142094	20q13.32	57282669	57290069	TUBBL	3.02E-165
III	NOTCH4	AL354836.13.1.141056	20q13.33	60601582	60607445	ADRM1	2.50E-09
III	NOTCH4	AL121673.41.1.151163	20q13.33	61421476	61437370	C20orf59	2.90E-07
I	DDR1	AL121829.30.1.113196	20q13.33	61996950	62006900	PTK6	3.20E-14
xII	ZNF297	AL121845.20.1.120917	20q13.33	62212439	62299987	Novel	3.90E-12
III	BAT5	AL118506.27.1.139505	20q13.33	62330271	62331761	C20ORF135	1.50E-117
III	HSPA1L	AF130358.2.1.197778	21q11.2	12307991	12372209	ABCC13	1.80E-44
III	HSPA1L	AF130249.1.1.97083	21q11.2	12405307	12417341	STCH	3.30E-46
III	CLIC1	AP000330.2.1.170377	21q22.12	32702115	32750955	CLIC6	2.60E-38
xII	B3GALT4	AF064860.2.1.170121	21q22.2	37690022	37690381	genscan	9.90E-32
III	BAT8	AP001615.1.1.124516	21q22.3	39783391	39784221	genscan	4.60E-17
xII	ZNF297	AP001620.1.1.95449	21q22.3	40039452	40061155	ZNF295	5.40E-10
xI	HIST1H2AC	AB001523.1.1.122638	21q22.3	42024993	42118907	TMEM1	7.20E-08
III	NOTCH4	AP001067.1.1.148845	21q22.3	42510546	42724266	C21orf29	2.10E-24
III	NOTCH4	AL163301.2.1.340000	21q22.3	43278171	43300571	C21orf80	5.40E-06
xII	COL11A2	AL163302.2.1.340000	21q22.3	43527445	43550695	SLC19A1	9.70E-12
III	BAT1	AP001604.1.1.186930	21q21.3	25401349	25401540	genscan	4.90E-09
I	TUBB	AC008079.23.1.170102	22q11.21	15544495	15554618	TUBA8	4.00E-10
xI	POM121L2	AC008103.27.1.98557	22q11.21	15773964	15776531	C22.2	5.60E-08
xI	POM121L2	AC000095.3.1.43728	22q11.21	15945799	15947779	C22.3	3.00E-10
III	NOTCH4	AC005500.2.1.192592	22q11.21	17480798	17552052	SRC2	1.00E-09
III	NOTCH4	AC007731.14.1.182617	22q11.21	17480798	17552052	SRC2	2.60E-09
xI	POM121L2	AC007050.25.1.163908	22q11.21	17742343	17744900	C22.3	1.50E-09
III	C6orf46/ZNF297	AP000557.2.1.150036	22q11.21	18470386	18504441	H1C2	6.00E-19
II	HLA-DRB3/1	D87023.1.1.40392	22q11.22	19936243	19936975	IGLC1	2.80E-06
xI	POM121L2	AP000354.1.1.164756	22q11.23	21343934	21357627	NM_014549	9.60E-10
xI	POM121L2	AP000356.1.1.163795	22q11.23	21749568	21750854	POM121L1	1.70E-10
I	DDR1	AL022329.9.1.221507	22q12.1	22656954	22816015	ADRBK2	6.00E-21
xI	BTN1A1/RNF	AC002059.3.1.173029	22q12.2	26530668	26534540	RFPL1	2.40E-26
III	RNF5	AC002073.1.1.128978	22q12.2	28252236	28299047	Q96GF1	8.10E-20
xI	HIST1H2AC	AL096701.14.1.168110	22q12.2	28613161	28613624	novel	2.50E-07
xI	BTN1A1/RFP	AL008723.8.1.154414	22q12.3	29282473	29295511	RFPL2	5.70E-26
xI	BTN1A1/RFP	AL021937.1.1.173354	22q12.3	29447342	29453195	RFPL3	5.70E-26
III	BAT1	Z97056.1.1.124990	22q13.1	35496212	35516829	DDX17	2.00E-05
II	BRD2	AL096765.12.1.13053	22q13.2	38102320	38190075	EP300	1.20E-05
III	CYP21A2	AL021878.1.1.114847	22q13.2	39138588	39142847	CYP2D6	3.30E-09
III	NOTCH4	Z98047.1.1.47542	22q13.31	42534087	42532337	FBLN1	5.30E-12
III	NOTCH4	AL031588.1.1.127168	22q13.31	43322995	43499263	CELSR1	1.20E-09
III	BAT1	AC117517.7.1.121628	Xp22.11	21589340	21590458	genscan	4.40E-37
xI	HIST1H2AC	AL121577.1.1.175531	Xq21.1	35832664	35878934	XK	2.40E-08
xI	HIST1H2AC	AL121578.1.1.337101	Xp11.4	36141216	36141440	genscan	2.80E-08

III	CLIC1	AL391259.15.1.163520	Xp11.4	38963014	38963730	Genscan	2.90E-07
III	BAT1	AL391647.16.1.60310	Xp11.4	39441330	39472404	DDX39	6.20E-07
xII	ZNF297	AL590223.12.1.40331	Xp11.3	45567068	45603010	ZNF41	4.60E-10
xII	ZNF297	Z98304.1.1.209618	Xp11.23	46096233	46103478	Q96QH7	4.70E-10
III	BAT8	AC115618.1.1.158455	Xp11.23	46756067	46756093	no gene	3.40E-15
III	BAT8	AF196970.1.1.112595	Xp11.23	46815791	46828063	SUV39H1	1.30E-15
III	BAT1	AL445236.22.1.149749	Xp11.22	50567899	50607906	Novel	9.90E-39
xII	HSD17B8	Z97054.1.1.132805	Xp11.22	51162495	51165605	HADH2	3.60E-06
xII	KIFC1	AL357752.19.1.178868	Xq13.1	66987254	67117990	KIF4A	2.70E-08
III	BAT1	AL359740.24.1.98104	Xq13.2	70525227	70526105	genscan	1.00E-13
II	TAP2/1	AL359545.12.1.127243	Xq13.3	71447688	71550705	ABCB7	2.90E-08
xII	KIFC1	AL021786.2.1.70665	Xq21.1	75540478	75573738	Novel	1.40E-15
I	POU5F1	Z82170.1.1.127247	Xq21.1	79839811	79841286	POU3F4	6.20E-39
III	BAT1	AL136362.10.1.135240	Xq21.31	88337615	88338796	EST gene	1.30E-57
I	TUBB	AL390840.17.1.197611	Xq21.32	88819807	88819917	genscan	1.50E-09
xII	COL11A2	AL136080.6.1.116106	Xq23	104474931	104758796	COL4A6	2.50E-15
xII	COL11A2	AL031622.1.1.104674	Xq23	104759239	105016860	COL4A5	1.80E-17
III	HSPA1L	AC004822.1.1.127824	Xq23	111134972	111136228	genscan	1.30E-208
III	VARS2	AC005000.2.1.107314	Xq23	112024040	112063337	Novel	3.30E-05
I	TUBB	AC003012.1.1.104810	Xq24	112252763	112252876	genscan	4.60E-12
xII	ZNF297	AC002086.1.1.112686	Xq24	116370267	116377851	NM_006777	6.50E-09
III	HSPA1L	AC002377.1.1.141779	Xq24	117230333	117231259	genscan	1.50E-221
III	HSPA1L	AL391241.21.1.157860	Xq25	120232224	120232373	genscan	9.40E-65
III	NOTCH4	AL627231.9.1.146366	Xq25	121306552	121307673	Novel	6.10E-12
xII	ZNF297	AL590282.6.1.139296	Xq26.3	131228672	131323794	ZNF75	3.70E-10
III	C6orf46	U82670.3.1.279526	Xq28	149081175	149084483	ZNF275	5.30E-12
III	CLIC1	AL356738.14.1.174693	Xq28	150871755	150929271	CLIC2	8.80E-52
xI	HIST1H2AC	AC019175.4.37111.45694	Xq28	151078382	151078898	H2AFB	1.80E-13
xI	HIST1H2AC	AL592156.4.1.134995	Xq21.1	35423125	35423349	genscan	2.00E-11
III	BAT1	AC010129.3.1.44145	Yp11.2	5171386	5172558	Novel	6.30E-59
III	BAT1	AC004474.1.1.148280	Yq11.21	14326902	14356562	DBY	4.00E-05

Appendix 3

Primers used to amplify a paralogue specific probe for use in Northern blot, Dot blot and Southern blot analyses. 'T' stands for the annealing temperature.

Gene	*Primer*	*Sequence*	*T (°C)*	*Size (bp)*
AIF1	F	TGACCATGCTGATGTATGAGGAAAAAGCGA	62	200
	R	GATCTGGAGGAGGGGGTAAT		
AIF1-L	F	TGACCATGTTAAGGGAGGAGAGCAAGCA	62	251
	R	CTGAGCCCTTAGCCAGAGAA		
BRD2	F	TGACCATGGAGGGATGCAGGGACATTT	62	411
	R	AACAAAGACAGTCCAGGGGA		
BRDT	F	TGACCATGGGGTACCATTGATATGACCCTT	62	199
	R	CTGTTTAATCATTTTAGAGCAGTCA		
BRD3	F	TGACCATGGACAGATGGATGTCGCACAC	62	425
	R	CAAATGACAAGGACAATGCG		
BRD4	F	TGACCATGGTGAAAGGGACAGGACTCCA	65	508
	R	CAGTGAGAAGCATGCTGTGG		
C4	F	TGACCATGAGAGATGACTCCGCGTCTGT	65	395
	R	ATTCTCCTTCTGCCCCAGAT		
C3	F	TGACCATGCATTCCCCCACTCCAGATAA	65	214
	R	ACATGAAGGTGAGGCAGGTC		
C5	F	TGACCATGTTGCACTTATGGACTCCTGTTG	65	352
	R	GATCAGTTTCCTGTTCCTTGGT		
CLIC1	F	TGACCATGAAGTACCGGGGATTCACCAT	65	310
	R	CTTTCCCTCATCCCCTCTTC		
CLIC4	F	TGACCATGGGAGATTGGAGTCTGAATGGA	65	384
	R	AATGGGTTTAAGGGCACAGA		
CLIC3	F	TGACCATGGTACGCCGCTACCTGGAC	65	153
	R	CCCGACAAAGATGCCTTTATT		
CLIC5	F	TGACCATGTGTTGATGCCAAAATACCCA	65	427
	R	GACCACCTCCTAAATGTGGC		
CLIC6	F	TGACCATGTGTGGCCAAGAAGTACAGAGAT	65	146
	R	TTGCAACATCTGAATATGCG		
CLIC2	F	TGACCATGGAATTCTCAGGAGTCTGGCG	65	350
	R	GCAGTGGTTTGCCATACAGA		
GPX5	F	TGACCATGTAGCAATGGGGTCACAGTCA	65	277
	R	TCCTCTCCAGGTGCCATAAC		
GPX4	F	TGACCATGTCCACAAGTGTGTGGCCC	65	186
	R	CACAAGGTAGCCAGGGGTG		
GPX3	F	TGACCATGAACCCAAAGGAAAAACCAGC	62	451
	R	GAGTCTCAAGCCAGTGGACC		
GPX1	F	TGACCATGCTCTTCGAGAAGTGCGAGGT	65	439
	R	ACTGGGATCAACAGGACCAG		
GPX2	F	TGACCATGTCCTACTCCATCCAGTCCTGA	62	256
	R	CTTCACGCCTCTCAGACACC		
NOTCH4	F	TGACCATGCATTAAAAGGCAGGCTGGAA	65	475
	R	CATCCCCACAGTGGAGTTCT		
NOTCH2	F	TGACCATGATGAGGAGGACAACACTGCC	65	395
	R	GCATCACAGCCAATTGCTTA		
NOTCH1	F	TGACCATGCAATACTGCATCCATGGCCT	65	244

	R	GTCCCTGAGCAACCATCTGT		
NOTCH3	F	TGACCATGATGTTCCATAGCCTTGCTGG	65	294
	R	GGGAATTCAGCTACACAGGG		
PBX2	F	TGACCATGGCAGGGCTGGACTCAGTAAT	62	409
	R	CACTTCCAACCTGTCCCAGT		
PBX1	F	TGACCATGCAGGAGGGAGGGTTTCTCTC	62	267
	R	TCAGTGATATGAGAGAGGGCG		
PBX3	F	TGACCATGCGAGTGTGGAAACATTGGGT	62	325
	R	TCAATCCAGGGTGTAATCCA		
PBX4	F	TGACCATGGTTTGGGGGATAAGCAGGAA	62	286
	R	GAAAATCTGTGCCCAGTCCT		
RXRB	F	TGACCATGAAGAAATGCCAGTGGTGGAG	62	263
	R	AAAGGAGCCCCAAAGAGAAG		
RXRG	F	TGACCATGTCCTGACTAATCCCAGAGGG	62	215
	R	CATAGCCTGCGGGAAACTT		
RXRA	F	TGACCATGTATACTTGGATATGGCGGGG	65	299
	R	CGGAGAAGCCACTTCACAGT		
TUBB_6p21.3	F	TGACCATGAGAGCAACATGAACGACCTG	65	200
	R	TGGAGGGAGATTGAAAGTGG		
TUBB2_18p11.3	F	TGACCATG TTCCTTCTTGAACCCTGGTG	65	225
	R	TTTATTTTGTGGCCCCTCAG		
TUBB5_19p13.3	F	TGACCATGCTGAATCCCCTCTGACTCCA	65	293
	R	CCTCTCTTCCTCACAGGCAC		
TUBB4QL_10p15.3	F	TGACCATGACAGCATCTGGTTTTGCCTC	65	130
	R	CCACTGGAATGCTTGTTCCT		
TUBB4_16q24.3	F	TGACCATGCAGCTGGAGTGAGAGGCAG	65	201
	R	GCCTGGAGCTGCAATAAGAC		
TUBB1_20q13.3	F	TGACCATGTGCACTCACCATTAGCTTCG	65	396
	R	TAGTCAGGCACCTGGCTCTT		

Appendix 4

Primers used to generate paralogue specific PCR products for each paralogue. The products were used to spot on to the microarrays and were also labelled and used to hybridise to the 'Paralogue Microarray'. 'T' stands for the annealing temperature. They were also used in the RT-PCR experiments.

Gene	*Primer*	*Sequence*	*T (°C)*	*Size (bp)*
AIF1	F	TGACCATGCTGATGTATGAGGAAAAAGCGA	62.5	200
	R	GATCTGGAGGAGGGGGTAAT		
AIF1-L	F	TGACCATGTTAAGGGAGGAGAGCAAGCA	62.5	251
	R	CTGAGCCCTTAGCCAGAGAA		
BRD2	F	TGACCATGGAGGGATGCAGGGACATTT	62.5	411
	R	AACAAAGACAGTCCAGGGGA		
BRDT	F	TGACCATGGGGTACCATTGATATGACCCTT	62.5	199
	R	CTGTTTAATCATTTTAGAGCAGTCA		
BRD3	F	TGACCATGGACAGATGGATGTCGCACAC	62.5	425
	R	CAAATGACAAGGACAATGCG		
BRD4	F	TGACCATGGTGAAAGGGACAGGACTCCA	65	508
	R	CAGTGAGAAGCATGCTGTGG		
C4	F	TGACCATGAGAGATGACTCCGCGTCTGT	65	395
	R	ATTCTCCTTCTGCCCCAGAT		
C3	F	TGACCATGCATTCCCCCACTCCAGATAA	65	214
	R	ACATGAAGGTGAGGCAGGTC		
C5	F	TGACCATGTTGCACTTATGGACTCCTGTTG	65	352
	R	GATCAGTTTCCTGTTCCTTGGT		
CLIC1	F	TGACCATGAAGTACCGGGGATTCACCAT	62.5	310
	R	CTTTCCCTCATCCCCTCTTC		
CLIC4	F	TGACCATGGGAGATTGGAGTCTGAATGGA	65	384
	R	AATGGGTTTAAGGGCACAGA		
CLIC3	F	TGACCATGGTACGCCGCTACCTGGAC	65	153
	R	CCCGACAAAGATGCCTTTATT		
CLIC5	F	TGACCATGTGTTGATGCCAAAATACCCA	65	427
	R	GACCACCTCCTAAATGTGGC		
CLIC6	F	TGACCATGTGTGGCCAAGAAGTACAGAGAT	65	146
	R	TTGCAACATCTGAATATGCG		
CLIC2	F	TGACCATGGAATTCTCAGGAGTCTGGCG	65	350
	R	GCAGTGGTTTGCCATACAGA		
GPX5	F	TGACCATGTAGCAATGGGGTCACAGTCA	62.5	277
	R	TCCTCTCCAGGTGCCATAAC		
GPX4	F	TGACCATGTCCACAAGTGTGTGGCCC	62.5	186
	R	CACAAGGTAGCCAGGGGTG		
GPX3	F	TGACCATGTCTGGGTCTACCACACTCCC	62.5	329
	R	GAGTCTCAAGCCAGTGGACC		
GPX1	F	TGACCATGCTCTTCGAGAAGTGCGAGGT	62.5	439
	R	ACTGGGATCAACAGGACCAG		
GPX2	F	TGACCATGTCCTACTCCATCCAGTCCTGA	62.5	256
	R	CTTCACGCCTCTCAGACACC		
NOTCH4	F	TGACCATGCATTAAAAGGCAGGCTGGAA	62.5	475
	R	CATCCCCACAGTGGAGTTCT		
NOTCH2	F	TGACCATGATGAGGAGGACAACACTGCC	65	395
	R	GCATCACAGCCAATTGCTTA		

NOTCH1	F	TGACCATGCAATACTGCATCCATGGCCT	65	244
	R	GTCCCTGAGCAACCATCTGT		
NOTCH3	F	TGACCATGATGTTCCATAGCCTTGCTGG	65	294
	R	GGGAATTCAGCTACACAGGG		
PBX2	F	TGACCATGGCAGGGCTGGACTCAGTAAT	62.5	409
	R	CACTTCCAACCTGTCCCAGT		
PBX1	F	TGACCATGCAGGAGGGAGGGTTTCTCTC	62.5	267
	R	TCAGTGATATGAGAGAGGGCG		
PBX3	F	TGACCATGACCGAGTGTGGAAACATTGG	62.5	328
	R	TTCAATCCAGGGTGTAATCCA		
PBX4	F	TGACCATGAAGTTTGGGGGATAAGCAGG	62.5	288
	R	GAAAATCTGTGCCCAGTCCTA		
RXRB	F	TGACCATGGCCTTCCTCCTCTCAAACCT	62.5	263
	R	CTCCACCACTGGCATTTCTT		
RXRG	F	TGACCATGCGATCTAGAGGCAGATTCCTGA	62.5	231
	R	CATAGCCTGCGGGAAACTT		
RXRA	F	TGACCATGTATACTTGGATATGGCGGGG	65	299
	R	CGGAGAAGCCACTTCACAGT		
TUBB_6p21.3	F	TGACCATGACCAACCAGGTGCTGAAAAC	65	242
	R	TGGAGGGAGATTGAAAGTGG		
TUBB2_18p11.3	F	TGACCATG TTCCTTCTTGAACCCTGGTG	65	225
	R	TTTATTTTGTGGCCCCTCAG		
TUBB5_19p13.3	F	TGACCATGCTGAATCCCCTCTGACTCCA	62.5	293
	R	CCTCTCTTCCTCACAGGCAC		
TUBB4QL_10p15.3	F	TGACCATGACAGCATCTGGTTTTGCCTC	65	130
	R	CCACTGGAATGCTTGTTCCT		
TUBB4_16q24.3	F	TGACCATGCAGCTGGAGTGAGAGGCAG	65	201
	R	GCCTGGAGCTGCAATAAGAC		
TUBB1_20q13.3	F	TGACCATGTGCACTCACCATTAGCTTCG	65	396
	R	TAGTCAGGCACCTGGCTCTT		

Appendix 5

Summary of *in-silico* results.

Gene	***Brain(whole)***	***Ear***	***Eye***	***Nervous_normal***	***Heart***	***Aorta***	***Pharynx***	***Oesophagus***	***Stomach***	***Liver***	***Pancreas***	***Intestine***	***Colon***	***Gallbladder***	***Kidney***	***Bladder***	***Prostate***	***Genitourinary***	***Endometrium***	***Uterus***	***Cervix***	***Cervical carcinoma cell-line, Hela S3***	***Ovary***	***Breast***
AIF1_6p21.33	0	0	1	1	1	1	0	0	1	1	0	0	1	0	1	0	1	0	0	1	0	0	1	1
AIF1-L_9q34.12	1	0	1	1	1	0	0	0	1	1	1	0	1	0	1	0	1	0	0	1	1	0	1	1
BRD2_6p21.32	1	0	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1
BRDT_1p22.1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BRD3_9q34.2	1	0	1	1	0	1	0	0	1	0	1	1	1	0	1	0	0	0	0	1	1	0	1	1
BRD4_19p13.12	1	0	1	1	1	1	0	1	1	1	1	0	1	0	1	1	1	0	0	1	1	0	1	1
C4_6p21.33	1	0	1	1	1	0	0	1	1	1	0	0	1	1	1	1	1	0	1	0	0	0	1	1
C5_9q33.2	1	0	1	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	0	0	1	0	0	0
C3_19p13.3	1	0	1	1	1	0	0	0	1	1	1	0	1	1	1	0	1	0	1	1	1	1	1	1
CLIC1_6p21.33	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1
CLIC4_1p35.3	1	1	1	1	1	1	0	0	1	1	1	1	0	1	1	1	1	0	0	1	0	0	1	1
CLIC3_9q34.3	1	1	0	0	1	0	0	0	0	1	1	0	1	0	1	1	1	0	0	1	1	0	0	0
CLIC5_6p21.1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	1
CLIC6_21q22.12	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
CLIC2_Xq28	1	0	0	0	0	0	0	0	0	1	1	0	0	0	1	1	1	0	0	1	0	0	1	1
GPX5_6p22.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GPX4_19p13.3	1	0	1	1	1	1	0	0	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1
GPX1_3p21.31	1	0	1	1	1	1	0	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1
GPX3_5q33.1	1	0	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1
GPX2_14q23.3	1	0	0	0	0	0	0	1	1	1	1	0	1	1	1	1	1	0	0	1	0	0	1	1
NOTCH4_6p21.33	1	0	0	1	1	0	0	0	0	1	1	0	1	0	1	0	1	0	0	1	0	0	1	1
NOTCH2_1p11.2	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1
NOTCH1_9q34.3	1	0	0	1	1	0	0	1	1	0	1	1	1	0	1	0	0	0	0	1	0	0	0	1
NOTCH3_19p13.12	1	1	1	1	1	0	0	1	1	1	1	1	1	0	1	0	1	1	0	1	0	0	1	1
PBX2_6p21.33	0	0	1	1	0	0	0	0	1	1	1	1	1	0	1	0	1	0	0	1	0	1	1	0
PBX1_1q23.3	1	0	1	0	1	0	0	0	1	1	0	0	1	0	1	0	1	1	0	1	0	0	1	1
PBX3_9q33.3	1	1	1	1	1	0	0	0	1	1	1	1	1	0	1	1	1	0	0	0	0	1	1	1
PBX4_19p13.11	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
RXRB_6p21.32	1	0	1	1	1	0	0	0	1	1	1	1	1	0	1	1	1	0	1	1	1	0	1	1
RXRG_1q23.3	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
RXRA_9q34.2	1	0	1	1	1	0	0	0	1	1	1	0	1	0	1	1	1	0	0	1	1	0	1	1
TUBB_6p21.3	1	1	1	1	1	1	0	0	1	1	1	1	1	0	1	1	1	0	1	1	1	0	1	1
TUBB4_16q24.3	1	0	1	1	1	0	0	0	1	1	1	0	1	0	1	0	0	0	0	1	0	0	1	1
TUBBL_18p11.3	1	1	1	0	1	1	0	0	1	1	1	0	0	0	1	0	1	0	1	1	1	0	0	1
TUBB5_19p13.3	1	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1
TUBB1_20q13.3	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0

In-silico results continued (part 2 of 3).

Gene	***Testis***	***Epididymis***	***Placenta***	***Germ cell***	***Amnion_normal***	***Spleen***	***Thymus***	***Leukocyte***	***Lymph node***	***Lymphatic***	***Bone marrow***	***B cell***	***T cell***	***Macrophage***	***Monocyte***	***Blood***	***Nose***	***Trachea***	***Lung***	***Adrenal gland***	***Parathyroid***	***Thyroid gland***	***Pineal***	***Pituitary***
AIF1_6p21.33	1	0	1	0	0	1	1	0	0	0	1	0	0	0	0	1	0	0	1	1	1	0	0	0
AIF1-L_9q34.12	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0
BRD2_6p21.32	1	0	1	0	0	1	0	0	0	1	1	1	1	1	0	1	0	0	1	1	1	1	1	1
BRDT_1p22.1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BRD3_9q34.2	1	0	0	0	1	0	0	0	0	1	1	1	1	0	0	0	0	0	1	1	0	0	0	0
BRD4_19p13.12	1	1	1	1	0	1	1	1	0	1	1	1	1	0	0	1	1	0	1	1	1	1	0	0
C4_6p21.33	0	1	1	1	0	1	0	0	1	0	1	1	1	0	0	1	1	0	1	1	1	0	0	1
C5_9q33.2	0	0	1	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0
C3_19p13.3	1	1	1	1	0	1	0	0	1	0	1	0	1	1	0	0	1	0	1	1	1	1	0	0
CLIC1_6p21.33	1	1	1	1	1	1	1	0	1	0	1	1	1	0	0	1	0	0	1	1	1	1	0	1
CLIC4_1p35.3	1	0	1	1	1	0	0	0	1	0	1	1	1	0	0	1	0	0	1	1	1	1	1	0
CLIC3_9q34.3	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
CLIC5_6p21.1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
CLIC6_21q22.12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
CLIC2_Xq28	0	0	1	1	0	1	0	0	1	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0
GPX5_6p22.1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GPX4_19p13.3	1	1	1	1	0	0	0	1	1	0	1	1	1	0	0	1	0	0	1	1	1	0	1	0
GPX1_3p21.31	1	1	1	1	0	1	1	1	1	0	1	1	1	0	0	1	1	0	1	1	1	1	1	1
GPX3_5q33.1	1	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	1
GPX2_14q23.3	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
NOTCH4_6p21.33	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
NOTCH2_1p11.2	1	1	1	1	1	0	0	0	1	0	1	1	1	1	1	1	1	0	1	1	1	0	1	0
NOTCH1_9q34.3	1	0	1	1	0	0	0	1	1	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0
NOTCH3_19p13.12	1	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0	1	0	0	0	0	0
PBX2_6p21.33	1	1	1	1	0	0	1	1	1	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0
PBX1_1q23.3	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
PBX3_9q33.3	1	0	0	1	0	0	0	1	1	0	0	1	0	0	0	0	0	0	1	1	1	1	0	1
PBX4_19p13.11	1	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
RXRB_6p21.32	1	1	1	1	0	1	0	1	1	0	0	1	0	0	0	1	0	0	1	0	1	1	0	0
RXRG_1q23.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
RXRA_9q34.2	1	0	1	0	0	0	0	1	1	0	1	1	0	0	0	0	0	0	1	0	1	0	0	0
TUBB_6p21.3	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	1	1	0	1	1	1
TUBB4_16q24.3	1	0	1	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	1	0	0	0	1
TUBBL_18p11.3	0	0	1	1	0	1	0	0	1	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0
TUBB5_19p13.3	1	1	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0
TUBB1_20q13.3	1	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0

In-silico results continued (part 3 of 3).

