# MAMMALIAN-WIDE INTERSPERSED REPEATS (MIRs) AND THEIR ROLE IN MAMMALIAN GENE FUNCTION AND EVOLUTION

SARA MARIE CROFT

A thesis submitted in partial fulfilment of the requirements of Nottingham Trent University for the degree of Doctor of Philosophy

September 2009

# ABSTRACT

Transposable elements (TEs) are ubiquitous components of plant and animal genomes and constitute more than ~45% of the human genome. Though originally considered as ‘parasitic’ or ’junk’ DNA, TEs are now thought to have played a role in shaping genomes during evolution, contributing to genome plasticity and diversity. All classes of retrotransposons accumulate in the genome via a process termed retrotransposition, wherein the elements are reverse transcribed into RNA and inserted into the genome as DNA. Exaptation of these elements can provide additional or novel function for endogenous genes. Mammalian-wide interspersed repeats (MIRs) are short interspersed nuclear elements (SINEs), belonging to the non-autonomous class of retroelements and are found in all mammals. The recruitment of an MIR element by a gene may provide insight into mammalian evolution and gene function.

The human genome was screened for genes that have exaptated MIR elements and the compiled dataset was analysed to determine any commonality which may suggest conserved function(s). Subsequently 1359 genes were identified that have exaptated MIR elements, constituting 5% of the total genes in the human genome. MIR elements may be multifunctional, as 1% of the total human genes contain MIRs that are spliced and/or are contributing to protein coding sequences. Subsequently sequence motifs were identified in the MIR consensus sequences which resemble canonical mammalian splice sites; therefore MIR elements recruited in the 5’-UTR and coding sequence may be a result of the exonisation of intronic elements. The MIR-containing transcripts are frequently expressed in neurological tissue, suggesting a role in neuronal function. Moreover a number of MIR-containing mRNA transcripts are known to be localised to the dendritic compartment of the neurone, and ciliated region of photoreceptors. Some of the localised mRNAs contain putative microRNA binding sites within the MIR sequence, and possible dsRNA structures were noted between MIR elements. It is proposed that exaptated MIR elements may be a source of *cis*-acting regulatory elements, involved in post-transcriptional control of gene expression, including localisation of mRNA to distinct intracellular compartments.

# ACKNOWLEDGEMENTS

I would like to begin by thanking my supervisor, Dr David Hughes, for his constant guidance, motivation, and support during my undergraduate and postgraduate studies. I would not have reached this point without his insight, experience and knowledge during this project. I would also like to thank fellow lab member Amy Poole, whom assisted me in gaining laboratory experience and knowledge. I also acknowledge fellow members of the Signal Transduction laboratory including Rania, Ash, Wayne, Alan and David who have at some point assisted me in the two years we have shared the lab together.

I would like to thank Dr Eli Verderio-Edwards, my internal assessor during my PhD studies, who became a welcomed collaborator throughout the entire study process. I also acknowledge fellow PhD students, Laurice, Amy, Ash and Rania for being faithful friends, who have made the whole experience more memorable and enjoyable.

I would like to especially show my appreciation and thanks to my family, particularly Rich, for their love, support and patience in all my efforts. And last but not least I especially thank my son Jake, my greatest achievement, for being a very understanding, comical, loving and mature little man, who each day encourages me to focus and be the person I am, without which I would not have had the commitment or dedication to be where I am today.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ABBREVIATIONS

| | |
|---|---|
| AFC | African cichlid fish family of SINEs |
| BLAST | basic local alignment search tool |
| BLS | bare lymphocyte syndrome |
| cDNA | complementary deoxyribonucleic acid |
| CDS | coding sequence |
| CGD | chronic granulomatous disease |
| CML | chronic myeloid leukaemia |
| CR | cys-rich repeat |
| DMSO | dimethyl sulphoxide |
| DNA | deoxyribonucleic acid |
| dsRNA | double stranded RNA |
| DTE | dendritic targeting element |
| DTT | dithiothreitol |
| EDTA | ethylenediaminetetraacetic acid |
| EST | expressed sequence tag |
| FBS | foetal bovine serum |
| FDR | false discovery rate |
| GAD | genetic association database |
| gDNA | genomic DNA |
| GLUD | Glutamate dehydrogenase |
| GO | gene ontology |
| GPCR | G-protein coupled receptor |
| IPTG | isopropyl β-D-1-thiogalactopyranoside |
| JBTS | Joubert syndrome |
| KEGG | Kyoto encyclopaedia of genes and genomes |
| LB | Luria Bertani |
| LINE | long interspersed nuclear elements |
| LRRs | leucine rich repeats |
| LTR | long terminal repeat |
| MHC | major histocompatibility complex |
| MIR | mammalian-wide interspersed repeats |
| miRNA | microRNA |
| mRNA | messenger RNA |
| myr | Million years |
| NCBI | the National Centre for Biotechnology Information |

# ABBREVIATIONS CONTINUED

| | |
|---|---|
| ncDNA | non-coding DNA |
| ncRNA | non-coding RNA |
| NMD | nonsense-mediated decay |
| OMIM | online mendelian inheritance in man |
| ORF | open reading frame |
| PBS | phosphate-buffered saline |
| Ref Seq | reference sequence |
| RISC | RNA-induced silencing complex |
| RNA | ribonucleic acid |
| RNAi | RNA interference |
| RNP | ribonucleoprotein complex |
| RP | retinitis pigmentosa |
| rRNA | ribosomal RNA |
| RT-PCR | reverse transcriptase polymerase chain reaction |
| SINE | short interspersed nuclear elements |
| siRNA | small interfering RNA |
| SNP | single nucleotide polymorphism |
| SOC | super optimal broth with catabolite repression |
| ssDNA | single stranded DNA |
| SVA | acronym for SINE-R, VNTR and Alu |
| TEs | transposable elements |
| TPRT | target primed reverse transcription |
| tRNA | transfer RNA |
| TSD | target site duplication |
| UTR | untranslated region |
| VNTR | variable-number-of-tandem-repeats |
| X-GAL | 5-bromo-4-chloro-3-indolyl-b-D-galactopyranoside |

# 1. INTRODUCTION

## 1.1. Early history of non-coding and repetitive DNA

The phenomenon of transposition was first suggested by Barbara McClintock in the late 1940s; when observing mosaic colour patterning in maize (*Zea mays*); she noted a variation in the pigmentation of individual maize kernels. She identified specific loci (*Ds* and *As*) responsible for unstable inheritance of mosaicism between generations, and later suggested *As* could change the chromosome position (transpose) of *Ds,* resulting in the mosaic patterning (McClintock, 1944; McClintock, 1950). These *Ds* and *As* elements were often referred to as 'jumping genes' or 'controlling elements'; however McClintock's hypothesis received scepticism from the scientific community and 15 years elapsed until McClintock revisited her original ideas (McClintock, 1961). Both *Ds* and *As* were later discovered to be a class of DNA transposons (Peterson, 1981; Fedoroff, 1989) and Barbara McClintock went on to be the first women to receive an unshared Nobel Prize for her efforts, 30 years after discovering mobile genetic elements.

The pioneers in repetitive DNA research were Roy Britten and his colleagues. In the 1960s a DNA-agar procedure was developed, which enabled the measurement of DNA hybridisation between the DNA of different species (Marmur and Doty, 1961; McCarthy and Bolton, 1963). Britten and Kohne (1968) used this technique to demonstrate the hybridisation of large sections of DNA between divergent species, and noticed the frequent hybridisation of regions of 200-300 base pairs (bp). They suggested that the hybridisation was a consequence of DNA sequence homology and the re-association of interspersed repeat elements, and was not the result of non-specific aggregation as previously thought (Britten and Kohne, 1968). Britten further suggested that non-coding repetitive DNA may play a fundamental role in gene expression, regulation and arrangement and the Britten-Davidson model was proposed (Britten and Davidson, 1969). They hypothesed that genome complexity involves the coordination and regulation of unrelated genes via a regulatory element, which could target specific unlinked genes, and they suggested the regulatory sequences were the previously described interspersed repeat elements (Britten and Davidson, 1969). Britten and Kohn further suggested that repetitive DNA may be a core component in phenotype diversity, and provide a means of modifying genomes during speciation and evolution (Britten and Kohn 1970; Britten and Kohn 1971). However this model was received with some scepticism and shortly after Ohno (1972) proposed that non-coding DNA (ncDNA) was “junk” or “parasitic” material with little or no biological function. It is worth noting that Ohno argued that there may be an advantage to having non-coding intronic and intergenic DNA (Ohno, 1972).

Similar questions over the significance of repeat sequences are raised by the c-value paradox, whereby there appears to be no correlation between genome size and organism complexity or diversity (Zuckerkandl, 1976; Gregory, 2001; Petrov, 2001; Kidwell, 2002; Patrushev and Minkevich, 2008). In vertebrates a large proportion of ncDNA is composed of transposable elements (TEs), many of which are propagated by the mechanism of retrotransposition (Deininger *et al.,* 2003). In the past 15 years there has been renewed interest in TEs and the importance of these elements has been reassessed, as a result the term “junk DNA” is now viewed with a different understanding. For example, everyday household junk may be stored in the event it will be of some use in the future whereas rubbish is thrown away and discarded, and ncDNA is now considered more as a genomic scrap yard (Brenner, 1990; Makałowski, 2000).

A large portion of ncDNA will be non-functional, representing fossils of past evolutionary events or ‘natures experiments’, which have been maintained merely due to the lack of selective constraints. Likewise there are regions of ncDNA which have become vital components of the functional genome through exaptation (for a review see Mattick and Makunin, 2006). Following the recruitment of a repeat element there will be additional nucleotides available which may allow for the modification of gene function and/or expression, and contribute to alternate splicing, polyadenylation features and additional protein coding information (Lev-Maor *et al.,* 2003; Lee *et al.,* 2008; Nekrutenko and Li, 2001). It is now accepted that modern genomic DNA (gDNA) may have evolved in close association with TEs, with retrotransposition providing additional raw genetic material, which may be utilised to adapt new phenotypes, shaping genomes during evolution (Roy-Engel *et al.,* 2002).

## 1.2. Classification of transposable elements (TEs)

The understanding of repetitive DNA and transposons in particular has improved greatly following the sequencing of the human genome (Lander *et al.,* 2001). Approximately 1.2% of the total human genome encodes for protein, whereas almost half is derived from TEs (Lander *et al.,* 2001). In comparison 37% of the mouse genome is repeat-derived (Waterson *et al.,* 2002), and in both the mouse and human genomes 60% of these repeats are retained within intronic sequences (Sela *et al.,* 2007). However it is worth noting that not all TEs will have been positively identified, and the estimations outlined in figure 1.1 are likely to be an underestimate of the total number of TEs in the human genome. Most transposons are inactive and have been for millions of years, therefore many have fragmented and diverged to a point where they are no longer distinguishable, creating problems when using current algorithms. When predictions are made using the evolutionary-based 'repeat probability cloud' detection method as described by Gu *et al.,* (2008) it is suggested that only half of the actual transposons may have been annotated.

TEs fall into three categories; DNA transposons, long terminal repeat (LTR) retrotransposons and non-LTR retrotransposons (Deininger and Batzer, 2002). The retrotransposons spread via an RNA intermediate and comprise the majority of mammalian TEs. In contrast, DNA transposons spread directly as DNA without transcription and constitute a small proportion of the total repeat elements in the human genome (Deininger *et al.,* 2003). DNA transposons are considered the least successful of the TEs due to a less efficient mode of transposition, as these elements propagate via a 'cut and paste' mechanism and do not increase in number.

Autonomous retrotransposons have the capacity to direct their own amplification; long interspersed nuclear elements (LINEs) and LTR retrotransposons are autonomous elements. In contrast the short interspersed nuclear elements (SINEs) are non-autonomous and rely on co-opting the retrotranspositional machinery encoded by other elements, predominantly LINEs (Dewannieux *et al.,* 2003). Retrotransposons vary considerably in their structure as outlined in figure 1.2.

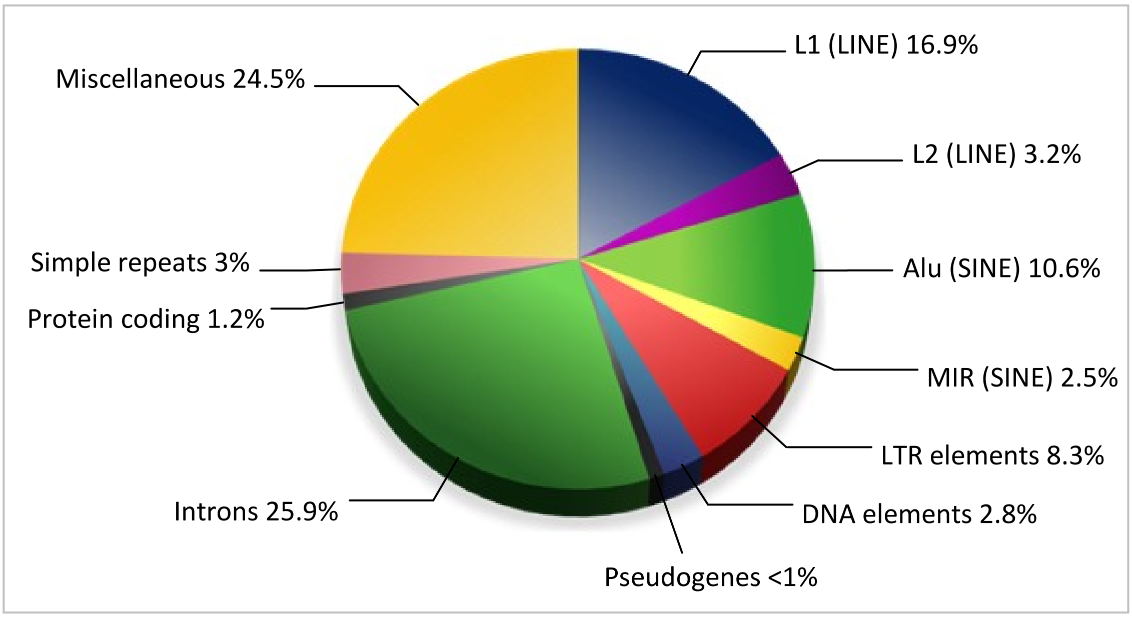


**Figure 1.1. The principal components of the human genome**

Almost 26% of the total human genome is composed of intronic sequence whereas only 1.2% encodes for protein. TEs comprise 44% of the total genome sequence, with the most abundant elements being the LINE and SINE repeats which comprise 21.1% and 13.1% respectively (Lander *et al.,* 2001).

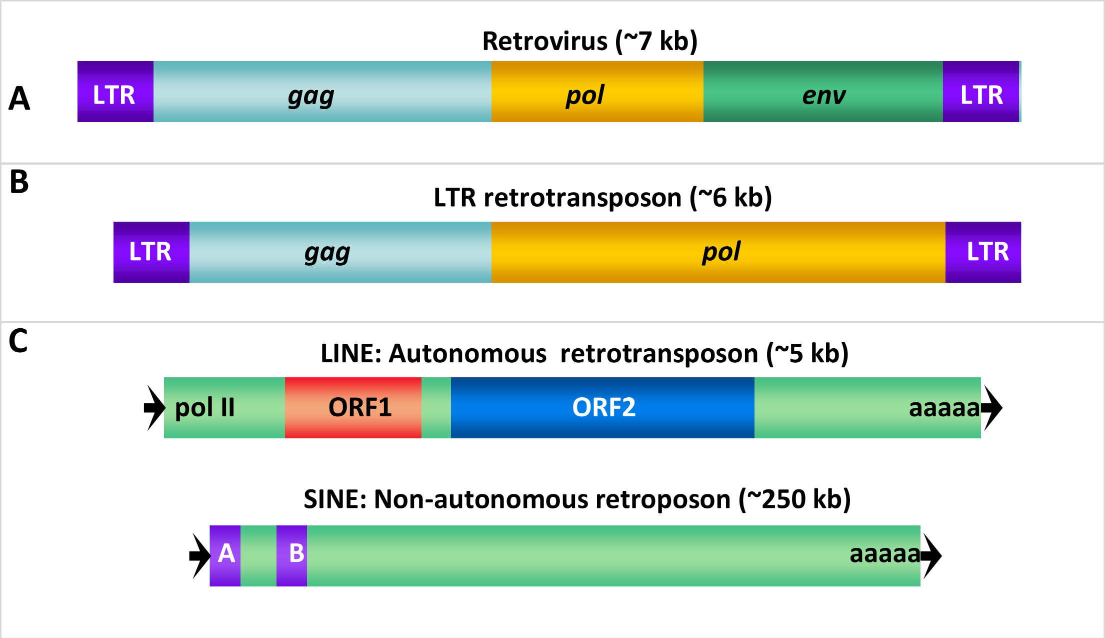

Retrovirus (~7 kb)
A
LTR
gag
pol
env
LTR
B
LTR retrotransposon (~6 kb)
LTR
gag
pol
LTR
C
LINE: Autonomous retrotransposon (~5 kb)
pol II
ORF1
ORF2
aaaaa
SINE: Non-autonomous retroposon (~250 kb)
A
B
aaaaa

**Figure 1.2. Schematic representation of the major categories of retrotransposons**

A, Retrovirus, the retroviral genes are in italics: B, LTR retrotransposon, similar in structure to the retrovirus though lacking the *env* gene; C, non-LTR retrotransposons, the promoter regions have been included as have the poly(A) tails. The LINE element contains two open reading frames (ORF1 and ORF2). The promoter boxes for the SINE elements are labelled A and B and the target site duplications are indicated by black arrows.

### 1.2.1. LTR retrotransposons and endogenous retroviruses

LTR retrotransposons are ubiquitous across species, ranging from single-celled organisms to humans, and resemble exogenous retroviruses in arrangement (Deininger *et al.,* 2003). This family of repeats contain flanking long terminal repeats (LTR), are of a similar size to retroviruses and encode for *gag* and *pol* genes, but lack a functional *env* gene (Havecker *et al.,* 2004). LTR retrotransposons comprise 8% of the human genome (see figure 1.1), the majority of which are immobile, with a small number of exceptions such as the Tf1/sushi group (Butler *et al.,* 2001) and human endogenous retroviruses (HERV) (Medstrand and Mager, 1998). In contrast, rodent genomes contain an abundance of active LTR retroelements including the intracisternal A-particles (IAPs), MaLR, VL30 and ETn elements (French and Norton, 1997; Gwynn *et al.,* 1998; Mager and Freeman, 2000; Baust *et al.,* 2002).

It has been argued that retroviruses may have originally evolved from LTR-retrotransposons following the acquisition of the envelope gene. However the most likely scenario is that retroviruses originated from a provirus, with the endogenous retrovirus losing the capacity to form the viral envelope, subsequently resulting in the early LTR-retrotransposon (Xiong and Eickbush, 1990; Hughes and Coffin, 2002).

### 1.2.2. Long interspersed nuclear elements (LINEs)

The most ancient and possibly the most successful class of retrotransposons are the LINEs, of which there are three distinctive families; (L1, L2 and L3) numbered according to their age, with the L1 elements being the youngest of the group (Lander *et al.,* 2001). The most abundant and only active LINE elements in the human genome are the L1 family, with over 500,000 copies, constituting almost 17% of the total genome sequence (figure 1.1).

LINE repeats are typically 6-7 kb when full length and contain two open reading frames (ORF). ORF1 is ~1 kb in length and encodes for a protein which has nucleic acid binding capacity and is hypothesised to bind to the LINE template RNA during retrotransposition (Kolosha and Martin, 1997). The larger ORF2 of ~4 kb encodes for a multifunctional protein with reverse transcriptase and endonuclease activity (McClure, 1991). LINE elements contain an internal polymerase II promoter, which is located downstream of the transcription start site; as a consequence the promoter is preserved following retrotransposition (Minakami *et al.,* 1992). L1 elements are non-LTR retrotransposons and therefore do not contain LTRs, however they do contain a poly(A) tail (Shedlock and Okada, 2000).

There are an estimated 80 to 100 L1s which remain active in the human genome (Brouha *et al.,* 2003); in contrast as many as 3000 L1 elements may be active within the mouse genome (Goodier *et al.,* 2001). The most active L1 element in humans is the Ta subfamily, estimated to be ~2 million years old, a third of which remain full length (Boissinot and Furano, 2001). The retrotransposition frequency of LINE elements in humans is relatively high, with an integration rate of 1:50 sperm and a new event is documented in every 10 to 250 human individuals born (Ostertag and Kazazian, 2001). L1 elements remain a constant source of mutation, and a number of human diseases are caused by *de novo* insertions of these elements (Ostertag *et al.,* 2003; Narita *et al.,* 1993; Deininger and Batzer, 1999).

### 1.2.3. Short interspersed nuclear elements (SINEs)

SINEs are short sequence elements ranging between 70-500bp in length, depending on the sub-class. Unlike LINE elements, SINEs are non-autonomous and lack the ability to encode any proteins such as reverse transcriptase and transposase, and propagate by utilising the enzymatic machinery encoded by the LINE elements (Eickbush, 1992). Three distinct SINE families exist in the human genome; the active primate specific Alu and SVA elements and the inactive ancient mammalian-wide interspersed repeats (MIR), each of which are comprised of several sub-families (Lander *et al.,* 2001; Ostertag *et al.,* 2003).

#### 1.2.3.1. The origin of SINE elements and the LINE/SINE relationship

SINE elements require enzymes encoded by LINEs to transpose and some SINE families may have acquired independent retrotranspositional activity by the fusion of the core tRNA-like sequence with the 3'-portion of a LINE repeat (Daniels and Deininger, 1985). Sharing the 3'-end with a LINE may provide SINEs with the enzymatic machinery critical to complete retrotransposition (Okada and Hamada, 1997). For instance the 3'-end of the MIR elements shares considerable sequence identity with L2 and L3 elements (Gilbert and Labuda, 1999; Okada and Hamada 1997). Other examples of LINE-derived 3'-ends of SINE repeats include the Anolis sauria SINE and Bov-B LINE (Piskurek *et al.,* 2008) and the SINE Bov-tA and LINE Bov-B (Gilbert and Labuda, 1999).

SINEs are divided into three classes according to their origin. The major class are those derived from transfer RNA (tRNA). The majority of SINEs, including the mammalian MIRs, rodent B2 elements and sauria SINEs are derived from tRNA genes (Labuda *et al.,* 1991; Deininger *et al.,* 2003; Piskurek *et al.,* 2006), with tRNA$^{Lys}$ being the most common ancestor (Shedlock and Okada, 2000). The minor classes of SINEs include those derived from the signal recognition particle RNA (7SL RNA) and ribosomal RNA (5S RNA). Examples of 7SL RNA-derived SINEs include the primate specific Alu repeats and the rodent B1 elements. These elements actively transpose via a close relationship with the active L1 elements. The 5S RNA derived class of SINEs are the relatively newly identified SINE3 elements, which are thus far exclusive to fish and insect genomes (Kapitonov and Jurka, 2003; Kohany *et al.,* 2006). There is shared homology of the 3'-end of SINE3s with that of the CR1-like non-LTR retrotransposons and it is thought that SINE3 elements actively transposed by utilising the necessary enzymes encoded by these CR1-like elements (Kapitonov and Jurka, 2003).

The polymerase promoters of SINE elements are situated within the tRNA-related region however in the case of the non-tRNA elements such as the Alu and B1 repeats the promoter region is within the left monomer which is not descended from the 7SL related sequence and shares some sequence homology to a tRNA structure (Kriegs *et al.,* 2007). Suggesting all SINEs may be descendant from transfer RNA.

### 1.2.3.2. The active primate SINEs; Alus and SVA elements

The Alu repeats are the most abundant SINEs in the human genome and similar sequences are also present in other species for example the rodent B1 elements (Labuda *et al.,* 1991). Alus are short sequences of ~300bp, accounting for 10.6% of the total human gDNA sequence, with an estimated 1.2 million copies (Jurka, 2004). These elements derive their name from an endonuclease site located within the middle of the consensus sequence (Houck *et al.,* 1979). Alus are specific to primate genomes and first appeared ~65 million years (myr) ago and the AluY sub-family are actively propagating today (Bennet *et al.,* 2008). The species specificity of Alus is a useful tool in understanding primate lineages and has been utilised in understanding complex phylogenies including the great apes and old world monkeys (Salem *et al.,* 2003).

The SVA family of repeats have been described relatively recently and are suggested to be the youngest of the SINEs in primate genomes. SVA elements are estimated to be ~20 myr old and as such the copy number in the human genome is much less than other SINEs, due to the short activity time (Wang *et al.,* 2005, Ostertag *et al.,* 2003). The SVA repeats were originally designated SINE-R elements by Ono *et al.,* (1987) who suggested they were descendants of a retroviral fragment which was later confirmed as a HERV. The SINE-R repeat was subsequently identified in the C2 gene, which contains a variable-number-of-tandem-repeats (VNTR) locus (Zhu *et al.,* 1992). Shen *et al.,* (1994) developed this further and noted that the SINE-R.C2 element was associated with an antisense Alu sequence and were the first group to refer to these composite repeats as SVA elements, which is the acronym for the three components described (SINE-R, VNTR and Alu). SVA elements are still actively propagating within the human genome and like Alus require active L1 repeats (Ostertag *et al.,* 2003).

### 1.2.4. Mammalian-wide interspersed repeats (MIRs)

One evolutionary ancient group of SINEs and the focus of this study are the MIR elements which are identifiable in all mammals including marsupials and monotremes. MIRs comprise a family of at least seven repeat elements, which vary slightly in their consensus sequence (table 2.1; figure 3.2). All MIR families have a distinct structural arrangement and are composed of a 5'- tRNA related sequence containing the RNA polymerase III promoter, a highly conserved core-SINE and a LINE related sequence located at the 3'-end. There is a short AT rich region at the 3'-end which serves as the priming region during retrotransposition (figure 1.3). It is thought that the MIR may have arisen following the fusion of a tRNA molecule with the 3'-end of an existing LINE (Tulko *et al.,* 1997; Terai *et al.,* 1998).

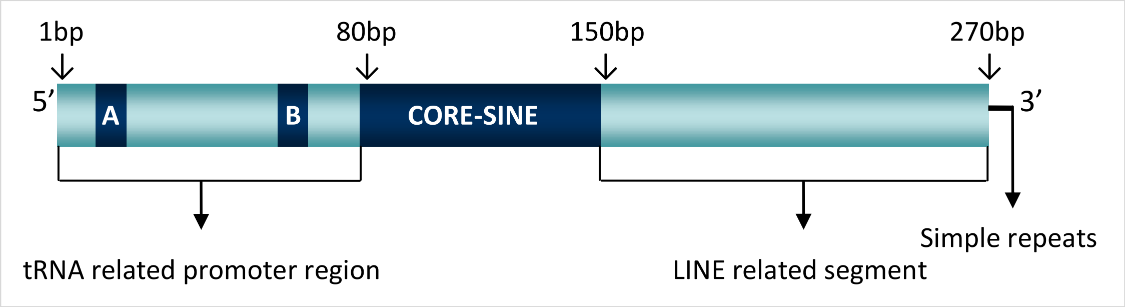


**Figure 1.3. Representation of the structure of the ancient MIR element**

The A and B box promoter regions are situated within the tRNA related region. There is a highly conserved central core-SINE sequence of ~ 70bp. At the 3'-end the MIR element shares homology to a LINE2 repeat and at the terminal end there is an AT rich reagion.

#### 1.2.4.1. Evolution and conservation of MIR elements

MIRs were first identified by Degen and Davie (1987), with a consensus sequence of ~70bp (Degan and Davie, 1987; Donehower *et al.,* 1989; Armour *et al.,* 1989). These early elements were subsequently named the mirror B1, or MB1 element, as the 70bp fragment was similar to the reverse sequence of the rodent B1 repeat, which is homologous to the primate-specific Alus (Korotkov, 1991). Subsequent analysis demonstrated that the full length MIR element was ~260bp (Smit and Riggs, 1995). The original 70bp region is currently known as the core-SINE which is highly conserved across mammals explaining the original assumption that it was the full length of the MIR.

One intriguing observation with these ancient elements is the level of conservation of the core-SINE between mammalian genomes. Considering that MIRs are inactive and have been for ~150 myr, the accumulation of mutation and divergence is less than anticipated, and the core-SINE of the MIRs has been shown to acquire mutations at a slower rate than that of neutral DNA (Sironi *et al.,* 2006). The reduced mutation rate is currently unexplained; however, this observation has lead to the premise that the core region may be providing some functionality to mammalian genomes (Smit and Riggs, 1995).

MIR elements were actively propagating prior to the radiation of mammals and before placental mammals separated; therefore the age of the MIRs was originally estimated at ~130 myr (Smit and Riggs, 1995; Jurka *et al.*, 1995). However Chaley and Korotkov (2001) suggested that the core-SINE may have originated ~550 myr ago, due to the similarities observed between the placental MIR consensus Ter-1 and the SINEs of non-mammals such as reptiles, birds and octopuses (Gilbert and Labuda, 1999).

On average 10% of mammalian genomes are derived from SINEs, with 2% being of MIR origin (table 1.1). MIR elements would have been present in a common ancestor of all mammals, as orthologue sequences are detectable in the human genome and the distant monotremes. This degree of negative selection suggests that the ancient MIR elements may confer a yet to be determined advantage to all mammalian species.

| Species | SINEs genome coverage (%) | MIR genome coverage (%) | Publication |
|---|---|---|---|
| Human | 13.63 | 2.91 | Lander *et al.,* 2001 |
| Dog | 10.57 | 2.7 | Lindblad-Toh *et al.,* 2005 |
| Cat | 11.20 | 3.1 | Pontius *et al.*, 2007 |
| Mouse | 7.96 | 0.58 | Sela *et al.,* 2007 |
| Opossum | 10.43 | 2.2 | Mikkelsen *et al.,* 2007 |
| Wooly mammoth | 6.95 | 0.71 | Zhao *et al.,* 2009 |

**Table 1.1. Percentage of mammalian genomes derived from MIR elements.**

The percentage genome coverage of MIR elements and SINEs are those estimated following the initial sequencing of the listed genome. The human genome comprises the largest percentage of SINEs with ~3% of the human, dog and cat genome being derived from MIR elements. The mouse genome has recruited the least number of MIR elements.

### 1.2.4.2. Functional role of MIR elements in mammalian genomes

There are few documented examples of MIR elements playing a specific role in gene regulation, though validated examples are slowly emerging. Previous studies have largely focussed upon the young retroelements, namely the primate-specific Alus and L1 family of elements. There are occasional validated observations of older elements, such as the MIRs, providing a function to specific genes. The gene proopiomelanocortin (POMC) encodes a neuronal hormone precursor polypeptide. In mammals, hypothalamic expression of POMC is regulated by two upstream nPE enhancer sequences (nPE1 and nPE2). The nPE2 enhancer is highly conserved in all mammals including marsupials and monotremes, and was later noted to be derived from an MIR element (Santangelo *et al.,* 2007). Smith *et al.,* (2008) also identified an MIR element 1kb upstream of the stem cell leukaemia (TAL1) transcription factor enhancer. The MIR element has been demonstrated to both regulate and significantly increase TAL1 enhancer activity. There is also a reported example of a mutation detected within an MIR element activating a cryptic splice site in the gene CYBB, leading to the human genetic disorder chronic granulomatous disease (Rump *et al.,* 2006).

One area of research which has been shown particular interest in recent years is the premise that ancient retrotransposons are a source of microRNA (miRNA) precursors and/or target sites. It has been previously hypothesised that MIR elements have the potential to form dsRNAs due to the homology between individual MIR insertions (Hughes, 2000). These dsRNA may then be cleaved into small interfering RNAs (siRNA; Zeng and Cullen, 2005). SINEs have also been suggested as a source of miRNA precursors and the miRNAs; miR-95 and miR-151* display sequence complementary to the MIR/L2 elements located within the 3’-UTRs of mammalian mRNAs (Smalheiser and Torvik, 2005). In a similar study Piriyapongsa *et al.,* (2007) identified 18 human miRNAs derived from transposable elements, 14 of which are related to the ancient L2 and MIR families.

Similar observations have been made in the dog genome, with five MIR-derived miRNA precursors formed from adjacent MIR elements recruited in opposite orientations (Zhou *et al.,* 2002). Overall it appears that the recruitment of the more ancient TEs, such as the MIRs and L2s may have played a crucial role during mammalian evolution by generating miRNA target sites and as miRNA precursors. Furthermore, the mechanism described by Smalheiser and Torvik (2005) appears to be a phenomenon exclusive to the expansion of mammalian genomes, as no similar miRNA precursors were detected in chicken, *D.* melanogaster or *C.elegans* (Smalheiser and Torvik, 2005).

#### 1.2.4.3. Distribution of MIR elements

The global distribution of MIR elements in the human genome is unknown, though they are documented to reside predominantly in GC-rich regions (Medstrand *et al.,* 2002). The number of positively identified MIRs in the human genome is regularly increasing, Smit and Riggs (1995) who first identified the full length MIR estimated there to be >300,000 copies. Following the analysis and sequencing of the human genome the figure rose to ~446,000 MIR sequences in the whole genome (Lander *et al.,* 2001). A more recent study identified ~548,000 copies in humans and ~116,000 in the mouse genome (Sela *et al.,* 2007). These figures are likely to be underestimated due to the methods of detection and it has been suggested that less than half of the total genomic MIR elements have been identified using conventional methods (Gu *et al.,* 2008).

The number of genes which have exaptated MIR elements still remains unclear. Chaley and Korotkov (2001) reported MIR sequences present in the coding sequence (CDS) of 254 human transcripts, however this number does not represent single genes and several genes are represented by multiple transcripts with others exaptated within the untranslated region (UTR). Sela *et al.,* (2007) more recently identified 181 MIRs located within the UTR and CDS of human genes, of which 78 have been recruited in the CDS (Sela *et al.,* 2007).

## 1.3. Mechanism of retrotransposition

The mechanism of how transposons replicate and move through the genome varies between classes. The DNA transposon move through a ‘cut and paste’ process and as such the element is transposed to a different genomic region, retrotransposons however propagate via a ‘copy and paste mechanism’ known as target-primed-reverse-transcription (TPRT) (Cost *et al.,* 2002). TPRT is an accumulative process as the original copy is not removed from the genome; secondly the internal promoter, poly(A) tail and ORFs are contained within the newly integrated elements. The ability of retroelements to maintain the transcriptional machinery explains the abundance of these elements within vertebrate genomes compared to the DNA transposons.

During TPRT the initial step is the transcription of the LINE into RNA via the internal RNA polymerase promoter, located within the 5’-UTR (Roy *et al.,* 2000). The RNA molecule is then transported to the cytoplasm where both ORFs are translated. These LINE-encoded proteins then bind to the RNA template to form a cytoplasmic ribonucleoprotein (RNP) complex. The RNP complex then travels to the nucleus where insertion takes place (Finnegan, 1997). The insertion process involves the cleavage of the first strand of the target gDNA, generating a 3’-OH nick, facilitated by the LINE-encoded endonuclease. There appears to be a preference for a gDNA target site of ~15bp, which contains the conserved cleavage site, 5’ TT|AAT on the antisense strand (Feng *et al.,* 1996). The 5'-end of the template RNA which contains the poly(A) tail, or in the case of MIR elements the AT rich region, anneals with the cleavage target site by base pairing with the nicked gDNA. The exposed 3’ hydroxyl serves as a primer for first strand synthesis (Luan *et al.,* 1993), this process is facilitated by the reverse transcriptase encoded by the retroelement as demonstrated in figure 1.4.

SINEs do not encode a reverse transcriptase, so in the case of SINEs, such as the MIR elements, it is thought that the reverse transcriptase of LINEs is utilised, possibly encoded by L2 and L3 elements, for their amplification (Deininger and Batzer, 2002; Terai *et al.,* 1998).

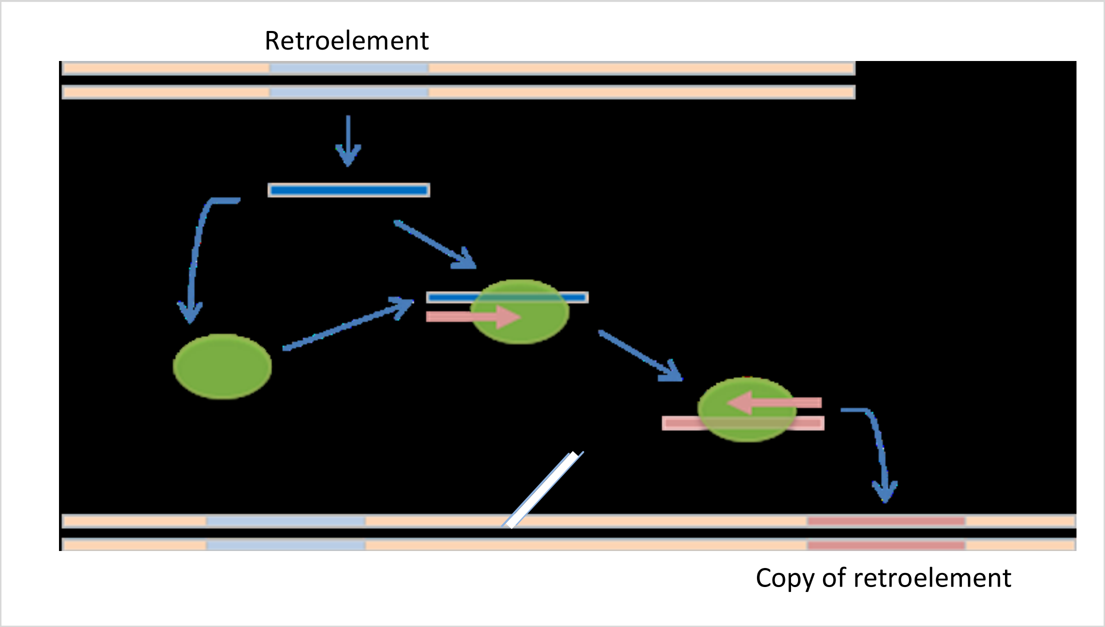


**Figure 1.4. The retroelement transposes via an RNA intermediate**

The retrotransposon is transcribed to generate the template RNA which is reversed transcribed to produce a DNA copy of the original element. The target genomic sequence is nicked which acts as a primer for target primed reverse transcription, following which the dsDNA copy is reinserted into a unique region of the genome.

Finally a second DNA strand is synthesised to form a RNA/DNA hybrid and the second gDNA strand is cleaved. The target site is filled with a double-stranded DNA (dsDNA) copy of the original retroelement. When TPRT is completed the single-stranded DNA (ssDNA) remaining at the nicked target site is filled, producing target site duplications (TSD; figure 1.5). In younger elements, these flanking TSD can be a ‘signature’ of the insertion of a retroelement into the genome (Jurka and Klonowski, 1996). Some insertions have been noted which do not contain TSD (Han *et al.,* 2005); however the TSD of older elements will contain indels and may no longer be recognisable. Several of the mechanisms involved in retrotransposition remain to be clarified. The process by which the 5’-end of the newly transcribed template RNA associates with the cleaved target DNA is unclear. Similarly the process of the degradation of the target RNA, second strand synthesis and the mechanism of target site duplication is also speculative (Cost *et al.,* 2002).

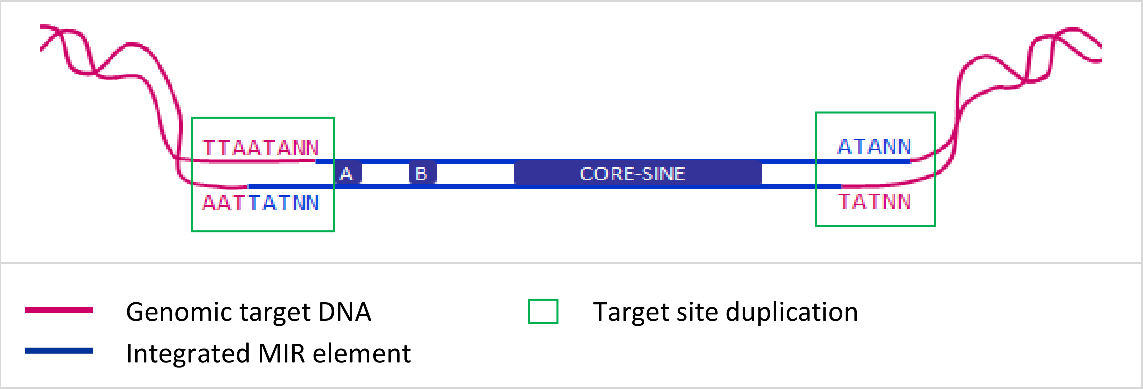


**Figure 1.5. Target site duplications flanking an integrated MIR element**

Genomic DNA is represented in pink and the MIR element in blue. The genomic target site is nicked at the consensus TTAAT which then anneals to the AT rich region at the 3'-end of the MIR consensus. When target primed reverse transcription is complete the nick is filled producing target site duplications (boxed green).

## 1.4. Retroelements, human disease, and the process of exonisation

When a retroelement has been transposed to a new genomic region it will predominantly reside within intronic sequences or intergenic regions (Sela *et al.,* 2007), with no discernible impact on gene expression patterns and protein function. However retrotransposons may be transposed directly into an exon of a gene and behave as insertional mutagens by disrupting reading frames and regulatory elements.

The mutagenic potential of retrotransposons is clearly observed following *de novo* insertions of active elements. Current retroelement activity in the human genome is largely due to the Alu and L1 families of retrotransposons. Chen *et al.,* (2005) who studied Alu and SVA repeats, noted 40 insertional events into genes which resulted in human disease (Chen *et al.,* 2005). Alu elements have been suggested to play a role in the development of 0.3% of human genetic disease (Deininger and Batzer, 1999) and an estimated 0.1% of all human germline disease is due to insertion of L1 elements (Li *et al.,* 2001). For example the insertion of an L1 into exon 14 of the factor VIII gene has been demonstrated in an haemophilia A patient (Kazazian *et al.,* 1988). L1 insertions in the dystrophin gene (DMD) have been shown to cause a frameshift and skipping of coding exon 44; responsible for the development of Duchenne muscular dystrophy (Narita *et al.,* 1993). Similar has been observed following the integration of an Alu element in the NF1 (neurofibromatosis 1) gene which causes exon skipping and shifts the reading frame resulting in neurofibromatosis type 1 (Wallace *et al.,* 1991). The *de novo* insertion of an SVA element in the α-spectrin gene (SPTA1) also results in exon skipping in cases of hereditary elliptocytosis (Hassoun *et al.,* 1994). The retrotranspositional insertion rate varies between mammals, for example retrotransposons account for 0.2% of spontaneous mutations in humans but comprise >10% in rodents (Ostertag *et al.,* 2003). However the genomes of some mammals such as the South American rodent Oryzomys appear to contain no active retroelements (Casavant *et al.,* 2000).

Most exonic integrations will be selected against and as such will no longer be detectable, and the observed exonic retrotransposons will most likely have integrated in intronic regions and were incorporated into the CDS via exonisation (Lei *et al.,* 2003). Exonisation of a retroelement occurs when the repeat contains sequences which are recognised as splicing motifs, or alternatively mutations within the recruited element may generate cryptic splice sites or optimise an existing splicing signal. In the majority of cases, the recruitment of a retroelement will introduce an alternative reading frame, which may ultimately result in a truncated protein of either no function or with altered expression levels (Krull *et al.,* 2005). Premature termination codons may trigger nonsense-mediated decay (NMD) of the mRNA transcript if a polyadenylation signal is not in close proximity (Chang *et al.,* 2007).

One of the earliest observations of the exonisation of a retrotransposon occurred more that 15 years ago. Mitchell *et al.,* (1991) studying mutations in the ornithine-aminotransferase gene (OAT), demonstrated that a point mutation within an Alu element (residing within an intron) produced a cryptic splice site. This aberrant splicing generated an alternative reading frame and thus a premature in-frame stop codon (Mitchell *et al.,* 1991). The truncated protein is a contributing factor in ornithine aminotransferase deficiency, the causative agent of the autosomal recessive eye disease, Gyrate Atrophy (Simell and Takki, 1973).

A further example of Alu exonisation is noted following a deletion located within an Alu element, residing within an intron of the glucuronidase-beta gene (GUSB). The deletion produces a cryptic donor splice site and as a consequence individuals develop

Mucopolysaccharidosis type VII (Sly syndrome). This mutation produces an alternative exon which results in the skipping of a constitutive exon, therefore causing a mild form of Sly syndrome (Vervoort *et al.,* 1998). Numerous other cases of Alu-based exonisation have been documented (for a review see Sorek *et al.,* 2002).

There is one reported example where an inserted MIR element contributes to a human genetic disorder. In a chronic granulomatous disease (CGD) patient, Rump *et al.,* (2006) identified a mutation within the CYBB gene (cytochrome b-245, beta polypeptide) which activated a cryptic splice site in the middle of an intron. Inclusion of this novel 56bp exon results in a reading frame with a termination codon, which encodes a truncated protein. Previous investigations have reported the exonisation of LTR elements, Alus and MIR repeats (Piriyapongsa *et al.,* 2007; Sela *et al.,* 2007; Krull *et al.,* 2007), and the Alus have been shown to contain motifs in the repeat consensus sequence resembling classical splice sites (Sorek *et al.,* 2002).

## 1.5. Transposable elements as markers for evolution

SINE elements are useful tools for defining monophyletic clades and assist in understanding complex lineages. SINEs help identify ancient radiations, which would otherwise be undetectable due to the accumulation of mutations over time. The majority of SINEs insert into a genome with little or no impact and therefore provide a record of biological history and as such are considered as neutral genetic markers (Shedlock and Okada, 2000).

The use of SINEs can prove more effective at studying evolutionary relationships between species than single-nucleotide-polymorphisms (SNPs) and nucleotide substitutions, as they may have integrated following numerous independent mutations and are not inherited from a common distant ancestor. SINEs are thought to represent identity by descent and the probability of two individual elements integrating independently at the exact location of a chromosome is small (Salem *et al.,* 2005). Therefore the problem of parallel mutation and ambiguous results due to outgroup selections is less likely (Shedlock and Okada, 2000). Transposable elements are a useful tool at deciphering complex and/or closely related lineages, due to the mode of evolution being unidirectional, as there is currently no known mechanism for the removal of SINEs from a genome following recruitment (Salem *et al.,* 2005; Shedlock *et al.,* 2004). However there are some caveats when conducting phylogenetic analysis using transposable elements, for example it is possible that insertion homoplasy may occur between species of different ancestry over time. Furthermore repeat sequences can be lost by deletion (Salem *et al.,* 2005)

### 1.5.1. Retrotransposons and complex lineages

The African cichlid fish (AFC) family are found in Lakes Victoria, Malawi, and Tanganyika, and have captured the interest of researchers for many years. This endemic species exhibits extraordinary levels of diversity for each lake as a consequence of explosive adaptive radiation that occurred independently and in parallel (Terai *et al.,* 2004; Terai *et al.,* 2003; Takahashi and Okada 2002). AFC SINE elements have been used as genetic markers for phylogenetic analysis to delineate the complex radiation of the cichlid fishes. Similarly, analysis of Alu lineages has been used to assess relationships between the primate lineage of humans, gorillas and chimpanzees (Salem *et al.,* 2003; Roy-Engel *et al.,* 2002). The study assisted in understanding primate phylogeny and to reconstruct the demographics of human populations. The sister relationship between humans and chimpanzees is clearly distinguishable following the application of SINEs (Salem *et al.,* 2003). SINE elements have been used to elucidate the origin of whales. Shimamura *et al.,* (1997) determined that the order Cetacea (whales, porpoises and dolphins) form a monophyletic group with ruminants and hippopotamuses and not Atriodactyla, as had been previously suggested. This was supported with the systematic use of SINEs by Nikaido *et al.,* (2001).

## 1.6. Functionality of retrotransposons

The integration of a retroelement will provide additional genomic material and therefore may have contributed to genome plasticity, allowing for accelerated evolutionary changes in protein structure and gene expression (Häsler and Strub, 2006; Szmulewicz *et al.,* 1998). There is increasing evidence to support the idea that retrotransposons allow for the modification of gene function and/or expression by providing regulatory elements, alternate splice sites, polyadenylation signals and additional protein coding information (Nekrutenko and Li, 2001; Lev-Maor *et al.*, 2003; Baertsch *et al.,* 2008).

### 1.6.1. Pseudogenes

Pseudogenes represent copies of functional genes which have been inserted into a novel location within the genome, and generally have no protein coding capacity, due to the failure of transcription or translation. The nucleotide sequence will be highly similar to the parental gene but may contain mutations or lack functional sequences, as a consequence the reading frame may be interrupted and regulatory elements rendered inactive. There are two classes of pseudogene; processed pseudogenes are generated by reverse transcription of RNA whereas non-process pseudogenes are derived directly by an increase in DNA content. Non-processed pseudogenes usually spread via gene duplication, so are commonly clustered and situated near the original functional gene. This class of pseudogenes contain introns, flanking sequences, including the upstream promoter and will usually contain multiple in-frame stop codons. DNA-derived pseudogenes theoretically have the potential to be functional if the original promoter is included; however the majority are present as redundant copies which have accumulated mutations.

Processed pseudogenes are generated by the reverse transcription of mRNA, tRNA and ribosomal RNA (rRNA) and are also termed retrogenes (Krasnov *et al.,* 2005). This category of pseudogenes are thought to have relied on retrotransposons for their amplification (Esnault *et al.,* 2000), and many have flanking retrotransposons and share the poly(A) tails and TSD of the repeat elements (Deininger and Batzer, 2002). These pseudogenes will contain the features of the original RNA so are intronless and commonly contain the original poly(A) tails. Processed pseudogenes are usually truncated at the 5’-end and as such lack upstream promoters, and processed pseudogenes are generally considered as “dead” following integration (Graur *et al.,* 1989). Even though processed pseudogenes are considered “dead on arrival”, some remain as protein-coding intronless retrogenes (Marques *et al.,* 2005). Retrogenes may not encode for the gene product of the original paralogue gene as there is no need to maintain two genes of identical function, secondly processed pseudogenes are usually truncated at the 5’-end resulting in an alternative reading frame and hence a different peptide sequence. Some pseudogenes may remain as functional ncRNAs (ncRNA) (Sasidharan and Gerstein, 2008). With all retrotransposons being derived from RNA genes it is thought that the original ‘master’ repeat elements may have been a successful pseudogene which continued to retrotranspose effectively (Shen *et al.,* 1991; Deininger *et al.,* 1992).

Functional processed pseudogenes are often derived from X-linked genes, for example phosphoglycerate-kinase (PGK)2. PGK2 is a glycolytic enzyme which is encoded by an intronless gene derived from the intron-containing paralogue PGK1 (McCarrey *et al.,* 1996). PGK1 has a similar function to PGK2, but PGK2 is expressed exclusively in spermatocytes (McCarrey *et al.,* 1996). Uechi *et al.,* (2002) identified a new family of ribosomal protein genes designated RPL10L, PRPL36AL and RPL39L, which appear to have originated through the retrotransposition of an X-linked ribosomal gene. These ribosomal retrogenes are primarily expressed in the testes and are thought, as with PGK1 to have evolved as a means of compensating for the reduced dosage of X-linked genes, which may be silenced during spermatogenesis (Uechi *et al.,* 2002). This mechanism would presumably allow for the processed copy of the gene to specialise in testes function while the parental gene maintains the original function in somatic tissues (Shiao *et al.,* 2007). Other pseudogenes, which seem to have evolved as a means of compensating in this manner, include glucose-dehydrogenase (G6pd), X-linked XAP-5 (FAM50B) and the testes-specific pyruvate-dehydrogenase subunit (PDHA2) (Dahl *et al.,* 1990; Hendrikson *et al.,* 1997; Sedlacek *et al.,* 1999).

Retrotransposition has also contributed to the evolution of salivary amylase. The amylase gene family consists of two pancreatic amylases and three salivary amylases. Salivary amylase (AMY1) is thought to have evolved from pancreatic amylase (AMY2), a process which was facilitated by the insertion of a processed gamma-actin sequence into the promoter of the original ancestral amylase gene (Ting *et al.,* 1992). This event is likely to have been selected for due to the need for salivary amylase when there was an increase in the ingestion of starch in the human diet following the development of agriculture (Caldwell *et al.,* 2004).

### 1.6.2. Retrogenes and mammalian evolution

More than 90% of G protein-coupled receptors (GPCRs) are encoded by an intronless gene. It has been suggested that these genes share a common ancestor, which may have been a retrogene (Gentles and Karlin, 1999). Being intronless suggests that a gene may be derived following retrotransposition as processed-pseudogenes are transposed from mature mRNA. Eliminating post-transcriptional splicing would result in a faster rate of protein expression, which may be advantageous to GPCRs. This improved efficiency would be beneficial when there is a need for rapid protein expression, such as in the central nervous system, leading to the idea that retrotransposition was important in the evolution of increased cognitive capacity in humans (Gentles and Karlin, 1999).

Glutamate dehydrogenase (GLUD)1 is a mitochondrial matrix enzyme, which functions in glutamate and nitrogen metabolism (Smith *et al.,* 2001). GLUD2 is an intronless pseudogene of GLUD1, derived from GLUD1 by retrotransposition during a time that coincides with a period of a notable acceleration of evolution and increase in brain size of hominoids (Burki and Kaessmann, 2004). It is thought that higher neural activity co-evolved with greater brain size, and may have been aided by retroelement activity (Burki and Kaessmann, 2004). Retrotransposition has also been implicated in other areas of primate evolution. Marques *et al.,* (2005) estimate that there are ~75 retrogenes which emerged during and after the primate burst of retrotransposition. It was also noted that the majority of these retrogenes were expressed in the testes and originated from X-linked genes, and have evolved functional roles in spermatogenesis. It is suggested by Marques *et al.,* (2005) that the compensation of X-linked genes through retrotransposition may have assisted human evolution, as autosomal substitution during X inactivation would primarily enhance male germline function allowing for rapid primate evolution (Marques *et al.,* 2005).

The total number of human retrogenes currently identified is 163 and similar numbers are observed across mammals (Pan and Zhang, 2009). There also appears to be a burst of young retrogenes appearing in mammals occurring prior to that discussed with primates, which may have aided mammalian evolution (Pan and Zhang, 2009).

### 1.6.3. Imprinting, epigenetics and microRNAs

Genetic imprinting occurs when a gene is differentially expressed, depending on the origin of the inherited allele, either maternally or paternally. The differential expression is referred to as the parent-of-origin effect. There are ~80 human imprinted genes, at least 11 of which are retrogenes (Morison *et al.,* 2005). Several others display the features of retrotransposed genes and many retrogenes are located within imprinting regions (Walter and Paulsen, 2003). Due to this commonality of retrogenes with imprinted genes it is thought that imprinting may have occurred as a secondary effect following the defence against foreign DNA, as active repeat elements are rendered inactive by DNA methylation (Wood *et al.,* 2007).

It has been suggested that a large number of microRNAs (miRNA) may have evolved from TEs, as the genome responds to a retrotranspositional event by attempting to deactivate and degrade the TE via DNA methylation and RNA degradation (Suzuki *et al.,* 2007). Silencing of TEs by genomes is mediated by siRNAs, which are derived from the repeat element target itself (Piriyapongsa *et al.,* 2007; Buchon and Vaury, 2006; Vaughn and Martienssen, 2005). For example the epigenetic machinery present in plants is thought to have evolved in this manner (Zilberman and Henikoff, 2005), and miRNAs have been demonstrated to be encoded by a short DNA transposon known as a MITE in *A.thaliana* and *O.sativa* (Piriyapongsa and Jordan, 2008). MiRNAs can also be transcribed by the RNA polymerase III promoter of the Alu elements (Borchert *et al.,* 2006).

Evidence is increasing to support the idea that miRNA precursors are derived from TEs; including MIR/L2 repeats (Smalheiser and Torvik, 2005), L1 elements (Devor *et al.,* 2009) and Alu repeats (Smalheiser and Torvik, 2006; Devor, 2006). Piriyapongsa *et al.,* (2007) identified 14 human miRNAs which appear derived from L2 and MIR elements. Similar observations have been made with the dog genome. Five MIR-derived miRNA precursors were identified by Zhou *et al.,* (2002), who noted that stem-loop structures are formed between adjacent MIR elements when recruited in opposite orientations.

## 1.7. Concluding remarks

The majority of studies have focussed on the analysis of active retrotransposons such as Alus and L1 repeat sequences. There has been limited investigation of the exaptation of MIRs (Krull *et al.,* 2007; Sela *et al.,* 2007; Chaley and Korotkov, 2001). MIR elements have been resident in the genome since the mammalian radiation and are ubiquitous across mammals; therefore they have the potential to provide insight in mammalian gene evolution through history. Furthermore the highly conserved core-SINE region is not only conserved between different genes of a single species but between divergent species. However to date few groups focus on this particular class of elements, possibly due to the level of divergence observed with respect to the age of the elements.

Previous publications have identified MIR-containing genes (Chaley and Korotkov, 2001; Sela *et al.,* 2007), although the number of coding transcripts remains small and none have systematically looked for a functional role of these ancient elements. A preliminary analysis of the human genes which have recruited MIRs identified a possible link with dendritic localisations of mRNA transcripts to the neurone (Hughes, 2000; Hughes, unpublished data). This localisation has been implicated in synaptic plasticity, and a role in neuronal function has been supported by the discovery of mutations in some of these genes in cases of mental retardation.

## 1.8. Aims and objectives of the study

The primary aim of the project is to identify MIR-containing genes, which following bioinformatics and laboratory analysis may provide insight into the potential role of MIR elements in mammalian gene evolution, and to determine the functional significance of the exaptation of this family of repeats.

The key areas are as follows:

- To identify human genes which have exaptated MIR elements by the use of bioinformatics and data mining tools.

- Following acquisition of the MIR-containing genes to outline the distribution of these repeat sequences by determining if there is a preference to recruit MIR elements in a particular orientation, gene/genomic region or of a particular MIR sub-family.

- To establish if there is a commonality in function between the MIR-containing genes, such as protein function, interactions and biological pathways.

- Determine if exonisation of the MIRs has occurred to produce transcript variants which may contribute to human disease, differential expression or gene function.

# 2. MATERIALS AND METHODS

## 2.1. Bioinformatics analysis

### 2.1.1. Acquisition of MIR-containing genes

Retrotransposon consensus sequences were obtained from the Repbase Update which is provided by the Genetic Information Research Institute (GIRI; table 2.1). The Repbase Update is a reference source and database of repetitive DNA for numerous eukaryotic genomes (Jurka *et al.,* 2005; http://www.girinst.org/repbase/update/index.html).

| Repeat Type | Repeat Sub-type | Aliases | Size (bp) | Species origin |
|---|---|---|---|---|
| SINE | MIR | MB1, MER24, MIR1 | 262 | *Mammalia* |
| SINE | MIRc | - | 268 | *Mammalia* |
| SINE | MIR3 | L3, MIR | 224 | *Mammalia* |
| SINE | MIRb | SINE2 | 268 | *Mammalia* |
| SINE | MIRm | MON1, THER1, THER2 | 273 | *Monotremata* |
| SINE | MIR_Mars | THER1, THER2 | 263 | *Metatheria* |
| SINE | THER1_MD | - | 275 | *Didelphidae* |

**Table 2.1. MIR element sub-families contained in RepBase**

The MIR element subfamilies are catagorised according to the RepBase annotations. The size of the consensus sequences are in the range of 224-275bp. The earliest mammal the MIR element is thought to have been actively transposing is listed if known.

Potential MIR containing nucleotide sequences were collected by screening human genes for homology to the MIR consensus using the 'BLAST' analysis program (Altschul *et al.,* 1990; http://www.ncbi.nlm.nih.gov/BLAST/). Repeat elements were alternatively identified by searching the 'Genomic ScrapYard', which is a database compiled of transcripts that are known to contain all families of transposable elements including SINEs, LINEs, DNA and LTR transposons (Makalowski, 2000; http://warta.bio.psu.edu/ScrapYard/). Only a proportion of the MIR-containing genes identified in this study were available in the Genomic ScrapYard; however, it was a useful resource.

### 2.1.2. Validation and annotation of MIR elements

'RepeatMasker' was the primary tool utilised to confirm the presence of an MIR element. The 'RepeatMasker' is a software resource which screens query sequences for interspersed repeat elements (http://www.repeatmasker.org/) using the transposons consensus sequences in the Repbase. The orientation of the repeat sequence within the transcript can be determined, as can the coordinates, size and the repeat sub-family. Transcripts which were noted to have recruited an MIR element were validated by assigning the sequence to a specific gene (http://www.ncbi.nlm.nih.gov/gene/) and confirming that the sequence was either a current reference sequence or by searching for ESTs (expressed sequence tags) which correspond to the nucleotide and by noting published examples. Only annotated 'named' genes were included in this study and hypothetical sequences discarded as they are unconfirmed sequences which may represent sequencing errors, pseudogenes or non-functional 'fossil' genes.

### 2.1.3. Repeat sequence location

The position of the MIR element within a gene was determined using the sequence alignment program 'Spidey', and 'Entrez Gene'. Both sites are located at NCBI (http://www.ncbi.nlm.nih.gov/IEB/Research/Ostell/Spidey/; http://www.ncbi.nlm.nih.gov/gene/). Spidey is an alignment tool which when aligning mRNA to the corresponding gDNA determines the exonic arrangement and provides the genomic and mRNA coordinates for each exon determined. Splice sites were then scored using the online GENIE program available at the Zhang laboratory website (http://rulai.cshl.edu/new_alt_exon_db2/HTML/score.html; Reese *et al.,* 1997; Zhang, 1998).

The genomic locations for all of the human genes identified were mapped using 'KaryoView' at Ensembl (http://www.ensembl.org/Homo_sapiens/karyoview). The density of human genes for a given chromosome was calculated according to the NCBI 36.2 assembly of the human genome, which is currently used in the Ensembl release 43. The coordinates for all of the MIR subfamilies were plotted using the SUMPRODUCT function available in excel, which calculated the number of total transcripts that corresponds to each nucleotide of the MIR consensus sequence.

### 2.1.4. Sequence homology

Multiple sequence alignments were generated using the EBI ClustalW2 server (http://www.ebi.ac.uk/Tools/clustalw2/index.html; Larkin *et al.,* 2007). Clustal alignments of >20 sequences were generated using the WebLogo browser application (Crooks *et al.,* 2004; http://weblogo.threeplusone.com/). WebLogo provides a clear graphical description of sequence conservation. The alignment generates a stack of four letters representing each nucleic acid (A, C, G, T) with the overall height signifying the frequency of each nucleotide. The top letter in the alignment is most conserved, allowing for the identification of patterns otherwise difficult to perceive when aligning large numbers of sequences with ClustalW2. Sequence conservation between species was viewed using the ECR browser (http://ecrbrowser.dcode.org/; Ovcharenko *et al.,* 2004), which displays evolutionary conserved genomic regions.

### 2.1.5. Protein function and peptide sequences

Nucleotide sequences were translated into the six potential reading frames using the translator feature at JustBio (http://www.justbio.com/tools.php) and conserved protein domains were obtained from the Pfam database of protein families (http://pfam.sanger.ac.uk/; Finn *et al.,* 2008) and the Conserved Domains and Protein Classification (CDD) database (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml).

### 2.1.6. Functional enrichment analysis

The significance of the MIR dataset of genes was further investigated using the web-based functional annotation tool D.A.V.I.D (Database for Annotation, Visualisation and Integrated Discovery; Huang *et al.,* 2007; http://david.abcc.ncifcrf.gov/home.jsp). The gene identifiers for the MIR-containing genes were uploaded to D.A.V.I.D and screened against a large number of databases for functional enrichment. Databases accessible through D.A.V.I.D, used during the investigation include OMIM (Online Mendelian Inheritance in Man; http://www.ncbi.nlm.nih.gov/omim/*),* the Genetic Association Database (GAD; http://geneticassociationdb.nih.gov/), the Gene Ontology database (http://www.geneontology.org; section 2.1.6.1), SwissProt (http://www.expasy.ch/cgi-bin/keywlist.pl), KEGG for pathway maps (http://www.kegg.com; Kanehisa *et al.,* 2008; section 2.1.6.2) and the U133A/GNF1H tissue expression panel (http://wombat.gnf.org/; Su *et al.,* 2004). The U133A /GNF1H human expression data includes a panel of 79 tissue typesand the corresponding data for >40,000 human genes.

### 2.1.6.1. Gene Ontology

Gene ontology (GO) data was retrieved from the GO consortium and EBI gene ontology annotation service (http://www.ebi.ac.uk/GOA/; http://www.geneontology.org). The gene ontology is a controlled vocabulary used to describe gene function, products and sub-cellular locations. The total number of human genes which have recruited MIRs that have been assigned a specific GO term were determined using the data mining tool, 'MartVeiw' located at Ensembl (http://www.ensembl.org/Multi/martview) and FatiGO at Babelomics v3.1 (Al-Shahrour *et al.,* 2007; http://babelomics.bioinfo.cipf.es). The frequencies of MIR-containing genes which have a particular GO term assigned were compared to the frequency of the total number of genes in the human genome with the same GO accession. If a GO term was overrepresented for the MIR group of genes compared to the genome may suggest function, statistical significance was calculated using D.A.V.I.D.

