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.

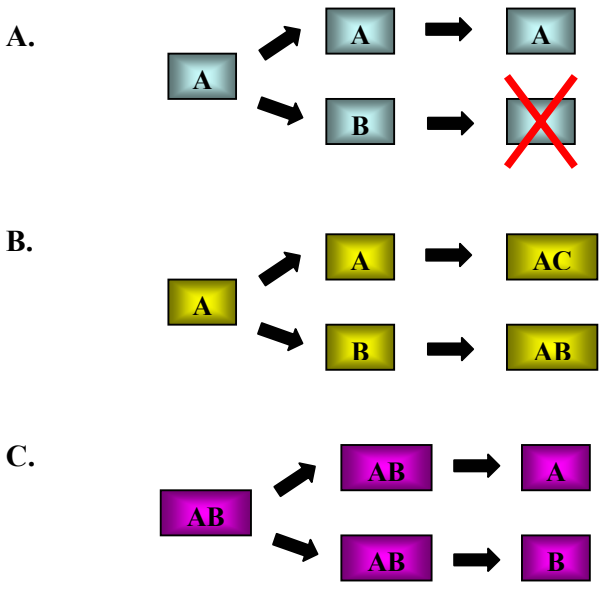


Figure 6.1 Fates of duplicated genes (adapted from Mazet and Shimeld, 2002). (A) non-functionalisation, in which one copy degenerates after duplication, (B) neo-functionalisation, when initially identical duplicates with function *A* diverge by acquiring new functions *B* and *C* and (C) sub-functionalisation,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

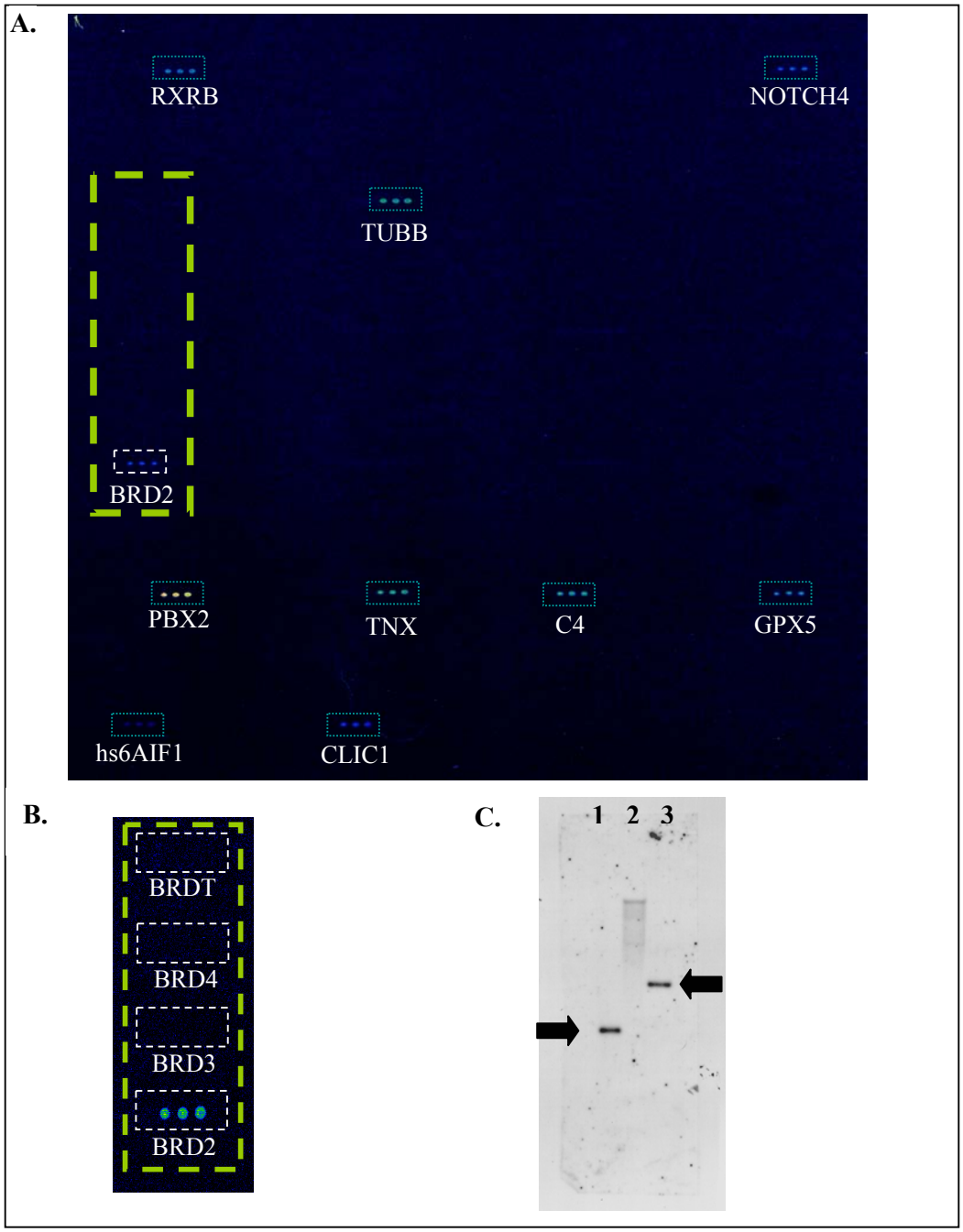


Figure 6.2 Verification of probe specificity. (A) represents the results of the 'Paralogue Microarray' hybridisation using the 10 pooled probes. The corresponding probes and targets have hybridised and the spots (in triplicate) are visible (they are boxed and the name of the extended MHC gene given). 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.