Gene	***Salivary gland***	***Mammary gland***	***Skin***	***Bone***	***Adipose***	***Connective***	***Fibroblast***	***Cartilage***	***Muscle***	***Tongue***	***Synovial membrane***	***Mixed***	***Unknown***
AIF1_6p21.33	0	0	1	1	0	0	0	0	1	0	0	1	1
AIF1-L_9q34.12	0	0	1	1	0	0	0	0	1	0	0	1	1
BRD2_6p21.32	0	0	1	1	1	1	1	0	1	1	1	1	1
BRDT_1p22.1	0	0	0	0	0	0	0	0	0	0	0	1	1
BRD3_9q34.2	0	0	1	1	0	0	0	1	1	0	0	1	1
BRD4_19p13.12	1	0	0	1	1	0	1	1	1	1	0	1	1
C4_6p21.33	0	0	1	0	1	0	0	1	0	1	0	1	1
C5_9q33.2	0	0	1	0	0	0	0	1	0	1	0	1	1
C3_19p13.3	0	0	1	1	1	0	1	1	1	1	0	1	1
CLIC1_6p21.33	0	0	1	1	1	0	1	1	1	0	0	1	1
CLIC4_1p35.3	0	0	1	1	1	0	1	1	1	0	0	1	1
CLIC3_9q34.3	0	0	0	0	0	0	1	0	0	0	0	1	0
CLIC5_6p21.1	0	0	0	0	0	0	0	0	0	0	0	1	1
CLIC6_21q22.12	0	0	0	1	0	0	0	0	0	0	0	1	1
CLIC2_Xq28	0	0	0	1	0	0	0	0	1	0	0	1	0
GPX5_6p22.1	0	0	0	0	0	0	0	0	0	0	0	0	0
GPX4_19p13.3	0	0	1	1	1	0	1	1	1	0	0	1	1
GPX1_3p21.31	1	0	1	1	0	0	1	1	1	1	1	1	1
GPX3_5q33.1	1	0	1	0	1	0	0	1	1	1	1	1	1
GPX2_14q23.3	0	0	1	0	0	0	0	0	0	0	0	1	1
NOTCH4_6p21.33	0	0	1	0	0	0	0	0	0	0	0	1	1
NOTCH2_1p11.2	0	0	1	1	0	1	1	1	1	0	1	1	1
NOTCH1_9q34.3	0	0	0	0	0	0	0	0	0	0	0	1	1
NOTCH3_19p13.12	0	0	1	0	0	0	1	0	0	1	0	1	1
PBX2_6p21.33	0	0	1	0	0	0	0	0	1	0	0	1	1
PBX1_1q23.3	0	0	0	0	0	0	0	1	0	0	0	1	0
PBX3_9q33.3	0	0	0	1	0	0	1	1	0	0	0	1	1
PBX4_19p13.11	0	0	1	0	0	0	0	0	0	0	0	1	0
RXRB_6p21.32	0	0	1	0	0	0	1	1	0	0	0	1	1
RXRG_1q23.3	0	0	1	0	0	0	0	0	1	0	0	1	1
RXRA_9q34.2	0	0	1	1	0	0	1	1	1	0	0	1	1
TUBB_6p21.3	0	1	1	1	1	0	1	1	1	0	1	1	1
TUBB4_16q24.3	0	0	1	1	0	0	1	1	0	0	0	1	1
TUBBL_18p11.3	0	1	1	1	1	0	1	0	1	0	0	1	1
TUBB5_19p13.3	0	1	0	0	0	0	0	0	0	0	0	1	1
TUBB1_20q13.3	0	0	0	0	0	0	0	0	0	0	0	1	1

Appendix 6

Summary of dot blot results

Gene	*Brain*	*Cerebral cortex*	*Frontal lobe*	*Parietal lobe*	*Occipital lobe*	*Temporal lobe*	*Paracentral gyrus of cerebral cortex*	*Pons*	*Cerebellum, left*	*Cerebellum, right*	*Corpus callosum*	*Amygdala*	*Caudate nucleus*	*Hippocampus*	*Medulla oblongata*	*Putamen*	*Accumbens nucleus*	*Thalamus*	*Heart*	*Aorta*	*Atrium, left*	*Atrium, right*	*Ventricle, left*	*Ventrcle, right*
AIF1_6p21.33	0	0	0	0	1	0	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
AIF1-L_9q34.12	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BRD2_6p21.32	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BRDT_1p22.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BRD3_9q34.2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BRD4_19p13.12	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C4_6p21.33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C5_9q33.2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1
C3_19p13.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
CLIC1_6p21.33	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
CLIC4_1p35.3	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1
CLIC3_9q34.3	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
CLIC5_6p21.1	0	1	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0
CLIC6_21q22.12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CLIC2_Xq28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GPX5_6p22.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GPX4_19p13.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
GPX1_3p21.31	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
GPX3_5q33.1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
GPX2_14q23.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1
NOTCH4_6p21.33	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	1	1	1
NOTCH2_1p11.2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
NOTCH1_9q34.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
NOTCH3_19p13.12	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
PBX2_6p21.33	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
PBX1_1q23.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
PBX3_9q33.3	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
PBX4_19p13.11	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
RXRB_6p21.32	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
RXRG_1q23.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RXRA_9q34.2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
TUBB_6p21.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
TUBB4QL_10p15.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0
TUBB4_16q24.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
TUBBL_18p11.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
TUBB5_19p13.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1
TUBB1_20q13.3	0	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0

Dot blot results continued (2 of 3).

Gene	*Interventricular septum*	*Apex of heart*	*Oesophagus*	*Stomach*	*Duodenum*	*Jejunum*	*Ileum*	*Ileocecum*	*Appendix*	*Colon, ascending*	*Colon, transverse*	*Colon, descending*	*Rectum*	*Kidney*	*Skeletalmuscle*	*Spleen*	*Thymus*	*Pripheral blood leukocyte*	*Lymph node*	*Bone marrow*	*Trachea*	*Lung*	*Placenta*	*Bladder*
AIF1_6p21.33	1	1	0	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1
AIF1-L_9q34.12	1	1	1	0	0	1	0	1	1	1	0	0	1	1	0	1	0	1	1	0	1	0	1	0
BRD2_6p21.32	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BRDT_1p22.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BRD3_9q34.2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BRD4_19p13.12	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1
C4_6p21.33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
C5_9q33.2	1	0	0	1	1	1	1	1	1	0	0	0	0	1	1	1	0	0	1	1	0	0	1	1
C3_19p13.3	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1
CLIC1_6p21.33	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
CLIC4_1p35.3	1	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	1	1	0
CLIC3_9q34.3	1	1	1	1	0	1	0	0	0	0	1	0	0	0	1	0	0	1	1	0	0	1	1	1
CLIC5_6p21.1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
CLIC6_21q22.12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
CLIC2_Xq28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GPX5_6p22.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GPX4_19p13.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
GPX1_3p21.31	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
GPX3_5q33.1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
GPX2_14q23.3	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
NOTCH4_6p21.33	1	1	0	1	1	1	1	1	1	0	0	1	1	0	0	0	1	0	1	0	0	1	1	1
NOTCH2_1p11.2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
NOTCH1_9q34.3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
NOTCH3_19p13.12	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
PBX2_6p21.33	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
PBX1_1q23.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	1	1	1	1	1
PBX3_9q33.3	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
PBX4_19p13.11	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	0
RXRB_6p21.32	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
RXRG_1q23.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
RXRA_9q34.2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
TUBB_6p21.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
TUBB4QL_10p15.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TUBB4_16q24.3	1	1	1	1	1	1	1	1	1	0	0	0	0	1	0	0	0	1	1	1	1	1	1	1
TUBBL_18p11.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
TUBB5_19p13.3	1	1	0	1	1	1	1	1	1	0	1	1	1	0	0	0	1	1	1	1	1	0	0	1
TUBB1_20q13.3	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0	1	0

Dot blot results continued (3 of 3).

Gene	***Uterus***	***Prostate***	***Testis***	***Ovary***	***Liver***	***Pancreas***	***Adrenalgland***	***Thyroid gland***	***Salivary gland***	***Leukemia, HL-60***	***HeLa S3***	***Leukemia, HK-562***	***Molt4 (T cell)***	***Raji (B cell)***	***Burkitt's lymphoma, Daudi***	***Colorectal adenocarcinoma***	***Lung carcinoma, A549***	***Foetal brain***	***Foetal heart***	***Foetal kidney***	***Foetal liver***	***Foetal spleen***	***Foetal thymus***	***Foetal lung***
AIF1_6p21.33	1	1	0	0	1	1	1	1	1	0	0	0	1	0	0	0	0	0	1	1	1	1	1	1
AIF1-L_9q34.12	1	1	1	1	0	0	0	1	1	0	0	0	0	0	0	1	0	1	0	1	0	1	0	0
BRD2_6p21.32	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BRDT_1p22.1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BRD3_9q34.2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
BRD4_19p13.12	1	1	1	1	1	1	1	1	1	0	0	0	0	1	0	0	0	1	0	1	1	1	1	1
C4_6p21.33	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C5_9q33.2	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1
C3_19p13.3	1	1	1	0	1	1	1	1	1	0	1	1	1	0	0	0	0	1	1	1	1	1	1	1
CLIC1_6p21.33	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
CLIC4_1p35.3	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CLIC3_9q34.3	1	1	0	0	1	1	0	1	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	1
CLIC5_6p21.1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1
CLIC6_21q22.12	0	0	0	0	0	1	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
CLIC2_Xq28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GPX5_6p22.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GPX4_19p13.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
GPX1_3p21.31	1	1	1	1	1	1	1	1	1	0	1	0	1	1	0	1	0	1	1	1	1	1	1	1
GPX3_5q33.1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1
GPX2_14q23.3	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
NOTCH4_6p21.33	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NOTCH2_1p11.2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
NOTCH1_9q34.3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NOTCH3_19p13.12	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
PBX2_6p21.33	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1
PBX1_1q23.3	1	1	1	1	1	1	1	1	1	0	0	1	0	1	0	0	1	1	1	1	1	1	1	1
PBX3_9q33.3	1	1	1	1	1	1	1	1	1	0	1	1	0	1	0	0	0	1	1	1	1	1	1	1
PBX4_19p13.11	1	1	1	1	1	1	0	1	1	1	1	1	0	0	0	0	0	1	1	1	1	1	1	1
RXRB_6p21.32	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
RXRG_1q23.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RXRA_9q34.2	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1
TUBB_6p21.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
TUBB4QL_10p15.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
TUBB4_16q24.3	1	1	1	1	1	1	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1
TUBBL_18p11.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
TUBB5_19p13.3	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	0	1	0	1	1	1	1	0
TUBB1_20q13.3	0	0	0	0	0	0	0	1	0	1	1	1	0	0	0	0	1	0	0	0	0	0	0	0

Appendix 7

Northern blot results and transcript sizes (kb). The most dominant transcripts are in bold.

Gene	*heart*	*brain*	*placenta*	*lung*	*liver*	*skeletal muscle*	*kidney*	*pancreas*
AIF1_6p21.33	- 3.0 1.25 0.6	- 3.0 - -	- 3.0 1.25 -	- 3.0 - -	- - 1.25 0.6	5.0 3.0 - 0.6	- 3.0 - 0.6	- 3.0 - 0.6
AIF1-L_9q34.12	3.4	3.4	3.4	0	0	0	**3.4**	0
BRD2_6p21.32	**4.6** **3.8**	4.6 3.8	4.6 3.8	4.6 3.8	4.6 3.8	4.6 3.8	4.6 3.8	**4.6** **3.8**
BRDT_1p22.1	7.0 - -	7.0 - -	7.0 - -	0	7.0 - -	7.0 - -	7.0 - -	7.0 4.0 3.5
BRD3_9q34.2	6.5	6.5	6.5	6.5	6.5	**6.5**	6.5	6.5
BRD4_19p13.12	6.0 4.4	6.0 4.4	6.0 4.4	6.0 4.4	6.0 4.4	**6.0** **4.4**	6.0 4.4	6.0 4.4
C4_6p21.33	0	0	0	0	5.4	0	5.4	0
C5_9q33.2	0	0	0	0	6.0 **5.0** 4.2 1.6 1.0	0	- 5.0 - - -	- 5.0 4.2 - -
C3_19p13.3	0	0	0	0	5.0	0	0	0
CLIC1_6p21.33	1.25 1.1	0	**1.25** **1.1**	1.251.1	1.25 1.1	1.25 1.1	1.25 1.1	1.25 1.1
CLIC4_1p35.3	4.0	0	4.0	4.0	4.0	4.0	4.0	0
CLIC3_9q34.3	- 4.4 - -	0	5.5 - 2.6 **0.7**	- - - 0.7	- - - 0.7	- 4.4 - -	- 4.4 - -	- 4.4 - -
CLIC5_6p21.1	0	**2.7** **2.4**	0	2.7 -	2.7 -	0	**2.7** **2.4**	2.7 -
CLIC6_21q22.12	6.0 **3.8** 3.0 2.3	0	0	- 3.8 3.0 -	- - - 0	6.0 **3.8** - -	- - - 0	- - 3.0 -
CLIC2_Xq28	1.7 1.25	0	1.7 1.25	1.7 1.25	1.7 1.25	1.7 1.25	1.7 1.25	1.7 1.25
GPX5_6p22.1	- 0.8 0.6	- 0.8 -	- - 0.6	0	- 0.8 0.6	3.8 0.8 0.6	0	- 0.8 0.6
GPX4_19p13.3	4.4 2.6 2.0 **0.9**	- - - 0.9	- 2.6 2.0 **0.9**	- - - 0.9	4.4 2.6 2.0 **0.9**	4.4 - 2.0 **0.9**	- - 2.0 **0.9**	- - - 0.9
GPX1_3p21.31	**0.8**	**0.8**	**0.8**	**0.8**	**0.8**	**0.8**	**0.8**	**0.8**
GPX3_5q33.1	- 0.7	0	- 0.7	- 0.7	1.8 **0.7**	1.8 0.7	- 0.7	- 0.7
GPX2_14q23.3	0	0	0	0	0	0	**0.9**	0
NOTCH4_6p21.33	7.8 6.8 2.4	- - -	7.5 6.8 -	7.5 6.8 -	- - 2.4	- 6.8 2.4	- 6.8 -	- 6.8 2.4

	- 0.8	1.6 -	- 0.8	- -	- 0.8	- 0.8	- -	1.6 0.8
NOTCH2_1p11.2	**10** 1.9	10 -	**10** 1.9	10 1.9	10 1.9	**10** 1.9	**10** 1.9	10 1.9
NOTCH1_9q34.3	9.3	0	9.3	0	9.3	9.3	9.3	0
NOTCH3_19p13.12	8	8	8	0	8	8	8	8
PBX2_6p21.33	**3.2**	3.2	**3.2**	3.2	3.2	3.2	3.2	3.2
PBX1_1q23.3	7.5 - 2.9 -	7.5 - - -	7.5 - 2.9 -	7.5 - 2.9 -	7.5 - 2.9 -	**7.5** 4.4 2.9 2.6	**7.5** - 2.9 -	**7.5** 4.4 2.9 -
PBX3_9q33.3	**2.5** 2.2 1.4	2.5 - -	2.5 - -	2.5 - -	2.5 - -	2.5 2.2 -	0	**2.5** 2.2 -
PBX4_19p13.11	- 4.5 1.4 1.2	- 4.5 - -	- 4.5 1.4 1.2	0	- 4.5 1.4 1.2	6.0 4.5 1.4 -	- 4.5 - -	- 4.5 1.4 -
RXRB_6p21.32	7.7 2.8 1.7	- 2.8 -	- 2.8 -	0	- - 1.7	7.7 **2.8** 1.7	- 2.8 -	- 2.8 -
RXRG_1q23.3	1.7	0	2.8 0.9	0	1.7 1.5	**1.7**	0	1.7 1.5
RXRA_9q34.2	5.4	5.4	5.4	0	5.4	5.4	0	0
TUBB_6p21.3	- 4.8 4.0 2.5 1.7 -	- - - 2.5 **1.7** -	5.1 - - - 1.7 -	- - - - 1.7 -	- - - 1.7 -	- - 4.8 4.0 1.7 1.35	- - - - 1.7 -	- - - - 1.7 1.35
TUBB4QL_10p15.3	0	2.8	0	0	0	0	0	0
TUBB4_16q24.3	1.7 1.3	1.7 1.3	1.7 1.3	1.7 1.3	1.7 1.3	1.7 1.3	1.7 1.3	1.7 1.3
TUBBL_18p11.1	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3
TUBB5_19p13.3	0	2.3	0	0	0	0	0	2.3
TUBB1_20q13.32	0	0	0	0	0	0	0	0

Appendix 8

Summary of microarray results.

Gene	*Adrenal gland*	*Brain*	*Skeletal muscle*	*Spleen*	*Testis*	*293T (kidney)*	*Jurkat*	*Raji*	*THP1*	*U937 (lung)*
AIF1_6p21.33	0	0	0	1	0	0	1	0	0	1
AIF1-L_9q34.12	0	1	0	1	0	0	0	0	0	0
BRD2_6p21.32	1	1	1	1	1	1	1	1	1	1
BRDT_1p22.1	0	0	0	0	1	0	0	0	0	0
BRD3_9q34.2	1	1	0	0	1	0	1	1	0	0
BRD4_19p13.12	0	1	1	0	0	0	0	1	0	0
C4_6p21.33	0	0	0	0	0	0	0	0	0	0
C5_9q33.2	0	0	0	0	0	0	0	0	0	0
C3_19p13.3	0	0	0	0	0	0	0	0	0	0
CLIC1_6p21.33	1	1	1	1	1	1	1	1	1	1
CLIC4_1p35.3	0	1	1	1	1	0	1	1	0	1
CLIC3_9q34.3	0	0	1	0	0	0	0	0	0	0
CLIC5_6p21.1	0	0	1	1	0	0	0	0	0	0
CLIC6_21q22.12	0	0	0	0	0	0	0	0	0	0
CLIC2_Xq28	0	0	0	1	0	0	0	0	0	0
GPX5_6p22.1	0	0	0	0	0	0	0	0	0	0
GPX4_19p13.3	1	1	1	1	1	1	1	1	1	1
GPX1_3p21.31	1	1	1	1	1	1	1	1	1	1
GPX3_5q33.1	1	1	1	1	1	0	0	0	0	0
GPX2_14q23.3	0	0	0	0	0	0	0	0	0	0
NOTCH4_6p21.33	0	0	0	0	0	0	0	0	0	0
NOTCH2_1p11.2	0	1	0	1	1	0	0	1	0	0
NOTCH1_9q34.3	0	0	0	0	0	0	0	0	0	0
NOTCH3_19p13.12	1	1	0	1	1	0	1	0	0	0
PBX2_6p21.33	0	0	0	0	0	0	0	0	0	0
PBX1_1q23.3	1	1	0	0	1	1	0	1	0	0
PBX3_9q33.3	0	0	0	0	1	0	0	1	0	0
PBX4_19p13.11	0	1	0	0	1	0	1	1	0	0
RXRB_6p21.32	1	1	1	1	1	0	1	1	1	1
RXRG_1q23.3	0	0	1	0	0	0	0	0	0	0
RXRA_9q34.2	1	1	0	1	1	0	0	0	1	1
TUBB_6p21.3	1	1	1	1	1	1	1	1	0	1
TUBB4QL_10p15.3	0	1	1	0	1	0	0	1	0	0
TUBB4_16q24.3	0	1	0	0	1	0	0	0	0	0
TUBBL_18p11.3	1	1	1	1	1	0	0	1	1	0
TUBB5_19p13.3	1	1	0	0	1	0	0	0	0	0
TUBB1_20q13.3	0	0	0	1	0	0	0	0	0	0

Appendix 9

Comparison of three methods used to generate the expression profiles for nine MHC paralogous gene families. The differences between the three methods are highlighted in yellow for the nine tissues common to each method.

Gene	*Adrenal gland*			*Brain*			*Skeletal muscle*			*Spleen*			*Testis*			*Kidney*			*T cell*			*B cell*			*Lung*		
	M	**D**	**S**	**M**	**D**	**S**	**M**	**D**	**S**	**M**	**D**	**S**	**M**	**D**	**S**	**M**	**D**	**S**	**M**	**D**	**S**	**M**	**D**	**S**	**M**	**D**	**S**
AIF1_6p21.33	0	1	0	0	0	1	0	1	1	1	1	1	0	0	1	0	1	1	1	1	0	0	0	0	1	1	1
AIF1-L_9q34.12	0	0	1	1	1	1	0	0	1	1	1	1	0	1	0	0	1	1	0	0	0	0	0	0	0	0	1
BRD2_6p21.32	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BRDT_1p22.1	0	0	0	0	0	1	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
BRD3_9q34.2	1	1	1	1	1	1	0	1	1	0	1	0	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1
BRD4_19p13.12	0	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1	1	0	0	1	1	1	1	0	1	1
C4_6p21.33	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	1	0	0	1
C5_9q33.2	0	1	0	0	1	1	0	1	0	0	1	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	1
C3_19p13.3	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	0	0	0	1	1
CLIC1_6p21.33	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
CLIC4_1p35.3	0	0	1	1	1	1	1	1	1	1	1	0	1	0	1	0	1	1	1	0	1	1	0	1	1	1	1
CLIC3_9q34.3	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1
CLIC5_6p21.1	0	0	1	0	0	0	1	1	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0
CLIC6_21q22.12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1
CLIC2_Xq28	0	0	0	0	0	1	0	0	1	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
GPX5_6p22.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
GPX4_19p13.3	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
GPX1_3p21.31	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
GPX3_5q33.1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	0	0	0	0	0	1	1
GPX2_14q23.3	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1
NOTCH4_6p21.33	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1
NOTCH2_1p11.2	0	1	1	1	1	1	0	1	1	1	1	0	1	1	1	0	1	1	0	1	1	1	1	1	0	1	1
NOTCH1_9q34.3	0	0	0	0	1	1	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0	1	0	0	1
NOTCH3_19p13.12	1	1	0	1	1	1	0	1	0	1	1	0	1	1	1	0	1	1	1	1	0	0	1	1	0	1	1
PBX2_6p21.33	0	1	0	0	1	0	0	1	1	0	1	0	0	1	1	0	1	1	0	0	0	0	0	0	0	1	1
PBX1_1q23.3	1	1	1	1	1	1	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	1	1	0	0	1	1
PBX3_9q33.3	0	1	1	0	0	1	0	1	0	0	1	0	1	1	1	0	1	1	0	0	0	1	1	1	0	1	1
PBX4_19p13.11	0	0	0	1	1	1	0	1	0	0	0	0	1	1	1	0	1	0	1	0	0	1	0	0	0	1	1
RXRB_6p21.32	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	1
RXRG_1q23.3	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
RXRA_9q34.2	1	1	0	1	1	1	0	1	1	1	1	0	1	1	1	0	1	1	0	0	0	0	0	1	1	1	1
TUBB_6p21.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
TUBB5_19p13.3	0	0	1	1	1	1	1	0	0	0	0	0	1	0	1	0	0	1	0	0	0	1	0	1	0	0	1
TUBB4_16q24.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	1	0	1	1	0	0	1	1
TUBB2_18p11.1	1	1	0	1	1	1	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1
TUBB1_20q13.32	0	0	0	0	0	1	0	0	0	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
Total differences	**5**	**3**	**9**	**4**	**1**	**8**	**5**	**4**	**5**	**3**	**5**	**7**	**4**	**4**	**5**	**16**	**1**	**8**	**3**	**2**	**5**	**2**	**2**	**5**	**15**	**0**	**10**
% difference	**14**	**8**	**25**	**11**	**3**	**22**	**14**	**11**	**14**	**8**	**14**	**19**	**11**	**11**	**14**	**44**	**3**	**22**	**8**	**6**	**14**	**6**	**6**	**14**	**42**	**0**	**28**

Appendix 1

Summary of the annotation of the chromosomal region 9q32 to 9q34.3. Putative paralogues are in bold text.