### 2.1.6.2. Regulatory pathways

The involvement of the MIR-containing genes in metabolic and signalling pathways was determined by comparing to the known regulatory pathways listed in the Kyoto Encyclopaedia of Genes and Genomes (KEGG) database (Kanehisa *et al.,* 2008; http://www.kegg.com). The significance of the MIR-containing genes involved in each pathway was determined using the web-based analysis tool D.A.V.I.D.

### 2.1.6.3. MicroRNAs (miRNA) and miRNA binding sites

Potential microRNA (miRNA) binding sites were identified by accessing the data available in the 'miRBase', the miRNA registry available at the Sanger Institute, which contains >5000 miRNA gene loci (http://microrna.sanger.ac.uk/). The most effective approach is to use 'miRBase Targets', firstly searching for a gene of interest, then identifying if the transcript with the MIR is present and finally to establish if the target site is located within the MIR sequence, this can be further verified using ‘Spidey’. Alternatively miRNA target sites can be detected using ‘TargetScan’ (Lewis *et al.,* 2003; http://www.targetscan.org/).

### 2.1.6.4. Regulatory RNA, motifs and elements

A query sequence was screened for regulatory RNAs such as splice enhancers, splice silencers, motifs within the 3’- and 5’-UTRs using the ‘RegRNA’ software (Huang *et al.,* 2006; http://regrna.mbc.nctu.edu.tw/html/prediction.html). Alternatively 5’- and 3’-UTR motifs were identified in a query sequence using UTRScan (Mignone *et al.,* 2005; http://www2.ba.itb.cnr.it/UTRSite/). Motifs and elements were only accepted as true if there were publications to verify the sequence.

### 2.1.6.5. Disease associations

The involvement of gene transcripts in human disease was determined by visiting OMIM at NCBI (http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim) and AceView also at NCBI (http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/index.html), details were confirmed by current journal articles. Disease association of the MIR-containing genes was also determined using the Genetic Association Database (GAD; http://geneticassociationdb.nih.gov/) accessible through D.A.V.I.D. Databases for specific human diseases were also visited, including Retina International (http://www.retina-international.com/sci-news/disloci.htm) and the Retinal Information Network (RetNet$^{\mathrm{TM}}$; http://www.sph.uth.tmc.edu/retnet/).

### 2.1.7. Statistical analysis

Calculating a $P$ value which entails multiple tests may produce false positives, therefore the Fisher's exact test was adjusted using the false discovery rate (FDR) corrected method (Al-Shahrour *et al.,* 2007; Benjamini and Hochberg, 1995). $P \leq 0.05$ (95% confidence limit) was deemed significant and $P < 0.01$ (99% confidence limit) was considered extremely significant. The unadjusted Fisher’s exact test was used to provide a general indication of any trends which may suggest a potential function of the MIR-containing genes, though without statistical significance.

## 2.2. Tissue preparation, homogenisation and RNA extraction

### 2.2.1. Tissue and cell culture preparation

Adult male and female Wistar rats were sacrificed by cervical dislocation in accordance with the Animals (Scientific Procedures) Act 1986, UK. Tissues were rapidly excised in a sterile environment, rinsed in 0.1% (v/v) DEPC-treated $H_2O$ and flash-frozen in liquid nitrogen. The rat tissues used in this study are detailed in table 2.2.

| Tissue | Source (Adult Winster Rats) | Experimental Procedure |
|---|---|---|
| Brain | Male x 3; Female x 2 | RT-PCR |
| Brain | Male x 2; Female x 3 | Radioactive *in situ* hybridisation |
| Heart | Male x 3; Female x 3 | RT-PCR |
| Kidney | Male x 3; Female x 3 | RT-PCR |
| Liver | Male x 3; Female x 3 | RT-PCR |
| Lung | Male x 3; Female x 3 | RT-PCR |
| Pancreas | Male x 3; Female x 2 | RT-PCR |
| Spleen | Male x 3; Female x 3 | RT-PCR |
| Testis | Male x 3 | RT-PCR |

**Table 2.2. Dissected rat tissues used for RT-PCR and *in situ* hybridisation experiments**

Whole dissected rat tissues have been listed with the number of animals sacrificed, the sex of the animal and the tissue excised for each procedure.

### 2.2.2. Extraction of total RNA

The tissues were weighed and diced prior to homogenisation. 1ml of TRI reagent® (Ambion; containing phenol) was added per 100mg of tissue and per 5-10 x $10^6$ of cultured cells. Cells were homogenised by pipette mixing and animal tissue was homogenised for a period of 30 seconds to a few minutes for hard or fibrous tissue using a manual glass-teflon homogeniser (Scientific Laboratory supplies Ltd). C6 cells were donated by Dr A Hargreaves (Nottingham Trent University) and RNA was isolated (section 2.2.2). The BRIN-BD11 cell line was donated by Dr E Verderio-Edwards (Nottingham Trent University) and grown in culture (section 2.8).

The homogenised cells/tissues were aliquoted 1ml per 1.5ml Eppendorf tube and incubated at room temperature in the fume cupboard for 5 min. 200µl of chloroform was added to each suspension, and gently vortex mixed (Vortex-Genie® 2; Sigma) for 10 sec and incubated on ice for 5 min. The samples were then centrifuged at 12000 *x g* for 20 min at 4°C (Mikro 20 microcentrifuge (Hettich) for all instances unless stated otherwise). Following centrifugation the suspension consists of three layers; protein, fat/lipids and nucleic acids, the upper clear aqueous layer containing the RNA/DNA mix was transferred to a fresh Eppendorf tube. If the intermediate phase was disturbed the solution was re-centrifuged. 500µl of isopropanol was added to each tube containing the supernatant, vortex mixed briefly and incubated on ice for a further 10min. The samples were centrifuged at 12000 *x g* for 10 min at 4$^{o}$C. The supernatant was removed by gentle aspiration using a micropipette and the pellet washed using 1ml of 75% (v/v) ethanol. The pellet was centrifuged at 7500 *x g* for 10 min at 4$^{o}$C, following which the ethanol was removed by gentle aspiration and the pellet was left to air dry in the fume cupboard for 2-5 min until almost dry. The isolated RNA was re-suspended in 20-100µl of RNase and DNase $H_2O$ depending on the pellet size. For large pellets re-suspension was aided by heating at 50$^{o}$C for 5 min. The concentration of the RNA was calculated using a spectrophotometer (section 2.4.3). The RNA was used directly for cDNA synthesis or stored in 2.5 volume absolute ethanol at -80$^{o}$C until required.

## 2.3. The storage, recovery, concentration and quantification of nucleic acids

### 2.3.1. Storage and ethanol precipitation of nucleic acids

Isolated RNA and cDNA were recovered from storage and concentrated by precipitating with ethanol. The nucleic acids were precipitated in 0.1 volume of sodium acetate (3M, pH5.2) and 2.5 volume absolute ethanol. The precipitating RNA was inverted to mix and stored at -80$^{o}$C for 30min. The suspension mix was centrifuged at 12000 *x g* for 20 min at 4$^{o}$C. For high concentrations of RNA (>150µg) the suspension was centrifuged for 30 sec and half of the solution aliquoted to a fresh Eppendorf tube and centrifugation of both continued for 20 min. The supernatant was removed by gentle aspiration using a micropipette; 1ml of 75% (v/v) ethanol was added to the pellet and re-centrifuged at 12000 *x g* for 5 min at 4$^{o}$C. The supernatant was carefully removed using a micropipette and the pellet was left to dry in the fume cupboard for 3-5 min. The nucleic acids were re-suspended in RNase and DNase free $H_2O$, for stock which may be subject to numerous freeze thaw cycles 1X TE buffer (10mM Tris-HCl, 1mM EDTA; pH 8.0) was used as an alternative.

### 2.3.2. RNA storage, purification using FTA cards

Homogenised tissue and cells suspended in phosphate-buffered saline (PBS; 137mM NaCl, 2.7mM KCl, 10mM $Na_2HPO_4$, 1.8mM $KH_2PO_4$; pH 7.6) were alternatively stored on FTA micro cards (Whatman®). The tissue culture cells were applied to the FTA card evenly (125µl) with a minimum concentration of 100 cells/µl. The card was then left to dry for 1hr at room temperature after which the FTA card was stored at -20°C to extend the storage time. Tissues which did not contain a high content of fibrous or extracellular material were homogenised in PBS and also applied to FTA cards and stored in the same manner. The RNA processing buffer required for extraction of the RNA from the FTA card was made up as follows:

| | |
|---|---|
| Tris-HCl (10mM) to final volume of | 400µl |
| Glycogen (200µg/ml; Sigma) | 40µl |
| EDTA (0.1mM) | 20µl |
| RNasin (40U/µl; Promega) | 4µl |
| DTT (dithiothreitol, 2mM; Promega) | 2µl |

A sample disc was punched from the FTA card using a 2mm Harris Micro-Punch (Whatman®) over a cutting mat. The FTA disc was ejected into a sterile 0.5µl Eppendorf tube containing the 400µl of RNA processing buffer and the sample was incubated on ice for 15 min and vortex mixed at 5 min intervals. The RNA was precipitated from the wash solution by adding 0.1 volume of NaAc (3M; pH 5.2) and 1 volume of absolute isopropanol. The sample was then incubated at -20°C for 1 hr following which the sample was centrifuged at 12000 *x g* for 5 min at 4°C and the supernatant removed. The pellet and disc were washed by adding 500µl of 75% (v/v) ethanol and centrifuged at 12000 *x g* for 5 min at 4°C. The supernatant was removed and pellet air dried in the fume cupboard for 3-5 min. The RNA pellet was then resuspended in 25µl of TE buffer (10mM Tris-HCl, 0.1mM EDTA; pH 8.0), less EDTA was used to reduce the interaction of the EDTA with the the FTA disc. The total RNA yield per extraction was approximately 12ng/µl.

### 2.3.3. Quantification and purity of RNA

RNA solutions were measured at absorbance 260nm and 280nm using a DU® 530 UV/vis spectrophotometer (Beckman Coulter). Samples were diluted 1:100 with 0.1% (v/v) DEPC-treated $H_2O$, total 0.1% (v/v) DEPC-treated $H_2O$ was used to calibrate the spectrophotometer. The purity of the RNA sample was determined by the ratio of $A_{260nm}$ /$A_{280nm.}$ The RNA concentration (µg/ml) was calculated as follows:

$$A_{260nm} \text{ x dilution factor x } 40\mu g/ml.$$

## 2.4. First strand synthesis of complementary DNA

Complementary DNA (cDNA) was synthesised from 1µg of total RNA, isolated from several rat tissues (section 2.2.1). A human multiple tissue cDNA panel was purchased from BD Clontech, UK (Human MTC Panel I). All human cDNAs in the panel are free of genomic DNA have been fully normalised to four house keeping genes.

### 2.4.1. Removal of genomic DNA (gDNA) and phenol /chloroform precipitation

Before cDNA synthesis could commence gDNA had to be broken down and removed. The following components (RQ1 components from Promega) were added to a 1.5ml eppendorf tube and incubated at 37°C for 1 min:

| | |
|---|---|
| Total RNA (25µg) | 20-30µl |
| 10x RQ1 RNase-free DNase buffer (Promega) | 5µl |
| RQ1 RNase-free DNase (1U/µg RNA; Promega) | 25U |
| RNase and DNase free $H_2O$ to final volume of | 50µl |

Following incubation, 5µl of RQ1 DNase stop solution (20mM EDTA (pH 8.0); Promega) were added and the sample incubated further at 65 °C for 10 min.

### 2.4.2. Phenol chloroform precipitation and wash

The RNA was purified to remove the digested DNA, protein and salt. An equal volume of phenol/chloroform/isoamyl alcohol (PCI; ratio of 25:24:1; pH 4.5) solution was added to the tube. The RNA solution was manually shaken for 30 sec and centrifuged at 12000 *x g* for 5 min at 4°C. The upper aqueous phase was transferred to a fresh 1.5ml Eppendorf tube and an equal volume of chloroform/isoamyl alcohol (CI; ratio of 24:1) added. The mixture was shaken for a further 30 sec and centrifuged at 12000 *x g* for 5 min at 4°C. The upper aqueous phase was transferred to a fresh Eppendorf tube and an ethanol precipitation was performed (section 2.4.1). The purified RNA pellet was resuspended in 50μl of 0.1% (v/v) DEPC-treated $H_2O$ and incubated at 65 °C for 5 min.

### 2.4.3. cDNA synthesis with RNA isolated from tissue and cells

Complementary DNA was synthesised by adding the following components to a sterile 0.5ml Eppendorf tube:

| | |
|---|---|
| Total RNA | 1μg |
| Oligo(dT)$_{15}$ primers (0.5μg/μl; Promega) | 1μl |
| RNase and DNase free $H_2O$ to final volume of | 10μl |

The RNA suspension was incubated at 70°C for 5 min to allow for primer annealing, after which the mix was cooled on ice and the remaining components were added to the RNA mix:

| | |
|---|---|
| 5x M-MLV buffer (Promega) | 5μl |
| deoxyNTP mixture (10mM; Promega) | 1μl |
| RNasin (40U/μl; Promega) | 0.5μl |
| RNase and DNase free $H_2O$ to final volume of | 25μl |

The sample was incubated at 37°C for 5 min, cooled on ice and 1μl of M-MLV reverse transcriptase (200U/μl; Promega) was added. The reaction mixture was further incubated at 37°C for 70 min and cooled on ice. The sample was then vortex mixed briefly and heated at 90°C for 10 min to stop the reverse transcriptase activity. The cDNA was stored at -20°C or used directly for RT-PCR. The cDNA panel of rat tissue was normalised for subsequent RT-PCR experiments. All cDNA synthesis reactions were performed with 1µg of total RNA which was quantified spectrophotometrically (section 2.4.4). Following quantification RT-PCR was performed using primers specific to a fragment of the house-keeping gene GAPDH, to visualise the purity and concentration of each cDNA template (figure 2.1).

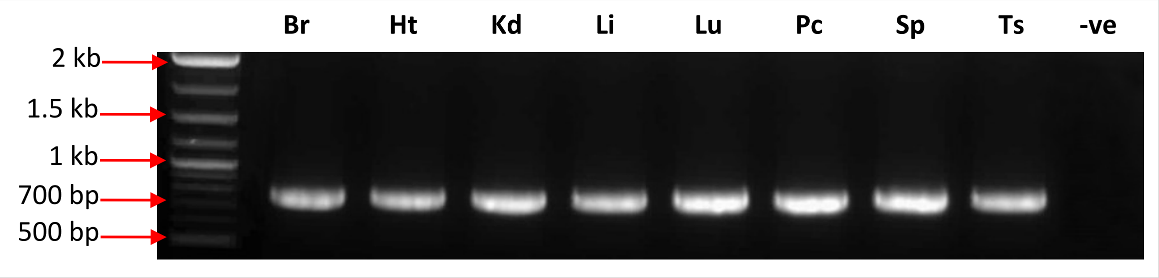


**Figure 2.1. Expression analysis of GAPDH amplified from a rat cDNA tissue panel**

Rat cDNA synthesised from 100ng/µl total RNA were used as template to amplify a 738bp fragment of GAPDH. The PCR reaction was run for 35 cycles (95°C-1 min; 59°C-30 sec; 72°C-1 min) and the products visualised on a 1.5% agarose gel with 1X SYBRsafe. Tissues are abbreviated as follows: Br, brain; Ht, heart; Kd, kidney; Li, liver; Lu, lung; Pc, pancreas; Sp, spleen and Ts, testis. RNase and DNase free $H_2O$ has been used a replacement for cDNA in the negative control.

### 2.4.4. Quantification of DNA

Both dsDNA and ssDNA were quantified using the Quant-iT™ assay kit (Quant-iT™ dsDNA BR and ssDNA Assay Kit; Invitrogen). 200µl of Quant-iT™ working solution was prepared for each sample by diluting the Quant-iT™ reagent 1:200 volume in Quant-iT™ DNA buffer. 1-10µl of DNA sample (depending on the concentration) was made upto 200µl with the working solution and vortex mixed for 3 sec. The sample was left to incubate at room temperature for 2 min. The machine was first calibrated by using the standards provided with the kit, following which a reading can be taken. The Qubit™ fluorometer provides a reading in µg/ml, which represents the concentration after the sample was diluted in the working solution. To calculate the concentration of the sample the following equation was used: QF value x (200/×). Where the QF value is that given by the Qubit™ flurometer and × is the volume of DNA sample added to the working solution.

## 2.5. Polymerase Chain Reaction (PCR)

### 2.5.1. Primer design and PCR Reaction mix

Primers were designed which would target several MIR-containing genes (table 2.3). The primers were generated so the forward oligonucleotide corresponds to the sense strand of the target sequence and the reverse primer to the reverse complement. All reverse primers either cross exon splice sites of the mRNA or are in downstream exons to the sense primer to eliminate gDNA amplification. The criteria for primer design were ~50% GC content, with a melting temperature between 57-64°C with a minimum sequence length of 19bp.

PCR reaction mixes were prepared in a 0.2ml thin walled PCR tube (Dutscher Scientific) and the following constituents were added:

| | |
|---|---|
| 5x GoTaq Flexi buffer (Promega) | 10µl |
| $MgCl_2$ (25mM; Promega) | 5µl |
| deoxyNTP mixture (10mM; Promega) | 1µl |
| Forward primer x (20mM; Sigma) | 0.5µl |
| Reverse primer y (20mM; Sigma) | 0.5µl |
| GoTaq DNA polymerase (5U/µl) | 0.5µl |
| cDNA template (20ng) | 5µl |
| RNase and DNase free $H_2O$ to final volume of | 50µl |

Negative and positive controls were also prepared. The cDNA template was substituted with RNase and DNase free $H_2O$ in the negative control and for the positive control the primers were replaced with those of housekeeping genes (table 2.4). Following preparation of the PCR master mix and adding of the DNA polymerase the samples were placed in a PCR thermal cycler (Labnet MultiGene II). All samples were denatured once for 2 min following which three steps (denaturation, annealing and extension) were repeated 40 times. Denaturing steps were at 96°C, annealing was for 30 sec at varying temperatures and the extension stage was performed at 72°C for 1 min/kb of amplicon. The annealing temperatures for each primer combination have been included in table 2.3. Following the repetition of the 3 steps the samples were further heated to 72°C for 5 min to ensure full extension of the PCR products.

| Sp. | Primer name | Primer sequence (5'-3') | Transcript details | % GC content | Melting temp (°C) | Annealing temp (°C) | Product (bp) |
|---|---|---|---|---|---|---|---|
| Hs | AHI1 F1 | TTGGAACCCAGAAACAGGAG | Paired with both Abelson helper integration site 1 (AHI1) reverse primers | 50 | 64 | 60 | - |
| Hs | AHI1 R1 | GCTGCCATACCACCAGTCTT | Full length transcript of AHI1 | 55 | 64 | 60 | 594 |
| Hs | AHI1 RT | CAGGTCGGCTCAGTTCTTCT | Read-through transcript of AHI1 | 50 | 64 | 60 | 670 |
| Hs | CIITA F1 | AGCTGTGTCACCCGTTTCAG | Paired with the class II, major histocompatibility complex, transactivator (CIITA) reverse primers | 55 | 67 | 60 | - |
| Hs | CIITA Sh | TTTGAGCCCAGGTCAGTCTC | Read-through transcript of CIITA | 55 | 65 | 60 | 303 |
| Hs | CIITA Lg | TTTCCCAGGTCTTCCACATC | Full length transcript of the CIITA | 50 | 64 | 60 | 161 |
| Hs | NRL F1 | ACCCACCTTCAGTGAACCG | Paired with Neural retina leucine zipper (NRL) reverse primers | 55 | 64 | 62 | - |
| Hs | NRL Sh | TTTTTTGCTAATTACAGCTTTAA | Alternative truncated 3-UTR | 22 | 57 | 62 | 966 |
| Hs | NRL Lg | AGGACCCAGGTTTCCAGACT | Reference sequence full length NRL | 55 | 64 | 62 | 1198 |
| Hs | GSG1L F1 | AATGTCGCAGCTTCATTGAC | Paired with the GSG1-like (GSG1L) reverse primers | 45 | 63 | 59 | - |
| Hs | GSG1L RX | CCTCTCTGTGCCTCGATTTG | Specific to the MIR-derived exon of GSG1L | 55 | 65 | 59 | 489 |
| Hs | GSG1L R1 | CCCTCTTCTCCATCCTCTCC | Reference sequence full length GSG1L | 60 | 64 | 59 | 499 |

**Table 2.3. Primer sequences and thermal cycling conditions for human and rat RT-PCR experiments.**

All primers used in this study have been listed, including the melting temperature calculated from the primer sequence and the annealing temperature used during thermal cycling. Where the primers are species specific the species is indicated as follows: Hs, *homo sapiens*; Rn, *rattus norvegicus.*

| Sp. | Primer name | Primer sequence (5’-3’) | Transcript details | % GC content | Melting temp ($^{o}$C) | Annealing temp ($^{o}$C) | Product (bp) |
|---|---|---|---|---|---|---|---|
| Rn | 5TG | ACTTTGACGTGTTTGCCCAC | Paired with 3TG-S and 3TG-L | 50 | 57 | 59 | - |
| Rn | 3TG-S | GCTGAGTCTGGGTGAAGACACAG | Tissue transglutaminase splice variant TGM2_dLg (Tolentino *et al.,* 2002) | 56 | 63 | 59 | 414 |
| Rn | 3TG-L | CAATATCAGTCGGGAACAGGTC | Tissue transglutaminase splice variant TGM2_wLg (Tolentino *et al.,* 2002) | 50 | 59 | 59 | 513 |
| Rn | TG2 F Vs | GGAACTTTGGGCAGTTTGAG | Paired with TG2 Rn R Vs | 50 | 57 | 59 | - |
| Rn | TG2 R Vs | TTCAGGGTATGGAACTCATGG | Tissue transglutaminase splice variant TGM2_Vs | 48 | 58 | 59 | 488 |
| Rn | TG2 F Sh | CAGGAGAAGAGCGAAGGAAC | Paired with TG2 Rn R Sh | 55 | 59 | 59 | - |
| Rn | TG2 R Sh | CAGGCAGAGCCTTACCAGAG | Tissue transglutaminase splice variant TGM2_Vs | 60 | 61 | 59 | 544 |
| Hs | TG2_1 | CCTAGACATCTGCCTGATCC | Paired with TG2_Vs (Hs) and within exon 5 | 55 | 62 | 59 | - |
| Hs | TG2_Vs | TCATGACCCACATCCCAGC | Tissue transglutaminase splice variant TGM2_Vs (Hs) | 58 | 60 | 59 | 323 |
| Hs | TG2_2 | CTGAGCACCAAGTACGATGC | Paired with TG2_Sh (Hs) and within exon 9 | 55 | 62 | 59 | - |
| Hs | TG2_Sh | AACACAGGGCTTTACCAGAG | Tissue transglutaminase splice variant TGM2_Sh | 50 | 60 | 59 | 481 |
| Hs | TG2 Hs F Tc | CACATTCCTTCCCCTCTCTG | 5’-truncated TGM2 splice variant, paired with TGM2_R Tc | 55 | 64 | 59 | - |
| Hs | TG2 Hs R Tc | CACAGAGCATTCCTCACAGC | Tissue transglutaminase splice variant TGM2_Tc | 55 | 62 | 59 | 482 |

**Table 2.3. Primer sequences and thermal cycling conditions for human and rat RT-PCR experiments continued.**

Primers used to amplify TGM2_wLg and TGM2_dLg were designed by Tolentine *et al.,* (2002).

| Species | Gene | Primers (5'-3') | Product (bp) |
|---|---|---|---|
| *Homo sapiens* | G3PDH | TGAAGGTCGGAGTCAACGGATTTGGT<br>CATGTGGGCCATGAGGTCCACCAC | 983 |
| *Rattus norvegicus* | Gapdh | GGCTGCCTTCTCTTGTGAC<br>GGCCGCCTGCTTCACCAC | 738 |

**Table 2.4. Primer sequences for the house keeping gene used as a positive control.**

The house keeping gene glyceraldehyde-3-phosphate dehydrogenase (GAPDH) has been used as a positive control for RT-PCR analyses.

### 2.5.2. Gel electrophoresis

All PCR amplification products were mixed with a 6x loading dye (0.25% (w/v) bromophenol blue, 40% (w/v) sucrose, 4ml (v/v) TE buffer, 10.8ml 0.1% (v/v) DEPC-treated $H_2O$). Samples were separated in a 2% agarose gel, prepared with agarose powder (Bioline) dissolved in 100ml of 1x TAE buffer (40mM Tris-base, 20mM glacial acetic acid, 5mM EDTA; pH 8.0), at 100V for 1hr. Once the agarose solution reached 55-60$^{o}$C 10µl of the DNA gel stain SYBR Safe (Invitrogen) was added. A molecular marker was included to monitor the migration distance (HyperLadder™ IV, Bioline; 1kb ladder, Promega) and was visualised using a transilluminator emitting UV light (Syngene).

### 2.5.3. Sample purification and sequencing

The PCR products were purified directly when a single amplicon is present, the gel band was excised if there were large primer dimers, smearing or if a sample was to be used in cloning techniques. All DNA fragments were purified using the Wizard® SV Gel and PCR Clean-Up System (Promega) following the manufacturer's instructions. The kit functions by binding the DNA to silica membranes of a spin column and impurities removed by a series of washes before eluting the DNA. Purified DNA fragments were then confirmed as the sequence of interest: 10ng of purified product was mixed with 3.2pmol of reverse primer and made up to 10µl using RNase and DNase free water. Reaction mixes were sent to the Functional Genomics and Proteomics unit at the University of Birmingham for sequencing using a 3730 DNA analyser.

## 2.6. Molecular cloning using the pGEM®-T Easy vector

PCR products with low yields and poor sequence data were cloned using the pGEM-T Easy cloning vector (Promega). The vector contains two RNA polymerase promoters, SP6 and T7 flanking a multiple cloning region within the coding region of β-galactosidase (*lacZ)*. Cloning within this region allows for colour screening of recombinant DNA.

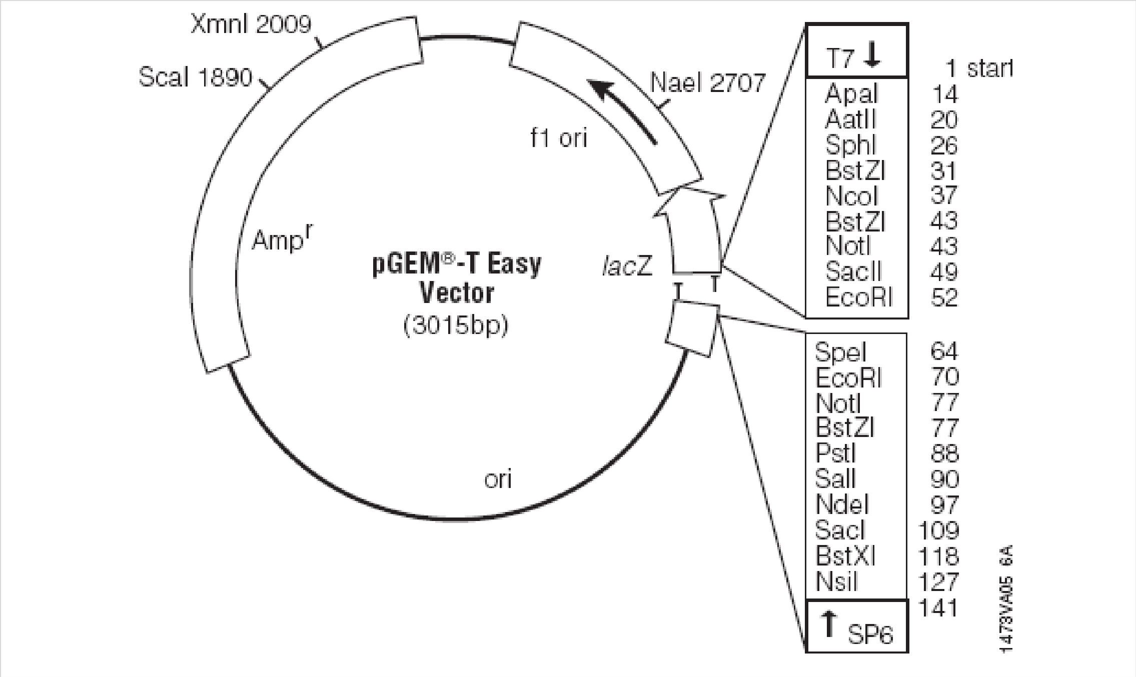


**Figure 2.2. Map of the pGEM-T Easy cloning vector with sequence reference points**

The pGEM®-T Vector is derived from the pGEM®-5Zf (+) Vector (GenBank accession number X65308). The vector contains two RNA polymerase promoters, SP6 and T7 flanking the coding region of β-galactosidase (*lacZ)*. The Amp$^r$ gene allows for the identification of ampicillin resistant *E.coli* colonies. The positions of the Ampicillin resistance region (restriction sites (1-141bp) and the phage f1 region (f1 ori) have been indicated.

### 2.6.1. Preparation of *E. coli* XL1-blue cells

A colony of *E. coli* XL1-blue cells (genotype: *rec*A1, *end*A1, *gyr*A96, *thi*-1, *hsdR*17 (rk-, mk+), *supE*44, *rel*A1, *lac* [F' *proAB*, *lacI*$^q$ZΔM15::Tn10(tet$^r$)]) were selected from an LB agar plate (1% (w/v) tryptone, 0.5% (w/v) yeast extract, 1% (w/v) NaCl, 1.5% (w/v) agar (Becton Dickinson Company)) containing tetracycline (5µg/ml) and placed in 5ml of LB medium (1% (w/v) tryptone, 0.5% (w/v) yeast extract, 1% (w/v) NaCl) containing 0.1% tetracycline (5µg/ml). The cell suspension was incubated at 37°C in an orbital shaker (Gallenkamp) at 200rpm overnight.

4ml of the overnight culture was added to 100ml of LB medium tetracycline (5µg/ml). A 1.5ml aliquot of the inoculated medium was added to a 1ml plastic cuvette (Sarstedt) to calibrate the spectrophotometer. The remainder was incubated at 37°C on a shaking platform (200 rpm; Luckham R100) for 3hr. The absorbance was then measured at $_{A}660$nm every 30 min thereafter until an optical density of 0.5 was obtained (approximately 4.5hr). The inoculated LB medium was split into two 50ml tubes (Sarstedt) and centrifuged at 3000 *x g* at 4°C for 15 min. The supernatant was discarded and each 50ml tube of pelleted cells were resuspended in 20ml of filter sterilised (0.22nm) ice-cold TFBI solution (30mM KAc, 50mM $MgCl_2$, 100mM KCl, 10mM $CaCl_2$, 15% (v/v) glycerol; pH 5.8) and briefly vortex mixed and incubated on ice for 30 min. The cells were then centrifuged at 3000 *x g* at 4°C for 20 min. The supernatant was discarded and the cells were resuspended in 2ml of filter sterilised ice-cold TFBII solution (10mM Na-MOPS, 75mM $CaCl_2$, 10mM KCl, 15% (v/v) glycerol; pH 6.5) per tube and the cells were gently resuspended. 100µl aliquots of the competent cells were prepared and stored at $^{-}80$°C until required.

### 2.6.2. Ligation and transformation using the pGEM T-Easy vector

The purified DNA fragments (insert) were ligated into the pGEM-T® Easy cloning vector. The volume of insert (ng) required for each reaction was calculated with a insert/vector molar ratio of 3:1.

$$\frac{\text{vector (ng) x insert (kb)}}{\text{vector (kb)}} \text{ x insert:vector molar ratio}$$

The following reagents were added to 0.5ml Eppendorf tubes.

| | |
|---|---|
| 2x rapid ligation buffer (Promega) | 5µl |
| pGEM T-Easy vector (50ng/µl; Promega) | 1µl |
| Purified insert (50ng/µl) | 3µl |
| T4 DNA ligase (1U/µl; Roche) | 1µl |

The sample was incubated at room temperature for 1hr. The ligation product of the insert and vector were then transformed into *E. coli* XL1-blue cells. The ligation mixture was added to a 2ml Eppendorf tube with 100µl of competent XL1-blue cells. The recombinant plasmid cell suspension was then transformed into incubated *E.coli* XL1-blue cells via 'heat shock' by placing on ice for 45 min then at 42°C for 1 min. The cells were then incubated on ice for 2 min and 150µl of SOC medium (1% (w/v) tryptone, 0.5% (w/v) yeast extract, 10mM NaCl, 2.5mM KCl, 10mM $MgCl_2$, 20mM $MgSO_4$, 20mM glucose; pH 7.0), was added and incubated at 37°C in an orbital shaker at 200rpm for 1 hr.

The presence of the β-galactosidase encoding gene (*lacZ*) in the cloning vector allows for the positive identification of colonies containing the insert, therefore 100µl of the *lacZ* inducer IPTG (isopropyl-β-D-thiogalactopyranoside; 200mg/ml; Bioline) and 60µl of the artificial substrate X-GAL (5-bromo-4-chloro-3-indolyl-β-D-galactoside; 20mg/ml; Bioline) were spread onto LB agar plates (containing 0.1% (v/v) ampicillin 100) using glass beads. 20µl of the transformed cells were then spread in the same manner onto the agar plate and incubated in a moist chamber at 37°C overnight. Single white colonies were picked and subcultured in liquid LB medium overnight (see 2.6.1).

### 2.6.3. Isolation of plasmid DNA

Plasmid DNA Mini-preparations were prepared, according to the manufacturer's instructions of (Wizard ® *plus* SV minipreps DNA purification system; Promega). The kit comprised a series of solutions to lyse bacterial membranes, neutralise the reaction mixture and remove impurities including by using spin columns. The columns bind the plasmid DNA to a matrix until it was eluted and resuspended in RNase and DNase free water. A final ethanol precipitation was conducted to concentrate the purified plasmid DNA (section 2.3.1). Following plasmid isolation the purified plasmid was loaded onto a 2% agarose gel to confirm the presence of the insert; following confirmation, the purified plasmid DNA was used as a template for PCR, using the original primers and cycling conditions of the insert. The PCR product was then purified (section 2.5.3) and sent for sequencing.

## 2.7. Radioactive *in situ* hybridisation

Radioactive *in situ* hybridisation was used to localise the transcript variants of tissue transglutaminase in the adult rat (Wistar) brain.

### 2.7.1. Probe selection and design

The read-through transcripts for rat tissue transglutaminase (TGM2) were predicted by generating composite transcripts from the gDNA sequence (Accession NC_005102.2, base 148832866 to 148862385). A 45bp antisense oligonucleotide was designed specific to each isoform with a GC content of 50-60%; necessary to ensure sufficient probe binding with the minimal background interference.

A 45mer oligonucleotide was designed to recognise the shortest transcript variant (TGM2_Vs) which is generated following the skipping the donor splice site of exon 6, and reading through into the consecutive intron which contains an in-frame stop codon.

5’-CACAAGTTGGTGGTGAACTTCCAGTGTGACAAGCTGAAGTCGGTC-3’

A sequence was selected which is specific to rar TGM2_Sh mRNA, an isoform also created by reading-through the donor splice site of common exon 10, and similarly a stop codon is recruited.

5’-GGATGAGACCATTCCGAGACGGACACGCACGGACCTTGTAGGTCC-3’

The isoform which has been called TGM2_dLg is a splice variant generated via an alternative donor and acceptor spice site in constitutive exons 12 and 13. As a consequence 210bp of mRNA is excluded from the full length transcript. The atypical splicing generates an alternative reading frame and subsequently an alternative stop codon. The isoform specific oligonucleotide sequence for this transcript is as follows:

5’-GACCGACTTCAGCTTGTCACACTGGAAGTTCACCACCAACTTGTG-3’

An oligonucleotide was designed which was specific to the full length reference sequence (RefSeq) for tissue transglutaminase (TGM2_wLg).

5’-CTGGCAGGTGATGGGCTGAGTCTGGGTGAAGACACAGTCATACAG-3’

### 2.7.2. Silanisation of slides and cryostat sectioning

Microscope slides (VWR) were silanised to aid the fixing of brain sections. Pre-cleaned slides were baked in an oven (Sakura) at 180$^o$C for 2 hr following which they were transferred to glass racks and subject to a selection of baths: acetone for 1-2 min; acetone and silane (3-aminopropyltriethoxysilane) mix of 49:1 for 5 min; acetone for 1 min twice and in 0.1% (v/v) DEPC-treated water for 1 min three times. The slides were air dried for 1 hr and baked at 180$^o$C for a further 2 hr.

Frozen brains were mounted, over dry ice, onto a mounting plate using Tissue Tek (Sakura) and left to equilibrate in the cryostat for 1 hr. 10μm coronal brain sections were cut using a cryostat (CM1900, Leica) at –16$^o$C and mounted onto the slides. Samples were stored at –80$^o$C until required to preserve the RNA.

### 2.7.3 Tissue fixation

The tissue sections were fixed to the microscope slides via a sequence of baths. The slides were bathed in chilled 2% (w/v) paraformaldehyde (5g PFA in 200ml 0.1% (v/v) DEPC-treated water and 50ml 5X PBS; pH 7.4) for 10 min. The sections were then bathed twice in 1x PBS. Following which the tissue sections were dehydrated by bathing once in 70% (v/v) ethanol, once in absolute ethanol and air dried for 30 min.

### 2.7.4. Labeling and purification of radiolabeled probe

The oligonucleotide probes were labeled with the radioisotope sulphur-35 [α-$^{35}$S] by adding the following constituents to a sterile 0.5ml Eppendorf tube:

| | |
|---|---:|
| RNAse and DNAse-free water | 16.25μl |
| 5x terminal deoxynucleotidyl transferase (TdT) reaction buffer | 5μl |
| Antisense oligonucleotide (5ng/μl) | 1μl |
| [α-$^{35}$S] dATP (12.5mCi/ml; PerkinElmer Life Sciences Inc.) | 1.25μl |
| rTdT (20U/ml; Promega) | 1.5μl |

The un-purified probe was incubated at 37°C for 75 min in a designated water bath. Residual [$\alpha$-$^{35}$S] dATP was removed using Sephadex columns (Amersham Pharmacia). The column was centrifuged at 735 *x g* for 4 min to remove the matrix which was discarded. 1µl of the incubated labelling reaction was added to 5ml of scintillation fluid (Ultima Gold; Packard Bioscience) to determine the activity of the [$\alpha$-$^{35}$S] dATP labelled to the oligonucleotide. The remaining labeling reaction was placed in the Sephadex column, and centrifuged at 735 *x g* for 4 min. The columns contain tiny beads which bind to any unattached nucleotides and only labelled probe is eluted. 1µl of elute was added to 5ml of scintillation fluid to compare the activity to that prior to purification. The radioactivity of the samples was measured using a Liquid Scintillation Analyser (Tri-Carb 2250CA; Packard).

### 2.7.5. Hybridisation

Each slide required 200µl of hybridisation buffer which was prepared by adding the following to a 50ml sterile tube: 4x standard saline citrate (SSC; 0.3M tri-sodium citrate, 3M NaCl (pH 7)), 5x Denhardt's solution (1% (w/v) polyvinylpyrrolidone, 1% (w/v) bovine serum albumin (BSA; fraction V), 1% (w/v) ficoll), 25mM sodium phosphate (0.5M $Na_2HPO_4$ (pH 9) and 0.5M $NaH_2PO_4$ (pH 4); pH 7), 1mM sodium pyrophosphate (0.6M $Na_4P_2O_7$; pH 10.4) and 10% (w/v) dextran sulphate. The buffer was vortex mixed for 10 min and the remainder added: 10mM DTT, 120µg/ml heparin (Grade I-A; from porcine intestinal mucosa), 100µg/ml polyadenylic acid, 200µg/ml denatured DNA from salmon testes, 50% (v/v) formamide. The mixture was adjusted to 50ml using 0.1% (v/v) DEPC-treated water. For each slide 100,000cpm of the purified labelled probe was mixed in 200µl of hybridisation buffer. The hybridisation buffer was evenly distributed across the sections and parafilm (Pechiney Plastic Packaging) was positioned over the sections, which were then placed in a home-made humidified case. The cases were incubated in a hybridisation oven (Bibby Stewart Scientific) at 42°C overnight. Negative control hybridisations were prepared by adding a 200-fold excess of the same unlabelled oligonucleotide to the 100,000cpm radiolabeled probe. The 200x excess is sufficient to compete with the labelled probe if specific hybridisation occurs. Unspecific binding of the oligonucleotide has occurred if a signal is present in the negative control

### 2.7.6. Washing

Sections were washed to remove the hybridisation buffer and excess radiolabeled oligonucleotide. The parafilm was removed from the sections by bathing the slides in 1x SSC for 30 sec at room temperature. Following which the slides were then washed twice in 1x SSC for 30 min at 55$^o$C using a water bath shaker (Clifton) at 200rpm. The slides were then transferred to 1x SSC at room temperature for 30 sec and bathed in 0.1x SSC for 30sec. The brain sections were dehydrated by bathing once in 70% (v/v) ethanol, once in absolute ethanol and air dried for 30 min

### 2.7.7. Film exposure and visualisation

The radiolabeled slides were exposed to X-ray film (Kodak Biomax MR, PerkinElmer Life Sciences Inc) for 4-6wks and developed using an automatic developing machine (Compact X4, Xograph). The film was visualised by being placed on a light box (UVP Inc) and images were captured using a CCD camera system and associated software (Euresys Multicam). Rat brain regions were mapped using the interactive databases BrainMaps (http://brainmaps.org/index.php) and the Allen Brain Atlas (http://www.brain-map.org/).

## 2.8. Cell culture

### 2.8.1. Culturing and maintenance of BRIN-BD11 cells

BRIN-BD11 cells were kindly donated by Dr Verderio-Edwards, Nottingham Trent University. BRIN-BD11 cells are a clonal insulin-secreting rat cell line, which were cultured for the study of tissue transglutaminase expression by reverse transcriptase-PCR (RT-PCR). Cell culture was performed using aseptic techniques within a class II medical safety cabinet (Walker). The cells are routinely stored in the gaseous phase of liquid nitrogen, suspended in freezing medium containing 95% (v/v) foetal bovine serum (FBS; Cambrex) and 5% (v/v) sterile DMSO (Sigma Aldrich). A vial of BRIN-BD11 cells (Passage 28; 4 x 10$^6$) were recovered from storage in liquid nitrogen and rapidly thawed in a 37°C water bath. The cell suspension was transferred to a sterile

75cm$^2$ (T-75) tissue culture flask (Sarstedt) containing 15ml of fresh RPMI-1640 growth medium, supplemented with 10% (v/v) heat-inactivated foetal bovine serum (FBS; Cambrex), 2mM L-glutamine (Cambrex), penicillin (100U/ml) and streptomycin (100µg/ml; Cambrex).

Cells were grown in an incubator (Sanyo) at 37$^o$C with a humidified atmosphere of 95% (v/v) air/5% (v/v) carbon dioxide and cells were assessed for viability via an inverted light microscope (Nikon). The cells were incubated for 24h to allow for attachement, following which the cell monolayer was rinsed with 5ml of complete growth medium to remove residual freezing medium. The cells were further incubated in 15ml of complete growth medium to obtain ~80% confluency.

### 2.8.2. Sub-culture and detachment

When a sufficient number of cells were present (~80% confluent) the culture medium was removed from the flask and the cell monolayer rinsed with serum-free medium to remove traces of serum. The cell monolayer was detached using trypsin-EDTA (trypsin 0.5mg/ml; EDTA 5mM; 1ml/25cm$^2$). After 5 min, or when sufficient detachment had occurred, the trypsin was deactivated by adding complete serum equal to 2x the volume of trypsin. The cell suspension was transferred to 15ml tubes and centrifuged at 300 *x g* for 4 min (Harriot 18 (MSE)). The supernatant was carefully discarded and the pellet resuspended in 4ml of complete medium. The cells were added to a T-175 culture flask with 26ml of complete medium and incubated at 37$^o$C with a humidified atmosphere of 5% carbon dioxide until ~80% confluent, following which the cells were counted and RNA extracted (section 2.2.2).

### 2.8.3. Cell counting and viability

The cells were counted using a Neubauer glass haemocytometer. A 10µl sample was transferred to the 0.1mm deep well of the haemocytometer and visualised using an inverted microscope. The cells were counted in four fields of 1mm$^2$ at 100x magnification. The cell density/ml was calculated as follows:

$$\text{Mean cell number x } 10^4 \text{ x dilution factor}$$

The cells were then suspended in PBS and used for RNA extraction (section 2.2.1).

# 3. THE NUMBER, DISTRIBUTION AND CONSERVATON OF MIR ELEMENTS WITHIN THE HUMAN GENOME

## 3.1. Introduction

The number of positively identified MIRs in the human genome is regularly increasing, with the latest estimate in excess of 548,000 copies (Sela *et al.,* 2007). Fewer MIRs have been identified in the mouse genome, (~116,000 copies; Sela *et al.,* 2007). However the number of genes which have recruited MIR elements still remains unclear. Chaley and Korotkov (2001) reported MIR elements present in the CDS of 254 transcripts corresponding to 50 known human genes, and suggested that the remaining MIR elements are likely to be situated within introns and intergenic DNA. A recent investigation, describes a total of 181 MIRs located within both the UTR and CDS of human genes of which 78 have been recruited in the CDS (Sela *et al.,* 2007).

The distribution of MIR elements in the human genome has not been documented, however SINEs, including Alus and MIRs, are known to reside predominantly in GC-rich regions (Medstrand *et al.,* 2002). Younger SINEs, such as the AluY family (<5 myr old) are most frequently located in AT-rich regions (Medstrand *et al.,* 2002), whereas LINE elements are integrated and maintained in AT-rich regions of the genome, regardless of age (Jurka *et al.,* 2004). The global distribution of MIR elements is unknown, whereas other retrotransposons such as the Alus are suggested to be distributed non-randomly throughout the genome (Grover *et al.,* 2004).

To determine the role of MIR elements in mammalian gene function and evolution, it was necessary to first generate a dataset of genes which have exaptated at least one MIR element. Furthermore the distribution and conservation of exaptated MIR elements within the human genome has not previously been investigated; therefore the global distribution and degree of conservation of MIRs has been determined.

## 3.2. Identification of MIR-containing human genes within the human genome

Genes which have recruited MIRs were detected and analysed as follows:

- **Identification**

Human MIR-containing genes were identified through three processes:

1. BLAST searches against non-redundant nucleotide sequence databases using the MIR core-SINE consensus sequence.
2. Genes which are known to have recruited a SINE element were obtained from the Genomic ScrapYard database and screened using RepeatMasker for MIR elements.
3. Genes reported in journal articles to have recruited MIR elements were noted.

- **Verification**

The presence of an MIR element was confirmed with the RepeatMasker tool following which:

- The position of the repeat in the mRNA sequence was recorded.
- The coordinates of the MIR consensus retained, repeat orientation and MIR sub-family were noted.

- **Annotation**

Following the verification of the recruitment of an MIR element in a sequence the corresponding gene was determined:

- The gene symbol and description was noted and alternative transcript variants screened for MIR elements using RepeatMasker.
- The gene region which the MIR element has been recruited was noted by determining the exonic composition; regions recorded as the 5'- or 3'-untranslated region or the CDS.
- The position of the MIRs and the gene regions which they are located was determined using Spidey which aligns mRNA to genomic DNA providing the exonic organisation.
- Genes without a confirmed open reading frame or protein sequence listed in the database were considered as ncRNA.

The total number of human genes identified in this study which have exaptated MIR repeats is 1575, 1359 of which are annotated and validated genes (appendix 9.1). The remaining 216 transcripts are currently annotated in GenBank as either hypothetical proteins, or genes of unknown function (appendix 9.2). Only the named genes were included in subsequent analyses. The MIR elements identified were catagorised according to the MIR sub-family. RepeatMasker annotates the MIR sub-type by comparing the element to the concensus sequences listed in Repbase (table 2.1). The orientation and location of the MIR within a nucleotide sequence was determined with the alignment programme Spidey and subsequent sequence analysis (table 3.1). Several human genes contain more than one MIR element, therefore a total of 1854 MIRs have been listed. It appears that there is no preference to recruit an MIR element in the direct or inverse orientation in the mRNA sense strand. The majority of the repeat elements are located within the 3'-UTR (75%), with 14% in the 5'-UTR and 9% in the CDS of the gene. There are also 35 MIR elements located within transcripts which are recently annotated and have no corresponding peptide sequence, so have been listed as ncRNA (see table 3.1). MIR elements are least frequent in the CDS, however they do provide initiating methionines and stop codons for 86 genes, resulting in transcript variants, frameshifts, truncated proteins and nonsense mediated degradation.

| | 5' UTR | | 3' UTR | | ATG | | TAG | | CDS | | ncRNA | | TOTAL | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | [+] | [-] | [+] | [-] | [+] | [-] | [+] | [-] | [+] | [-] | [+] | [-] | [+] | [-] |
| **MIR** | 48 | 68 | 217 | 156 | 2 | 4 | 12 | 13 | 5 | 24 | 9 | 9 | 293 | 274 |
| **MIRb** | 33 | 50 | 289 | 267 | 1 | 5 | 12 | 10 | 12 | 14 | 3 | 7 | 350 | 353 |
| **MIR3** | 18 | 24 | 170 | 145 | 3 | 4 | 6 | 6 | 5 | 8 | 0 | 3 | 202 | 190 |
| **MIRm** | 5 | 4 | 34 | 25 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 42 | 32 |
| **THER1_MD** | 4 | 4 | 19 | 16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 23 | 21 |
| **MIR_mars** | 0 | 2 | 17 | 19 | 1 | 0 | 1 | 0 | 2 | 0 | 0 | 0 | 21 | 21 |
| **MIRc** | 0 | 2 | 8 | 13 | 1 | 0 | 2 | 1 | 1 | 3 | 0 | 1 | 12 | 20 |
| **TOTAL** | 108 | 154 | 754 | 641 | 8 | 13 | 34 | 31 | 26 | 50 | 13 | 22 | 943 | 911 |
| **TOTAL** | 262 | | 1395 | | 21 | | 65 | | 76 | | 35 | | 1854 | |

**Table 3.1. Distribution of MIR sub-families in human genes**

MIRs have been organised according to the MIR sub-type, the orientation (+, sense; -, antisense) and the gene region to which it has been recruited. 5'-untranslated regions (5'-UTR), 3'-untranslated regions (3'-UTR) and protein coding sequence (CDS). The CDS is further divided accordingly; ATG represents MIRs which provide an initiating methionine and TAG represents MIRs which provide a stop codon.

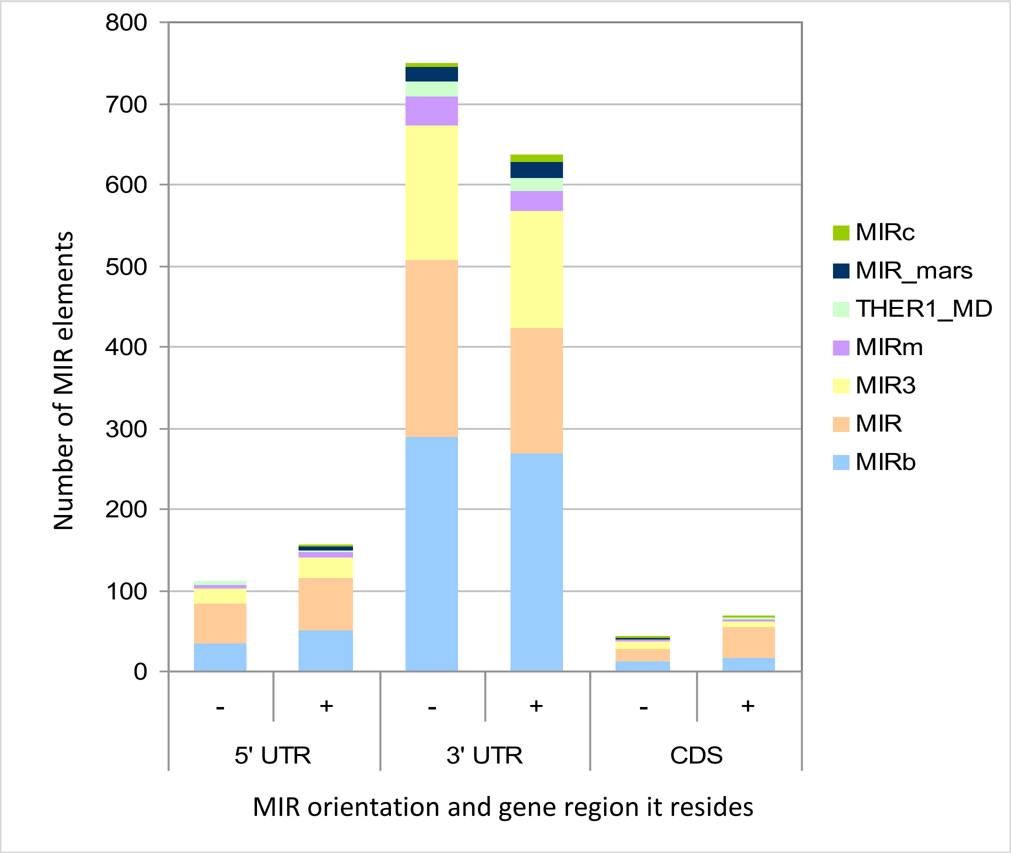


**Figure 3.1. Distribution of MIR elements within the human genome**

Each MIR sub-family are grouped according to the gene region they reside which includes both the 3’- and 5’-UTRs and protein CDS. The MIRs are also organised as either in the sense orientation (+) or antisense (-) orientation.

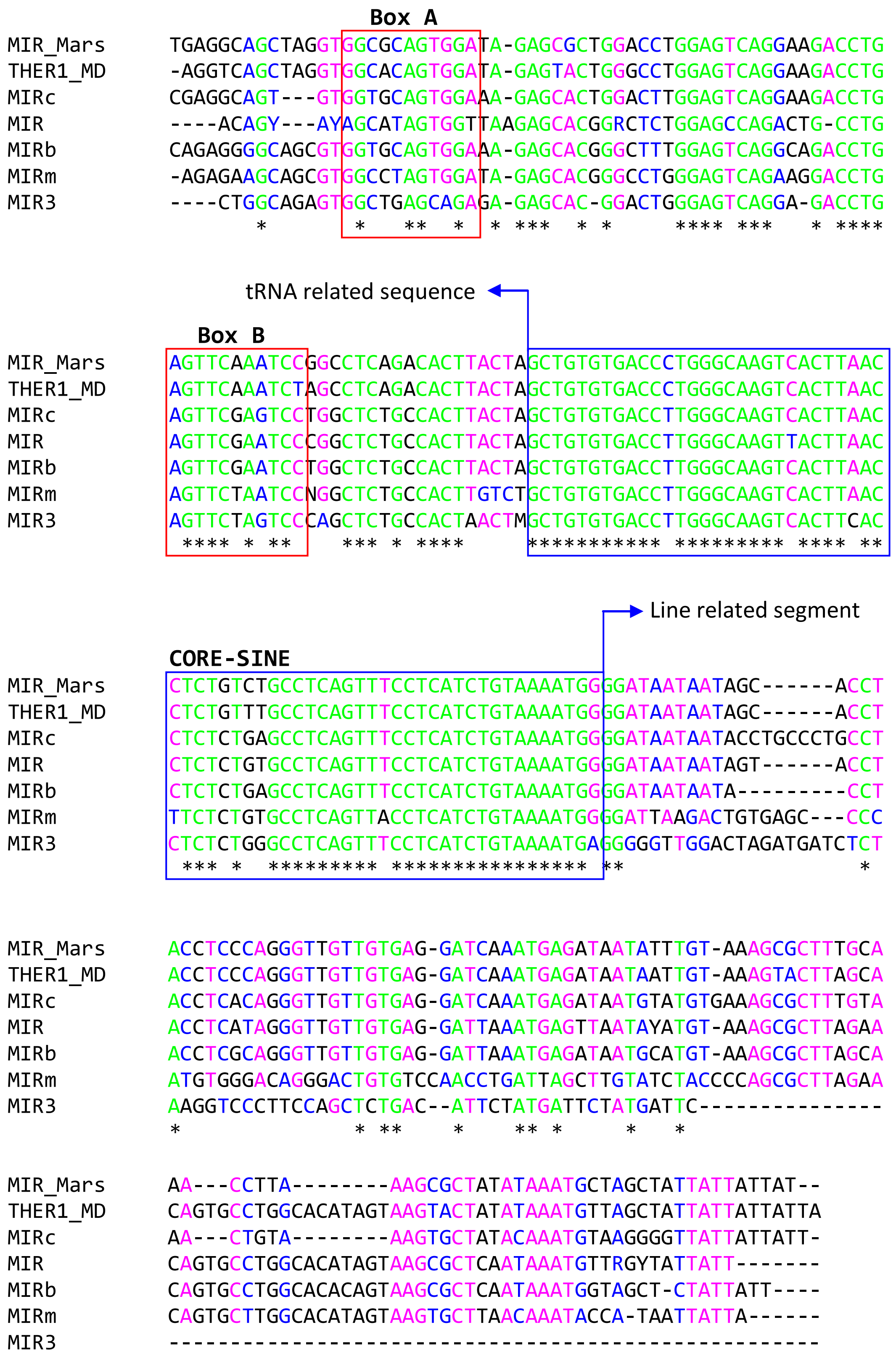
The MIR elements identified were catagorised according to the sub-families listed in RepBase (table 2.1). The predominant MIR elements identified in the human gene dataset are the MIRb and MIR sub-families, with no apparent preference for the repeat sequence to be recruited in either the direct or inverse orientation. Several MIR sub-families exist due to sequence changes of the consensus during evolution, with each sub-family having a different evolutionary age. It is thought that few repeats remain active at any one time, with several master elements being the source of amplification (Deininger *et al.,* 1992). MIR elements were actively transposing for >350 myr, therefore if only a single sub-family is active at any one time, sequence drift and divergence over time may occur, generating the numerous sub-families observed today. As such the most abundant elements identified, the MIRb and MIR sub-families, are the youngest of the repeats with the MIRm and MIR_mars elements representing the more ancient of the MIRs.

## 3.3. Conservation of human MIR elements

MIRs comprise a family of at least seven repeat elements, which vary slightly in their consensus sequence and there is a high degree of sequence conservation between the core-SINE region (65bp) of all MIR-family members (figure 3.2). Core-SINE sequences for a number of species were obtained following performing BLAST searches against genomic DNA databases. After which it can be determined that conservation is maintained equally in intergenic DNA and mRNA sequences, with 94% sequence homology between the core-SINE of human and montremata intergenic DNA and 91% between human and montremata mRNA (figure 3.3).

In order to determine whether there is selection for any part of the MIR sequence, the nucleotide positions of the exaptated MIRs were determined for each transcript. The percentage of MIRs which correspond to each nucleotide of the consensus sequence was calculated (figure 3.4), demonstrating that the core-SINE is the most highly conserved region, coinciding with previous observations made by Chaley and Korotkov (2001). There is a steady decline in both the tRNA and LINE regions, yet you would anticipate a uniform decline of the consensus sequence if there was a steady rate of divergence, unless there was a positive selective pressure maintaining this core sequence. Once again there appears to be no preference to maintain the MIR repeat in a particular orientation; a similar pattern of conservation was also detected when plotting the conservation of the MIRs recruited in the CDS or UTRs independently (appendix 9.3).

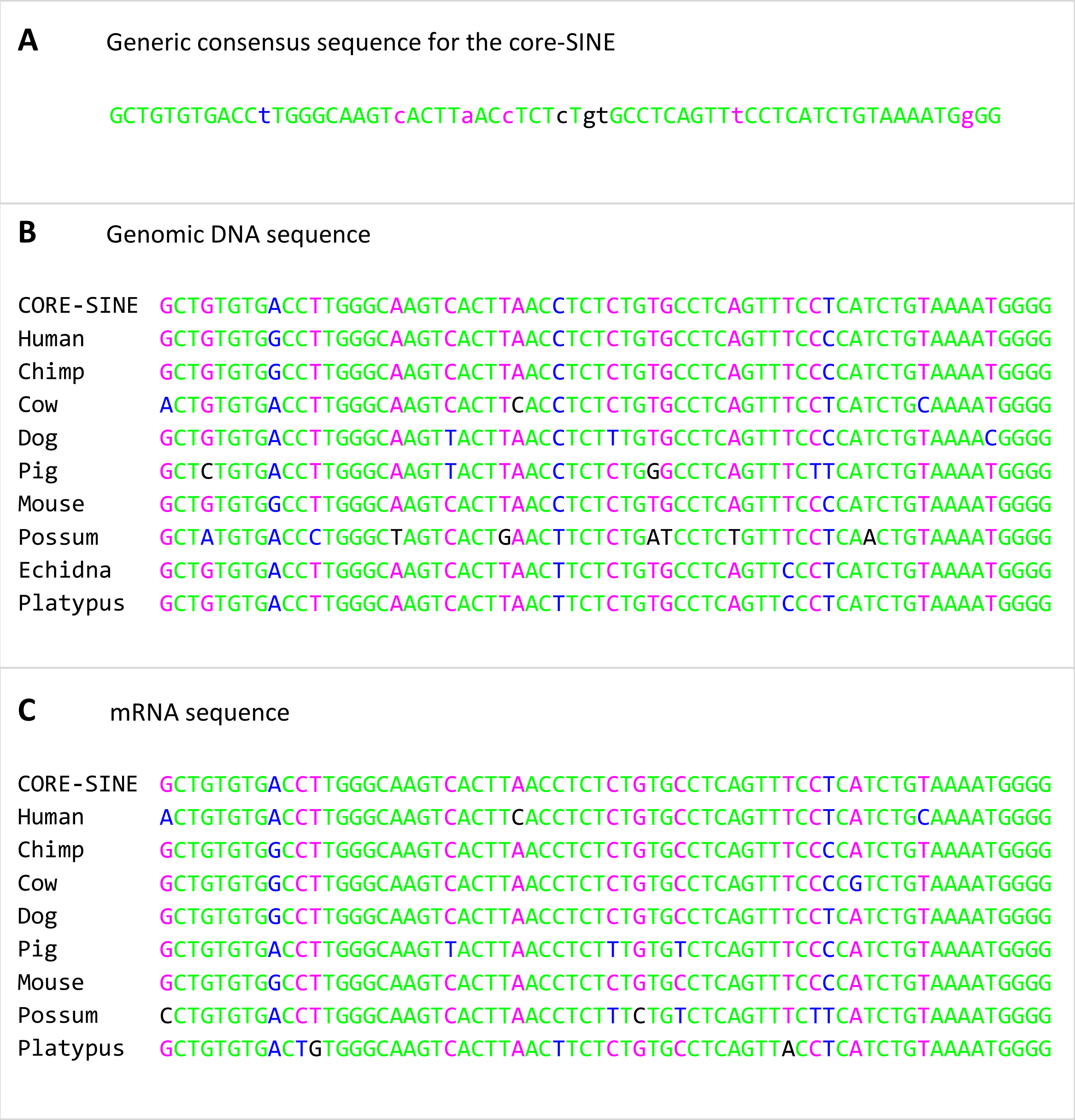
Some MIR elements may be spliced and as a consequence only the exonised portion of the MIR element will be included in the mature mRNA sequence, with the remaining fragment being intronic. However only a small number (<1%) of the MIR-containing genes have recruited MIR elements in the coding sequence and, as such, these will not be reflected in this analysis (figure 3.4). MIR-mediated exonisation is discussed further in chapter 5.



```
                                    Box A
MIR_Mars        TGAGGCAGCTAGGTGGCGCAGTGGATA-GAGCGCTGGACCTGGAGTCAGGAAGACCTG
THER1_MD        -AGGTCAGCTAGGTGGCACAGTGGATA-GAGTACTGGGCCTGGAGTCAGGAAGACCTG
MIRc            CGAGGCAGT---GTGGTGCAGTGGAAA-GAGCACTGGACTTGGAGTCAGGAAGACCTG
MIR             ----ACAGY---AYAGCATAGTGGTTAAGAGCACGGRCTCTGGAGCCAGACTG-CCTG
MIRb            CAGAGGGGCAGCGTGGTGCAGTGGAAA-GAGCACGGGCTTTGGAGTCAGGCAGACCTG
MIRm            -AGAGAAGCAGCGTGGCCTAGTGGATA-GAGCACGGGCCTGGGAGTCAGAAGGACCTG
MIR3            ----CTGGCAGAGTGGCTGAGCAGAGA-GAGCAC-GGACTGGGAGTCAGGA-GACCTG
                       *       *   **  *  * ***  * *     **** ***   * ****

                      tRNA related sequence

                  Box B
MIR_Mars        AGTTCAAATCCGGCCTCAGACACTTACTAGCTGTGTGACCCTGGGCAAGTCACTTAAC
THER1_MD        AGTTCAAATCTAGCCTCAGACACTTACTAGCTGTGTGACCCTGGGCAAGTCACTTAAC
MIRc            AGTTCGAGTCCTGGCTCTGCCACTTACTAGCTGTGTGACCTTGGGCAAGTCACTTAAC
MIR             AGTTCGAATCCCGGCTCTGCCACTTACTAGCTGTGTGACCTTGGGCAAGTTACTTAAC
MIRb            AGTTCGAATCCTGGCTCTGCCACTTACTAGCTGTGTGACCTTGGGCAAGTCACTTAAC
MIRm            AGTTCTAATCCNGGCTCTGCCACTTGTCTGCTGTGTGACCTTGGGCAAGTCACTTAAC
MIR3            AGTTCTAGTCCCAGCTCTGCCACTAACTMGCTGTGTGACCTTGGGCAAGTCACTTCAC
                 **** * **    *** * ****     *********** ********* **** **

                                                        Line related segment

                CORE-SINE
MIR_Mars        CTCTGTCTGCCTCAGTTTCCTCATCTGTAAAATGGGGATAATAATAGC------ACCT
THER1_MD        CTCTGTTTGCCTCAGTTTCCTCATCTGTAAAATGGGGATAATAATAGC------ACCT
MIRc            CTCTCTGAGCCTCAGTTTCCTCATCTGTAAAATGGGGATAATAATACCTGCCCTGCCT
MIR             CTCTCTGTGCCTCAGTTTCCTCATCTGTAAAATGGGGATAATAATAGT------ACCT
MIRb            CTCTCTGAGCCTCAGTTTCCTCATCTGTAAAATGGGGATAATAATA---------CCT
MIRm            TTCTCTGTGCCTCAGTTACCTCATCTGTAAAATGGGGATTAAGACTGTGAGC---CCC
MIR3            CTCTCTGGGCCTCAGTTTCCTCATCTGTAAAATGAGGGGTTGGACTAGATGATCTCT
                 *** *  ********* **************** **                   *

MIR_Mars        ACCTCCCAGGGTTGTTGTGAG-GATCAAATGAGATAATATTTGT-AAAGCGCTTTGCA
THER1_MD        ACCTCCCAGGGTTGTTGTGAG-GATCAAATGAGATAATAATTGT-AAAGTACTTAGCA
MIRc            ACCTCACAGGGTTGTTGTGAG-GATCAAATGAGATAATGTATGTGAAAGCGCTTTGTA
MIR             ACCTCATAGGGTTGTTGTGAG-GATTAAATGAGTTAATAYATGT-AAAGCGCTTAGAA
MIRb            ACCTCGCAGGGTTGTTGTGAG-GATTAAATGAGATAATGCATGT-AAAGCGCTTAGCA
MIRm            ATGTGGGACAGGGACTGTGTCCAACCTGATTAGCTTGTATCTACCCCAGCGCTTAGAA
MIR3            AAGGTCCCTTCCAGCTCTGAC--ATTCTATGATTCTATGATTC---------------
                *              * **    *    ** *     *   *

MIR_Mars        AA---CCTTA--------AAGCGCTATATAAATGCTAGCTATTATTATTAT--
THER1_MD        CAGTGCCTGGCACATAGTAAGTACTATATAAATGTTAGCTATTATTATTATTA
MIRc            AA---CTGTA--------AAGTGCTATACAAATGTAAGGGGTTATTATTATT-
MIR             CAGTGCCTGGCACATAGTAAGCGCTCAATAAATGTTRGYTATTATT-------
MIRb            CAGTGCCTGGCACACAGTAAGCGCTCAATAAATGGTAGCT-CTATTATT----
MIRm            CAGTGCTTGGCACATAGTAAGTGCTTAACAAATACCA-TAATTATTA------
MIR3            -----------------------------------------------------
```

**Figure 3.2. Multiple sequence alignment of the MIR family of repeat elements**

MIR consensus sequences were obtained from RepBase, the primary reference database of prototypic repetitive DNA sequences. The MIR sub-families are names according to the nomenclature provided by Repbase (table 2.1). Completely conserved sequences are in green, identical residues in pink, similar residues in blue and different residues in black. The promoter boxes A and B have been highlighted in red and the core-SINE in blue.