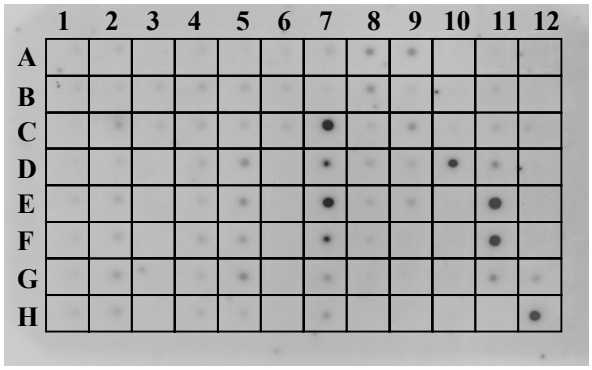
Figure 6.4 summarises the results of the dot-blot analysis of the Allograft inflammatory factor 1 gene (AIF1; figure 6.4.A) and the paralogue on 9q34.12 (AIF1-

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.

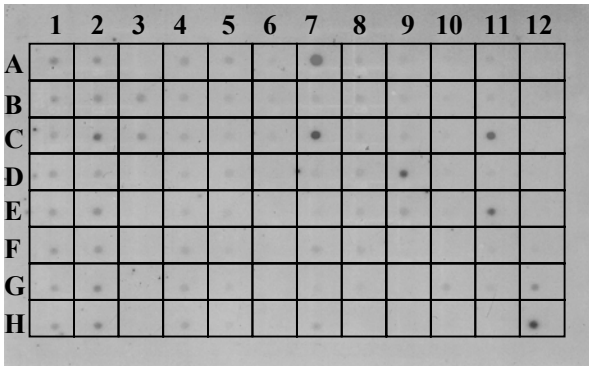
There are other examples of co-expression and divergence of expression. In particular, both are expressed in areas of the brain. AIF1-L is expressed in whole brain as well as the 17 different sections of the brain individually spotted on the blot, whereas AIF1 is more selective and is expressed in 10 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.

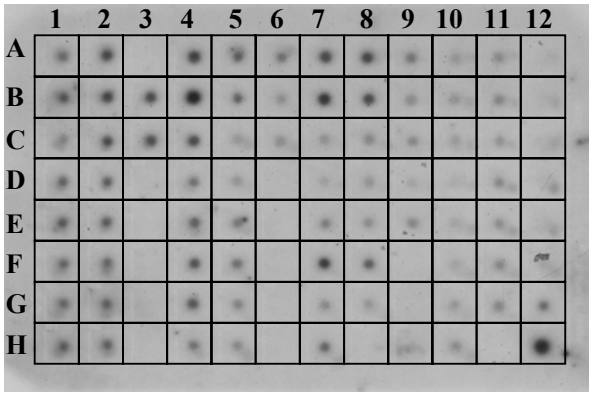
A. AIF1 (6p21.33)



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.