Clone		*Ensembl Symbol*	*SPTR, Refseq or Ensembl entry*	*Transcript Start*	*Transcript end*	*Protein length (aa)*	*Description*
AL160275	**q32**	**ATP6V1G1**	**O75348**	**109082196**	**109092823**	**118**	**Vacuolar ATP synthase subunit**
	q32	NM_153045	Q8ND43	109118880	109140873	281	Unknown
AL390240	**q32**	**TNFSF15**	**O95150**	**109283763**	**109300582**	**251**	**Tumor necrosis factor ligand**
AL133412	**q32**	**TNFSF8**	**P32971**	**109398384**	**109424753**	**234**	**Tumor necrosis factor ligand**
AL162425	**q33.1**	**TNC**	**P24821**	**109514975**	**109612609**	**2201**	**Tenascin precursor**
AL355601	q33.1	NM_017418	O17418	109636267	109897093	95	Deleted in esophageal cancer 1
AL731824	q33.1	No genes					
AL714001	q33.1	No genes					
AL691420	q33.1	No genes					
AL731897	q33.1	No genes					
AL731813	q33.1	No genes					
AL732367	q33.1	EST-YD1	Q9P2X8	110398837	110399090	84	EST-YD1 protein
AL691426	q33.1	No genes					
AL353141	q33.1	No genes					
AL137024	q33.1	PAPPA	Q13219	110677328	110808083	716	Pregnancy associated plasma protein A
AL669963	q33.1	No genes					
AL133282	q33.1	ASTN2	O75129	110836698	111823883	1321	Astrotactin 1
	q33.1	Novel	ENSG00000179990	110963536	110981401	73	Unknown
AL133284	q33.1	TRIM32	Q13049	111098800	111112220	653	Zing finger protein HT2A
AL157829	q33.1	Novel	ENSG00000136913	111196159	111196227	23	Unknown
	q33.1	FLJ20958	Q9BQ00	111205755	111205826	24	Unknown
AL392085	q33.1	No genes					
AL354981	q33.1	No genes					
AL355608	q33.1	No genes					
AL358792	q33.1	No genes					
AL445644	q33.1	No genes					
AL161630	q33.1	Novel	ENSG00000179956	112057445	112065871	72	Unknown
AL160272	q33.1	TLR4	O00206	112113140	112124614	839	TOLL-like receptor 4 precursor
AL354754	q33.1	No genes					
AL445663	q33.1	No genes					
AL158831	q33.1	No genes					
AL365195	q33.1	No genes					
AL445440	q33.1	No genes					
AL355592	q33.1	No genes					
AL589703	q33.1	No genes					

AL157780	q33.1	No genes					
AL512602	q33.1	No genes					
AL445310	q33.1	No genes					
AL353773	q33.1	DCBBR1	O14618	113575169	113778264	761	Deleted in bladder cancer chromosome region
AL138894	q33.1	No genes					
AL353630	q33.1	No genes					
AC006288	q33.1	No genes					
AL445683	q33.1	No genes					
AL354931	q33.1	No genes					
AL441989	q33.1	No genes					
AL355589	q33.1	No genes					
AL592549	q33.1	No genes					
AL353736	q33.1	No genes					
AL391870	q33.2	CDK5RAP2	Q9NV90	114797706	114988994	943	CDK5 regulatory subunit associated protein 2
AL590642	q33.2	No genes					
AL138836	q33.2	EGFL5	Q9H1U4	115009649	115068409	401	EGF like domain multiple 5 protein
	q33.2	Novel	ENSG00000176341	115098762	115122988	236	Unknown
AL161911	q33.2	FBXW2	Q9UKT8	115160814	115202253	454	F-BOX/WD-repeat protein 2
	q33.2	PSMD5	Q16401	115224889	115251748	504	26S proteasome non-ATPase regulatory subunit 5
	q33.2	Novel	ENSG00000180095	115251902	115263209	105	Unknown
AL354792	q33.2	Q9UFS9	Q9UFS9	115264535	115286029	473	Transcription factor
AC006430	q33.2	PRO1995	Q9P1F7	115299108	115300075	105	Unknown
	q33.2	TRAF1	EBI6	115311227	115337603	350	TNF receptor associated factor 1
	q33.2	**C5**	**P01031**	**115361172**	**11549110**	**1676**	**Complement C5 precursor**
AL137068	q33.2	Novel	ENSG00000171635	115499108	115516745	219	Testis specfic
	q33.2	CEP1	O07018	115521379	115586444	1800	Centrosomal protein 1
	q33.2	RAB14	P35287	115586971	115610724	215	Ras-related protein
AL513122	q33.2	Novel	ENSG00000180552	115647510	115648753	409	Unknown
	q33.2	MOST2	Q9NRJ2	115689602	115694364	209	MOST2 protein
	q33.2	GSN	O06396	115708681	115741676	782	Gelosin precursor, plasma
AL161784	q33.2	EPB72	P27105	115747913	115779060	288	Erythrocyte band 7 integral membrane protein
AL359644	q33.2	Novel	ENSG00000165196	115868813	115888025	174	Unknown
AL357936	q33.2	No genes					
AL365274	q33.2	DAP2IP	Q8TDL2	115974718	116194365	964	DOC-2/DAB2 interactive protein
AL450285	q33.2	Novel					
AL596244	q33.2	Novel					
AL445587	q33.2	Novel	ENSG00000171539	116362696	116383814	140	Unknown
AL442634	q33.2	Q8NHH0	Q8NHH0	116397745	116502441	538	Unknown
AL162423	q33.2	NDUFA8	P51970	116552893	116568579	172	NADH-Ubiquinone oxidoreductase subunit
AL162424	q33.2	LHX6	Q9UPM6	116611414	11667540	363	LIM/Homeobox protein
	q33.2	NM_033117	Q96H35	116648845	11667367	190	Unknown
	q33.2	NM_138777	Q9BU92	116679683	116732299	262	RIKEN cDNA D02
	q33.2	PTGS1	P23219	116779785	116804538	596	Prostaglandin G/H synthase 1 precursor
AL359636	**q33.2**	**OR**	**Q8NGS3**	**116885796**	**116886761**	**322**	**Olfactory receptor**
		OR	**OR1J5**	**116919637**	**116920575**	**313**	**Olfactory receptor**
		OR	**Q8NGS1**	**116927976**	**116928914**	**313**	**Olfactory receptor**
		OR	**OR1N1**	**116935199**	**116936125**	**309**	**Olfactory receptor**
		OR	**Q8NGR9**	**116960247**	**116962994**	**316**	**Olfactory receptor**
		OR	**Q8NGR8**	**116976386**	**116977312**	**309**	**Olfactory receptor**
Al162254	**q33.2**	**OR**	**Q9UDD7**	**117016871**	**117017632**	**254**	**Olfactory receptor**
		OR	**OR1Q1**	**117023573**	**117024514**	**314**	**Olfactory receptor**
AC006313	**q33.2**	**OR**	**Q8NGR6**	**117037417**	**117038367**	**295**	**Olfactory receptor**
		OR	**Q8NH94**	**117070551**	**117071480**	**317**	**Olfactory receptor**
		OR	**Q8NH93**	**117083965**	**117084936**	**324**	**Olfactory receptor**
		OR	**Q8NGR5**	**117132825**	**117133757**	**311**	**Olfactory receptor**

		OR	**Q96R80**	**117158887**	**117159534**	**216**	**Olfactory receptor**
		OR	**Q8WVK7**	**117170703**	**117171065**	**121**	**Olfactory receptor**
AL359512	**q33.2**	**OR**	**Q8NGR3**	**117208958**	**117209905**	**316**	**Olfactory receptor**
	q33.2	PDCL	Q13371	117226985	117237466	301	Phosdusin-like protein
	q33.2	MNAB	O18835	117253391	117314050	1191	Membrane associated binding protein
AL731645	q33.2	ZID	Q15916	117317457	117320910	424	Zinc finger protein
	q33.2	**BIOR**	**Q9HCK0**	**117326928**	**117340094**	**441**	**Zinc finger protein**
	q33.2	GAPCenA	Q9Y3P9	117349852	117513707	997	RAB6 GTPase activating protein
AL358946	q33.3	No genes					
AL365338	q33.2	NM_030814	Q9H2N8	117518337	117522309	167	Unknown
	q33.2	STRBP	Q96S19	117533552	117593141	658	RNA binding protein
AL365504	q33.3	PRO226	Q9P180	117674928	117675167	80	Unknown
AL445489	q33.3	FLJ38464	Q8N930	117782642	117787587	215	Unknown
	q33.3	FLJ00224	Q8TEH3	117788495	118338969	896	Unknown
AL161790	q33.3	No genes					
AL390774	q33.3	No genes					
AL158208	q33.3	No genes					
AC006450	q33.3	LHX2	P50458	118420439	118441992	406	LIM/Homeobox protein
AL158052	q33.3	No genes					
Al445284	q33.3	No genes					
AL162724	q33.3	NEK6	Q9HC98	118666435	118761271	338	Serine-threonine protein kinase
AL137846	**q33.3**	**PSMB7**	**Q99436**	**118762294**	**118824271**	**277**	**Proteasome subunit beta type 7**
	q33.3	Q8NH12	Q8NH12	118874053	118892161	984	Seven transmembrane helix receptor
AL354979	q33.3	NR5A1	Q13285	118890062	118916249	461	Steroidogenic factor 1
	q33.3	NR6A1	Q15406	118928956	119180139	476	Orphan nuclear receptor
AL669818	q33.3	No genes					
AL158075	q33.3	No genes					
AL354928	q33.3	FLJ90228	Q8NCI9	119186100	119223707	318	Unknown
	q33.3	Novel	ENSG00000136918	119263034	119266399	233	Unknown
	q33.3	RPL35	P42766	119266713	119270796	141	60S ribosomal protein L35
	q33.3	NM_030978	Q9BPX5	119278120	119286561	159	Actin related protein
	q33.3	GOLGA1	Q92805	119287196	119349928	767	Golgin 97,gap junction protein
AL451125	q33.3	FLJ40705	Q8N1I4	119360938	119552351	629	Unknown
AL445930	q33.3	PPP6C	O00743	119557957	119598620	305	Serine/threonine protein phospatase 6
	q33.3	Novel	ENSG00000173602	119603477	119604650	282	40S ribosomal protein
AL354710	q33.3	P40	O00568	119609374	119642831	372	RAB9 effector P40
	q33.3	**HSPA5**	**P11021**	**119643682**	**119650159**	**654**	**78 KDA Glucose regulated protein**
	q33.3	Novel	ENSG00000176094	119672414	119672919	158	40S ribosomal protein
AL627223	q33.3	FLJ20119	Q9NXQ1	119670661	119773833	833	Unknown
AL359632	q33.3	MAPKAP1	Q9BPZ7	119846225	120116031	486	MAP kinase interacting protein 1
AL162584	q33.3	Novel	ENSG00000178022	120004502	120005454	316	Unknown
AL358074	q33.3	SIN1		119846225	120116031	522	SAPK interacting protein 1
	q33.3	NM_016158	Q9UN39	120153187	120155380	129	Erythrocyte transmembrane protein
AL627303	q33.3	No genes					
AL445186	**q33.3**	**PBX3**	**P40426**	**120156161**	**120376205**	**434**	**Pre-B-cell leukaemia transcription factor 3**
AL589923	q33.3	No genes					
AL445664	q33.3	No genes					
AL162391	q33.3	No genes					
AC006443	q33.3	FLJ00022	Q9H7P6	120735668	120915859	344	Unknown
AL356309	q33.3	No genes					
AL161908	q33.3	No genes					
AL161731	q33.3	LMX1B	O60663	121023337	121105270	379	LIM/Homeobox protein
	q33.3	Q8N243	Q8N243	121212014	121214492	115	Unknown
	q33.3	**ZNF297B**	**O43298**	**121213845**	**121244041**	**467**	**Zinc finger 297B**
AL354944	q33.3	KIAA1993	Q8NCN2	121269484	121289729	532	Unknown

AL160169	q33.3	No genes					
AL356862	q33.3	RalGPS1A	Q8WUV7	121323592	121631982	590	Ral guanine nuceotide factor
	q33.3	ANGPTL2	Q9UKU9	121496150	121531456	493	Angiopoietin-related protein 2
AL357623	q33.3	No genes					
AL450263	q33.3	Novel	ENSG00000176889	121633429	121650133	78	Unknown
	q33.3	NM_032293	Q9BQH6	121673400	121802357	802	Unknown
AL445222	q33.3	SLC2A8	Q9NY64	121806004	121816716	477	Solute carrier family 2
	q33.3	**RPL12**	**P30050**	**128564940**	**121860226**	**164**	**60S ribosomal protein L12**
	q33.3	FLJ31641	Q96MZ7	121860323	121912317	696	Unknown
	q33.3	Novel	ENSG00000176217	121901370	121904622	109	Unknown
	q33.3	NBL_HUMAN	Q96TA1	121914157	121987798	733	Niban-like protein.
AL390116	q33.3	No genes					
AL162426	q34.11	STXBP1	Q64320	122021098	122101525	594	Syntaxin binding protein 1
	q34.11	Novel	ENSG00000160401	122108029	122124506	785	Unknown
	q34.11	FLJ00176	Q8TEL7	122124875	122140409	867	Unknown
	q34.11	TOR2A	Q96LSL7	122140333	122144087	253	Torsin family 2, member A
	q34.11	SH2D3C	Q9Y2X5	122147126	122187504	703	SH2 Domain containing protein 3
AL162586	q34.11	CDK9	P50750	122194861	122198896	372	Cell division protein kinase 9
	q34.11	Novel	ENSG00000177953	122199888	122201861	85	Unknown
	q34.11	FPGS	Q05932	122211733	122222873	587	Folypolyglutamate synthase
	q34.11	FLJ33157	Q96LW6	122225388	122225972	195	Unknown
	q34.11	ENG	P17813	122224494	122263514	658	Endoglin procursor
AL157935	q34.11	AK1	P00568	122275199	122286472	194	Adenylate kinase isoenzyme 1
	q34.11	FLJ13838	Q9H8A2	122294130	122308407	352	Beta-N-Acetylgalactosaminide
	q34.11	SIAT7D	Q9H4F1	122316695	122325831	298	Sialyltransferase
	q34.11	Novel	ENSG00000167103	122330677	122339810	471	Kinase
	q34.11	Novel	ENSG00000136908	122343908	122347297	162	Unknown
	q34.11	FLJ00179	Q8TEL4	122349392	122356974	194	Unknown
	q34.11	NM_018033	Q9NW83	122374689	122375144	152	Unknown
AL360268	q34.11	Q8WU12	Q8WU12	122472894	122475975	172	Unknown
AL590708	q34.11	KIAA1896	Q96PZ1	122500273	122518054	568	Mitochondrial solute carrier
	q34.11	PTGES2	Q9H7Z7	122529502	122537271	379	Prostaglandin E synthase 2
	q34.11	Q9N1Y9	Q9N1Y9	122537454	122538008	185	Unknown
	q34.11	LCN2	P80188	122558275	122562260	192	Lipocalin
	q34.11	C9orf16	Q9BUW7	122569160	122572735	83	Unknown
	q34.11	CIZ1	Q9ULV3	122574874	122613197	967	Zinc finger protein
	q34.11	DNM1	Q05193	122612217	122664055	864	Dynamin-1
	q34.11	GOLGA2	Q08379	122665504	122684750	1008	Golgin-95
AL590722	q34.11	Q8N2W6	Q8N2W6	122685026	122697798	209	Unknown
	q34.11	FLJ21673	Q9H6Y8	122709723	122710145	423	Unknown
AL359091	q34.11	FLJ11094	O95900	122717908	122731245	331	Unknown
	q34.11	C0Q4	Q9Y3A0	122731344	122742880	265	Coenzyme Q biosynthesis protein 4
	q34.11	SLC27A4	O95186	122749454	122770025	640	Fatty acid transport protein 4
	q34.11	NM_030914	Q9BTM9	122780169	122799542	101	Unknown
	q34.11	KIAA1502	Q9P226	122820574	122846159	560	Cerebral cell adhesion molecule
	q34.11	ODF2	O14721	122864961	122909768	638	Outer dense fibre of sperm tails 2
AL445287	q34.11	GLE1L	O75458	122913513	122951096	698	Gle-1 like RNA export mediator
AL356481	q34.11	SPTAN1	Q13813	122961411	123042401	2474	Spectrin alpha chain
	q34.11	NM_052844	Q9BV46	123042469	123065595	522	Unknown
	q34.11	SET	Q01105	123092703	123105196	290	SET (HLA-DR associated protein II)
AL359678	q34.11	No genes					
AL441992	q34.11	PKNbeta	O13355	123111331	123129414	889	Protein kinase
	q34.11	ZDHHC12	O32799	123129677	123132930	267	Zinc finger protein
	q34.11	ZYG	O00156	123138594	123180701	766	ZYG homologue
	q34.11	FLJ10743	Q9NVG8	123196140	123219240	275	Unknown

	q34.11	ENDOG	Q14249	123227276	123231484	297	Endonuclease G
	q34.11	HSPC109	Q9P041	123228459	123238629	384	Unknown
	q34.11	CCBL1	Q16773	123241750	123290846	422	Cyoplasmic cysteine
AL672142	q34.11	KIAA1437	Q9P2B1	123294854	12336844	811	Unknown
	q34.11	Q96GM4	Q96GM4	123329703	123351210	206	Unknown
	q34.11	KIAA1094	Q9UPQ8	123354339	123356427	538	Unknown
AL592211	q34.11	KIAA0169	Q14675	123357991	123415903	1739	Unknown
	q34.11	SH3GLB2	Q9BRZ5	123416600	123437108	130	SH3-containing protein
	q34.11	FLJ00199	Q9TEJ6	123445452	123480880	383	Unknown
	q34.11	Q96GF8	Q96GF8	123489937	123499243	237	Unknown
AL158151	q34.11	CRAT	P43155	123503602	123519612	626	Cartinine o-acetyltransferase
	q34.11	PPP2R4	Q15257	123519773	123557754	358	Protein phosphatase 2A
	q34.11	Novel	ENSG00000167133	123585649	123586860	290	Unknown
AL161785	q34.11	FLJ35269	Q8NAJ2	123729824	123733713	232	Unknown
AL353803	q34.11	FLJ34873	Q8NAS2	123745332	123747103	144	Unknown
AL391056	q34.11	Novel	ENSG00000179068	123897948	123913240	98	Unknown
	q34.11	FLJ35803	Q8NA65	124021035	124029584	377	Unknown
AL590369	q34.11	AD003	Q9UI28	124034981	124044744	223	Adrenal gland protein
	q34.11	ASB6	Q9NWX5	124043412	124050973	421	Ankyrin repeat containing protein
	q34.11	PMX2	Q99811	12074444	124131482	253	Paired mesoderm protein
	q34.11	PTGES	O14684	124147139	124161855	152	Prostaglandin E synthase
AL592219	q34.11	No genes					
AL158207	q34.11	TOR1B	O14657	124211961	124220092	336	Torsin B precursor
	q34.11	DYT1	Q96CA0	124221751	124232942	336	Torsin A precursor
	q34.11	HSPC220	Q9NZ63	124236100	124244083	289	Unknown
	q34.11	USP20	Q9Y2K6	124244254	12490636	914	Ubiquitin carboxyl-terminal hydrolase
	q34.11	FNBP1	Q96RU3	124295995	124451976	672	Thyroid receptor interacting protein
AL136141	q34.11	GPR107	Q96T26	124462729	124548972	416	G Protein-coupled receptor
AL392105	q34.11	No genes					
AL360004	q34.11	FREQ	P36610	124581381	124645435	190	Neuronal calcium sensor 1
	q34.11	Novel	ENSG00000178890	124674690	124718869	822	Unknown
50 kb Gap							
AL354898	q34.11	Q8NDA2	Q8NDA2	124808356	124841498	1187	Unknown
	q34.11	FLJ23816	Q8TCI8	124852198	124856039	220	Unknown
	q34.11	ASS	P00966	124866845	124923190	412	Argininosuccinate synthase
AL353695	q34.11	No genes					
AL359092	q34.11	FUBP3	Q92946	125001544	125060268	542	Fuse binding protein 3
	q34.12	PRDM12	Q9H4Q4	125086510	125104913	367	PR domain containing protein 12
	q34.12	RRP4	Q13868	125115687	125126785	293	Exosome complex exonuclease RRP4
AL161733	q34.12	ABL1	P00519	125136236	125309589	1130	Abelson murine leukaemia viral oncogene
	q34.12	FLJ14810	Q96SJ7	125324358	125360767	198	Unknown
AL583807	q34.12	LAMC3	Q9Y6N6	125431028	125516389	1575	Laminin gamma-3 chain precursor
AL355872	q34.12	No genes					
AL157938	**q34.12**	**AIF1L**	**Q9BQI0**	**125518441**	**125545061**	**150**	**Ionised cacium binding adaptor molecule 2**
	q34.13	NUP214	P35658	125547506	125656586	2140	Nuclear pore complex protein
	q34.13	Q8N2W3	Q8N2W3	125679994	125698463	191	Unknown
AL354855	q34.13	FLJ90726	Q8NBV4	125711653	12731177	271	Unknown
	q34.13	Novel	ENSG00000130710	125729227	125729298	24	Unknown
AL358781	**q34.13**	**BAT2L**	**Q9BU62**	**125852061**	**125869120**	**325**	**HLA-B associated transcript**
	q34.13	LQFBS-1	O95209	125921328	125922066	245	Unknown
	q34.13	POMT1	Q9UNT2	125924841	125945722	747	Protein-o-mannosyltransferase 1
	q34.13	UCK1_HUMAN	Q9HA47	125945717	125953181	201	Uridine cytidine kinase 1
AL160276	q34.13	GRF2	Q13905	126000707	126159454	1077	Guanine nucleotide releasing factor 2
AL160271	q34.13	CRSP8	O95401	126282028	126512112	273	Cofactor required transcriptional activation
AL603649	q34.13	No genes					

AL713892	q34.13	No genes					
AL691506	q34.13	No genes					
AL513102	q34.13	No genes					
AL353631	q34.13	No genes					
AL159997	q34.13	KIAA1857	Q96JH0	126594193	126675069	541	Netrin G2
	q34.13	KIAA0625	Q8WX33	126693686	126761056	915	Unknown
AL353701	q34.13	TTF1	Q15361	126808230	126835074	882	Transcription termination factor
AL354735	q34.13	Novel	ENSG00000178595	126930940	126975292	179	Unknown
	q34.13	BARHL1	Q9BZE3	127014851	127022519	327	BARH (Drosophila)-like 1
AL160165	q34.13	DDX31	Q96NY2	127026534	127102646	851	DEAD/H Box Helicase
	q34.13	GTF3C4	Q9UKN8	127102586	127122695	822	General transcription factor
AL445645	q34.13	FLJ32704	Q96MA6	127157823	127310564	479	Unknown
	q34.13	C9orf9	Q96E40	127310608	127322275	222	Unknown
	q34.13	TSC1	Q92574	127323595	127376866	1164	Tuberous sclerosis 1 gene
	q34.13	Novel	ENSG00000176140	127379314	127383801	47	Unknown
AL593851	q34.13	GFI1B	O95270	127418923	127426295	330	Growth factor independent 1B
AL162417	q34.2	GTF3C5	Q9H4P2	127462958	12790748	528	General transcription factor
	q34.2	CEL	P19835	127494229	127504006	756	Carboxyl ester lipase
	q34.2	NM_173692		127513557	127514144	196	Unknown
	q34.2	CELL	Q14018	127514780	127519600	59	Carboxyl ester lipase-like
	q34.2	**RALGDS**	**Q12967**	**127529965**	**127553410**	**914**	**Ral Guanine nucleotide**
	q34.2	FRS	Q9UKI5	127585198	127596144	347	Forssman synthetase
AL732364	q34.2	OBPIIB	Q9NPH6	127637537	127641486	170	Odorant binding protein 2B
AL158826	q34.2	ABO	P16442	127687850	127694413	287	ABO blood group system
	q34.2	SURF6	O75683	127754393	127759885	361	SURFEIT locus protein 6
	q34.2	SURF5	Q15528	127764596	127771813	200	SURFEIT locus protein 5
	q34.2	SURF3	P11518	127771906	127775122	265	SURFEIT locus protein 3
	q34.2	Q9H3B2	Q9H3B2	127774377	127775089	101	Unknown
	q34.2	SURF1	Q15526	127775504	127780202	300	SURFEIT locus protein 1
	q34.2	SURF2	Q15527	127780269	127784875	256	SURFEIT locus protein 2
	q34.2	SURF4	O15260	127785181	127799817	269	SURFEIT locus protein 4
	q34.2	Q8NE28	Q8NE28	127800125	127828061	651	Unknown
	q34.2	Novel	ENSG00000175977	127821374	127824619	97	Unknown
	q34.2	XPMC2H	Q9GZR2	127828027	127840010	422	Prevents mitotic catastrophe 2
	q34.2	ADAMTS13	Q96L37	127843961	127881349	1427	Von Willebrand factor-cleaving protease
AL593848	q34.2	C9orf7	Q9UGQ2	127881962	127892726	172	Unknown
	q34.2	SLC2A6	Q8NCC2	127893058	127901068	515	Solute carrier family 2
BX324209	q34.2	No genes					
AC002321	q34.2	No genes					
<5 kb Gap							
AC002101	q34.2	No genes					
AL365494	q34.2	DBH	P09172	128007020	128030001	603	Dopamine beta-monooxygenase precursor
	q34.2	SARDH	Q9UL10	128056341	128124046	832	Sarcosine dehydrogenase
	q34.2	PP3781	Q8WY83	123131956	128132329	124	Unknown
AL590710	q34.2	Novel	ENSG00000176983	128211430	128251481	432	Unknown
	q34.2	SARDH	Q9UL10	128252348	128272381	396	Unknown
AL357934	q34.2	VAV2	P52735	128297977	12826304	878	Oncogene VAV-2 protein
AL445931	q34.2	Novel	ENSG00000179483	128559910	128562122	119	Unknown
	q34.2	**BRD3**	**Q15059**	**128566862**	**128602533**	**726**	**Bromodomain containing protein3**
	q34.2	Novel	ENSG00000179457	128588311	128592826	216	Unknown
AL591386	q34.2	No genes					
200kb Gap							
AL354796	q34.2	No genes					
AL683798	q34.2	No genes					
13 kb Gap							

AL669970	**q34.2**	**RXRA**	**P19793**	**129062693**	**129101647**	**453**	**Retinoid X receptor, alpha**
AL591890	**q34.3**	**COL5A1**	**Q96HC0**	**129302868**	**129503955**	**590**	**Collagen alpha 1 (V) chain precursor**
AL603650	q34.3	FCN2	Q15485	129541874	129548582	313	Ficolin 2 precursor
AL353611	q34.3	FCN1	O00602	129570647	129579025	326	Ficolin 1 precursor
AL159992	q34.3	No genes					
AL390778	q34.3	OLFM1	Q9BWJ9	129736484	129782241	467	Olfactomedin related ER localised protein
AL353615	q34.3	NM_173520	Q8N4C0	130006614	130009927	152	Unknown
AL161452	q34.3	Novel	ENSG00000178197	130131516	130144083	96	Unknown
	q34.3	NM_014811	Q9Y4D3	130146060	130152258	1209	Unknown
	q34.3	NM_144654	Q8WU44	130158546	130165104	92	Unknown
	q34.3	MRPS2	Q9Y399	130164060	130168038	296	Mitochondrial ribosomal protein S2
	q34.3	LCN1	P31025	130184820	130189897	176	Lipocalin 1
	q34.3	OBPIIA	Q9NY56	130209504	130213321	170	Odorant-binding protein 2A
AL354761	q34.3	PAEP	P09466	130225123	130230141	157	Progestagen associated endometrial
	q34.3	Novel	ENSG00000176541	130238316	130250477	104	Unknown
AL158822	q34.3	MUPL	Q8WX39	130326687	130329116	172	Putative MUP-like lipocalin
	q34.3	Q8NEE3	Q8NEE3	130356772	130362893	348	Unknown
	q34.3	KCNT1	Q9WX41	130365551	130455523	1151	Unknown
AL353636	q34.3	NM_018627	Q9WX42	130472946	130570503	1298	Unknown
AL355574	q34.3	GPDR1	Q9BSL1	130596334	130624745	405	Glialblastoma related protein
	q34.3	NM_144653	Q96BF6	130674721	130713948	587	Unknown
AL591038	q34.3	Novel	ENSG00000180858	130715097	130738981	273	Unknown
AL138781	q34.3	Q96GU2	Q96GU2	130777953	130782228	29	Unknown
	q34.3	Q8N3G2	Q8N3G2	130869698	130887479	541	Unknown
	q34.3	LHX3	Q9UBR4	130859621	130868480	397	LIM Homeobox gene 3
30 kb Gap							
AL603784	q34.3	AGS3	Q9UFS8	130939617	130943166	530	Unknown
AL592301	q34.3	CARD9	Q9H257	130947895	130957602	536	Caspase recruitment protein
	q34.3	SNAPC4	Q9Y6P7	130959516	130982736	1469	Small nuclear RNA activating complex
	q34.3	SDCCAG3	O60525	130985862	130994412	192	Serologically defined colon cancer antigen
	q34.3	INPP5E	Q10713	130994603	131007700	525	Mitochondrial processing peptidase subunit
	q34.3	PPI5PIV	Q9NRR6	131012558	131023761	644	Phosphatidylinositol (4,5) bisphosphate 5-phophatase
	q34.3	KIAA0310	Q96HP1	131024036	131059908	1433	Unknown
	q34.3	NM_152571	Q8N9P6	131067482	131070005	203	Unknown
	q34.3	**NOTCH1**	**P46531**	**131078383**	**131129726**	**2559**	**Neurogenic locus NOTCH homologue protein**
AL590226	q34.3	Novel	ENSG00000180360	131211094	131230721	251	Unknown
	q34.3	Q9P058	Q9P058	131232664	131244366	146	Unknown
	q34.3	ZNEU1	Q9UHF1	131242795	131256617	273	ZNEU1/NEU1 protein
	q34.3	**AGPAT2**	**O15120**	**131257082**	**131271362**	**278**	**Acylglycerol-phophate-acyltransferase 2**
	q34.3	NM_152421	Q8WYU5	131296511	131307989	431	Unknown
AL355987	q34.3	NM_032887	Q96IC0	131309173	131312123	37	Unknown
	q34.3	Novel	ENSG00000169672	131313093	131332422	726	Unknown
	q34.3	FLJ33328	Q8NBE9	131338327	131341477	333	Unknown, has IG_MHC domain
	q34.3	FLJ10101	Q96BU21	131383198	131425126	307	Unknown
	q34.3	FLJ30985	Q96NE7	131387866	131392787	197	Unknown
	q34.3	Nov-01	ENSG00000054148	131433033	131434977	186	Unknown
	q34.3	Novel	ENSG00000148406	131436356	131438789	350	Unknown
	q34.3	Novel	ENSG00000179285	131439193	131440765	264	Unknown
	q34.3	Q8NCX7	Q8NCX7	131440944	131444739	236	Unknown
	q34.3	EDF1	O60869	131446058	131450225	148	Endothelial differentiation-related factor 1
AL449425	q34.3	TRAF2	TRA2_HUMAN	131482651	131510546	501	TNF receptor associated factor 2
AL807752	q34.3	NM_018998	Q969U6	131524374	131528545	566	Unknown
	q34.3	C8G	P07360	131529200	131530906	198	Complement component 8, gamma subunit
	q34.3	PTGDS	P41222	131561509	131565680	190	Prostaglandin D2 synthase
	q34.3	Novel	ENSG00000176785	131567941	131570349	137	Unknown