**A** Generic consensus sequence for the core-SINE

```
GCTGTGTGACCtTGGGCAAGTcACTTaACcTCTcTgtGCCTCAGTTtCCTCATCTGTAAAATGgGG
```

**B** Genomic DNA sequence

```
CORE-SINE  GCTGTGTGACCTTGGGCAAGTCACTTAACCTCTCTGTGCCTCAGTTTCCTCATCTGTAAAATGGGG
Human      GCTGTGTGGCCTTGGGCAAGTCACTTAACCTCTCTGTGCCTCAGTTTCCCCATCTGTAAAATGGGG
Chimp      GCTGTGTGGCCTTGGGCAAGTCACTTAACCTCTCTGTGCCTCAGTTTCCCCATCTGTAAAATGGGG
Cow        ACTGTGTGACCTTGGGCAAGTCACTTCACCTCTCTGTGCCTCAGTTTCCTCATCTGCAAAATGGGG
Dog        GCTGTGTGACCTTGGGCAAGTTACTTAACCTCTTTGTGCCTCAGTTTCCCCATCTGTAAAACGGGG
Pig        GCTCTGTGACCTTGGGCAAGTTACTTAACCTCTCTGGGCCTCAGTTTCTTCATCTGTAAAATGGGG
Mouse      GCTGTGTGGCCTTGGGCAAGTCACTTAACCTCTCTGTGCCTCAGTTTCCCCATCTGTAAAATGGGG
Possum     GCTATGTGACCCTGGGCTAGTCACTGAACTTCTCTGATCCTCTGTTTCCTCAACTGTAAAATGGGG
Echidna    GCTGTGTGACCTTGGGCAAGTCACTTAACTTCTCTGTGCCTCAGTTCCCTCATCTGTAAAATGGGG
Platypus   GCTGTGTGACCTTGGGCAAGTCACTTAACTTCTCTGTGCCTCAGTTCCCTCATCTGTAAAATGGGG
```

**C** mRNA sequence

```
CORE-SINE  GCTGTGTGACCTTGGGCAAGTCACTTAACCTCTCTGTGCCTCAGTTTCCTCATCTGTAAAATGGGG
Human      ACTGTGTGACCTTGGGCAAGTCACTTCACCTCTCTGTGCCTCAGTTTCCTCATCTGCAAAATGGGG
Chimp      GCTGTGTGGCCTTGGGCAAGTCACTTAACCTCTCTGTGCCTCAGTTTCCCCATCTGTAAAATGGGG
Cow        GCTGTGTGGCCTTGGGCAAGTCACTTAACCTCTCTGTGCCTCAGTTTCCCCGTCTGTAAAATGGGG
Dog        GCTGTGTGGCCTTGGGCAAGTCACTTAACCTCTCTGTGCCTCAGTTTCCTCATCTGTAAAATGGGG
Pig        GCTGTGTGACCTTGGGCAAGTTACTTAACCTCTTTGTGTCTCAGTTTCCCCATCTGTAAAATGGGG
Mouse      GCTGTGTGGCCTTGGGCAAGTCACTTAACCTCTCTGTGCCTCAGTTTCCCCATCTGTAAAATGGGG
Possum     CCTGTGTGACCTTGGGCAAGTCACTTAACCTCTTTCTGTCTCAGTTTCTTCATCTGTAAAATGGGG
Platypus   GCTGTGTGACTGTGGGCAAGTCACTTAACTTCTCTGTGCCTCAGTTACCTCATCTGTAAAATGGGG
```

**Figure 3.3. Multiple sequence alignment of MIR elements from a number of mammalian species**

**A)** A generic core-SINE consensus sequence was predicted from the multiple sequence alignment of the MIR element sub-families shown in figure 3.2. **B)** MIR elements detected within intergenic DNA sequences of mammalian genomes (RefSeq_genomic); **C)** MIR elements located within cDNA clones and EST sequences (dbESTs, nonRefSeq_RNA). Completely conserved sequences are in green, identical residues in pink, similar residues in blue and different residues in black. Sequence conservation was determined using the ClustalW2 alignment tool (http://www.ebi.ac.uk/Tools/clustalw2/index.html; accession numbers appendix 9.3). All available mammalian sequences have been included in the alignments shown. The generic core-SINE consensus was then used to detect homology for mammalian genomic DNA and mRNA sequences using BLASTn.

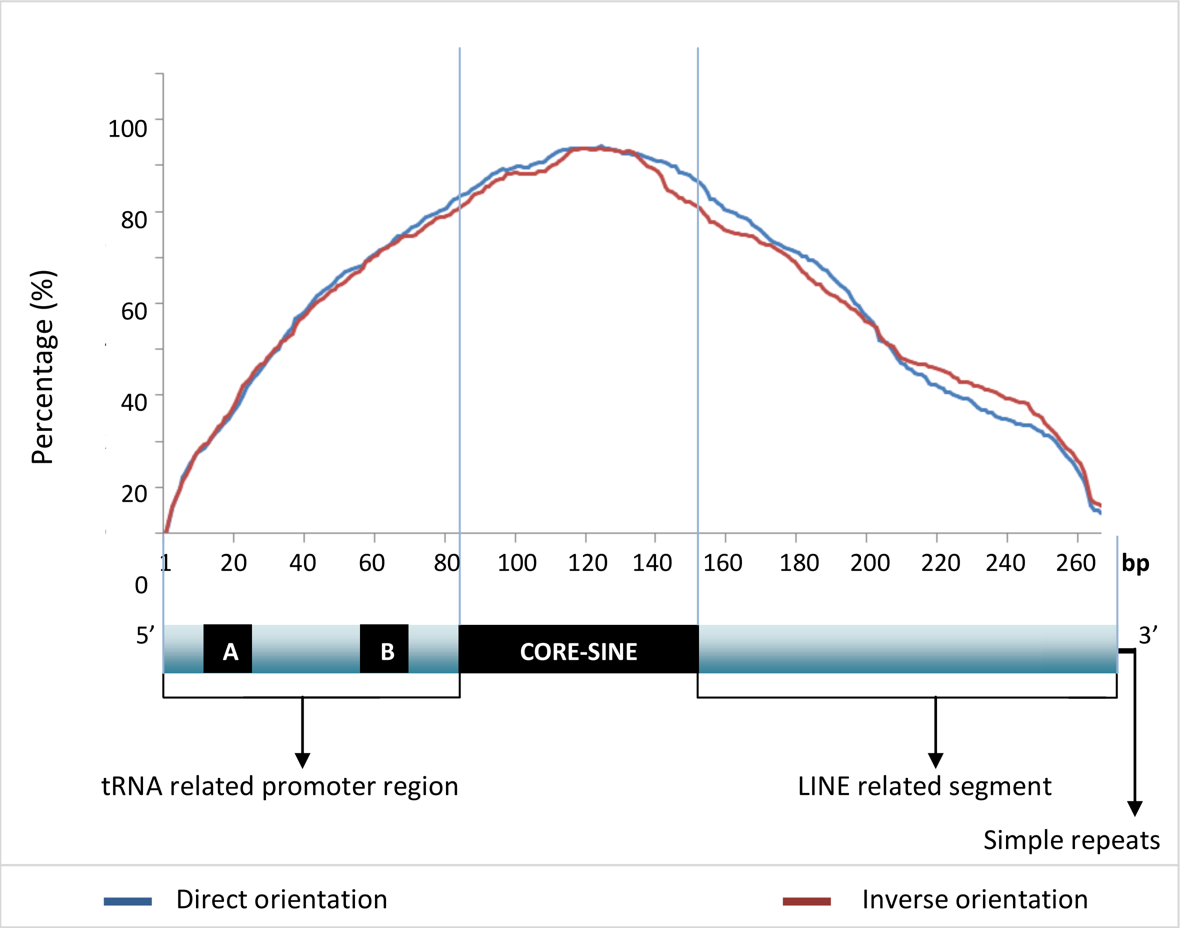


**Figure 3.4. Conservation of all exonic human MIR elements from all gene regions**

The number of MIR elements which correspond to each nucleotide of the generic consensus sequence has been plotted. All of the MIR sub-families located within all gene regions including the 3'-UTR, 5'-UTR and CDS were analysed. A similar pattern of conservation was detected when plotting the MIRs recruited in the CDS or UTRs independently (appendix 9.3).

## 3.4. Distributions of MIR elements

Approximately 5% of the genes in the human genome have exaptated MIRs, although it is not clear whether this is entirely random. Determining the chromosomal distribution of the MIR-containing genes will indicate if there is clustering of the MIR elements, or likewise if the MIRs are exaptated randomly.

### 3.4.1. Chromosomal locations of the MIR-containing genes

The number of genes which have recruited MIRs was determined for each chromosome and the percentage of the total MIR-containing genes per chromosome compared to the total number of genes in the human genome for each corresponding genomic region. Assuming that MIR elements have inserted randomly within the genome, it would be expected that the number of MIR-exaptated genes on a chromosome would be proportional to the number of genes on that chromosome.

| Chromosome | MIR-containing genes (%) | Total genes in the genome (%) | Normalised percentage |
|---|---|---|---|
| 1 | 10.9 | 9.60 | 2.14 |
| 2 | 5.4 | 6.56 | 1.82 |
| 3 | 6.4 | 5.09 | 2.26 |
| 4 | 3.3 | 4.01 | 1.82 |
| 5 | 4.7 | 4.41 | 2.07 |
| 6 | 5.6 | 5.19 | 2.08 |
| 7 | 3.0 | 5.02 | 1.60 |
| 8 | 2.9 | 3.41 | 1.85 |
| 9 | 5.8 | 3.96 | 2.46 |
| 10 | 3.6 | 3.85 | 1.94 |
| 11 | 7.4 | 6.44 | 2.15 |
| 12 | 5.0 | 4.78 | 2.05 |
| 13 | 1.5 | 1.90 | 1.79 |
| 14 | 3.4 | 4.44 | 1.77 |
| 15 | 4.7 | 3.21 | 2.46 |
| 16 | 3.5 | 3.85 | 1.91 |
| 17 | 4.9 | 4.92 | 2.00 |
| 18 | 1.1 | 1.51 | 1.73 |
| 19 | 6.5 | 5.87 | 2.11 |
| 20 | 4.1 | 2.57 | 2.60 |
| 21 | 1.1 | 1.16 | 1.95 |
| 22 | 2.0 | 2.58 | 1.78 |
| X | 3.3 | 4.59 | 1.72 |
| Y | 0.8 | 1.07 | 1.75 |

**Table 3.2. Percentage of MIR-containing genes for each human chromosome**

The percentage of genes which have recruited MIRs for each chromosome compared to the percentage of the total human genes for all chromosomes. The percentage of MIR-containing genes are normalised: (% MIR-genes + % genes in genome)/% genes in genome.

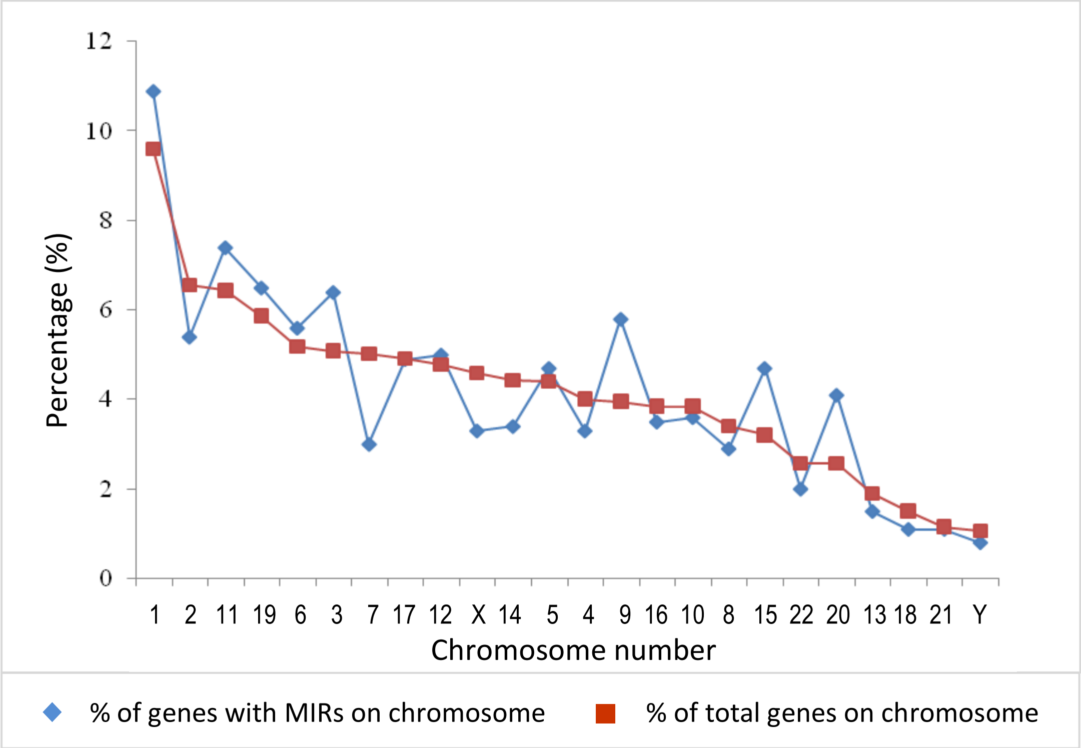
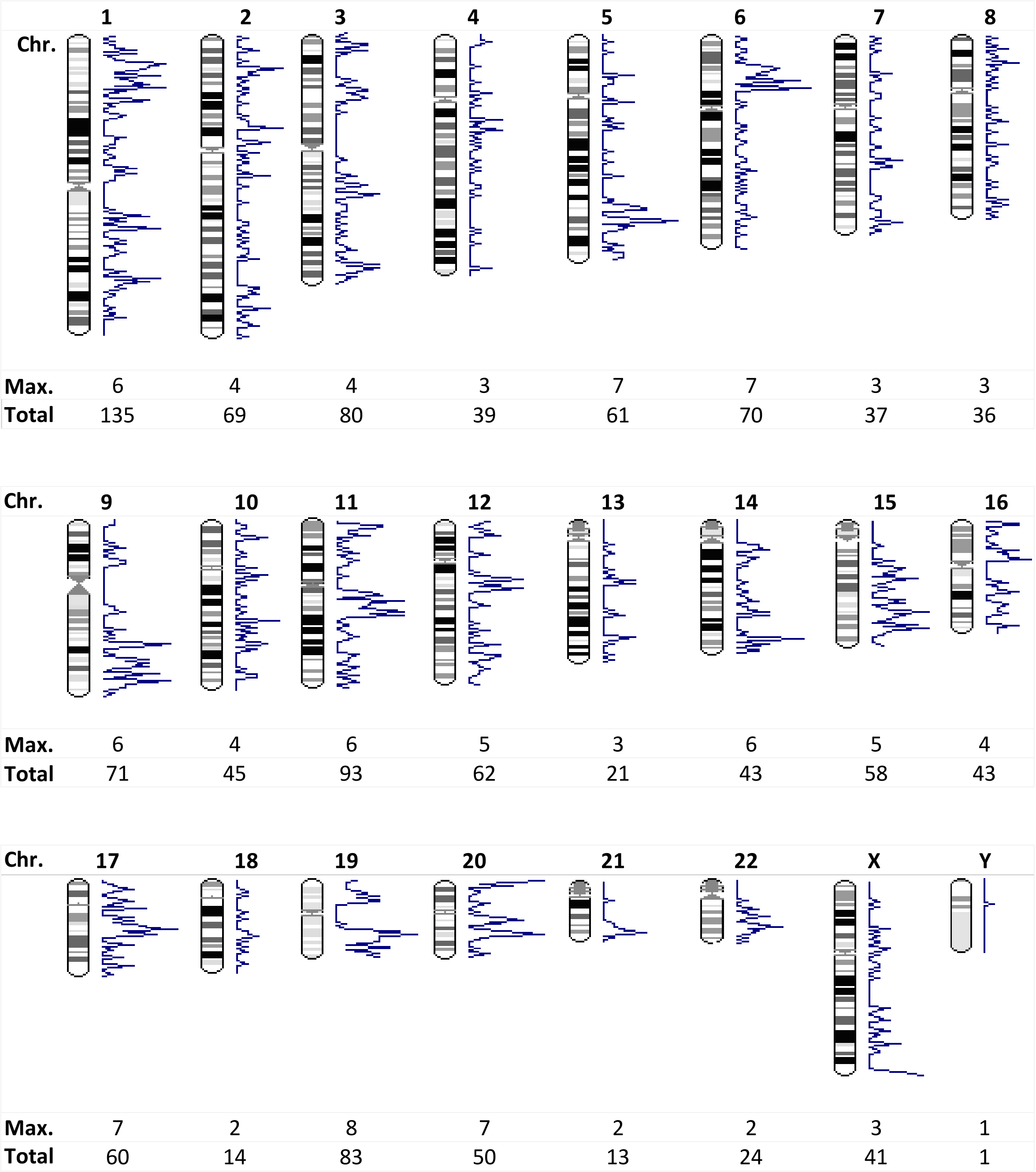


**Figure 3.5. Percentage of MIR-containing genes for each human chromosome**

The percentage of genes which have recruited MIRs for each chromosome is shown in blue. The percentage of the total human genes for all chromosomes has been provided as a reference (red), and the values are plotted according to the chromosome size in base pairs.

Overall it appears that the exaptation of MIRs is a random process (table 3.2, figure 3.5). There is a slight over-representation for the MIR-genes on chromosomes 3, 9, 15 and 20. However, the normalised percentage demonstrates that the MIR-containing genes are equally distributed between all chromosomes (table 3.2). It has previously been reported that Alu repeats are not distributed randomly throughout the human genome. However, these data represent the distribution of all genomic Alu elements not those which have been exaptated by human genes (Grover *et al.,* 2004).

The sub-chromosomal locations of the genes which have recruited MIRs were also investigated to determine if there is any clustering of these genes (figure 3.6). There is apparent clustering of MIR-containing genes on several chromosomes including chromosomes 11, 14, 15, 19 and 20. These chromosomes are known to contain imprinted domains, the MIR elements of which have been further investigated in section 3.4.3.



| Chr. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Max. | 6 | 4 | 4 | 3 | 7 | 7 | 3 | 3 |
| Total | 135 | 69 | 80 | 39 | 61 | 70 | 37 | 36 |

| Chr. | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|
| Max. | 6 | 4 | 6 | 5 | 3 | 6 | 5 | 4 |
| Total | 71 | 45 | 93 | 62 | 21 | 43 | 58 | 43 |

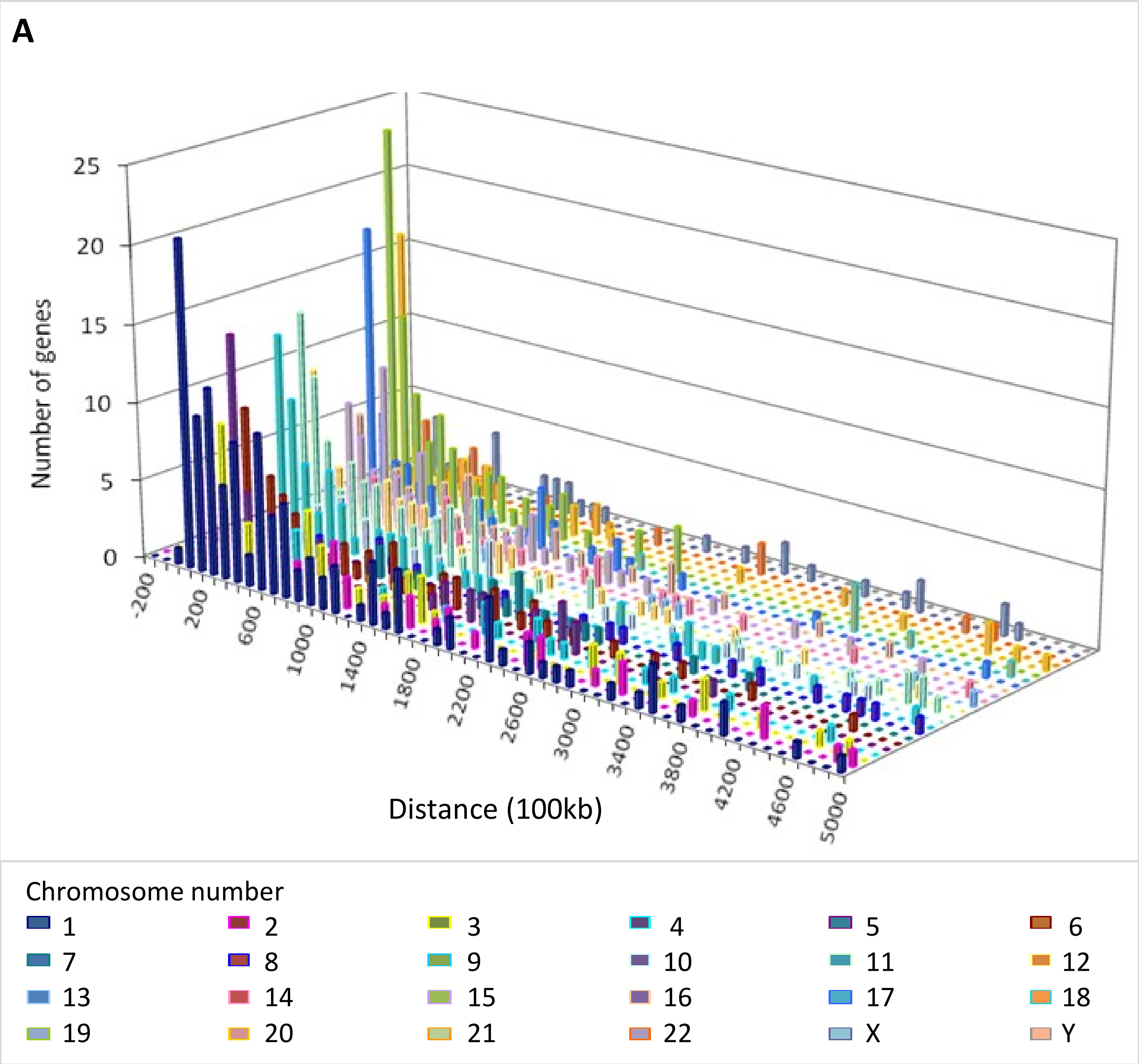
| Chr. | 17 | 18 | 19 | 20 | 21 | 22 | X | Y |
|---|---|---|---|---|---|---|---|---|
| Max. | 7 | 2 | 8 | 7 | 2 | 2 | 3 | 1 |
| Total | 60 | 14 | 83 | 50 | 13 | 24 | 41 | 1 |

**Figure 3.6. A schematic representation of the density of genes which have exaptated MIRs for each chromosomal location**

The blue represents the frequency of MIR elements for each loci, the maximum number of genes in any one plot line has been supplied for reference along with the total number of MIR genes for each chromosome. The chromosome banding (grey) indicates the standard human karyotype, stained with Giemsa dye.

### 3.4.2. Gene distances between MIR-containing genes

The distances between the MIR-containing genes were noted for each chromosome to further illustrate if the genes are randomly distributed or clustered within the human genome (figure 3.7a). As a comparison the average distance between the MIR-containing genes was compared to the average distance between genes of the human genome (Figure 3.7b). Genes that have exaptated MIRs are most commonly located within the range of 0-200kb, which is similar to that of the remainder of the genes in the genome. This similarity in distribution, on first impressions, may suggest that clustering is occurring in gene-rich regions. However, when comparing the MIR-gene dataset to a random collection of human genes (figure 3.7c) a similar trend to the MIR-containing genes is observed, confirming that MIRs may have been randomly exaptated.



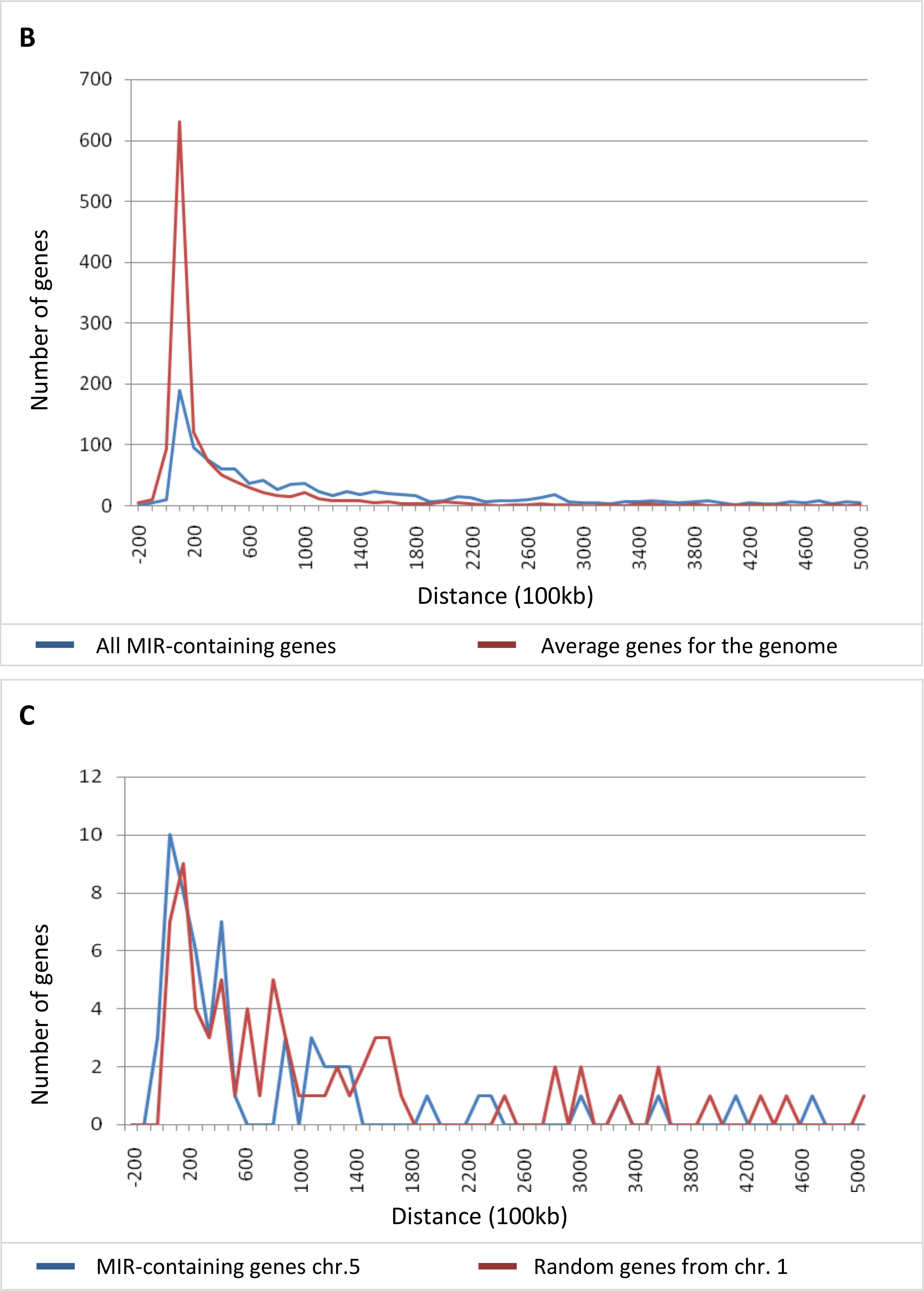


**Figure 3.7. The distance between MIR-containing genes in the human genome**

**A)** The percentage of MIR-containing genes has been calculated for each human chromosome for distance ranges in multiples of 100kb; **B)** The average distance between MIR-containing genes (blue) and the distance between all genes on chromosome 5 is provided as a comparison (red); **C)** The frequency of MIR-containing genes (blue) for chromosome 5 compared to the same number of genes chosen randomly from human chromosome 1 (red).

### 3.4.3. MIRs within imprinting regions and disease loci

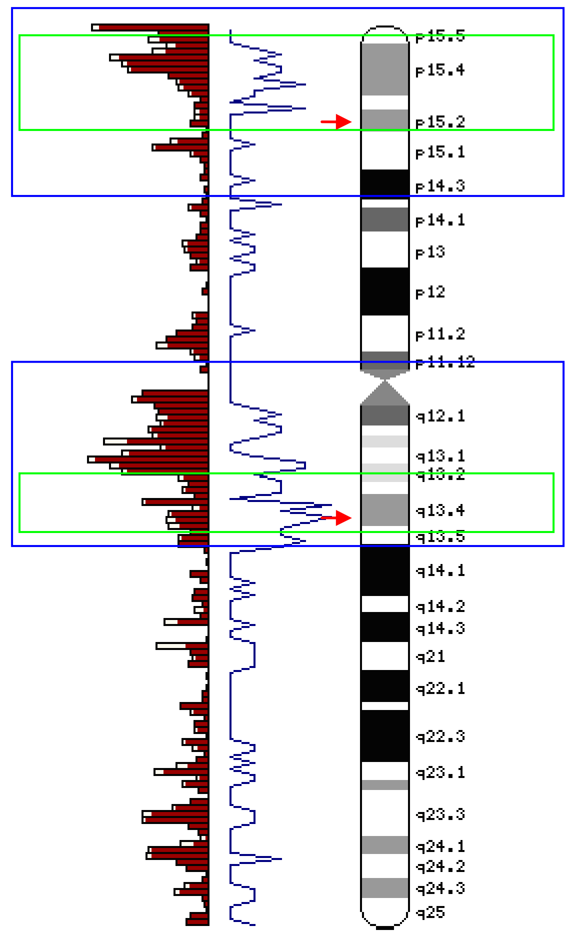
It has been noted previously that imprinted regions are rich in repetitive sequence and it has been suggested that they may act as imprinting markers (Walter *et al.,* 2006). MIR elements have not been specifically implicated in imprinting. However, following analysis of the frequency of MIR-containing genes for each chromosomal location (section 3.4.2), dense regions are notable in chromosomes which possess large imprinting domains. There are 41 known imprinted genes (Morison *et al.,* 2005) and 106 further predicted for the human genome (Luedi *et al.,* 2007; http://www.geneimprint.org/). There are 11 MIR-containing genes listed in the GeneImprint database, 8 of which are known to be imprinted and the remainder being predicted (table 3.3). The density of MIR-containing genes in imprinted regions and the disease loci of chromosomes 11, 14, 19, 20 and 21 are detailed in figure 3.8.

| Gene symbol | Gene name | Location | Allele | ICR | Reference |
|---|---|---|---|---|---|
| CCBL2* | cysteine conjugate-beta lyase 2 | 1p22.2 | M | - | Luedi *et al.,* 2007 |
| PLAGL1 | pleiomorphic adenoma gene-like 1 | 6q24-q25 | P | - | Arima *et al.,* 2006 |
| FAM50B* | Family with sequence similarity 50, member B | 6p25 | M | - | Luedi *et al.,* 2007 |
| DHCR7 | 7-dehydrocholesterol reductase | 11q13.4 | M | - | Schulz *et al.,* 2006 |
| AMPD3 | adenosine monophosphate deaminase 3 | 11p15.4 | M | - | Schulz *et al.,* 2006 |
| MEG3 | maternally expressed 3 | 14q32 | M | P<br>DMR | Miyoshi *et al.,* 2000 |
| RASGRF1 | RAS protein-specific guanine nucleotide-releasing factor 1 | 15q24 | P | P | de la Puente *et al.,* 2002 |
| ZNF597 | Zinc finger protein 597 | 16p13 | M | - | Pant *et al.,* 2006 |
| HM13 | histocompatibility (minor) 13 | 20q11.21 | M | - | Wood *et al.,* 2007 |
| L3MBTL | l(3)mbt-like | 20q13.12 | P | DMR | Li *et al.,* 2004 |
| SIM2* | single-minded homolog 2 (Drosophila) | 21q22.2 | P | - | Luedi *et al.,* 2007 |

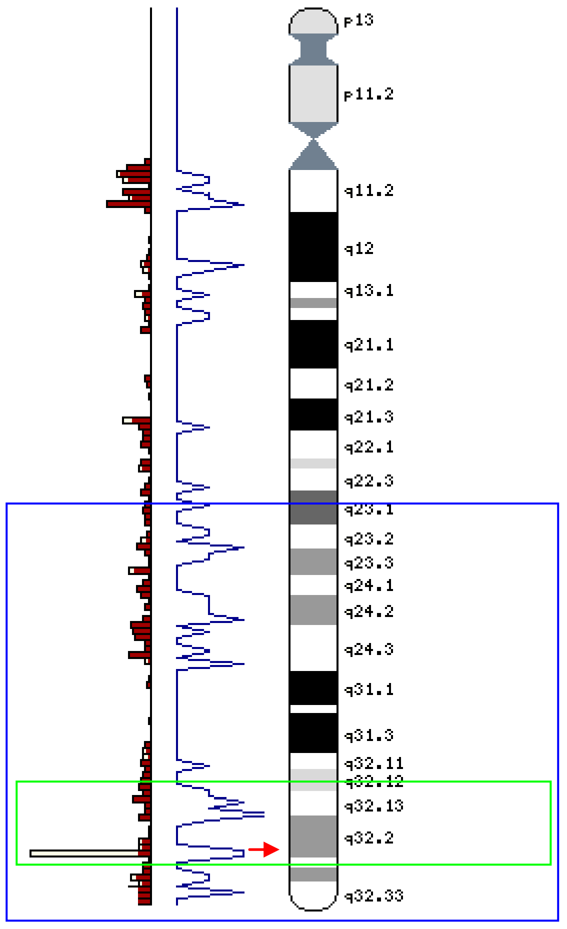
**Table 3.3. Imprinted genes which have recruited MIR elements**

Expression is from either the paternal (P) or maternal (M) allele. The imprinting control region (ICR) has been included if known with two paternal examples and one differentially methylated region (DMR). Genes with the symbol * are listed as predicted in the Gene-Imprint database (Luedi *et al.,* 2007).

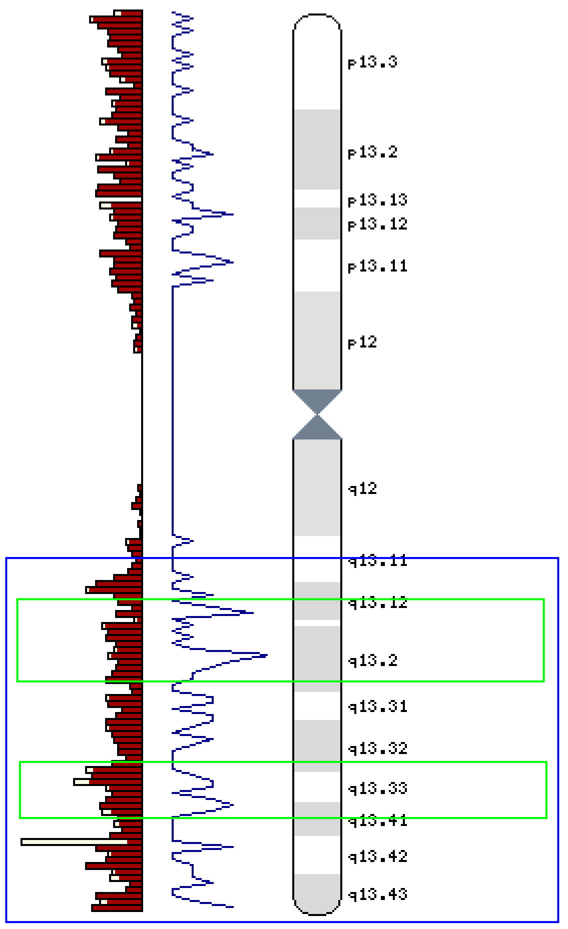
Chromosome 11



Chromosome 14



Chromosome 19



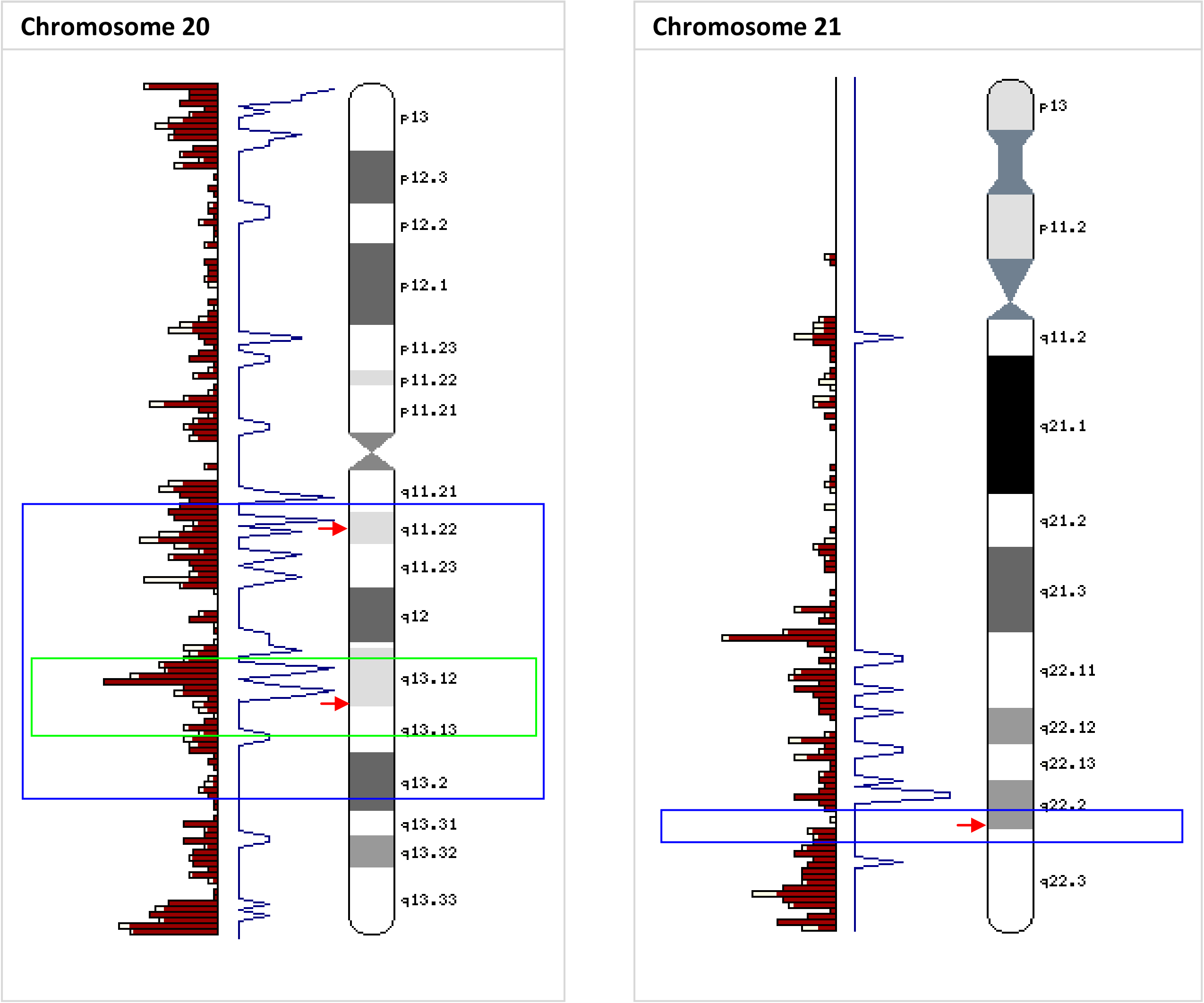


**Figure 3.8. The density of genes which have exaptated MIR elements for chromosomes that are known to contain imprinting domains.**

Recognised imprinting regions and chromosomal areas dense in imprinted genes are outlined in blue. Sections where the densities of MIR-containing genes appear greater than the frequency of genes in the genome are boxed (green). The location of the known imprinting genes which have exaptated MIR elements; DCHR7, AMPD3, MEG3, HM13, L3MBTL and SIM2, are illustrated with red arrows.

The red histogram indicates the frequency of known genes for each chromosomal region and the blue trace the frequency of MIR-containing genes. The chromosome banding (grey) indicates the standard human karyotype, stained with Giemsa dye.

▟ Density of human genes

⋛ Density of MIR-containing genes

□ Recognised imprinting regions

🟩 Concentration of MIR-containing genes

→ Known imprinting genes (with MIR)

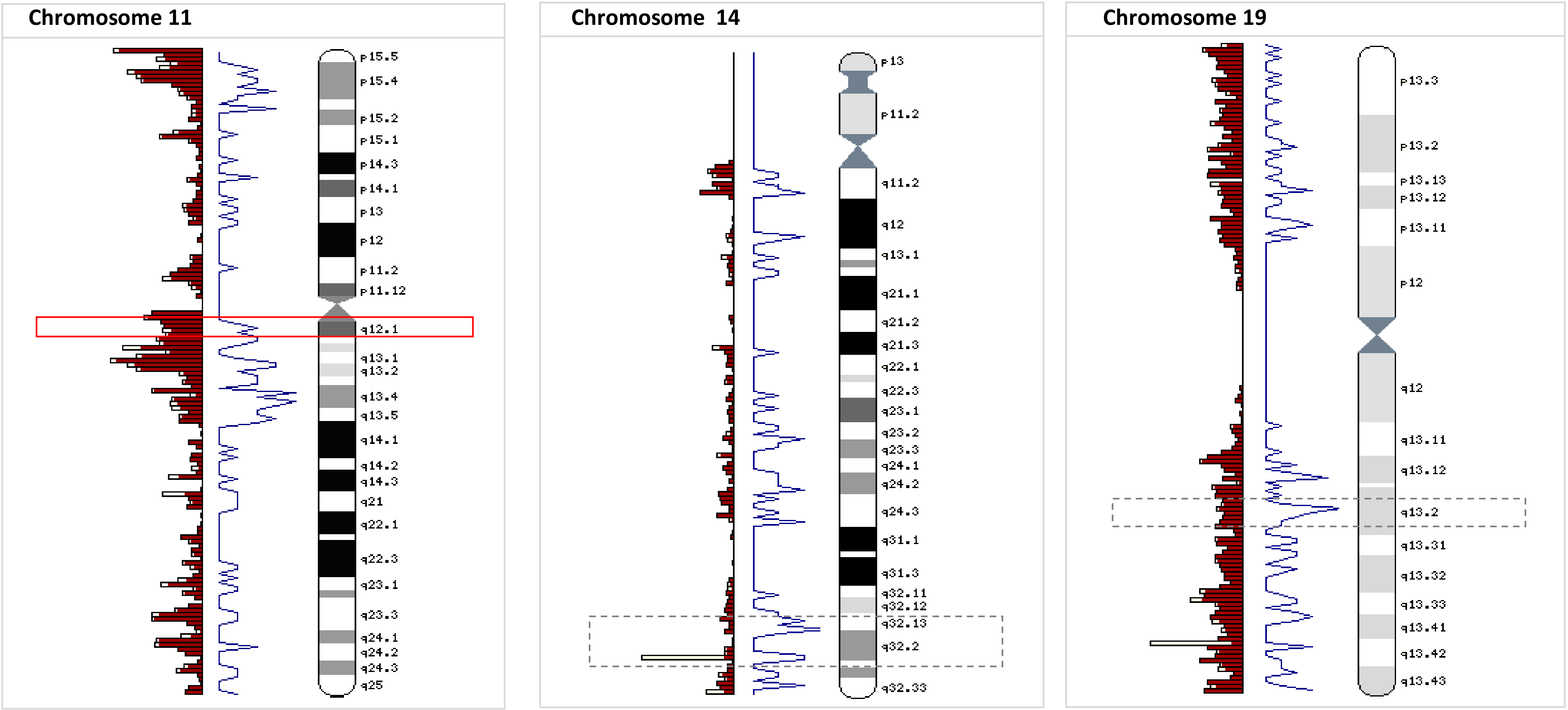


**Figure 3.9. The density of human genes which have exaptated MIR elements for disease loci.**

Known chromosome regions associated with Alzheimer's disease are boxed (grey dashed). The red box signifies a group of six genes in a genomic region associated with Joubert syndrome. The red histogram indicates the frequency of known genes for each chromosomal region and the blue trace the frequency of MIR-containing genes. The chromosome banding (grey) indicates the standard human karyotype, stained with Giemsa dye.

- ▮ Density of human genes
- ≋ Density of MIR-containing genes
- ⬚ Regions associated with Alzheimer's disease
- □ Regions associated with Joubert syndrome

## 3.6. Discussion

Compilation of a dataset of human genes which have recruited MIRs, identified ~5% of the genes in the genome that have exaptated one or more MIR element. This corresponds to 1359 known and validated genes that have recruited MIRs in exons which are either protein-coding or within untranslated regions. This figure is likely to be an under representation of the true number of genes which have recruited MIRs, as many elements may be difficult to detect due to sequence drift and divergence.

The majority of the MIR elements are within the 3'-UTR (75%), and the remainder within the CDS (9%) and 5'-UTR (14%). The prevalence of MIR element within the 3'-UTR was expected as on average the 3'-UTR is six times larger than the 5'-UTR (Zhang, 1998). Secondly the MIR elements have integrated less commonly within the CDS, as a novel insertional event would likely disrupt open reading frames and/or splice sites. MIRs within the 3'-UTRs are more tolerable as protein synthesis can continue undisrupted. Furthermore the frequency of MIR elements in each gene region (5'-UTR, CDS and 3'-UTR) is comparable to that observed for other TEs, with the exception of LTR elements (Piriyapongsa *et al.,* 2007; Jordon *et al.,* 2003; Nekrutenko and Li, 2001; Makałowski, 2000). The abundance of MIRs in the 3'-UTR may merely reflect a lack of selection against the integration of the element due to it being less disruptive. However, it is also possible that some of these MIRs may have been co-opted to produce *cis*-acting regulatory sequences which may play a role in post-translational processes. Likewise MIR elements exaptated in the 5'-UTR and CDS could be providing functional sequences. It has been shown previously that due to the sequence composition some repeat elements, such as Alus can form double-stranded RNAs (dsRNA) which may then be targets for RNA interference (RNAi) pathways, and A-to-I editing (Kim *et al.,* 2004), similar has been suggested for MIR elements (Hughes, 2000). Further roles of the 3'-UTR MIR elements may include regulating translational rates, possibly via miRNA target sites, and altering mRNA stability (Muotri *et al*., 2007; Smalheiser and Torvik, 2005).

There are a number of MIR elements which reside within the CDS of the genes identified; however, it is likely that rather than integrating within a coding exon an intronic repeat underwent sequence changes to produce a cryptic splice site, thus generating alternative transcript variants (see chapter 5). There are several examples where the MIR is providing translational signals such as initiating methionine codons and stop codons thus contributing alternative transcript variants. Following collection of the dataset it also appears that there is no preference to recruit an MIR element in the direct or indirect consensus in relation to the mRNA orientation.

The conservation of the exaptated MIR elements was investigated and demonstrated that the core-SINE region is highly conserved, similar to a previous study (Gilbert and Labuda, 1999). The conservation of the MIR elements located in both the UTRs and CDS were studied independently, as were each MIR family member, and there appears to be a similar trend to conserve the core-SINE irrespective of MIR sub-type and orientation, or the gene region the element is exaptated. The core-SINE is not only highly conserved between human genes but also between species, with a surprising 94% sequence similarity between the core region of human intergenic MIRs and those retainined in intergenic regions of the platypus genome, a similar level of conservation of core-SINEs exaptated within mRNA sequences is also high between these two species (91%). However if MIR elements are non-functional it might be expected that there would be a constant rate of divergence along the whole consensus, yet the core-SINE is highly conserved not only between genes of a single genome but also between species suggesting that the MIR may be functional and maintained due to a selective advantage.

All retrotransposons are thought to have integrated randomly throughout the genome, and various investigations were performed to determine the distribution of the human genes which have exaptated MIR elements. This was achieved by noting the frequency of MIRs for each chromosome and secondly the distribution along each chromosome and comparing to the general distribution of all human genes. If the MIR elements have exaptated randomly within the genome then it would be expected that the number of MIR-containing genes would be proportional to the total number of genes on each chromosome, which is what was observed for the human exaptated MIR elements. When observing the distribution of the MIR-containing genes in more detail there are notable areas where there are high concentrations of the MIR-gene dataset compared to the frequency of the genes in the rest of the genome, specifically in known imprinted regions. There is increasing evidence to support the idea that retrotransposons may contribute to genomic imprinting (Suzuki *et al.,* 2007; Walter *et al.,* 2006; Walter and

Paulsen, 2003), and at least eight imprinted genes appear to have arisen through retrotransposition (Morison *et al.,* 2005; Lander *et al.,* 2001). Secondly retrotransposons which are located either within gene regions or in close proximity have been shown to epigenetically control phenotypic variations in mammalian species (Lippman *et al.,* 2004; Morgan *et al.,* 1999). Furthermore whilst little is known about the epigenetic mechanisms that occur in imprinted domains, it has been suggested that repeat elements may carry the signature sequences which distinguish imprinted genes from other non-imprinted genes or genomic regions (Walter *et al.,* 2006). These signature elements, potentially carried by retrotransposons may then direct the machinery necessary for epigenetic modifications to these areas. It has further been suggested that epigenetic regulation of genes and DNA methylation may have evolved originally as a defence mechanism against the intrusion of a TE and has been demonstrated with *A. thaliana* (Zilberman and Henikoff, 2005; Chan *et al.,* 2005; Lippman *et al.,* 2004).

One of the major chromosomes involved in imprinting is chromosome 11, and there are fourteen genes which have recruited MIRs in the imprinting region 11p15.2-15.5 (appendix 9.5). Secondly eleven of the MIR-containing genes identified are currently known to be imprinted or are predicted imprinted genes (Luedi *et al.,* 2007). In addition there are several chromosomal areas which display a high concentration of MIR-containing genes in regions where evidence supports the parent-of-origin effect in human disease. For example high levels of hypermethylation were observed in patients with the uniparental disomy (UPD) -like phenotype [upd(14)mat] in the DMR region of the MIR-containing gene, maternally expressed 3 (MEG3) (Zechner *et al.,* 2007; Hosoki *et al.,* 2008). There are also dense regions of genes which have exaptated MIRs in chromosome loci with linkage to specific disease. For instance there is a cluster of 24 MIR-containing genes in the known Joubert syndrome (JBTS) locus, 11p12-q13.3 defined by Valente *et al.,* (2005; appendix 9.5). Moreover there is an MIR-gene rich region located at position 14q32.2-13 where there is linkage to Alzheimer's disease and UPD-like phenotype [upd(14)mat] (Zechner *et al.,* 2009; Lee *et al.,* 2007). A further Alzheimer's disease locus has been mapped to 19q13.2, a region which contains 14 MIR-containing genes, including periaxin (PRX) which is essential for the maintenance and stabilisation of peripheral nerve myelin (Williams and Brophy, 2002).

## 3.7. Conclusion

In summary, 5% of the total human genes have exaptated at least one exonic MIR sequence in either the UTRs (91%) or CDS (9%), with more than 1850 exaptated MIR elements identified. There appears to be no preference to recruit an MIR in the direct or indirect orientation and the majority (75%) are retained within 3'-UTR exons. The core-SINE of the MIR consensus sequence is highly conserved between mammalian species, including the ancient orders monotremata and marsupialis. Sequence identity of >90% is detected for core-SINEs integrated in intergenic DNA and mRNA across mammals. The MIR elements appear to be distributed randomly in genes throughout the human genome, as indicated by the density of MIR-containing genes for each chromosome and the sub-chromosomal locations.

# 4. THE FUNCTIONAL SIGNIFICANCE OF THE EXAPTATION OF AN MIR ELEMENT

## 4.1. Introduction

TEs are ubiquitous component of all eukaryote genomes and constitute a large portion of the host genomic sequence. As such, TEs are suggested to play a major role in the evolution of eukaryotic complexity (Jurka, 2008; Medstrand *et al.,* 2005; Bowen and Jordan, 2002). It has been suggested that the immune system may have evolved by co-opting transposons (Schatz, 1999). Adaptive immunity has developed a means of precisely targeting specific pathogens through the production of a myriad of immunoglobulin (Ig) proteins. Ig proteins are encoded by three gene families; V (variable), D (density) and J (joining) which reside within the V(D)J locus. During lymphocyte differentiation somatic recombination occurs of the V(D)J gene region producing a composite gene made up of segments of the three families. V(D)J recombination is fundamental in generating a diverse collection of Ig proteins and T cell receptors, and is catalysed by a number of proteins including RAG1 and RAG2 (Lewis and Wu, 1997; Schatz *et al.,* 1989). The activation mechanism is not dissimilar to the 'cut and paste' mode of transposition observed in DNA transposons and RAG1 is now known to be derived from the *Transib* DNA transposase (Kapitinov and Jurka, 2005; Spanopoulou *et al.,* 1996), RAG2 is also thought to be TE-derived (Schatz, 1999). It is worth noting that the human RAG1 gene has recruited an MIR element in the 3'-UTR (appendix 9.1).

Taking together the similarities between retrotransposition and V(D)J recombination and that RAG1 is TE-derived supports the role of TEs in the evolution of the vertebrate immune system (Schatz, 1999; Bowen and Jordan, 2002), and there is increasing evidence to suggest a role of TEs in other aspects of immune responses. For example the expression of type III interferons (IFN) is thought to be regulated by NF-kappaB, through a cluster of NF-kappaB binding sites located in the IFN promoter region. These binding sites are derived from Alus and LTR elements (Thomson *et al.,* 2009). Similarly the promoter region of human IFN-γ also contains Alu-derived binding sites for NFAT (nuclear factor of activated T-cells) and NF-kappaB (Ackerman *et al.,* 2002).

TEs have the potential to provide *cis*-acting regulatory elements which may then be involved in the regulation of gene expression. At least 10% of human transcription factor binding sites are TE-derived (Polavarapu *et al.,* 2008). MIR elements which have been transposed into gene regions predominantly reside within the 3'-UTRs (chapter 3), which is the site most miRNAs are known to target. The introduction of a novel TE into a gene region may be deleterious to the organism and numerous examples exists whereby human disease has been attributed to an integrated retrotransposon (reviewed in Deininger and Batzer, 1999; Ostertag *et al.,* 2003; Callinan and Batzer, 2006). There are few publications which demonstrate similar activity occurring with inserted MIR elements, with the exception of the example discussed in section 1.4 and the alternative splicing of gene CYBB. The aim of this chapter is to categorise all of the MIR-containing genes which are implicated in the development of disease. The MIR elements may not necessarily be involved in the pathology as such, but sorting in this manner will highlight the functional importance of these genes and may provide a possible indication or hint as to the role of these elements. Similarly the overall functions of the MIR-gene dataset will be sorted to screen for any themes or commonalities, following which specific functional roles will be discussed, such as the involvement in dsRNA-mediated gene expression and RNAi.

## 4.2. Data mining to search for commonalities in the MIR-containing genes

The functional significance of the exaptation of MIR elements was unravelled using a number of sources, aiming to distinguish enriched biological themes and functionally related genes, specifically using Gene Ontology (section 4.2.2) and tissue expression data (section 4.2.3). The involvement of the MIR-dataset of genes in biological regulatory pathways and gene-disease associations were also investigated. The MIR-containing genes were first analysed using D.A.V.I.D (Database for Annotation, Visualisation and Integrated Discovery; Huang *et al.,* 2007), to search for significant terms, collected from multiple databases. This preliminary analysis screened all of the available databases accessible through D.A.V.I.D including GAD, OMIM, NCBI, SWISSPORT and KEGG. The overrepresented significant key terms ($P$ < 0.005) include alternative splicing, glycoprotein, the membrane, interleukin activity and metal binding (table 4.1).

| Keyword | Number of MIR-containing genes | Ratio | FDR adjusted *P* value |
|---|---|---|---|
| Alternative splicing | 486 | 37.67% | <0.001 |
| Glycoprotein | 306 | 23.72% | <0.001 |
| Membrane part | 446 | 34.57% | <0.001 |
| Interleukin activity | 13 | 1.01% | 0.001 |
| Metal-binding | 212 | 16.43% | 0.004 |
| Neurotransmitter transport | 12 | 0.85% | 0.007 |
| Disease mutation | 122 | 9.46% | 0.008 |
| Vision | 20 | 1.55% | 0.02 |
| Zinc-finger | 136 | 10.54% | 0.02 |
| Transmembrane protein | 62 | 4.81% | 0.04 |
| Retinitis pigmentosa | 12 | 0.85% | 0.09 |
| Signal-anchor | 39 | 3.02% | 0.12 |
| Immunoglobulin domain | 54 | 4.19% | 0.19 |
| Symport | 14 | 1.09% | 0.28 |
| Palmitate | 21 | 1.63% | 0.46 |
| Tyrosine-protein kinase | 15 | 1.16% | 0.46 |

**Table 4.1. Significant key terms detected for the MIR-containing gene dataset**

Keywords were detected using D.A.V.I.D (Huang *et al.,* 2007). The number and ratio of the total MIR-containing genes is listed for each term. Statistical significance is calculated using the Benjamini-Hochberg (1995) method for false discovery rate control.

### 4.2.1. Metabolic and signalling pathways

Investigating the involvement of MIR-containing genes in regulatory pathways may highlight common metabolic and signalling processes. The pathway information was collected from the Kyoto Encyclopaedia of Genes and Genomes (KEGG) database (Kanehisa *et al.,* 2008). The number of MIR-containing genes for each pathway was noted and the frequency compared to the total number of genes reported. Filtering the MIR dataset of genes in this manner may indicate a shared function and subsequently a potential role for the MIR elements.

The significance of these results was determined using FDR-corrected method calculated by the functional enrichment tool FatiGO at Babelomics v3.1 (Al-Shahrour *et al.,* 2007). A total of 70 different KEGG pathways were noted to involve genes which have recruited MIR elements (figure 4.1; appendix 9.6). However no pathways were deemed significant ($P$ < 0.05) and all groups gave a $P$ value of 1. The Fisher's exact test highlighted four pathways which have an uncorrected $P$ value of < 0.05 (table 4.2). These results have not been calculated using the Benjamini-Hochberg correction method, so can not be considered statistically significant. The pathways fall into two groups: 1) Cellular communication; including cytokine receptor interaction and Jak-STAT signalling and 2) neuronal development. The neuronal pathways are axon guidance and the development of Amyotrophic lateral sclerosis (ALS), also known as Maladie de Charcot, which is the most common form of motor neuron disease in young adults (Mitchell and Borasio, 2007).

| | Pathway name | KEGG ID | Number of MIR-genes | % (norm) | Uncorrected *P* value |
|---|---|---|---|---|---|
| 1 | Cytokine-cytokine receptor interaction | hsa04060 | 12 | 66.57 | 0.02 |
| | Amyotrophic lateral sclerosis | hsa05014 | 2 | 87.35 | 0.05 |
| 2 | Jak-STAT signalling pathway | hsa04630 | 7 | 70.31 | 0.03 |
| | Axon guidance | hsa04360 | 7 | 77.27 | 0.02 |

**Table 4.2. KEGG pathways which have a large number of MIR-containing genes**

The KEGG pathway name and the identification numbers are listed. The total genes involved in the pathways identified are; 66/1359 MIR-containing genes and 1139/31524 genes of the total genome. *P* values (Fisher's exact test) are calculated using the functional enrichment tool Babelomics (Al-Shahrour *et al.,* 2007). The percentage of MIR-containing genes have been normalised: (% MIR-genes + % genes in genome)/% genes in genome.

### 4.2.2. Gene Ontology

The Gene ontology (GO) is the primary database of controlled terms used to annotate gene functions, products and sub-cellular locations. The GO database is made up of three main ontologies (trees); biological process (BP), molecular function (MF) and cellular component (CC). Each ‘tree’ is structured in a hierarchical manner of multiple branches (nodes), which are organised by parent–daughter relationships between ontologies, with each node branching to a further more detailed GO annotations.

This controlled vocabulary of GO terms is a useful resource which can assist in identifying if there are any common functions shared between the dataset of MIR-containing genes. Accessing the GO may assist in highlighting specific functional groups; however, the GO is an incomplete and growing database, with many genes having yet to be assigned GO terms. Secondly, many terms have few genes assigned. These small sample sizes may result in the FDR-correction method missing potentially significant groups. Therefore the Fisher’s exact test was used to reveal trends which may highlight common themes and functional areas which may warrant further investigation. Revisiting the GO in the future, when the database has expanded may provide more significant results.