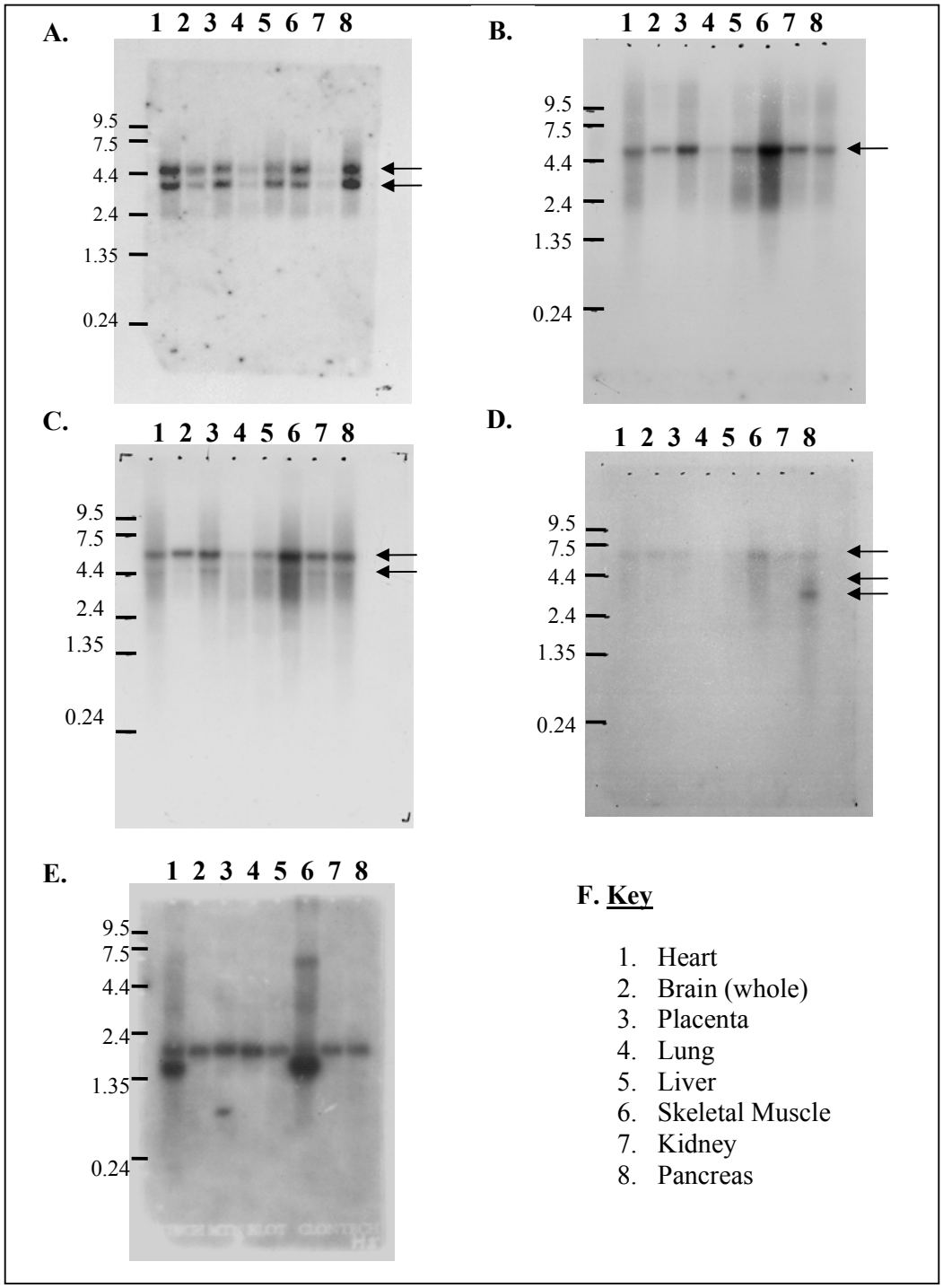


Figure 6.5 Transcription pattern and splice variants of the BRD2 (A), BRD3 (B) BRD4 (C), BRDT (D) and β-actin control (E) genes after hybridisation with specific probes to a Northern blot with eight different tissues. (F) Tissue/source key of the Northern blot as supplied by the manufacturer (Clontech). 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).

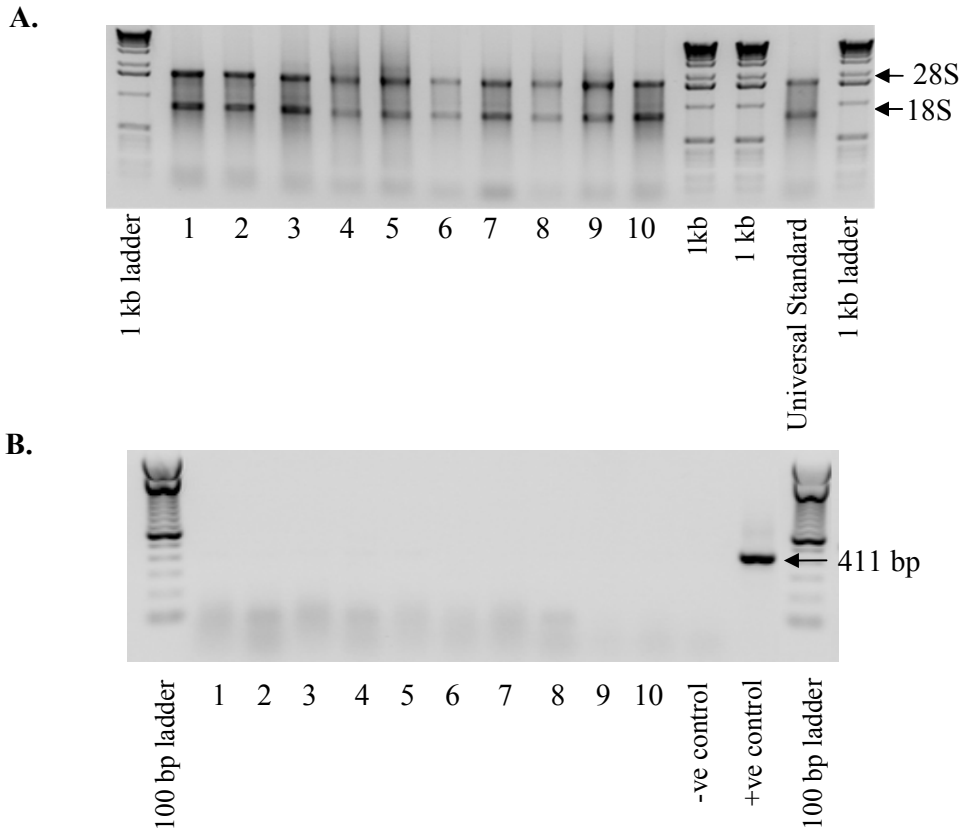


Figure 6.6 (A) Assessment of the quality of the eleven RNAs used in the expression microarray experiments. The RNA is of good quality and has the two distinct bands corresponding to 28S and 18S ribosomal RNA, indicated by arrows. (B) RNA was checked for DNA contamination using 'no RT-PCR'. The primers 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.