	q34.3	**CLIC3**	**O95833**	**131578574**	**131580807**	**207**	**Chloride intracellular channel protein 3**
	q34.3	**ABCA2**	**Q9BZC7**	**131591173**	**131612250**	**2440**	**ATP-binding cassette, subfamily A, member 2**
	q34.3	Q9BUH6	Q9BUH6	131576357	131577915	212	Unknown
	q34.3	FUT7	Q11130	131614669	131615685	339	Fucosyltransferase 7
AL929554	q34.3	Q8N224	Q8N224	131618602	131619006	135	Unknown
	q34.3	NPDC1	Q9NQX5	131623413	131630157	325	Neuronal proliferation protein 1
	q34.3	ENTPD2	Q9Y5L3	131632616	131638290	494	Ectonucleoside triphosphate diphosphydrolase 2
	q34.3	Q8TEI1	Q8TEI1	131662043	131670905	270	Unknown
	q34.3	Q8WUC7	Q8WUC7	131669250	131671114	104	Unknown
	q34.3	MAN1B1	Q9UKM7	131671245	131695076	699	Endoplasmic reticulum mannosidase
	q34.3	Novel	ENSG00000179395	131687948	131691411	989	Unknown
	q34.3	**DPP7**	**Q9UHL4**	**131696546**	**131699393**	**331**	**Dipeptidyl-peptidase**
	q34.3	GRIN1	Q05586	131724505	131753424	928	Glutamate receptor subunit zeta 1
	q34.3	NM_013366	Q9UJX6	131759296	131773049	822	Anaphase-promoting complex subunit 2
	q34.3	SSNA1	O43805	131773159	131774882	119	Sjorgen's syndrome nuclear autoantigen 1
	q34.3	FLJ90254	Q8NCH2	131776627	131784420	433	Unknown
	q34.3	NM_053045	Q969S6	131788473	131790029	136	Unknown
BX255925	q34.3	Unfinished					
BX322799	q34.3	Unfinished					
AL365502	q34.3	AD038	Q96F01	132035782	132039264	205	AD038 protein, function unknown
	q34.3	NM_152285	Q8N5I2	132039236	132048941	433	Unknown
	q34.3	MZIP	Q96E35	132064147	132072376	227	Melanin-concentrating hormone receptor 1
	q34.3	NM_138778	Q9BTV6	132075663	132096389	484	Unknown
	q34.3	MRPL41	NM_032477	132098739	132099406	137	Mitochondrial ribosomal protein L41
	q34.3	NTE-L	Q8TAY5	132153193	132190541	702	Neuropathy Target Esterase
	q34.3	FLJ14568		132192950	132200987	327	Unknown
	q34.3	Q9NTU2		13295160	132200984	130	Unknown
AL590627	q34.3	Novel	ENSG00000181090	132289873	132374235	436	Unknown
AL611925	**q34.3**	**HMT1**	**Q9H9B1**	**132395728**	**132553855**	**1247**	**Histone methyltransferase**
AL772363	q34.3	CACNA1B	Q00975	132562084	132806329	2357	Calcium channel voltage-dependent
AL591424	q34.3	IL9R	ENSG00000165830	132821281	132832099	216	Interleukin 9 receptor IL 9R
	q34.3	**Novel**	**ENSG00000159247**	**132859209**	**132861370**	**425**	**Tubulin pseudogene**
	q34.3	Novel	ENSG00000179338	132868297	132868767	157	LINE 1 Reverse Transcriptase Homologue
AL954642	q34.3	No genes					

Appendix 2

Table of results the whole-genome survey. The P-values are coloured according to the level of confidence; black are L0-paralogues, green L1-paralogues, blue L2-paralogues and red are L3-paralogues.

Class	*MHC gene*	*Clone*	*Locus*	*Start*	*End*	*BLAST Match*	*P-value*
III	NOTCH4	AL390719.31.31331.88824	1p36.33	708136	744003	AGRN	1.50E-23
xII	B3GALT4	AL162741.35.1.111409	1p26.33	919626	921479	B3GALT6	1.40E-08
III	NOTCH4	AL391244.11.1.67923	1p36.33	1073552	1087162	NM_030937	9.10E-09
III	BAT8	AL391244.11.1.67923	1p36.33	1089751	1108843	Novel	7.60E-12
III	NOTCH4	AL512413.21.1.101803	1p36.32	2844459	2886399	EGFL3	2.20E-42
xI	BTN1A1	AL662907.11.1.64693	1p35.1	3343046	33466609	Q9BVG3	1.10E-18
III	NOTCH4	AL513320.27.1.132592	1p36.32	3364015	3405344	EGFL3	2.00E-25
III	C2	AL109811.40.1.112769	1p36.22	10930224	10950939	MASP2	1.30E-05
xI	BTN1A1	AC074003.3.23107.35032	1p36.13	14652770	14656331	Novel	4.20E-14
III	C6orf46/ZNF297	AL034555.2.1.86897	1p36.13	15447174	15481304	ZNF151	2.40E-14
I	DDR1	AL451042.10.1.88098	1p36.13	15564788	15596501	EPHA2	1.10E-17
xII	KIFC1	AL663074.13.1.8581	1p36.12	20071643	20079710	Novel	2.60E-24
III	NOTCH4	AL590103.12.1.175162	1p36.12	21183949	21258020	HSPG2	1.10E-16
I	DDR1	AL035703.21.1.160705	1p36.12	21925207	21965283	EPHA8	1.40E-15
I	DDR1	AL035704.9.1.113956	1p36.12	22143606	22277501	EPHB2	3.10E-20
xII	LYPLA2L	AL031295.1.1.124001	1p36.11	23188619	23193014	LYPLA2	9.00E-75
III	CLIC1	AL662924.15.1.121762	1p35.3	24141534	24239539	CLIC4	3.80E-27
I	DDR1	AL031729.16.1.125287	1p36.11	26912974	26924738	FGR	3.20E-15
xII	COL11A2	AC114488.1.90406.184673	1p35.2	31102432	31154233	COL16A1	2.10E-11
xII	ZNF297	AL033529.25.1.147167	1p35.1	31870816	32006551	NM_144621	1.00E-13
xI	RFP	AL662907.11.1.64693	1p35.1	32565928	32602196	NM_018207	7.60E-46
III	C2	AC115285.1.63883.124348	1p35.1	33035623	33094251	Q96Q03	9.80E-08
III	BF	AC115285.1.63883.124348	1p35.1	33799053	33862505	Q9H4W4	9.50E-08
I	POU5F1	AL139158.11.1.115614	1p34.3	37518563	37519168	no gene	4.50E-44
III	HSPA1L	AL354702.7.1.107422	1p34.3	38182110	38184069	Novel	4.30E-158
xII	ZNF297	AL356379.10.1.64960	1p34.2	40000812	40017418	NM_152373	7.80E-09
xI	RFP/MOG/BTNL2	AL512353.16.1.81704	1p34.2	42286363	42314156	ERMAP	6.60E-49
I	DDR1	AC093420.1.127596.194462	1p34.2	42767145	42789215	TIE	4.80E-05
II	BTNL2	AL109659.20.1.181678	1p33	47521181	181364180	genscan	1.10E-24
xI	BTN1A1	AL109659.20.1.181678	1p33	48243776	48288207	no gene	3.40E-29
xI	MOG	AL109659.20.1.181678	1p33	48253780	48253842	genscan	4.00E-22
xI	GPX5	AL356976.30.1.64323	1p32.3	51957064	51963742	NM_015696	8.30E-09
I	TUBB	AL445183.19.1.193774	1p32.3	53049002	53152605	SCP2	8.90E-12
I	DDR1	AL445205.14.1.115936	1p31.3	63525514	63564266	EST gene	3.10E-45
I	DDR1	AC093427.2.1.131877	1p31.3	64219769	64351748	JAK1	2.50E-09
III	C6orf29	AC107627.2.1.90513	1p31.1	74605198	75011843	NM_152697	8.70E-49
III	MSH5	AL445464.9.1.103097	1p31.1	75197713	75313965	MSH4	4.30E-05
III	DDAH2	AL078459.8.1.83946	1p22.3	84926028	85072691	DDAH1	3.50E-15
xII	COL11A2	AL356059.27.1.76418	1p22.3	85351949	85791154	NM_152890	1.90E-05
II	BRD2	AC004798.1.1.42497	1p22.1	91625907	91677611	BRDT	6.50E-101
III	C9orf29	AC093429.2.1.182165	1p21.3	94491811	94566684	NM_152369	1.30E-08

xII	COL11A2	AC093150.2.1.189945	1p21.1	104172486	104410718	COL11A1	4.10E-42
III	NOTCH4	AL390252.9.1.169241	1p13.3	108785784	108869934	SORT1	6.40E-22
III	BAT1	AL445483.13.1.164008	1p13.2	112714422	112726367	DDX20	3.00E-14
xI	MOG/BTN1A1	AL391476.20.1.171595	1p13.1	118145476	118212809	NM_024626	9.30E-05
III	NOTCH4	AL359752.11.1.137955	1p11.2	119292037	119450143	NOTCH2	2.30E-61
III	NOTCH4	AL592307.24.14836.157830	1q21.1	141809266	141890630	Novel	3.20E-27
III	NOTCH4	AC018381.3.23653.71017	1q21.1	141956297	141968135	Novel	3.70E-28
xI	HIST1H2AC	AL591493.13.1.113370	1q21.2	145553544	145555199	Histones	3.30E-48
III	BAT8	AL590133.32.1.192096	1q21.3	146646159	146684482	SETDB1	1.90E-12
xII	ZNF297	AL451085.20.1.182166	1q22	150706074	150720422	ZFP67	5.40E-14
I	POU5F1	AL139410.20.1.166288	1q22	151143602	151144684	Q9BZW0	4.10E-105
III	NOTCH4	AL158169.17.1.99802	1q23.1	152575891	152593889	INSRR	6.10E-12
I	DDR1	AL158169.17.1.99802	1q23.1	152595949	152616657	NTRK1	1.70E-30
xI	HFE	AL138899.23.1.134137	1q23.1	153882029	153886978	CD1D	2.50E-05
III	NOTCH4	AL356104.6.1.96693	1q23.1	154745243	154755113	Q8TEK2	4.20E-22
III	HSPA1L	AL590385.22.1.110781	1q23.3	157226591	157228843	HSPA6	6.20E-263
III	CREBL1	AL391825.15.1.211662	1q23.3	157386942	157579736	ATF6	4.30E-25
I	DDR1	AL445197.4.1.117040	1q23.3	158253102	158401084	DDR2	1.50E-131
III	PBX2	AL357568.14.1.71359	1q23.3	160179905	160302546	PBX1	3.40E-88
I	POU5F1	AL136984.20.1.169627	1q24.2	162840972	163036179	POU2F1	5.70E-19
xII	RXRB	AL160058.8.1.155369	1q23.3	163005239	163049202	RXRG	3.00E-63
xII	RPS18	AL031733.3.1.215861	1q24.2	163211499	163211834	no gene	7.20E-14
III	BAT2	AL021579.1.1.99886	1q24.3	167083736	167191695	BAT2-ISO	1.60E-63
III	TNF	Z96050.1.1.85811	1q24.3	168257203	168265061	TNFSF6	5.90E-06
III	TNXB	Z94055.1.1.134539	1q25.1	170720587	170804587	TNR	6.00E-07
III	C6orf46	AL136170.12.1.127541	1q25.1	171377563	171488882	NM_032522	2.10E-33
I	DDR1	AL139132.16.1.157866	1q25.2	174505391	174627374	ABL2	7.90E-12
I	HLA-A/HLA-E	AL162431.17.1.139006	1q25.3	176341488	176402911	STX6	1.60E-20
xI	HLA Class I and II	AL356267.27.1.181808	1q25.3	176431452	176452972	HLALS	3.50E-31
III	BAT1	AL049557.19.1.128379	1q25.2	176601093	176723077	ABL2	3.40E-10
III	BAT8	AL138776.10.1.100549	1q25.3	177972880	177984295	RNASEL	3.30E-05
xII	RAB2L	AL590422.14.1.198210	1q25.3	179033557	179326003	RGL1	8.30E-17
xII	RING1	AL109865.36.1.201823	1q25.3	180442880	180499976	RNF2	4.20E-50
I	DHX16	AL355999.9.1.76504	1q31.1	185516454	185517347	genscan	1.10E-07
xII	B3GALT4	AL390863.9.1.122864	1q31.2	188662474	188670039	B3GALT2	2.20E-21
III	NOTCH4	AL513325.13.1.212888	1q31.3	192695849	192906025	CRB1	4.20E-52
III	ATP6V1G2	AL157402.19.1.210331	1q31.3	193950307	193968515	ATP6V1G3	2.40E-09
xII	RXRB	AC096633.2.1.178152	1q32.1	195455200	195604973	NR5A2	1.50E-15
xII	KIFC1	AL445483.13.1.164008	1q32.1	195979729	196048406	KIF14	7.50E-10
III	BAT1	AL512326.24.1.189269	1q32.1	198323462	198369154	NM_031306	4.00E-06
xII	RPS18	AL606462.5.1.112401	1q42.13	223356887	223357135	genscan	1.20E-40
xI	RFP	AL139288.15.1.151563	1q42.13	224319970	224333114	TRIM11	2.10E-64
xI	HIST1H2AC	AL139288.15.1.151563	1q42.13	224383273	224384153	H2AFL	4.80E-49
xI	BTN1A1	AL139288.15.1.151563	1q42.13	224436665	224436976	genscan	4.50E-64
II	BTNL2	AL139288.15.1.151563	1q42.13	224436716	224436991	genscan	3.70E-19
II	TAP2/1	AL121990.33.1.147913	1q42.13	225346509	225388622	ABCB10	5.10E-24
III	PBX2	AL359255.12.1.20809	1q42.13	226893037	227045628	OBSCN	5.40E-11
xI	BTN1A1	AC026657.4.97959.109520	1q42.13	227053394	227055850	TRIM11	1.20E-16
xI	MOG	AL139288.15.1.151563	1q42.13	227195475	227197055	Novel	9.50E-16
xI	RFP	AL591686.9.1.150680	1q43	237498626	237934567	NM_152666	2.40E-11
III	C6orf46/ZNF297	AL590483.25.118180.187060	1q44	239460483	239466676	ZNF238	9.20E-16
III	HSPA1L	AL390728.34.1.206255	1q44	242630342	242630950	no gene	5.80E-218
xI	RFP	AC099571.1.86529.165648	1q44	243256848	243277448	NM_015431	1.50E-65
xI	BTN1A1	AC099571.1.86529.165648	1q44	245957656	245974557	Q9Y4N9	3.50E-37
III	NOTCH4	AC105450.1.1.163782	2p25.3	1491150	1620306	TP0	2.10E-07
xII	KIFC1	AC013449.8.1.120997	2q23.3	26242053	26297524	KIF3C	3.30E-37

III	BAT1	AL121658.4.1.162692	2p22.3	32334563	32390404	Q96NC3	6.30E-05
xI	HIST1H2AC/NOTCH4	AL133244.1.1.200368	2p22.3	33311500	33772726	LTBP1	6.50E-18
III	CYP21A2	AC009229.5.1.209156	2p22.2	38259590	38268136	CYP1B1	9.30E-15
I	DHX16	AC092833.4.1.143506	2p22.1	39171874	39237365	NM_145646	5.30E-10
I	DHX16	AC018693.8.1.164125	2p22.1	39282730	39282825	no gene	1.00E-09
xI	NEFAL	AC016722.9.1.149995	2p21	47297756	47311689	NM_139279	3.80E-32
III	MSH5	AC009600.19.1.215260	2p21	47798999	47879105	MSH2	5.00E-09
xI	UBD	AC079807.5.1.156175	2p16.3	48126446	48126718	genscan	5.80E-05
III	MSH5	AC006509.15.1.124015	2p16.3	48179027	48202837	MSH6	7.80E-08
xII	B3GALT4	AC093401.4.1.99088	2p15	62609435	62638037	B3GNT1	2.30E-13
I	DHX16	AC005041.2.1.191356	2p13.1	74958046	74966106	NM_133637	1.30E-10
III	BAT2	AC068279.6.1.135351	2p11.2	87959386	87959877	no gene	3.10E-05
III	BAT2	AC026106.12.36729.68570	2p11.2	90750748	90751398	no gene	6.50E-05
xII	ZNF297	AC092835.4.1.158404	2q-tel	94264885	94283773	ZNF2	3.70E-10
xI	RFP	AC018892.8.1.191055	2q11.2	96145683	96151491	Novel	1.60E-21
I	DDR1	AC016699.10.1.54480	2q11.2	96793061	96808883	ZAP70	3.30E-15
I	POU5F1	AC018730.7.1.154728	2q12.1	103925142	103926146	POU3F3	2.60E-44
xI	BTN1A1	AC005040.2.1.189949	2q12.3	106059831	106060295	genscan	3.40E-08
xII	ZNF297	AC013268.5.1.206457	2q13	109008786	109017957	NM_152518	4.40E-13
I	C6ORF18	AC018737.9.1.206454	2q14.3	120015449	120327116	CLASP1	9.70E-06
xII	RPS18	AC018737.9.1.206454	2q14.3	120421203	120421337	no gene	9.80E-10
I	TUBB	AC018804.9.1.195514	2q21.1	128250367	128259530	genscan	3.00E-08
I	TUBB	AC073869.5.1.195280	2q21.2	129791442	129796126	TUBA2	1.80E-10
I	MRPS18B	AC012497.8.1.212104	2q22.1	138733844	138734416	genscan	1.20E-75
xII	RXRB	AC074099.6.1.143653	2q24.1	155709873	155718141	NR4A2	1.00E-06
xII	B3GALT4	AC016723.11.1.202001	2q24.3	167216857	167269041	B3GALT1	8.10E-30
II	TAP2/1	AC069137.6.1.108836	2q24.3	168321109	168292498	ABCB11	6.90E-21
xII	COL11A2	AC066694.7.1.120381	2q32.2	187975163	188013419	COL3A1	5.40E-32
III	HSPA1L	AC013409.8.1.195478	2q34	208662170	208666396	Novel	1.00E-116
I	TUBB	AC068946.4.1.172260	2q35	218133730	218152718	TUBA4	2.00E-10
I	DDR1	AC010899.8.1.210232	2q36.1	221011996	221158367	EPHA4	9.20E-15
III	HSPA1L	AC009302.2.1.180970	2q36.1	221548101	221549027	genscan	7.00E-113
xI	BTN1A1	AC104772.3.1.106526	2q36.1	221566190	222241888	SYFB	1.40E-19
xI	RFP	AC104772.3.1.106526	2q36.1	222156619	222241888	SYFB	4.40E-44
xI	PRSS16	AC008072.3.1.206177	2q36.1	223964489	223987825	NM_024785	4.20E-10
xII	COL11A2	AC073869.5.1.195280	2q36.3	226590998	226750349	COL4A4	3.50E-07
xII	COL11A2	AC097662.4.37779.206758	2q36.3	226750355	226900581	COL4A3	8.90E-12
III	NOTCH4	AC008273.2.1.151297	2q36.3	228954450	229310970	NM_139072	3.30E-23
xII	B3GALT4	AC017104.8.1.168880	2q37.1	230992101	230995454	B3GNT7	3.50E-15
III	NOTCH4	AC005237.2.1.175179	2q37.3	240214701	240251678	PASK	2.30E-53
xII	KIFC1	AC011298.6.31675.58437	2q37.3	240584194	240668241	ATSV	3.00E-06
III	BAT8	AC034191.5.1.172215	3p26.1	4284929	4298795	SETMAR	2.00E-25
xII	RXRB	AC090947.1.1.166043	3p25.2	12270465	12415723	PPARG	5.30E-09
III	NOTCH4	AC090509.1.1.165994	3p25.1	13551690	13619799	FBLN2	2.40E-09
xII	RXRB	AC090937.1.1.160696	3p25.1	15002136	15024392	NR2C2	6.60E-09
xI	HMGN4	AC027125.4.1.173836	3p25.1	15346179	15346391	no gene	1.90E-06
III	BAT8	AC090950.1.1.199282	3p25.1	15648627	15776696	Y379	3.60E-05
III	HSPA1L	AC097635.2.1.162887	3p24.3	19380503	19387078	Novel	2.90E-102
xII	ZNF297	AC006059.3.1.185161	3p22.1	41880829	41889031	NM_145166	1.10E-11
III	C6orf46	AC099669.2.1.217035	3p21.32	43776685	43804920	Novel	4.10E-13
III	C6orf46	AC124045.1.109944.135528	3p21.32	43934102	43945288	NM_033210	5.10E-12
I	DDR1	AC104439.2.1.197279	3p21.32	45363297	45363752	genscan	7.20E-14
I	DHX16	AC026318.7.1.19068	3p21.31	47135873	47174627	DDX30	5.00E-25
xII	COL11A2	AC005903.3.1.60660	3p21.31	47884518	47915700	COL7A1	5.00E-11
III	NOTCH4	AC005923.2.1.88326	3p21.31	47956918	47983375	CELSR3	3.60E-22
xI	GPX5	AC121247.1.77964.92674	3p21.31	48542852	48544273	GPX1	2.40E-34

III	NOTCH4	AC112215.1.181144.198956	3p21.31	51781935	51811071	STAB1	1.20E-16
I	ABCF1	AC021123.4.149752.161126	3q-tel	91147420	91149286	Novel	1.50E-63
I	DDR1	AC107028.4.1.185539	3q11.2	92010160	92148744	EPHA3	1.20E-15
xII	RPS18	AC108715.2.1.176462	3q11.2	94194400	94194576	no gene	2.40E-15
xII	RPS18	AC108695.2.1.190845	3q11.2	94194185	94194358	Novel	4.30E-24
xII	COL11A2	AC069222.23.1.117000	3q12.1	96016863	96137341	COL8A1	3.50E-23
I	POU5F1	AC117460.7.1.183595	3q12.1	97141764	97141988	EST gene	2.20E-18
I	TUBB	AC046144.15.1.188840	3q13.11	100051837	100051959	genscan	2.00E-11
xII	RPS18	AC073861.18.98940.165924	3q12.3	100725089	100725481	EST gene	4.20E-62
xII	ZNF297	AC084198.24.86116.155268	3q12.3	100797978	100825652	NM_014415	3.60E-12
I	POU5F1	AC079945.13.52386.74222	3q21.3	128590475	128590648	genscan	4.70E-55
III	NOTCH4	AC080007.26.1.168551	3q21.3	129795525	129875755	WDR10	8.50E-10
III	HSPA1L	AC020632.16.1.162029	3q22.1	133021872	133186190	NM_153240	1.90E-22
I	DDR1	AC092969.6.71736.203901	3q22.2	135625566	135694383	Novel	5.10E-21
III	HSPA1L	AC117478.3.1.77155	3q22.3	138390731	138391456	genscan	4.20E-215
III	PBX2	AC018450.26.1.191474	3q24	139874190	139875716	PBXP1	4.50E-161
III	C6orf46/ZNF297	AC010184.18.1.190580	3q23	141935447	141992349	Q8NAP3	6.00E-07
III	BAT1	AC112907.4.101183.186560	3q25.2	151048380	151140293	Novel	2.00E-24
xII	B3GALT4	AC021649.18.1.209521	3q25.33	161749105	161770594	B3GALT3	9.40E-19
III	BAT1	AC092946.7.1.115500	3q27.3	183276473	183282793	EIF4A2	8.70E-15
xII	B3GALT4	AC069417.16.47453.72677	3q27.1	183826771	183845940	B3GNT5	8.40E-28
I	ABCF1	AC048331.32.147427.232441	3q27.1	184898582	184906494	NM_018358	1.60E-37
I	GNL1	AC046143.20.1.180365	3q29	197722675	197769579	NM_018385	3.90E-07
III	NOTCH4	AC021118.6.1.194612	4p15.31	20338712	20704631	SLIT2	1.00E-25
I	DHX16	AC115110.2.23128.113396	4p15.2	24612486	24669565	DDX15	4.30E-58
III	C6orf46/ZNF297	AC105287.4.52544.192873	4p14	39946115	39947420	Novel	1.40E-163
II	HLA Class II	AC097451.2.1.146808	4p13	43676798	43677424	EST gene	1.10E-22
III	LSM2	AC108054.2.101121.147995	4p12	48652844	48737332	Q9P270	1.40E-20
I	DDR1	AC098587.1.9710.175365	4q12	55065587	55126000	PDGFRA	3.00E-05
xI	RFP	AC107058.4.1.126135	4q13.1	65884515	65884994	genscan	1.00E-24
III	BAT8	AC053527.8.1.233250	4q13.3	74230262	74364018	Q9H288	8.00E-05
xI	HIST1H2AC	AC097460.3.1.164370	4q23	101179583	101181853	H2AFZ	2.70E-15
III	BAT1	AC105460.4.1.185755	4q24	104893633	104894151	Genscan	4.00E-10
III	CYP21A2	AC096564.3.1.163317	4q25	109292353	109297744	Novel	2.10E-15
xII	LYPLA2L	AC004062.1.1.154252	4q25	112148688	112149080	no gene	7.40E-66
I	TUBB	AC093663.4.1.171745	4q25	113404522	113404635	genscan	1.20E-09
I	TUBB/RNF5	AC093816.3.1.170227	4q27	123297626	123297751	genscan	4.50E-36
III	BAT8/NOTCH4	AC105421.2.1.162793	4q28.1	125865162	125873003	YB23	2.50E-16
III	NOTCH4	AC092629.2.1.148673	4q28.1	126653380	126692458	NM_024582	2.20E-30
III	HSPA1L	AC093591.3.1.158758	4q28.1	128982992	129034039	OS94	2.00E-16
I	POU5F1	AC093887.3.1.192886	4q31.22	147935518	147939053	POU4F2	5.90E-28
II	TAP1	AC017037.10.1.186106	4q32.1	159166512	159166844	no gene	8.50E-05
III	HSPA1L	AC105250.3.1.70449	4q32.3	165537822	165539805	Novel	3.40E-147
I	IER3	AC106872.5.1.174535	4q32.3	166282549	166282689	no gene	1.10E-07
xI	RFP	AC106872.5.1.174535	4q32.3	166359709	166359709	Novel	8.10E-63
xI	BTN1A1/RFP	AC106872.5.1.174535	4q32.3	166381434	166391179	NM_152620	3.00E-31
xI	BTN1A1	AC108465.3.1.48677	4q32.3	166812464	166832879	Butyrophilin	5.20E-34
xI	RFP	AC080079.5.1.112516	4q32.3	167014840	167015103	genscan	7.60E-22
xII	HSD17B8	AC021151.8.1.175081	4q32.3	170330785	170353453	NM_032783	7.10E-06
III	NOTCH4	AC079226.7.1.184032	4q35.1	184330635	184402306	NM_018104	1.30E-11
III	NOTCH4	AC110761.3.1.153458	4q35.2	188188876	188324915	FAT	2.50E-13
xI	BTN1A1/RFP	AC108073.2.107221.165848	4q35.2	189692345	189706326	NM_173553	8.20E-19
xI	POM121L2	AC093308.2.104996.134015	5p14.3	21925895	21928486	Novel	1.10E-09
xI	SMA3L	AL157879.7.1.161460	5p13.3	34638648	34642503	Novel	2.00E-45
xI	SMA3L	AC114970.1.37696.100300	5p13.3	34737381	34778708	Novel	1.40E-46