GO terms and accessions were collected for all of the MIR-containing genes, corresponding to a total of 2669 ontology terms. It was necessary to categorise these terms into specific levels and relationships, as a gene may have multiple GO terms attributed which belong to the same parent-daughter relationship. The main top level ontologies (incorporating nodes 3 and above) are outlined in table 4.3. Of the top three nodes studied, none of the GO terms were found to be statistically significant (table 4.3a). When consulting the two-tailed Fisher’s exact test the MIR-containing genes appear to be involved primarily in ion binding and transport, neurotransmitter transport, immune responses, homeostasis, intracellular signalling, peptide binding and receptor activity (table 4.3b; figure 4.1).

**A**

| ONTOLOGY TERM | MIR-genes | Ratio (%) | % (norm) | FDR-correct *P* value |
|---|---|---|---|---|
| Receptor activity | 128 | 13.5 | 51.6 | 0.35 |
| Homeostasis | 35 | 4.0 | 59.4 | 0.51 |
| Cell communication | 273 | 30.3 | 51.2 | 0.54 |
| Signal transduction | 247 | 28.2 | 51.0 | 0.61 |
| Reproduction | 18 | 2.0 | 54.7 | 0.64 |
| Neurotransmitter transport | 12 | 1.5 | 78.6 | 0.65 |
| Peptide binding | 18 | 1.9 | 62.8 | 0.65 |
| Cell differentiation | 159 | 18.2 | 53.1 | 0.69 |
| Metabolism | 485 | 53.9 | 49.8 | 0.83 |
| Localisation | 178 | 19.8 | 51.9 | 1 |
| Ion binding | 321 | 33.9 | 55.0 | 1 |
| Immune response | 70 | 7.8 | 59.1 | 1 |
| Intracellular signalling cascade | 111 | 13.5 | 54.6 | 1 |
| Ion transport | 63 | 7.7 | 55.8 | 1 |
| Cellular development | 159 | 17.7 | 53.1 | 1 |
| Transcription | 163 | 19.8 | 50.9 | 1 |
| Apoptosis | 52 | 9.0 | 50.1 | 1 |
| DNA Replication / Modification | 9 | 11.9 | 52.9 | 1 |
| Nucleic acid binding | 211 | 22.3 | 48.8 | 1 |
| Protein localisation | 40 | 4.4 | 45.7 | 1 |
| Translation | 25 | 3.2 | 41.6 | 1 |

**B**

| ONTOLOGY TERM | MIR-genes | Ratio (%) | % (norm) | Uncorrected *P* value |
|---|---|---|---|---|
| Ion binding | 321 | 33.86 | 55.02 | <0.001 |
| Neurotransmitter transport | 12 | 1.46 | 78.58 | <0.001 |
| Immune response | 70 | 7.78 | 59.1 | 0.002 |
| Homeostasis | 35 | 4.00 | 59.39 | 0.02 |
| Intracellular signalling cascade | 111 | 13.49 | 54.61 | 0.03 |
| Peptide binding | 18 | 1.90 | 62.78 | 0.03 |
| Ion transport | 63 | 7.65 | 55.78 | 0.04 |

**Table 4.3. Gene ontology categories for all of the MIR-containing genes**

The gene ontology terms are from the top three nodes of the main ontology trees (BP, CC and MF). The percentage of MIR-containing genes has been normalised: (% MIR-genes + % genes in genome)/% genes in genome. Statistical significance was calculated using the Babelomics resource. **A)** The FDR-correction method demonstrates that none of the top ontology terms are statistically significant ($P < 0.05$). **B)** The Fisher's exact method reveals seven GO terms with a $P < 0.05$.

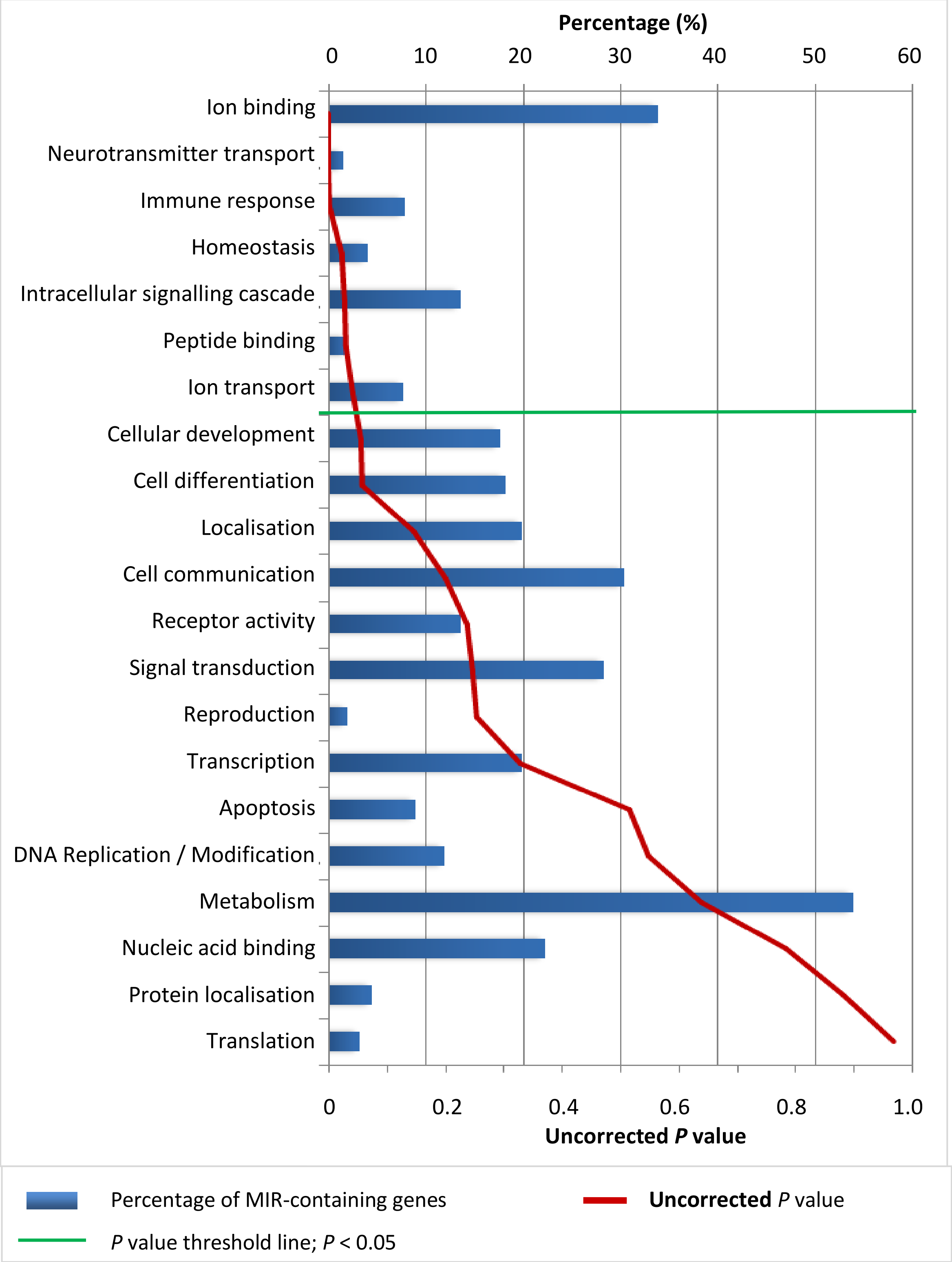


**Figure 4.1. Gene ontology categories for MIR-containing genes and all genes of the genome**

The categories are derived from the Gene Ontology functional classification system. Top level ontology terms are included from all main ontologies; cell component, biological process and molecular function. Blue bars represent the percentage of the total data set of MIR-containing genes which has the GO term assigned and the red line represents the *P* value. The green threshold line signifies the uncorrected *P* < 0.05.

The full collection of gene ontology terms were grouped into specific groups to facilitate the analysis (Nodes 4 and below). The functional groups included growth and development, mammalian reproduction, neurological function, cell component, immune responses and binding activity (figures 4.2-4.7; appendix 9.7). Of these categories only four terms were considered statistically significant (FDR-corrected *P* < 0.05); which include the plasma membrane, metal ion binding, cation binding and the golgi apparatus (table 4.4a). The Fisher's exact test reveals three trends (table 4.4b): 1) Neuronal function; neuronal-related terms include the dendrite and neurone projection, GPCR activity and the photoreceptor. Photoreceptors are a specialised type of neurone specifically located in the retina, which are involved in phototransduction and vision. Note that key terms listed in table 4.1 include retinitis pigmentosa and vision. 2) Immune responses; predominantly the response to wound injury by cytokines, Interleukin activity was identified as a significant key term in table 4.1 (*P* < 0.001). 3) Growth and development; specifically related to skeletal and organ development.

**A**

| ONTOLOGY TERM | MIR-genes | Ratio (%) | % (norm) | FDR-correct *P* value |
|---|---|---|---|---|
| Plasma membrane | 164 | 20 | 56.2 | 0.01 |
| Metal ion binding | 317 | 37.8 | 54.8 | 0.02 |
| Cation binding | 295 | 35.2 | 54.7 | 0.02 |
| Golgi apparatus | 40 | 6.8 | 63.2 | 0.05 |

**B**

| | ONTOLOGY TERM | MIR-genes | Ratio (%) | % (norm) | Uncorrected *P* value |
|---|---|---|---|---|---|
| **1** | Neurotransmitter transport | 12 | 1.5 | 78.6 | <0.001 |
| | GPCR activity | 14 | 1.8 | 65.9 | 0.02 |
| | Dendrite | 6 | 0.8 | 73.5 | 0.03 |
| | Neuron projection | 11 | 1.4 | 66.1 | 0.04 |
| | Photoreceptor inner segment | 2 | 0.2 | 90.5 | 0.04 |
| **2** | Cytokine receptor activity | 71 | 1.6 | 82.6 | <0.001 |
| | Immune response | 7 | 7.8 | 59.1 | 0.002 |
| | Wound healing | 7 | 1.9 | 68.1 | 0.006 |
| **3** | Skeletal development | 28 | 3.6 | 66.1 | 0.001 |
| | Organ development | 97 | 11.8 | 55.1 | 0.03 |
| | Organ morphogenesis | 34 | 4.4 | 58.2 | 0.04 |

**Table 4.4. Gene ontology categories for the MIR-containing gene dataset**

The GO terms are from nodes 4 and below. The percentage of MIR- genes are normalised: (% MIR-genes + % genes in genome)/% genes in genome. **A)** The FDR-correction method (*P* < 0.05). **B)** The uncorrected Fisher's exact method.

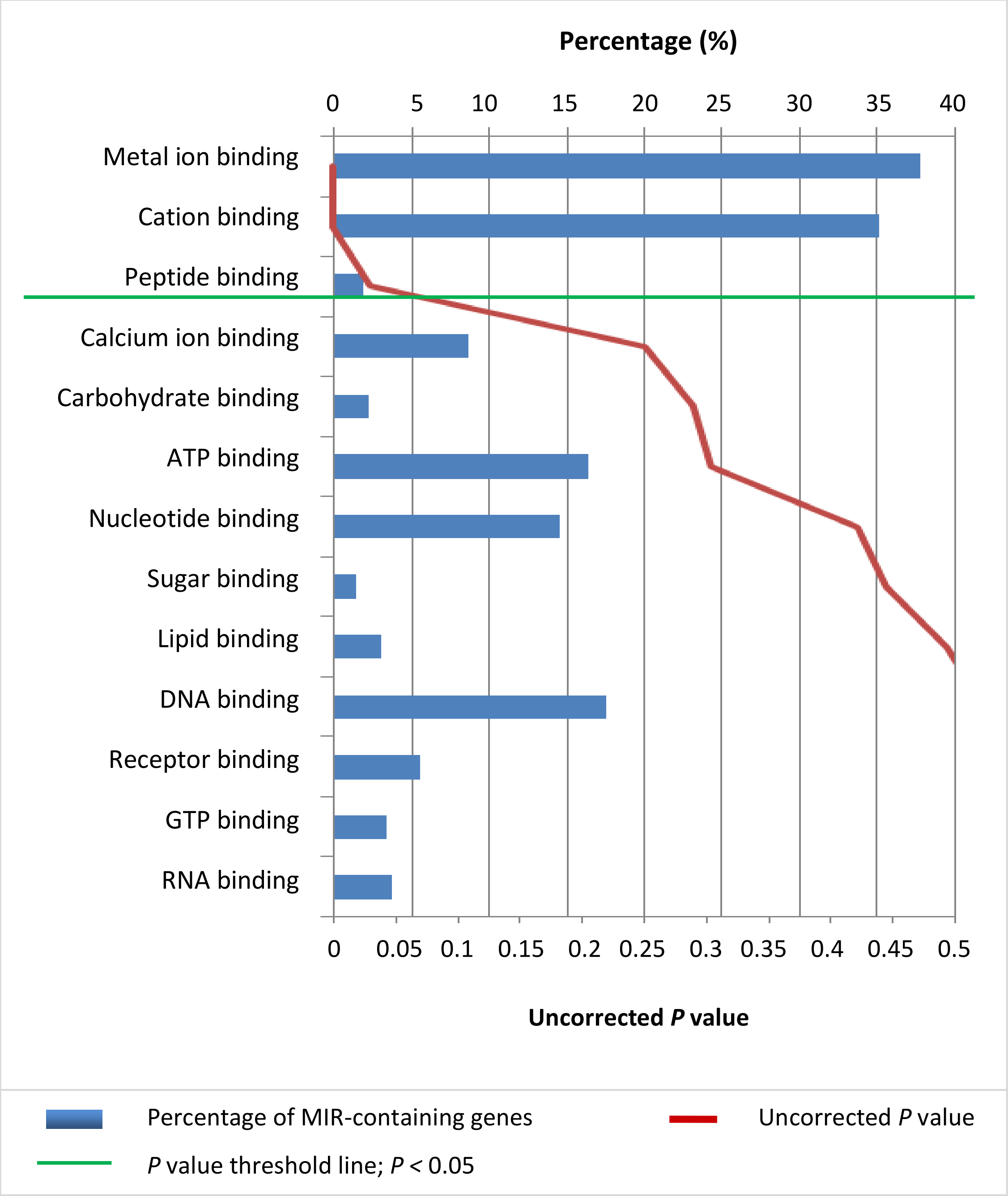


**Figure 4.2. MIR-containing genes for the gene ontology terms in the functional group protein binding**

Gene ontology terms are from all Ontologies (BP, CC and MF) nodes 4 and below. The blue bar represents the percentage of MIR-containing genes assigned each ontology term. The red line signifies the uncorrected *P* value. The green threshold line represents a *P* value of 0.05. The top three terms (Uncorrected *P* < 0.05) are metal ion binding, cation binding and peptide binding.

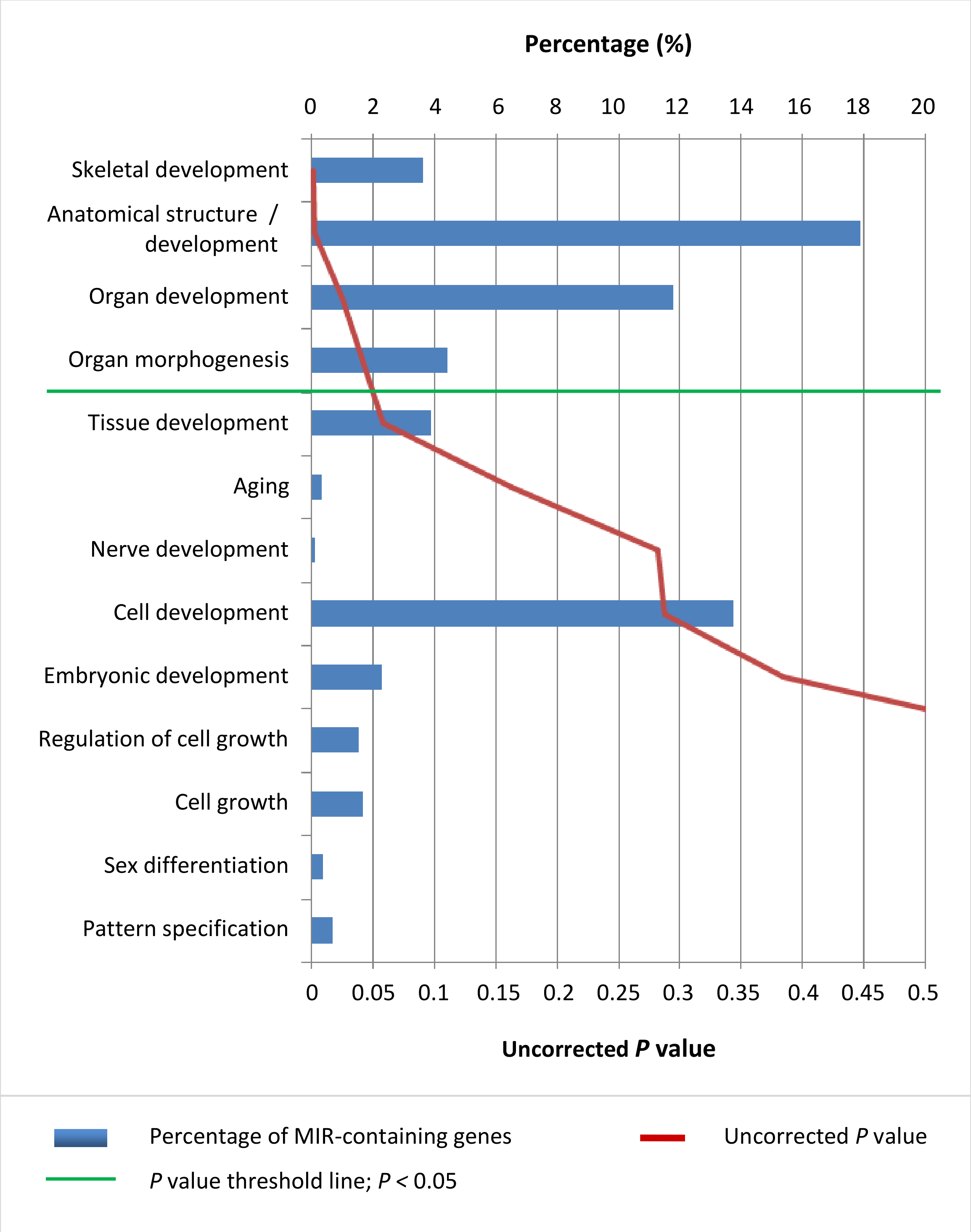


**Figure 4.3. MIR-containing genes for the gene ontology terms in the functional group growth and development**

Gene ontology terms are from all Ontologies (BP, CC and MF) nodes 4 and below. The blue bar represents the percentage of MIR-containing genes assigned each ontology term. The red line signifies the uncorrected *P* value. The green threshold line represents a *P* value of 0.05. The top four terms (Uncorrected *P* < 0.05) are skeletal development, anatomical structure and development, organ development and organ morphogenesis.

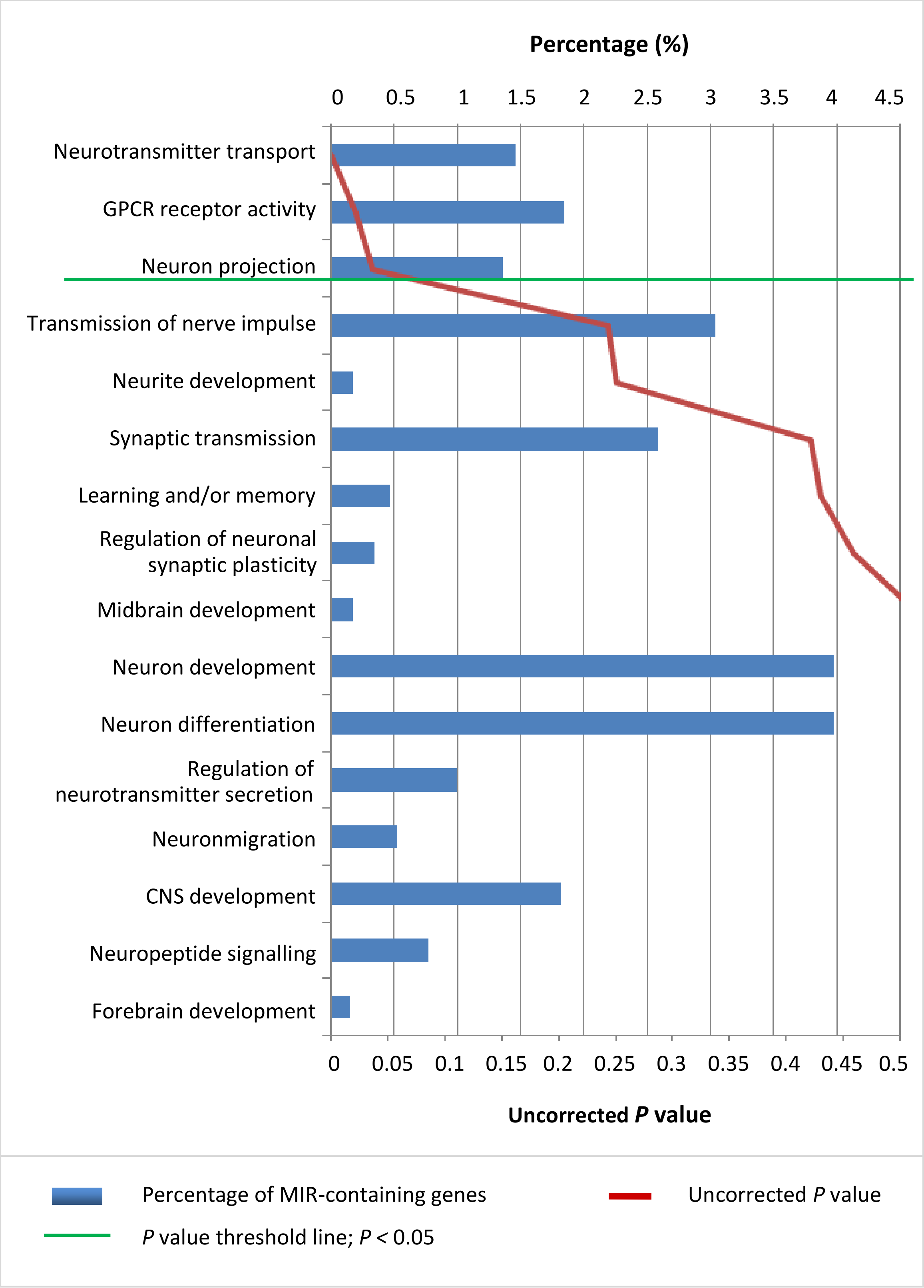


**Figure 4.4. MIR-containing genes for the gene ontology terms in the group neuronal function**

Gene ontology terms are from all Ontologies (BP, CC and MF) nodes 4 and below. The blue bar represents the percentage of MIR-containing genes assigned each ontology term. The red line signifies the uncorrected *P* value. The green threshold line represents a *P* value of 0.05. The top three terms (Uncorrected *P* < 0.05) are neurotransmitter transport, GPCR receptor activity and neuron projection.

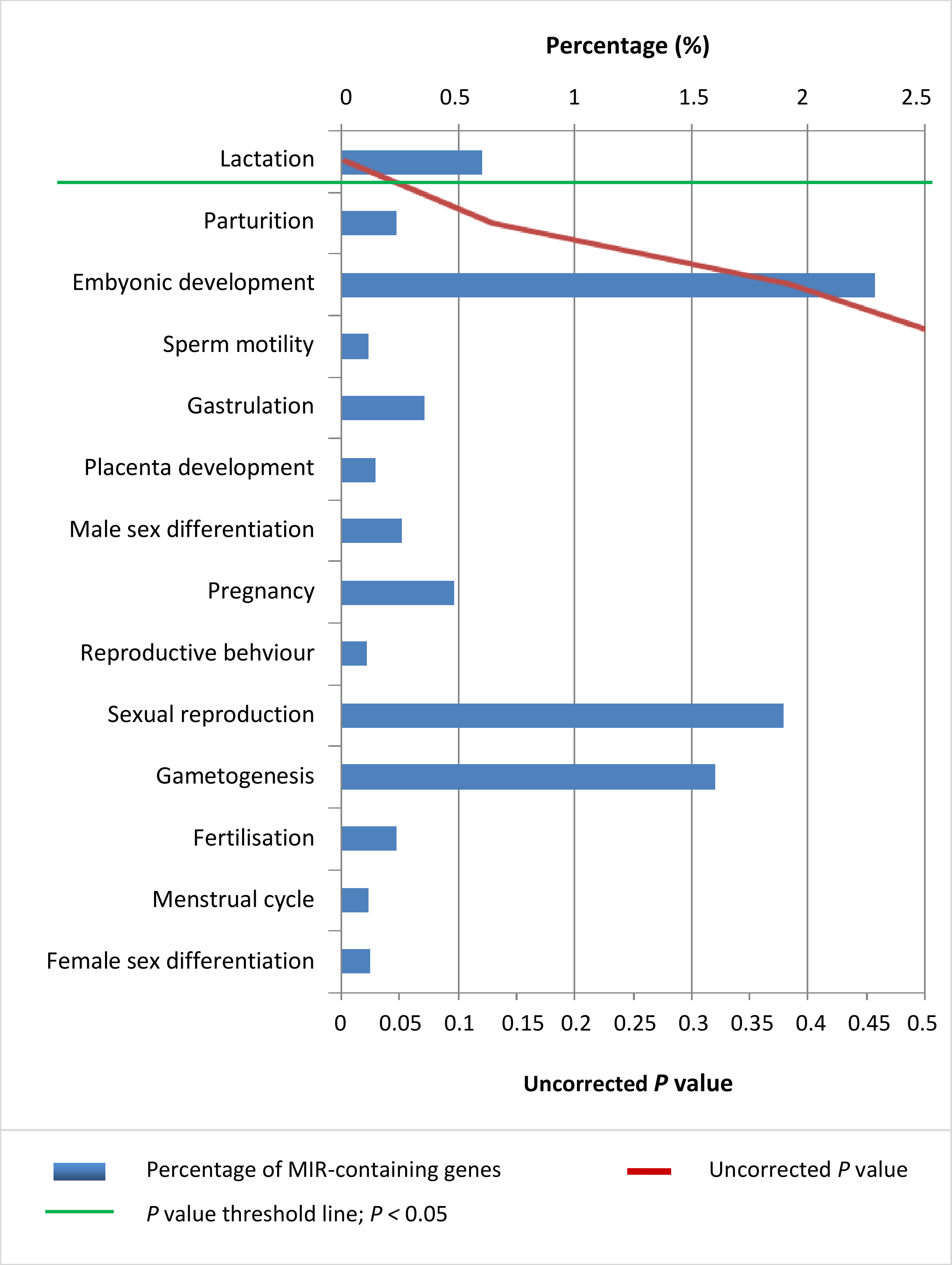


**Figure 4.5. MIR-containing genes for the gene ontology terms in the functional group mammalian reproduction**

Gene ontology terms are from all Ontologies (BP, CC and MF) nodes 4 and below. The blue bar represents the percentage of MIR-containing genes assigned each ontology term. The red line signifies the uncorrected *P* value. The green threshold line represents a *P* value of 0.05. Lactation is the only ontology terms with an uncorrected *P* < 0.05 suggesting MIR elements do not play a role in mammalian reproduction.

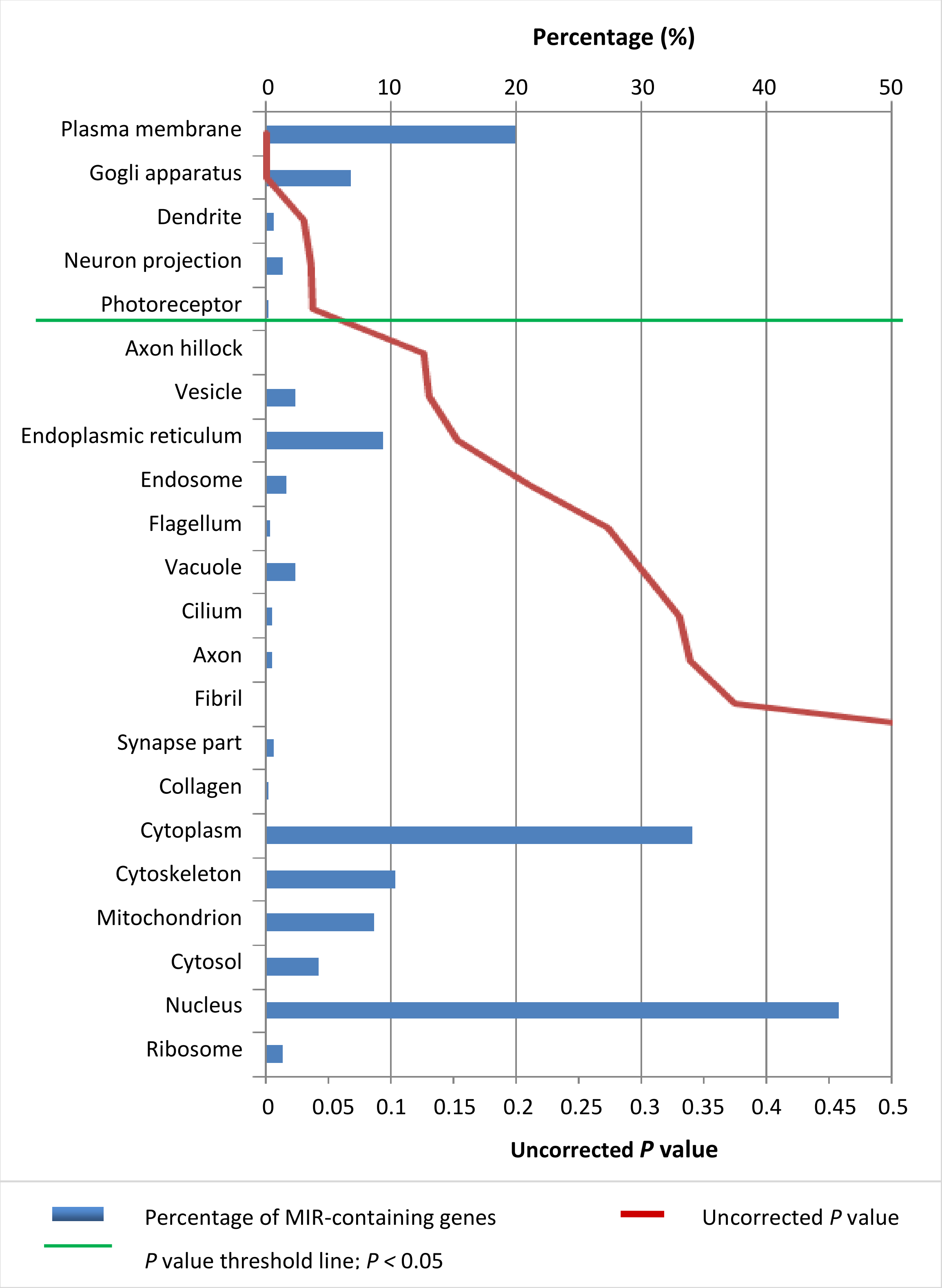


**Figure 4.6. MIR-containing genes for the gene ontology terms in the functional group cell compartment**

Gene ontology terms are from all Ontologies (BP, CC and MF) nodes 4 and below. The blue bar represents the percentage of MIR-containing genes assigned each ontology term. The red line signifies the uncorrected *P* value. The green threshold line represents a *P* value of 0.05. The top ontology terms (*P* < 0.05) are the plasma membrane, golgi apparatus, the dendrite, neurone projections and photoreceptors.

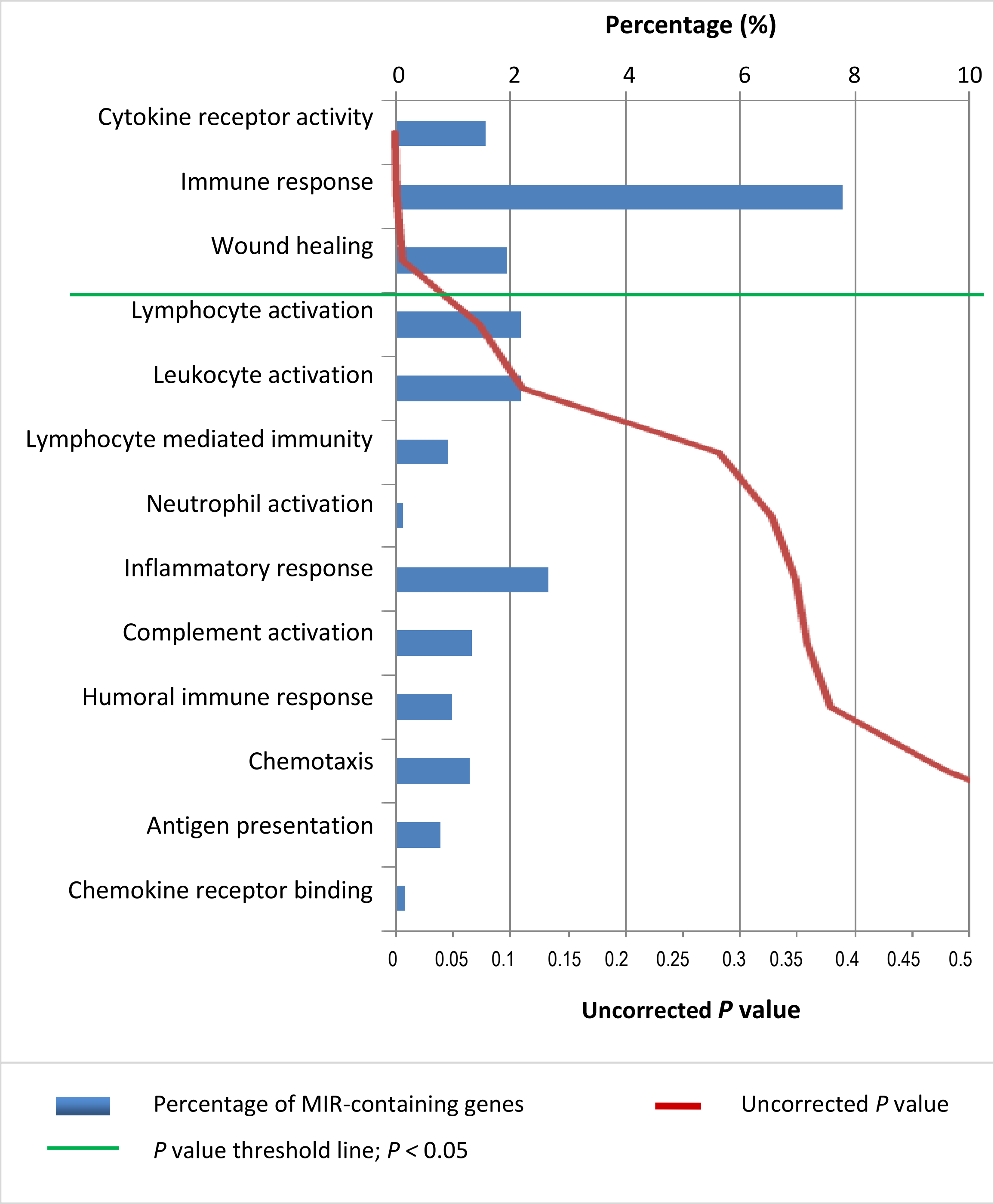


**Figure 4.7. MIR-containing genes for the gene ontology terms in the functional group immune responses**

Gene ontology terms are from all Ontologies (BP, CC and MF) nodes 4 and below. The blue bar represents the percentage of MIR-containing genes assigned each ontology term. The red line signifies the uncorrected *P* value. The green threshold line represents a *P* value of 0.05. The top three terms ($P < 0.05$) are cytokine receptor activity, the immune response amd wound healing.

### 4.2.3. Tissue expression profile of the MIR-containing genes reveals a potential role in neuronal function and mRNA localisation

Studying the tissue-specific expression pattern of mRNA may provide an indication of a shared function of the MIR dataset of genes. Expression data was collected from the human tissue atlas datasets U133A and GNF1H, (Su *et al.,* 2004) which includes whole-genome gene expression arrays and expression profiles for >40,000 human genes (annotated and predicted) across a panel of 79 tissue types.

The tissue expression profiles of the MIR-containing genes were examined to determine if the dataset is overrepresented for any of the specific tissue types. The expression data was analysed using the D.A.V.I.D functional enrichment analysis tool (Huang *et al.,* 2007) to determine if the MIR-dataset of genes are expressed in a particular tissue more frequently than the total genes included in the tissue atlas. The MIR-containing genes were noted to be expressed in 29 of the 79 tissues with a significance value of $P < 0.05$. Interestingly half of these tissues are rich in neurones; brain, spinal cord and the eye including the ciliary ganglion which is a type of neurone located at the posterior orbit of the eye (figure 4.8; table 4.5)). Other highly significant ($P < 0.05$) tissues include the testis, thymus, whole blood, kidney, prostate, uterus and bone marrow.

The identified tissues were organised into functional groups with four groups being apparent (table 4.5). Overall it appears that the MIRs are preferentially recruited and maintained in genes expressed in tissues rich in neurones (48%), mammalian reproductive tissues (15%), muscular tissue (12%) and tissues critical in the production of immune cells (9%).

| | Tissue Type | Number of MIR-genes | Ratio (%) | FDR-correct *P* value |
|---|---|---|---|---|
| **1** | Amygdala | 331 | 25.7 | <0.001 |
| | Ciliary ganglion | 360 | 27.9 | <0.001 |
| | Pituitary | 292 | 22.6 | <0.001 |
| | Occipital Lobe | 286 | 22.2 | <0.001 |
| | Olfactory bulb | 258 | 20 | <0.001 |
| | Prefrontal Cortex | 273 | 21.2 | <0.001 |
| | Cerebellum | 697 | 54 | 0.001 |
| | Temporal lobe | 294 | 22.8 | 0.005 |
| | Sub-thalamic nucleus | 216 | 16.7 | 0.01 |
| | Cingulate cortex | 179 | 13.9 | 0.02 |
| | Pons | 410 | 31.8 | 0.02 |
| | Cerebellum peduncles | 125 | 9.7 | 0.03 |
| | Spinal cord | 166 | 12.9 | 0.04 |
| | Whole eye | 367 | 28.5 | 0.04 |
| **2** | Testis | 268 | 20.8 | <0.001 |
| | Prostate | 673 | 52.2 | <0.001 |
| | Uterus | 227 | 17.6 | <0.001 |
| | Ovary | 207 | 16.1 | 0.003 |
| **3** | Smooth muscle | 330 | 25.6 | <0.001 |
| | Tongue | 276 | 21.4 | 0.002 |
| | Skeletal muscle | 487 | 37.8 | 0.03 |
| **4** | Whole blood | 415 | 32.2 | <0.001 |
| | Thymus | 229 | 17.8 | <0.001 |
| | Bone marrow | 182 | 14.1 | 0.002 |
| | Colorectal Adenocarcinoma | 321 | 24.9 | <0.001 |
| | Atrioventricular node | 132 | 10.2 | <0.001 |
| | Kidney | 460 | 35.7 | <0.001 |
| | Bronchial epithelia | 146 | 11.3 | 0.003 |
| | Appendix | 274 | 21.2 | 0.02 |

**Table 4.5. Tissue expression profile for all of the MIR-containing genes from the human tissue atlas (U133A and GNF1H).**

The experession profiles of the MIR-containing genes was calculated according to the human tissue atlas datasets U133A and GNF1H The MIR elements are predominately expressed in four tissue groups: **1)** Tissues rich in neurones; **2)** tissues of the mammalian reproductive system; **3)** tissues rich in immune cells and **4)** muscular tissue. Statistical significance was determined by the FDR-correction method calculated using the Babelomics resource.

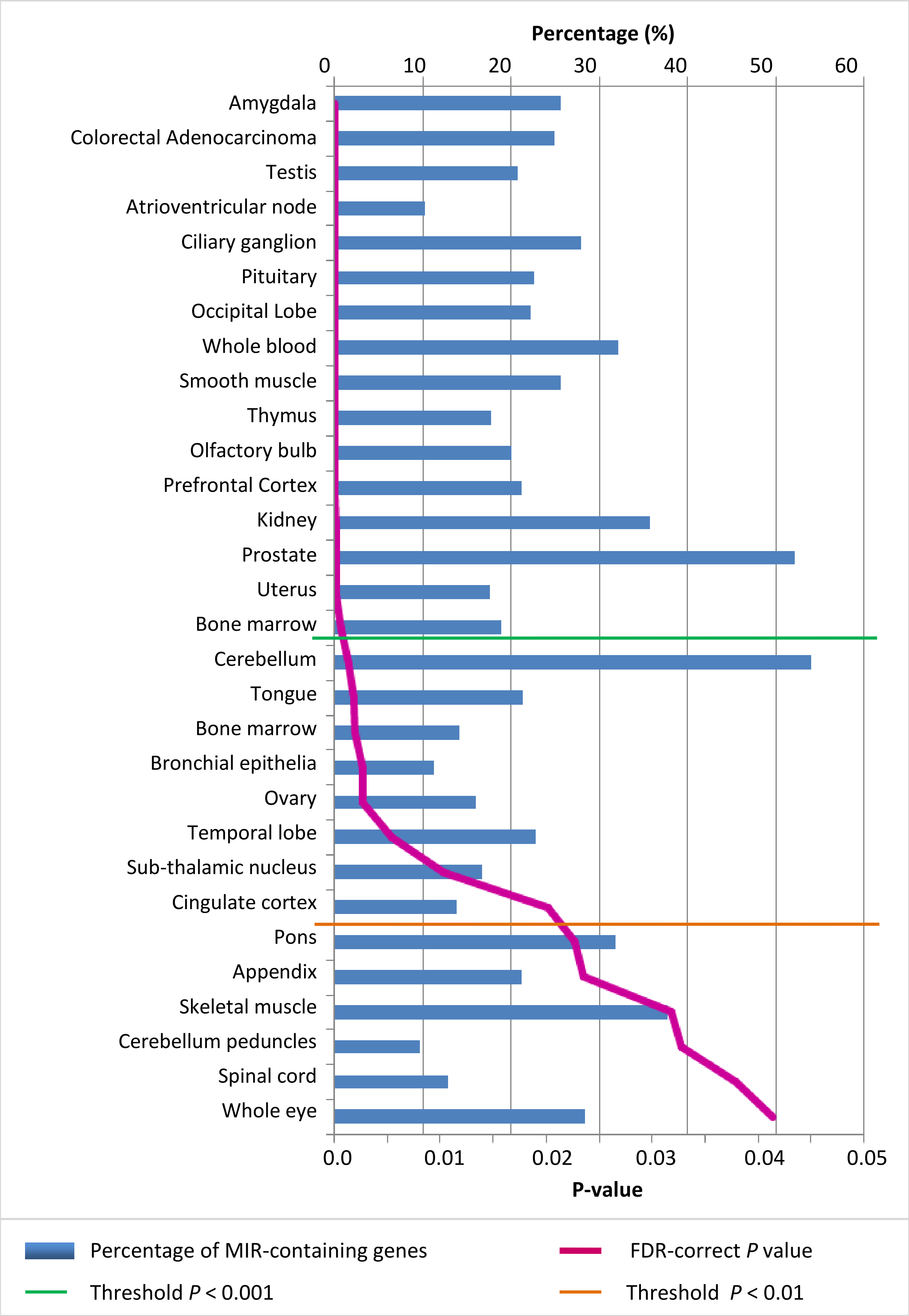


**Figure 4.8. Tissue expression profile of the MIR-containing genes**

The blue bar represents the frequency of MIR-containing genes expressed in each tissue type. The pink line represents the FDR-corrected *P* value; orange line *P* < 0.001 and the green line *P* < 0.01.

Zhong *et al.,* (2006) identified genes which were overrepresented in the dendritic compartment (compared to the cell body) and annotated 154 genes where the mRNA is localised to the dendritic compartment. Of these genes 18 are noted to have exaptated MIR elements (table 4.6). Therefore of the genes described which demonstrate dendritic localisation, 10.5% have been found to contain MIRs (for comparison the total number of genes in the human genome which have recruited these repeats is ~5%). Of these genes none have MIR elements recruited in the coding sequence and the majority have exaptated MIRs in the 3'-UTR (89%) with three having recruited MIR elements in the 5'-UTR.

| HGNC Symbol | Gene Description | Transcript Accession | Position |
|---|---|---|---|
| APLN | Apelin, AGTRL1 ligand | NM_017413 | 3'-UTR |
| CAMK2A | Calcium/calmodulin-dependent protein kinase II alpha | NM_015981<br>NM_171825 | 3'-UTR<br>3'-UTR |
| CD4 | CD48 antigen | NM_000616 | 3'-UTR<br>3'-UTR |
| CD59 | CD59 antigen | NM_000611<br>NM_203329<br>NM_203330<br>NM_203331 | 3'-UTR x5<br>3'-UTR x5<br>3'-UTR x5<br>3'-UTR x5 |
| CLDN10 | Claudin 10 | NM_182848<br>NM_006984 | 3'-UTR<br>3'-UTR |
| DDN | Dendrin | NM_015086 | 3'-UTR |
| EDG6 | Endothelial differentiation, lysophosphatidic acid GPCR 2 | NM_003775 | 3'-UTR |
| FILIP1 | Filamin A interacting protein 1 | NM_015687 | 3'-UTR |
| IGSF1 | Immunoglobulin superfamily, member 1 | NM_205833 | 3'-UTR |
| KCNK3 | Potassium channel, subfamily K, member 3 | NM_002246 | 3'-UTR |
| MAP3K8 | Mitogen-activated protein kinase kinase kinase 8 | NM_005204 | 3'-UTR |
| NEURL | Neuralized-like (Drosophila) | NM_004210 | 3'-UTR |
| NTSR2 | Neurotensin receptor 2 | NM_012344 | 3'-UTR |
| PCBP4 | Poly(rC) binding protein 4 | NM_020418<br>NM_033009 | 5'-UTR<br>5'-UTR |
| RNASE4 | Ribonuclease, RNase A family 4 | NM_194430 | 5'-UTR |
| RPL28 | Ribosomal protein L28 | NM_000991 | 3'-UTR<br>3'-UTR |
| TSPAN2 | Tetraspanin 2 | NM_005725 | 3'-UTR |
| WIT1 | Wilms tumour upstream neighbour 1 | NM_015855 | 5'-UTR<br>5'-UTR<br>5'-UTR |

**Table 4.6. Genes showing dendritic localisation which have recruited MIR elements**

Genes where the mRNA is localised to the dendritic compartment as identified by Zhong *et al.,* 2006 and Pinkstaff *et al.,* 2001. The gene symbol, protein description and GenBank accession number is included. The gene region the MIR has exaptated is listed as CDS (coding sequence) or UTR (untranslated region).

## 4.3. MIR elements and human disease

The functional importance of a gene in a particular tissue is clearly demonstrated if mutations in that gene disrupt normal tissue function and thus disease. The MIR-containing genes which are involved in human diseases were identified and grouped to assist in further understanding the commonality in function of the dataset. The MIR elements may not necessarily be involved directly in the pathology of a disorder, but sorting in this manner will highlight the importance of these genes in specific tissue functions. The disease data was collected from AceView at NCBI, which contains sequences of human genes involved in disease, either directly via mutation, or indirectly. Therefore information was only included in the analysis if peer reviewed publications implicated the gene directly.

It became apparent that a number of MIR-containing genes are implicated and/or mutated in a number of human neurological disorders such as Huntington's disease, mental retardation, schizophrenia and Alzheimer's disease (table 4.7). Other key disease groups were noted including, diabetes and insulin resistance, deafness, inflammation, rheumatoid arthritis and hypertension. A total of 28 genes which have recruited MIRs were noted to be mutated in retinal degenerative disorders and related syndromes (table 4.7). This retinal disease category was selected for further analysis, due to the previous observations of MIR elements being significantly ($P$ < 0.001) expressed in neuronal tissues, including the whole eye and the occipital lobe (visual processing centre of the brain; figure 4.5). Retinitis pigmentosa was also a key term observed in table 4.1, and the photoreceptor and photoreceptor inner segment where two ontology terms noted previously (table 4.4, figure 4.6). Finally photoreceptors are a specialised polarised cell type, with at least one example of an MIR-containing mRNA being localised to the photoreceptor outer segement, critical in photoreceptor development (Gomi *et al.,* 2000).

The retinal disease category was sorted into particular conditions such as cone/rod dystrophies and retinal degeneration, retinitis pigmentosa and Leber congenital amaurosis (LCA) which is an autosomal recessive disorder resulting in blindness due to abnormal photoreceptor development (table 4.8). Furthermore some of the conditions such as retinal degeneration and mental retardation are characteristics of particular syndromes such as Joubert syndrome and Usher syndrome.

| Disorder | MIR-containing gene |
|---|---|
| Alzheimer's disease | ADAM19, BCAS2, C1RL, CCR5, CD59, DENND2C, DHCR24, ESR1, HFE, IL1A, IREB2, KLC1, MAOB, MAPKAPK2, MPO, PSEN2, SERPINA5, TAP2, TGFB1, TGM2, WIT1, YWHAZ |
| Deafness and hearing loss | ATP6V1B1, BSND, CATSPER2P1, FXC1, GJB3, KCNK6, IGF1, LHFPL5, MYO15A, MYO7A, PIK3AP1, POLG, SCARB2, SLC45A2, TMC1, TP53, TRIOBP, UBE3B, USH2A, ZFAND5 |
| Huntington's disease | CCBL2, HAP1, IGF1, IL12RB2, IL8, MAOB, MSX1, SH3BP2, TGM2, TMEM139, TP53 |
| Mental retardation | ARHGEF6, CHL1, CLN8, DCX, FGD1, FKRP, FREQ, IGF1, JARID1B, LRRC48, OPHN1, PIK3AP1, SMC1A, SRGAP3 |
| Retinal diseases | AHI1, AIPL1, ATM, BBS7, CACNA2D4, CLN5, CRB1, CX3CR1, CYP4V2, EFEMP2, GUCA1B, HPS1, KCNJ13, LRAT, MC1R, MYO7A, NPHP1, NR2E3, NRL, RD3, RGS9BP, RHO, RIMS1, SAG, SLC45A2, TGFB1, TPP1, USH2A |
| Schizophrenia | ADRA1A, AHI1, CHL1, CNTF, DBH, DISC1, DPSL2, ERBB3, FREQ, GPR78, HTR4, IGF1, MAOB, PLXNA2, SH3TC2, SLC1A2, SYNGR1, TGM2, TP53, VDR, ZFP91 |
| Diabetes (Type I, II and MODY) | ACE2, ADIPOQ, BDKRB2, BTC, CASR, CCR5, CD4, CD48, CPM, DPP4, ENSA, FABP2, FXN, HFE, HK2, HNF4A, IGF1, IL1A, ITGA2, ITGB3, LEP, MAP4K5, MBL2, MICA, MYCBP, OAS1, PLAGL1, SAPS3, SCD, SERPINA5, SLC2A5, SLC2A10, SLC9A1, SLC11A1, TAP2, TCF2, TGFB1, TGFBR2, TGM2, TNRC6A, TP53, VDR |
| Insulin resistance | ADIPOQ, BCAS2, BDKRB2, DENND2C, FABP2, HFE, HNF4A, LEP, SEPN1 |
| Obesity | C11orf57, DBH, HFE, IL8, IREB2, MAOB, MAP3K14, POLG, SERPINA5, SFXN5, SYT11, TOR1A, VDR |
| Hypertension | ACE2, ADD2, ADIPOQ, ADRA1A, AVPR2, BDKRB2, BMPR2, CAPN5, CCR5, DBH, ESR1, IGF1, IL8, LEP, NFXL1, PRRX1, PTGIS, SCNN1G, SERPINA5, SRC, TGFB1, VDR |
| Rheumatoid arthritis | ATP6V1G2, BCAS2, C1RL, CASP10, CDKN2A, CIITA, CUL1, DENND2C, ESR1, FSTL1, GAB2, GALNT10, HFE, MFAP3, MBL2, MICA, OSM, RUNX1, SERPINA5, SLC11A1, TAP2, TGFB1, TP53, VDR |

**Table 4.7. MIR-containing genes which are known to be mutated and/or implicated in human disease**

The MIR-containing genes are implicated in a large number of human disorders, those conditions which appear more commonly have been listed. The gene symbol is included, consult appendix 9.1 for details regarding the MIR coordinates and gene position it has been recruited.

### 4.3.1. Usher syndrome

Usher syndrome is an autosomal recessive condition of which there are three types, numbered type I-III, according to the severity, with type III being the least severe (Pennings *et al.,* 2003; Möller *et al.,* 1989). The condition is characterised by congenital deafness and blindness including retinitis pigmentosa. Mutations in MYO7A (Myosin VIIA) are associated with the type I syndrome (40% of cases) and mutations in USH2A (Usher syndrome 2A) are responsible for most cases of type II Usher syndrome (60%; Maubaret *et al.,* 2005). Both of these genes express mRNA transcripts which have recruited MIR elements in the coding sequence. A splice variant of MYO7A where the stop codon is provided by the MIR element, results in a truncated protein product (appendix 9.8), of which the MIR element is providing the stop codon. Two transcript variants of USH2A have also been described and an MIR element has been exaptated in the middle of protein-coding exon 45 of the full length isoform.

### 4.3.2. Joubert syndrome

Joubert syndrome (JBTS) is an autosomal recessive disorder characterised by multiple symptoms including retinitis pigmentosa, renal disease, physical deformities and malformation of the brain and subsequently ataxia and mental retardation (Valente *et al.,* 2008). Genes which have recruited MIR elements and are mutated in JBTS are AHI1 (Abelson helper integration site 1) and NPHP1 (nephronophthisis 1 (juvenile)). There is an alternative truncated transcript of AHI1 in which an MIR element is providing the stop codon; this is discussed further in chapter 5. NPHP1 has exaptated an MIR element in the 3'-UTR and mutations in NPHP1 are known to be responsible for Joubert syndrome, furthermore NPHP1 is the causative agent in the majority of juvenile nephronophthisis cases, a type of autosomal recessive kidney disease and a symptom developed in Joubert syndrome cases (Simms *et al.,* 2009).

| Gene Symbol | Gene description | MIR location | Disorder details |
|---|---|---|---|
| AHI1 | Abelson helper integration site 1 | TAG | RD; Recessive JS (includes RP and LCA). |
| AIPL1 | Aryl hydrocarbon receptor interacting protein-like 1 | 3'-UTR | Dominant CRD; Recessive LCA and other early onset severe retinal dystrophies. |
| BBS7 | Bardet-Biedl syndrome 7 | 3'-UTR | Recessive Bardet-Biedl syndrome. |
| CACNA2D4 | Calcium channel, voltage-dependent, alpha 2/delta subunit 4 | 3'-UTR<br>3'-UTR<br>3'-UTR | RD; Recessive CRD. |
| CLN5 | Ceroid-lipofuscinosis, neuronal 5 | 3'-UTR | Neuronal ceroid lipofuscinosis. |
| CRB1 | Crumbs homolog 1 (Drosophila) | 3'-UTR | Recessive RP; Recessive LCA; Dominant pigmented paravenous chorioretinal atrophy; Early onset severe retinal dystrophies; Blindness. |
| CYP4V2 | Cytochrome P450, family 4, subfamily V, polypeptide 2 | 3'-UTR | RD; Recessive Bietti crystalline corneoretinal dystrophy. |
| GUCA1B | Guanylate cyclase activator 1B (retina) | 3'-UTR | RD; Dominant RP; Dominant MD. |
| HPS1 | Hermansky-Pudlak syndrome 1 | 3'-UTR | Oculocutaneous albinism. |
| KCNJ13 | Potassium inwardly-rectifying channel, subfamily J, member 13 | 3'-UTR | Dominant vitreoretinal degeneration, snowflake. |
| LRAT | Lecithin retinol acyltransferase (phosphatidylcholine-retinol O-acyltransferase) | 3'-UTR | RD; Blindness; Recessive RP; Recessive LCA. |
| MC1R | Melanocortin 1 receptor (alpha melanocyte stimulating hormone receptor) | 3'-UTR | Oculocutaneous albinism. |
| MYO7A | Myosin VIIA | TAG | Recessive US type 1b; Recessive congenital deafness without RP. |
| NPHP1 | Nephronophthisis 1 (juvenile) | 3'-UTR | Recessive Senior-Loken syndrome; Recessive nephronophthisis, juvenile; Recessive JS; RD; RP. |
| NR2E3 | Nuclear receptor subfamily 2, group E, member 3 | 5'-UTR | Recessive enhanced S-cone syndrome; Recessive RP; dominant RP; RD; MD; Retinal dysplasia; Retinal dystrophies. |
| NRL | Neural retina leucine zipper | 3'-UTR | Dominant and recessive RP; RD; Blindness. |

| Gene Symbol | Gene description | MIR location | Disorder details |
|---|---|---|---|
| RD3 | Retinal degeneration 3 | 3'-UTR | RD; Recessive LCA and other early onset severe retinal dystrophies. |
| RGS9BP | Regulator of G protein signalling 9 binding protein | 5'-UTR | Recessive delayed cone adaptation; Brandyopsia; Retinal and MD. |
| RHO | Rhodopsin (opsin 2, rod pigment) (retinitis pigmentosa 4, autosomal dominant) | 3'-UTR<br>3'-UTR | Dominant RP; Dominant congenital stationary night blindness; Recessive RP; Retinal dystrophies; RD; MD |
| RIMS1 | Regulating synaptic membrane exocytosis 1 | 3'-UTR | RP; Dominant CRD. |
| RPGR | Retinitis pigmentosa GTPase regulator | 3'-UTR | Recessive; X-linked RP, Dominant; X-linked CRD 1; Recessive X-linked atrophic MD. |
| SAG | S-antigen; retina and pineal gland (arrestin) | 3'-UTR | Recessive Oguchi disease; Recessive RP; Congenital stationary night blindness. |
| SLC45A2 | Solute carrier family 45, member 2 | 3'-UTR<br>3'-UTR | Oculocutaneous albinism. |
| TPP1 | Tripeptidyl peptidase I | 3'-UTR | Neuronal ceroid lipofuscinosis. |
| USH2A | Usher syndrome 2A (autosomal recessive, mild) | CDS | Recessive Usher syndrome, type 2a; Recessive RP. |

**Table 4.8. The involvement of MIR-containing genes in specific retinal disease and associated syndromes**

The HGNC gene symbol, protein description and position of the exaptated MIR element are listed; CDS, coding sequence; UTR, untranslated region; TAG, MIR element provides termination codon sequence. Retinal diseases are abbreviated as follows: CRD, cone rod dystrophies; JS, Joubert syndrome; LCA, Leber congenital amaurosis; MD, macular dystrophies; RD, Retinal degeneration; RP, Retinitis pigemntosa; US, Usher syndrome.

When analysing the MIR-gene dataset a total of 12 genes were noted to be mutated in recessive, dominant and X-linked retinitis pigmentosa cases (table 4.8). For example RPGR (retinitis pigmentosa GTPase regulator) and NRL (neural retina leucine zipper) have recruited MIR elements in the 3'-UTR. Mutations in RPGR are responsible for the majority of cases of the X-linked form of retinitis pigmentosa (RP), and there are two splice variants of RPGR with the full length transcript being composed of 19 exons (Khanna *et al.,* 2005). A second transcript referred to as RPGR-ORF15 is generated by reading through the donor splice site of exon 15 into the intronic region which contains the exptated MIR element; translation terminates at an alternative in-frame stop codon (figure 4.9). This novel region of RPGR-ORF15 is mostly comprised of a GAA triplet repeat sequence and as such is a source of glutamic acid, the function of which is undetermined.

The significance of this splice variant is demonstrated by the mutations observed in the novel coding region, of which 80% of the documented RPGR mutations are located and 60% are specifically linked to X-linked RP (Jin *et al.,* 2007; Vervoort and Wright, 2002; Vervoort *et al.*, 2000). When looking specifically at the exaptated MIR element of RPGR-ORF15 the MIR sequence demonstrated sequence identity to miRNA target sites (section 4.4), which may be a means of regulating or suppressing the expression of the isoform predominantly responsible for the retinitis pigmentosa cases; however the non-mutated form of the shorter transcript may be significant as once expressed will be a rich source of glutamate, essential in neuronal excitability.

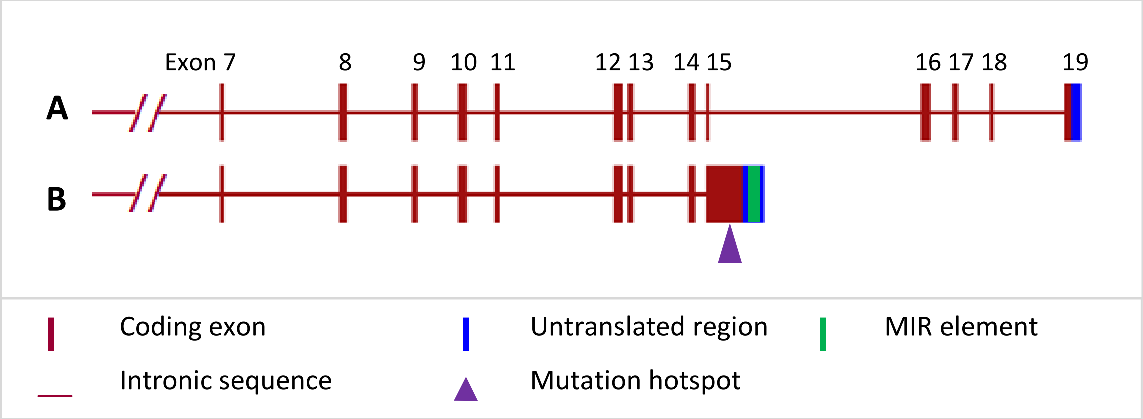


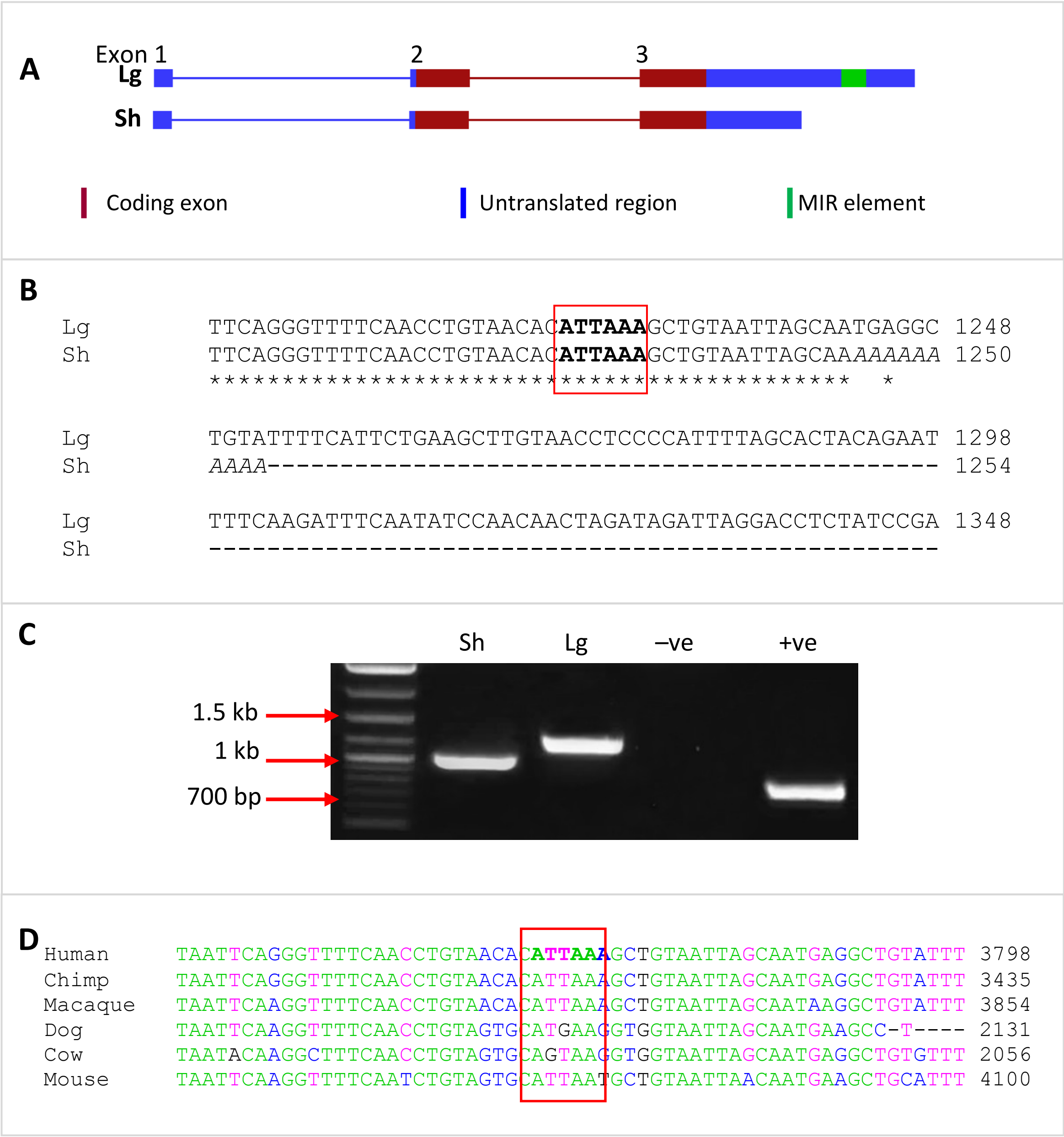
**Figure 4.9. Major isoforms of human retinitis pigmentosa GTPase regulator (RPGR)**

The exonic composition begins at constitutive exon 7. **A)** The full length RPGR mRNA transcript composed of 19 exons; **B)** Truncated transcript of RPGR due to reading-through intron 15 which contains 80% of the RPGR mutations, an MIR element is recruited in the 3'-UTR (green).