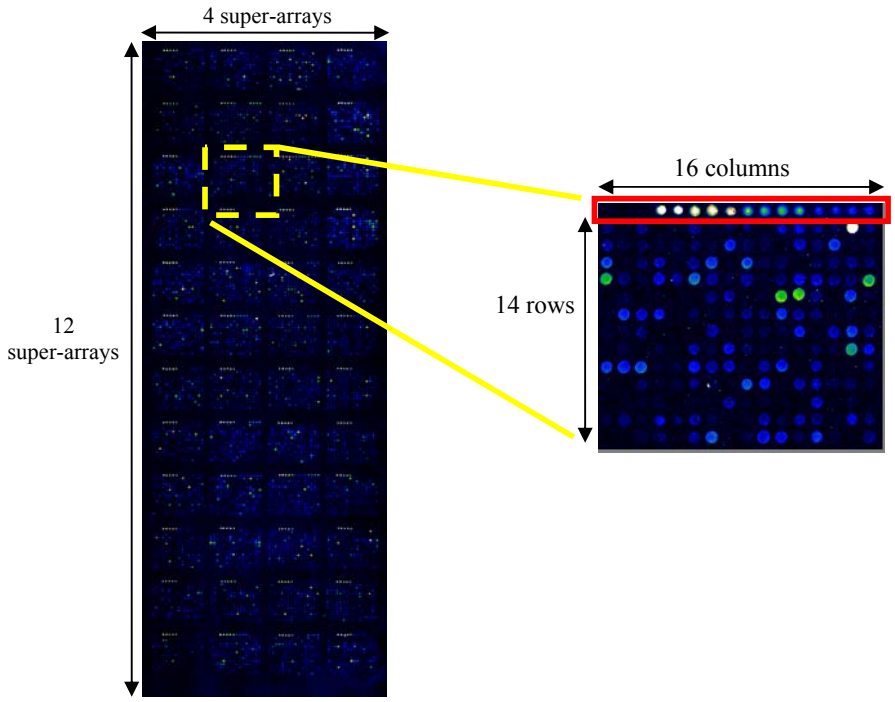


Figure 6.7 Results of a hybridisation with the standard Stratagene RNA to the '10K/Paralogue Microarray'. The layout of the 48 sub-arrays in 12 x 4 super-arrays is visible after hybridisation and the sub-array boxed in yellow is expanded. The first row of the sub-array is boxed in red. Columns 1 to 8 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.

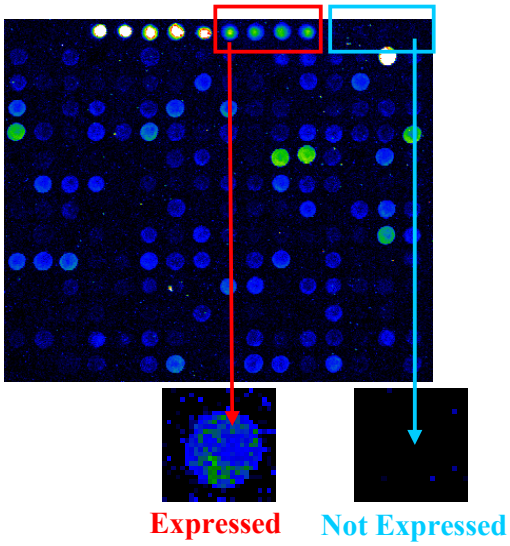


Figure 6.8 One of the 48 sub-arrays of the '10K/Paralogue Microarray' after hybridisation using the Stratagene standard RNA. The four spots boxed in red correspond to one paralogue which is expressed in the standard RNA. The area boxed in blue contains four DNA elements corresponding to one paralogue 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).