xI	POM121L2	AC114970.1.37696.100300	5p13.3	34761244	34778854	Q9H1S5	1.30E-10
I	TUBB	AC106800.1.2657.73135	5p12	43538759	43538860	genscan	8.00E-05
III	BAT1	AC016632.6.1.176784	5q11.2	55195823	55274550	DDX4	6.10E-09
xI	GPX5	AC091977.3.1.183494	5q11.2	55884790	55889596	Q8TED1	2.50E-05
I	DHX16	AC020728.4.1.201404	5q11.2	55980727	56032245	NM_019030	9.30E-13
III	SKIV2L	AC020728.4.1.201404	5q11.2	56032666	56149658	KIAA0052	9.80E-50
xI	SMA3L	AC108108.1.77200.116688	5q13.2	70698028	70703834	Novel (SMA3)	2.00E-48
III	MSH5	AC022493.12.1.153078	5q14.1	80189323	80411166	MSH3	3.90E-15
xII	RPS18	AC008799.6.1.123098	5q14.3	90941311	90941703	genscan	4.70E-49
xII	RXRB	AC106818.1.6855.31297	5q15	93388022	933399296	NRF1	6.60E-19
I	POU5F1	AC108102.2.1.161056	5q15	93545276	93545938	genscan	4.20E-48
III	BAT1	AC016567.8.1.159734	5q15	97361820	97362176	Genscan	1.30E-11
xI	POM121L2	AC114324.1.39743.91827	5q21.1	99294853	99295455	no gene	6.90E-07
xI	SMA3L	AC114324.1.39743.91827	5q21.1	99303064	99303249	genscan	8.30E-27
xI	SMA3L	AC092278.3.1.123469	5q21.1	99829117	99834848	Novel (SMA3)	1.30E-42
III	LSM2/DDR1	AC109481.3.1.20673	5q21.3	108524961	108964792	FER	9.90E-36
I	DHX16	AC093208.2.1.112115	5q22.2	113296938	113337241	Novel	7.10E-07
III	BAT8	AC010226.5.1.147140	5q22.3	115264395	115288381	FEM1C	7.20E-06
xI	HMGN4	AC109456.3.1.121848	5q23.1	115322895	115323107	no gene	5.80E-05
III	NOTCH4	AC008682.6.1.217221	5q23.2	127062260	127232628	NM_032446	1.30E-18
III	NOTCH4	AC010424.9.1.192282	5q23.2	127289102	127326544	NM_130809	9.10E-10
III	NOTCH4	AC025169.5.1.161920	5q23.2	128029494	128309063	FBN2	4.50E-35
III	HSPA1L	AC005373.1.1.112220	5q23.3	129911964	129912584	genscan	6.20E-159
xII	KIFC1	AC004237.1.1.38715	5q23.3	132479552	132517394	KIF3A	4.20E-18
III	HSPA1L	AC113410.2.1.123851	5q31.1	132834893	132887389	HSPA4	3.00E-19
III	BAT1	AC010301.7.1.155067	5q31.1	133745073	133815582	NM_014829	1.40E-09
xI	HIST1H2AC	AC026691.5.1.135062	5q31.1	135237074	135302580	H2AFY	2.10E-21
III	HSPA1L	AC011385.6.1.134599	5q31.2	138566125	138586223	HSPA9B	2.50E-72
I	POU5F1	AC011396.4.1.87692	5q32	146300493	146301825	POU4F3	5.50E-28
III	ATP6V1G2	AC008385.7.1.151712	5q33.1	150765251	150795891	Novel	1.50E-06
xI	GPX5	AC008666.5.1.99108	5q33.1	151002039	151004829	GPX3	4.20E-64
xII	KIFC1	AC008410.5.53020.91332	5q33.2	154988937	154992638	Novel	1.80E-38
III	NOTCH4	AC011369.4.1.141529	5q34	167774142	168258463	Novel	8.20E-10
III	NOTCH4	AC011365.4.1.81930	5q34	168684275	168685047	Novel	7.40E-29
xII	ZNF297	AC104117.1.114225.124230	5q35.3	179277145	179302355	Q8N9F8	5.60E-10
xI	BTN1A1/MOG	AC016572.6.1.143687	5q35.3	180455914	180507663	Butyrophilin	3.30E-54
xI	BTN1A1/MOG	AC091874.2.1.13312	5q35.3	180545631	180563128	BTNL3	4.10E-32
xI	BTN1A1	AC091874.2.13413.147570	5q35.3	180597001	180613745	Butyrophilin	9.10E-53
xI	BTN1A1	AC008443.9.1.120524	5q35.3	180699946	180712464	Q8WV44	4.00E-31
xI	BTN1A1/MOG/BTNL2	AC022413.4.1.166525	5q35.3	181312424	181364180	NM_024850	5.60E-26
II	BTNL2	AC091874.2.13413.147570	5q35.3	181402141	181419638	Butyrophilin	7.20E-23
II	BTNL2	AC091874.2.1.13312	5q35.3	181453511	151470537	NM_152547	2.70E-27
xI	RFP	AC008443.9.1.120524	5q35.3	181733455	181735630	TRIM7	3.80E-58
I	TUBB	AL031963.40.1.149546	6p25.2	3138899	3142759	TUBBL	2.40E-208
I	TUBB	AL445309.13.1.136587	6p25.2	3209729	3212968	TUBBL	2.40E-208
xII	RPS18	AL359643.27.1.166863	6p25.1	4964305	4964649	genscan	7.50E-35
III	C6orf46	AL161903.19.1.47104	6p21.32	33384295	33421772	SYNGAP1	3.30E-17
xII	ZNF297	AL161903.19.1.47104	6p21.32	33407966	33421769	NM_152735	2.90E-50
xII	RXRB	AL022721.1.1.170245	6p21.31	35306800	35392369	PPARD	5.50E-09
xII	KIFC1	AL590387.7.1.76075	6p21.2	39560349	39603922	Novel	1.20E-12
I	TUBB	AL136089.15.1.99479	6p21.2	39962644	39962751	genscan	1.00E-10
I	DDR1	AL355385.15.1.129884	6p21.1	43040485	43125871	PTK7	2.20E-17
xI	BTN1A1	AL512353.16.1.81704	6p34.2	43256510	43284362	ERMAP	2.60E-68
II	TAP2/1	AL359813.23.1.102892	6p21.1	43391706	43414579	ABCB10	2.00E-15
III	NOTCH4	AL359813.23.1.102892	6p21.1	43414506	43420784	NM_023932	5.70E-28
III	CLIC1	AL357057.19.1.58133	6p21.1	45865819	46044480	CLIC5	8.00E-53

xI	SMA3L	AL021368.1.1.188642	6p11.2	58109978	58115773	Novel(SMA3)	7.30E-41
III	NOTCH4	AL137007.9.1.105779	6q12	65571522	65587322	Q9H557	2.30E-24
xII	COL11A2	AL080275.20.1.113983	6q13	70892928	70979474	COL9A1	8.90E-08
xII	LYPLA2L	AL365267.11.1.49616	6q13	71841661	71842056	no gene	7.60E-71
III	C4B	AL590428.7.1.163577	6q13	74364147	74493108	NM_133493	4.10E-08
xII	RPS18	AL355796.11.1.152086	6q14.1	79954508	79954717	no gene	1.60E-22
I	DDR1	AL354857.13.1.199223	6q16.1	93918505	94095966	EPHA7	1.20E-14
I	POU5F1	AL022395.2.1.126882	6q16.2	99299451	99300710	POU3F2	8.00E-43
xII	HSD17B8	AL591803.10.1.90325	6q16.2	99639262	99640282	Novel	1.90E-20
xII	RXRB	AL078596.8.1.64183	6q21	108510126	108532877	NR2E1	2.00E-21
xII	ZNF297	AL109947.19.1.128960	6q21	109806583	109827304	Y441	3.50E-12
I	DDR1	Z97989.1.1.155937	6q21	112005349	112217491	FYN	2.00E-11
I	DDR1	AL357141.8.1.125184	6q22.1	116285557	115404785	FRK	9.00E-20
xII	COL11A2	AL121963.10.1.107553	6q22.1	116462985	116470164	COL10A1	4.00E-23
I	DDR1	Z98880.1.1.108260	6q22.1	117632394	117769882	ROS1	2.60E-14
xII	RPS18	AL357084.12.1.76042	6q24.1	141090774	141091166	genscan	1.60E-48
I	TUBB	AL031320.6.1.133574	6q24.2	143380844	143380963	genscan	1.10E-51
xII	RPS18	AL078581.11.1.102019	6q25.1	149763293	149791539	KATNA1	8.40E-51
III	HSPA1L	AL590413.18.1.104939	6q25.1	151621254	151668524	NM_017909	1.70E-16
xI	MAS1L	AL035691.17.1.129968	6q25.3	160201536	160202670	MAS1	3.10E-36
xII	KIFC1	AL589733.20.1.201088	6q27	168140261	168167477	KIF25	4.00E-07
III	NOTCH4	AL078605.30.1.119563	6q27	170231439	170329846	NM_032448	3.70E-41
III	CYP21A2	AC073957.7.1.196204	7p22.3	667370	673509	NM_017781	2.60E-09
III	BAT8	AC005995.3.1.80010	7p22.1	5716846	5720381	Novel	3.80E-07
III	C6orf46	AC073343.6.1.173967	7p22.1	6374466	6390938	Z325	3.10E-14
III	C4B	AC060834.8.1.113686	7p21.3	9410581	9410991	genscan	4.50E-06
III	HSPA1L	AC009945.2.1.75517	7p21.3	10135315	10136415	genscan	2.30E-231
III	NOTCH4	AC013470.10.1.170723	7p21.3	12014883	12054396	Q96SQ3	1.00E-08
II	TAP2/1	AC002486.1.1.79611	7p21.1	20327136	20342642	O14573	6.30E-24
II	TAP2/1	AC005060.3.1.120169	7p21.1	20365471	20439590	Novel	1.80E-20
I	POU5F1	AC005483.1.1.161667	7p14.1	39025428	39150485	NM_007252	1.80E-11
xI	HIST1H2AC	AC004854.3.1.98697	7p13	44512586	44533943	H2-like	7.00E-15
xI	POM121L2	AC074397.7.1.114576	7p12.1	52567227	52568144	Q8N7R1	8.00E-26
III	C6orf26	AC073057.6.1.178105	7p11.2	56890588	56890821	genscan	5.90E-05
xII	ZNF297	AC115220.1.1.115916	7q11.21	62137099	62202401	Novel	7.80E-08
xI	SMA3L	AC115220.1.1.115916	7q11.21	62213350	62213517	genscan	6.70E-11
xII	ZNF297	AC092685.2.1.183263	7q11.21	63492669	63505908	NM_152626	1.30E-10
xI	SMA3L	AC073261.8.1.93403	7q11.21	64065461	64086982	GUSB	1.00E-24
III	BAT2	AC073089.5.1.171788	7q11.21	65101611	65344287	NM_018264	3.10E-24
III	BAT2	AC091738.4.1.131928	7q11.23	70715527	70834138	Novel	2.00E-22
xI	POM121L2	AC005488.2.1.185737	7q11.23	70990215	71062258	POM121	5.30E-92
xI	POM121L2	AC073841.9.1.55588	7q11.23	71351007	71352683	EST gene	2.90E-72
xI	RFP	AC073841.9.1.55588	7q11.23	71357485	71363082	WBSCR20A	8.90E-39
xI	POM121L2	AC006014.3.1.127761	7q11.23	73578273	73650919	POM121	7.40E-71
II	TAP2/1	AC005045.2.1.123947	7q21.12	85566663	85640217	ABCB4	8.00E-22
II	TAP2/1	AC005068.2.1.98472	7q21.12	85668428	85877856	ABCB1	4.90E-25
xII	COL11A2	AC002528.1.1.141120	7q21.3	92559772	92595972	COL1A2	8.10E-24
III	HSPA1L	AC004957.1.1.160687	7q21.3	95968910	95969113	no gene	3.30E-93
xI	BTN1A1/RFP	AC011904.3.1.113879	7q22.1	98022613	98051761	TRIM4	2.30E-19
xI	HLA Class I	AC004522.2.1.100096	7q22.1	98099206	98108247	AZGP1	1.10E-24
I	DDR1	AC011895.4.1.172358	7q22.1	98934759	98959566	EPHB4	8.80E-21
I	HLA-E	AC006329.5.1.145253	7q22.1	99414126	99414392	genscan	1.10E-05
I	DDR1	AC004416.1.1.32173	7q31.2	114790113	114916094	MET	1.20E-13
xII	LYPLA2L	AC073054.2.1.154419	7q21.32	121348077	121348109	genscan	1.20E-37
III	VARS2	AC008038.1.1.202945	7q33	131061808	131063239	Novel	1.00E-11
I	TUBB	AC083874.2.1.186281	7q33	132762314	132762403	genscan	4.00E-09

I	DDR1	AC104597.3.1.161425	7q34	140878465	140894489	EPHB6	2.70E-09
I	DDR1	AC092214.3.1.72045	7q34	141413870	141431627	EPHA1	1.40E-16
III	C6orf46/ZNF297	AC073422.8.1.80743	7q36.1	147046333	147070069	Novel	2.00E-05
III	C6orf46/ZNF297	AC073314.4.1.73888	7q36.1	147374895	147404413	NM_015694	7.40E-05
II	TAP2/1	AC010973.6.1.222605	7q36.1	148971967	148989086	ABCB8	8.10E-24
III	BAT8	AC010973.6.1.222605	7q36.1	149119222	149130915	ABS10	3.60E-06
I	ABCF1	AC021097.5.1.35899	7q36.1	149151360	149170753	ABCF2	1.20E-36
III	NOTCH4	AC110288.6.1.84664	8p23.3	1266863	1277550	Novel	6.60E-11
III	BF	AC023296.6.1.189532	8p23.2	2642344	3292997	CSMD1	3.00E-06
III	C2	AC023296.6.1.189532	8p23.2	2953246	3468313	CSMD1	3.60E-05
III	BAT1	AC012119.7.1.154898	8p21.2	23396081	23396278	Genscan	3.10E-06
III	HSPA1L	AC090820.6.1.138021	8p12	30546063	30546623	genscan	7.10E-170
III	RNF5	AC069120.4.60762.91265	8p11.23	38073473	38073760	genscan	6.20E-73
III	ATP6V1G2	AC120036.3.58309.177413	8q-tel	46087951	46088247	genscan	5.50E-08
xII	RING1	AC016113.9.1.185419	8q11.23	53945099	53945155	no gene	2.20E-19
xII	LYPLA2L	AC060764.10.163950.183771	8q11.23	54898165	54953804	LYPLA1	2.10E-27
I	DDR1	AC046176.11.1.132489	8q12.1	56731370	56862134	LYN	2.10E-12
xI	HIST1H2AC	AC084251.13.1.181400	8q13.3	70956550	70956840	no gene	3.70E-18
xII	RPS18	AC022730.7.1.155468	8q13.3	71276927	71277244	genscan	1.30E-31
xII	RXRB	AC040917.6.1.158031	8q21.11	76308140	76464431	HNF4G	5.80E-28
xII	ZNF297	AC009812.17.1.155405	8q21.13	81446105	81481018	NM_023929	2.30E-10
I	TUBB	AC007992.12.1.146921	8q22.1	96147896	96147985	genscan	1.40E-09
I	POU5F1	AP002851.2.1.200610	8q22.3	103702017	103702310	no gene	1.70E-17
III	BF	AC007719.7.1.150831	8q23.3	112280606	112310596	Novel	3.10E-05
xI	HIST1H2AC	AC022360.23.1.171991	8q23.3	112695588	112695632	no gene	3.60E-05
III	C2	AC007719.7.1.150831	8q23.3	113296979	113550957	Q96PZ3	9.60E-05
I	DDR1	AC022239.14.1.171829	8q23.1	113344396	11414823	BLK	6.00E-10
xII	ZNF297	AC105210.4.125110.131923	8q24.3	145036223	145040537	Q96C28	1.20E-07
xII	HKE4	AC022505.17.153698.191078	8q24.3	145675676	145680153	SLC39A4	7.40E-08
xII	KIFC1	AC084125.8.1.197314	8q24.3	145729611	145737380	NM_145754	2.40E-17
III	NOTCH4	AC084125.4.1.26696	8q24.3	145753310	145765393	PPP1R16A	2.60E-12
III	C6orf46	AF235103.3.1.344150	8q24.3	146143706	146164020	ZNF64	1.10E-14
I	DDR1	AL161450.14.1.171146	9p24.1	5003251	5108156	JAK2	1.60E-06
xII	RING1	AL162411.23.1.59964	9p24.1	6650955	6651159	no gene	7.00E-39
III	CLIC1	AC017067.4.1.191373	9p21.3	23075247	23075813	genscan	1.30E-48
xI	NOL5B	AL445623.2.1.198637	9p21.3	23903514	23903843	genscan	1.10E-15
III	HSPA1L	AL353745.7.1.174850	9p21.1	31159147	31159635	genscan	5.50E-17
II	BRD2	AL589642.6.1.92982	9p21.1	32799398	32805026	TAF1L	2.80E-08
xII	ZNF297	AL158155.24.1.192336	9p13.2	37607553	37634838	NM_014872	2.30E-10
III	C6orf46	AL353770.18.1.130898	9p13.1	39619983	39630670	Q96M55	1.10E-14
III	CYP21A2	AL359997.8.1.169102	9q21.13	66702026	66702607	genscan	1.20E-14
xII	LYPLA2L	AL353637.16.1.133212	9q21.2	71442203	71442490	no gene	3.50E-59
III	BAT1	AL158047.9.1.201629	9q21.32	75746587	75747222	Genscan	2.30E-19
xII	KIFC1	AL354733.15.1.189579	9q21.32	78368024	78383954	Novel	5.50E-27
I	DDR1	AL445532.8.1.171629	9q21.33	79138134	79490850	NTRK2	1.80E-35
xII	ZNF297	AL136981.22.1.182280	9q22.31	87336127	87368066	Q9H559	2.80E-10
xI	GABBR1	AL445495.5.1.155837	9q22.33	92793899	93215009	GPR51	1.60E-09
xII	COL11A2	AL354923.12.1.134965	9q22.33	93449719	93576588	COL15A1	3.10E-12
III	C6orf29	AL450265.11.1.68871	9q31.1	99804995	99897178	CTL1	1.40E-12
III	C9orf29	AL450265.11.1.68871	9q31.1	99804995	99897178	NM_022109	1.40E-12
III	BF	AL158158.14.1.194835	9q31.3	103971543	103990925	Novel	1.70E-11
III	C2	AL158158.14.1.194835	9q31.3	104860730	104923764	NM_153366	2.50E-15
III	NOTCH4	AL354982.12.1.119077	9q31.3	104936930	105074331	Novel	1.20E-33
I	DDR1	AL157881.14.1.162726	9q31.3	105163309	105295448	MUSK	1.80E-41
III	C6orf46	AL159168.15.1.129010	9q31.3	106019666	106038882	Q8TF39	1.80E-14

III	BAT2	AL354877.25.1.116236	9q31.3	106181071	106289393	NM_173521	3.30E-17
xII	ZNF297	AL162588.22.1.76606	9q31.3	107536344	107551166	ZFP37	7.80E-09
III	NOTCH4	AL162425.15.1.177728	9q31.3	108581009	108678635	Novel	6.30E-28
xII	COL11A2	AL445543.20.1.140327	9q31.3	108662161	108805998	Q96JF7	2.60E-13
III	ATP6V1G2	AL160275.14.1.189709	9q32	109082196	109092823	ATP6V1G1	1.50E-12
III	TNF	AL390240.18.1.93876	9q32	109283763	109300763	TNFSF15	3.80E-10
III	TNXB	AL162425.15.1.177728	9q33.1	109514975	109612609	TNC	7.80E-25
I	TUBB	AL589703.6.1.48697	9q33.1	112631091	112631204	genscan	1.60E-10
xII	B3GALT4	AL161911.17.1.109176	9q33.2	115207621	115207998	genscan	6.40E-31
III	C4B	AC006430.22.1.194799	9q33.2	115361172	115459110	C5	5.70E-20
III	C6orf46	AC007066.4.1.190815	9q33.2	117340094	117340094	BIOR	9.80E-65
III	NOTCH4	AL445489.10.1.175869	9q33.3	117788495	118338969	NM_024820	2.20E-76
xII	RXRB	AL354979.17.1.85997	9q33.3	118928956	119180139	NR6A1	9.70E-13
III	HSPA1L	AL354710.17.1.131708	9q33.3	119643682	119650159	HSPA5	8.1E-161
III	PBX2	AL445186.4.1.156124	9q33.3	120156161	120376205	PBX3	9.10E-85
III	C6orf46/ZNF297	AL161731.20.1.182452	9q33.3	121213845	121244041	ZNF297B	7.30E-33
xII	C6orf46/ZNF297	AL354944.22.1.49144	9q33.3	121269484	121289729	Q8NCN2	9.50E-39
I	DDR1	AL161733.20.1.176466	9q34.12	125136236	125309589	ABL1	8.50E-12
III	AIF1	AL157938.22.1.197019	9q34.12	125518441	125545061	NM_031426	1.20E-56
III	BAT2	AL358781.19.1.147492	9q34.13	125852061	125869120	NM_032640	2.20E-65
xII	RAB2L	AL162417.23.1.152863	9q34.2	127529965	127553410	RALGDS	3.70E-29
II	BRD2	AL445931.29.1.175033	9q34.2	128566862	128602533	BRD3	1.90E-133
xII	RXRB	AL669970.6.1.58552	9q34.2	129062693	129101647	RXRA	6.30E-88
xII	COL11A2	AL603650.10.1.131466	9q34.3	129302868	129503955	COL5A1	2.10E-43
III	NOTCH4	AL390778.30.1.221373	9q34.3	129736484	129782241	OLFM1	4.60E-11
III	NOTCH4	AL353615.27.1.37093	9q34.3	130006614	130009927	NM_173520	6.00E-10
xII	ZNF297	AL591038.9.1.51295	9q34.3	130674721	130713948	NM_144653	7.90E-15
III	NOTCH4	AL592301.14.1.188462	9q34.3	131078383	131129726	NOTCH1	2.40E-224
III	EGFL8	AL590226.23.1.149567	9q34.3	131242795	131256617	ZNEU1	9.70E-12
III	AGPAT1	AL590226.23.1.149567	9q34.3	131257082	131271362	AGPAT2	2.50E-29
III	CLIC1	AC068451.2.53215.58850	9q34.3	131578574	131580507	CLIC3	3.30E-35
III	BAT8	AL611925.20.31668.168509	9q34.3	132395728	132553855	HMT1	7.30E-142
I	TUBB	AL713922.8.1.121218	10p15.3	33000	35178	TUBBL	2.70E-190
III	NOTCH4	AL513304.27.1.163243	10p15.3	1427949	1428792	no gene	3.20E-06
xII	KIFC1	AL161932.15.1.143423	10p11.22	3016668	32061904	KIF5B	1.00E-06
III	HSPA1L	AC069544.9.1.214866	10p13	14843884	14877306	NM_016299	1.00E-15
III	BAT8	AC069544.9.1.214866	10p13	14884428	14909880	SU92	1.40E-17
III	NOTCH4	AL133415.12.1.179912	10p13	17152297	17207242	DNMT2	1.10E-15
xII	HKE4	AL590111.14.1.41069	10p12.33	17957452	18048843	NM_152725	7.10E-08
xII	RPS18	AL513128.11.1.184685	10p12.2	22534350	22534484	no gene	1.20E-05
xII	ZNF297	AL117337.25.1.161452	10q11.21	37982417	38009185	ZNF25	1.00E-07
xII	ZNF297	AL161931.13.1.19853	10q11.21	38043229	38099906	ZNF33A	1.30E-10
xII	ZNF297	AL022345.2.1.146328	10q11.21	42553027	42602464	ZNF11B	1.30E-10
III	C6orf46	AL353801.13.1.222490	10q11.21	44964852	44969243	ZNF22	4.10E-14
III	BAT8	AL359377.18.1.172177	10q21.2	60813515	61174843	ANK3	6.70E-05
I	POU5F1	AL356741.11.1.87244	10q21.3	68768389	68953381	Q9HCH9	1.40E-75
III	BAT1	AL359844.15.1.171364	10q22.1	69527140	69554693	DDX21	1.20E-05
III	BAT1	AC016394.13.1.149726	10q22.2	73713068	73782589	Q9Y2I0	1.90E-10
xI	HIST1H2AC	AL391421.27.1.168239	10q22.3	78922958	78923248	no gene	1.00E-14
III	BAT1	AL365434.12.1.158357	10q23.31	91694792	91695643	Genscan	7.30E-17
III	BAT1	AL731553.9.1.161141	10q23.31	91759982	91760050	Genscan	7.40E-17
III	BAT1	AL158040.13.1.213648	10q23.32	92777959	92779325	Novel	1.50E-24
xII	KIFC1	AL356128.27.1.191935	10q23.33	93574641	93636806	KIF11	1.50E-11
III	CYP21A2	AL359672.19.1.143181	10q23.33	95689793	95722511	CYP2C8	3.20E-05
III	NOTCH4	AL442123.12.1.96660	10q24.1	97651046	97838934	SLIT1	6.90E-33
xII	ZNF297	AL135791.12.1.66975	10q24.1	97820294	3785029	Q9NQN2	7.80E-09