NRL is an intrinsic regulator of mammalian photoreceptor differentiation and function, and missense mutations are associated with the development and progression of recessive and dominant RP (Bessant *et al.,* 2003). The common reference nucleotide sequence of NRL is composed of three exons with an MIR element recruited in the 3'-UTR. The NRL transcript is <2 kb with a protein product of ~26 kDa. A further non-reference cDNA clone was identified in the sequence databases (accession number BC012395), which is truncated due to an alternative polyadenylation (poly(A)) site (figure 4.10a).

There is a canonical poly(A) site (ATTAAA) 20bp upstream of the cleavage site (figure 4.10b). To confirm the expression of the shorter transcript RP-PCR was performed to amplify fragments of the two transcript variants from human retina cDNA. The antisense primer of the shorter transcript incorporated a portion of the poly(A) tail so as to differentiate between the two isoforms, following which both transcripts were amplified (figure 4.10c). The poly(A) site of the shorter transcript appears to be conserved only in primates with similar but incomplete sequences in earlier mammals (figure 4.10d).

Intriguingly the protein-coding sequence will be retained in both the NRL products with the only notable difference being the addition of 720bp of nucleotide sequence in the 3'-UTR of the longer transcript (with the MIR). If these isoforms display differences in function, stability or localisation it is possible that the additional nucleotide sequence is involved in this control. A similar observation has previously been made with mammalian insulin-like growth factor 1 (IGF1), whereby a splice variant encodes an identical protein but with a larger MIR-containing 3'-UTR exon (Hughes, 2000). The larger IGF1 mRNA has been demonstrated to have a significantly higher rate of decay compared with the other IGF1 isoforms (Hepler *et al.,* 1990), and it has been suggested that the MIR element may play a part in the reduced half-life (Hughes, 2000).



```
Lg          TTCAGGGTTTTCAACCTGTAACACATTAAAGCTGTAATTAGCAATGAGGC 1248
Sh          TTCAGGGTTTTCAACCTGTAACACATTAAAGCTGTAATTAGCAAAAAAAA 1250
            ******************************************   *

Lg          TGTATTTTCATTCTGAAGCTTGTAACCTCCCCATTTTAGCACTACAGAAT 1298
Sh          AAAA---------------------------------------------- 1254

Lg          TTTCAAGATTTCAATATCCAACAACTAGATAGATTAGGACCTCTATCCGA 1348
Sh          --------------------------------------------------
```

```
Human      TAATTCAGGGTTTTCAACCTGTAACACATTAAAGCTGTAATTAGCAATGAGGCTGTATTT 3798
Chimp      TAATTCAGGGTTTTCAACCTGTAACACATTAAAGCTGTAATTAGCAATGAGGCTGTATTT 3435
Macaque    TAATTCAAGGTTTTCAACCTGTAACACATTAAAGCTGTAATTAGCAATAAGGCTGTATTT 3854
Dog        TAATTCAAGGTTTTCAACCTGTAGTGCATGAAGGTGGTAATTAGCAATGAAGCC-T---- 2131
Cow        TAATACAAGGCTTTCAACCTGTAGTGCAGTAAGGTGGTAATTAGCAATGAGGCTGTGTTT 2056
Mouse      TAATTCAAGGTTTTCAATCTGTAGTGCATTAATGCTGTAATTAACAATGAAGCTGCATTT 4100
```

**Figure 4.10. Transcript variants of human neural retina leucine zipper (NRL)**

Abbreviations: Lg, full length NRL mRNA transcript; Sh, Isoform with a shorter 3'-UTR exon due to an earlier poly(A) site (accession: Sh, BC01239; Lg, NM_006177). **A)** Exonic arrangement of NRL outlining the alternative polyadenylation sites, the MIR element (green) is located in the full length transcript only. Coding regions are in red and UTR exons in blue. **B)** Pairwise sequence alignment of the full length transcript and the shorter cDNA sequence highlighting the position of the poly(A) site (boxed red) and the poly(A) tract (italics). **C)** Gel electrophoresis image of fragments amplified using RT-PCR confirming the expression of the alternative transcripts. Visualised on a 2% agarose gel with 1X SybrSafe. DNase and RNase-free water was used as a negative control and GAPDH (738bp) as a positive control. Fragment sizes are 966bp (Sh) and 1198bp (Lg). **D)** Multiple sequence alignment the NRL cDNA sequence highlighting the conservation of the poly(A) site (boxed red) for a selection of mammals. Green sequences are completely conserved, pink identical residues, blue similar and black completely different. Sequence conservation was determined using the ClustalW2 multiple alignment tool (http://www.ebi.ac.uk/Tools/clustalw2/index.html; accession numbers appendix 9.3).

## 4.4. The role of MIR elements in the localisation of mRNA

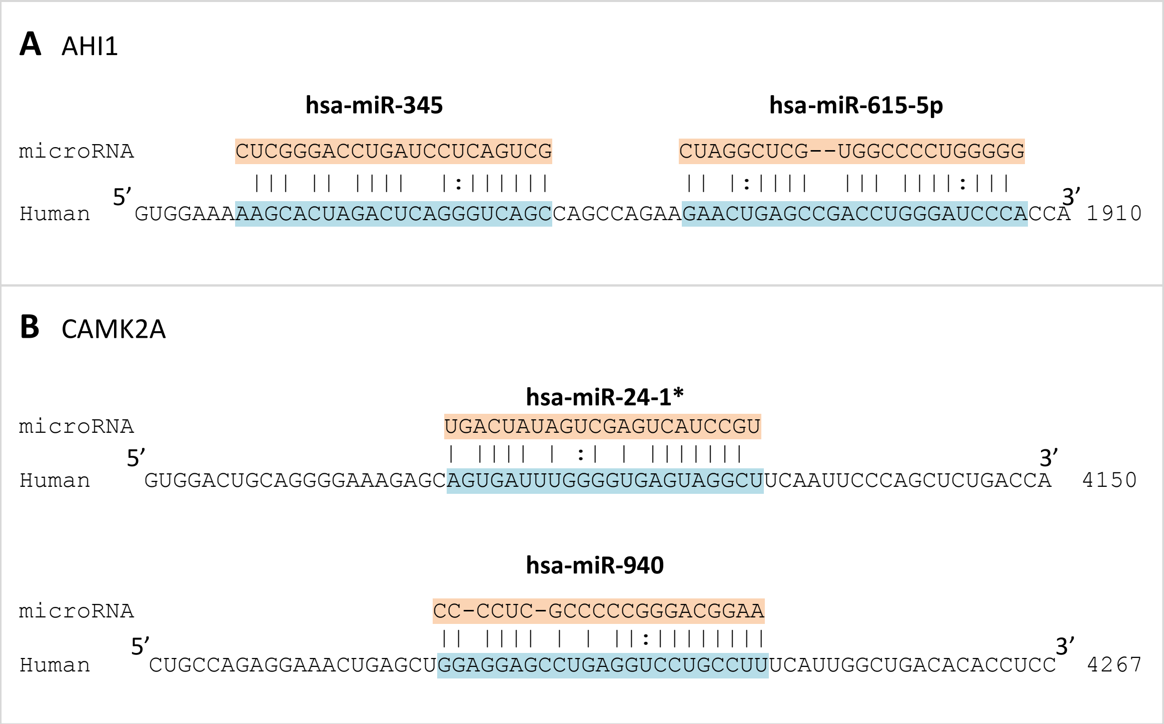
MiRNAs have been shown to block protein translation and alter mRNA stability of target genes (Bartel, 2004). The ability to suppress translation allows for mRNA translocation to specific cellular compartments and spatial expression of protein (Svoboda and Cara, 2006). It has been suggested that retrotransposons may be precursors for miRNAs (Devor *et al.,* 2009; Piriyapongsa *et al.,* 2007; Smalheiser and Torvik, 2006; Smalheiser and Torvik, 2005). Given that several of the MIR containing mRNAs are known to be dendritically localised (potentially via miRNAs) a number of genes were selected and screened for putative miRNA binding sites and precursor stem-loops. Due to the size of the dataset of genes which have recruited MIR elements (appendix 9.1) genes were selected specifically if mRNA localisation had been previously suggested or if the genes were known to be functional in polarised cells where mRNAs are frequently reported to be localised. Using this approach eight transcripts were found to contain one or more miRNA target sites located within the exaptated MIR sequence (table 4.8; figure 4.8).

Of the genes investigated in detail, AHI1, NRL, RHO and RPGR have been described earlier in this chapter, as they are known to be mutated in retinal disease (table 4.7; Valente *et al.,* 2006; Bessant *et al.,* 2003; Liu *et al.,* 2009; Vervoort *et al.,* 2000). The mRNA transcripts for CAMK2A, DDN and NEURL are localised to the dendritic compartment of hippocampal neurones and are suggested to be involved in synaptic function and plasticity (table 4.5; Kremerskothen *et al.,* 2006; Timmusk *et al.,* 2002; Mayford *et al.* 1996) and CD59 is a leukocyte antigen involved in the complement cascade and the mRNA is also noted to be dendritically localised (table 4.5).

| Gene symbol | MicroRNA Name | Type | mRNA (bp) | MIR (bp) | Details |
|---|---|---|---|---|---|
| AHI1 | hsa-miR-345<br>hsa-miR-615-5p | Target<br>Target | 1851 1873<br>1882 1905 | 1835 1947 | Photoreceptor |
| CAMK2A | hsa-miR-24-1*<br>hsa-miR-940 | Target<br>Target | 4108 4130<br>4225 4247 | 4087 4250 | Dendritically localised |
| CD59 | hsa-miR-175<br>hsa-miR-640 | Target<br>Stem-loop | 5474 5497<br>5951 6012 | 5456 5560<br>5900 6098 | Dendritically localised |
| DDN | hsa-let-7a<br>hsa-miR-466 | Target<br>Target | 3624 3645<br>3662 3684 | 3548 3730 | Dendritically localised |
| NEURL | hsa-miR-1291 | Target | 3835 3866 | 3833 3952 | Dendritically localised |
| NRL | hsa-miR-1207<br>hsa-miR-1266<br>hsa-miR-615-5p | Target<br>Target<br>Target | 1486 1510<br>1501 1518<br>1479 1504 | 1441 1603 | Photoreceptor |
| RPGR | hsa-miR-136<br>hsa-miR-101* | Target<br>Target | 4375 4396<br>4397 4420 | 4351 4435 | Photoreceptor |
| RHO | hsa-miR-20b<br>hsa- let-7a-2* | Target<br>Target | 1744 1757<br>1758 1779 | 1610 1781 | Photoreceptor |

**Table 4.9. MIR-containing genes which have sequence homology to known miRNA targets**

The miRNA name has been included and the co-ordinates of the mRNA where the homology has been detected. The position the MIR element has been recruited is included as a comparison.


A AHI1
hsa-miR-345
hsa-miR-615-5p
microRNA CUCGGGACCUGAUCCUCAGUCG CUAGGCUCG--UGGCCCCUGGGGG
Human 5' GUGGAAAAAGCACUAGACUCAGGGUCAGCCAGCCAGAAGAACUGAGCCGACCUGGGAUCCCACCA 3' 1910
B CAMK2A
hsa-miR-24-1*
microRNA UGACUAUAGUCGAGUCAUCCGU
Human 5' GUGGACUGCAGGGGAAAGAGCAGUGAUUUGGGGUGAGUAGGCUUCAAUUCCCAGCUCUGACCA 3' 4150
hsa-miR-940
microRNA CC-CCUC-GCCCCCGGGACGGAA
Human 5' CUGCCAGAGGAAACUGAGCUGGAGGAGCCUGAGGUCCUGCCUUUCAUUGGCUGACACACCUCC 3' 4267

**C** CD59

**hsa-mir-640**

```
Hairpin         GUGACCCUG-GGCAAGUUCCUGAAGAUCAGACACAUCAGAUCCCUUAUCUGUAAAAUGGGCA   61
                | ||||||  |||||||| || ||  | | |  | |||| |  || ||||| |||||| | |
CD59            GCGACCCUCAGGCAAGUUACUUAACCUUACAUGCCUCAGUUUUCUCAUCUGGAAAAUGAGAA   6012
             5'                                                              3'
```

**hsa-miR-175**

```
microRNA                      UCGAUCAGUU-CGCGCCAAC-ACCCCC
                               ||||||::| || |||||| |||||
Human        5' GCCCAAGGUAGCAUGGCUAGUUGAUGCCGGUUGAUGGGGCUUAAACCCAGCUCCCUCAUCU 3'  5521
```

**D** DDN

**hsa-leU-7a** **mmu-miR-466**

```
                UUGAUAUGUUGGAUGAUGGAGU                cAGAAUACACACGCACAUACAUa
             5' :: | ||:|| |  |||||||:                 |:| ||| |    ||||||||  3'
Human           GGGUUUAUAAACACCUACCUCGCAGGGUUGUUGUGAGGAUUUAAUGCGAUAAUGUAUGUAA   3684
```

**E** NEURL

**hsa-miR-1291**

```
microRNA                       UGACGACCAGA---A--GUCAGUCCCGGU
                               |||||   |||   |  ||| ||||||
Human        5' UUGCAUCCAUCAGAGACUGCACCUCUGUGUGGCAGGCAGGGCAUGGGUUUUAGU 3' 3874
```

**F** NRL

**hsa-miR-1207**

```
microRNA                         GGGGAGGGU---C-GGAGGGACGGU
                                  |:||:|||   | :| |||||:||
Human        5' CCUGGGUCCUAGUCCCAGCUCUUCCAUGGGAUCCCCUGUCACCCUGAGCAAAUCAGU 3'  1526
```

**hsa-miR-1266**

```
microRNA                                        UCGGGACAAGAUGUCGGGACUCC
                                                  ||||||    || |||||||
Human        5' CCUGGGUCCUAGUCCCAGCUCUUCCAUGGGAUCCCCCUGU----CA-CCCUGAGCAAAU 3'  1522
```

**hsa-miR-615-5p**

```
microRNA                  CUAGGCUCG---UGG-CCCCUGGGGG
                            ||| |||    ||  ||||:||||
Human        5' CCUGGGUCCUAGUCCCAGCUCUUCCAUGGGAUCCCCCUGUCACCCUGAGCAAAUCAGU 3'  1526
```

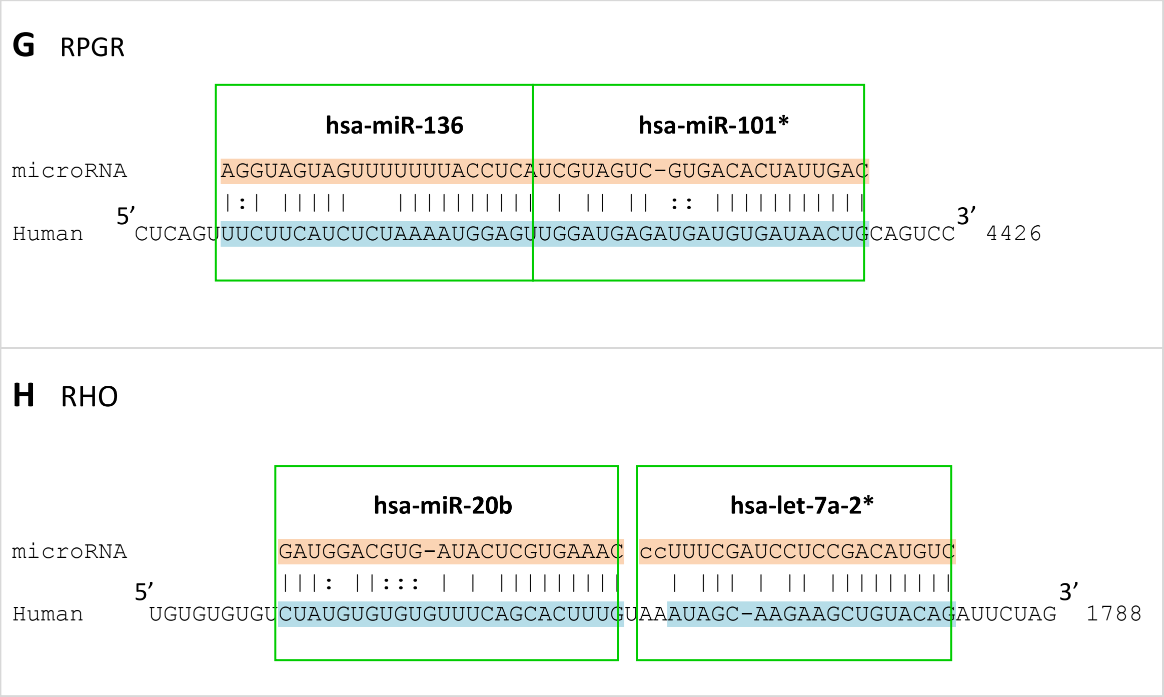

G RPGR
hsa-miR-136
hsa-miR-101*
microRNA AGGUAGUAGUUUUUUUACCUCAUCGUAGUC-GUGACACUAUUGAC
Human 5' CUCAGUUUCUUCAUCUCUAAAAUGGAGUUGGAUGAGAUGAUGUGAUAACUGCAGUCC 3' 4426
H RHO
hsa-miR-20b
hsa-let-7a-2*
microRNA GAUGGACGUG-AUACUCGUGAAAC ccUUUCGAUCCUCCGACAUGUC
Human 5' UGUGUGUGUCUAUGUGUGUGUUUCAGCACUUUGUAAAUAGC-AAGAAGCUGUACAGAUUCUAG 3' 1788

**Figure 4.11. MiRNA targets or stem-loop precursor sequences identified within recruited MIR sequences**

Potential miRNA binding sites were identified by accessing the data available in the miRBase, the miRNA registry available at the Sanger Institute, which contains >5000 miRNA gene loci (http://microrna.sanger.ac.uk/). All predicted miRNA target sites outlined (orange) reside on the sense strand and within MIR sequences which have been recruited in the 3'-UTR of the transcript. **A)** A truncated transcript of the gene Abelson helper integration site 1 (AHI1) has two miRNA target sites (accession NM_001134832). **B)** The reference sequence of the CD59 antigen, complement regulatory protein (CD59) has recruited a total of five MIR elements and has 18 different miRNA target sites and a putative precursor miRNA sequence had also been detected (accession NM_000611). **C)** Calcium/calmodulin-dependent protein kinase IIα (CAMK2A) has two predicted miRNA target sites which is present in both reference sequences (accession NM_015981). **D)** Dendrin (DDN) whose function is unknown, but is found to be localised to the dendritic compartment, has two miRNA binding sites in the reference sequence (accession NM_015086). **E)** Neuralized homolog (Drosophila) (NEURL) has a miRNA target site also in the reference sequence (accession NM_004210). **F)** Neural retina leucine zipper (NRL) contains three miRNA target sites which are not present in non-reference sequence with the shorted 3'-UTR (accession NM_006177). **G)** The truncated transcript of Retinitis pigmentosa GTPase regulator (RPGR) has two potential miRNA target sites (accession NM_001034853). **H)** Rhodopsin (RHO) has two miRNA target sites positioned within an MIRb element which has been exaptated in the 3'-UTR of the RefSeq transcript (accession NM_014850).

## 4.5. Discussion

A large area of research has focussed on the 'parasitic' role of TEs, and the deleterious and mutagenic impact on the host following insertion (Kidwell and Lisch, 2001). Ultimately many TEs behave as mutagens due to the abundance of repeat elements in all genomes; however it is also inevitable that some may become ‘domesticated’, given sufficient time (Sinzelle *et al.,* 2009; Volff, 2006; Miller *et al.*, 1999). Over the past 15 years attention has focussed primarily on the role of TEs in shaping genomes during evolution, and TEs have long been known to play a part in the evolution and function of plant genomes (Zilberman and Henikoff, 2005; Piriyapongsa and Jordon, 2008). Less information is available for a general functional role of retrotransposons in vertebrates; even though TEs would have provided a source of raw genetic material, which may have assisted in the evolution of many species (Makałowski, 2000). Most TEs are located predominantly within introns, suggesting that the elements are providing *cis*-acting regulatory sequences, which may be involved in mediating gene expression (Jurka, 2008; Sela *et al.,* 2007; Britten, 2006; Nekrutenko and Li, 2001).

Approximately 1359 human genes (5%) have recruited MIRs, although this is likely to be an underestimate, as there may be as yet undiscovered MIR-containing transcripts, particularly ncRNAs. It is not clear whether exaptation is a random event, or whether there is any selection for the exaptation of MIRs on a genome-wide scale. It is likely that for any one gene there may be a selective advantage, or at least no selective disadvantage in exaptation. Particularly when retained in coding sequence or when providing alternative splicing, which may have participated in the modification and development of novel gene function, shaping the mammalian genome and transcriptome.

The dataset of human genes which have recruited MIR elements were screened for commonality in function by collecting and examining gene ontology data, tissue expression profiles, potential miRNA targets sites or precursors and the involvement of the genes in human disorders. Throughout this process specific functional roles were consistently noted, predominantly neuronal function, vision and immune responses. For example examining gene expression profiles revealed a significant proportion of the MIR-containing genes being expression in neuronal tissues specifically brain regions.

Significant overrepresentation ($P < 0.001$) was noted for the amygdala, which is involved in fear, emotion and aggressive behaviour; the olfactory bulb which is the site of sensory perception and smell, and the occipital lobe which is the visual processing region in the mammalian brain, it is also the area where epileptic seizures originate.

The Gene Ontology (GO) was a useful resource when collecting functional information and a number of ontology terms were overrepresented for the MIR-containing genes when comparing to the rest of the genome, including the dendrite, photoreceptor, and in neurotransmitter transport. Further significant ontology terms were noted relating to immunological responses such as wound healing and cytokine receptor activity. When collating the known disorders which the MIR-containing genes are implicated or mutated, several groups included neurological conditions; specifically Alzheimer's disease, retinal degeneration, Parkinson's disease, schizophrenia and mental retardation.

### 4.5.1. MIR elements and the mammalian visual system

The involvement of the MIR-containing genes in the mammalian visual system and in retinal disease was a recurrent observation. Preliminary screening of the genes for significant keywords revealed 'vision' and 'retinitis pigmentosa' as two top terms. Furthermore when examining tissue expression data for the complete set of the genes which have recruited MIRs, three were significantly overrepresented ($P < 0.01$); the ciliary ganglion, occipital lobe (vision processing centre) and the whole eye. The ciliary ganglion is a structure situated behind the orbit of the eye and is rich in sensory neurones and regulates the constriction of the pupil when exposed to light (Thakker *et al.,* 2008).

Vision in mammals is a complex sequence of events, involving multiple cell types. The retina consists of photoreceptors, retinal ganglion cells, amacrine cells, müller cells and bipolar cells (Jeon *et al.,* 1999). Photoreceptors are a specialised neurone which function exclusively in phototransduction and perceive an image as a pattern of light. The light is converted into neuronal signals and transmitted to the brain via the retinal ganglion cells. Positioned at the distal end of the photoreceptor, away from the nucleus is the outer segment which is composed of a collection of stacked membrane disks. The outer segment contains visual pigments and other phototransduction components and is the region where the visual transduction cascade occurs (figure 4.9; Pepe, 2001). In vertebrates there are two classes of photoreceptor cells, rods and cones. Rods are the most abundant cell type in the retina and make up at least 70% of the total cells (Carter-Dawson and LaVail, 1979). Rods mediate vision during dim light and produce achromatic vision (black and white) and as such contain only one photopigment (Oh *et al.,* 2008). Photoreceptor cones are less sensitive to light and are responsible for vision during bright light. In humans and closely related primates there are three types of photopigments in cone cells, which detect different wavelengths of light, the information perceived by both types of photoreceptor is processed in the occipital lobe; a region noted where the MIR-containing genes are frequently expressed ($P < 0.001$).

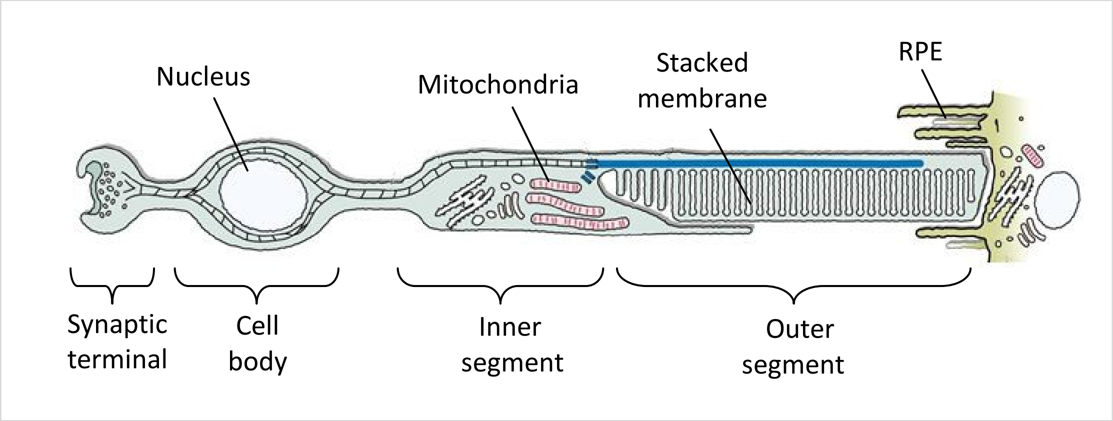


**Figure 4.12. Anatomy of a rod photoreceptor cell**

The inner segment mainly houses mitochondria and ribosomes. The outer segment consists of a tightly packed membrane discs bearing the photopigment (rhodopsin). The synaptic terminal connects to bipolar cells and the outer segment is connected to the retinal pigment epithelium (RPE) which is situated just outside the retina (Image adapted from Liu *et al.,* 2007)

The single photopigment present in rod cells is encoded by the GPCR rhodopsin (RHO), which is expressed almost exclusively in this cell type. RHO detects electromagnetic radiation when bound to a derivative of vitamin A and when activated undergoes a conformational change triggering the visual transduction cascade. When photons of light decrease RHO is phosporylated by rhodopsin kinase. Phosphorylated RHO is a substrate for arrestin (SAG), which renders it fully deactivated. Both RHO and SAG have recruited MIR elements in the 3'-UTR. Two further genes which have also exaptated MIR repeats in the 3'-UTR are the transcription factors neural-retina-leucine-zipper (NRL) and NR2E3, which is a photoreceptor-specific orphan nuclear receptor. Both NRL and NR2E3 are key components of mammalian photoreceptor differentiation and retinal development (Hendrickson *et al.,* 2008). For example NRL determines the fate of precursor photoreceptor cells, by mediating the formation of rods at the expense of cone development, partly by modulating the expression of NR2E3 by directly bind to the promoter region (Oh *et al.,* 2008). In transgenic mice lacking NRL expression, rod cells are converted to cones and NR2E3 expression is abolished (Mears *et al.,* 2001). NRL is also noted to regulate the expression of RHO by binding to the promoter region (Mitton *et al.,* 2000).

Other genes have recruited MIRs which are important in retinal function but are not mutated in retinal disease. Ciliary neurotrophic factor (CNTF), is encoded by a single transcript with an MIR element in the 3'-UTR. CNTF promotes rod photoreceptor survival, reduces RHO protein levels and upregulates SAG expression *in vivo* (Wen *et al.,* 2006). CNTF is generally involved in the survival of neurons and as such mutations are noted in a number of neurological disorders including schizophrenia, Huntington's disease and multiple sclerosis (Lin and Tsai, 2004; Emerich *et al.*, 1997; Giess *et al.,* 2002). The retina specific protein PAL (LRIT1) is expressed specifically in photoreceptors and has been shown to be significant in the development of the outer segment, as the mRNA is specifically localised to this region of photoreceptors (Gomi *et al.,* 2000). The retina-specific gene F379 (confirmed by RT-PCR and northern blot) is composed almost entirely of TEs, with an Alu and MIR element comprising the majority of the predicted coding sequence and an LTR element in the 3'-UTR (Mah *et al.,* 2001). F379 is uncharacterised and there is currently no evidence to support the translation of a protein product, either in the sequence databases or in publications. Sequence homology detected to other proteins is due to the recruitment of other Alus and/or MIR elements and it is possible that F379 functions as an ncRNA.

### 4.5.2. MIR elements and the translocation of mRNA in neuronal cells

It is clear that a number of genes which have recruited MIR elements are expressed in polarised cell types including neurones of the brain, photoreceptors and lymphocytes. The process of translocating mRNA in asymmetric polarised cells has long been observed in drosophila during oogenesis (Berleth *et al.,* 1988; St Johnston, 1995) and the mechanisms in vertebrates are being unravelled. It appears that mRNA localisation is achieved by a number of means, including diffusion to a localised anchor, active transport and via miRNAs; processes directed by *cis*-acting elements, which are usually in the 3'-UTR (Mohr and Richter, 2001; Ben Fredj *et al.,* 2004). Fine tuning protein expression and post-transcriptional control may be crucial when rapid responses and changes in protein levels are required. The full activities of c*is*-acting elements in the 3'-UTRs of vertebrates remain to be determined; however the formation of dsRNA structures and the binding of complementary ncRNAs are known components (Mohr and Richter, 2001; Le and Maizel, 2007).

Many human transcripts do not encode for protein such as tRNA, rRNA and small-nuclear RNA (snRNA) and the role of other ncRNAs in gene regulation is gradually becoming clear with many being known gene regulators. ncRNA has been shown to mediate processes such as mRNA metabolism, stability, RNA editing, imprinting and long-term memory (Claverie, 2005; Mattick and Makunin, 2006; Storz *et al.,* 2004; Royo and Cavaillé, 2008; Mercer *et al.,* 2008). A widely studied class of ncRNAs are a family of small (20-22nt) RNAs which are divided into the small-interfering RNAs (siRNA) and miRNAs (Ambros and Chen, 2007). It is also apparent that human gene expression may be regulated by miRNAs and Friedman *et al.,* (2009) predicted miRNA target sites in the 3'-UTRs of >45,000 human transcripts.

The transport of RNA to the dendritic compartment is an essential process in synaptic plasticity, and is linked to learning and memory (Sossin and DesGroseillers, 2006). As RNA transport to post-synaptic sites and translational suppression allows for local, selective protein synthesised on demand, in response to synaptic stimulation. As a consequence synapses have individual molecular identities and regulate their own morphology and efficiency (reviewed by Dahm *et al.,* 2007). Calcium/calmodulin-dependent protein kinase II alpha (CAMK2A) is known to be involved in synaptic plasticity; the mRNA of dendrin (DDN), a gene of unknown function, is also dendritically localised suggesting a role in synaptic function and plasticity (Kremerskothen *et al.,* 2006; Pinkstaff *et al.,* 2001). Both of these genes have recruited two independent MIR elements in the 3'UTR.

CAMK2 is encoded by four genes α, β, γ and δ, of which CAMK2A is the only subunit which is translocated to the dendrite for translation (Tobimatsu and Fujisawa, 1989). Mori and colleagues (2000) identified two dendritic-targeting elements (DTE) within the 3'-UTR and a third DTE was identified in a different region of the 3'-UTR by Blichenberg *et al*., (2001). CAMK2A has recruited two MIR elements one of which is located within the DTE region of CAMK2A identified by Mori *et al.,* (2000), the second MIR element was noted to contain two predicted miRNA target sites following screening against miRNA databases. DDN has a DTE situated within a 1kb region of the 3'-UTR (Kremerskothen *et al.,* 2006), an MIR element resides directly within the central region of this control element. The second MIR is situated 200bp upstream of the DTE and further investigation revealed two putative miRNA binding sites located within this MIR element. One detail to consider is that all neuronal *cis*-acting localisation elements, including DTEs and miRNAs targets are situated exclusively in the 3'-UTR, and almost 75% of the MIRs have been integrated in this untranslated gene region.

Recent studies reveal that miRNA-mediated translational control appears to be a fundamental process in synaptic plasticity. Several components of the RNAi interference pathway are found to be enriched in dendritic spines, compared to the cell body; Dicer the endonuclease which cleaves pre-miRNA and initiates the formation of the RNA-induced-silencing-complex (RISC; Bernstein *et al.,* 2001) and eIF2c a component of the RISC (Lugli *et al.,* 2005). MiRNAs can silence translation and assist in translocating mRNA to the dendritic compartment, and upon deactivation the suppression is released targeting protein expression to the synapse (Ashraf and Klunes, 2006; Schratt *et al.,* 2006; Kosik, 2006).

Ashraf and colleagues (Ashraf *et al.,* 2006) noted persistent mRNA localisation in the antennal lobe synapses of the fruit fly following olfactory-avoidance learning. The team further observed a 25-fold increase in CAMK2A expression in the brains of *dicer*-2 null mice, suggesting that RNAi pathways may participate in mRNA localisation and translational regulation. Ashraf *et al.,* (2006) further identified two putative miRNA target sites in the 3'-UTR of CAMK2A and concluded that miRNAs and the RISC pathway play a crucial role in regulating localised protein synthesis in the synapse during long-term memory formation.

### 4.5.3. MIR elements as a source of miRNA targets sites and precursor sequences

It has been previously hypothesised that MIR elements have the potential to form dsRNA due to the homology between individual MIR insertions (Hughes, 2000). These dsRNA may then be cleaved into siRNAs (Zeng and Cullen, 2005). SINEs may also be a source of miRNA precursors and the miRNAs; miR-95 and miR-151* display sequence complementary to the MIR-LINE2 elements located within the 3'-UTRs of mammalian mRNAs (Smalheiser and Torvik, 2005). Piriyapongsa *et al.,* (2007) identified 18 miRNAs derived from transposable elements, 14 of which are related to the LINE2 and MIR families. Overall it appears that the recruitment of TEs may have played a crucial role during mammalian evolution by generating miRNA target sites and possibly as miRNA precursors. Furthermore, the mechanism described by Smalheiser and Torvik appears to be a phenomenon exclusive to the expansion of mammalian genomes as no similar miRNA precursors were detected in chicken, Drosophila or *C.elegans* (Smalheiser and Torvik, 2005).

Putative miRNA target sites were identified within the exaptated MIR sequence of eight human genes; including AHI1, CD59, NEURL, RPGR and the recently discussed NRL, RHO, DDN and CAMK2A. All eight genes demonstrate localisation of mRNA to specific subcellular compartments, such as the dendritic compartment of neurones or the outer segment of rod photoreceptor cells (Zhong *et al.,* 2006; Timmusk *et al.,* 2002; Pinkstaff *et al.,* 2001). It is possible that the predicted miRNA target site located within the recruited MIR sequence may be suppressing translation during this process. CD59 is a leukocyte antigen involved in the complement pathway which is expressed in a number of tissues and CD59 deficiency is involved in the development of several disorders including Alzheimer's disease and neurodegeneration (Yang *et al.,* 2000). CD59 has also been suggested to protect neurones against complement-mediated damage (Pedersen *et al.,* 2007). Neuralized homolog (NEURL) is thought to be involved in neurogenesis and is suggested to be a tumour suppressor (Nakamura *et al.,* 1998). A further example includes the gene retinitis pigmentosa GTPase regulator (RPGR), which is mutated in X-linked retinitis pigmentosa. The truncated protein RPGR-ORF15 is known to interact with two chromosome-associated proteins, SMC1 and SMC3, and RPGR-ORF15 is involved in regulating the transport of SMC proteins to the ciliated region of the photoreceptor outer segment (Khanna *et al.,* 2005).

One gene which belongs to the SMC family, SMC1A, is encoded by a single transcript, and has recruited seven independent MIR elements in the 3'-UTR. Mutations in SMC1A have been found in patients with the X-linked disorder Cornelia de Lange syndrome, which is characterised primarily by mental retardation (Deardorff *et al.,* 2007; Musio *et al.,* 2006). The SMC1A protein can be phosphorylated by ATM (Ataxia-telangiectasia mutated) kinase, the ATM transcript has recruited two MIR repeats also in the 3'-UTR (Kim *et al.,* 2002). ATM is mutated in Ataxia telangiectasia; a neurodegenerative disease which affects the development of the cerebellum, causes immunodeficiency and a predisposition to malignant neoplasms (Mavrou *et al.,* 2008). ATM is also linked to the development of ocular telangiectasia, a form of vascular disease of the retina (Mauget-Faÿsse *et al.,* 2003).

As previously discussed miRNAs are central in fine tuning protein expression in neurones, and thus may play a role in synaptic plasticity. MiRNAs are also reported to be abundant in the cells of the retina, particularly photoreceptors. For instance the miRNAs miR-96 and miR-183 are upregulated in P347S-Rhodopsin transgenic mice, a model for retinitis pigmentosa (Loscher *et al.,* 2007; Loscher *et al.,* 2008) and photoreceptor degeneration has been observed in retinal Dicer knockout mice (Damiani *et al.,* 2008). These observations confirm the significance of miRNAs in retinal function, and given the apparent significance of MIR elements in photoreceptors, it is possible these repeats are involved in the localisation of mRNAs, to the inner and/or outer photoreceptor segments. The MIR elements may be acting as miRNA targets during photoreceptor plasticity, in a similar manner to that suggested for hippocampal neurones, leading to suppression of translation until there is a demand for protein expression upon phototransduction.

### 4.5.4. MIR elements and human disease

Several of the genes which have exaptated MIRs are known to be mutated in neurological disease and syndromic conditions, including Alzheimer's disease, schizophrenia and Joubert syndrome (JBTS). Both kinesin light chain 1 (KLC1) and tissue transglutaminase (TGM2) are implicated in Alzheimer's disease, and encode several splicoforms due to the exonisation of MIR elements (discussed further in chapters 5 and 6). KCL1 is responsible for anterograde transport of proteins such as the microtubule-associated protein *tau* (Utton *et al.,* 2005), and abnormal aggregations of *tau* are found in the neuronal cytoskeleton of Alzheimer's patients (Andersson *et al.,* 2007). *Tau* has been shown to be an excellent substrate for TGM2, and lesions observed in Alzheimer's patients have been linked to abnormal *tau* cross-linking by TGM2 (Wang *et al.,* 2008).

KLC1 is also noted to associate with Huntington-associated protein 1 (HAP1; McGuire *et al.,* 2006), which has exaptated an MIR element in the 3'-UTR. HAP1 binds to the Huntington gene (HTT) and is enriched in neurones, suggesting a link with the pathology of Huntington's disease (Wu and Zhou, 2009). HAP1 is also known to interact with the MIR-containing gene AHI1 which is mutated in Joubert syndrome, and there is a significant reduction in AHI1 expression in a HAP1 knock-out strain of mice (Sheng *et al.,* 2008). Furthermore studies of TGM2 knock-out mice cross bred with Huntington's disease mouse models (R6/2 and R6/1), demonstrate a reduction in neuronal cell death, motor dysfunction and abnormal protein aggregates, implicating TGM2 in the pathology of Huntington's disease (Mastroberardino *et al.,* 2002; Bailey and Johnson, 2005).

At least 12 MIR-containing genes are mutated in distinct retinal disorders and associated syndromes (including the previously discussed JBTS, AH1I, NR3E2, NRL, RHO, RPGR and SAG. JBTS is characterised by a malformed brain stem and the complete or partial absence of the cerebellar vermis. These characteristics result in motor and mental retardation, accompanied by numerous physiological characteristics (Parisi *et al.,* 2007). AHI1, also termed Jouberin, and NPHP1 (Nephronophthisis 1 (juvenile)) are mutated in JBTS (Tory *et al.,* 2007) and both genes have exaptated MIRs in the 3'-UTR. There are a number of splice variants of AHI1 which encode two protein products and the MIR element is providing an in-frame stop codon for the truncated AHI1 isoform (chapter 5). In addition, AHI1 is a candidate for schizophrenia susceptibility (Amann-Zalcenstein *et al.,* 2006).

In the JBTS1 (Joubert syndrome 1) chromosomal region, 9q34, resides the MIR-containing gene EHMT1 (Euchromatic histone-lysine N-methyltransferase 1). This gene has not yet been associated with JBTS cases however EHMT1 is found to be involved in subtelomeric 9q34 deletion syndrome (Kleefstra *et al.,* 2006), a region associated with both mental retardation and JBTS. Furthermore within the JBTS locus 11p12-q13.3 defined by Valente *et al.,* (2005) there are 24 genes which have recruited MIR elements (appendix 9.5).

## 4.6. Conclusion

Overall it appears that MIR elements may indeed be important multifunctional components in mammalian biological processes. Given the sequence conservation and abundance of the MIR repeats in the human genome, it is possible that these elements may have been co-opted during evolution in the development and enhancement of mammalian neuronal function, specifically in synaptic plasticity and photoreceptor activity. When taking together the functional enrichment data it becomes apparent that the MIR-containing genes are functionally significant in neuronal cells and possibly the adaptive immune system. Furthermore MIR elements may be playing a role in translocating mRNA and regulating localised protein synthesis by the formation of dsRNAs, miRNA precursors or miRNAs binding sites.

# 5. MIR ELEMENTS MAY CONTRIBUTE TO THE TRANSCRIPTOME VIA EXONISATION

## 5.1. Introduction

The majority of the human genome is comprised of ncDNA (98.5%), of which at least 45% is made up of TEs (Lander *et al.,* 2001). It has been estimated that 60% of TEs are located in introns in both human and mouse genomes, yet intronic sequences comprise <25% of the total human gDNA (Sela *et al.,* 2007). This abundance of TEs in introns compared to intergenic DNA suggests preferential insertion and/or retention within gene regions. Most intronic TEs will not disrupt gene expression and as a consequence are less likely to be subject to natural selection than exonic TEs; however this would also be the case for intergenic TEs. A collection of intronic TEs have been demonstrated to play a role in gene expression, for instance the insertion of *Mu* DNA transposons in maize intronic sequences have long been known to play a part in gene regulation and phenotypic variation (Greene *et al.,* 1994; Luehrsen and Walbot, 1990; Bennetzen *et al.,* 1984). In mammals intronic SINEs have also been reported to generate cryptic splice sites resulting in alternative splicing (Sela *et al.,* 2007; Krull *et al.,* 2005; Sorek *et al.,* 2002).

TEs which have been recruited within intronic and exonic regions may provide spice sites and splice enhancers/silencers, thereby providing an assortment of gene products. Exonic TEs may contribute polyadenylation signals and peptide sequences including methionines and translational termination codons. Exonic TEs are unlikely to be retained and most will be selected against, as newly integrated TEs are potentially insertional mutagens, disrupting reading frames (Sela *et al.,* 2007); therefore most TEs within coding regions are most likely a direct result of the exonisation of intronic elements. Novel exons arise through three mechanisms; gene duplication, tandem exon duplication and exonisation (Kondrashov and Koonin 2001; Corvelo and Eyras 2008). The latter process occurs when TE-derived intronic and/or genomic sequences are recognised by the spliceosome and are exonised, allowing for a gene to adopt additional mRNA sequences.

A number of studies investigate the exonisation potential of intronic TEs in primate and rodent genomes, including Alu repeats, MIR elements and LTRs (Piriyapongsa *et al.,* 2007; Sela *et al.,* 2007; Krull *et al.,* 2007; Sorek *et al.,* 2002; Nekrutenko and Li, 2001). The most widely studied class of transposons are the Alu elements which appear to play a prominent role in the creation of *de novo* exons in primate species (Krull *et al.,* 2005; Lev-Maor *et al.,* 2003; Sorek *et al.,* 2002; Makałowski 2000). Alu repeats are frequently exonised when in the inverse orientation as the consensus sequences include motifs which resemble acceptor and donor splice sites. Mutations or RNA editing of the Alu sequence may allow for the elements to be actively spliced and exonised (Lev-Maor *et al.,* 2003; Sorek *et al.,* 2002; Singer *et al.,* 2004). 95% of A-to-I editing is reported to occur within Alu sequences located within ncDNA, following the formation of dsRNAs (Levanon *et al.,* 2004).

Recent research has focussed on the exaptation and exonisation of MIR elements; Mersch *et al.,* (2007) identified the exonisation of 86 MIR elements in the human genome; Sela *et al.,* 2007 identified 181 MIRs exonisation events within the UTRs and CDS of human genes (78 within the CDS) and Krull *et al.,* (2007) identified 107 MIR exonisation events. As indicated in chapter 3.1, this study has identified a greater number of MIR elements located in protein-coding regions (176), however not all of these MIR elements are spliced, for instance many are providing initiating methionines and stop codons without splicing. Furthermore it is possible that MIR elements may not only contribute to splicing in the CDS but exonised MIR elements may also be present in the UTRs. This chapter will detail the investigation of the potential for MIRs to provide functional splice sites, transcriptional control codons and the role of MIRs in shaping the transcriptome during evolution will be discussed.

## 5.2. Exonised MIR elements in the human genome

The integration of an MIR element into a gene region may have one of several consequences to the mRNA produced (figure 5.1). In the majority of integration events the presence of the MIR will be of no consequence to the corresponding gene. If an MIR element is retained within an intronic or coding sequence it may, following mutation, provide a functional splice site. Intronic MIR elements may be exonised generating whole cassette exons, which may be alternatively or constitutively expressed, exon skipping may also occur. If the MIR is retained or exonised within the coding and UTR exons it is possible over time that they may acquire transcriptional and/or translational signals such as initiating methionine codons, splice sites and polyadenylation signals (see chapter 3.2).

The exonic arrangement of all mRNA transcripts which have recruited MIR elements was studied, and the positions of the repeat sequences determined. Those MIRs which contribute splice sites were noted, including CDS and UTR exons (table 5.1). The exonised MIR sequences were collected and the corresponding pre-mRNA splice site sequences recorded (appendix 9.8). The presence of initiating methionines and stop codons was noted and the orientation of the exonised MIR elements was determined. MIR element which have transposed in the middle of an exon, with no splice sites being provided were also included (table 5.1).

Of the total MIR-containing genes none contained spliced MIR elements in the 3'-UTR; however there are an equal number of MIR-mediated splicing events noted for CDS and 5'-UTR exons. The majority of the splice sites produce alternative splice variants (62%), however 25 acceptor and 23 donor constitutive splice site sequences are provided by a recruited MIR element. Likewise most of the exons containing the methionine and stop codons are alternatively expressed; 62% and 74% respectively. A total of 65 termination codons were identified in exons derived from MIRs of which 20 are alternative stop codons. There are 26 examples of genes where whole cassette exons are spliced due to both the acceptor and donor splice sites being generated by a single exaptated MIR element (table 5.2). A total of 73 EST sequences, assigned to known genes, were noted to contain spliced MIRs. There are no corresponding full length cDNA clones detected to support the expression of the EST sequences and as such these examples were not included in the analysis (appendix 9.9).

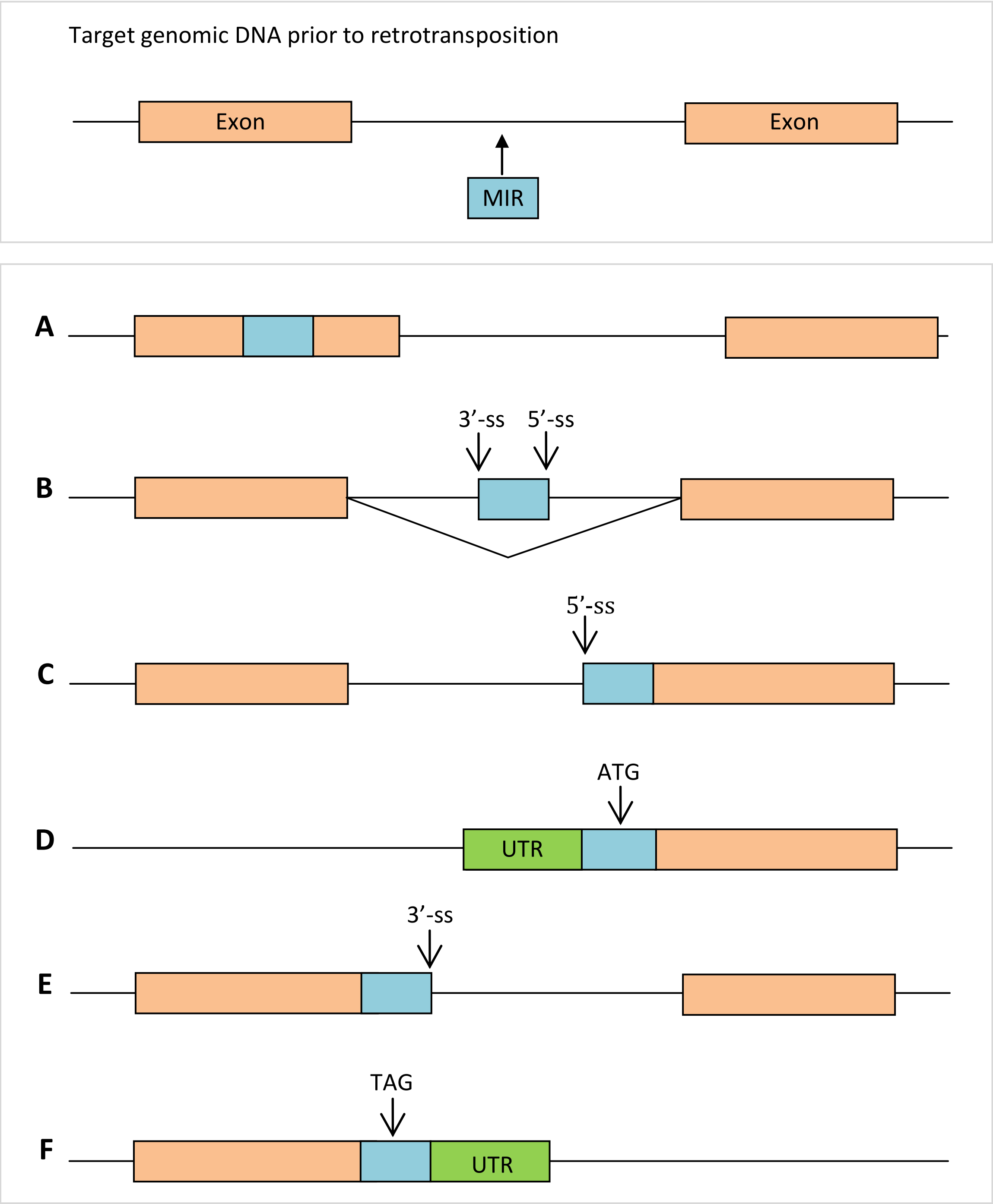


**Figure 5.1. Modes of exonisation following the integration of an MIR element within the coding sequence**

Coding exons are depicted by orange boxes, the MIR element is in blue and untranslated regions are green. Acceptor (5') and donor (3') splice sites, initiating methionines (ATG) and stop codons (TAG) have been included. **A)** Integration of an MIR in the middle of a coding exon can provide additional protein coding sequence and alter the reading frame. **B)** Insertion into an intron can generate an internal cassette exon which may be alternatively spliced or constitutively expressed. **C)** MIR can contribute an alternative 5'-splice site (ss). **D)** The MIR element can provide a functional initiating methionine residue. **E)** The MIR contributes an alternative 5'-ss. **F)** Reading through of 3'-ss into the consecutive intron and the MIR carries an in-frame stop codon, resulting in a truncated transcript.

**A**

| MIR Family | Entire CDS | | CDS Acc SS | | CDS Only | | CDS Don SS | | 5’-UTR Acc SS | | 5’-UTR Don SS | | 3’-UTR Acc SS | | 3’-UTR Don SS | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | + | - | + | - | + | - | + | - | + | - | + | - | + | - | + | - |
| MIR | 3 | 6 | 2 | 14 | 1 | 4 | 2 | 6 | 5 | 8 | 7 | 4 | 0 | 0 | 0 | 0 |
| MIRb | 2 | 9 | 3 | 8 | 4 | 0 | 5 | 6 | 2 | 5 | 6 | 8 | 0 | 0 | 0 | 0 |
| MIRc | 0 | 1 | 0 | 2 | 0 | 0 | 1 | 1 | 0 | 2 | 0 | 1 | 0 | 0 | 0 | 0 |
| MIRm | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 2 | 0 | 1 | 0 | 0 | 0 | 0 |
| MIR_mars | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| MIR3 | 2 | 2 | 1 | 5 | 1 | 1 | 3 | 2 | 2 | 4 | 2 | 3 | 0 | 0 | 0 | 0 |
| THER1_MD | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| TOTAL | 7 | 19 | 6 | 30 | 8 | 5 | 12 | 15 | 9 | 21 | 15 | 17 | 0 | 0 | 0 | 0 |
| | 26 | | 36 | | 13 | | 27 | | 30 | | 32 | | 0 | | 0 | |

**B**

| MIR Family | Alt Acc Ss | | Alt Don Ss | | Con Acc SS | | Con Don SS | | Alt ATG | | Con ATG | | Alt TAG | | Con TAG | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | + | - | + | - | + | - | + | - | + | - | + | - | + | - | + | - |
| MIR | 6 | 12 | 6 | 5 | 1 | 10 | 3 | 5 | 1 | 3 | 1 | 1 | 8 | 10 | 4 | 3 |
| MIRb | 4 | 8 | 8 | 9 | 1 | 5 | 3 | 5 | 0 | 2 | 1 | 3 | 11 | 4 | 1 | 6 |
| MIRc | 1 | 3 | 0 | 2 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 |
| MIRm | 0 | 0 | 0 | 0 | 0 | 3 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| MIR_mars | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| MIR3 | 2 | 6 | 4 | 2 | 1 | 3 | 1 | 3 | 3 | 2 | 0 | 2 | 5 | 6 | 1 | 0 |
| THER1_MD | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| TOTAL | 13 | 29 | 18 | 18 | 3 | 22 | 9 | 14 | 6 | 7 | 2 | 6 | 27 | 21 | 7 | 10 |
| | 42 | | 36 | | 25 | | 23 | | 13 | | 8 | | 48 | | 17 | |

**Table 5.1. Human genes which have exonised MIR elements within the coding sequence and untranslated regions**

The MIR elements have been grouped into family members, whether they are in the sense (+) or antisense (-) orientation, and the number of MIRs involved in each exonisation event recorded. **A)** Acceptor (Acc) and donor (Don) splice sites (SS) have been noted for coding and UTR exons. **B)** Whether the splice sites, methionines (ATG) and stop codons (TAG) are alternative (Alt) or constitutive (Con) has been determined.

| Symbol | Gene name | MIR type | | Gene Pos. | Con Pos. | Splice event | Type of exonisation | Sequence information |
|---|---|---|---|---|---|---|---|---|
| ABHD2 | alpha/beta hydrolase domain containing protein | MIR | + | 5' | Int 1 | Alt | Alternative 5'-UTR | tttcctcacctctaaaacagAAAAGGgtaagc |
| ANKRD9 | ankyrin repeat domain 9 | MIR3 | - | 5' | Ex 2 | Con | Constitutive 5'-UTR | acccctcccctttcttacagATACAGgtaggt |
| C11orf51 | chromosome 11 orf 51 | MIRb | - | 5' | Ex 2 | Con | Constitutive 5'-UTR | ttattatccccattttacagATGCAGgtctgt |
| C7orf65 | chromosome 7 orf 65 | MIRb | - | CDS | Ex 2 | Con | Constitutive CDS | tgcctgctcttattgtgcagGTAGAGgtaata |
| CHRD | chordin | MIR | - | TAG | Int 2 | Alt | Alternative CDS with TAG | attattatccccattttccaGACAGGgtaagtcttg |
| CHRDL2 | chordin-like 2 | MIR | - | CDS | Ex 10 | Con | Constitutive CDS | ttattatccccattttacagATGCCAgtaagt |
| CPA5 | carboxypeptidase A5 | MIRc | - | 5' | Int 1 | Alt | Alternative 5'-UTR | taatacattgcattttgcagGCTGAGgtaggg |
| CSF3R | colony stimulating factor 3 receptor (granulocyte) | MIR3 | - | 5' | Ex 2 | Con | Constitutive 5'-UTR | atgttatttctttcccacagATGCTGgtaagt |
| DMWD | dystrophia myotonica-containing WD repeat motif | MIR | - | CDS | Ex 4 | Con | Constitutive CDS | agacccctctgtctccgtagTTCCAAgtcagt |
| EIF4G3 | eukaryotic translation initiation factor 4 gamma, 3 | MIRb | - | 5' | Int 1 | Alt | Alternative 5'-UTR | attttgctaattctttccagCAAAAGgtaatc |
| FXYD4 | FXYD domain containing ion transport regulator 4 | MIR | - | 5' | Ex 2 | Con | Constitutive 5'-UTR | tgatgatcctccttttacagGACTTGgtaagt |
| GDPD5 | glycerophosphodiester phosphodiesterase domain | MIR | + | 5' | Int 2 | Alt | Alternative 5'-UTR | tgcaaatccttccttcctagGGCCAGgtttgt |
| GIPC1 | GIPC PDZ domain containing family, member 1 | MIR | - | 5' | Int 1 | Alt | Alternative 5'-UTR | gaattccacccatttttcagATGCTGgtaagt |
| GSG1L | germ cell associated 1-like | MIR | - | CDS | Int 4 | Alt | Alternative CDS | ccaaagtgatgggattacagGTGCTGgtaagt |

| Symbol | Gene name | MIR type | | Gene Pos. | Con Pos. | Splice event | Type of exonisation | Sequence information |
|---|---|---|---|---|---|---|---|---|
| IL6ST | interleukin 6 signal transducer (gp130, oncostatin M receptor) | MIRb | + | 5’ | Ex 2 | Con | Constitutive 5’-UTR | agactctcctctctttacagTGA TGGgtaagt |
| KLC1 | kinesin light chain 1 | MIRb | - | TAG | Int 13 | Alt | Alternative CDS with TAG | gtcctccccacattttaaagATG (AAC***TGA***CTT)ATGgtgagt |
| LAS1L | LAS1-like (S. cerevisiae) | MIR3 | - | CDS | Ex 9 | Con | Constitutive CDS | ccaacatcctcatgttacagATG ACAgtgagt |
| MORF4L2 | mortality factor 4 like 2 | MIR | + | 5' | Int 2 | Alt | Alternative 5’-UTR | taatagttcctacttcatagGAT CATgtaagt |
| MUC15 | mucin 15 | MIRb | + | 5' | Int 1 | Alt | Alternative 5’-UTR | ttgcctcttccattttccagCTT GTTgtaagt |
| NEK11 | NIMA (never in mitosis gene a)- related kinase 11 | MIR3 | + | 5’ | Int 1 | Alt | Alternative 5’-UTR | attctgtccttctctaaaagGAA TCTgtaagt |
| NUMB | numb homolog isoform 1 | MIRb | - | 5’ | Ex 3 | Con | Constitutive 5’-UTR | ttattattctcatttacagATG CAGgtctgt |
| PELI3 | pellino homolog 3 | MIRb | - | CDS | Int 2 | Alt | Alternative CDS | tcatttcaatcttcacaaagATG CTGgtaagt |
| RHBDD3 | rhomboid domain containing 3 | MIRm | - | 5' | Ex 2 | Con | Constitutive 5’-UTR | gatatccttggtttctacagAAG AAGgtaggc |
| SGIP1 | SH3-domain GRB2-like (endophilin) interacting protein 1 | MIRb | - | CDS | Int 15 | Alt | Alternative CDS | ttatctttccctttttacagATG CAGgtctgt |
| TP53I11 | p53-inducible protein 11 | MIRb | - | 5' | Int 1 | Alt | Alternative 5’-UTR | ttttgccccgggctccctagGAA GAGgtaagg |
| UBE2V1 | ubiquitin-conjugating enzyme E2 variant 1 | MIRb | - | ATG | Int 1 | Alt | Alternative ATG with CDS | ggaaagcattttatctccacAGC (AAG***ATG***GCA)AAGgtgagt |

**Table 5.2. Cassette exons with both the 3’ and 5’ splice sites provided by an MIR sequence**

MIR-mediated exonisation generating constitutive (Con) and alternative (Alt) cassette exons. The MIR sub-type has been listed along with the orientation (+, sense; -, antisense). The gene symbol and full description has been provided and the constitutive position of the exon is noted as either exonic (Ex) or intronic (Int). The splice site sequence is included, exonic sequences are in capitals and intronic sequences lower case.

## 5.3. Whole exons generated by MIR elements

In the human genome twenty six exons have been generated by slicing both in and out of a recruited MIR element. Eleven internal exons are constitutively spliced, of which four are protein-coding exons and the remaining seven within 5'-UTRs. Of the alternatively spliced exons, six are protein coding exons and nine are in 5'-UTR exons (table 5.2). The genes Kinesin light chain 1 (KLC1) and Chordin (CHRD) were selected for further sequence analysis, as these are the only examples where the MIR-derived exon carries an alternative translational termination codon.

Multiple alternate splicing event have been described for KLC1 resulting in nineteen mRNA transcript variants, encoding four different proteins of related function (McCart *et al.,* 2003). Five of the nineteen mRNA transcripts encode a truncated protein, due to splicing of the MIR-derived exon which carries an in-frame stop codon (figure 5.2). The full length CHRD protein contains four cys-rich repeats (CR1-4) which bind to bone morphogenetic proteins (BMP), and binding to any one of the four CRs is shown to significantly inhibit BMP signalling (Larrain *et al.,* 2000). Millet *et al.,* (2001) studied the expression of six CHRD splice variants, two of which contain the MIR-derived exon identified in table 5.2. Millet *et al.,* (2001) noted that the MIR-containing transcript lacks the CR2-4 binding regions and contains only a partion of CR1, they further demonstrated that this transcript was not a BMP antagonist. The alternative CHRD mRNA transcript contains at least six downstream 3'-UTR exons and as such is a candidate for nonsense-mediated decay (NMD), as NMD is a cellular surveillance mechanism which recognises and degrades mRNA with premature termination codons (Chang *et al.,* 2007). The splicing of the multiple 3'-UTR exons of CHRD is supported by two cDNA clones and four ESTs in the sequence databases, with none containing a poly(A) tract (accessions; AF209930, AF209929, DA092462, DA096654, DA336666, DA106058). Millet *et al.,* (2001) successfully amplified the MIR-containing CHRD splicoforms by northern blot analysis and RT-PCR in foetal liver, but did not study the expression of the truncated protein. Pairwise sequence alignments were generated of the alternative MIR-derived exons of KLC1 and CHRD, including the flanking intronic sequence and the MIR consensus sequence (figure 5.2). This revealed that in both the KLC1 and CHRD exons the stop codon and the acceptor and donor splice sites are at the same position in the MIR core-SINE region (see section 5.4).