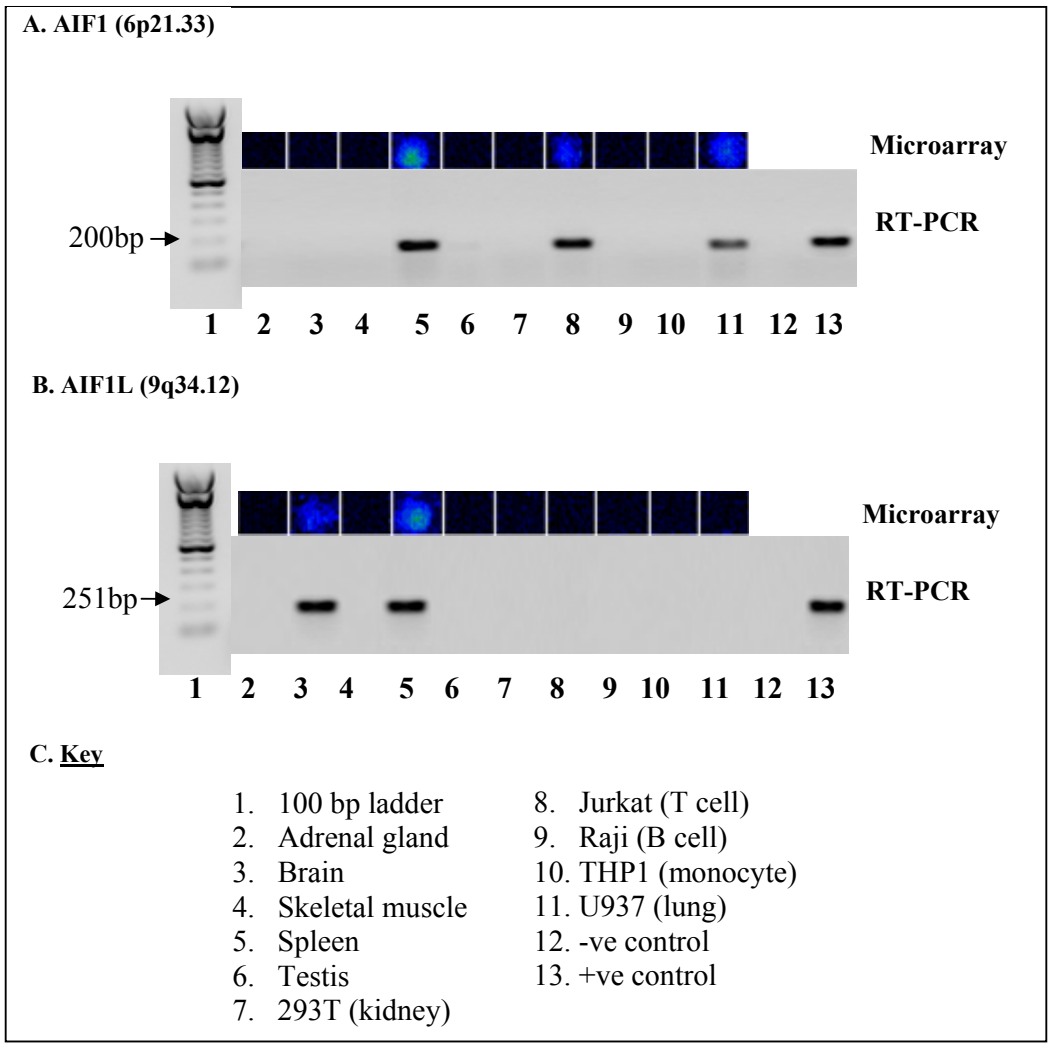


Figure 6.9 Microarray results confirmed by RT-PCR. (A) RT-PCR (using the primers AIF1.F1 and AIF1.R1 summarised in appendix 4) and (B) RT-PCR results (using AIF1-L.F and AIF1-L.R see Appendix 4) and microarray results of AIF1-L. The negative (-ve) control used water as a template and the positive (+ve)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

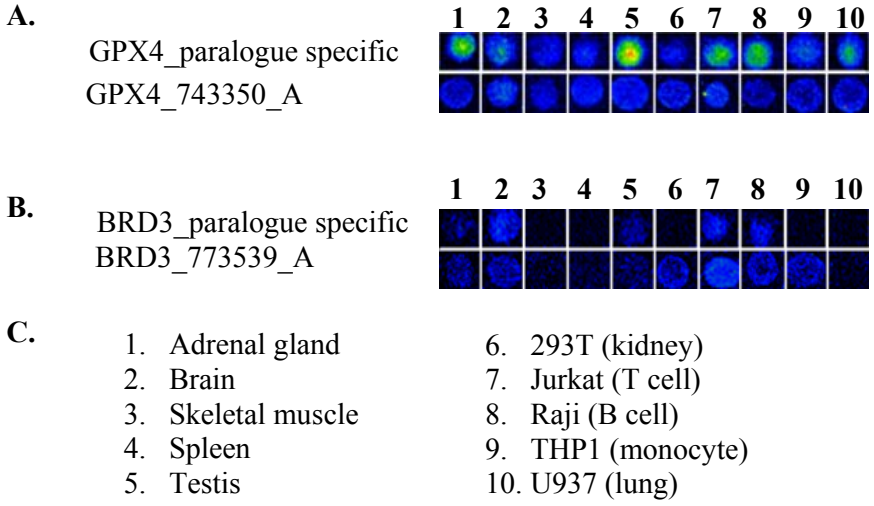


Figure 6.10 Comparison of the expression profiles of the paralogue specific PCR products designed in this thesis and those already on the standard Sanger Institute 10K microarray corresponding to (A) GPX4 and (B) BRD3 genes. (C) is the key to the tissues and cell-lines used.

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

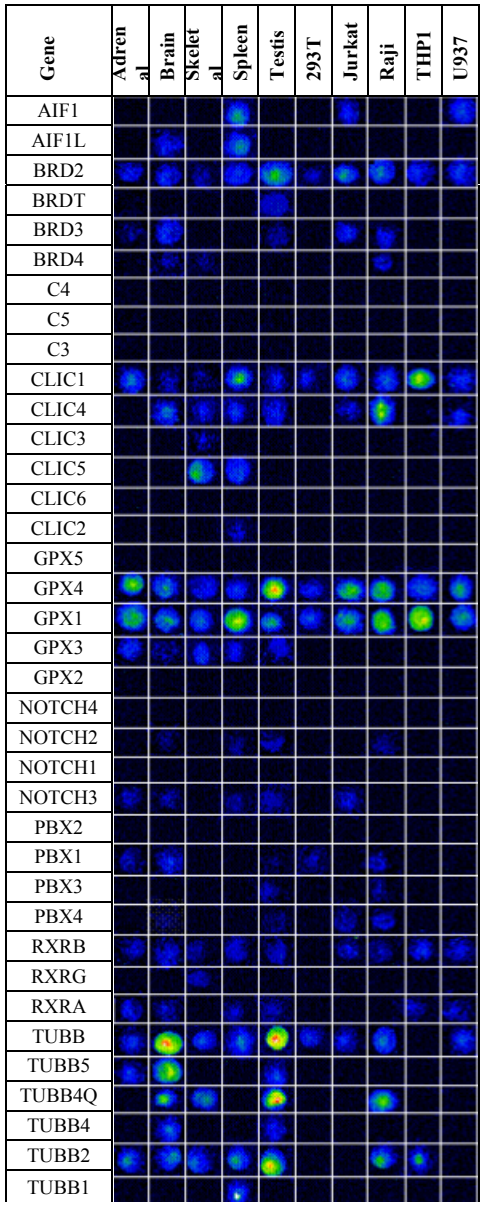
two out of the three experiments showed expression the gene in that tissue was assigned the value of 1.