II	TAP2	AL392107.16.1.94970	10q24.2	100776166	100845227	ABCC2	7.70E-05
III	CYP21A2	AL358790.22.1.131753	10q24.32	103483494	103490378	CYP17	4.20E-12
xI	RFP	AL391121.29.1.166600	10q24.32	103638012	103651712	TRIM8	1.90E-07
I	DHX16	AL360176.22.1.155699	10q26.2	126728551	126773529	DDX32	1.10E-21
III	CYP21A2	AL161645.14.1.161644	10q26.3	134255131	134266884	CYP2E	2.00E-07
xI	MAS1L	AC108448.5.135208.198047	11p15.4	3499638	3500522	Novel	5.70E-29
xI	RFP/RING1	AC009758.8.1.141485	11p15.4	4708183	4716972	SSA1	9.40E-71
xI	RFP	AC090719.8.1.179177	11p15.4	4921932	4931481	NM_018073	6.50E-72
xI	BTN1A1	AC009758.8.1.141485	11p15.4	4972222	4981011	SSA1	6.10E-38
xI	BTN1A1	AC090719.8.1.179177	11p15.4	5185974	5195520	Novel	1.40E-35
xI	RFP	AC015691.6.1.203036	11p15.4	5919434	5967728	TRIM6	1.30E-45
xI	RFP	AC109341.7.1.202761	11p15.4	5986894	6008393	TRIM5	9.10E-44
xI	BTN1A1	AC015691.6.1.203036	11p15.4	6372568	6413722	TRIM34	1.70E-16
xI	BTN1A1	AC109341.7.1.202761	11p15.4	6459104	6478807	TRIM22	1.70E-16
III	CYP21A2	AC018795.10.1.187836	11p15.2	15932556	15932621	no gene	2.90E-18
III	CYP21A2	AC090835.6.82428.167443	11p15.2	16080713	16094757	Novel	3.10E-19
xI	MAS1L	AC090099.10.28570.173306	11p15.1	19101039	19102007	MRGX3	1.70E-39
xI	MAS1L	AC107948.7.1.156839	11p15.1	19137100	19138068	MRGX4	3.30E-38
xI	MAS1L	AC023078.9.1.163718	11p15.1	19899153	19900121	MRGX1	2.00E-40
xI	MAS1L	AC023078.9.1.163718	11p15.1	19926836	19991360	Novel	2.40E-17
xI	MAS1L	AC027026.9.1.155376	11p15.1	20020747	20021739	MRGX2	1.10E-35
xII	KIFC1	AC023206.6.1.208561	11p14.1	28817128	28904683	NM_031217	5.40E-07
III	BF	AL354921.12.1.106657	11p13	37005965	37099743	Q96JW2	2.60E-06
III	NOTCH4	AC061999.6.1.182549	11p12	37290819	37297112	RAG2	3.90E-26
xII	RXRB	AC090589.8.1.190017	11p11.2	48157255	48168103	NR1H3	1.40E-06
xII	RXRB	AC018410.19.7721.155276	11p11.2	48168666	48229300	MADD	1.10E-06
xII	HKE4	AC090559.5.26090.106816	11p11.2	48306544	48315768	NM_152264	8.60E-14
xII	RXRB	AP001453.4.1.166300	11q13.1	65754594	65765769	ESRRA	9.70E-18
III	NOTCH4	AP000769.4.1.114794	11q13.1	66974259	66987851	SCYL1	1.20E-19
III	NOTCH4	AP001362.5.1.211382	11q13.1	67025483	67041797	Novel	1.70E-22
xI	MAS1L	AP000808.4.1.176380	11q13.3	70444307	70445269	Q8TDS7	7.50E-35
xI	MAS1L	AP003071.2.1.192759	11q13.3	70468658	70477508	MRGF	4.00E-27
xI	HIST1H2AC	AP002336.3.1.112484	11q13.3	71639671	71753357	PPFIA1	4.50E-14
III	NOTCH4	AP000867.4.1.199996	11q13.4	72870983	72871905	Q8NH65	3.20E-07
I	C6ORF18	AP000719.4.1.196424	11q13.4	73253747	73331398	NUMA1	1.90E-06
xII	B3GALT4	AP000752.4.1.194140	11q13.5	78289628	78290785	NM_138706	5.70E-25
III	NOTCH4	AP002768.3.1.186084	11q14.1	79903349	79952134	Q9P2P4	1.50E-09
xI	PRSS16	AP001646.4.1.182328	11q14.1	84074231	84150294	PRCP	4.10E-06
III	BAT1	AP003390.1.1.221091	11q23.3	91694792	91695643	Genscan	2.10E-106
I	TUBB	AP002364.3.1.165702	11q14.3	92344989	92345090	genscan	1.50E-11
I	TUBB	AP002799.3.1.177564	11q14.3	94012104	94012214	genscan	2.30E-14
III	NOTCH4	AP003171.2.1.137000	11q14.3	94081540	94134035	Q8TDW7	2.90E-10
III	BAT8	AP000786.4.1.75440	11q21	95734866	95740452	NM_017704	7.20E-06
III	BAT8	AP002840.2.1.177034	11q23.2	114770675	114783056	Q98NFD2	3.30E-15
III	NOTCH4	AP002840.2.1.177034	11q23.1	114792387	114857963	DRD2	2.10E-10
III	BAT8	AP001267.4.1.194310	11q23.3	119819041	119907224	MLL	2.50E-05
xI	HIST1H2AC	AP003391.1.1.46239	11q23.3	120476378	120477968	H2AFX	2.80E-42
III	BAT1	AP000713.2.1.11316	11q23.3	120964857	120979716	DDX6	1.20E-06
I	POU5F1	AP001150.4.1.157282	11q23.3	121622699	121702405	POU2F3	6.50E-21
III	BAT1	AP001994.4.1.167376	11q23.3	121762568	121762732	No gene	1.50E-105
III	HSPA1L	AP000926.5.1.196973	11q24.1	124441468	124446116	HSPA8	9.10E-243
III	BAT1	AP000842.4.1.179369	11q24.2	127774844	147793446	DDX25	1.80E-10
III	C6orf46/ZNF297	AP001183.4.1.174526	11q24.3	131612944	131697119	NM_014155	7.80E-15
III	BAT1	AC019227.4.1.190314	11q24.3	132166105	132166485	Genscan	1.50E-13
I	PPP1R10	AP000824.4.1.186920	11q24.3	132172516	132173124	genscan	7.20E-17
I	PPP1R10	AP003486.2.1.217488	11q24.3	132258421	132299008	SNXJ	9.60E-17

III	BAT1	AP000435.5.1.124067	11q12.1	60403473	60403808	Genscan	3.00E-12
III	HSPA1L	AC007207.22.1.191877	12p13.32	4100374	4101084	genscan	1.80E-215
xII	TAPBP	AC005840.2.1.140026	12p13.31	6535709	6550143	TAPBP-R	3.30E-05
III	BF	AC006512.12.1.157115	12p13.31	7087044	7087106	no gene	3.10E-06
III	C2	AC006512.12.1.157115	12p13.31	7243215	7245179	C10	6.10E-09
III	C4B	AC006581.16.1.172931	12p13.31	8712438	8724661	Novel	2.80E-05
III	C4B	AC007436.1.1.163881	12p13.31	8928367	8976544	A2M	3.20E-09
III	C4B	AC010175.4.1.127277	12p13.31	9009493	9069023	PZP	6.20E-09
III	BAT1	AC007215.43.2235.65215	12p13.2	12695938	12712573	NM_016355	1.60E-11
xI	HIST1H2AC	AC010168.6.1.104926	12p12.3	15072102	15073039	H2AFJ	1.10E-44
xII	COL11A2	AC004801.1.1.193561	12q13.11	48379436	48410949	COL2A1	2.80E-38
III	BAT1	AC025557.4.146238.171945	12q13.13	49116936	49138712	NM_004818	1.10E-14
I	TUBB	AC011603.33.10243.45426	12q13.12	49484144	49487748	TUBA1	9.00E-11
I	TUBB	AC010173.22.67252.90665	12q13.12	49586964	49595224	TUBA6	2.20E-12
I	TUBB	AC010173.22.160578.20759	12q13.12	49655804	49660085	TUBA1	1.00E-11
xII	RXRB	AC025259.48.1.210158	12q13.13	52474503	52482549	NR4A1	1.10E-07
III	BAT1	AC055716.24.1.110819	12q13.13	53260460	53260801	Genscan	5.40E-27
III	BAT1	AC068988.19.27848.161382	12q13.13	53292388	53292489	Genscan	7.70E-27
III	BAT1	AC073573.27.1.157807	12q13.13	53706088	53706432	Genscan	4.30E-35
III	BAT8	AC073896.29.107190.140910	12q13.2	56645899	56649310	NM_173594	2.30E-05
xII	ZNF297	AC026120.33.1.171998	12q13.3	57617302	57624914	Y352	1.10E-11
III	BAT1	AC117498.1.134066.149599	12q14.1	61110934	61180767	Novel	1.50E-15
xI	NOL5B	AC027288.26.1.177080	12q12.2	80316560	80316925	genscan	6.20E-24
xI	BTN1A1	AC009771.13.122068.178104	12q23.3	107466485	107466640	no gene	2.10E-30
xI	MAS1L	AC063957.22.1.71430	12q23.3	108566149	108613888	CMKLR1	8.80E-05
III	HSPA1L	AC005805.9.96579.142875	12q24.11	111438197	111439237	Novel	1.20E-134
III	CLIC1	AC078875.25.5011.18452	12q24.31	120214564	120215259	Novel	1.60E-14
xII	B3GALT4	AC048338.22.82693.113969	12q24.31	122667939	122671768	B3GNT4	2.10E-23
II	TAP2/1	AC026362.34.74237.162900	12q24.31	123114813	123152304	ABCB9	5.50E-45
I	DHX16	AC093719.6.127047.199959	12q24.31	125177591	125220663	DDX37	8.20E-17
xII	ZNF297	AC026786.5.1.160615	12q24.33	133310141	133338979	ZNF10	6.10E-10
xII	RXRB	AL359457.12.1.129779	13q12.11	14106712	14124427	ESRRAP	3.40E-27
xII	RXRB	AL158032.32.1.172004	13q12.11	15813801	15815722	Novel	4.50E-27
I	TUBB	AL139327.18.1.149559	13q12.11	17727916	17735936	TUBA2	1.10E-09
III	BAT1	AL354828.12.1.168114	13q12.12	21259290	21260180	Genscan	1.10E-111
I	DDR1	AL591024.14.1.76721	13q12.2	22557753	22654705	FLT3	9.00E-05
xI	POM121L2	AL359741.9.1.139877	13q12.3	23332361	23332804	genscan	1.20E-19
xI	POM121L2	AL596092.8.1.153841	13q12.3	23579447	24059956	O94872	2.70E-15
III	HSPA1L	AL137142.20.1.113850	13q12.3	25697387	25722697	H105	8.20E-22
III	BAT1	AL138822.13.1.126502	13q12.3	27152990	27154150	Genscan	1.50E-50
xI	HIST1H2AC	AL159980.14.1.162044	13q13.3	31050674	31050910	genscan	5.60E-20
III	BAT1	AL138706.9.1.195032	13q13.3	35499252	35499851	Genscan	1.40E-45
III	BAT8	AL136218.26.1.159863	13q14.2	44005950	44053746	C13ORF4	3.20E-06
xII	RXRB	AL138997.18.1.172342	13q21.1	50568900	50569076	genscan	5.20E-14
III	BAT1	AL161901.18.1.150054	13q21.2	59178906	59179631	Genscan	1.90E-31
xI	RFP	AL136145.23.1.83809	13q21.32	60838247	60838498	genscan	9.90E-09
I	DHX16	AC001226.1.1.106988	13q22.3	71554239	71557564	Novel	4.40E-10
I	POU5F1	AL445209.4.1.157302	13q31.1	73168139	73172615	POU4F1	7.40E-28
II	TAP2/1	AL157818.12.1.182485	13q32.1	90059271	90340865	ABCC4	7.60E-05
xI	HIST1H2AC	AL160155.19.1.149478	13q32.3	94254802	94255185	H2A-like	1.60E-11
xII	COL11A2	AL390755.5.1.186120	13q34	105188574	105346678	COL4A1	3.30E-17
xII	COL11A2	AL159153.17.1.102319	13q34	105346805	105553028	COL4A2	2.40E-16
III	NOTCH4	AL137002.19.1.132933	13q34	108397279	108411519	F7	2.60E-18
III	NOTCH4	AL161774.49.1.162296	13q34	110946220	111086222	RASA3	9.60E-09

II	PSMB8	AL132780.5.1.191946	14q11.2	17282361	17291410	PSMB5	2.40E-67
xI	HMGN4	AL163052.4.1.181905	14q12	23108107	23108319	no gene	8.20E-06
I	TUBB	AL445383.5.1.172914	14q21.2	41065751	41065876	genscan	7.00E-10
xII	RXRB	AL161756.6.1.176257	14q23.2	58487079	58598473	ESR2	8.90E-05
III	HSPA1L	AL049869.6.1.195840	14q23.3	58801222	58803614	HSPA2	4.80E-266
xI	GPX5	AL139022.4.1.190517	14q23.3	59199531	59203136	GPX2	1.50E-27
III	BAT1	AL391262.3.1.171296	14q24.1	67056068	67056841	Genscan	1.00E-18
III	NOTCH4	AC005479.2.1.140425	14q24.3	68764077	68777511	NPC2	3.70E-22
xII	RXRB	AC008050.6.1.176975	14q24.3	70654841	70785295	ESRRB	5.10E-09
xII	RPS18	AL122020.5.1.149904	14q32.11	85048919	85049456	Novel	4.40E-53
III	NOTCH4	AL132711.4.1.184924	14q32.2	95016286	95017418	no gene	6.90E-22
III	C6orf46/ZNF297	AL590327.3.1.59297	14q32.33	99258499	99259764	Novel	3.20E-14
III	NOTCH4	AL512356.5.1.158468	14q32.33	99443586	99452823	C14orf79	1.60E-59
III	NOTCH4	AL512355.5.1.196132	14q32.33	99772045	99852517	O60342	1.90E-54
III	ATP6V1G2	AL122127.6.1.169802	14q32.33	103287219	103287536	no gene	1.00E-07
II	TAP1	AC116165.3.1.90200	15q11.2	16305068	16351483	Novel	1.70E-35
II	TAP2	AC116165.3.1.90200	15q11.2	16305068	16351483	Novel	7.20E-22
II	TAP1	AC016033.7.99902.141149	15q11.2	16392186	16403533	Novel	7.70E-36
II	TAP2	AC016033.7.99902.141149	15q11.2	16392186	16403533	Novel	7.50E-23
xI	POM121L2	AC090983.10.101166.203171	15q11.2	17724332	17724397	no gene	2.20E-11
II	TAP2	AC091304.12.1.179219	15q13.1	21513802	21524551	Novel	1.50E-44
II	TAP1	AC091304.12.1.179219	15q13.1	21513802	21524551	Novel	2.30E-40
xI	HMGN4	AC022613.13.1.188117	15q13.1	25526308	25530507	HMG17	6.40E-06
III	NOTCH4	AC020661.8.1.191655	15q15.1	34166830	34304183	Q9ULG1	1.70E-42
II	HLA Class II	AC025270.6.1.128484	15q21.1	37899544	37906166	B2M	7.10E-05
xI	BTN1A1/RFP	AC018901.8.1.199503	15q21.1	37924579	37955869	RNF36	1.20E-29
II	HLA-DPB1	AC018901.8.1.199503	15q21.1	38056731	38056793	no gene	6.90E-06
III	CYP21A2	AC020705.4.136565.149466	15q21.1	38739399	68813334	CYP1A2	1.10E-12
III	NOTCH4	AC022467.7.1.193703	15q21.1	41748164	41983082	FBN1	8.10E-27
III	BAT1	AC091700.4.1.97653	15q22.2	55722396	55722959	Genscan	4.20E-08
III	NOTCH4	AC009433.11.1.169638	15q22.31	59286605	59645098	NM_032445	4.90E-28
III	BAT8	AC067837.6.1.173919	15q23	61668952	61687004	FEM1B	1.70E-09
III	BAT8	AC021553.14.1.185596	15q23	61692853	61823052	ITGA11	2.50E-07
xII	RXRB	AC104938.2.66191.114293	15q23	65200660	65208271	NR2E3	3.40E-22
III	RNF5	AC048383.8.169960.172969	15q23	66533763	66534050	genscan	2.30E-74
II	BTNL2	AC022188.7.15746.68046	15q24.1	67107924	67122957	NM_025240	5.80E-18
III	CYP21A2	AC020705.4.92855.102206	15q24.1	68848252	68854305	CYP1A1	4.90E-15
I	DDR1	AC027243.13.89123.218680	15q24.2	69598117	69693731	ETFA	1.90E-13
xI	MOG	AC022188.7.15746.68046	15q24.1	70012753	70027796	NM_025240	9.70E-10
III	CYP21A2	AC091230.8.108454.128536	15q24.1	70970184	70976444	Novel	9.00E-12
I	DDR1	AC011966.7.1.167862	15q25.3	81649264	82028917	NTRK3	7.20E-24
xII	KIFC1	AC079075.5.54114.209978	15q26.1	83693866	83715674	ANPEP	1.10E-15
xI	HIST1H2AC	AC091544.9.1.126968	15q26.1	87091764	87110331	H2 -like	9.40E-17
xII	RXRB	AC016251.9.1.182943	15q26.2	90630767	90631006	no gene	4.20E-43
I	DDR1	AC069029.9.1.191018	15q26.3	93033248	93342019	IGF1R	3.10E-07
xI	BTN1A1/RFP	AJ003147.1.1.239566	16p13.3	3325667	3340266	MEFV	2.70E-35
II	BRD2	AC004651.1.1.42016	16p13.3	3810213	3964357	CREBBP	1.80E-05
xII	B3GALT4	AC040160.4.1.209574	16q22.1	6761048	67659124	FHOD1	3.70E-19
II	TAP2	AC025778.7.1.207614	16p13.12	15526117	15599260	ABCB6	1.30E-06
III	NOTCH4	AC106796.1.45233.67716	16p12.3	19774406	19794065	UMOD	2.00E-11
xI	HMGN4	AC093509.2.1.120576	16p12.1	25470563	25470814	Q96C64	3.90E-06
xII	KIFC1	AC023831.8.22510.115251	16p11.2	30014062	30032892	QPRT	2.00E-09
III	C6orf46	AC002310.1.1.120955	16p11.2	31053573	31058082	NM_033410	5.30E-09
III	C6orf46	AC093249.3.1.185664	16p11.2	31102367	31110170	Q96CS4	5.30E-08
xII	ZNF297	AC106886.2.20127.148471	16p11.2	31281559	31287006	Q9UEG4	6.00E-10

xI	RFP	AC009088.7.1.233305	16p11.2	31722146	31734564	Q8N4X6	1.40E-33
II	PSMB8	AC007494.7.1.206113	16q12.1	47130249	47435934	PSMB10	3.80E-43
II	TAP1	AC096996.1.1.194627	16q12.1	48249418	48329904	ABCC11	2.80E-05
xI	UBD	AC026473.7.1.170393	16q21	51486489	51486719	genscan	6.60E-05
xII	KIFC1	AC092118.2.1.148401	16q13	57876138	57920423	KIFC3	1.20E-25
III	BAT1	AC004531.1.1.191565	16q22.1	58750455	58752074	DDX28	2.00E-18
xII	B3GALT4	AC074143.4.1.152953	16q22.1	67560936	67562065	NM_033309	2.70E-19
I	DHX16	AC009087.4.1.174933	16q22.2	72851029	72870003	DDX38	2.90E-50
xII	ZNF297	AC009078.6.1.176926	16q23.1	76142879	76166439	NM_153688	7.80E-09
III	BAT1	AC093491.2.1.162178	16q24.1	76485011	76485421	Genscan	5.20E-16
III	CLIC1	AC092327.3.1.189757	16q24.1	77844902	77845267	Genscan	7.10E-08
xII	ZNF297	AC009113.5.61390.188481	16q24.3	90272052	90283297	Q96MU6	3.70E-10
I	TUBB	AC092143.3.1.183047	16q24.3	90971498	90989716	TUBBL	9.20E-195
xII	ZNF297	AC090617.7.1.169947	17p13.3	2300029	2302554	HIC1	4.70E-10
III	BAT1	AC015799.7.1.66824	17p13.3	2858331	2858387	genscan	1.70E-20
I	DDR1	AC087742.7.63895.97713	17p13.2	4468643	4469263	EST gene	4.80E-11
II	PSMB9	AC027820.9.1.56340	17p13.2	5044936	5047269	PSMB6	3.50E-26
xII	KIFC1	AC004771.1.1.91927	17q13.2	5239168	5273857	KIF1C	8.70E-05
xII	ZNF297	AC087500.12.1.136618	17p13.2	5420893	5425754	Q96JF6	6.10E-10
I	DHX16	AC004148.1.1.118276	17p13.2	5683637	5710243	DDX33	2.10E-45
I	DDR1	AC113189.3.50089.71700	17p13.1	8014004	8022650	TNK1	1.10E-16
III	BAT1	AC016876.5.1.48645	17p13.1	8215901	8221709	EIF4A1	3.20E-18
III	BAT1	AC007421.12.1.95240	17p13.1	8215901	8221709	EIF4A1	1.30E-05
xII	RPS18	AC013248.5.1.66571	17p12	15723309	15723764	Novel	5.60E-60
xII	ZNF297	AC005324.1.1.176643	17p12	16717966	16767419	ZNF386	2.20E-10
III	C6orf46	AJ009612.5.1.148978	17p11.2	17565367	17583130	ZNF287	4.10E-13
III	C6orf46	AC005822.1.1.169931	17p11.2	17634716	17638760	YD49	4.10E-14
xII	ZNF297	AC026271.6.1.171978	17p11.2	20226123	20246008	Novel	1.70E-10
III	PBX2	AC087499.8.20079.65528	17p11.2	20668666	20668839	genscan	2.80E-05
xI	UBD	AC087575.3.156902.181085	17q	24638783	24639010	UBB	1.30E-12
I	FLOT1	AC024267.9.50190.98519	17q11.2	29105687	29123905	FLOT2	1.70E-31
xII	RXRB	AC068669.4.36251.62842	17q21.1	40302169	40309811	NR1D1	6.20E-10
xII	RXRB	AC080112.4.61535.75578	17q21.2	40640652	40689179	RARA	7.90E-13
III	NOTCH4	AC006070.1.1.161987	17q21.2	41487285	41488289	KRTAP9-9	4.80E-18
III	NOTCH4	AC003958.1.1.127834	17q21.2	41624135	41630437	KRTHA3B	6.70E-10
I	DHX16	AC068675.9.124153.141665	17q21.31	43738377	43778728	DDX8	4.80E-52
xII	KIFC1	AC015936.7.29291.133312	17q21.31	45189863	45189988	no gene	1.50E-05
xII	COL11A2	AC015909.8.44136.121814	17q21.33	47864916	47882452	COL1A1	4.00E-32
xII	KIFC1	AC019315.9.1.152057	17q22	54381662	54383970	NM_032559	6.40E-24
xI	RFP	AC004584.1.1.104871	17q23.2	57450065	57472912	ZNF147	3.60E-19
I	DHX16	AC004167.1.1.124876	17q23.2	60092562	60135284	NM_024612	4.60E-25
I	DHX16	AC005702.1.1.147686	17q23.2	60503679	60528304	Novel	7.50E-15
III	BAT1	AC015651.18.1.191583	17q23.3	64290823	64323079	NM_007372	1.20E-08
III	CLIC1	AC004805.1.1.184263	17q24.1	64773117	64773701	Novel	5.70E-08
III	BAT1	AC009994.6.166827.180372	17q24.2	68976215	68982889	DDX5	2.00E-07
III	BAT1	AC087741.2.60294.77121	17q25.3	81412654	81424538	IF4N	6.80E-06
I	TUBB	AP001005.5.1.137000	18p11.32	35028	37159	TUBBL	2.80E-184
III	BAT1	AP002449.2.169334.172757	18p11.21	12998814	12998903	genscan	7.60E-17
II	C6orf10	AC006238.1.1.211945	18q11.2	23904177	23904497	genscan	2.10E-06
xII	ZNF297	AC105101.6.1.172381	18q12.1	45351393	45363238	O75453	5.40E-18
xII	ZNF297	AC006130.1.1.84984	19p13.3	2936627	2947795	NM_024967	4.90E-11
I	DDR1	AC005777.1.1.43190	19p13.3	3847245	3871088	MATX	9.30E-07
xII	ZNF297	AC016586.7.116093.145761	19p13.3	4117018	4136099	O00456	5.40E-16
III	BAT8	AC005523.1.1.41468	19p13.3	4860388	4864189	FEM1A	2.70E-09

I	TUBB	AC010503.8.1.141295	19p13.3	6562943	6570948	TUBBL	1.30E-202
III	TNF	AC008760.7.1.200167	19p13.3	6733175	673234	TNFSF14	6.20E-10
III	C4B	AC008760.7.1.200167	19p13.3	6746489	6789295	C3	3.50E-27
III	NOTCH4	AC020895.8.1.139846	19p13.3	6959105	7022006	EMR1	1.00E-08
I	DDR1	AC010311.9.1.91172	19p13.2	7254547	7432507	INSR	3.10E-12
xII	LYPLA2L	AC010336.7.1.160769	19p13.2	8042351	8049685	Novel	3.30E-96
III	NOTCH4	AC022146.6.66353.150193	19p13.2	8235200	8317297	FBN3	5.30E-34
III	EGFL8	AC022146.6.66353.150193	19p13.2	8275726	8322330	NM_032447	2.60E-08
xII	COL11A2	AC008742.8.1.194623	19p13.2	10191743	10242653	COL5A3	3.10E-67
I	DDR1	AC011557.6.1.30505	19p13.2	10684031	10714039	TYK2	2.20E-07
III	LSM2	AC011475.6.1.179953	19p13.2	10932292	10932453	no gene	5.70E-23
III	C6orf29	AC011475.6.1.179953	19p13.2	10959135	10978061	CTL2	5.20E-59
xII	RAB2L	AC024575.6.1.119638	19p13.2	11718017	11752815	Q8TEP0	1.30E-24
xII	ZNF297	AC011446.6.1.115932	19p13.2	13622357	13628643	STX10	2.20E-14
III	BAT1	AC008569.7.1.227245	19p13.13	14887089	14897635	DDX39	4.10E-98
III	NOTCH4	AC005327.1.1.37988	19p13.12	15236788	15282936	EMR2	5.70E-05
III	EGFL8	AC004663.1.1.41150	19p13.12	15649643	15690991	NOTCH3	6.00E-05
III	NOTCH4	AC004663.1.1.41150	19p13.12	15664050	15705404	NOTCH3	7.70E-227
II	BRD2	AC114486.2.1.179070	19p13.12	15741907	15784868	BRD4	2.30E-90
xII	RXRB	AC010646.5.1.41461	19p13.12	17734984	17748449	NR2F6	7.00E-41
xII	B3GALT4	AC008761.7.1.226170	19p13.12	18106912	18149110	Q9UPW8	1.10E-12
xII	B3GALT4	AC005952.1.1.39976	19p13.11	18298235	18315904	B3GNT3	2.50E-14
xII	RPS18	AC020904.7.1.148824	19p13.11	18551604	18551837	EST gene	5.30E-46
III	BAT1	AC002985.1.1.38041	19p13.11	19422473	19431417	NM_019070	1.50E-05
III	PBX2	AC011448.4.1.165122	19p13.11	20063771	20120711	PBX4	2.10E-68
xII	ZNF297	AC008751.6.1.169089	19p13.11	21436099	21452779	ZNF85	2.20E-10
III	C6orf46	AC016628.6.1.41153	19p13.11	23871643	23887148	Novel	6.70E-16
xII	ZNF297	AC020910.7.1.203201	19q13.12	35697932	35713073	Q96NL3	2.30E-11
III	BAT8	AD000671.1.1.46251	19q13.12	36657876	36678735	TRX2	1.70E-05
xII	ZNF297	AC092295.2.1.165566	19q13.12	37465837	37479151	EST gene	1.00E-10
III	C6orf46	AC008806.4.1.135173	19q13.13	38293492	38349772	NM_152484	5.30E-10
xII	ZNF297	AC022148.5.1.198751	19q13.13	38430631	38431506	Q8N3U1	1.00E-10
III	NOTCH4	AC011500.7.1.200430	19q13.2	40327364	40358466	SUPT5H	1.10E-27
III	NOTCH4	AC010412.8.1.155085	19q13.2	41494937	41527447	LTBP4	1.50E-33
III	CYP21A2	AC008537.5.1.169089	19q13.2	41988964	41996369	CYP2A6	1.60E-09
I	DDR1	AC011510.7.1.129402	19q13.2	42116547	42159395	AXL	5.40E-09
III	CYP21A2	AC008962.9.1.154169	19q13.2	42259850	42273778	CYP2F1	2.90E-08
xII	B3GALT4	AC011526.7.1.40887	19q13.2	42323217	42324407	Novel	6.40E-20
III	CYP21A2	AC011510.7.1.129402	19p13.2	42338667	42352612	CYP2S1	9.30E-06
I	POU5F1	AC024076.4.1.39443	19q13.2	42986837	43028331	POU2F2	4.30E-23
III	NOTCH4	AC011497.6.1.168586	19q13.2	43248294	43273290	EGFL4	4.00E-09
I	DHX16	AC008754.8.1.66792	19p13.32	48246893	48270700	DDX34	7.80E-27
I	DHX16	AC073548.4.1.66051	19q13.32	48322806	48366009	SLC8A2	2.20E-17
xI	HLA Class I	AC010619.7.1.179394	19q13.33	50384629	50397727	FCGRT	8.50E-10
xII	RXRB	AC008655.7.1.123149	19p13.33	51241040	51247541	NR1H2	1.10E-06
xI	MAS1L	AC005946.1.1.37392	19q13.33	52688362	52689423	FPRL2	9.40E-06
III	C6orf46	AC010320.9.1.220458	19q13.41	53262462	53283017	Q96JK0	3.60E-15
III	C6orf46	AC022150.6.1.228156	19q13.41	53461403	53462023	ZNF137	9.00E-16
xII	ZNF297	AC013256.1.1.36095	19q13.43	57406251	57442939	NM_022103	1.40E-10
xII	ZNF297	AC005498.1.1.37321	19q13.43	57504539	57522397	ZFP28	2.80E-10
xII	ZNF297	AC003682.1.1.153875	19q13.43	58536894	58544255	Q9BWM5	1.10E-11
III	C6orf46	AC003682.1.1.153875	19q13.43	58579587	58587258	ZNF134	1.60E-15
III	C6orf46	AC003006.1.1.84114	19q13.43	58734983	58745942	NM_017652	5.20E-12
III	C6orf46	AC012313.7.1.185417	19q13.43	59398153	59405560	ZNF132	1.70E-25
xII	ZNF297	AC012313.7.1.185417	19q13.43	59478868	59485159	NM_032792	5.20E-29
xI	NOL5B	AL049712.12.1.159272	20p13	2580791	2587039	NOL5A	4.00E-29