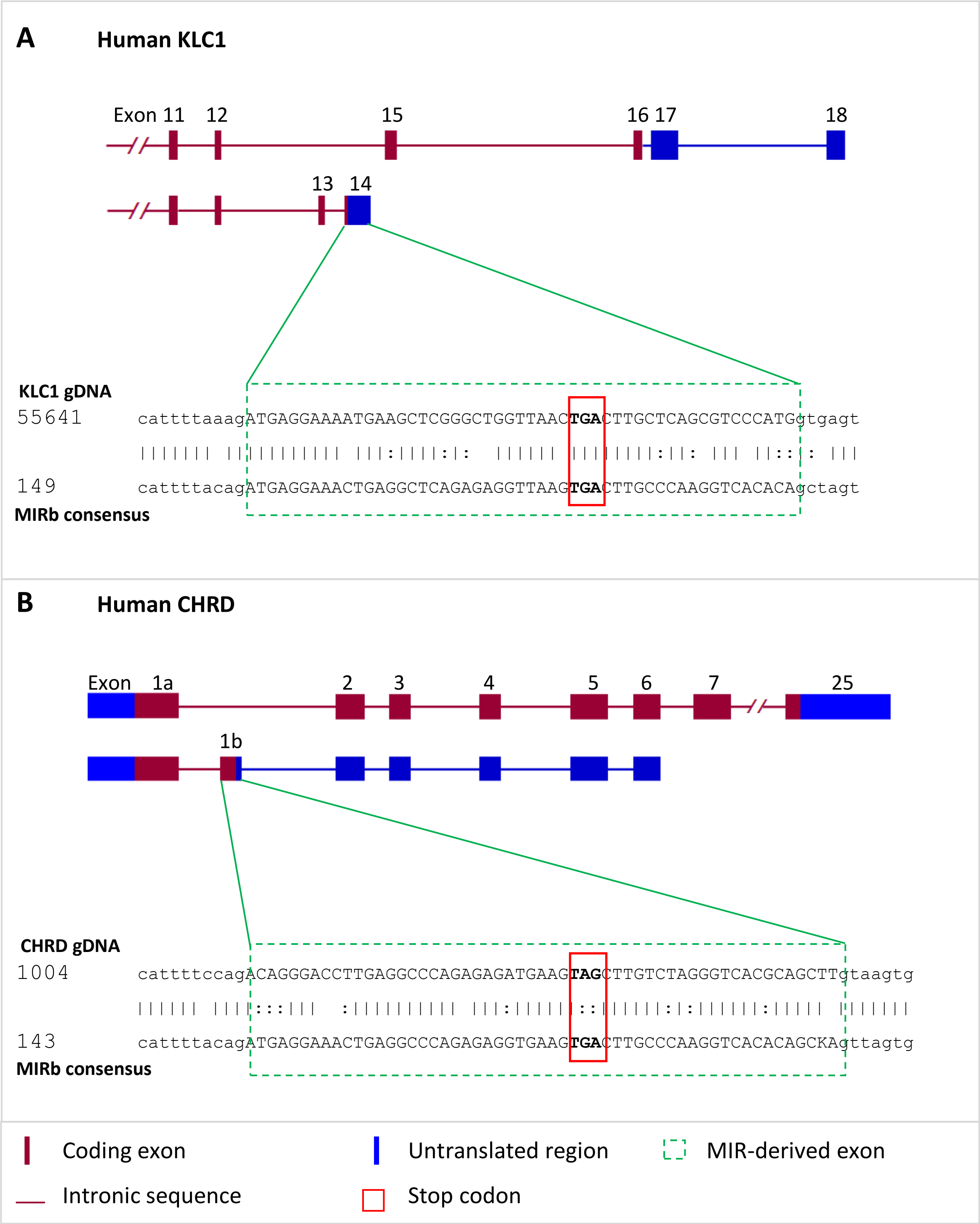
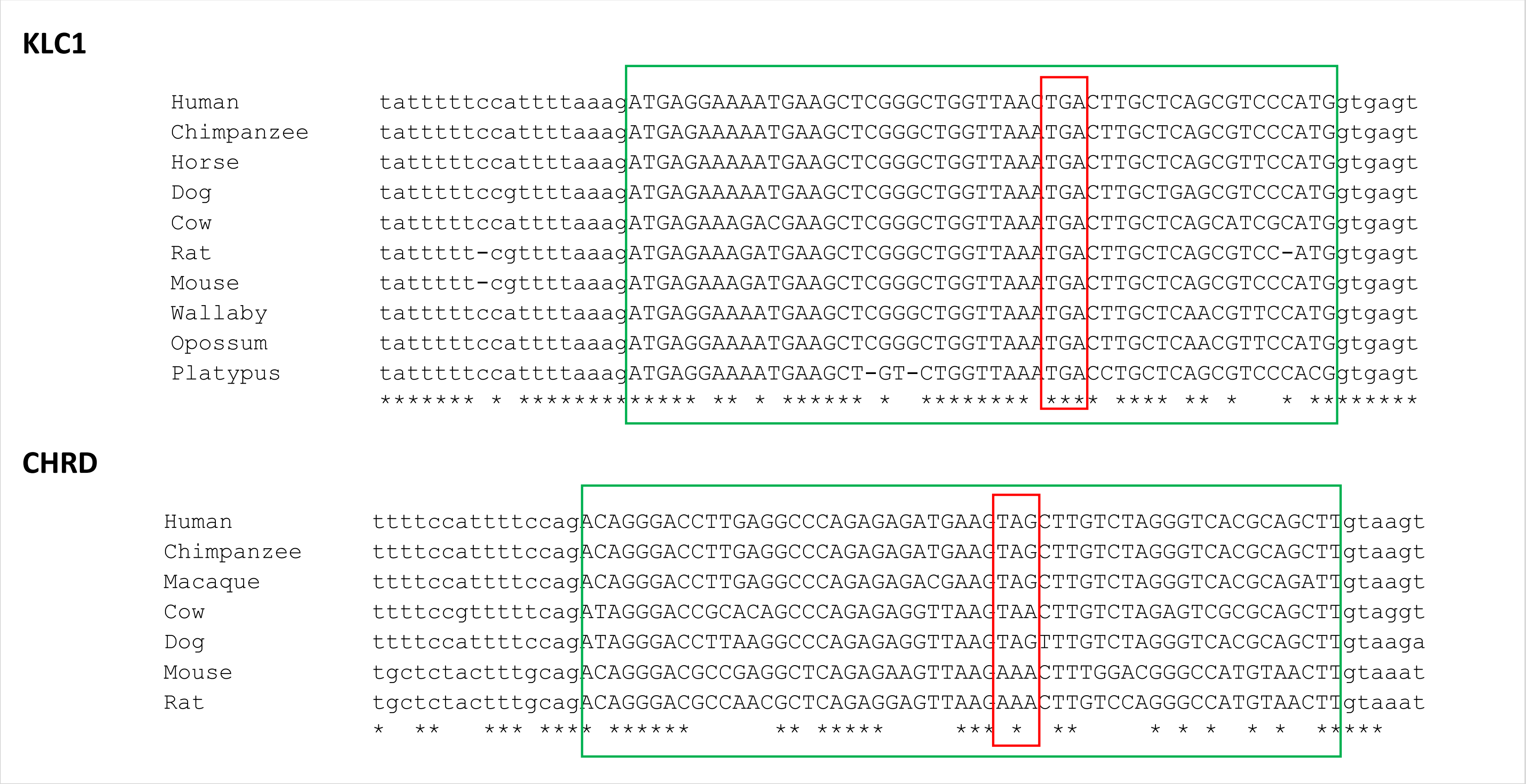

A
Human KLC1
Exon 11
12
15
16 17
18
13
14
KLC1 gDNA
55641
cattttaaagATGAGGAAAATGAAGCTCGGGCTGGTTAACTGACTTGCTCAGCGTCCCATGgtgagt
149
cattttacagATGAGGAAACTGAGGCTCAGAGAGGTTAAGTGACTTGCCCAAGGTCACACAgctagt
MIRb consensus
B
Human CHRD
Exon
1a
2
3
4
5
6
7
25
1b
CHRD gDNA
1004
cattttccagACAGGGACCTTGAGGCCCAGAGAGATGAAGTAGCTTGTCTAGGGTCACGCAGCTTgtaagtg
143
cattttacagATGAGGAAACTGAGGCCCAGAGAGGTGAAGTGACTTGCCCAAGGTCACACAGCKAgttagtg
MIRb consensus
Coding exon
Untranslated region
MIR-derived exon
Intronic sequence
Stop codon

**Figure 5.2. Pairwise sequence alignments of whole cassette exons with alternative stop codons generated by the exonisation of an MIR element**

**A**) Alternative exon 14 of KLC1 is MIR-derived and will encode a truncated protein; the MIR sequence provides the translation termination codon (accession NM_005552.4). **B)** CHRD exon 1b is derived from a MIRb element and contains numerous 3'-UTR exons. Note that the stop codon and the acceptor and donor splice sites are at the same position for both MIR sequences (accession AF209930.1). Sequence conservation was determined using the ClustalW2 alignment tool (http://www.ebi.ac.uk/Tools/clustalw2/index.html).



KLC1

```
Human          tattttttccattttaaagATGAGGAAAATGAAGCTCGGGCTGGTTAACTGACTTGCTCAGCGTCCCATGgtgagt
Chimpanzee     tattttttccattttaaagATGAGAAAAATGAAGCTCGGGCTGGTTAAATGACTTGCTCAGCGTCCCATGgtgagt
Horse          tattttttccattttaaagATGAGAAAAATGAAGCTCGGGCTGGTTAAATGACTTGCTCAGCGTTCCATGgtgagt
Dog            tattttttccgttttaaagATGAGAAAAATGAAGCTCGGGCTGGTTAAATGACTTGCTGAGCGTCCCATGgtgagt
Cow            tattttttccattttaaagATGAGAAAGACGAAGCTCGGGCTGGTTAAATGACTTGCTCAGCATCGCATGgtgagt
Rat            tattttt-cgttttaaagATGAGAAAGATGAAGCTCGGGCTGGTTAAATGACTTGCTCAGCGTCC-ATGgtgagt
Mouse          tattttt-cgttttaaagATGAGAAAGATGAAGCTCGGGCTGGTTAAATGACTTGCTCAGCGTCCCATGgtgagt
Wallaby        tattttttccattttaaagATGAGGAAAATGAAGCTCGGGCTGGTTAAATGACTTGCTCAACGTTCCATGgtgagt
Opossum        tattttttccattttaaagATGAGGAAAATGAAGCTCGGGCTGGTTAAATGACTTGCTCAACGTTCCATGgtgagt
Platypus       tattttttccattttaaagATGAGGAAAATGAAGCT-GT-CTGGTTAAATGACCTGCTCAGCGTCCCACGgtgagt
               ******* * ************* ** * ****** *  ******** **** **** ** *   * ********
```

CHRD

```
Human          ttttccattttccagACAGGGACCTTGAGGCCCAGAGAGATGAAGTAGCTTGTCTAGGGTCACGCAGCTTgtaagt
Chimpanzee     ttttccattttccagACAGGGACCTTGAGGCCCAGAGAGATGAAGTAGCTTGTCTAGGGTCACGCAGCTTgtaagt
Macaque        ttttccattttccagACAGGGACCTTGAGGCCCAGAGAGACGAAGTAGCTTGTCTAGGGTCACGCAGATTgtaagt
Cow            ttttccgtttttcagATAGGGACCGCACAGCCCAGAGAGGTTAAGTAACTTGTCTAGAGTCGCGCAGCTTgtaggt
Dog            ttttccattttccagATAGGGACCTTAAGGCCCAGAGAGGTTAAGTAGTTTGTCTAGGGTCACGCAGCTTgtaaga
Mouse          tgctctactttgcagACAGGGACGCCGAGGCTCAGAGAAGTTAAGAAACTTTGGACGGGCCATGTAACTTgtaaat
Rat            tgctctactttgcagACAGGGACGCCAACGCTCAGAGGAGTTAAGAAACTTGTCCAGGGCCATGTAACTTgtaaat
               *  **   *** **** ******      ** *****     *** *  **     * * *  * *  *****
```

**Figure 5.3. Multiple sequence alignment of the MIR elements for KLC1 and CHRD for a number of mammalian species**

Genomic DNA containing the exaptated MIR elements has been aligned from a number of species. The exonic region is boxed green and the stop codon boxed red. Intronic regions are in lower case and exonic sequence is capitalised. Sequence conservation was determined using the ClustalW2 alignment tool (http://www.ebi.ac.uk/Tools/clustalw2/index.html; accession numbers appendix 9.3).

| KLC1 | Acceptor SS | Score | Donor SS | Score |
|---|---|---|---|---|
| Human | tccattttaaagA | 4.2 | ATGgtgagt | 10.4 |
| Macaque | tccattttaaagA | 4.2 | ATGgtgagt | 10.4 |
| Chimpanzee | tccattttaaagA | 4.2 | ATGgtgagt | 10.4 |
| Cow | tccattttaaagA | 4.2 | ATGgtgagt | 10.4 |
| Dog | tccgttttaaagA | 5.0 | ATGgtgagt | 10.4 |
| Mouse | ttcgttttaaagA | 5.4 | ATGgtgagt | 10.4 |
| Rat | ttcgttttaaagA | 5.4 | ATGgtgagt | 10.4 |
| Wallaby | tccattttaaagA | 4.2 | ATGgtgagt | 10.4 |
| Opossum | tccattttaaagA | 4.2 | ATGgtgagt | 10.4 |
| Platypus | tccattttaaagA | 4.2 | ATGgtgagt | 10.4 |
| **CHRD** | **Acceptor SS** | **Score** | **Donor SS** | **Score** |
| Human | ttttccagacagG | 8.3 | CTTgtaagt | 6.9 |
| Macaque | ttttccagacagG | 8.3 | CTTgtaagt | 6.9 |
| Chimpanzee | ttttccagacagG | 8.3 | ATTgtaagt | 6.9 |
| Cow | tttttcagatagG | 6.6 | CTTgtaggt | 3.9 |
| Dog | ttttccagatagG | 6.6 | CTTgtaaga | 5.6 |
| Mouse | ctttgcagacagG | 6.2 | CTTgtaaat | 3.5 |
| Rat | ctttgcagacagG | 6.2 | CTTgtaaat | 3.5 |

**Table 5.3. Acceptor and donor splice site strength for the MIR-derived exon of KLC1 and CHRD for a number of mammals**

The splice site strengths were scored using GENIE. Intronic regions are in lowercase and exonic sequences capitalised.

Multiple sequence alignments were generated with the gDNA of KLC1 and CHRD from a selection of species. Conservation of the MIR-derived exon was determined by aligning the flanking constitutive exons with the intronic sequence to determine if the MIR element and splice sites are conserved in other mammals (figure 5.3). Exon 14 of human KLC1 appears to be highly conserved in mammals including marsupials and monotremes. The donor splice site of KLC1 was conserved in all species and scored highly at 10.4 using the GENIE program (Reese *et al.,* 1997). The acceptor splice site however scored lower, with a mean average of 4.5 suggesting minor splicing of this isoform.

The CHRD isoform with multiple 3'-UTR exons appears to be conserved in other mammalian species, there is a corresponding dog cDNA sequence in the GenBank database (accession number XM_854602) and sequence homology of the alternative MIR-derived for other primates and rodents (figure 5.3). The splice donor site is completely conserved between primates and scores highly with GENIE (mean 6.4) in other mammals. However the conservation observed may be a consequence of the exonic sequence and splice sites being derived from the core-SINE, which by default is highly conserved.

## 5.4. Splice sites within MIR elements

Analysis of the MIR-derived alternative exons of KLC1 and CHRD identified a specific region of the MIR element providing the acceptor and donor splice sites (figure 5.2). Subsequently all of the splice sites (including the exon/intron boundary) residing within MIR elements were aligned, and the co-ordinates in relation to the MIR consensus sequence were recorded to determine if the same MIR region was being exonised. Following which an mRNA consensus sequence for the acceptor and donor splice sites was determined; ttacagATG for the acceptor splice site and CAGgtaagt for the donor splice site (figure 5.4). The acceptor and donor consensus sequences are highly conserved in the inverse orientation of the MIR sequence. The acceptor and donor splice sites which are located in MIR elements in the direct orientation are distributed randomly across the consensus, being generated following multiple nucleotide changes.

The putative acceptor splice site sequence has perfect sequence homology to the MIR consensus sequence at position 138-129nt (figure 5.4), these co-ordinates fall into the core-SINE region and the splice site is 100% conserved between all MIR sub-families. The strength of the acceptor splice site was calculated using GENIE and the consensus sequence gives a score of 8.4 (maximum possible of 14.2, average of 7.9 for constitutive exons; Zhang 1998; Reese *et al.,* 1997). The degree of conservation and the splice site strength suggests that no sequence changes are necessary from the original MIR sequence to provide a functional acceptor splice site for all sub-types, and subsequent sequence changes may merely strengthen or weaken the expression.

A donor splice site was identified at position 80-71nt of the MIR inverse consensus sequence and is located just outside of the core-SINE region. The putative donor splice site is predictably less conserved between the MIR elements, due to this MIR region demonstrating reduced conservation (see section 3.3). The donor splice site consensus determined, following aligning the pre-mRNA sequences, scores exceptionally high at 12.4, much higher than the average for constitutive exons (8.1). However sequence changes at 78-79nt (T/A and A/G) from the MIR consensus are necessary to achieve this splicing strength, as the score for the homologous region in the MIR consensus is weaker at 7.4 for sub-types MIR, MIRb, THER1_MD and MIR_Mars; MIR3 gives a weaker score of 3.1 and MIRm is not recognisable as a functional splice site with a score of -9.0 (table 5.3).

It was noted that splicing predominantly occurs in MIR and MIRb elements and no splicing events were detected for THER1_MD (table 5.1). Given that the acceptor splice site is completely conserved it would be anticipated that acceptor splice sites be observed for all of the MIR element sub-types; however when looking at the dataset as a whole (chapter 3) a very small proportion of the total MIRs are the older elements including THER1_MD, and those identified are predominantly in the 3'-UTR.

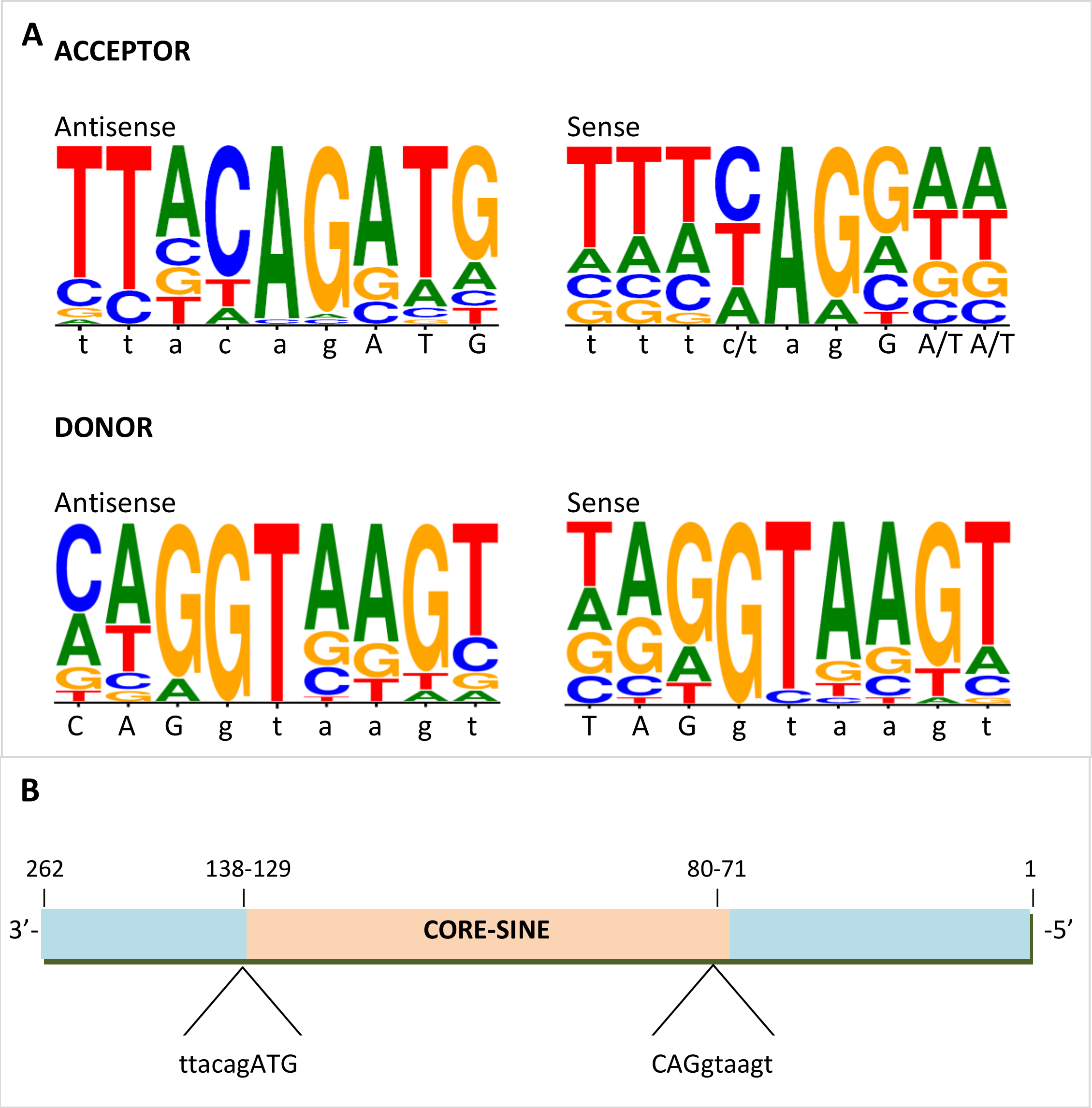
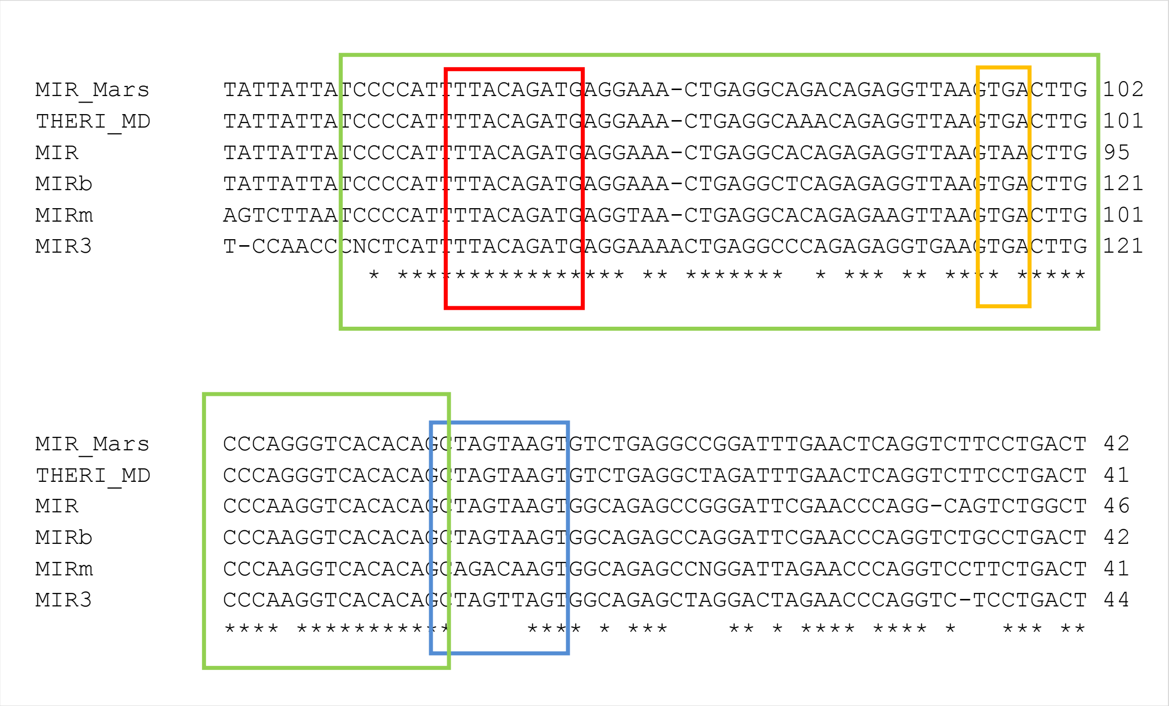


**Figure 5.4. Putative splice site for the MIR consensus sequences**

**A)** The MIR-derived splice site sequences were sorted according to whether the MIR was in the sense or antisense orientation. The splice site consensus sequence was predicted using WebLogo. The image demonstrates the probability of a nucleotide being present for each position with the height of the stacked letters signifying the frequency of each nucleotide. The top letter in the alignment is most conserved (Crooks *et al.,* 2004). **B)** The co-ordinates of the acceptor and donor slice sites identified within the antisense MIR sequence.



```
MIR_Mars        TATTATTATCCCCATTTTACAGATGAGGAAA-CTGAGGCAGACAGAGGTTAAGTGACTTG 102
THERI_MD        TATTATTATCCCCATTTTACAGATGAGGAAA-CTGAGGCAAACAGAGGTTAAGTGACTTG 101
MIR             TATTATTATCCCCATTTTACAGATGAGGAAA-CTGAGGCACAGAGAGGTTAAGTAACTTG 95
MIRb            TATTATTATCCCCATTTTACAGATGAGGAAA-CTGAGGCTCAGAGAGGTTAAGTGACTTG 121
MIRm            AGTCTTAATCCCCATTTTACAGATGAGGTAA-CTGAGGCACAGAGAAGTTAAGTGACTTG 101
MIR3            T-CCAACCCNCTCATTTTACAGATGAGGAAAACTGAGGCCCAGAGAGGTGAAGTGACTTG 121
                          * **************** ** *******  * *** ** **** *****


MIR_Mars        CCCAGGGTCACACAGCTAGTAAGTGTCTGAGGCCGGATTTGAACTCAGGTCTTCCTGACT 42
THERI_MD        CCCAGGGTCACACAGCTAGTAAGTGTCTGAGGCTAGATTTGAACTCAGGTCTTCCTGACT 41
MIR             CCCAAGGTCACACAGCTAGTAAGTGGCAGAGCCGGGATTCGAACCCAGG-CAGTCTGGCT 46
MIRb            CCCAAGGTCACACAGCTAGTAAGTGGCAGAGCCAGGATTCGAACCCAGGTCTGCCTGACT 42
MIRm            CCCAAGGTCACACAGCAGACAAGTGGCAGAGCCNGGATTAGAACCCAGGTCCTTCTGACT 41
MIR3            CCCAAGGTCACACAGCTAGTTAGTGGCAGAGCTAGGACTAGAACCCAGGTC-TCCTGACT 44
                **** ***********     **** * ***    ** * **** **** *   *** **
```

**Figure 5.5. Putative splice site for the MIR consensus sequences**

The MIR consensus sequences have been aligned in the antisense orientation outlining the sequence conservation for each MIR sub-family. The acceptor (red) and donor (blue) splice sites are boxed; the core-SINE is highlighted green and the stop codon observed with KLC1 and CHRD boxed orange. Sequence conservation was determined using the ClustalW2 alignment tool (http://www.ebi.ac.uk/Tools/clustalw2/index.html). MIR consensus sequences were obtained from RepBase, the primary reference database of prototypic repetitive DNA sequences. The MIR sub-families are names according to the nomenclature provided by Repbase (table 2.1).

| Consensus sequence | Nucleotide sequence | Splice site strength |
|---|---|---|
| Perfect | AAGgtaagt | 12.6 |
| Aligned mRNAs | CAGgtaagt | 12.4 |
| MIR | CTAgtaagt | 7.4 |
| MIRb | CTAgtaagt | 7.4 |
| THERI_MD | CTAgtaagt | 7.4 |
| MIR_Mars | CTAgtaagt | 7.4 |
| MIR3 | CTAgttagt | 3.1 |
| MIRm | CAGacaagt | -9.0 |

**Table 5.4. Putative splice site for the MIR consensus sequences**

The putative consensus donor splice site and the corresponding sequences of the MIR sub-families are scored. A ‘perfect’ donor splice site has been included as a reference. The splice site strengths were scored using the GENIE program.

## 5.5. Tissue specific expression of exonised MIR elements in the human genome

Several of the human genes which have exaptated MIR elements in coding exons were chosen for further analysis. It was not possible to investigate all of the genes in the dataset, therefore a sample set of the human genes which have recruited MIR elements were chosen. Genes were selected that are within the three main function groups observed in chapter 4 (neuronal function, immune responses and mammalian reproduction). The genes were further filtered if they demonstrated alternative splicing due to the presence of an MIR sequence. Abelson helper integration site 1 (AHI1) is essential in brain development and was selected due to the MIR element carrying a stop codon within a novel exon, resulting in the synthesis of a truncated putative protein product. An alternative transcript variant of the Class II, major histocompatibility complex, transactivator (CIITA) was also chosen, as it encodes a splice variant generated by the skipping of a donor splice site. The incorporated intronic sequence has recruited an MIR element which carries an alternative stop codon. Finally the gene Germ cell associated 1 like (GSG1L) was selected as a novel internal exon may be expressed due to splicing in and out of an intronic MIR element.

### 5.5.1. Expression profile of the splice variants of AHI1

In humans there are four transcript variants for AHI1, three differ in the 5’-UTR whereas the fourth is truncated at constitutive exon 25 (figure 5.6). The fourth transcript is generated by splicing into intron 25 of the full length reference-sequence (RefSeq) resulting in an alternative terminal exon. The novel region of the read-through exon contains an MIR element which is located 27bp downstream of the acceptor splice site and provides an in-frame stop codon (figure 5.6). The splice site is strong (score 10.4) and is only conserved in primates, MIR elements could also not be detected using RepeatMasker. Genomic sequences of AHI1 were aligned to detect the presence of the MIR repeats in other species. Demonstrating the recruitment of the MIR element in chimp, macaque, dog and cattle sequences and could not be detected in rodent species, and is most likely lost by deletion. The stop codon observed is only in-frame in the human splice variant due to indels; however translation terminates in an upstream stop codon within the MIR sequence for the other primates (figure 5.8). The other mammalian species studied (cow, dog) have recruited an MIR element, however there are no detectable in-frame stop codons present in the intron, and as such similar translation to that of primates could not occur. AHI1 is thought to be essential in brain development, and mutations are associated with autism, schizophrenia and symptoms linked to JBST 3; such as retinal dystrophy and central nervous system abnormalities (Alvarez-Retuerto *et al.,* 2008; Ingason *et al.,* 2007; Valente *et al.,* 2006). Considering the pathology of AHI1, high levels of mRNA expression would be expected in the retina, kidney and brain. RT-PCR was performed on human cDNA samples with isoform specific primers (figure 5.7) to determine if there is tissue specific expression of the AHI1 variants. The full length transcript is ubiquitously expressed in all tissues analysed with low levels of expression detected in the kidney and placenta (figure 5.7a). There appears to be restricted expression of the truncated AHI1 human mRNA, with expression detected in the lung, pancreas and testes (figure 5.7b).

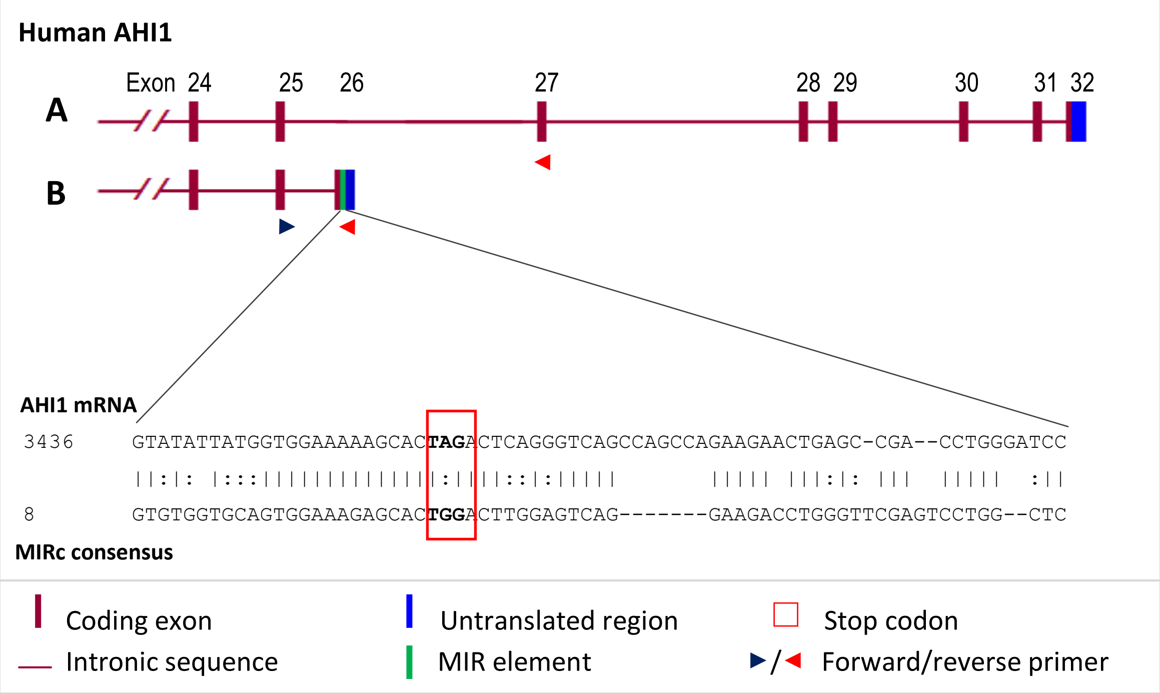


**Figure 5.6. Exonic organisation of Abelson helper integration site 1 (AHI1)**

Exonic organisation of AHI1 beginning at constitutive exon 24. Coding sequences are in red and UTRs in blue. The MIR element is integrated within an alternative terminal exon of the short variant (B) and carries an in-frame stop codon. The region with the stop codon has been aligned with the MIRc sense consensus sequence. The mRNA coordinates have been included and the MIR element spans the region 3436 – 3548bp. Primer sequences are listed in table 2.3 and the position of the forward (blue) and reverse (red) primers are indicated by arrows. Sequence conservation was determined using the ClustalW2 alignment tool (http://www.ebi.ac.uk/Tools/clustalw2/index.html; accession numbers appendix 9.3).

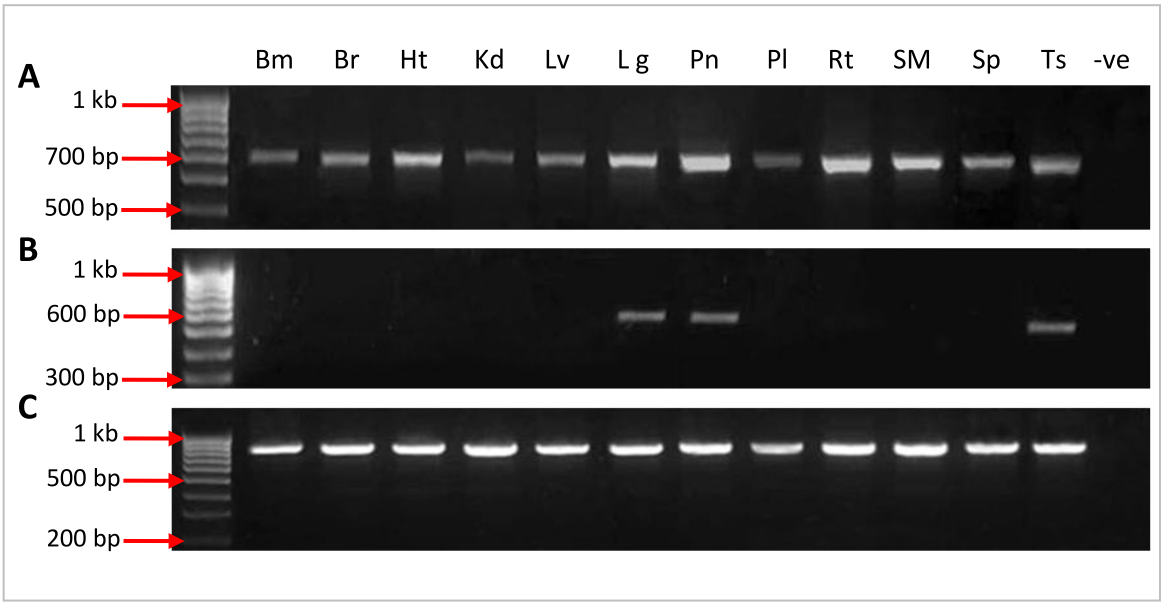


**Figure 5.7. Amplified AHI1 fragments from a number of human tissues**

**A)** Amplification of the full length AHI1 transcript, 594 bp; **B)** MIR-containing short AHI1 variant 670 bp; **C)** GAPDH positive control, 983 bp. All fragments were analysed on a 2% agarose gel with 1X SybrSafe. The fragments are of the anticipated size and were confirmed by sequencing. DNase and RNase-free water was used as a negative control. Tissues are abbreviated as follows; Bm, bone marrow; Br, brain; Ht, heart; Kd, kidney; Lv, liver; Lg, lung; Pl, placenta; Pn, pancreas; Rt, retina; SM, skeletal muscle; TSp, spleen and Ts, testis.

### 5.5.2. Expression profile of the splice variants of CIITA

CIITA is a transcriptional co-activator of major histocompatibility complex (MHC) class II gene transcription, and restores MHC class II gene expression in mutant cells. CIITA mutations are associated with bare lymphocyte syndrome (BLS) type II (Steimle *et al.*, 2003), and a SNP within the promoter region results in susceptibility to rheumatoid arthritis, multiple sclerosis, myocardial infarction (Swanberg *et al.*, 2005).

A novel transcript was identified for CIITA which encodes a truncated protein. The RefSeq is comprised of 18 exons (figure 5.9a), and the shorter transcript (figure 5.9b) is expressed as a result of skipping the donor splice site of constitutive exon 11 and the inclusion of the consecutive intron. The novel read-through region contains an MIRb element which carries an in-frame stop codon is located at nt2796. The MIR is located 141bp downstream of the splice site and as such the transcript encodes a novel 47 amino acid residues. The MIR-containing transcript also skips exon 6 of the RefSeq following which the reading frame is maintained. The first 11 exons of the mRNA of CIITA were aligned including the read-through sequence, for a number of mammals to determine if the transcript is conserved, and whether the MIR element has been exaptated in any other mammalian species (figure 5.8). The MIR element was detected by RepeatMasker for all of the species including chimp, macaque, rodents and canine but the stop codon is only in-frame in primate species. The skipped splice site is completely conserved (CAGgtgggg) in all animals and scores slightly lower than average at 7.7.

RT-PCR was performed to determine if there is tissue-specific expression of the two CIITA isoforms. The full length CIITA mRNA was found to be ubiquitously expressed (figure 5.10b), whereas the MIR-containing transcript demonstrates tissue-specific expression (figure 5.10a). High levels of expression were noted in the retina, spleen and thymus and lower levels in the pancreas and testis. The short form of CIITA was undetectable in bone marrow and skeletal muscle.

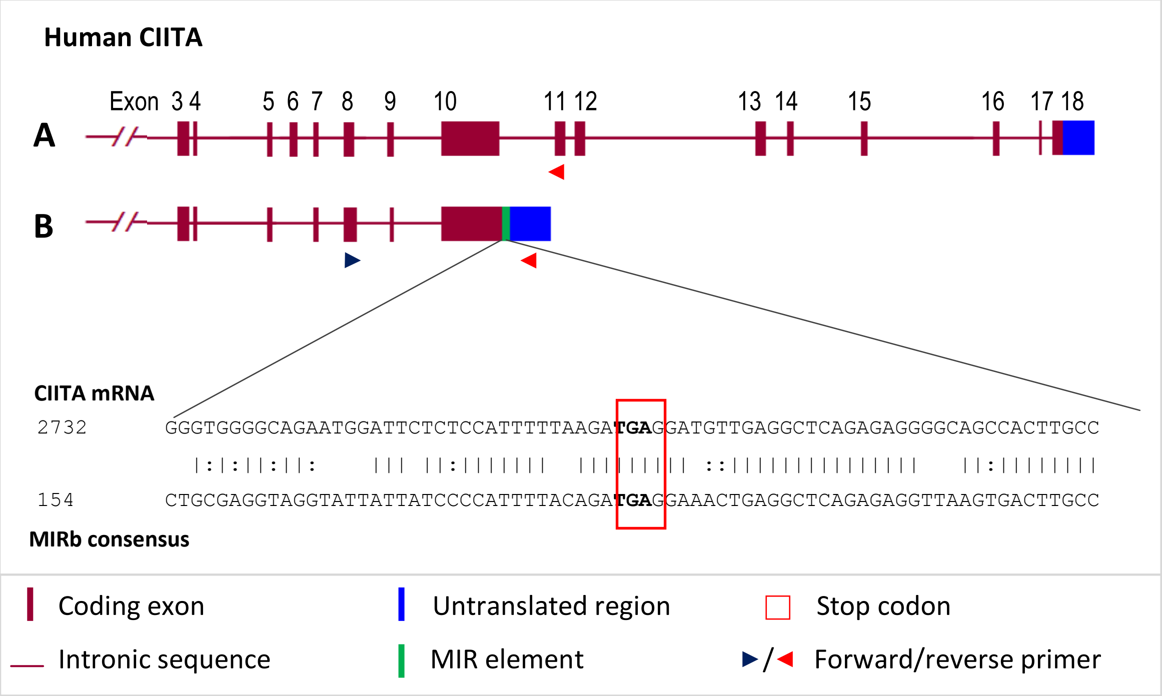

Human CIITA
Exon 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
A
B
CIITA mRNA
2732 GGGTGGGGCAGAATGGATTCTCTCCATTTTTAAGATGAGGATGTTGAGGCTCAGAGAGGGGCAGCCACTTGCC
154 CTGCGAGGTAGGTATTATTATCCCCATTTTACAGATGAGGAAACTGAGGCTCAGAGAGGTTAAGTGACTTGCC
MIRb consensus
Coding exon
Untranslated region
Stop codon
Intronic sequence
MIR element
Forward/reverse primer

**Figure 5.8. Exonic organisation of CIITA**

Exonic composition of CIITA beginning at constitutive exon 3. Coding sequences are in red and UTRs in blue. The MIR element is integrated within the terminal exon of the short variant (B) and carries an in-frame stop codon. The region with the stop codon has been aligned with the MIRb antisense consensus sequence. The mRNA coordinates have been included and the MIR element spans the region 2648 – 2844bp. Primer details are listed in table 2.3 and the position of the forward (blue) and reverse (red) primers are indicated by arrows. Sequence conservation was determined using the ClustalW2 alignment tool (http://www.ebi.ac.uk/Tools/clustalw2/index.html; accession numbers appendix 9.3).

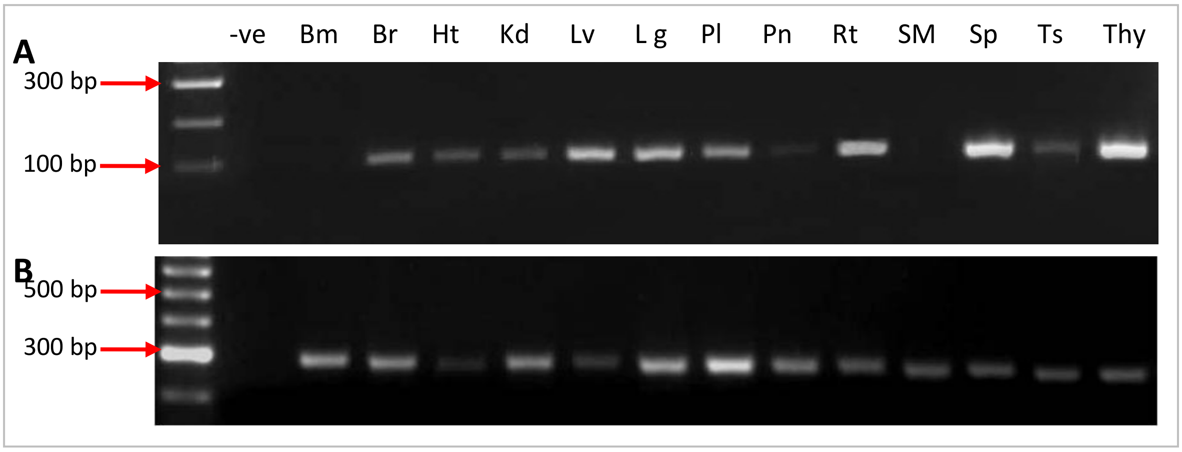


**Figure 5.9. Amplified CIITA fragments from a number of human tissues**

**A)** Amplification of a fragment of the truncated MIR-containing CIITA variant, 161 bp; **B)** the full length CIITA transcript, 303 bp. All fragments were analysed on a 2% agarose gel with 1X SybrSafe. The fragments are of the anticipated size and were confirmed by sequencing. DNase and RNase-free water was used as a negative control. Tissues are abbreviated as follows; Bm, bone marrow; Br, brain; Ht, heart; Kd, kidney; Lv, liver; Lg, lung; Pl, placenta; Pn, pancreas; Rt, retina; SM, skeletal muscle; Sp, spleen; Ts, testis and Thy, thymus.

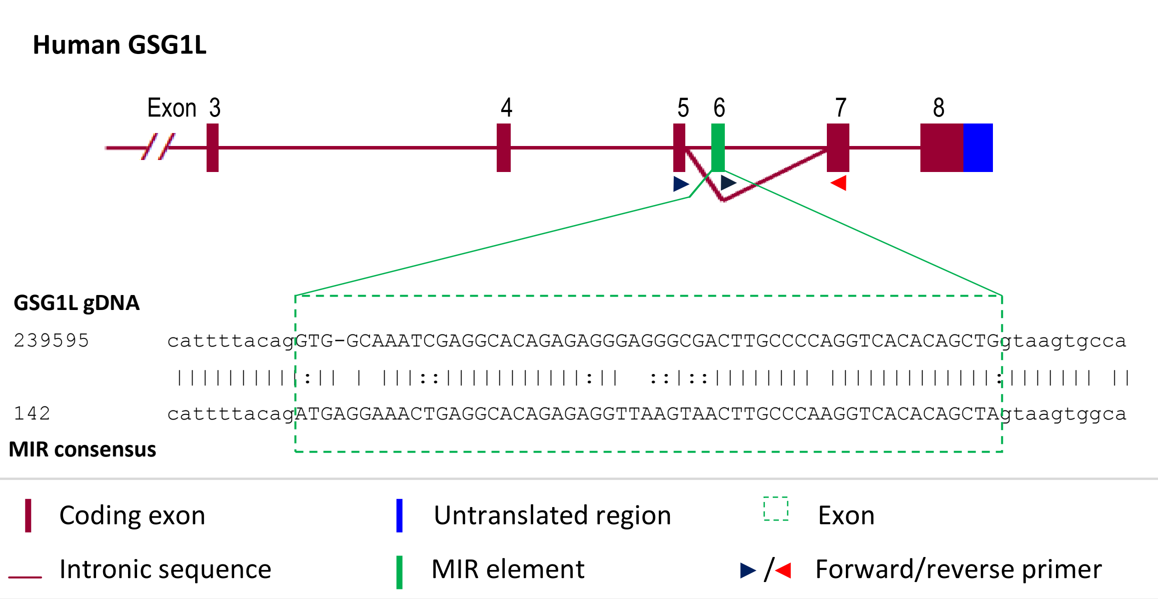
### 5.5.3. Expression profile of the splice variants of GSG1L

An alternative splice variant of GSG1L is generated by the inclusion of an alternative exon in intron 5; following which the reading frame is maintained for the remainder of the constitutive exons spliced (figure 5.11). Both the acceptor and donor splice sites are derived from an MIR element in the antisense orientation. The MIR element was aligned with the GSG1L gDNA sequence, revealing that the acceptor and donor splice sites are at the same region of the MIR consensus previously identified (section 5.4), and both the acceptor and donor splice site scored highly at 9.4 and 10.4 respectively. In the MIR-derived exons of KLC1 and CHRD a specific region of the core-SINE (108-106nt) was noted to provide the stop codon sequence, for the alternative GSG1L exon the termination codon is neither conserved nor in-frame. Multiple sequence alignments were generated to determine if the MIR is recruited in other species, and if there is conservation of the splice sites consistent with expression of the alternative exon (figure 5.8). RepeatMasker provided evidence of the retention of the MIR by chimpanzee, macaque, rat, mouse and dog; however the acceptor splice site is only sufficiently conserved in primates (humans, chimpanzee and macaque; table 5.5). A weaker splice site is present for dog and the sequence in rodents has diverged from the original MIR consensus and is not longer recognisable as a splice site.

| Source | Acceptor SS | Score | Donor SS | Score |
|---|---|---|---|---|
| MIR consensus | cccattttacagATG | 8.4 | CTAgtaagt | 7.4 |
| Human | tccattttacagGTG | 9.4 | CTGgtaagt | 7.4 |
| Chimp | tccattttacagATG | 8.6 | CTGgtaagt | 7.4 |
| Macaque | cccattttacagATG | 8.4 | CTGgtaagt | 7.4 |
| Dog | cccattttatagATG | 6.8 | CCGgtaaat | 6.9 |
| Rat | cctactttacatATG | -1.9 | CTGgctagt | -4.7 |
| Mouse | cctactttacatATG | -1.9 | CTGgctagt | -4.7 |

**Table 5.5. Acceptor and donor splice site strength for the MIR-derived exon of GSG1L for a number of mammals**

All mammalian sequences available in the nucleotide databases were included in the analysis. The splice site (SS) strengths were scored using the web base calculator GENIE. Intronic regions are in lowercase and exonic sequences capitalised. As a reference the average score for human constitutive exons is 7.9 and 8.1 for acceptor and donor splice sites respectively (Reese *et al.*, 1997).



```
239595       cattttacagGTG-GCAAATCGAGGCACAGAGAGGGAGGGCGACTTGCCCCAGGTCACACAGCTGgtaagtgcca
             ||||||||||:|| | |||::|||||||||||:||  ::|::|||||||| |||||||||||||:||||||| ||
142          cattttacagATGAGGAAACTGAGGCACAGAGAGGTTAAGTAACTTGCCCAAGGTCACACAGCTAgtaagtggca
```

**Figure 5.10. Exonic organisation of GSG1-like (GSG1L)**

Exonic composition of GSG1L beginning at constitutive exon 3. Coding sequences are in red and UTRs in blue. The MIR element is provided an internal cassette exon which is alternatively spliced. The exon has been aligned with the MIR antisense consensus sequence. The gDNA co-ordinates have been included. The intronic sequence including the splice sites is shown (lower case). Primer details are listed in table 2.3 and the position of the forward (blue) and reverse (red) primers are indicated by arrows. Sequence conservation was determined using the ClustalW2 alignment tool (http://www.ebi.ac.uk/Tools/clustalw2/index.html; accession numbers appendix 9.3).

Little is known regarding the molecular function of GSG1L, although expression appears to be restricted to the brain, eye and testis (UniGene cluster Hs.91910; http://www.ncbi.nlm.nih.gov/UniGene/clust.cgi?ORG=Hs&CID=91910). RT-PCR was performed to confirm the presence of the exonised MIR and to determine if there is differential expression of the two variants (figure 5.12). The RefSeq variant of GSG1L is ubiquitously expressed, with high levels of expression in the brain, heart and retina, with weak expression noted in bone marrow and liver. The MIR-derived isoform displays a different expression pattern and was undetectable in bone marrow and liver, and expressed at low levels in the kidney, suggesting a potentially different function for this splicoform.

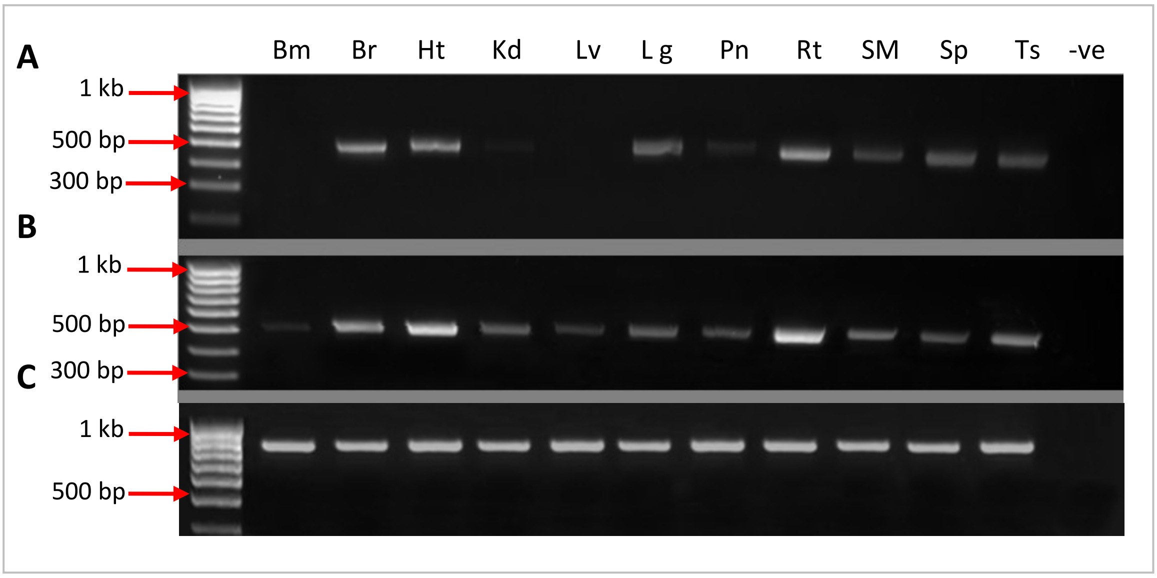
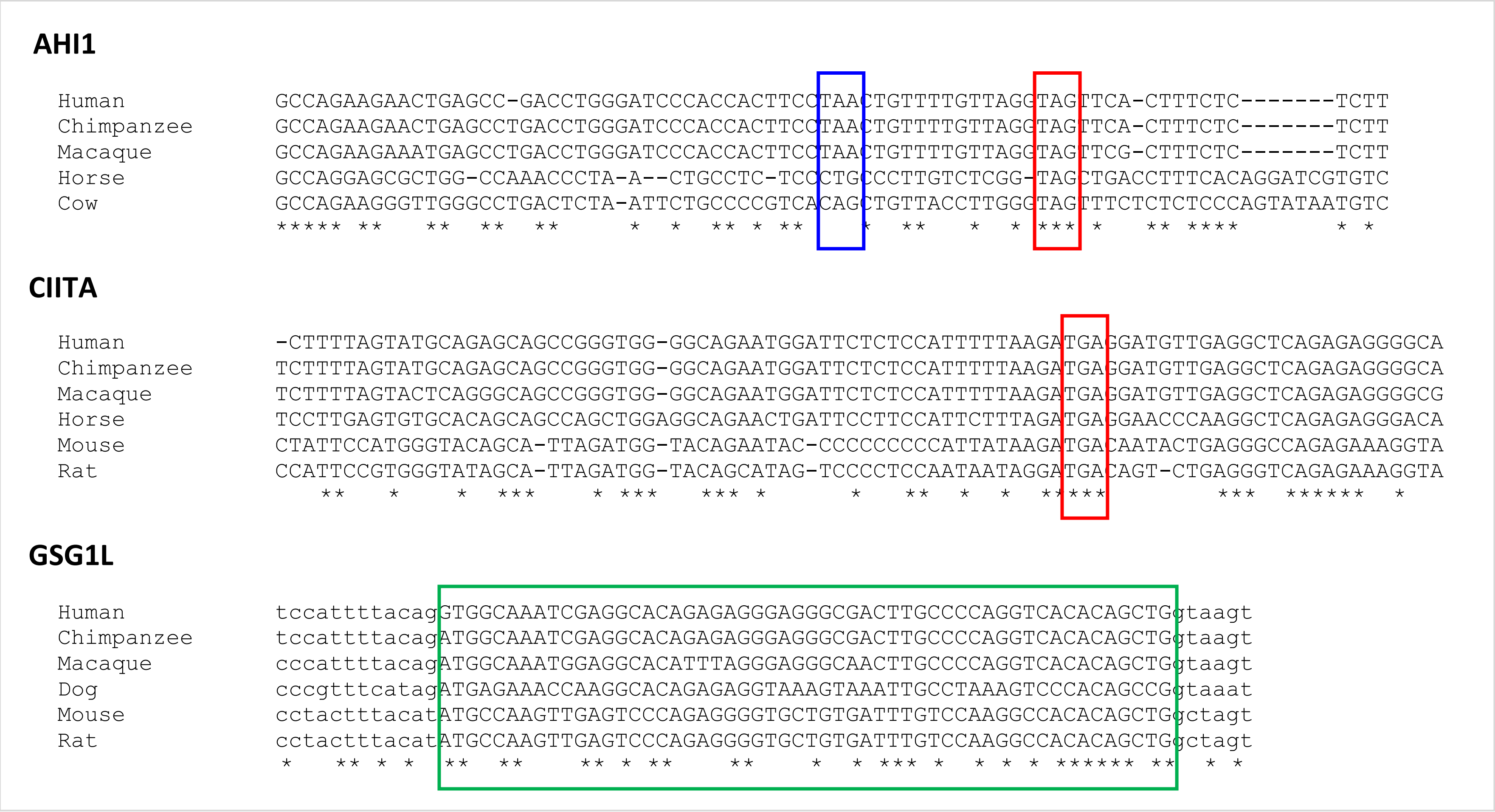


**Figure 5.11. Amplified GSG1L fragments from a number of human tissues**

**A)** Amplification of the short MIR-containing GSG1L variant, 499 bp; **B)** the full length GSG1L transcript, 489 bp; **C)** GAPDH positive control, 983 bp. All fragments were analysed on a 2% agarose gel with 1X SybrSafe. The fragments are of the anticipated size and were confirmed by sequencing. DNase and RNase-free water was used as a negative control Tissues are abbreviated as follows; Bm, bone marrow; Br, brain; Ht, heart; Kd, kidney; Lv, liver; Lg, lung; Pn, pancreas; Rt, retina; SM, skeletal muscle; Sp, spleen and Ts, testis.



**AHI1**

```
Human           GCCAGAAGAACTGAGCC-GACCTGGGATCCCACCACTTCCTAACTGTTTTGTTAGGTAGTTCA-CTTTCTC-------TCTT
Chimpanzee      GCCAGAAGAACTGAGCCTGACCTGGGATCCCACCACTTCCTAACTGTTTTGTTAGGTAGTTCA-CTTTCTC-------TCTT
Macaque         GCCAGAAGAAATGAGCCTGACCTGGGATCCCACCACTTCCTAACTGTTTTGTTAGGTAGTTCG-CTTTCTC-------TCTT
Horse           GCCAGGAGCGCTGG-CCAAACCCTA-A--CTGCCTC-TCCCTGCCCTTGTCTCGG-TAGCTGACCTTTCACAGGATCGTGTC
Cow             GCCAGAAGGGTTGGGCCTGACTCTA-ATTCTGCCCCGTCACAGCTGTTACCTTGGGTAGTTTCTCTCTCCCAGTATAATGTC
                ***** **   **  **  **     *  *  ** * **    *  **   *  * *** *   ** ****       * *
```

**CIITA**

```
Human           -CTTTTAGTATGCAGAGCAGCCGGGTGG-GGCAGAATGGATTCTCTCCATTTTTAAGATGAGGATGTTGAGGCTCAGAGAGGGGCA
Chimpanzee      TCTTTTAGTATGCAGAGCAGCCGGGTGG-GGCAGAATGGATTCTCTCCATTTTTAAGATGAGGATGTTGAGGCTCAGAGAGGGGCA
Macaque         TCTTTTAGTACTCAGGGCAGCCGGGTGG-GGCAGAATGGATTCTCTCCATTTTTAAGATGAGGATGTTGAGGCTCAGAGAGGGGCG
Horse           TCCTTGAGTGTGCACAGCAGCCAGCTGGAGGCAGAACTGATTCCTTCCATTCTTTAGATGAGGAACCCAAGGCTCAGAGAGGGACA
Mouse           CTATTCCATGGGTACAGCA-TTAGATGG-TACAGAATAC-CCCCCCCCCATTATAAGATGACAATACTGAGGGCCAGAGAAAGGTA
Rat             CCATTCCGTGGGTATAGCA-TTAGATGG-TACAGCATAG-TCCCCTCCAATAATAGGATGACAGT-CTGAGGGTCAGAGAAAGGTA
                   **   *    *  ***    * ***   *** *      *   **  *  *  *****        ***  ******  *
```

**GSG1L**

```
Human           tccattttacagGTGGCAAATCGAGGCACAGAGAGGGAGGGCGACTTGCCCCAGGTCACACAGCTGgtaagt
Chimpanzee      tccattttacagATGGCAAATCGAGGCACAGAGAGGGAGGGCGACTTGCCCCAGGTCACACAGCTGgtaagt
Macaque         cccattttacagATGGCAAATGGAGGCACATTTAGGGAGGGCAACTTGCCCCAGGTCACACAGCTGgtaagt
Dog             cccgtttcatagATGAGAAACCAAGGCACAGAGAGGTAAAGTAAATTGCCTAAAGTCCCACAGCCGgtaaat
Mouse           cctactttacatATGCCAAGTTGAGTCCCAGAGGGGTGCTGTGATTTGTCCAAGGCCACACAGCTGgctagt
Rat             cctactttacatATGCCAAGTTGAGTCCCAGAGGGGTGCTGTGATTTGTCCAAGGCCACACAGCTGgctagt
                 *   ** * *  **  **    ** * **    **    *  * *** *  * * * ****** **  * *
```

**Figure 5.12. Multiple sequence alignment of the MIR elements identified for CIITA, GSG1L and AHI1 for a number of mammalian species**

The exonic region of GSG1L is boxed green, intronic regions are in lowercase and the exonic sequence is in capitals. The human stop codon region is highlighted in red and the alternative primate in-frame primate stop codon for AHI1 is blue. Sequence conservation was determined using the ClustalW2 alignment tool (http://www.ebi.ac.uk/Tools/clustalw2/index.html; accession numbers appendix 9.3).

## 5.6. Discussion

The creation of *de novo* exons via exonisation provides a means of generating alternative splice variants, and subsequently a range of diverse gene products. Nascent exons are most commonly expressed at lower levels than constitutive exons. Expressing new exons at low levels will allow for the adaptation and generation of novel functions, with minimal consequence to the existing gene function, as the original splicoform remains predominant (Nurtdinov *et al.,* 2009; Kim *et al.,* 2008). Numerous examples exist where Alus are exonised thus increasing the transcriptome, however these examples are limited to primate genomes. Furthermore Alus are relatively young retrotransposon (<120myr) compared to MIR elements and the inclusion of a *de novo* exon cannot be attributed to splice site strength but rather to exon age (Kim *et al.,* 2008). Older exons are more strongly spliced due to having sufficient time to acquire regulatory sequences such as splice enhancers, allowing for a higher inclusion strength and greater establishment (Ast, 2004).

In chapter 3 MIR elements were noted to be contributing protein coding sequences in 117 transcripts and many are alternatively spliced. There has been previous interest into the exonisation of TEs including Alus, MIRs and LTR elements (Krull *et al.,* 2005; Sela *et al.,* 2007; Krull *et al.,* 2007; Lin *et al.,* 2009; Piriyapongsa *et al.,* 2007). Therefore the dataset of genes identified in chapter 3 were screened for spliced MIR elements and translational control codons, not only to compare with other studies but to give a more complete picture of the distribution of these elements in human genes.

### 5.6.1. Exaptated MIR elements may provide functional splice sites and contribute to alternative splicing

Overall a total of 126 splicing events take place within an exaptated MIR element, of which 78 generate alternative transcripts whilst the remainder are constitutively spliced. Of the exonised MIR elements, there appears to be no obvious preference to exonise at either 3'- or 5'- splice sites and a similar number of acceptor and donor splice sites are located within MIR sequences (67 and 59 respectively). There are also equal numbers of 5'-UTR acceptor and donor splicing events observed (62 and 63 respectively). There were no examples of MIRs recruited in the 3'-UTRs providing functional splice sites. MIR elements were also noted to contribute initiating methionine codons and translational termination codons (22 and 65 respectively). The majority of splicing events occur primarily in the younger MIR elements, predominantly the MIR and MIRb sub-types, with no splicing noted in the retained sequence of the sub-family THER1_MD and few observed for MIRm and MIR_Mars. However this can be partially explained by the data obtained in chapter 3, as the vast majority of the total MIR elements identified in human genes being the younger elements (MIR, MIRb and MIR3) with the older elements constituting only 9% of the total exaptated. Furthermore the majority of the older MIR repeats are recruited in the 3'-UTRs. For example eight THER1_MD elements are found in 5'-UTRs of human genes with the remainder in the 3'-UTR (note that MIRs in the 3'-UTRs have not been exonised).

EST sequences with exonised MIR elements corresponding to 73 different annotated genes were identified, making a total of 243 human genes which have exonised MIR elements and/or have the repeats contributing to protein coding sequences. There were no corresponding full length cDNA clones detected to support the expression of the EST sequences, and as such the reading frames and complete exonic composition can not be predicted, therefore these examples were not included in the analysis (appendix 9.9), however taking into account the annotated genes with exonised EST sequences, it is possible that almost 1% of the total human genes in the genome contain MIR-derived protein coding sequences, corresponding to 25% of the genes with TEs in protein-coding regions (Nekrutenko and Li, 2001).

The majority of the acceptor splice sites are present when the MIR element is in the inverse orientation (76%), but there is no apparent orientation preference for donor splicing, with 53% occurring when in the MIR has integrated in the inverse orientation. There were 26 distinct whole internal cassette exons which are derived from a single MIR element, of which 42% are constitutively spliced. Of the MIR-derived exons, 12 are protein-coding with the remainder being in the 5'-UTR; with the majority of the MIR repeats in the inverse orientation (73%). Putative acceptor and donor splice sites were apparent in the MIR inverse consensus sequence following the generation of multiple sequence alignments of all of the exonised nucleotide sequences. The acceptor splice site is highly similar to the vertebrate splice consensus sequence and a *de novo* MIR element theoretically could have been strongly spliced following integration. There would be no need for sequence changes from the consensus, as the putative splice sequence is completely conserved between MIR sub-families, with being in the core-

SINE region. The putative donor splice region is less conserved between the MIR elements and at least two base changes are necessary. Aligning the nucleotide sequences of these splice sites (including the intronic sequence) demonstrates the consensus sequence of all of the splice sites generated through MIR recruitment. Both the acceptor and donor splice site consensus sequences generated score exceptionally high (8.4 and 12.4 respectively), much higher than average constitutive splice sites, demonstrating that following mutation an MIR could potentially generate a complete exon of ~50bp in all mammalian genomes. However if this was the case most antisense MIR elements would be exonised, considering the degree of homology between different MIR elements, suggesting further regulatory sequences are necessary if the splice site is to be actively maintained. As such most of the dormant putative splice sites of exaptated MIR elements in the inverse orientation have diverged and no longer resemble canonical splice sites.