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).

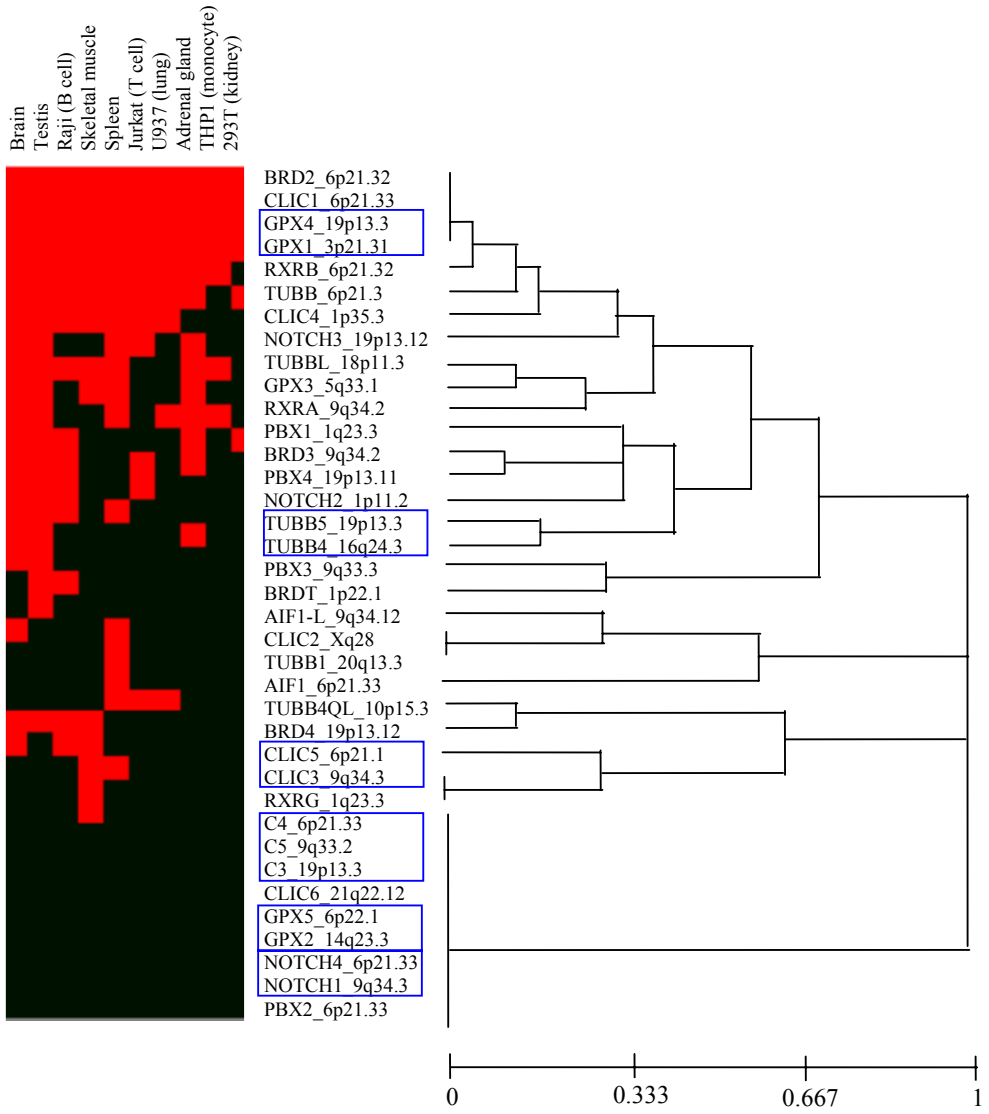
In addition to highlighting the relationships between members of the same paralogous gene families, clustering the data reveals the relationships between all the paralogues used in the expression analysis (summarised in figure 6.12.B). It is interesting to note that the four paralogues that are expressed 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.