III	NOTCH4	AL035456.26.1.125952	20p12.2	10566334	10602636	JAG1	1.30E-49
xII	KIFC1	AL049794.16.1.119696	20p12.1	16200749	16502021	C20ORF23	1.30E-09
III	NOTCH4	AL049651.2.1.97912	20p11.21	22964121	22965287	SSTR4	4.00E-10
III	EGFL8	AL118508.27.1.123832	20p11.21	23048052	23055034	C1QR1	1.90E-05
III	NOTCH4	AL118508.27.1.123832	20p11.21	23054616	23054911	Q8WY72	1.90E-16
xI	BTN1A1	AL080312.14.1.94664	20p11.21	25027858	25028403	genscan	1.60E-08
I	DDR1	AL049539.21.1.111694	20q11.21	30388101	30437940	HCK	6.80E-11
xII	KIFC1	AL121897.32.1.145414	20q11.21	30613467	30669435	KIF3B	3.10E-46
I	DDR1	AL133293.28.1.68662	20q11.23	35700500	35722250	SRC	3.00E-15
I	DHX16	AL023803.3.1.155379	20q11.23	37279429	37356793	DDX35	2.10E-19
xII	RXRB	AL132772.14.1.83798	20q13.12	42718338	42747410	HNFA4	4.50E-26
xII	ZNF297	AL354745.11.1.13535	20q13.12	44818128	44830619	ZNF334	1.60E-05
III	BAT1	AL049766.14.1.110293	20q13.13	47524305	47549031	DDX27	1.70E-10
I	TUBB	AL109840.24.1.142094	20q13.32	57282669	57290069	TUBBL	3.02E-165
III	NOTCH4	AL354836.13.1.141056	20q13.33	60601582	60607445	ADRM1	2.50E-09
III	NOTCH4	AL121673.41.1.151163	20q13.33	61421476	61437370	C20orf59	2.90E-07
I	DDR1	AL121829.30.1.113196	20q13.33	61996950	62006900	PTK6	3.20E-14
xII	ZNF297	AL121845.20.1.120917	20q13.33	62212439	62299987	Novel	3.90E-12
III	BAT5	AL118506.27.1.139505	20q13.33	62330271	62331761	C20ORF135	1.50E-117
III	HSPA1L	AF130358.2.1.197778	21q11.2	12307991	12372209	ABCC13	1.80E-44
III	HSPA1L	AF130249.1.1.97083	21q11.2	12405307	12417341	STCH	3.30E-46
III	CLIC1	AP000330.2.1.170377	21q22.12	32702115	32750955	CLIC6	2.60E-38
xII	B3GALT4	AF064860.2.1.170121	21q22.2	37690022	37690381	genscan	9.90E-32
III	BAT8	AP001615.1.1.124516	21q22.3	39783391	39784221	genscan	4.60E-17
xII	ZNF297	AP001620.1.1.95449	21q22.3	40039452	40061155	ZNF295	5.40E-10
xI	HIST1H2AC	AB001523.1.1.122638	21q22.3	42024993	42118907	TMEM1	7.20E-08
III	NOTCH4	AP001067.1.1.148845	21q22.3	42510546	42724266	C21orf29	2.10E-24
III	NOTCH4	AL163301.2.1.340000	21q22.3	43278171	43300571	C21orf80	5.40E-06
xII	COL11A2	AL163302.2.1.340000	21q22.3	43527445	43550695	SLC19A1	9.70E-12
III	BAT1	AP001604.1.1.186930	21q21.3	25401349	25401540	genscan	4.90E-09
I	TUBB	AC008079.23.1.170102	22q11.21	15544495	15554618	TUBA8	4.00E-10
xI	POM121L2	AC008103.27.1.98557	22q11.21	15773964	15776531	C22.2	5.60E-08
xI	POM121L2	AC000095.3.1.43728	22q11.21	15945799	15947779	C22.3	3.00E-10
III	NOTCH4	AC005500.2.1.192592	22q11.21	17480798	17552052	SRC2	1.00E-09
III	NOTCH4	AC007731.14.1.182617	22q11.21	17480798	17552052	SRC2	2.60E-09
xI	POM121L2	AC007050.25.1.163908	22q11.21	17742343	17744900	C22.3	1.50E-09
III	C6orf46/ZNF297	AP000557.2.1.150036	22q11.21	18470386	18504441	H1C2	6.00E-19
II	HLA-DRB3/1	D87023.1.1.40392	22q11.22	19936243	19936975	IGLC1	2.80E-06
xI	POM121L2	AP000354.1.1.164756	22q11.23	21343934	21357627	NM_014549	9.60E-10
xI	POM121L2	AP000356.1.1.163795	22q11.23	21749568	21750854	POM121L1	1.70E-10
I	DDR1	AL022329.9.1.221507	22q12.1	22656954	22816015	ADRBK2	6.00E-21
xI	BTN1A1/RNF	AC002059.3.1.173029	22q12.2	26530668	26534540	RFPL1	2.40E-26
III	RNF5	AC002073.1.1.128978	22q12.2	28252236	28299047	Q96GF1	8.10E-20
xI	HIST1H2AC	AL096701.14.1.168110	22q12.2	28613161	28613624	novel	2.50E-07
xI	BTN1A1/RFP	AL008723.8.1.154414	22q12.3	29282473	29295511	RFPL2	5.70E-26
xI	BTN1A1/RFP	AL021937.1.1.173354	22q12.3	29447342	29453195	RFPL3	5.70E-26
III	BAT1	Z97056.1.1.124990	22q13.1	35496212	35516829	DDX17	2.00E-05
II	BRD2	AL096765.12.1.13053	22q13.2	38102320	38190075	EP300	1.20E-05
III	CYP21A2	AL021878.1.1.114847	22q13.2	39138588	39142847	CYP2D6	3.30E-09
III	NOTCH4	Z98047.1.1.47542	22q13.31	42534087	42532337	FBLN1	5.30E-12
III	NOTCH4	AL031588.1.1.127168	22q13.31	43322995	43499263	CELSR1	1.20E-09
III	BAT1	AC117517.7.1.121628	Xp22.11	21589340	21590458	genscan	4.40E-37
xI	HIST1H2AC	AL121577.1.1.175531	Xq21.1	35832664	35878934	XK	2.40E-08
xI	HIST1H2AC	AL121578.1.1.337101	Xp11.4	36141216	36141440	genscan	2.80E-08

III	CLIC1	AL391259.15.1.163520	Xp11.4	38963014	38963730	Genscan	2.90E-07
III	BAT1	AL391647.16.1.60310	Xp11.4	39441330	39472404	DDX39	6.20E-07
xII	ZNF297	AL590223.12.1.40331	Xp11.3	45567068	45603010	ZNF41	4.60E-10
xII	ZNF297	Z98304.1.1.209618	Xp11.23	46096233	46103478	Q96QH7	4.70E-10
III	BAT8	AC115618.1.1.158455	Xp11.23	46756067	46756093	no gene	3.40E-15
III	BAT8	AF196970.1.1.112595	Xp11.23	46815791	46828063	SUV39H1	1.30E-15
III	BAT1	AL445236.22.1.149749	Xp11.22	50567899	50607906	Novel	9.90E-39
xII	HSD17B8	Z97054.1.1.132805	Xp11.22	51162495	51165605	HADH2	3.60E-06
xII	KIFC1	AL357752.19.1.178868	Xq13.1	66987254	67117990	KIF4A	2.70E-08
III	BAT1	AL359740.24.1.98104	Xq13.2	70525227	70526105	genscan	1.00E-13
II	TAP2/1	AL359545.12.1.127243	Xq13.3	71447688	71550705	ABCB7	2.90E-08
xII	KIFC1	AL021786.2.1.70665	Xq21.1	75540478	75573738	Novel	1.40E-15
I	POU5F1	Z82170.1.1.127247	Xq21.1	79839811	79841286	POU3F4	6.20E-39
III	BAT1	AL136362.10.1.135240	Xq21.31	88337615	88338796	EST gene	1.30E-57
I	TUBB	AL390840.17.1.197611	Xq21.32	88819807	88819917	genscan	1.50E-09
xII	COL11A2	AL136080.6.1.116106	Xq23	104474931	104758796	COL4A6	2.50E-15
xII	COL11A2	AL031622.1.1.104674	Xq23	104759239	105016860	COL4A5	1.80E-17
III	HSPA1L	AC004822.1.1.127824	Xq23	111134972	111136228	genscan	1.30E-208
III	VARS2	AC005000.2.1.107314	Xq23	112024040	112063337	Novel	3.30E-05
I	TUBB	AC003012.1.1.104810	Xq24	112252763	112252876	genscan	4.60E-12
xII	ZNF297	AC002086.1.1.112686	Xq24	116370267	116377851	NM_006777	6.50E-09
III	HSPA1L	AC002377.1.1.141779	Xq24	117230333	117231259	genscan	1.50E-221
III	HSPA1L	AL391241.21.1.157860	Xq25	120232224	120232373	genscan	9.40E-65
III	NOTCH4	AL627231.9.1.146366	Xq25	121306552	121307673	Novel	6.10E-12
xII	ZNF297	AL590282.6.1.139296	Xq26.3	131228672	131323794	ZNF75	3.70E-10
III	C6orf46	U82670.3.1.279526	Xq28	149081175	149084483	ZNF275	5.30E-12
III	CLIC1	AL356738.14.1.174693	Xq28	150871755	150929271	CLIC2	8.80E-52
xI	HIST1H2AC	AC019175.4.37111.45694	Xq28	151078382	151078898	H2AFB	1.80E-13
xI	HIST1H2AC	AL592156.4.1.134995	Xq21.1	35423125	35423349	genscan	2.00E-11
III	BAT1	AC010129.3.1.44145	Yp11.2	5171386	5172558	Novel	6.30E-59
III	BAT1	AC004474.1.1.148280	Yq11.21	14326902	14356562	DBY	4.00E-05

Appendix 3

Primers used to amplify a paralogue specific probe for use in Northern blot, Dot blot and Southern blot analyses. 'T' stands for the annealing temperature.

Gene	*Primer*	*Sequence*	*T (°C)*	*Size (bp)*
AIF1	F	TGACCATGCTGATGTATGAGGAAAAAGCGA	62	200
	R	GATCTGGAGGAGGGGGTAAT		
AIF1-L	F	TGACCATGTTAAGGGAGGAGAGCAAGCA	62	251
	R	CTGAGCCCTTAGCCAGAGAA		
BRD2	F	TGACCATGGAGGGATGCAGGGACATTT	62	411
	R	AACAAAGACAGTCCAGGGGA		
BRDT	F	TGACCATGGGGTACCATTGATATGACCCTT	62	199
	R	CTGTTTAATCATTTTAGAGCAGTCA		
BRD3	F	TGACCATGGACAGATGGATGTCGCACAC	62	425
	R	CAAATGACAAGGACAATGCG		
BRD4	F	TGACCATGGTGAAAGGGACAGGACTCCA	65	508
	R	CAGTGAGAAGCATGCTGTGG		
C4	F	TGACCATGAGAGATGACTCCGCGTCTGT	65	395
	R	ATTCTCCTTCTGCCCCAGAT		
C3	F	TGACCATGCATTCCCCCACTCCAGATAA	65	214
	R	ACATGAAGGTGAGGCAGGTC		
C5	F	TGACCATGTTGCACTTATGGACTCCTGTTG	65	352
	R	GATCAGTTTCCTGTTCCTTGGT		
CLIC1	F	TGACCATGAAGTACCGGGGATTCACCAT	65	310
	R	CTTTCCCTCATCCCCTCTTC		
CLIC4	F	TGACCATGGGAGATTGGAGTCTGAATGGA	65	384
	R	AATGGGTTTAAGGGCACAGA		
CLIC3	F	TGACCATGGTACGCCGCTACCTGGAC	65	153
	R	CCCGACAAAGATGCCTTTATT		
CLIC5	F	TGACCATGTGTTGATGCCAAAATACCCA	65	427
	R	GACCACCTCCTAAATGTGGC		
CLIC6	F	TGACCATGTGTGGCCAAGAAGTACAGAGAT	65	146
	R	TTGCAACATCTGAATATGCG		
CLIC2	F	TGACCATGGAATTCTCAGGAGTCTGGCG	65	350
	R	GCAGTGGTTTGCCATACAGA		
GPX5	F	TGACCATGTAGCAATGGGGTCACAGTCA	65	277
	R	TCCTCTCCAGGTGCCATAAC		
GPX4	F	TGACCATGTCCACAAGTGTGTGGCCC	65	186
	R	CACAAGGTAGCCAGGGGTG		
GPX3	F	TGACCATGAACCCAAAGGAAAAACCAGC	62	451
	R	GAGTCTCAAGCCAGTGGACC		
GPX1	F	TGACCATGCTCTTCGAGAAGTGCGAGGT	65	439
	R	ACTGGGATCAACAGGACCAG		
GPX2	F	TGACCATGTCCTACTCCATCCAGTCCTGA	62	256
	R	CTTCACGCCTCTCAGACACC		
NOTCH4	F	TGACCATGCATTAAAAGGCAGGCTGGAA	65	475
	R	CATCCCCACAGTGGAGTTCT		
NOTCH2	F	TGACCATGATGAGGAGGACAACACTGCC	65	395
	R	GCATCACAGCCAATTGCTTA		
NOTCH1	F	TGACCATGCAATACTGCATCCATGGCCT	65	244

	R	GTCCCTGAGCAACCATCTGT		
NOTCH3	F	TGACCATGATGTTCCATAGCCTTGCTGG	65	294
	R	GGGAATTCAGCTACACAGGG		
PBX2	F	TGACCATGGCAGGGCTGGACTCAGTAAT	62	409
	R	CACTTCCAACCTGTCCCAGT		
PBX1	F	TGACCATGCAGGAGGGAGGGTTTCTCTC	62	267
	R	TCAGTGATATGAGAGAGGGCG		
PBX3	F	TGACCATGCGAGTGTGGAAACATTGGGT	62	325
	R	TCAATCCAGGGTGTAATCCA		
PBX4	F	TGACCATGGTTTGGGGGATAAGCAGGAA	62	286
	R	GAAAATCTGTGCCCAGTCCT		
RXRB	F	TGACCATGAAGAAATGCCAGTGGTGGAG	62	263
	R	AAAGGAGCCCCAAAGAGAAG		
RXRG	F	TGACCATGTCCTGACTAATCCCAGAGGG	62	215
	R	CATAGCCTGCGGGAAACTT		
RXRA	F	TGACCATGTATACTTGGATATGGCGGGG	65	299
	R	CGGAGAAGCCACTTCACAGT		
TUBB_6p21.3	F	TGACCATGAGAGCAACATGAACGACCTG	65	200
	R	TGGAGGGAGATTGAAAGTGG		
TUBB2_18p11.3	F	TGACCATG TTCCTTCTTGAACCCTGGTG	65	225
	R	TTTATTTTGTGGCCCCTCAG		
TUBB5_19p13.3	F	TGACCATGCTGAATCCCCTCTGACTCCA	65	293
	R	CCTCTCTTCCTCACAGGCAC		
TUBB4QL_10p15.3	F	TGACCATGACAGCATCTGGTTTTGCCTC	65	130
	R	CCACTGGAATGCTTGTTCCT		
TUBB4_16q24.3	F	TGACCATGCAGCTGGAGTGAGAGGCAG	65	201
	R	GCCTGGAGCTGCAATAAGAC		
TUBB1_20q13.3	F	TGACCATGTGCACTCACCATTAGCTTCG	65	396
	R	TAGTCAGGCACCTGGCTCTT		

Appendix 4

Primers used to generate paralogue specific PCR products for each paralogue. The products were used to spot on to the microarrays and were also labelled and used to hybridise to the 'Paralogue Microarray'. 'T' stands for the annealing temperature. They were also used in the RT-PCR experiments.

Gene	*Primer*	*Sequence*	*T (°C)*	*Size (bp)*
AIF1	F	TGACCATGCTGATGTATGAGGAAAAAGCGA	62.5	200
	R	GATCTGGAGGAGGGGGTAAT		
AIF1-L	F	TGACCATGTTAAGGGAGGAGAGCAAGCA	62.5	251
	R	CTGAGCCCTTAGCCAGAGAA		
BRD2	F	TGACCATGGAGGGATGCAGGGACATTT	62.5	411
	R	AACAAAGACAGTCCAGGGGA		
BRDT	F	TGACCATGGGGTACCATTGATATGACCCTT	62.5	199
	R	CTGTTTAATCATTTTAGAGCAGTCA		
BRD3	F	TGACCATGGACAGATGGATGTCGCACAC	62.5	425
	R	CAAATGACAAGGACAATGCG		
BRD4	F	TGACCATGGTGAAAGGGACAGGACTCCA	65	508
	R	CAGTGAGAAGCATGCTGTGG		
C4	F	TGACCATGAGAGATGACTCCGCGTCTGT	65	395
	R	ATTCTCCTTCTGCCCCAGAT		
C3	F	TGACCATGCATTCCCCCACTCCAGATAA	65	214
	R	ACATGAAGGTGAGGCAGGTC		
C5	F	TGACCATGTTGCACTTATGGACTCCTGTTG	65	352
	R	GATCAGTTTCCTGTTCCTTGGT		
CLIC1	F	TGACCATGAAGTACCGGGGATTCACCAT	62.5	310
	R	CTTTCCCTCATCCCCTCTTC		
CLIC4	F	TGACCATGGGAGATTGGAGTCTGAATGGA	65	384
	R	AATGGGTTTAAGGGCACAGA		
CLIC3	F	TGACCATGGTACGCCGCTACCTGGAC	65	153
	R	CCCGACAAAGATGCCTTTATT		
CLIC5	F	TGACCATGTGTTGATGCCAAAATACCCA	65	427
	R	GACCACCTCCTAAATGTGGC		
CLIC6	F	TGACCATGTGTGGCCAAGAAGTACAGAGAT	65	146
	R	TTGCAACATCTGAATATGCG		
CLIC2	F	TGACCATGGAATTCTCAGGAGTCTGGCG	65	350
	R	GCAGTGGTTTGCCATACAGA		
GPX5	F	TGACCATGTAGCAATGGGGTCACAGTCA	62.5	277
	R	TCCTCTCCAGGTGCCATAAC		
GPX4	F	TGACCATGTCCACAAGTGTGTGGCCC	62.5	186
	R	CACAAGGTAGCCAGGGGTG		
GPX3	F	TGACCATGTCTGGGTCTACCACACTCCC	62.5	329
	R	GAGTCTCAAGCCAGTGGACC		
GPX1	F	TGACCATGCTCTTCGAGAAGTGCGAGGT	62.5	439
	R	ACTGGGATCAACAGGACCAG		
GPX2	F	TGACCATGTCCTACTCCATCCAGTCCTGA	62.5	256
	R	CTTCACGCCTCTCAGACACC		
NOTCH4	F	TGACCATGCATTAAAAGGCAGGCTGGAA	62.5	475
	R	CATCCCCACAGTGGAGTTCT		
NOTCH2	F	TGACCATGATGAGGAGGACAACACTGCC	65	395
	R	GCATCACAGCCAATTGCTTA		

NOTCH1	F	TGACCATGCAATACTGCATCCATGGCCT	65	244
	R	GTCCCTGAGCAACCATCTGT		
NOTCH3	F	TGACCATGATGTTCCATAGCCTTGCTGG	65	294
	R	GGGAATTCAGCTACACAGGG		
PBX2	F	TGACCATGGCAGGGCTGGACTCAGTAAT	62.5	409
	R	CACTTCCAACCTGTCCCAGT		
PBX1	F	TGACCATGCAGGAGGGAGGGTTTCTCTC	62.5	267
	R	TCAGTGATATGAGAGAGGGCG		
PBX3	F	TGACCATGACCGAGTGTGGAAACATTGG	62.5	328
	R	TTCAATCCAGGGTGTAATCCA		
PBX4	F	TGACCATGAAGTTTGGGGGATAAGCAGG	62.5	288
	R	GAAAATCTGTGCCCAGTCCTA		
RXRB	F	TGACCATGGCCTTCCTCCTCTCAAACCT	62.5	263
	R	CTCCACCACTGGCATTTCTT		
RXRG	F	TGACCATGCGATCTAGAGGCAGATTCCTGA	62.5	231
	R	CATAGCCTGCGGGAAACTT		
RXRA	F	TGACCATGTATACTTGGATATGGCGGGG	65	299
	R	CGGAGAAGCCACTTCACAGT		
TUBB_6p21.3	F	TGACCATGACCAACCAGGTGCTGAAAAC	65	242
	R	TGGAGGGAGATTGAAAGTGG		
TUBB2_18p11.3	F	TGACCATG TTCCTTCTTGAACCCTGGTG	65	225
	R	TTTATTTTGTGGCCCCTCAG		
TUBB5_19p13.3	F	TGACCATGCTGAATCCCCTCTGACTCCA	62.5	293
	R	CCTCTCTTCCTCACAGGCAC		
TUBB4QL_10p15.3	F	TGACCATGACAGCATCTGGTTTTGCCTC	65	130
	R	CCACTGGAATGCTTGTTCCT		
TUBB4_16q24.3	F	TGACCATGCAGCTGGAGTGAGAGGCAG	65	201
	R	GCCTGGAGCTGCAATAAGAC		
TUBB1_20q13.3	F	TGACCATGTGCACTCACCATTAGCTTCG	65	396
	R	TAGTCAGGCACCTGGCTCTT		

Appendix 5

Summary of *in-silico* results.

Gene	***Brain(whole)***	***Ear***	***Eye***	***Nervous_normal***	***Heart***	***Aorta***	***Pharynx***	***Oesophagus***	***Stomach***	***Liver***	***Pancreas***	***Intestine***	***Colon***	***Gallbladder***	***Kidney***	***Bladder***	***Prostate***	***Genitourinary***	***Endometrium***	***Uterus***	***Cervix***	***Cervical carcinoma cell-line, Hela S3***	***Ovary***	***Breast***
AIF1_6p21.33	0	0	1	1	1	1	0	0	1	1	0	0	1	0	1	0	1	0	0	1	0	0	1	1
AIF1-L_9q34.12	1	0	1	1	1	0	0	0	1	1	1	0	1	0	1	0	1	0	0	1	1	0	1	1
BRD2_6p21.32	1	0	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1
BRDT_1p22.1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BRD3_9q34.2	1	0	1	1	0	1	0	0	1	0	1	1	1	0	1	0	0	0	0	1	1	0	1	1
BRD4_19p13.12	1	0	1	1	1	1	0	1	1	1	1	0	1	0	1	1	1	0	0	1	1	0	1	1
C4_6p21.33	1	0	1	1	1	0	0	1	1	1	0	0	1	1	1	1	1	0	1	0	0	0	1	1
C5_9q33.2	1	0	1	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	0	0	1	0	0	0
C3_19p13.3	1	0	1	1	1	0	0	0	1	1	1	0	1	1	1	0	1	0	1	1	1	1	1	1
CLIC1_6p21.33	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1
CLIC4_1p35.3	1	1	1	1	1	1	0	0	1	1	1	1	0	1	1	1	1	0	0	1	0	0	1	1
CLIC3_9q34.3	1	1	0	0	1	0	0	0	0	1	1	0	1	0	1	1	1	0	0	1	1	0	0	0
CLIC5_6p21.1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	1
CLIC6_21q22.12	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
CLIC2_Xq28	1	0	0	0	0	0	0	0	0	1	1	0	0	0	1	1	1	0	0	1	0	0	1	1
GPX5_6p22.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GPX4_19p13.3	1	0	1	1	1	1	0	0	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1
GPX1_3p21.31	1	0	1	1	1	1	0	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1
GPX3_5q33.1	1	0	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1
GPX2_14q23.3	1	0	0	0	0	0	0	1	1	1	1	0	1	1	1	1	1	0	0	1	0	0	1	1
NOTCH4_6p21.33	1	0	0	1	1	0	0	0	0	1	1	0	1	0	1	0	1	0	0	1	0	0	1	1
NOTCH2_1p11.2	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1
NOTCH1_9q34.3	1	0	0	1	1	0	0	1	1	0	1	1	1	0	1	0	0	0	0	1	0	0	0	1
NOTCH3_19p13.12	1	1	1	1	1	0	0	1	1	1	1	1	1	0	1	0	1	1	0	1	0	0	1	1
PBX2_6p21.33	0	0	1	1	0	0	0	0	1	1	1	1	1	0	1	0	1	0	0	1	0	1	1	0
PBX1_1q23.3	1	0	1	0	1	0	0	0	1	1	0	0	1	0	1	0	1	1	0	1	0	0	1	1
PBX3_9q33.3	1	1	1	1	1	0	0	0	1	1	1	1	1	0	1	1	1	0	0	0	0	1	1	1
PBX4_19p13.11	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
RXRB_6p21.32	1	0	1	1	1	0	0	0	1	1	1	1	1	0	1	1	1	0	1	1	1	0	1	1
RXRG_1q23.3	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
RXRA_9q34.2	1	0	1	1	1	0	0	0	1	1	1	0	1	0	1	1	1	0	0	1	1	0	1	1
TUBB_6p21.3	1	1	1	1	1	1	0	0	1	1	1	1	1	0	1	1	1	0	1	1	1	0	1	1
TUBB4_16q24.3	1	0	1	1	1	0	0	0	1	1	1	0	1	0	1	0	0	0	0	1	0	0	1	1
TUBBL_18p11.3	1	1	1	0	1	1	0	0	1	1	1	0	0	0	1	0	1	0	1	1	1	0	0	1
TUBB5_19p13.3	1	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1
TUBB1_20q13.3	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0

In-silico results continued (part 2 of 3).