The majority of exonised MIR elements observed in this investigation are a direct result of splicing into intronic MIR repeats. These elements may be actively spliced following mutation or the acquisition of further regulatory sequences, resulting in the spliceosome recognising the sequence motifs which resemble the splice sites consensus. Essentially intronic MIR elements may have provided a way for genomes to 'experiment' during evolution, by expressing novel alternative splice variants at low levels at no expense to normal gene function.

### 5.6.2. Tissue specific expression of transcript variants which have exonised MIR elements and human disease

171 human genes contain exonised MIR elements, of which many demonstrate alternative splicing. A number of genes were therefore selected and analysed for tissue specificity and validated using RT-PCR. AHI1 is critical for normal brain development and is mutated in neurodegenerative related disorders, learning difficulties and psychoses, including autism, Joubert syndrome and schizophrenia (Alvarez Retuerto *et al.,* 2008; Tory *et al.,* 2007; Valente *et al.,* 2006; Amann-Zalcenstein *et al.,* 2007). The gene GSG1L was selected as there are two alternatively spliced transcripts, differing by the presence of an internal coding exon which is derived entirely from a single MIR element. Little is known regarding the function of GSG1L however expression is

restricted to the brain, whole eye and testis (UniGene cluster Hs.91910) , presenting the gene as an interesting candidate for further analysis, due to the putative functional role of MIR elements in neuronal function discussed previously. The final gene to be investigated was the MHC class II transactivator CIITA, which is mutated in immunodeficiency syndromes such as bare lymphocyte syndrome type II (BLS), leukaemias, multiple sclerosis and rheumatoid arthritis (Steimle *et al.,* 1993; Swanberg *et al.,* 2005).

All three of these genes contain alternative splice variants as a direct consequence of the recruitment of an MIR element. Both AHI1 and CIITA express a truncated protein product due to a stop codon in an MIR element in a read-through exon and GSG1L contains an internal MIR-derived coding exon. Of these splice variants, all three demonstrate tissue-specific expression, compared to the the full length transcripts which have not recruited an MIR element. The truncated AHI1 transcript, which appears to be conserved only in primate orders, displayed a restricted expression pattern and could be detected only in the lung, pancreas and testes. This transcript is not expressed in the brain and retina; key tissues when considering the pathological role of this gene, such as in the development of Joubert syndrome, schizophrenia and kidney disease. The tissue specificity of the shorter AHI1 mRNA transcript suggests that the splice variants may have distinctly different functional roles in these tissues. It is possible that the read-through transcript may be redundantly expressed and non-functional, the splice site being skipped due to the presence of weak splice signals; however this is unlikely considering the tissue-specific expression of the MIR-containing transcript.

The MIR-derived splice variant of GSG1L displayed tissue specific expression with no detectable expression in bone marrow and liver, and low expression levels in the kidney. The full length transcript is expressed in all of the tissues analysed, though detectable at low levels in the bone marrow, liver and spleen. Both of the GSG1L isoforms are highly expressed in brain, heart and retina. GSG1L encodes a protein of unknown function, when analysing the conserved domains listed in the Pfam database there is sequence similarity with a domain conserved in proteins specifically expressed in germ line testicular cells (Matsui *et al.,* 1997). A further conserved region is shared with the Claudin family of proteins, which are involved in the formation of tight junctions, and have been known to localise to the cilia of the retinal pigment epithelium (Rahner *et al.,* 2004). GSG1L is a recently annotated gene and these results demonstrate differential expression and potentially different functional roles of the two splicoforms.

The main area of discussion during chapter 3 primarily focussed on the role of MIR elements in neuronal gene function; however it is highly likely that MIRs may play a role in other biological processes. For example aspects of the immune system such as cytokine function appeared to be significant when consulting the Gene Ontology. The Joubert syndrome loci, AHI1 and NPHP1, are also associated with non-neurological disorders such as coeliac disease, leukaemia and kidney disease (Maiuri *et al.,* 2005; Jiang *et al.,* 2004; Ala-Mello *et al.,* 1998). Therefore a gene known to function in the immune system was selected.

CIITA regulates MHC class II gene transcription and loss of CIITA expression has been associated with chronic myeloid leukaemia (CML) and bare lymphocyte syndrome, type II (Day *et al.,* 2003; Dziembowska *et al.*, 2002). CIITA encodes several transcripts variants, of which one is truncated due to skipping a donor splice site resulting in the incorporation of an intronic MIR element. The exaptated MIR carries the premature stop codon. Following RT-PCR the full length sequence was detected in all tissues and the truncated MIR-containing isoform could not be detected in bone marrow or skeletal muscle. The C-terminus of the full length CIITA transcript contains two leucine rich repeats (LRRs), which are essential for the interaction with MHC class II promoter-binding proteins (Sisk *et al.,* 2001). The MIR-containing transcript lacks the LRRs and as such the truncated CIITA protein may be unable to revert MHC class II-negative cells to express class II antigens (Day *et al.,* 2003). A single nucleotide mutation in the C-terminal end of the full length CIITA isoform have been attributed to BLS, and it is possible that the CIITA isoform which has exaptated the MIR element may contribute to the progression of CML and BLS (Day *et al.,* 2003; Dziembowska *et al.*, 2002). Expression of the short form of CIITA was not detected in normal bone marrow by RT-PCR, which is interesting as expression of the truncated CIITA protein has been detected in K-562 (myelogenous leukaemia) and Rajii (Burkitt's lymphoma-derived) cell lines (Day *et al.,* 2003; Riley *et al.,* 1995). This further suggests that upregulation of the MIR-containing transcript may be involved in the progression of leukaemias and BLS.

## 5.7. Conclusion

The exonisation of MIR elements in the human genome suggests that MIRs may play a significant role in mammalian genome evolution by shaping the transcriptome. Overall it appears that 1% of the total human genes have MIR-derived splice sites or coding sequences. Furthermore 38% of the splicing events observed generate constitutive exons, in contrast to the distribution of exonised Alu elements which are predominantly alternatively spliced (Sorek *et al.,* 2002; Sela *et al.,* 2007). The MIR family are more ancient and sufficient time may have passed for the MIR elements to be accepted as fully functional and highly expressed exons, whereas the younger exonised Alus may still be in their infancy.

# 6. MIR ELEMENTS AND THE ALTERNATIVE SPLICING OF TISSUE TRANSGLUTAMINASE (TGM2)

## 6.1. Introduction

TGM2 is of particular interest to this study as all of the documented isoforms have recruited independent MIR elements. A further truncated transcript was identified following *in silico* analysis which has exaptated six MIRs, recruited in the coding sequencece and both the 5'- and 3'- UTRs (appendix 9.1). Secondly, TGM2 dysfunction is implicated in several of the key disorders discussed in chapter 5, such as the neurodegenerative conditions; schizophrenia, Alzheimer's disease and Huntington's disease and both Type-2 diabetes and MODY (maturity onset diabetes of the young; table 4.7).

Tranglutaminases (TGases) comprise a large family of structurally similar enzymes, which catalyse protein cross-linking via specific ε-(γ-glutamyl) lysine isopeptide bonds in a $Ca^{2+}$ dependent manner (Aeschlimann and Paulsson, 1991). In humans there are seven distinct TGase genes, which all post-translationally modify proteins (Aeschlimann *et al.,* 1998; Chen and Mehta, 1999). TGases differ in their biological function and peptide sequence; however they share a common active site and all display transamidation activity. Tissue transglutaminase (TGM2) is a multifunctional enzyme which differs from its relatives, as in addition to TGase activity and protein cross-linking TGM2 has the ability to bind and hydrolyse GTP (Lee *et al.,* 1989). TGM2 GTPase activity is similar to the α subunits of large heterotrimeric G proteins (Mhaouty-Kodja, 2004).

TGM2 is ubiquitously expressed and functions as a monomer (Lin and Ting, 2006). In humans the full length RefSeq protein comprises 687 amino-acids (aa), the nucleotide sequence, including the UTRs, is ~4 kb in length and is composed of 13 coding exons. The full length TGM2 protein encompasses a number of binding sites and activity domains. The N-terminus consists of a β-sandwich domain which contains fibronectin and integrin binding sites. The catalytic core contains a triad of papain-like catalytic residues, calcium binding sites and a conserved tryptophan (Trp) residue (Murthy *et al.,* 2002). The C-terminus is composed of two β-barrels, the first containing a collection of guanine binding sites and the second β-barrel interacts with phospholipase C-δ1 (PLCδ1), α1-adrenergic receptors (α1-AR) and the G-protein coupled receptor GPR56 (Xu and Hynes, 2007). The tertiary structure of the TGM2 protein varies; when bound to GTP, at normal cellular levels of free $Ca^{2+}$, the TGase active sites are blocked and TGM2 functions as a guanine nucleotide-binding protein (G-protein) in receptor signalling, and is not involved in TGase activity. When GTP is hydrolysed the three dimensional structure changes and binding to $Ca^{2+}$ exposes two Trp residues, which control the access of the substrate to the active site (Fesus and Piacentini, 2002). It is thought that GTP may act as a molecular 'switch' between the two TGM2 activities (Monsonego *et al.,* 1998; Festoff *et al.,* 2002).

TGM2 has been implicated in complex and diverse biological functions including cell differentiation and growth, active responses to injury, cell adhesion and protein aggregation (Collighan and Griffin, 2009; Verderio *et al.,* 2005). TGM2 disruption is implicated in the pathogenesis of numerous neurological conditions including Alzheimer Disease, Huntington's disease, Parkinson's disease and schizophrenia (Ikura *et al.,* 1993; Lesort *et al.,* 2002; Andringa *et al.,* 2004; Bradford *et al.,* 2009). TGM2 has also been implicated in inflammation, rheumatoid arthritis, diabetes (Type I, Type II and MODY) cystic fibrosis and has been identified as an autoantigen in coeliac disease (Kim, 2006; Maiuri *et al.,* 2005; Dieterich *et al.,* 1997; Hummel *et al.,* 2007; Porzio *et al., 2007;* Bernassola *et al.,* 2002; Maiuri *et al.,* 2008).

There are two widely documented TGM2 isoforms which differ not only in structure but also in their function. The full length RefSeq transcript demonstrates protective activity against apoptotic events and subsequent cell death, whereas the short form is shown to be upregulated prior to cell death and has been suggested to promote apoptotic responses (Antonyak *et al*., 2006). The shorter TGM2 peptide is 548aa in length and is encoded by a transcript of ~1.9 kb. The truncated isoform lacks the C-terminal region which contains the guanine nucleotide binding sites and is unable to bind GTP (Antonyak *et al*., 2006). A third protein product (TGM2_Vs) has also been reported of approximately 38.7 kDa, detected in the human erythroleukemia (HEL) cell line, induced by retinoic acid (Fraij and Gonzales, 1996). Finally an aberrant splicing event has been reported at RefSeq exons 12 and 13, generating a transcript with a shorter UTR and lacking the terminal 63-residues (TGM2_dLg) (Lai *et al.,* 2007).

## 6.2. The identified TGM2 isoforms and the exaptation of MIR elements

As MIR elements were actively transposing prior to the radiation of the placental mammals it is possible that the MIRs may have been recruited in TGM2 of other mammalian species. Nucleotide databases were screened for trace nucleotide sequences, ESTs and cDNA clones which correspond to the genomic region of TGM2 for a number of species. ESTs and trace sequences were accepted if a splice site was present which conforms to the consensus sequences described by Zhang (1998), as any non-spliced ESTs could represent splicing errors, natural antisense transcripts or nRNA. Where necessery composite mRNA sequences were generated and all identified splice variants were screened for the presence of MIR elements.

A total of five human TGM2 splice variants were identified (figure 6.1); four of which have been published (Lai *et al.,* 2007; Antonyak *et al.,* 2006; Festoff *et al.,* 2002; Fraij and Gonzales, 1996) and a novel fifth transcript (TGM2_Tc). All five isoforms have recruited MIR elements (table 6.2), the RefSeq transcript (TGM2_wLg) has recruited three MIR elements within the 3'-UTR located 559, 670 and 885bp downstream of the termination codon; TGM2_Tc encodes the same three terminal exons (11-13) as TGM2_wLg and as such has recruited the same three MIR elements. TGM2_Sh, shares the initial 538-residues with the RefSeq TGM2. The protein is truncated at constitutive exon 10, due to reading through the donor splice site, resulting in the retention of intronic sequence encoding 10 novel residues and an alternative stop codon. TGM2_Sh contains an unrelated MIR element in the 3'-UTR sequence, 64bp downstream of the stop codon. The third transcript, TGM2_Vs, is generated in a similar manner to TGM2_Sh, with skipping of the donor splice site occurring at RefSeq exon 6; a further independent MIR element is present in the 3'-UTR, 91bp downstream of the termination codon.

A further transcript designated TGM2_dLg is generated by a rare splicing event at the consensus CT/YCAC, which does not appear to conform to conventional splicing mechanisms, and is discussed in more detail in section 6.3 (Lai *et al.,* 2007). This transcript contains the same MIR elements noted in the 3'-UTR of the full length transcript; however due to the altered reading frame an alternative stop codon is carried by the first of the three MIR elements.

The final transcript, TGM2_Tc, is the only one of the five isoforms regulated by an alternative promoter as the initiating methionine is located within RefSeq intron 10. The protein encodes 147 novel N-terminal residues followed by the common TGM2 C-terminus. TGM2_Tc has exaptated seven MIR elements, one in the 5'-UTR, one which contains the methionine codon and contributes 51 residues of peptide sequence, one within the coding sequence and the remaining three within the 3'-UTR, at the same positions seen with the full length transcript (table 6.1).

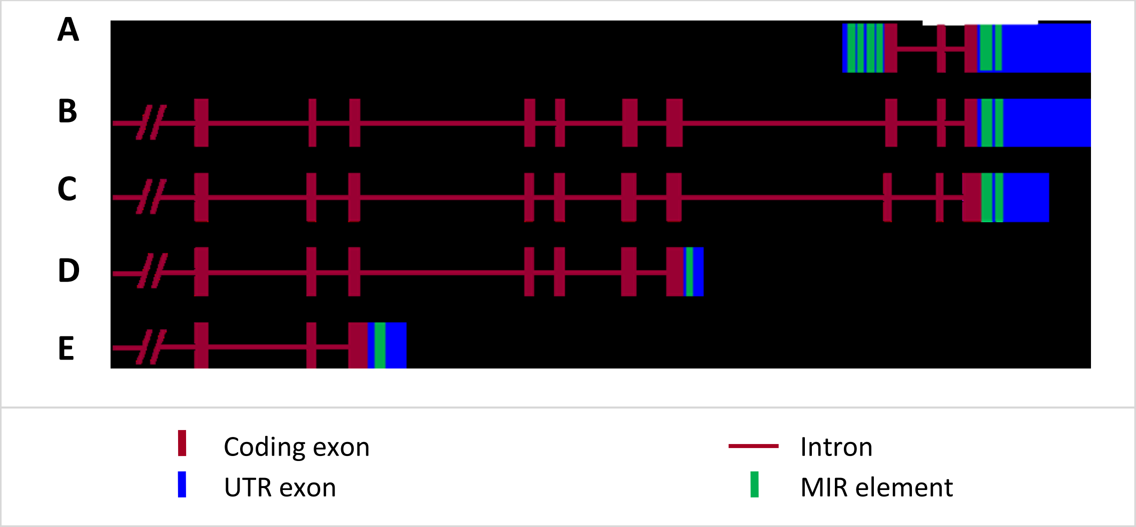


**Figure 6.1. Exonic composition of the multiple splice variants of TGM2**

The exonic composition of the TGM2 isoforms begins at canonical exon 4. **A)** TGM2_Tc, the transcript start site is located within common intron 10 and the initiating methionine is within a recruited MIR element: **B)** TGM2_wLg is the full length RefSeq mRNA of 13 exons: **C)** TGM2_dLg demonstrates variation in splicing between canonical exons 11 and 12, an alternative reading frame is present and the in-frame stop codon is within an MIR element: **D)** TGM2_Sh is truncated due to reading through the donor splice site at exon 10 and contains an MIR in the 3'-UTR: **E)** TGM2_Vs is truncated at exon 6 and has recruited an MIR element in the 3'-UTR. The position of the last common residue V286 for TGM2_Vs and S538 for TGM2_Sh has been included as a reference. S538 is also the first amino acid which is shared with TGM2_wLg.

Since MIR elements are present in all mammalian genomes, the genomic sequences of a number of mammals were studied to determine if the TGM2 isoforms are conserved or expressed, and if the MIR elements have been recruited (table 6.1). The TGM2 isoforms were identified in all species studied with the exception of TGM2_Tc, which could only be predicted in primates, the initating methionine sequence is absent in other species due to the deletion of the MIR elements in the novel region. TGM2_dLg could only be predicted for primates and rat due to either limited gDNA sequences in the databases or a lack of conservation of the cryptic splice site. An MIR element could not be detected for the rodent full length TGM2 using RepeatMasker.

| **SPECIES** | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Isoform ID** | **Human** | **Chimp** | **Macaque** | **Dog** | **Cow** | **Rat** | **Mouse** |
| **TGM2_Vs** | 3'-UTR | 3'-UTR | 3'-UTR | 3'-UTR | 3'-UTR | 3'-UTR | 3'-UTR |
| **TGM2_Sh** | 3'-UTR | 3'-UTR | 3'-UTR | 3'-UTR | 3'-UTR | 3'-UTR | NM |
| **TGM2_dLg** | TAG,<br>3'-UTR<br>3'-UTR | TAG,<br>3'-UTR<br>3'-UTR* | TAG,<br>3'-UTR<br>3'-UTR* | NE | NE | -<br>3'-UTR<br>3'-UTR* | NE |
| **TGM2_wLg** | 3'-UTR<br>3'-UTR<br>3'-UTR | 3'-UTR*<br>3'-UTR<br>3'-UTR | 3'-UTR*<br>3'-UTR<br>3'-UTR | 3'-UTR*<br>3'-UTR*<br>3'-UTR | 3'-UTR*<br>3'-UTR*<br>3'-UTR | -<br>3'-UTR *<br>3'-UTR * | -<br>3'-UTR *<br>3'-UTR * |
| **TGM2_Tc** | 5'-UTR<br>ATG<br>CDS<br>3'-UTR<br>3'-UTR<br>3'-UTR | 5'-UTR<br>ATG<br>CDS<br>3'-UTR<br>3'-UTR<br>3'-UTR* | 5'-UTR<br>ATG<br>CDS<br>3'-UTR<br>3'-UTR<br>3'-UTR* | NE | NE | NE | NE |

**Table 6.1. TGM2 splice variants and the presence of MIR elements for a number of mammalian species**

Abbreviation: NE) No evidence of the isoform or limited sequence data available; NM) There is evidence for the transcript but no MIR element is detected; *) No MIR can be detected using RepeatMasker but there is evidence of an MIR element when aligning with the human mRNA. The gene region which the MIR element is detected has been included for all isoforms and species (ATG, methionine; TAG, stop codon; UTR, untranslated region).

## 6.3. The functional domains and binding sites of TGM2 varies between transcript variants

The full length TGM2 protein contains a collection of documented binding sites and activity domains (Chen and Mehta, 1999; figure 6.2). There is a β-sandwich domain at the N-terminus which contains fibronectin binding sites. The catalytic core contains a triad of catalytic residues and calcium binding sites and the there are two β-barrels at the C-terminus which containing binding sequences for PLC-δ1 and α1-AR. The TGM2 isoforms identified differ in their peptide sequence and as such the presence of the described binding sites and domains was determined for each of the TGM2 isoforms.

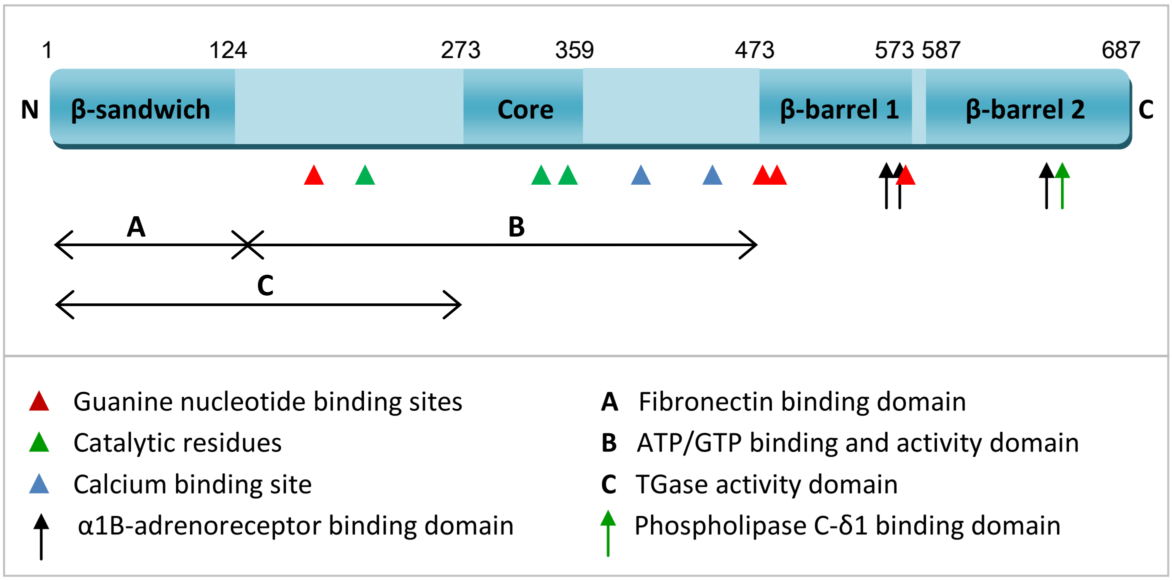

1
124
273
359
473
573 587
687
N
β-sandwich
Core
β-barrel 1
β-barrel 2
C
A
B
C
▲ Guanine nucleotide binding sites
▲ Catalytic residues
▲ Calcium binding site
↑ α1B-adrenoreceptor binding domain
A Fibronectin binding domain
B ATP/GTP binding and activity domain
C TGase activity domain
↑ Phospholipase C-δ1 binding domain

**Figure 6.2. Schematic representation of the known catalytic sites and binding domains of the full length TGM2 protein**

The Guanine nucleotide binding sites are positioned as follows; K173, F174, R478, V479, M483, R580 and Y583. Catalytic active sites are at C227, H335 and D358, the active site region common to all TGases is Y274-W278. Calcium binding sites are N398, D400, E447 and E452. The α1B-adrenoreceptor binding domains are L547-I561, R564-D581, Q633-E646 and phospholipase C-δ1 binding domain V665-K672 (Chen and Mehta, 1999). The activity and binding domains are **A)** The fibronectin domain overlapping with the TGase activity site at amino-acid position 1-272; **B)** The ATPase/GTPase activity domain and ATP/GTP binding domain located at 138-471 and **C)** TGase activity domain positioned at 1-138.

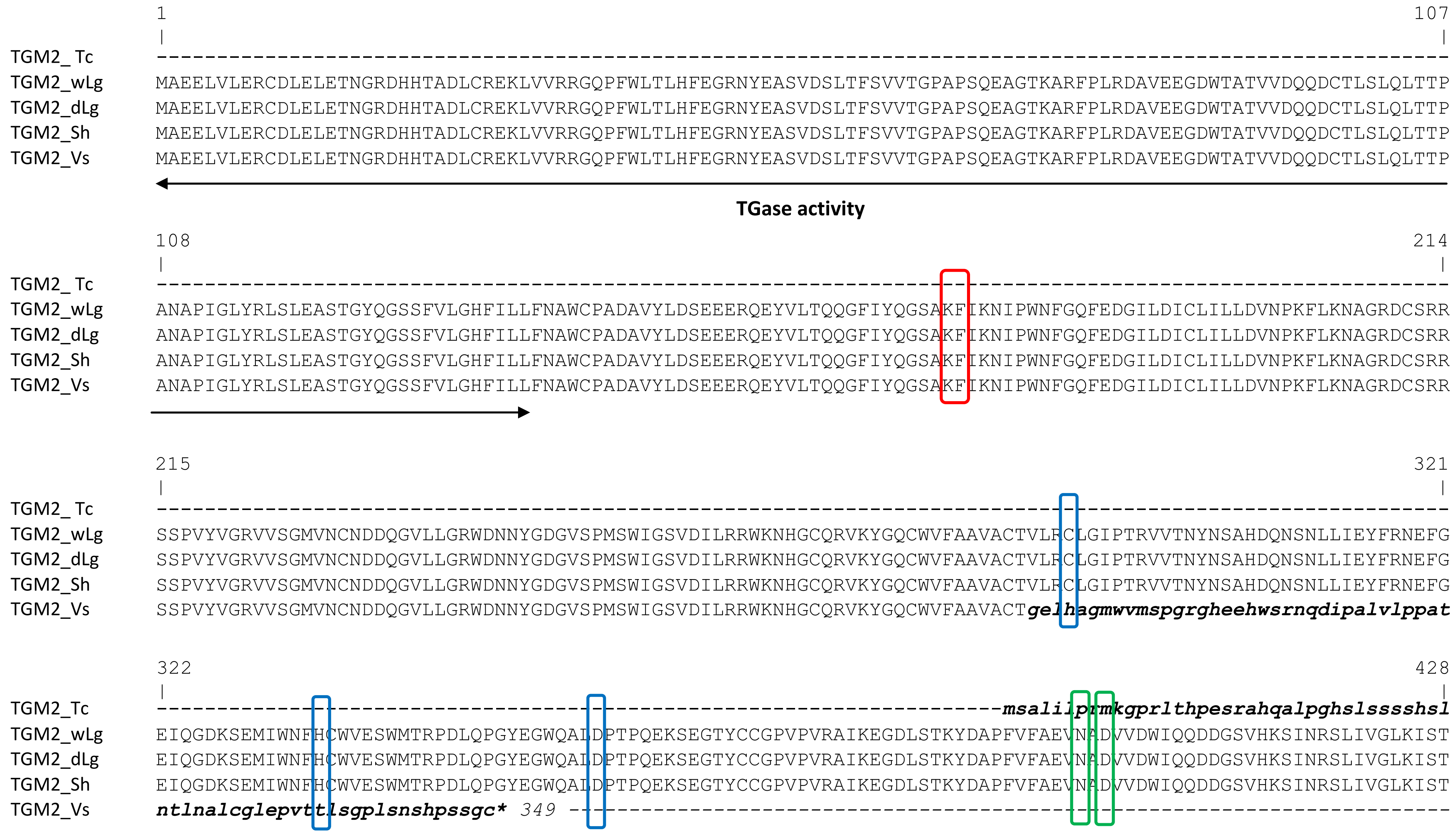
Following the clarification of all known binding and activity domains, it was possible to determine the potential impact on the function of the splice variants of TGM2 (table 6.2; figure 6.3). It appears that all of the transcripts, with the exception of TGM2_Tc may have the ability to bind fibronectin and possess varying levels of TGase activity due to the absence of catalytic residues and disruption of the GTP pocket. Both TGM2_dLg and TGM2_Sh contain all of the catalytic active sites, TGM2_Vs however maintains only the active site common in all TGase genes at C227. The transcripts will also have varying levels of G-protein activity as only TGM2_wLg and TGM2_dLg contain all of the known guanine-binding residues. TGM2 has been reported to function as a protein kinase, with the activity modulated by ATP and $Ca^{2+}$ (Mishra *et al.,* 2007). The exact phosphorylation sites of TGM2 remain to be determined and the kinase activity of the TGM2 isoforms can not be predicted.

Of the five TGM2 isoforms, only the RefSeq transcript (TGM2_wLg) and TGM2_Tc contain complete canonical binding domains for PLC-δ1 and α1-AR. TGM2_dLG is similar to the full length TGM2 in composition, as it contains 13 exons; however due to atypical splicing at the junction of exons 12/13 it lacks the PLC-δ1-binding domains, and maintains only one of the three α1-AR-binding regions, so may display a reduction in binding capacity or even fail to bind altogether. Both of the truncated read-through transcripts, TGM2_Vs and TGM2_Sh will lack the capacity to interact with both PLC-δ1 and α1-AR due to transcription terminating prior to reaching the known binding sequences. TGM2 possesses a further known binding domain for GPR56 (Xu and Hynes, 2007). The specific binding sequence is yet to be determined however the N-terminal end of GPR56 has been shown to interact with the C-terminus of TGM2, via the β-barrel domains. Xu and Hynes (2007) noted that TGM2 constructs containing just the β-barrels were sufficient to bind to GPR56, so it is likely that in addition to the full length peptide (TGM2_wLg ), the protein encoded by the TGM2_Tc isoform would also bind to GPR56 (Xu and Hynes*,* 2007). The novel polypeptide sequence encoded by TGM2_Tc is 147aa in length. When this novel protein sequence was screened against the Pfam database, no homology to known domains was detected. It is possible that the sequence may contain yet unidentified motifs, though unlikely as the novel sequence is a result of exon skipping and not a separate exon.

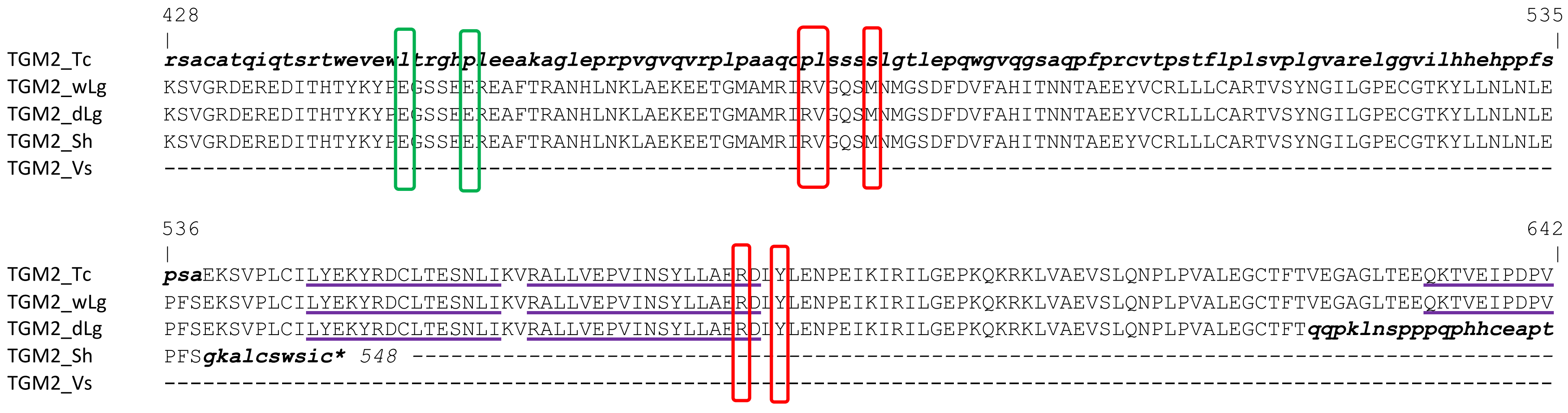
| Transcript details | | | | MIR elements | | | | Activity and binding domains | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Transcript name** | **Accession** | **Size (bp)** | **Size (aa)** | **MIR type** | **mRNA position** | **MIR** | **MIR position** | **TGase** | **Ca$^{2+}$ Activity** | **GTP/ ATPase** | **FN1** | **α1-AR** | **PLC-δ1** |
| TGM2_wLg | NM_004613 | 3937 | 687 | MIRb<br>MIRc<br>MIR3 | 2724 2859<br>2835 2984<br>3050 3222 | 256 116<br>137 6<br>17 203 | 3- UTR<br>3- UTR<br>3- UTR | + | + | + | + | + | + |
| TGM2_dLg | Appendix* | 2238* | 674 | MIRc<br>MIR3 | 1975 2074<br>2140 2312 | 137 6<br>17 203 | TAG<br>3- UTR | + | + | + | + | Q633-E646 only | - |
| TGM2_Sh | NM_198951 | 1879 | 548 | MIR | 1812 1879 | 251 184 | 3- UTR | + | + | Lacks R580 and Y583 | + | - | - |
| TGM2_Vs | CR604340 | 1476 | 349 | MIR | 1203 1333 | 47 180 | 3- UTR | + | C277 only | K173 and F174 only | + | - | - |
| TGM2_Tc | AK126508 | 3465 | 296 | MIR<br>MIR<br>MIRb<br>MIR<br>MIRb<br>MIRc<br>MIR3 | 391 488<br>614 832<br>948 1087<br>1099 1191<br>2270 2405<br>2381 2530<br>2596 2768 | 137 40<br>28 261<br>142 1<br>28 121<br>256 116<br>137 6<br>17 203 | 5- UTR<br>ATG<br>CDS<br>CDS<br>3- UTR<br>3- UTR<br>3- UTR | - | - | R580 and Y583 only | - | + | + |

**Table 6.2. Catalytic and binding domains for the human splicoforms of TGM2**

The transcript details including the accession numbers and transcript and peptide size have been supplied. * The transcript is a composite sequence predicted using previous publication data (Appendix 9.1; Lai *et al.,* 2007), transcript sizes which are predicted as the nucleotide sequence may not be full length as commonly the tail ends of UTRs are lacking following cloning. The MIR sub-family, position in the mRNA and the coordinated of the MIR consensus sequence has been included along with the region of the transcript which it resides. The presence of the activity domains and binding domains are also indicated as present (+) or lacking (-). Abbreviations as follows: ATG, initiating methionine; TAG, stop codon; UTR, untranslated region; CDS, protein coding sequence; FN1, fibronectin; α1-AR, α1-adrenoreceptor binding; PLC- δ1, phospholipase C-δ1 interaction.



```
           1                                                                                                        107
           |                                                                                                          |
TGM2_ Tc   -----------------------------------------------------------------------------------------------------------
TGM2_wLg   MAEELVLERCDLELETNGRDHHTADLCREKLVVRRGQPFWLTLHFEGRNYEASVDSLTFSVVTGPAPSQEAGTKARFPLRDAVEEGDWTATVVDQQDCTLSLQLTTP
TGM2_dLg   MAEELVLERCDLELETNGRDHHTADLCREKLVVRRGQPFWLTLHFEGRNYEASVDSLTFSVVTGPAPSQEAGTKARFPLRDAVEEGDWTATVVDQQDCTLSLQLTTP
TGM2_Sh    MAEELVLERCDLELETNGRDHHTADLCREKLVVRRGQPFWLTLHFEGRNYEASVDSLTFSVVTGPAPSQEAGTKARFPLRDAVEEGDWTATVVDQQDCTLSLQLTTP
TGM2_Vs    MAEELVLERCDLELETNGRDHHTADLCREKLVVRRGQPFWLTLHFEGRNYEASVDSLTFSVVTGPAPSQEAGTKARFPLRDAVEEGDWTATVVDQQDCTLSLQLTTP

                                                      TGase activity

           108                                                                                                      214
           |                                                                                                          |
TGM2_ Tc   -----------------------------------------------------------------------------------------------------------
TGM2_wLg   ANAPIGLYRLSLEASTGYQGSSFVLGHFILLFNAWCPADAVYLDSEEERQEYVLTQQGFIYQGSAKFIKNIPWNFGQFEDGILDICLILLDVNPKFLKNAGRDCSRR
TGM2_dLg   ANAPIGLYRLSLEASTGYQGSSFVLGHFILLFNAWCPADAVYLDSEEERQEYVLTQQGFIYQGSAKFIKNIPWNFGQFEDGILDICLILLDVNPKFLKNAGRDCSRR
TGM2_Sh    ANAPIGLYRLSLEASTGYQGSSFVLGHFILLFNAWCPADAVYLDSEEERQEYVLTQQGFIYQGSAKFIKNIPWNFGQFEDGILDICLILLDVNPKFLKNAGRDCSRR
TGM2_Vs    ANAPIGLYRLSLEASTGYQGSSFVLGHFILLFNAWCPADAVYLDSEEERQEYVLTQQGFIYQGSAKFIKNIPWNFGQFEDGILDICLILLDVNPKFLKNAGRDCSRR

           215                                                                                                      321
           |                                                                                                          |
TGM2_ Tc   -----------------------------------------------------------------------------------------------------------
TGM2_wLg   SSPVYVGRVVSGMVNCNDDQGVLLGRWDNNYGDGVSPMSWIGSVDILRRWKNHGCQRVKYGQCWVFAAVACTVLRCLGIPTRVVTNYNSAHDQNSNLLIEYFRNEFG
TGM2_dLg   SSPVYVGRVVSGMVNCNDDQGVLLGRWDNNYGDGVSPMSWIGSVDILRRWKNHGCQRVKYGQCWVFAAVACTVLRCLGIPTRVVTNYNSAHDQNSNLLIEYFRNEFG
TGM2_Sh    SSPVYVGRVVSGMVNCNDDQGVLLGRWDNNYGDGVSPMSWIGSVDILRRWKNHGCQRVKYGQCWVFAAVACTVLRCLGIPTRVVTNYNSAHDQNSNLLIEYFRNEFG
TGM2_Vs    SSPVYVGRVVSGMVNCNDDQGVLLGRWDNNYGDGVSPMSWIGSVDILRRWKNHGCQRVKYGQCWVFAAVACTgelhagmwvmspgrgheehwsrnqdipalvlppat

           322                                                                                                      428
           |                                                                                                          |
TGM2_Tc    ----------------------------------------------------------------------msalilprmkgprlthpesrahqalpghslsssshsl
TGM2_wLg   EIQGDKSEMIWNFHCWVESWMTRPDLQPGYEGWQALDPTPQEKSEGTYCCGPVPVRAIKEGDLSTKYDAPFVFAEVNADVVDWIQQDDGSVHKSINRSLIVGLKIST
TGM2_dLg   EIQGDKSEMIWNFHCWVESWMTRPDLQPGYEGWQALDPTPQEKSEGTYCCGPVPVRAIKEGDLSTKYDAPFVFAEVNADVVDWIQQDDGSVHKSINRSLIVGLKIST
TGM2_Sh    EIQGDKSEMIWNFHCWVESWMTRPDLQPGYEGWQALDPTPQEKSEGTYCCGPVPVRAIKEGDLSTKYDAPFVFAEVNADVVDWIQQDDGSVHKSINRSLIVGLKIST
TGM2_Vs    ntlnalcglepvttlsgplsnshpssgc* 349 -------------------------------------------------------------------------
```



```
              428                                                                                             535
              |                                                                                                 |
TGM2_Tc       rsacatqiqtsrtwevewltrghpleeakagleprpvgvqvrplpaaqcplsssslgtlepqwgvqgsaqpfprcvtpstflplsvplgvarelggvilhhehppfs
TGM2_wLg      KSVGRDEREDITHTYKYPEGSSEEREAFTRANHLNKLAEKEETGMAMRIRVGQSMNMGSDFDVFAHITNNTAEEYVCRLLLCARTVSYNGILGPECGTKYLLNLNLE
TGM2_dLg      KSVGRDEREDITHTYKYPEGSSEEREAFTRANHLNKLAEKEETGMAMRIRVGQSMNMGSDFDVFAHITNNTAEEYVCRLLLCARTVSYNGILGPECGTKYLLNLNLE
TGM2_Sh       KSVGRDEREDITHTYKYPEGSSEEREAFTRANHLNKLAEKEETGMAMRIRVGQSMNMGSDFDVFAHITNNTAEEYVCRLLLCARTVSYNGILGPECGTKYLLNLNLE
TGM2_Vs       -----------------------------------------------------------------------------------------------------------

              536                                                                                             642
              |                                                                                                 |
TGM2_Tc       psaEKSVPLCILYEKYRDCLTESNLIKVRALLVEPVINSYLLAERDIYLENPEIKIRILGEPKQKRKLVAEVSLQNPLPVALEGCTFTVEGAGLTEEQKTVEIPDPV
TGM2_wLg      PFSEKSVPLCILYEKYRDCLTESNLIKVRALLVEPVINSYLLAERDIYLENPEIKIRILGEPKQKRKLVAEVSLQNPLPVALEGCTFTVEGAGLTEEQKTVEIPDPV
TGM2_dLg      PFSEKSVPLCILYEKYRDCLTESNLIKVRALLVEPVINSYLLAERDIYLENPEIKIRILGEPKQKRKLVAEVSLQNPLPVALEGCTFTqqpklnspppqphhceapt
TGM2_Sh       PFSgkalcswsic* 548 ----------------------------------------------------------------------------------------
TGM2_Vs       -----------------------------------------------------------------------------------------------------------
```

```
              643                                         687
              |                                             |
TGM2_Tc       EAGEEVKVRMDLLPLHMGLHKLVVNFESDKLKAVKGFRNVIIGPA*
TGM2_wLg      EAGEEVKVRMDLLPLHMGLHKLVVNFESDKLKAVKGFRNVIIGPA*
TGM2_dLg      mcwvpptlagahgasnpfnhhgkllwallpg* 674 ---------
TGM2_Sh       ---------------------------------------------
TGM2_Vs       ---------------------------------------------
```

| | | | |
|---|---|---|---|
| 🟦 | Calcium binding | KLV | α1- AR binding domain |
| 🟩 | Catalytic site | KLV | PLC-δ1 binding domain |
| 🟥 | Guanine binding site | ***klv*** | Novel peptide sequence |

**Figure 6.3. Peptide sequences of the TGM2 isoforms**

The amino acid marker positions are included for each row as a reference when determining the positions of active sites and binding domains. Novel peptide sequences which are generated by the incorporation of intronic sequences, as a result of donor splicing being skipped are in bold italics. Binding sites and catalytic sites have been outlined as noted in the legend. All transcripts with the exception of TGM2_Tc bind fibronectin and possess TGase activity. TGM2_Tc contains only one of the GTP binding regions and has the ability to bind with PLC-δ1 and α1-AR. TGM2_dLg contains all binding and activity domains with the exception of one α1-AR binding region and lacks PLC-δ1 binding capacity. TGM2_Vs and TGM2_Sh both lack the PLC-δ1 and α1-AR binding domains and have restricted calcium binding and GTP/ATPase activity. Sequence conservation was determined using the ClustalW2 alignment tool (accession numbers appendix 9.3).

## 6.4. Splicing and stop codons generated by MIR elements for TGM2

Given that all of the TGM2 isoforms contain MIR elements the role of the repeats in the splicing and expression of the TGM2 transcripts was investigated. Translation of both TGM2_dLg and TGM2_Tc terminates at a stop codon located within an exaptated MIR element. TGM2_Tc contains seven MIR elements in all, and in addition to the stop codon the initiating methionine codon is carried within an MIR sequence. The alternative methionine sequence is undiverged from the original MIR consensus.

TGM2 genomic sequences were identified from a range of mammal to determine if TGM2_Tc is likely to be expressed in other species. Multiple cross-species alignments revealed that TGM2_Tc may be primate specific and a low level of conservation between the corresponding human TGM2 genomic sequence and other non-primate mammals is apparent (figure 6.5). However the information is limited to the sequence data available, and only rodents, dog, cow and opossum can be compared. The novel nucleotide sequence of TGM2_Tc is 100% conserved between human and chimpanzee, which is to be expected, and the methionine sequence is conserved and in-frame for the other two primate species studied within the order hominidae; gorilla and orangu-tan (figure 6.6). The complete conservation of TGM2_Tc does not extend to earlier primate orders such as the old world monkeys. For example, when comparing the genomic sequence of the rheusus macaque the methionine codon is conserved but is no longer in-frame, due to an indel generating a frame shift at position 997nt (figure 6.7b-c). A similar TGM2_Tc transcript may be present in the macaque genome however, as there is an in-frame methionine sequence, which conforms to the vertebrate Kozak consensus, located 32bp upstream of the methionine observed in the great apes (figure 6.7a).

Comparison of the sequences flanking the initiating methionine codon for TGM2_Tc to that of the Kozak consensus sequence (ACCATGG) indicates that the initiation site may be weak; however this appears to be typical of methionines located within MIR elements (appendix 9.8). To confirm this observation the MIR sequences containing the initiating methionine codons were aligned using WebLogo, revealing that overall the nucleotide sequence conforms to the Kozak consensus (figure 6.4) but it is not uncommon for the flanking 3bp upstream and downstream sequences to differ.

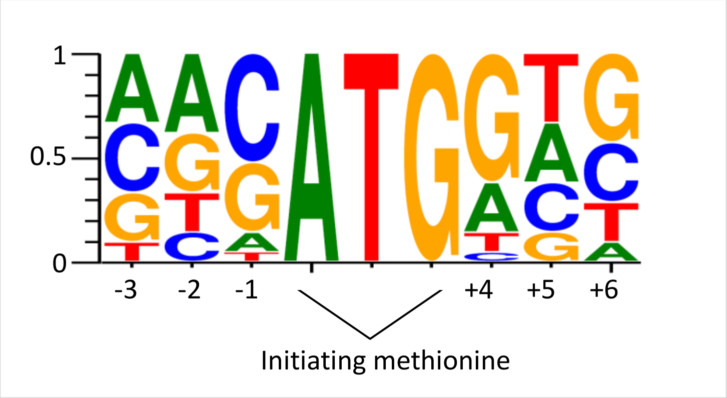


**Figure 6.4. Consensus sequence surrounding initiating methionines provided by MIR elements**

The image demonstrated the probability of a nucleotide being present for each position between -3 to +6. The image was created by aligning 21 MIR sequences which contain start codons using WebLogo. The height of the stacked letters represents the frequency of each nucleotide. The top letter in the alignment is most conserved (Crooks *et al.,* 2004).

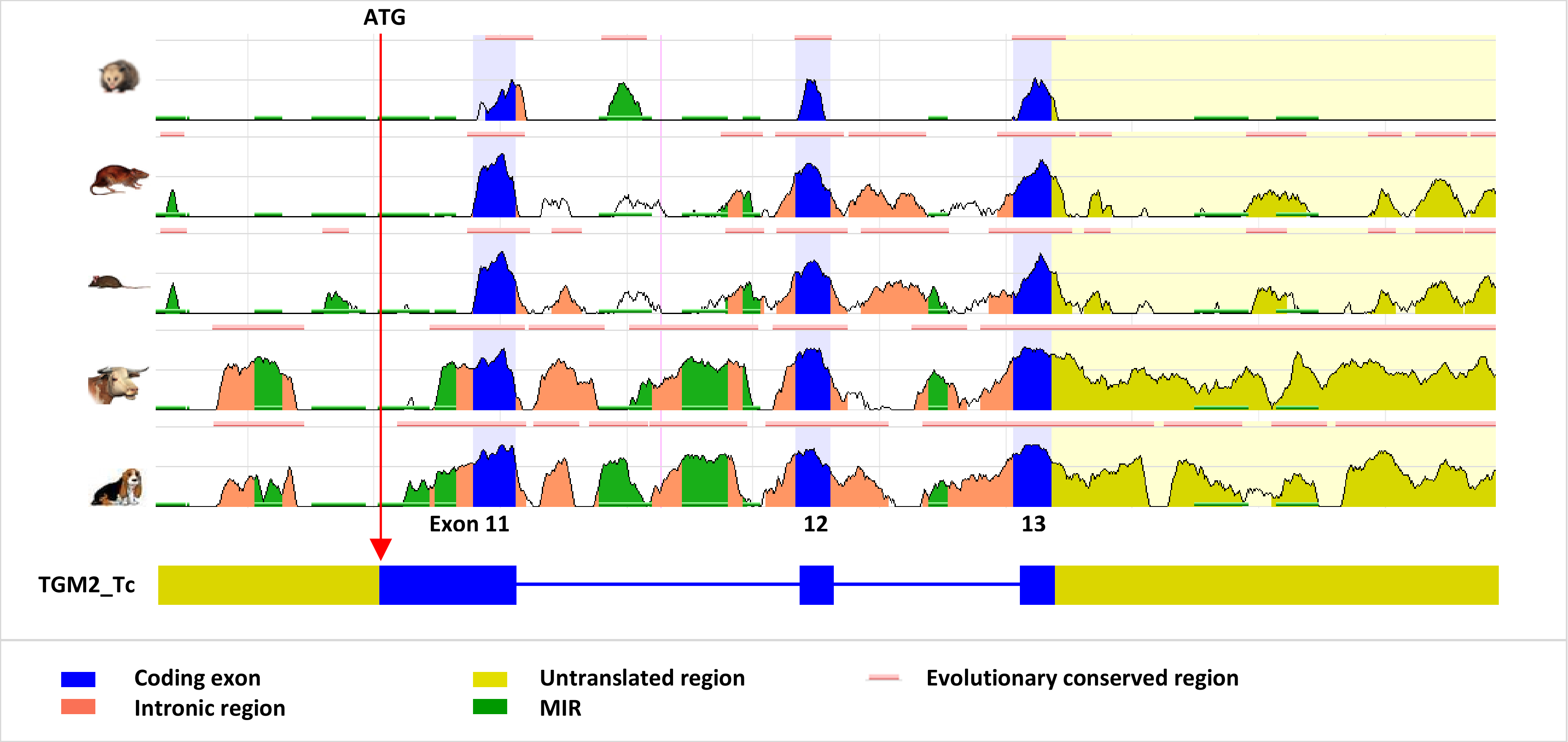


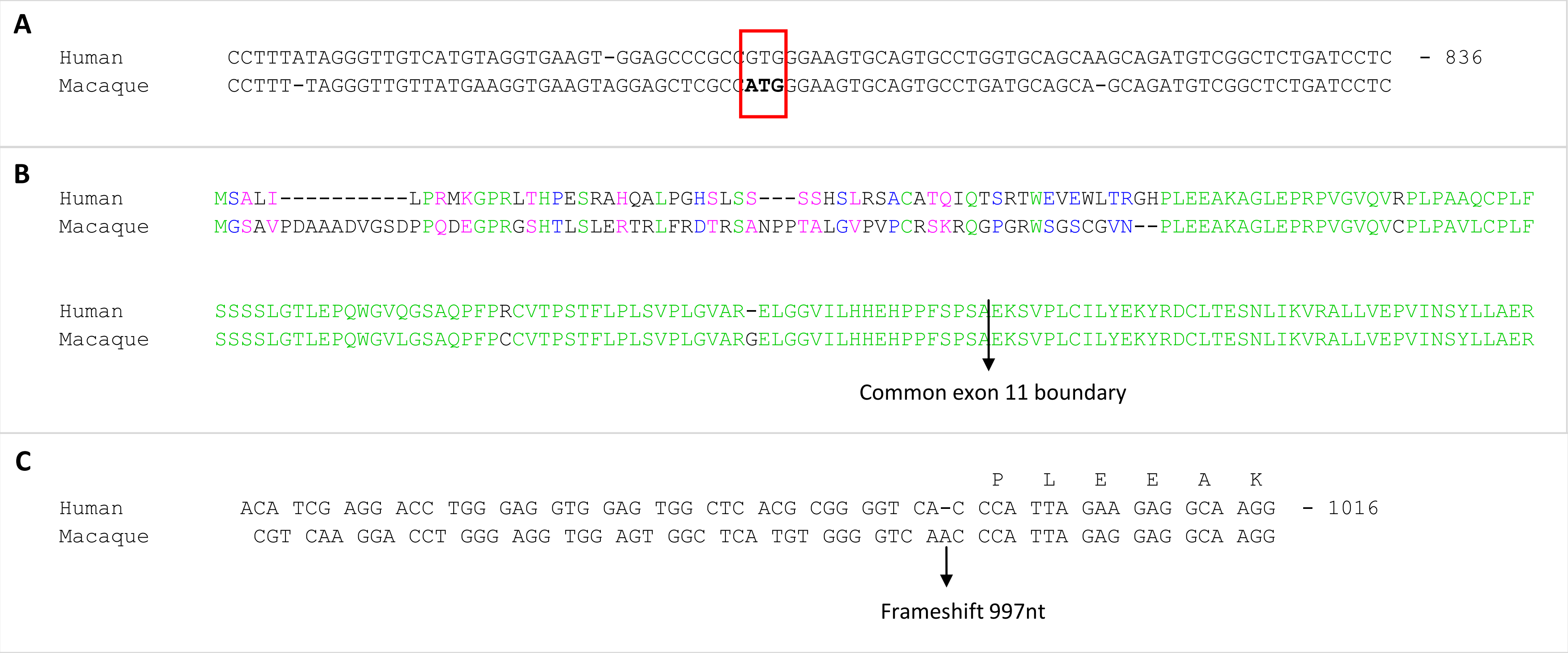
**Figure 6.5. Evolutionary conserved regions of a number of mammalian species for TGM2_Tc**

Image adapted from the ECR browser (http://ecrbrowser.dcode.org/) which displays evolutionary conserved genomic regions. The RefSeq coding exons have been numbered, 11-13. The exonic composition of TGM2_Tc is included as a reference. The species included are dog, cow, rat, mouse and opossum. The position of the initiating methionine (in the human transcript) has been included. The figure demonstrates the lack of conservation for the novel TGM2_Tc region for a number of mammals

```
Human         GGGAAGTGCAGTGCCTGGTGCAGCAAGCAGATGTCGGCTCTGATCCTCCCCAGGATGAAGGG-CCCGCGGCTCACACACCCTGAGTC  - 847
Gorilla       NNNNNNNNNNNNNNNNNNNNNNNNNNNNAGATGTCGGCTCTGATCCTCCCCAGGATGAAGGG-CCCGCGGCTCACACACCCTGAGTC
Chimpanzee    GGGAAGTGCAGTGCCTGGTGCAGCAAGCAGATGTCGGCTCTGATCCTCCCCAGGATGAAGGG-CCCGCGGCTCACACACCCTGAGTC
Orang-utan    NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNCCTCCCCAGGATGAAGGG-CCGGCGGCTCACACACCCCGAGTC
Macaque       GGGAAGTGCAGTGCCTGATGCAGCA-GCAGATGTCGGCTCTGATCCTCCCCAGGATGAAGGGCCCCGCGGCTCACACACCCTGAGTC


Human         CCGAGCGCACCAGGCTCTTCCGGGACACTCGCTCAGCTCATCCTCCCACAGCCTTAGGAGTGCCTGTGCCACGCAGATCCAAACAT   - 963
Gorilla       CCGAGCGCACCAGGCTCTTCCGGGACACTCGCTCAGCTCATCCTCCCACAGCCTTAGGAGTGCCTGTGCCACGNNNNNNNNNNNNN
Chimpanzee    CCGAGCGCACCAGGCTCTTCCGGGACACTCGCTCAGCTCATCCTCCCACAGCCTTAGGAGTGCCTGTGCCACGCAGATCCAAACAT
Orang-utan    CCGAGCACACCAGGCTCTTCCGGGACACTCGCTCAGCTCATCCTCCCACAGCCTTAGGAGTGCCTGTGCCANNNNNNNNNNNNNNN
Macaque       TGGAGCGCACCAGGCTCTTCCGGGACACTCGCTCAGCTAATCCTCCCACAGCCTTAGGAGTGCCTGTGCCATGCAGATCCAAACAT
```

**Figure 6.6. Multiple sequence alignment of the methionine containing region of TGM2_Tc for a number of primates**

Completely conserved sequences are in green, identical residues in pink, similar residues in blue and different residues in black. N signifies that the sequence was unobtainable in current databases. The ATG sequence, boxed red, is conserved in all species however in the macaque the codon is not in frame and therefore non-functional. Sequence conservation was determined using the ClustalW2 alignment tool (http://www.ebi.ac.uk/Tools/clustalw2/index.html; accession numbers appendix 9.3).



```
Human      CCTTTATAGGGTTGTCATGTAGGTGAAGT-GGAGCCCGCCGTGGGAAGTGCAGTGCCTGGTGCAGCAAGCAGATGTCGGCTCTGATCCTC  - 836
Macaque    CCTTT-TAGGGTTGTTATGAAGGTGAAGTAGGAGCTCGCCATGGGAAGTGCAGTGCCTGATGCAGCA-GCAGATGTCGGCTCTGATCCTC
```

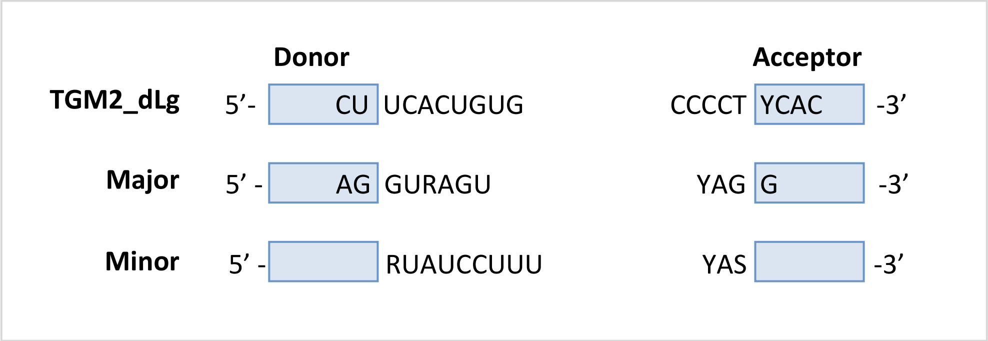
```
Human      MSALI-----------LPRMKGPRLTHPESRAHQALPGHSLSS---SSHSLRSACATQIQTSRTWEVEWLTRGHPLEEAKAGLEPRPVGVQVRPLPAAQCPLF
Macaque    MGSAVPDAAADVGSDPPQDEGPRGSHTLSLERTRLFRDTRSANPPTALGVPVPCRSKRQGPGRWSGSCGVN--PLEEAKAGLEPRPVGVQVCPLPAVLCPLF

Human      SSSSLGTLEPQWGVQGSAQPFPRCVTPSTFLPLSVPLGVAR-ELGGVILHHEHPPFSPSAEKSVPLCILYEKYRDCLTESNLIKVRALLVEPVINSYLLAER
Macaque    SSSSLGTLEPQWGVLGSAQPFPCCVTPSTFLPLSVPLGVARGELGGVILHHEHPPFSPSAEKSVPLCILYEKYRDCLTESNLIKVRALLVEPVINSYLLAER
```

```
                                                                 P   L   E   E   A   K
Human       ACA TCG AGG ACC TGG GAG GTG GAG TGG CTC ACG CGG GGT CA-C CCA TTA GAA GAG GCA AGG  - 1016
Macaque      CGT CAA GGA CCT GGG AGG TGG AGT GGC TCA TGT GGG GTC AAC CCA TTA GAG GAG GCA AGG
```

**Figure 6.7. Multiple sequence alignment of TGM2_Tc for human and macaque**

Completely conserved amino acids are in green, identical residues in pink, similar residues in blue and different residues in black. **A)** Alternative putative initiating methionine (boxed red) for the macaque which is not conserved in the human sequence. **B)** The peptide sequence is conserved until residue P61; the region where canonical exon 11 begins has been provided. **C)** Nucleotide triplet sequences for human and macaque outlining the region at nucleotide 997 where the frameshift occurs. Sequence conservation was determined using the ClustalW2 alignment tool (http://www.ebi.ac.uk/Tools/clustalw2/index.html; accession numbers appendix 9.3).

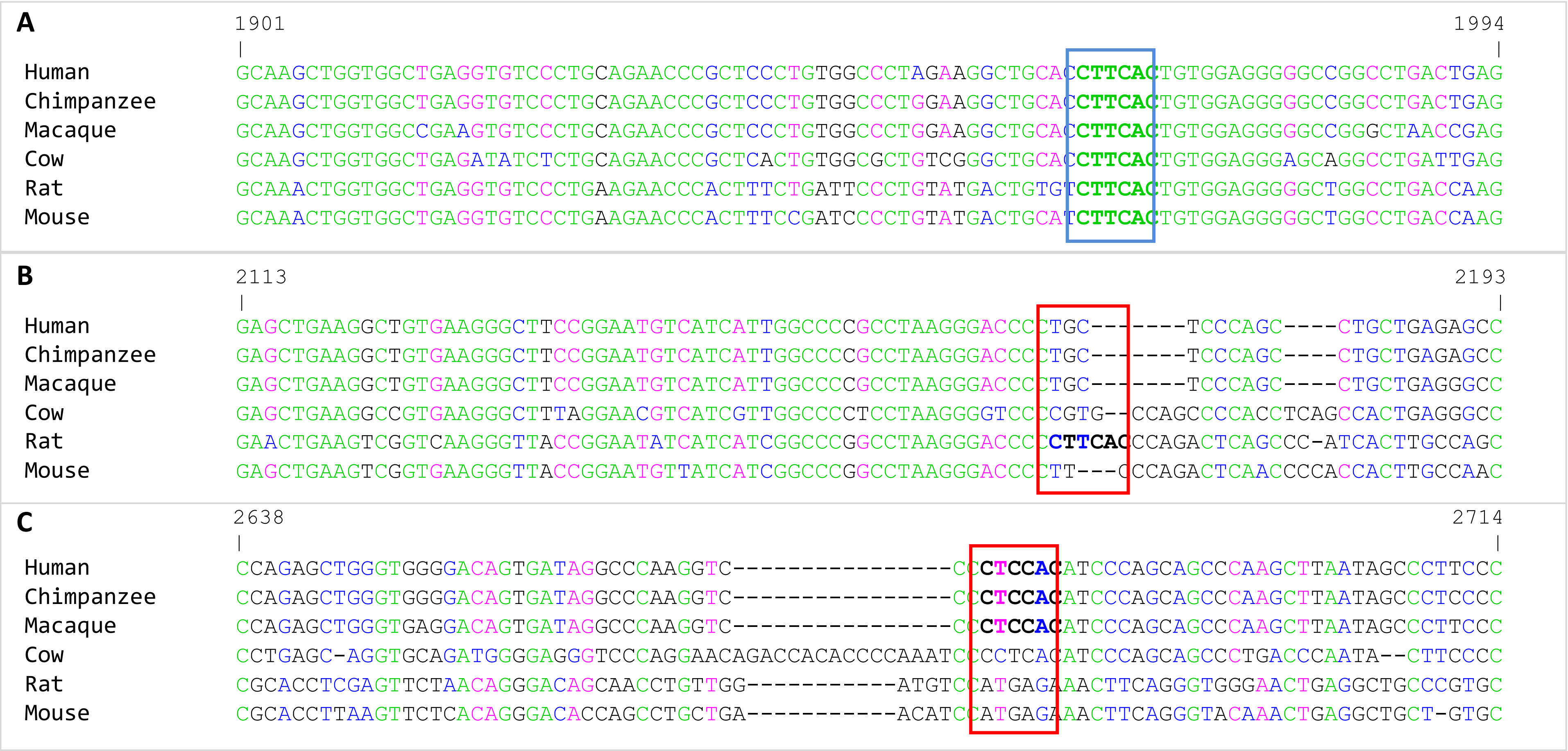
The termination codon of the TGM2_dLg transcript is situated within an MIR element (table 6.2). The same MIR is within the 3'-UTR of TGM2_wLg, but following the altered splicing an alternative reading frame is in use, allowing for translation to continue into the MIR sequence which contains the stop codon. The alternative C-terminal is made up of 51 novel amino acids. TGM2_dLg is generated through the use of atypical alternative acceptor and donor splice sites located in RefSeq exons 12 and 13. There are two known modes of splicing occurring in vertebrates which differ in the splice site consensus sequence, with the major spliceosome splicing at AG-GT. Splicing of TGM2_dLg occurs at the consensus CTYCAC, which is not a conventional major or minor splice site (figure 6.8). Previous studies have confirmed the expression of the isoform in human and rat tissues (Lai *et al.,* 2007; Tolentino *et al.,* 2004; Monsonego *et al.,* 1997). Both the acceptor and donor splicing occurs within a CTYCAC motif and as such it is not possible to distinguish the exact position where splicing occurs, although a putative sequence can be determined. Multiple sequence alignments were generated with the region containing both the atypical acceptor and donor splice sites for a number of mammalian species, to reveal the conservation of the atypical splicing (figure 6.9). The donor splice site was determined as (CUucacugug) and shown to be conserved in all mammals studied, with complete identity of the subsequent nine intronic nucleotides. The acceptor splice site described in human (ctCCAC) was only detected in primate species with a different 3' region noted in rats, both of which are noted to be proceeded by three cytosine nucleotides (figure 6.8 and 6.9). Surprisingly no similar sequence was detected in mice.



| | Donor | | | Acceptor | |
|---|---|---|---|---|---|
| **TGM2_dLg** | 5'- CU | UCACUGUG | CCCCT | YCAC | -3' |
| **Major** | 5' - AG | GURAGU | YAG | G | -3' |
| **Minor** | 5' - | RUAUCCUUU | YAS | | -3' |

**Figure 6.8. Comparison of major splicing, minor splicing and the CTYCAC motif of TGM2_dLg**

The putative non-canonical splice site of TGM2_dLg has been determined for primate and rat. Crucial exonic sequences are boxed blue and conserved intronic sequence outside the box. Major (U12 type intron) and minor (U2 type intron) splice site sequences have been included as a reference (Will and Lührmann, 2005). Fasta code Y = T / C; R = A / G; S = C / G.



```
A               1901                                                                                   1994
                |                                                                                         |
Human           GCAAGCTGGTGGCTGAGGTGTCCCTGCAGAACCCGCTCCCTGTGGCCCTAGAAGGCTGCACCTTCACTGTGGAGGGGGCCGGCCTGACTGAG
Chimpanzee      GCAAGCTGGTGGCTGAGGTGTCCCTGCAGAACCCGCTCCCTGTGGCCCTGGAAGGCTGCACCTTCACTGTGGAGGGGGCCGGCCTGACTGAG
Macaque         GCAAGCTGGTGGCCGAAGTGTCCCTGCAGAACCCGCTCCCTGTGGCCCTGGAAGGCTGCACCTTCACTGTGGAGGGGGCCGGGCTAACCGAG
Cow             GCAAGCTGGTGGCTGAGATATCTCTGCAGAACCCGCTCACTGTGGCGCTGTCGGGCTGCACCTTCACTGTGGAGGGAGCAGGCCTGATTGAG
Rat             GCAAACTGGTGGCTGAGGTGTCCCTGAAGAACCCACTTTCTGATTCCCTGTATGACTGTGTCTTCACTGTGGAGGGGGCTGGCCTGACCAAG
Mouse           GCAAACTGGTGGCTGAGGTGTCCCTGAAGAACCCACTTTCCGATCCCCTGTATGACTGCATCTTCACTGTGGAGGGGGCTGGCCTGACCAAG

B               2113                                                                             2193
                |                                                                                   |
Human           GAGCTGAAGGCTGTGAAGGGCTTCCGGAATGTCATCATTGGCCCCGCCTAAGGGACCCCTGC-------TCCCAGC----CTGCTGAGAGCC
Chimpanzee      GAGCTGAAGGCTGTGAAGGGCTTCCGGAATGTCATCATTGGCCCCGCCTAAGGGACCCCTGC-------TCCCAGC----CTGCTGAGAGCC
Macaque         GAGCTGAAGGCTGTGAAGGGCTTCCGGAATGTCATCATTGGCCCCGCCTAAGGGACCCCTGC-------TCCCAGC----CTGCTGAGGGCC
Cow             GAGCTGAAGGCCGTGAAGGGCTTTAGGAACGTCATCGTTGGCCCCTCCTAAGGGGTCCCCGTG--CCAGCCCACCTCAGCCACTGAGGGCC
Rat             GAACTGAAGTCGGTCAAGGGTTACCGGAATATCATCATCGGCCCGGCCTAAGGGACCCCCTTCACCCAGACTCAGCC-ATCACTTGCCAGC
Mouse           GAGCTGAAGTCGGTGAAGGGTTACCGGAATGTTATCATCGGCCCGGCCTAAGGGACCCCTT---CCCAGACTCAACCCCACCACTTGCCAAC

C               2638                                                                             2714
                |                                                                                   |
Human           CCAGAGCTGGGTGGGGACAGTGATAGGCCCAAGGTC----------------CCCTCCACATCCCAGCAGCCCAAGCTTAATAGCCCTTCCC
Chimpanzee      CCAGAGCTGGGTGGGGACAGTGATAGGCCCAAGGTC----------------CCCTCCACATCCCAGCAGCCCAAGCTTAATAGCCCTCCCC
Macaque         CCAGAGCTGGGTGAGGACAGTGATAGGCCCAAGGTC----------------CCCTCCACATCCCAGCAGCCCAAGCTTAATAGCCCTTCCC
Cow             CCTGAGC-AGGTGCAGATGGGGAGGGTCCCAGGAACAGACCACACCCCAAATCCCCTCACATCCCAGCAGCCCTGACCCAATA--CTTCCCC
Rat             CGCACCTCGAGTTCTAACAGGGACAGCAACCTGTTGG-----------ATGTCCATGAGAAACTTCAGGGTGGGAACTGAGGCTGCCCGTGC
Mouse           CGCACCTTAAGTTCTCACAGGGACACCAGCCTGCTGA-----------ACATCCATGAGAAACTTCAGGGTACAAACTGAGGCTGCT-GTGC
```

**Figure 6.9. Multiple sequence alignment of the CTYCAC motif of TGM2_dLg**

Completely conserved sequences are in green, identical residues in pink, similar residues in blue and different residues in black. The 'CTTCAC' acceptor Splice region is boxed blue and the donor splice sites red. The human mRNA co-ordinates are included as a reference. Sequence conservation was determined using the ClustalW2 alignment tool (http://www.ebi.ac.uk/Tools/clustalw2/index.html; accession numbers appendix 9.3).