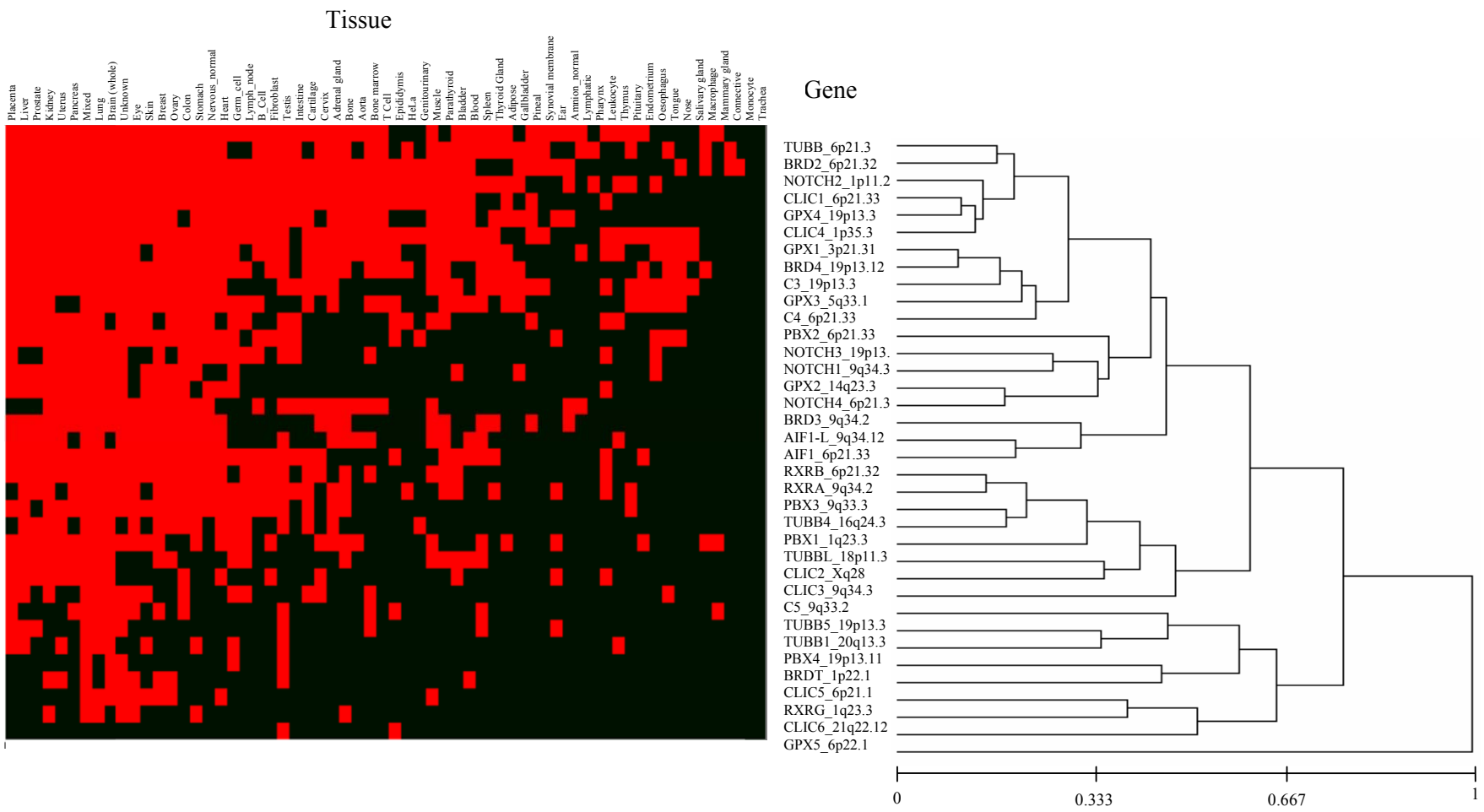

Brain
Testis
Raji (B cell)
Skeletal muscle
Spleen
Jurkat (T cell)
U937 (lung)
Adrenal gland
THP1 (monocyte)
293T (kidney)
BRD2_6p21.32
CLIC1_6p21.33
GPX4_19p13.3
GPX1_3p21.31
RXRB_6p21.32
TUBB_6p21.3
CLIC4_1p35.3
NOTCH3_19p13.12
TUBBL_18p11.3
GPX3_5q33.1
RXRA_9q34.2
PBX1_1q23.3
BRD3_9q34.2
PBX4_19p13.11
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