Gene	***Testis***	***Epididymis***	***Placenta***	***Germ cell***	***Amnion_normal***	***Spleen***	***Thymus***	***Leukocyte***	***Lymph node***	***Lymphatic***	***Bone marrow***	***B cell***	***T cell***	***Macrophage***	***Monocyte***	***Blood***	***Nose***	***Trachea***	***Lung***	***Adrenal gland***	***Parathyroid***	***Thyroid gland***	***Pineal***	***Pituitary***
AIF1_6p21.33	1	0	1	0	0	1	1	0	0	0	1	0	0	0	0	1	0	0	1	1	1	0	0	0
AIF1-L_9q34.12	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0
BRD2_6p21.32	1	0	1	0	0	1	0	0	0	1	1	1	1	1	0	1	0	0	1	1	1	1	1	1
BRDT_1p22.1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BRD3_9q34.2	1	0	0	0	1	0	0	0	0	1	1	1	1	0	0	0	0	0	1	1	0	0	0	0
BRD4_19p13.12	1	1	1	1	0	1	1	1	0	1	1	1	1	0	0	1	1	0	1	1	1	1	0	0
C4_6p21.33	0	1	1	1	0	1	0	0	1	0	1	1	1	0	0	1	1	0	1	1	1	0	0	1
C5_9q33.2	0	0	1	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0
C3_19p13.3	1	1	1	1	0	1	0	0	1	0	1	0	1	1	0	0	1	0	1	1	1	1	0	0
CLIC1_6p21.33	1	1	1	1	1	1	1	0	1	0	1	1	1	0	0	1	0	0	1	1	1	1	0	1
CLIC4_1p35.3	1	0	1	1	1	0	0	0	1	0	1	1	1	0	0	1	0	0	1	1	1	1	1	0
CLIC3_9q34.3	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
CLIC5_6p21.1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
CLIC6_21q22.12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
CLIC2_Xq28	0	0	1	1	0	1	0	0	1	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0
GPX5_6p22.1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GPX4_19p13.3	1	1	1	1	0	0	0	1	1	0	1	1	1	0	0	1	0	0	1	1	1	0	1	0
GPX1_3p21.31	1	1	1	1	0	1	1	1	1	0	1	1	1	0	0	1	1	0	1	1	1	1	1	1
GPX3_5q33.1	1	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	1
GPX2_14q23.3	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
NOTCH4_6p21.33	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
NOTCH2_1p11.2	1	1	1	1	1	0	0	0	1	0	1	1	1	1	1	1	1	0	1	1	1	0	1	0
NOTCH1_9q34.3	1	0	1	1	0	0	0	1	1	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0
NOTCH3_19p13.12	1	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0	1	0	0	0	0	0
PBX2_6p21.33	1	1	1	1	0	0	1	1	1	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0
PBX1_1q23.3	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
PBX3_9q33.3	1	0	0	1	0	0	0	1	1	0	0	1	0	0	0	0	0	0	1	1	1	1	0	1
PBX4_19p13.11	1	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
RXRB_6p21.32	1	1	1	1	0	1	0	1	1	0	0	1	0	0	0	1	0	0	1	0	1	1	0	0
RXRG_1q23.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
RXRA_9q34.2	1	0	1	0	0	0	0	1	1	0	1	1	0	0	0	0	0	0	1	0	1	0	0	0
TUBB_6p21.3	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	1	1	0	1	1	1
TUBB4_16q24.3	1	0	1	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	1	0	0	0	1
TUBBL_18p11.3	0	0	1	1	0	1	0	0	1	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0
TUBB5_19p13.3	1	1	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0
TUBB1_20q13.3	1	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0

In-silico results continued (part 3 of 3).

Gene	***Salivary gland***	***Mammary gland***	***Skin***	***Bone***	***Adipose***	***Connective***	***Fibroblast***	***Cartilage***	***Muscle***	***Tongue***	***Synovial membrane***	***Mixed***	***Unknown***
AIF1_6p21.33	0	0	1	1	0	0	0	0	1	0	0	1	1
AIF1-L_9q34.12	0	0	1	1	0	0	0	0	1	0	0	1	1
BRD2_6p21.32	0	0	1	1	1	1	1	0	1	1	1	1	1
BRDT_1p22.1	0	0	0	0	0	0	0	0	0	0	0	1	1
BRD3_9q34.2	0	0	1	1	0	0	0	1	1	0	0	1	1
BRD4_19p13.12	1	0	0	1	1	0	1	1	1	1	0	1	1
C4_6p21.33	0	0	1	0	1	0	0	1	0	1	0	1	1
C5_9q33.2	0	0	1	0	0	0	0	1	0	1	0	1	1
C3_19p13.3	0	0	1	1	1	0	1	1	1	1	0	1	1
CLIC1_6p21.33	0	0	1	1	1	0	1	1	1	0	0	1	1
CLIC4_1p35.3	0	0	1	1	1	0	1	1	1	0	0	1	1
CLIC3_9q34.3	0	0	0	0	0	0	1	0	0	0	0	1	0
CLIC5_6p21.1	0	0	0	0	0	0	0	0	0	0	0	1	1
CLIC6_21q22.12	0	0	0	1	0	0	0	0	0	0	0	1	1
CLIC2_Xq28	0	0	0	1	0	0	0	0	1	0	0	1	0
GPX5_6p22.1	0	0	0	0	0	0	0	0	0	0	0	0	0
GPX4_19p13.3	0	0	1	1	1	0	1	1	1	0	0	1	1
GPX1_3p21.31	1	0	1	1	0	0	1	1	1	1	1	1	1
GPX3_5q33.1	1	0	1	0	1	0	0	1	1	1	1	1	1
GPX2_14q23.3	0	0	1	0	0	0	0	0	0	0	0	1	1
NOTCH4_6p21.33	0	0	1	0	0	0	0	0	0	0	0	1	1
NOTCH2_1p11.2	0	0	1	1	0	1	1	1	1	0	1	1	1
NOTCH1_9q34.3	0	0	0	0	0	0	0	0	0	0	0	1	1
NOTCH3_19p13.12	0	0	1	0	0	0	1	0	0	1	0	1	1
PBX2_6p21.33	0	0	1	0	0	0	0	0	1	0	0	1	1
PBX1_1q23.3	0	0	0	0	0	0	0	1	0	0	0	1	0
PBX3_9q33.3	0	0	0	1	0	0	1	1	0	0	0	1	1
PBX4_19p13.11	0	0	1	0	0	0	0	0	0	0	0	1	0
RXRB_6p21.32	0	0	1	0	0	0	1	1	0	0	0	1	1
RXRG_1q23.3	0	0	1	0	0	0	0	0	1	0	0	1	1
RXRA_9q34.2	0	0	1	1	0	0	1	1	1	0	0	1	1
TUBB_6p21.3	0	1	1	1	1	0	1	1	1	0	1	1	1
TUBB4_16q24.3	0	0	1	1	0	0	1	1	0	0	0	1	1
TUBBL_18p11.3	0	1	1	1	1	0	1	0	1	0	0	1	1
TUBB5_19p13.3	0	1	0	0	0	0	0	0	0	0	0	1	1
TUBB1_20q13.3	0	0	0	0	0	0	0	0	0	0	0	1	1

Appendix 6

Summary of dot blot results

Gene	*Brain*	*Cerebral cortex*	*Frontal lobe*	*Parietal lobe*	*Occipital lobe*	*Temporal lobe*	*Paracentral gyrus of cerebral cortex*	*Pons*	*Cerebellum, left*	*Cerebellum, right*	*Corpus callosum*	*Amygdala*	*Caudate nucleus*	*Hippocampus*	*Medulla oblongata*	*Putamen*	*Accumbens nucleus*	*Thalamus*	*Heart*	*Aorta*	*Atrium, left*	*Atrium, right*	*Ventricle, left*	*Ventrcle, right*
AIF1_6p21.33	0	0	0	0	1	0	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
AIF1-L_9q34.12	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BRD2_6p21.32	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BRDT_1p22.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BRD3_9q34.2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BRD4_19p13.12	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C4_6p21.33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C5_9q33.2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1
C3_19p13.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
CLIC1_6p21.33	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
CLIC4_1p35.3	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1
CLIC3_9q34.3	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
CLIC5_6p21.1	0	1	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0
CLIC6_21q22.12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CLIC2_Xq28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GPX5_6p22.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GPX4_19p13.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
GPX1_3p21.31	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
GPX3_5q33.1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
GPX2_14q23.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1
NOTCH4_6p21.33	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	1	1	1
NOTCH2_1p11.2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
NOTCH1_9q34.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
NOTCH3_19p13.12	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
PBX2_6p21.33	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
PBX1_1q23.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
PBX3_9q33.3	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
PBX4_19p13.11	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
RXRB_6p21.32	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
RXRG_1q23.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RXRA_9q34.2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
TUBB_6p21.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
TUBB4QL_10p15.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0
TUBB4_16q24.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
TUBBL_18p11.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
TUBB5_19p13.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1
TUBB1_20q13.3	0	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0

Dot blot results continued (2 of 3).

Gene	*Interventricular septum*	*Apex of heart*	*Oesophagus*	*Stomach*	*Duodenum*	*Jejunum*	*Ileum*	*Ileocecum*	*Appendix*	*Colon, ascending*	*Colon, transverse*	*Colon, descending*	*Rectum*	*Kidney*	*Skeletalmuscle*	*Spleen*	*Thymus*	*Pripheral blood leukocyte*	*Lymph node*	*Bone marrow*	*Trachea*	*Lung*	*Placenta*	*Bladder*
AIF1_6p21.33	1	1	0	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1
AIF1-L_9q34.12	1	1	1	0	0	1	0	1	1	1	0	0	1	1	0	1	0	1	1	0	1	0	1	0
BRD2_6p21.32	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BRDT_1p22.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BRD3_9q34.2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BRD4_19p13.12	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1
C4_6p21.33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
C5_9q33.2	1	0	0	1	1	1	1	1	1	0	0	0	0	1	1	1	0	0	1	1	0	0	1	1
C3_19p13.3	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1
CLIC1_6p21.33	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
CLIC4_1p35.3	1	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	1	1	0
CLIC3_9q34.3	1	1	1	1	0	1	0	0	0	0	1	0	0	0	1	0	0	1	1	0	0	1	1	1
CLIC5_6p21.1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
CLIC6_21q22.12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
CLIC2_Xq28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GPX5_6p22.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GPX4_19p13.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
GPX1_3p21.31	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
GPX3_5q33.1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
GPX2_14q23.3	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
NOTCH4_6p21.33	1	1	0	1	1	1	1	1	1	0	0	1	1	0	0	0	1	0	1	0	0	1	1	1
NOTCH2_1p11.2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
NOTCH1_9q34.3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
NOTCH3_19p13.12	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
PBX2_6p21.33	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
PBX1_1q23.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	1	1	1	1	1
PBX3_9q33.3	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
PBX4_19p13.11	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	0
RXRB_6p21.32	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
RXRG_1q23.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
RXRA_9q34.2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
TUBB_6p21.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
TUBB4QL_10p15.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TUBB4_16q24.3	1	1	1	1	1	1	1	1	1	0	0	0	0	1	0	0	0	1	1	1	1	1	1	1
TUBBL_18p11.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
TUBB5_19p13.3	1	1	0	1	1	1	1	1	1	0	1	1	1	0	0	0	1	1	1	1	1	0	0	1
TUBB1_20q13.3	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0	1	0

Dot blot results continued (3 of 3).

Gene	***Uterus***	***Prostate***	***Testis***	***Ovary***	***Liver***	***Pancreas***	***Adrenalgland***	***Thyroid gland***	***Salivary gland***	***Leukemia, HL-60***	***HeLa S3***	***Leukemia, HK-562***	***Molt4 (T cell)***	***Raji (B cell)***	***Burkitt's lymphoma, Daudi***	***Colorectal adenocarcinoma***	***Lung carcinoma, A549***	***Foetal brain***	***Foetal heart***	***Foetal kidney***	***Foetal liver***	***Foetal spleen***	***Foetal thymus***	***Foetal lung***
AIF1_6p21.33	1	1	0	0	1	1	1	1	1	0	0	0	1	0	0	0	0	0	1	1	1	1	1	1
AIF1-L_9q34.12	1	1	1	1	0	0	0	1	1	0	0	0	0	0	0	1	0	1	0	1	0	1	0	0
BRD2_6p21.32	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BRDT_1p22.1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BRD3_9q34.2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
BRD4_19p13.12	1	1	1	1	1	1	1	1	1	0	0	0	0	1	0	0	0	1	0	1	1	1	1	1
C4_6p21.33	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C5_9q33.2	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1
C3_19p13.3	1	1	1	0	1	1	1	1	1	0	1	1	1	0	0	0	0	1	1	1	1	1	1	1
CLIC1_6p21.33	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
CLIC4_1p35.3	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CLIC3_9q34.3	1	1	0	0	1	1	0	1	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	1
CLIC5_6p21.1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1
CLIC6_21q22.12	0	0	0	0	0	1	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
CLIC2_Xq28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GPX5_6p22.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GPX4_19p13.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
GPX1_3p21.31	1	1	1	1	1	1	1	1	1	0	1	0	1	1	0	1	0	1	1	1	1	1	1	1
GPX3_5q33.1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1
GPX2_14q23.3	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
NOTCH4_6p21.33	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NOTCH2_1p11.2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
NOTCH1_9q34.3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NOTCH3_19p13.12	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
PBX2_6p21.33	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1
PBX1_1q23.3	1	1	1	1	1	1	1	1	1	0	0	1	0	1	0	0	1	1	1	1	1	1	1	1
PBX3_9q33.3	1	1	1	1	1	1	1	1	1	0	1	1	0	1	0	0	0	1	1	1	1	1	1	1
PBX4_19p13.11	1	1	1	1	1	1	0	1	1	1	1	1	0	0	0	0	0	1	1	1	1	1	1	1
RXRB_6p21.32	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
RXRG_1q23.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RXRA_9q34.2	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1
TUBB_6p21.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
TUBB4QL_10p15.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
TUBB4_16q24.3	1	1	1	1	1	1	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1
TUBBL_18p11.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
TUBB5_19p13.3	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	0	1	0	1	1	1	1	0
TUBB1_20q13.3	0	0	0	0	0	0	0	1	0	1	1	1	0	0	0	0	1	0	0	0	0	0	0	0

Appendix 7

Northern blot results and transcript sizes (kb). The most dominant transcripts are in bold.

Gene	*heart*	*brain*	*placenta*	*lung*	*liver*	*skeletal muscle*	*kidney*	*pancreas*
AIF1_6p21.33	- 3.0 1.25 0.6	- 3.0 - -	- 3.0 1.25 -	- 3.0 - -	- - 1.25 0.6	5.0 3.0 - 0.6	- 3.0 - 0.6	- 3.0 - 0.6
AIF1-L_9q34.12	3.4	3.4	3.4	0	0	0	**3.4**	0
BRD2_6p21.32	**4.6** **3.8**	4.6 3.8	4.6 3.8	4.6 3.8	4.6 3.8	4.6 3.8	4.6 3.8	**4.6** **3.8**
BRDT_1p22.1	7.0 - -	7.0 - -	7.0 - -	0	7.0 - -	7.0 - -	7.0 - -	7.0 4.0 3.5
BRD3_9q34.2	6.5	6.5	6.5	6.5	6.5	**6.5**	6.5	6.5
BRD4_19p13.12	6.0 4.4	6.0 4.4	6.0 4.4	6.0 4.4	6.0 4.4	**6.0** **4.4**	6.0 4.4	6.0 4.4
C4_6p21.33	0	0	0	0	5.4	0	5.4	0
C5_9q33.2	0	0	0	0	6.0 **5.0** 4.2 1.6 1.0	0	- 5.0 - - -	- 5.0 4.2 - -
C3_19p13.3	0	0	0	0	5.0	0	0	0
CLIC1_6p21.33	1.25 1.1	0	**1.25** **1.1**	1.251.1	1.25 1.1	1.25 1.1	1.25 1.1	1.25 1.1
CLIC4_1p35.3	4.0	0	4.0	4.0	4.0	4.0	4.0	0
CLIC3_9q34.3	- 4.4 - -	0	5.5 - 2.6 **0.7**	- - - 0.7	- - - 0.7	- 4.4 - -	- 4.4 - -	- 4.4 - -
CLIC5_6p21.1	0	**2.7** **2.4**	0	2.7 -	2.7 -	0	**2.7** **2.4**	2.7 -
CLIC6_21q22.12	6.0 **3.8** 3.0 2.3	0	0	- 3.8 3.0 -	- - - 0	6.0 **3.8** - -	- - - 0	- - 3.0 -
CLIC2_Xq28	1.7 1.25	0	1.7 1.25	1.7 1.25	1.7 1.25	1.7 1.25	1.7 1.25	1.7 1.25
GPX5_6p22.1	- 0.8 0.6	- 0.8 -	- - 0.6	0	- 0.8 0.6	3.8 0.8 0.6	0	- 0.8 0.6
GPX4_19p13.3	4.4 2.6 2.0 **0.9**	- - - 0.9	- 2.6 2.0 **0.9**	- - - 0.9	4.4 2.6 2.0 **0.9**	4.4 - 2.0 **0.9**	- - 2.0 **0.9**	- - - 0.9
GPX1_3p21.31	**0.8**	**0.8**	**0.8**	**0.8**	**0.8**	**0.8**	**0.8**	**0.8**
GPX3_5q33.1	- 0.7	0	- 0.7	- 0.7	1.8 **0.7**	1.8 0.7	- 0.7	- 0.7
GPX2_14q23.3	0	0	0	0	0	0	**0.9**	0
NOTCH4_6p21.33	7.8 6.8 2.4	- - -	7.5 6.8 -	7.5 6.8 -	- - 2.4	- 6.8 2.4	- 6.8 -	- 6.8 2.4

	- 0.8	1.6 -	- 0.8	- -	- 0.8	- 0.8	- -	1.6 0.8
NOTCH2_1p11.2	**10** 1.9	10 -	**10** 1.9	10 1.9	10 1.9	**10** 1.9	**10** 1.9	10 1.9
NOTCH1_9q34.3	9.3	0	9.3	0	9.3	9.3	9.3	0
NOTCH3_19p13.12	8	8	8	0	8	8	8	8
PBX2_6p21.33	**3.2**	3.2	**3.2**	3.2	3.2	3.2	3.2	3.2
PBX1_1q23.3	7.5 - 2.9 -	7.5 - - -	7.5 - 2.9 -	7.5 - 2.9 -	7.5 - 2.9 -	**7.5** 4.4 2.9 2.6	**7.5** - 2.9 -	**7.5** 4.4 2.9 -
PBX3_9q33.3	**2.5** 2.2 1.4	2.5 - -	2.5 - -	2.5 - -	2.5 - -	2.5 2.2 -	0	**2.5** 2.2 -
PBX4_19p13.11	- 4.5 1.4 1.2	- 4.5 - -	- 4.5 1.4 1.2	0	- 4.5 1.4 1.2	6.0 4.5 1.4 -	- 4.5 - -	- 4.5 1.4 -
RXRB_6p21.32	7.7 2.8 1.7	- 2.8 -	- 2.8 -	0	- - 1.7	7.7 **2.8** 1.7	- 2.8 -	- 2.8 -
RXRG_1q23.3	1.7	0	2.8 0.9	0	1.7 1.5	**1.7**	0	1.7 1.5
RXRA_9q34.2	5.4	5.4	5.4	0	5.4	5.4	0	0
TUBB_6p21.3	- 4.8 4.0 2.5 1.7 -	- - - 2.5 **1.7** -	5.1 - - - 1.7 -	- - - - 1.7 -	- - - 1.7 -	- - 4.8 4.0 1.7 1.35	- - - - 1.7 -	- - - - 1.7 1.35
TUBB4QL_10p15.3	0	2.8	0	0	0	0	0	0
TUBB4_16q24.3	1.7 1.3	1.7 1.3	1.7 1.3	1.7 1.3	1.7 1.3	1.7 1.3	1.7 1.3	1.7 1.3
TUBBL_18p11.1	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3
TUBB5_19p13.3	0	2.3	0	0	0	0	0	2.3
TUBB1_20q13.32	0	0	0	0	0	0	0	0

Appendix 8

Summary of microarray results.

Gene	*Adrenal gland*	*Brain*	*Skeletal muscle*	*Spleen*	*Testis*	*293T (kidney)*	*Jurkat*	*Raji*	*THP1*	*U937 (lung)*
AIF1_6p21.33	0	0	0	1	0	0	1	0	0	1
AIF1-L_9q34.12	0	1	0	1	0	0	0	0	0	0
BRD2_6p21.32	1	1	1	1	1	1	1	1	1	1
BRDT_1p22.1	0	0	0	0	1	0	0	0	0	0
BRD3_9q34.2	1	1	0	0	1	0	1	1	0	0
BRD4_19p13.12	0	1	1	0	0	0	0	1	0	0
C4_6p21.33	0	0	0	0	0	0	0	0	0	0
C5_9q33.2	0	0	0	0	0	0	0	0	0	0
C3_19p13.3	0	0	0	0	0	0	0	0	0	0
CLIC1_6p21.33	1	1	1	1	1	1	1	1	1	1
CLIC4_1p35.3	0	1	1	1	1	0	1	1	0	1
CLIC3_9q34.3	0	0	1	0	0	0	0	0	0	0
CLIC5_6p21.1	0	0	1	1	0	0	0	0	0	0
CLIC6_21q22.12	0	0	0	0	0	0	0	0	0	0
CLIC2_Xq28	0	0	0	1	0	0	0	0	0	0
GPX5_6p22.1	0	0	0	0	0	0	0	0	0	0
GPX4_19p13.3	1	1	1	1	1	1	1	1	1	1
GPX1_3p21.31	1	1	1	1	1	1	1	1	1	1
GPX3_5q33.1	1	1	1	1	1	0	0	0	0	0
GPX2_14q23.3	0	0	0	0	0	0	0	0	0	0
NOTCH4_6p21.33	0	0	0	0	0	0	0	0	0	0
NOTCH2_1p11.2	0	1	0	1	1	0	0	1	0	0
NOTCH1_9q34.3	0	0	0	0	0	0	0	0	0	0
NOTCH3_19p13.12	1	1	0	1	1	0	1	0	0	0
PBX2_6p21.33	0	0	0	0	0	0	0	0	0	0
PBX1_1q23.3	1	1	0	0	1	1	0	1	0	0
PBX3_9q33.3	0	0	0	0	1	0	0	1	0	0
PBX4_19p13.11	0	1	0	0	1	0	1	1	0	0
RXRB_6p21.32	1	1	1	1	1	0	1	1	1	1
RXRG_1q23.3	0	0	1	0	0	0	0	0	0	0
RXRA_9q34.2	1	1	0	1	1	0	0	0	1	1
TUBB_6p21.3	1	1	1	1	1	1	1	1	0	1
TUBB4QL_10p15.3	0	1	1	0	1	0	0	1	0	0
TUBB4_16q24.3	0	1	0	0	1	0	0	0	0	0
TUBBL_18p11.3	1	1	1	1	1	0	0	1	1	0
TUBB5_19p13.3	1	1	0	0	1	0	0	0	0	0
TUBB1_20q13.3	0	0	0	1	0	0	0	0	0	0

Appendix 9

Comparison of three methods used to generate the expression profiles for nine MHC paralogous gene families. The differences between the three methods are highlighted in yellow for the nine tissues common to each method.

Gene	*Adrenal gland*			*Brain*			*Skeletal muscle*			*Spleen*			*Testis*			*Kidney*			*T cell*			*B cell*			*Lung*		
	M	**D**	**S**	**M**	**D**	**S**	**M**	**D**	**S**	**M**	**D**	**S**	**M**	**D**	**S**	**M**	**D**	**S**	**M**	**D**	**S**	**M**	**D**	**S**	**M**	**D**	**S**
AIF1_6p21.33	0	1	0	0	0	1	0	1	1	1	1	1	0	0	1	0	1	1	1	1	0	0	0	0	1	1	1
AIF1-L_9q34.12	0	0	1	1	1	1	0	0	1	1	1	1	0	1	0	0	1	1	0	0	0	0	0	0	0	0	1
BRD2_6p21.32	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BRDT_1p22.1	0	0	0	0	0	1	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
BRD3_9q34.2	1	1	1	1	1	1	0	1	1	0	1	0	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1
BRD4_19p13.12	0	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1	1	0	0	1	1	1	1	0	1	1
C4_6p21.33	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	1	0	0	1
C5_9q33.2	0	1	0	0	1	1	0	1	0	0	1	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	1
C3_19p13.3	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	0	0	0	0	1	1
CLIC1_6p21.33	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
CLIC4_1p35.3	0	0	1	1	1	1	1	1	1	1	1	0	1	0	1	0	1	1	1	0	1	1	0	1	1	1	1
CLIC3_9q34.3	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1
CLIC5_6p21.1	0	0	1	0	0	0	1	1	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0
CLIC6_21q22.12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1
CLIC2_Xq28	0	0	0	0	0	1	0	0	1	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
GPX5_6p22.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
GPX4_19p13.3	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
GPX1_3p21.31	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
GPX3_5q33.1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	0	0	0	0	0	1	1
GPX2_14q23.3	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1
NOTCH4_6p21.33	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1
NOTCH2_1p11.2	0	1	1	1	1	1	0	1	1	1	1	0	1	1	1	0	1	1	0	1	1	1	1	1	0	1	1
NOTCH1_9q34.3	0	0	0	0	1	1	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0	1	0	0	1
NOTCH3_19p13.12	1	1	0	1	1	1	0	1	0	1	1	0	1	1	1	0	1	1	1	1	0	0	1	1	0	1	1
PBX2_6p21.33	0	1	0	0	1	0	0	1	1	0	1	0	0	1	1	0	1	1	0	0	0	0	0	0	0	1	1
PBX1_1q23.3	1	1	1	1	1	1	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	1	1	0	0	1	1
PBX3_9q33.3	0	1	1	0	0	1	0	1	0	0	1	0	1	1	1	0	1	1	0	0	0	1	1	1	0	1	1
PBX4_19p13.11	0	0	0	1	1	1	0	1	0	0	0	0	1	1	1	0	1	0	1	0	0	1	0	0	0	1	1
RXRB_6p21.32	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	1
RXRG_1q23.3	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
RXRA_9q34.2	1	1	0	1	1	1	0	1	1	1	1	0	1	1	1	0	1	1	0	0	0	0	0	1	1	1	1
TUBB_6p21.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
TUBB5_19p13.3	0	0	1	1	1	1	1	0	0	0	0	0	1	0	1	0	0	1	0	0	0	1	0	1	0	0	1
TUBB4_16q24.3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	1	0	1	1	0	0	1	1
TUBB2_18p11.1	1	1	0	1	1	1	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1
TUBB1_20q13.32	0	0	0	0	0	1	0	0	0	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
Total differences	**5**	**3**	**9**	**4**	**1**	**8**	**5**	**4**	**5**	**3**	**5**	**7**	**4**	**4**	**5**	**16**	**1**	**8**	**3**	**2**	**5**	**2**	**2**	**5**	**15**	**0**	**10**
% difference	**14**	**8**	**25**	**11**	**3**	**22**	**14**	**11**	**14**	**8**	**14**	**19**	**11**	**11**	**14**	**44**	**3**	**22**	**8**	**6**	**14**	**6**	**6**	**14**	**42**	**0**	**28**