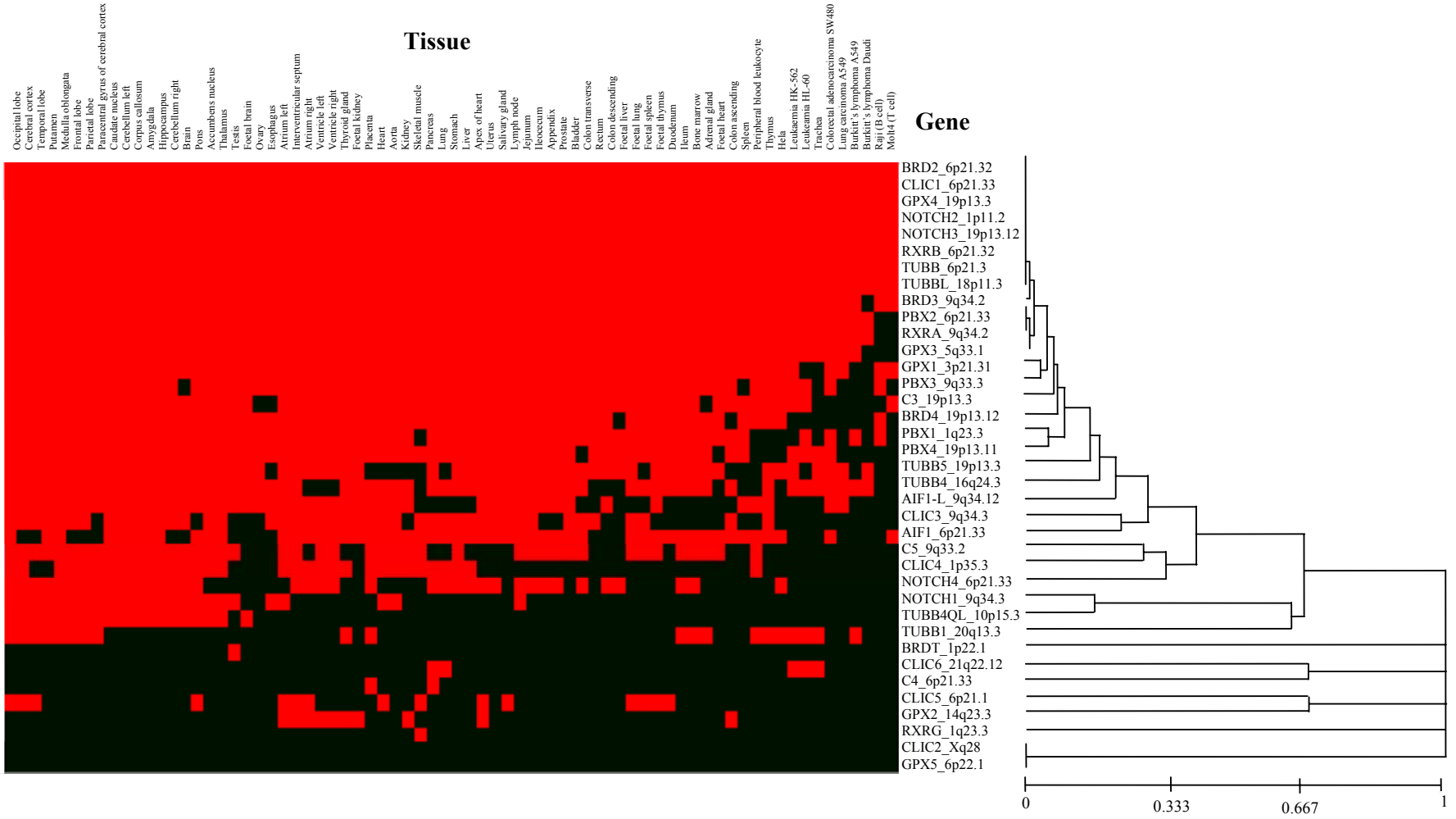
6.3.3.4 Dot-blot expression data

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).

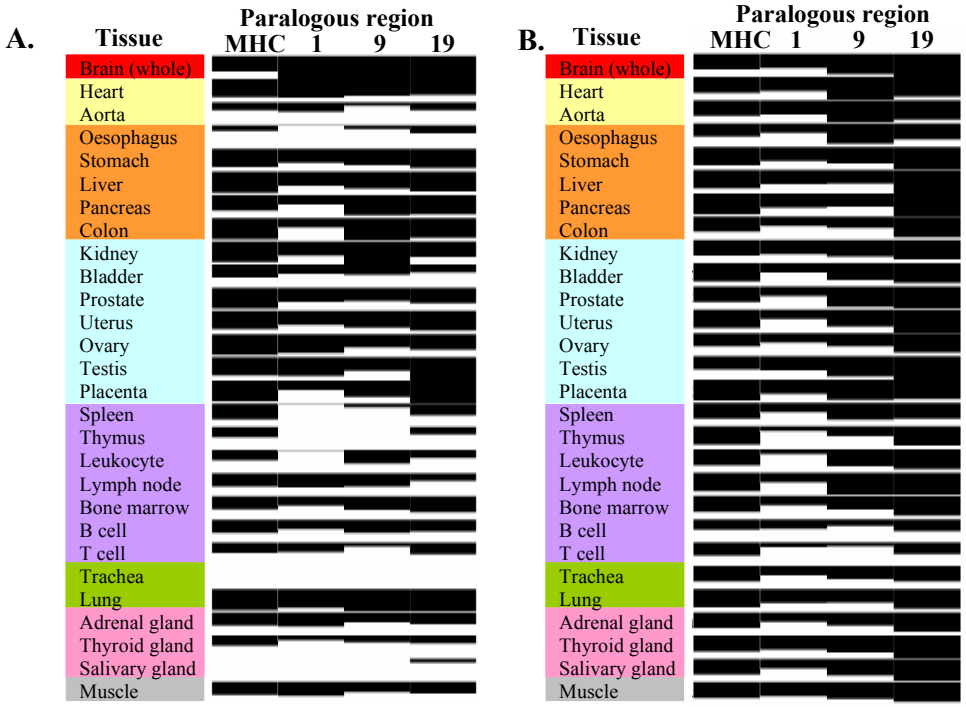


Figure 6.15 Comparison of the expression profiles of the paralogues located within the paralogous regions on chromosomes 1, 9 and 19 with the MHC genes using (A) *in-silico* and (B) dot blot analysis in 28 normal human tissues. The tissues are divided into eight systems of the human body; nervous (red), 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.