It is possible that the TGM2 isoforms which were detected *in silico* are artefactual and represent errors in splicing; specifically TGM2_dLg and TGM2_Tc, as there are no ESTs in the available databases which support the expression of these transcripts. The mRNA from human and rat tissues were analysed by RT-PCR to confirm the expression of the discussed TGM2 isoforms. Primers were used to amplify TGM2 (figure 6.10), and amplicons were purified, cloned and sequenced. All five TGM2 transcript variants were expressed in all tissues analysed (figure 6.11).

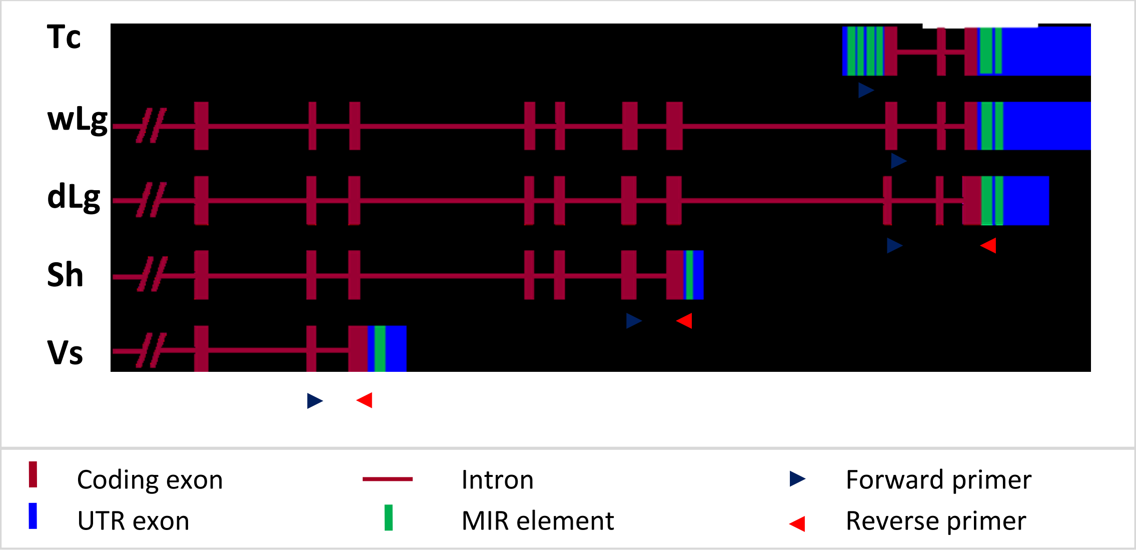


**Figure 6.10. Positions of the primers used in the RT-PCR experiments**

Exonic composition of the splice variants of TGM2 generated using the NCBI sequence alignment tool Spidey (http://www.ncbi.nlm.nih.gov/spidey/). Schematics begin at constitutive exon 4; displaying the MIR positions, and coding and non-coding exons. Primer details are listed in table 2.3 and the position of the forward (blue) and reverse (red) primers are indicated by arrows. The primers of TGM2_dLg and TGM2_wLg contain the splice site at the boundary of exon 12/13 to distinguish between the two splice variants. Transcripts TGM2_Tc, TGM2_wLg and TGM2_dLg use the same reverse primer annealing to the terminal exon.

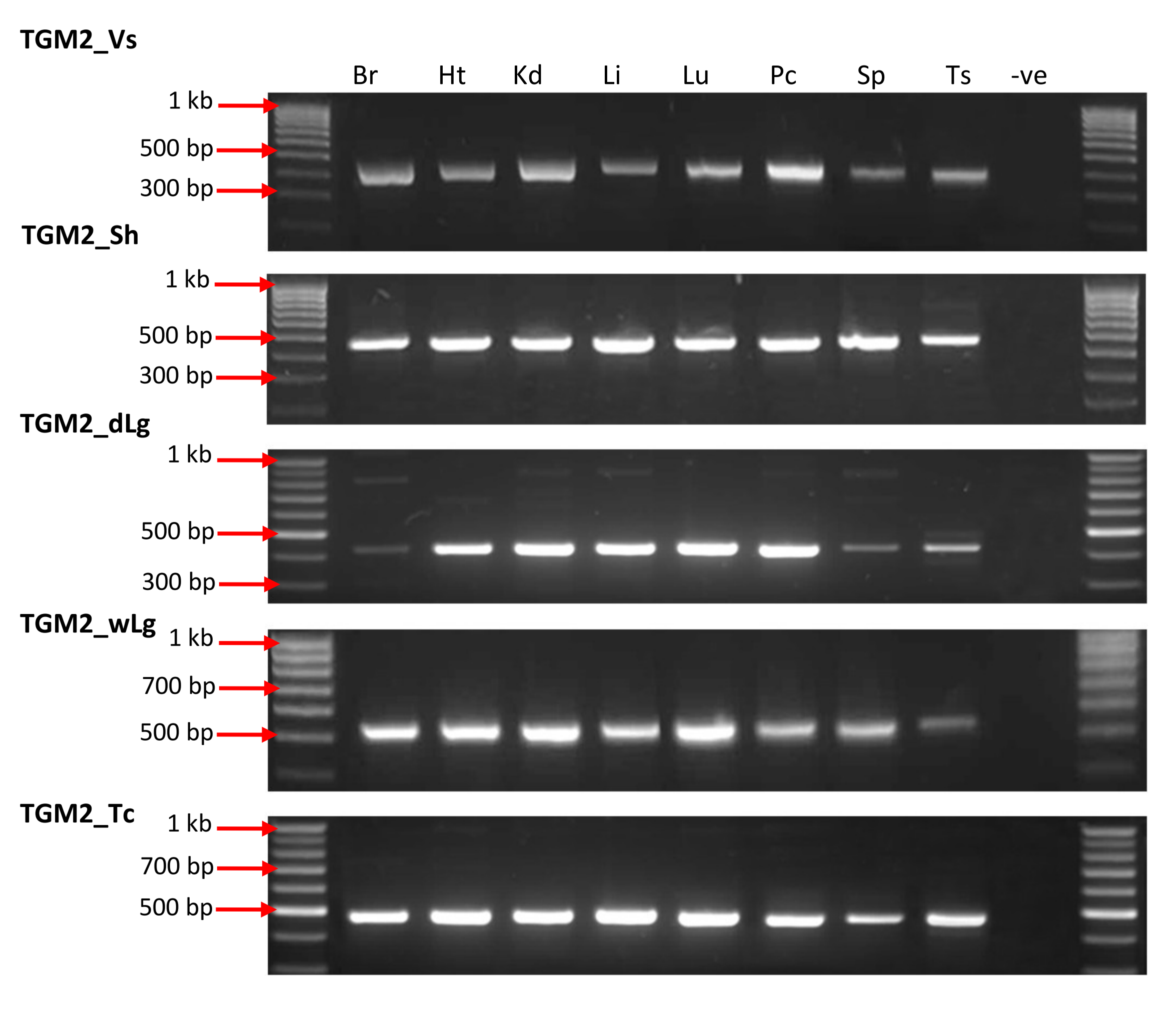


**Figure 6.11. Amplification of the TGM2 isoforms in multiple tissue types**

Amplified fragments of the TGM2 isoforms were analysed on a 2% agarose gel with 1X SybrSafe. Tissues are abbreviated as follows: Br, brain; Ht, heart; Kd, kidney; Li, liver; Lu, lung; Pc, pancreas; Sp, spleen and Ts, testis. TGM2_wLg (514 bp) and TGM2_dLg (414 bp) were amplified from rat cDNA normalised with GAPDH (figure 2.2), using primers designed elsewhere (Tolentino *et al.,* 2002). Published primers were used to ensure that only the targets were amplified, as the primer sequence must contain the CTTCAC motif of TGM2_dLg (figure 6.8) in order to differentiate between the two isoforms. TGM2_Vs (323 bp), TGM2_Sh (481 bp) and TGM2_Tc (482 bp) were amplified from a pre-synthesised human cDNA panel (Clontech; primer details table 2.1). All TGM2 fragments were confirmed by sequencing, DNase and RNase-free water was used as a template in the negative controls.

TGM2 has been implicated in neurodegenerative disease and diabetes (Ruan and Johnson, 2007; Porzio *et al.,* 2007; Bernassola *et al.,* 2002; Mastroberardino *et al.,* 2002), so to further expand the expression profile of TGM2, two rat cell lines were selected for RT-PCR analysis. C6 cells are derived from rat glioma cells (Brenda *et al.*, 1968) and BRIN-BD11 cells are glucosensitive insulin-secreting clonal pancreatic β-cells (Rasschaert *et al* 1996). cDNA was synthesised from C6 cells which were either differentiated or in the mitotic phase. Normal rat pancreatic islet cDNA was used as a comparison to the expression of the insulinoma BRIN-BD11 cells.

Fragments of all known rat TGM2 isoforms were amplified in both mitotic and differentiated C6 cells, BRIN-BD11 cells and rat islets (figure 6.12). Similar concentrations of the TGM2 isoforms were amplified from both mitotic and differentiated C6 cells compared to normal brain tissue (figure 6.11), however with higher levels of TGM2_wLg appearing to be amplified. The BRIN-BD11 cells and rat islets displayed lower levels of TGM2_wLg than expected with TGM2-Sh appearing to be preferentially expressed. This data however is not quantitative and real-time qRT-PCR would need to be performed to confirm these observations.

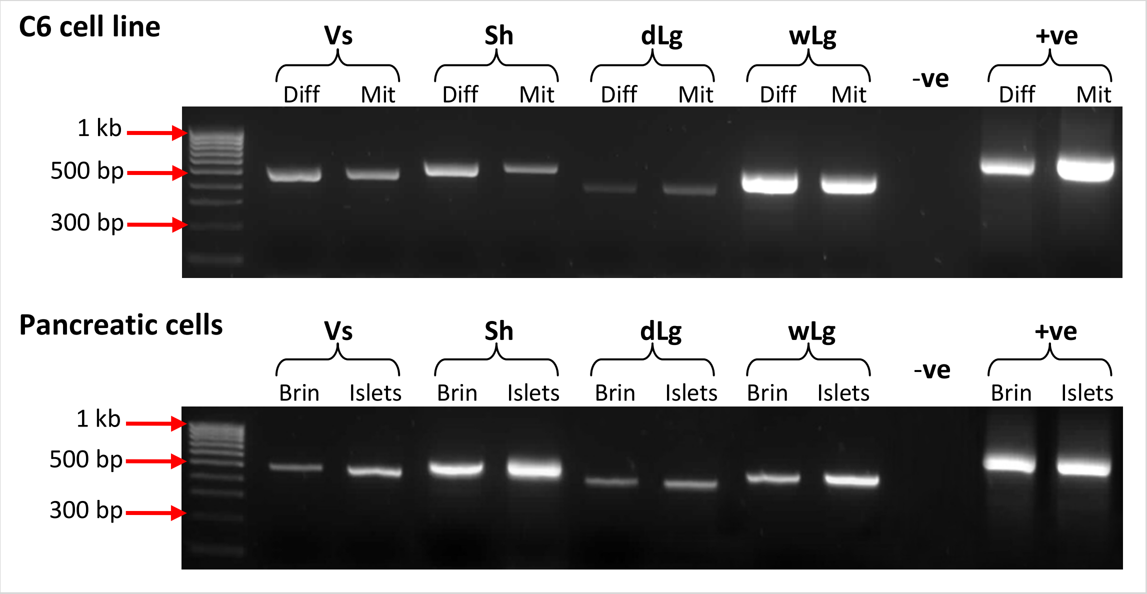


**Figure 6.12. Expression of TGM2 isoforms amplified from rat islets, C6 and BRIN-BD11 cells**

Amplified fragments of the TGM2 isoforms were analysed on a 2% agarose gel with 1X SybrSafe. Abbreviations: Diff, differentiated C6 cells; Mit, C6 cells in mitotic phase; Brin, BRIN-BD11 cells; Islets, normal rat islet of Langerhans. TGM2_wLg (514bp) and TGM2_dLg (414bp) were amplified from rat cDNA using primers designed elsewhere (Tolentino *et al.,* 2002). Primer details and cycling condition for TGM2_Vs (488bp) and TGM2_Sh (544bp) are in table 2.3. All TGM2 fragments were confirmed by sequencing, DNase and RNase-free water was used as a negative control and GAPDH (738bp) as a positive control.

## 6.5. *In situ* expression of the TGM2 isoforms in the adult rat brain

The TGM2 isoforms were successfully detected in the rat brain using RT-PCR; however the spatial distribution of these transcripts cannot be determined by PCR alone. Radioactive *in situ* hybridisation was performed on five adult Wistar rat brains to investigate the distribution of the TGM2 mRNA transcripts.

All four TGM2 isoforms were highly expressed in the corpus callosum and anterior commissure (figure 6.13); white matter regions connecting the left and right cerebral hemispheres, which consist of bundled axonal projections and facilitates communication between the two hemispheres. High levels of TGM2 expression was also detected in the limbic region; specifically the olfactory tract, which is important in odour perception and the hippocampus, including the dendrate gyrus and fimbria; areas fundamental in long-term learning and spatial awareness. TGM2_wLg appears to be expressed at lower levels in the olfactory tract than the other TGM2 isoforms. Finally high levels of expression was visualised in cells of the optic nerve, optic chiasma and optic tract for all four isoforms, consistent with a role in visual perception and memory. There appears to be little or no expression of the four TGM2 splice variants in the hypothalamus, the thalamus, the amygdala and the cuadate putamen. The spatial distribution of the primate-specific TGM2_Tc in the brain is unknown and cannot be compared to the rat *in situ* hybridisation data.

The four rat TGM2 isoforms are expressed at similar levels in the same brain regions, suggesting that expression all of transcripts variants is required for normal brain function. The transcription of these isoforms will be facilitated by a single promoter as these isoforms share the start exon and inititating methionine (figure 6.1). Any variation in protein expression levels may be due to differences in mRNA half-lives and stability. The expression of these spliceoforms may also be regulated by the MIR-derived putative miRNA target sites and precursors discussed in section 6.6.

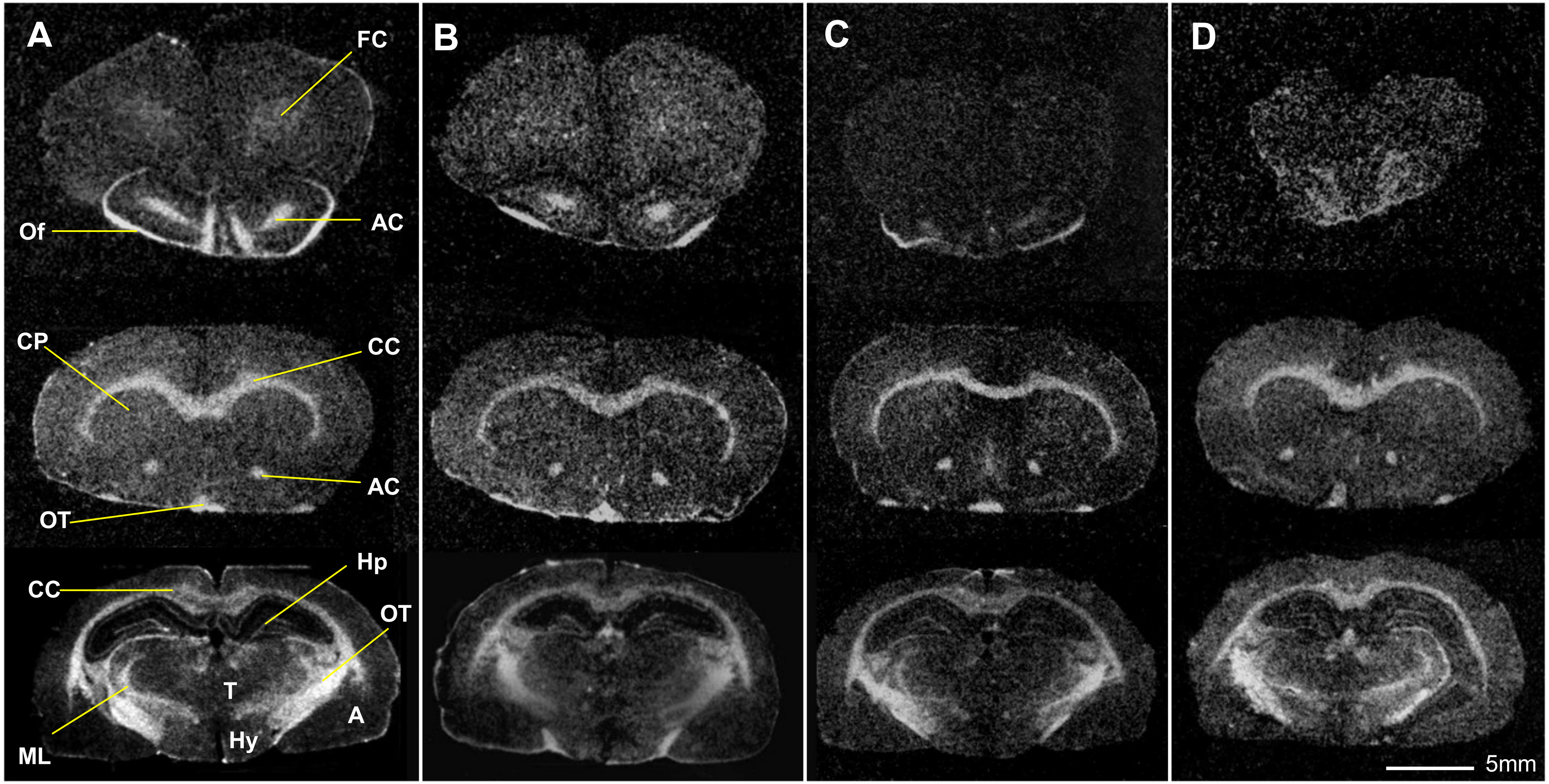


**Figure 6.13. Distribution of TGM2 mRNA expressing cells in the normal rat brain**

**A)** TGM2_Vs; **B)** TGM2_Sh; **C)** TGM2_dLg; **D)** TGM2_wLg. Coronal sections of the normal adult Wistar rat brain hybridised with $^{35}$S-labeled antisense RNA probes. Cells expressing TGM2 mRNA are visualised as light areas. High levels of TGM2 mRNA were detected within the white matter areas; (CC) corpus callosum and (AC) anterior commissure. High levels are also detected in sensory regions; optic nerve/chiasm/tract (OT) and the limbic region; (Hp) hippocampus and (Of) olfactory tract and (ML) media lemniscus. Low levels were detected in the (T) thalamus, (Hy) hypothalamus, (FC) frontal cortex and (CP) caudate putamen.

## 6.6. The role of MIR elements in the function of the TGM2 isoforms

Multiple TGM2 splice variants have been identified and subsequently the expression confirmed in a number of tissues; however the role of the MIR elements in regulating the function of these TGM2 isoforms was to be determined. It is clear that the MIRs are critical for the functional expression of TGM2_Tc and TGM2_dLg by providing methionine and/or stop codons.

In previous chapters the role of MIRs in post-transcriptional control of gene expression has been discussed, for example by the formation of dsRNA, precursor miRNAs and miRNA targets. TGM2 splice variants were screened for the presence of such sequences (table 6.3). There is sequence similarity between three known pre-miRNA stem-loop structures and the TGM2_Tc mRNA sequence (figure 6.14); two in the 5'-UTR and the third within the coding sequence. There is a further potential pre-miRNA stem-loop sequence within the 3'-UTR of TGM2_Vs which resides within the exaptated MIR element (figure 6.14). No evidence of similar structures was found in the other TGM2 isoforms (TGM2_Sh, TGM2_dLg and TGM2_wLg).

| MicroRNA name | Transcript name | mRNA location | Gene region | MIR | Strand | BLAST Score | E-value |
|---|---|---|---|---|---|---|---|
| hsa-mir-1285 | TGM2_Tc | 83 1 | 5'-UTR | N | - | 288 | 2.20E-08 |
| hsa-mir-619 | TGM2_Tc | 63 141 | 5'-UTR | N | + | 188 | 9.60E-04 |
| hsa-mir-220c | TGM2_Tc | 876 949 | CDS | N | + | 137 | 2.70E-01 |
| hsa-mir-640 | TGM2_Vs | 1240 1327 | 3'-UTR | Y | + | 143 | 1.20E-01 |
| NH | TGM2_Sh | . | . | . | . | . | . |
| NH | TGM2_dLg | . | . | . | . | . | . |
| NH | TGM2_wLg | . | . | . | . | . | . |

**Table 6.3. Predicted MiRNA precursor stem -loop sequences**

MiRNA precursor sequences have been predicted using the miRBase search engine. The gene region and co-ordinates which the hairpin is situated has been provided. Abbreviations: Y, the stem-loop sequence it within an exaptated MIR sequence; N, the predicted stem-loop is not located at the MIR region; NH, no hairpin structure detected. The miRNA name and the strand which it is located has been included (+, sense; -, antisense). The BLAST score and the expectation (E-) value has been included; the lower the E-value the more significant the score (Altschul *et al.,* 1990).

**TGM2_Tc**

```
   TGM2_Tc           83   UUUGGGAGGCUGAGGCAGGAGGAUUGCUUGAGCCCAGGAGUUUGAGACCAGCCUGAGCAACAUAGUGAGACC-CCGUCUCUAUA   1
                          |||||||||| ||||| || | ||  ||||||||||| | ||||||||||  ||| |||||| ||||||||| ||||||||| |
   hsa-mir-1285       1   UUUGGGAGGCCGAGGCUGGUGCAUCACUUGAGCCCAGCAAUUUGAGACCAAUCUGGGCAACAAAGUGAGACCUCCGUCUCUACA   84


   TGM2_Tc           63   CUCCUGCCUCAGCCUCCCAAAGUGCUGGGAUUACAGGCGCGAGCCACUACACCCAAC--U-ACUUGUAUUUAUUUACUGCUC   141
                          | ||  ||||||||||||||| ||||||||||||||||  |||||||| |   | ||  | || || |  | |||  ||| |
  hsa-mir-619         1   CGCCCACCUCAGCCUCCCAAAAUGCUGGGAUUACAGGCAUGAGCCACUGCGGUCGACCAUGACCUGGACAUGUUUG-UGCCC   81


   TGM2_Tc          876   GAGCGCACCAGGCUCUUCCGGGACACUCGCUCAGCUCAUCCUCCCACAGCCUU-AGGA--GUGCCUGUGC-CACG-CAG   949
                          || | |||  |||| ||  |  | ||||  | ||||||||| | ||||||||| || |  | ||||| || || | |||
   hsa-mir-220c       7   GACCACACAGGGCUGUUGUGA-AGACUCAGUGAGCUCAUCCCCACACAGCCUUCAGCACAGGGCCUG-GCUCAGGGCAG   83
```

**TGM2_Vs**

```
   TGM2_Vs    1240   GUGACCUUGAGCAAAUGACUUCUU-UCUGA-ACCUCAGUUUCCUCGUCUGGAAAAUGGGGACAACAUCAAG-A-CCUUCCUCCUAGAGUGG  1327
                     |||||| || |||| |  ||     || || || |||| | |||  |||| ||||||||  ||  ||| || | ||| |||| ||  || |
   hsa-mir-640   1   GUGACCCUGGGCAAGUUCCUGAAGAUCAGACACAUCAGAUCCCUUAUCUGUAAAAUGGG--CAUGAUCCAGGAACCUGCCUC-UACGGUUG  88
```

**Figure 6.14. Putative pre-miRNA stem-loop structures identified in the human mRNA sequences of TGM2_Vs and TGM2_Tc**

Sequence alignments identified using MiRBase demonstrating sequence homology between TGM2 mRNA and known human pre-miRNA stem-loops; mir-1285, mir-619, mir-220c and mir-640.

A further mechanism for the regulation of gene expression is via siRNAs derived from dsRNA structures (Ying *et al.,* 2008). These siRNA may then be involved in the RNAi pathway, and subsequently mRNA degradation, mRNA stability and suppressing translation. Two of the MIR elements within the 3'-UTR of the full length TGM2 are in opposite orientations (direct/inverse) and may potentially form dsRNAs (figure 6.15). These MIR elements are present in TGM2_dLg and TGM2_Tc, and may provide a means of regulating transcription and/or translation of all three isoforms. Likewise dsRNAs may form following the hybrisiation of an MIR element in the 5'-UTR and the MIR containing the methionine of the transcript TGM2_Tc.

**TGM2_Tc** (77% sequence identity)

```
  413 acagagaggggaggtcatctccccgggctcacacagccagtgagtggc 460
      |||||||||||| |  || | |  ||| |||||||||  | || ||||
 1186 tgtctctccccttccttacacgaaccccagtgtgtcgccccttaccg 1139
```

**TGM2_dLg** (84% sequence identity)

```
 2428 agtgg-cctcgtggttattagcaaggctgggtaatgtgaaggc 2469
      ||||| ||| |||||||||  ||||||||||  || | || ||
 2803 tcacctggaacaccaataaatgttccgacccggtatatttcg 2761
```

**Figure 6.15. Putative dsRNA structure formed by adjacent MIR sequences in opposite orientations**

There is 77% sequence homology for a potential dsRNA formed between an MIR located within the 5'-UTR and the protein-coding sequence of TGM2_Tc. Likewise there is 84% sequence homology between adjacent MIR elements which have been recruited in the following isoforms: TGM2_Lg, TGM2_dLg and TGM2_Tc. The mRNA coordinates for TGM2_Tc and TGM2_dLg have been provided as a reference.

### 6.6.1. Splice signals and termination codons

A potential role of the MIR elements in the expression of all of the transcripts has been proposed with the exception of TGM2_Sh. One observation with both TGM2_Sh and TGM2_Vs is the occurrence of read-through exons where the donor splice sites are skipped and the intron contains an MIR element. It could be that the MIRs are providing a splicing signal which silences the skipped splice site; therefore the read-through TGM2 isoform sequences were screened for possible exon splice silencer (ESS) sequences (table 6.4). ESSs were identified using the RegRNA web server and were considered authentic if previously demonstrated to be functional.

Putative ESS sequences were located within the MIRs of both TGM2_Sh and TGM2_Vs (table 6.4). It is worth noting that ESS sequences are small and may be abundant in both mRNAs; therefore they may represent artefacts, and cannot be confirmed as functional ESSs with these findings alone. However the possibility remains that the MIRs may be involved in the splicing of these transcripts via one or more of these ESS sequences.

| Transcript name | MIR position | ESS position | ESS sequence | Reference |
|---|---|---|---|---|
| TGM2_Vs | 1203 1333 | 1203 1206 | CAAG | Caputi *et al.,* 1995 |
| | | 1219 1226 | GAGGGAGG | Mandal *et al.,* 2004 |
| | | 1223 1230 | GTGGGAGG | Mandal *et al.,* 2004 |
| | | 1233 1236 | CAAG | Caputi *et al.,* 1995 |
| | | 1245 1249 | CAAGG | Staffa *et al.,* 1997 |
| | | 1246 1249 | CAAG | Caputi *et al.,* 1995 |
| | | 1259 1266 | GAAAGAAG | Mandal *et al.,* 2004 |
| | | 1306 1309 | CAAG | Caputi *et al.,* 1995 |
| | | 1312 1316 | GGAAG | Sun *et al.,* 1993 |
| | | 1318 1321 | TAGG | Del Gatto-Konczak *et al.,* 1999 |
| TGM2_Sh | 1812 1879 | 1819 1823 | CATGG | Uporova *et al.,* 1999 |
| | | 1849 1852 | TTAG | Si *et al.,* 1998 |
| | | 1814 1817 | AAGT | Mayeda *et al.,* 1999 |
| | | 1855 1858 | AAGT | Staffa and Cochrane, 1995 |
| | | 1874 1877 | AAGT | Del Gatto and Breathnach, 1995 |
| | | 1848 1851 | TAGG | Del Gatto-Konczak *et al.,* 1999 |
| | | 1849 1852 | TTAG | Tomonaga *et al.,* 2000 |

**Table 6.4. Exonic splice silencing sequences identified in the recruited MIR elements**

The mRNA co-ordinates for the recruited MIR elements and the predicted ESS have been supplied. References which report the sequences as being functional ESS in other genes have been listed.

## 6.7. Discussion

There are a number of reported alternative TGM2 transcripts which when screening with RepeatMakser were noted to contain MIR elements (Atonyak *et al.,* 2006; Fraij and Gonzales, 1996; Lai *et al.,* 2007). A further novel isoform was identified which has recruited seven MIR elements. TGM2 is also mutated in a number of human neurological disorders including Alzheimer's disease, schizophrenia and Huntington's disease. TGM2 was therefore selected for further analysis and discussion, to determine if the MIR elements and isoforms are conserved between mammals, and if the elements are playing a role in regulating the expression of the TGM2 transcript variants.

### 6.7.1. Alternative splicing of TGM2 and the exaptation of MIR elements

A total of five human TGM2 transcript variants were identified, four published and one novel (TGM2_Tc). Intriguingly all of the isoforms have recruited MIRs, with a total of nine independent exaptated MIR elements identified. The wild-type full length transcript (TGM2_wLg) has recruited three MIR elements within the 3'-UTR. Two of the TGM2 isoforms (TGM2_Sh and TGM2_Vs) encode truncated proteins due to skipping the donor splice site of exons 6 and 10 respectively. The shorter forms of TGM2 contain different MIR elements also in the 3'-UTR. A further TGM2 isoform (TGM2_dLg) is encoded following an atypical splicing event occurring in exons 12 and 13 of the full length transcript. Both the acceptor and donor splicing takes place within the consensus sequence CTYCAC, which is not a recognised classic splice site; however a similar atypical splicing event, containing a CTTC motif, has been reported for the G protein, beta-3 subunit (GNB3; Rosskopf *et al.,* 2003). TGM2_dLg shares the three MIR elements identified in the full length transcript, however the irregular splicing encodes a peptide from an alternative reading frame, with the stop codon located within the first of the recruited MIR. A fifth novel protein (TGM2_Tc) is truncated at the N-terminus, and is translated from an initiating methionine located within RefSeq intron 10, and as such will be regulated by an alternative promoter. TGM2_Tc has recruited seven MIR elements in total, two of which are contributing to the coding sequence, including the alternative initiating methionine codon, three are in the 3'-UTR and are at the same position observed with the full length isoform. The TGM2_Tc transcript shares the last three exons at the C-terminal of the wild-type TGM2 with a maintained reading frame.

Multiple cross-species sequence alignments revealed conservation of TGM2_Vs and TGM2_Sh with other primates, rodents, dog and cow. However conservation of TGM2_Tc could only be detected in primates, and there is no evidence of the integration of the MIRs in the genomic region. Given the age of the MIR elements all mammals should have integrated, and the repeats have been deleted in dog, cow and rodents, suggesting that TGM2_Tc may be involved in a recently developed function, possibly related to the known TGM2 dual-activities. TGM2_dLg is encoded following an atypical splicing event, and the splice site including the flanking intronic sequences are conserved in primates; with an alternative acceptor splice site being reported at a different position in rats (Monsonego *et al.,* 1997). Neither of the splicing motifs appears to be conserved in mouse, dog or cow, however the MIR elements have been exaptated but without the necessary sequence changes required to generate the splice motif. It appears therefore that the different TGM2 isoforms may have arisen at different times during evolution, possibly to modify the diverse enzyme activity of the TGM2 enzyme.

### 6.7.2. Expression analysis and conservation of the TGM2 splice variants

Expression analysis of the TGM2 isoforms revealed ubiquitous expression of the mRNA, with all five transcripts being expressed in a number of tissues, including the rat clonal beta cell line (BRIN-BD11) and the glioma (C6) rat cell line. The expression of the rat isoforms were observed in rat brain by radioactive *in situ* hybridisation, and all transcripts studied display similar spatial staining patterns. All TGM2 rat isoforms were expressed in the hippocampus, including the dendrate gyrus and fimbria; areas fundamental in long-term learning and spatial awareness. Memory loss is the earliest clinical observation in Alzheimer’s disease patients, and upregulation of TGM2_Sh has been observed in the hippocampal regions of such individuals. High levels of TGM2 expression was visualised in cells of the optic nerve, optic chiasma and optic tract for all four isoforms, consistent with a role in visual perception and memory. TGM2 has previously been linked to retinal dystrophy and apoptosis during photoreceptor degeneration in rats (Zhang *et al.,* 1996), but the role of TGM2 in visual pathways or the visual system of the brain has not previously been studied.

A high level of TGM2 mRNA staining was noted in the olfactory tract and olfactory bulb, which contains the sensory neurons responsible for processing and perceiving odours. Olfactory dysfunction is associated with early Parkinson's disease, and patients frequently experience odour hallucinations and anosmia (absence of smell sensation) (Diederich *et al.,* 2009; Wszolek *et al.,* 1998). Neurodegenerative anosmia is an early indicator of dementia and Alzheimer's disease, with all patients displaying olfactory impairment (Wilson *et al.,* 2009). Furthermore >70% of individuals over 80 years of age are reported to have a significant loss in odour sensation (Lafreniere *et al.,* 2009).

A strong signal for TGM2 mRNA was detected in the corpus callosum and anterior commissure, which are white matter regions connecting the left and right cerebral hemispheres; facilitating communication between the two hemispheres. The corpus callosum (CC) has been suggested to be involved in processing emotions and social interaction and as such is implicated in autism, with a reduction in the total CC size being observed in autistic individuals (Vidal *et al.,* 2006). Agenesis and/or atrophy or the CC has also been noted in Alzheimer's disease, Parkinson's disease, schizophrenia patients and established bipolar disorder cases (Serra *et al.,* 2009; Wszolek *et al.,* 1998; Walterfang *et al.,* 2009; Davis, 1994). TGM2 has not previously been implicated in autism or bipolar affective disorder; however TGM2 has been implicated in the development of schizophrenia following the identification of several TGM2 SNPs in schizophrenia cases and their parents (Bradford *et al.,* 2009). Furthermore abnormal protein cross-linking and amyloid plaques have been linked to TGM2 in the pathology of Alzheimer's disease and TGM2 has also been shown to be responsible for the formation of Lewy bodies in the hippocampus of Parkinson's disease patients (Andringa *et al.,* 2004).

### 6.7.3. The TGM2 isoforms will display differences in transamidation and GTP-binding activity

TGM2 is a multifunctional calcium-dependant enzyme which catalyses both intracellular and extracellular protein cross-linking, and functions as a G-protein when bound to GTP. TGM2 is composed of distinct activity and binding domains, the full length isoform contains several calcium binding sites within a catalytic core region, with the TGase function attributed to the core and N-terminus, and the C-terminus essential for G-protein activity.

There are numerous guanine binding sites positioned predominantly in the β-barrels at the C-terminal end, which are known to interact with α1-AR, PLC-δ1 and GPR56. There is a distinct relationship between the dual activities of TGM2. When bound to GTP the TGase activity is downregulated, whereas GTPase activity is inhibited by $Ca^{2+}$ binding. The two activities are known to possess different functional roles. When TGM2 is functioning as a TGase it has been shown to be involved in apoptosis, cell adhesion and protein cross-linking (Collighan and Griffin, 2009; Malion and Piacentini, 1998), whereas when TGM2 is acting as a G-protein it stimulates PLC-δ1 activity and regulates receptor-mediated signalling by coupling ligand-bound GPCRs to PLC-δ1, mainly α1-AR, thromboxane and oxytocin (Park *et al.,* 1998; Kang *et al.,* 2002). GTP-binding and subsequent PLC-δ1 stimulation has been implicated in anti-apoptotic cell survival (Citron *et al.,* 2002). Moreover there is a strong relationship between the interaction and stimulation of PLC-δ1 by TGM2 and the regulation of intercellular calcium levels. PLC-δ1 is stimulated upon binding to the GTP-bound TGM2 protein, thus inducing capacitative calcium entry, following which a constant level of intracellular $Ca^{2+}$ is maintained until the GTP is hydrolysed resulting in the inhibition of PLC-δ1 activity (Kang *et al.,* 2002). Overall the dual-function of TGM2 offers a means of switching between different enzyme activities, and alternative splicing of TGM2 may provide further regulation and control of the diverse protein function, allowing for TGM2 to adapt to different environmental conditions.

#### 6.7.3.1. Protein activity of TGM2_Sh and TGM2_Vs

Differential expression of the TGM2_Sh, TGM2_dLg and full length TGM2 has been previously demonstrated. Antonyak *et al.,* (2006) studied TGM2 mutants lacking 30aa at the C-terminal end which were unable to bind GTP, or to stimulate and/or interact with PLC-δ1, and displayed weak transamidation activity; possibly a consequence of failed TGM2-mediated receptor signalling and associated $Ca^{2+}$ entry. Antonyak *et al.,* (2006) further demonstrated a lack of GTP-binding and weak TGase activity of TGM2_Sh expressed in HeLa cells, and suggested that the full length isoform protects against apoptotic events and cell death, whereas the short isoform induces apoptotic responses. The apoptotic effect of TGM2 is suggested to be due to abnormal protein aggregations of TGM2, with TGM2_Sh forming large oligomers in cells, which is not attributed to TGase cross-linking but through an undetermined mechanism (Antonyak *et* *al.,* 2006). It is most likely that TGM2_Vs is also GTP-independent and may behave in a similar manner to that observed for TGM2_Sh due to the transcript encoding a peptide which lacks the C-terminal 149 residues. Furthermore it can be assumed that both the short isoforms will not interact with α1-AR, PLC-δ1 and the G-protein coupled receptor GPR56, due to the absence of G-protein function.

### 6.7.3.2. Protein activity of TGM2_dLG

TGM2_dLg is composed of 13 exons and is highly similar to the RefSeq, however the C-terminus differs due to atypical splicing, encoding a protein which contains all the guanine nucleotide binding sites but lacks the interaction sites for PLC-δ1 and α1-AR. Festoff *et al*., (2002) studied the expression of TGM2_dLG and TGM2_wLg following spinal cord injury (SCI) in rats. Prior to SCI they detected the full length TGM2, however 24 hrs post-SCI, both transcripts were detected with TGM2_dLg being upregulated. As such they proposed that following SCI, there is a switch in GTP-dependency of TGM2 with a preference for the GTP-independent isoform. Tolentino *et al.,* (2004) studied the expression of TGM2_dLg following cerebral ischemic attack (stroke), focussing particularly on the fluctuation of cytosolic $Ca^{2+}$. TGM2_dLg was noted to be upregulated and expressed at higher levels than TGM2_wLg at low calcium concentrations. Finally TGM2_dLg has been shown to have an altered affinity for $Ca^{2+}$ at normal calcium levels, TGase activity is less than 10% that observed in the full length form. Furthermore TGM2_dLg has also been demonstrated to enhance GTP hydrolysis and is insensitive to GTP inhibition when $Ca^{2+}$ levels increase (Lai *et al.,* 2007).

Expression of TGM2_dLg may allow for an enhanced response to subtle changes in calcium levels, whereas the full length transcript may not be receptive to such changes. It is possible that TGM2_dLg may be rapidly expressed following splicing of the full length mRNA transcript. For example, once TGM2_wLg has been exported to the cytoplasm, post-transcriptional splicing may occur, deleting the sequence flanked by the CTYCAC motif; resulting in nucleus-independent control of gene expression. A two step splicing event would allow for a rapid switch to the TGM2_dLg GTP-independent isoform, and downregulation of the full length transcript in response to environmental changes, such as following injury or ischemia. A similar means of post-transcriptional regulation has been suggested for the minor spliceosome, and it is thought that minor splicing (U12-type) may take place in the cytosol following nuclear export of partially spliced pre-mRNA, supported by the detection of the associated minor-class of small-nuclear-RNAs enriched in the cytoplasm (König *et al.,* 2007).

#### 6.7.3.3. Protein activity of TGM2_Tc

The final transcript identified (TGM2_Tc) is the only transcript that lacks the fibronectin domain and as such would not be involved in FN1-associated functions including wound healing, cell adhesion and blood plasma development (Akimov and Belkin, 2001; Telci and Griffin, 2006). TGM2_Tc shares the last three exons at the C-terminal end with the full length transcript and has intact binding sites for PLC-δ1 and α1-AR; however only four of the five guanine binding residues are encoded and as such the protein will not form a complete GTP-binding pocket. Consequently TGM2_Tc may have weak or null GTPase activity and as a result, limited PLC-δ1 and α1-AR interaction and stimulation.

TGM2_Tc is intriguing as it is the only isoform which will be regulated by an alternative promoter sequence. Furthermore the putative protein will most likely be calcium independent and will not exhibit TGase activity or the ‘normal’ GTP activity associated with the wild-type TGM2. The only common feature shared with the full length protein is the presence of the PLδ-C1, α1-AR and GPR56 interaction sites. It is possible that the previously discussed kinase activity of TGM2 may be attributed to the novel N-terminal polypeptide sequence or that the truncated protein functions solely as a G-protein. Alternatively the predicted dsRNAs formed by the hybridisation of adjacent MIR elements present both in the 5’-UTR and 3’-UTR may be a source of siRNAs that regulate the expression levels of the other isoforms, encoded from a single promoter. These suggestions however are speculative and require further investigation and experimental validation. Finally given that MIRs were active prior to the mammalian radiation, the MIR elements recruited by TGM2_Tc should be present in non-primate mammals, yet the repeats cannot be detected and have either integrated late in mammalian evolution or have been deleted. These finding suggesting that TGM2_Tc may have a recently adapted function, possibly involved in the known TGM2 dual-activities or an increase in cognitive ability, such as learning and memory.

Taking into account the potential variations in enzyme activity of the TGM2 isoforms, it appears that there is a necessity for ‘finely tuned’ gene expression in situations where there are fluctuations in intracellular [$Ca^{2+}$], for example when there is a reduction in free $Ca^{2+}$ following ischemia, as discussed previously. In addition, there may be a requirement for TGM2 to function as a G-protein in a $Ca^{2+}$-independent manner when there is a high level of $Ca^{2+}$ entry from intracellular stores. One such example being at the onset of and during parturition, as the primary trigger for human myometrial contractions is an increase in intracellular calcium (Tribe *et al.,* 2001). In such instances TGM2 may need to function solely as a G-protein to couple with α1-AR and oxytocin, the neurotransmitter responsible for contractions of the myometrium during labour and mammary tissue during lactation (Lee *et al.,* 2009). One study monitored TGM2 expression levels in rats during pregnancy (Dupuis *et al.,* 2004), and GTP-bound TGM2 was predominantly localised to the plasma membrane and upregulated during pregnancy. Furthermore myometrial α1-AR is known to participate in the initiation of uterine contractions (Limon-Boulez *et al.,* 1997), possibly through TGM2-mediated signalling with PLC-δ1.

### 6.7.4. The role of MIR elements in the differential expression and activity of TGM2

During previous chapters the primary observation has been that MIR elements may be a source of *cis*-acting regulatory sequences, such as miRNA, dsRNAs, cryptic splice sites and initiation and termination codons. The abundance of MIR elements recruited by TGM2 isoforms suggests that these elements may play a role in TGM2 function. Therefore the exaptated MIR sequences of TGM2 were studied to determine if the repeats are contributing to the expression of the TGM2 isoforms by the previously described mechanism, such as post-transcriptional control of gene expression via the formation of dsRNAs or by providing translational control codons.

The role of MIR elements in the protein expression of TGM2_Tc and TGM2_dLg is easily determined, as a repeat is providing an initiating methionine and a stop codon. However the role of the MIR elements in the expression of TGM2_Vs and TGM2_Sh is less clear and required further investigation. A putative miRNA precursor stem-loop structure was noted in the MIR sequence of TGM2_Vs, which could regulate expression by blocking translation of this protein product. TGM2_Vs and TGM2_Sh are both read-through transcripts where donor splicing is skipped and a truncated protein produced. It is possible that the MIR elements may be a source of splicing signals which may silence the skipped splice site, and several exonic splice silencer sequences were identified within the recruited MIRs of TGM2_Vs and TGM2_Sh, however these sequences have not been confirmed experimentally.

Finally three predicted pre-miRNA sequences were noted in the 5'-UTR and coding-sequence of TGM2_Tc. The stem-loops are not located within the MIR element, but may regulate the expression of the TGM2_Tc isoform, which is encoded from an MIR-derived methionine. Moreover a putative dsRNA structure was identified for TGM2_Tc, formed by the hybridisation of an MIR element in the 5'-UTR and a second inverted MIR element in the coding sequence. A second predicted dsRNA may also be generated between adjacent MIR elements present in the 3'-UTR of TGM2_wLg, TGM2_dLg and TGM2_Tc. The role of MIR elements in the localisation of mRNA in polarised cells, via the formation of dsRNA and miRNAs has previously been discussed (section 4.4). TGM2 is expressed in a number of polarised cells such as photoreceptors, neurones and epithelial cells (Zhang *et al.,* 1996; Festoff *et al.,* 2001; Treharne *et al.,* 2009). Furthermore TGM2 is known to be expressed in distinct cell compartments, for example G-protein activity occurs at the cell membrane and in the cytosol, yet when TGM2 is acting as TGase, protein activity takes place in the nucleus, extracellular space and cytosol (Fesus and Piacentini, 2002). Therefore MIR-mediated mRNA localisation and RNAi may play a role in regulating spatial protein expression of TGM2 in polarised cells.

## 6.8. Future investigations

To determine if the loss of the C-terminal residues has an effect on the overall activity of TGM2 a number of well documented protein and functional assays could be performed. TGase activity can be measured directly from cell lysates (Gnaccarini *et al.,* 2009; Slaughter *et al.,* 1992). Fibronectin binding, GTP binding and inhibition could also be measured by well established protocols (McEwen *et al.,* 2002; Datta *et al.,* 2006; Achyuthan *et al.,* 1996).

Determining the protein activity of the individual TGM2 isoforms would require the optimisation of several experiments in order to differentiate between the various splice variants and protein products. For example, the full length isoforms could be cloned into mammalian expression vectors, allowing for the transient transfection of mammalian cell lines. Isoform specific monoclonal antibodies may also be designed, specific to the novel peptide sequence of each protein product. These antibodies could then be used to measure protein levels by western blotting. Individual isoforms could also be knocked down in cultured cells by siRNAs.

The expression of the isoforms in transfected cells (plasmid or siRNA) could be measured by real-time qRT-PCR following exposure to cytotoxic or neuroprotective compounds modelling specific neurological conditions. For example treating with amyloid β-peptides that are known to accumulate and damage neurons in the brain of patients with Alzheimer disease (Wilhelmus *et al.,* 2009). Likewise the glucose responsiveness of transfected pancreatic β-cells could be studied and the isoform expression recorded in hyperglycaemic and/or lipotoxic conditions (Ball *et al.,* 2000).

## 6.9. Conclusion

It appears that MIR elements are involved in regulating the expression of TGM2 isoforms and may have provided a means of enhancing the multifunctional enzyme properties. There are five human TGM2 splice variants, all of which have recruited at least one MIR element. The isoforms appear to be conserved in other mammals with the exception of the primate specific TGM2_Tc and there are no TGM2 transcript variants identified for chicken and zebrafish. It appears that the MIR elements may play an important role in regulating the expression of the multiple transcript variants by providing *cis*-acting regulatory sequences such as stop codons, initiating methionines and miRNA target sites. And as such the repeats elements may be regulating the rate of protein expression, or mRNA stability by forming dsRNA and binding miRNA.

Overall it is possible that the transposition of MIR elements may have provided a pool of new genomic material that was utilised by the TGM2 gene to generate multiple protein products. This multitude of TGM2 proteins may have assisted in the adaptation of TGM2 function during mammalian evolution, producing the complexly regulated multi-activity of TGM2.

# 7. GENERAL DISCUSSION AND FURTHER STUDIES

## 7.1. MIR elements recruited in the 3'-UTRs may be involved in the post-transcriptional control of gene expression

The transcription of mature mRNA is a multi-factorial process involving a number of elements including RNA polymerase, promoters, transcription factors, enhancers and splice sites. UTRs have also been shown to play a role in post-transcriptional control of gene expression by a number of mechanisms including; mRNA transport and sub-cellular localisation, repression of translation and mRNA stability (Mignone *et al.,* 2003; Guhaniyogi and Brewer, 2001). The significance of the 3'-UTR in gene expression has previously been suggested, for example mutations and polymorphisms in the 3'-UTRs has been attributed to disease, such as bipolar disorder, retinal degeneration and systemic lupus erythematosus (Verardo *et al., 2*009; Pickard *et al.,* 2008; Citores *et al.,* 2004; Conne *et al.,* 2000).

In chapter 3 it was noted that 75% (1395) of the total MIR elements identified have been recruited in the 3'-UTRs of human genes, with a high level of sequence conservation of the core-SINE, as such these genes share a common feature and may contain sequence homology of ~70bp. The highly conserved property of the core-SINE suggests that maintaining the repeat sequence may be fortuitous. The abundance of MIR in the 3'-UTR may merely reflect the lack of selective pressure as these repeats are less likely to disrupt protein expression. Nonetheless it is conceivable to assume that some MIR elements may become domesticated, and that this tolerance of recruiting TEs in the 3'-UTR may provide sufficient time for the MIR to adopt a functional role without affecting the viability of the existing gene expression.

Sub-cellular localisation of mRNA is a means of targeting protein expression to a specific, usually distal intracellular region, and RNA is transported in the majority and possibly all polarised cell types (Mohr and Richter, 2001; Palacios and St Johnston, 2001). The translocation of mRNA in hippocampal neurones is regulated by miRNAs and has been implicated in learning and memory (Costa-Mattioli *et al.,* 2009). A number of putative miRNA target sites, precursor miRNA stem-loops and dsRNA structures were noted to reside within exaptated MIR elements in 3'-UTRs. Four of which the mRNA has previously been reported to be localised to the dendritic compartment of hippocampal neurones (CD59, CAMK2A, DDN and NEURL). Two of these genes, DDN and CAMK2A, are both suggested to be involved in synaptic plasticity, and learning and memory (Pinkstaff *et al.,* 2001; Kremerskothen *et al.,* 2006). Furthermore both DDN and CAMK2A contain regulatory sequences of ~1kb in the 3'-UTR, dendritic-targeting-elements (DTEs), which include recruited MIR elements. MiRNA target sites were also identified within exaptated 3'-UTR MIR elements of AHI1, NRL, RHO and RPGR; genes which are expressed in photoreceptor rod cells and have been implicated in retinal degenerative conditions.

## 7.2. MIR elements provide cryptic splice sites and many have been exonised

In chapter 4 it was noted that MIR elements are associated with 5% (1359) of the genes in the human genome, with 117 of these genes containing protein-coding sequences partly derived from MIRs. There are 26 distinct exons comprised entirely of an MIR element including both the 3' and 5' splice sites, further supporting the hypothesis that MIR elements are providing *cis*-regulatory sequences. A total of 126 splice sites were identified with a large proportion being alternative acceptor splice sites, when the MIR element is recruited in the antisense orientation. Two sequence motifs were identified in the MIR consensus sequences which resemble strong splice sites. The acceptor splice site could potentially be active immediately following integration without the need for change from the original MIR consensus, and all of the MIR sub-types contain an almost perfect canonical acceptor splice site in the consensus sequence. This degree of conservation suggests that further enhancer sequences are necessary if the splice site is to be maintained, otherwise most antisense MIR elements would be exonised.

The majority of MIR elements will have originated from intronic MIR elements, with most intact intronic MIRs behaving as pseudo-exons. As such any intronic MIR element could potentially be accessed by a gene during mammalian evolution when developing new phenotypes. A large number of MIR elements are regulating the expression of spliceoforms by providing alternative initiating methionines and stop codon sequences. MIR elements have provided a resource of raw genomic material that could be utilised by mammalian genes to fine tune gene function and to adapt or 'experiment' with a repertoire of protein products.

## 7.3. Further investigations

Many of the proposed functions of the MIR elements described in this study require experimental analysis. Validating predicted miRNA target sites and precursors is challenging and labour intensive, as miRNAs are small sequences (~22 nucleotides) and are usually expressed at low levels. The most effective method would be to use a transfection approach with recombinant clones. Firstly the miRNA precursor sequence will have to be cloned (from genomic DNA) into an expression vector and transfected in a chosen cell line, following which an increase in miRNA expression should be noted compared to the endogenous levels by real-time qRT-PCR. Secondly, constructs of the 3'-UTRs of the gene of interest, which harbours the MIR and predicted miRNA target, will have to be cloned into a luciferase reporter. A further construct of the 3'-UTR sequence could be generated which lacks the MIR element. The miRNA precursor expression vector and the 3'-UTR/luciferase reporter will then both have to be transfected in cultured cells. Following which the gene expression will be measured by real-time qRT-PC and/or western blotting, to validate the MIR-derived miRNA target sequence is regulating gene expression.

A similar approach could be used to determine if the MIR elements are involved in mRNA stability and degradation. The expression of the 3'-UTR/luceferase reporter with the MIR element could be compared to that which has the MIR element excised. The role of the MIR elements in translocating mRNA to the dendritic comapartment of neurones can be determined by cloning chosen full length mRNA contructs which either contain or lack the MIR element into mammalian expression vectors, tagged with green fluorescent protein (GFP). The recombinant clones will then have to be transfected into a neuronal cell line, and the spatial expression of the protein visualised with a fluorescent microscope to determine if the MIR is involved in regulating localised protein synthesis in specific cell compartments. Localisation of mRNA could also be detected by non-radiaoactive *in situ* hybridisation. Finally the tissue-specific expression of the MIR-containing spliceoforms will also have to be confirmed by quantitative real-time qRT-PCR.

One obvious constraint during these investigations is that the results presented are confined to the current annotated sequence databases, which are consistently updated with the verification of additional ESTs and cDNA clones daily. Likewise a large portion of mammalian genomic sequence data is incomplete. As such revisiting the databases at a later date could reveal further MIR-containing genes and provide additional insight in to the role of MIR elements is mammalian genomes. It is also possible that MIR elements may have other functional roles, for example mammalian imprinting, as a number of imprinted genes have recruited MIR elements. Evidence (chapter 4) suggests that MIR elements may be involved in immunological processes, both areas are of interest which may be investigated further.

Only exonic MIR elements were collected and analysed, corresponding to ~0.4% of the total MIR elements in the human genome. It is possible that intronic MIR elements may take part in controlling gene expression via undetermined mechanisms, and intronic mutations and SNPs have previously been attributed to disease. Likewise intergenic MIR elements may be involved in enhancing gene expression or contribute to promoter regions. For example a preliminary analysis using TranspoGene (http://transpogene.tau.ac.il/; Levy *et al.,* 2008) identified 489 annotated human genes which have putative promoters which contain MIR-derived sequences, these results are raw and are not included in this study, however the data may be revisited in the future. Furthermore it would be interesting to screen the complete dataset of MIR-containing genes for miRNA targets and dsRNA structures to determine if a particular miRNA family are prevalent. MIR elements actively accumulated prior to the radiation of mammals, and whilst this study has focussed on human genes specifically, it would be interesting to study the full global exaptation of the MIR elements across a large number of mammals, including the more distant relatives, monotremes and marsupials. All of these topics mentioned warrant further investigation and any one would provide a solid foundation for subsequent research projects.

## 7.4. Conclusion

In conclusion, MIR elements have the potential to provide a series of *cis*-regulatory sequences. Elements detected in 5'-UTRs and coding sequences are contributing to alternative splicing by carrying stop codons, translation initiation codons and motifs which resemble known splice sites. MIR elements recruited in the 3'-UTR appear to be playing a different but equally significant role in gene expression, and may be involved in post-transcriptional control by generating dsRNA structures, precursor miRNAs and providing miRNAs target sites. MIR elements may be involved in the regulation of expression in neuronal cells by localising mRNAs to distinct cellular compartments and by suppressing translation during transport until there is a need for rapid protein expression. It can also be considered that MIR elements may regulate RNA translocation in other polarised cell types, such as epithelial cells, cilia and fibroblasts; or may be involved in oogenesis. Noteworthy, it is possible that MIR elements may be a source of miRNA target sites when exaptated in the 5'-UTR or the coding sequence, though less common miRNAs have been reported to regulate gene expression and block translation and/or transcription when bound to these regions (Zhou *et al.,* 2009).

# 8. REFERENCES

Ackerman H, Udalova I, Hull J, Kwiatkowski D. (2002) Evolution of a polymorphic regulatory element in interferon-gamma through transposition and mutation. Mol Biol Evol; 19:884-890.

Aeschlimann D, Paulsson M. (1991) Cross-linking of laminin-nidogen complexes by tissue transglutaminase. A novel mechanism for basement membrane stabilization. J Biol Chem; 266: 15308-15317.

Akimov SS, Belkin AM. (2001) Cell-surface transglutaminase promotes fibronectin assembly via interaction with the gelatin-binding domain of fibronectin: a role in TGFbeta-dependent matrix deposition. J Cell Sci; 114:2989-3000.

Ala-Mello S, Sankila EM, Koskimies O, de la Chapelle A, Kaariainen H. (1998) Molecular studies in Finnish patients with familial juvenile nephronophthisis exclude a founder effect and support a common mutation causing mechanism. J Mol Biol; 35: 279-283.

Al-Shahrour F, Minguez P, Tárraga J, Montaner D, Alloza E, Vaquerizas JMM, Conde L, Blaschke C, Vera J, Dopazo J. (2006) BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. Nucleic Acids Res;34 (Web Server issue): W472-W476

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990) Basic local alignment search tool. J Mol Biol; 215: 403-410.

Alvarez-Retuerto AI, Cantor RM, Gleeson JG, Ustaszewska A, Schackwitz WS, Pennacchio LA, Geschwind DH. (2008) Association of common variants in the Joubert syndrome gene (AHI1) with autism. Hum Mol Genet; 17: 3887-3896.

Amann-Zalcenstein D, Avidan N, Kanyas K, Ebstein RP, Kohn Y, Hamdan A, Ben-Asher E, Karni O, Mujaheed M, Segman RH, Maier W, Macciardi F, Beckmann JS, Lancet D, Lerer B. (2006) AHI1, a pivotal neurodevelopmental gene, and C6orf217 are associated with susceptibility to schizophrenia. Eur J Hum Genet; 14: 1111-1119.

Ambros V, Chen X. (2007) The regulation of genes and genomes by small RNAs. Development; 134: 1635-41.

Andersson ME, Sjolander A, Andreasen N, Minthon L, Hansson O, Bogdanovic N, Jern C, Jood K, Wallin A, Blennow K, Zetterberg H. (2007) Kinesin gene variability may affect tau phosphorylation in early Alzheimer's disease. Int J Mol Med; 20: 233-239.

Andringa G, Lam KY, Chegary M, Wang X, Chase TN, Bennett MC. (2004) Tissue transglutaminase catalyzes the formation of alpha-synuclein crosslinks in Parkinson's disease. FASEB J; 18: 932-934.

Antonyak MA, Jansen JM, Miller AM, Ly TK, Endo M, Cerione RA. (2006) Two isoforms of tissue transglutaminase mediate opposing cellular fates. Proc Natl Acad Sci U S A; 103: 18609-18614.

Arima T, Yamasaki K, John RM, Kato K, Sakumi K, Nakabeppu Y, Wake N, Kono T. (2006) The human HYMAI/PLAGL1 differentially methylated region acts as an imprint control region in mice. Genomics; 88: 650-658.

Armour J A, Wong Z, Wilson V, Royle NJ, Jeffreys AJ. (1989) Sequences flanking the repeat arrays of human minisatellites: association with tandem and dispersed repeat elements. Nucleic Acids Res; 17: 4925-4935.

Ashraf SI, Kunes S. (2006) A trace of silence: memory and microRNA at the synapse. Curr Opin Neurobiol; 16: 535-539. Review.

Ast G. (2004) How did alternative splicing evolve? Nat Rev Genet; 5: 773-82. Review.

Baertsch R, Diekhans M, Kent WJ, Haussler D, Brosius J. (2008) Retrocopy contributions to the evolution of the human genome. BMC Genomics; 9:466.

Bailey CD, Johnson GV. (2005) Tissue transglutaminase contributes to disease progression in the R6/2 Huntington's disease mouse model via aggregate-independent mechanisms.J Neurochem; 92:83-92.

Ball AJ, Flatt PR, McClenaghan NH. (2000) Stimulation of insulin secretion in clonal BRIN-BD11 cells by the imidazoline derivatives KU14r and RX801080. Pharmacol Res; 42: 575-579.

Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. Cell; 116: 281–297.

Baust C, Baillie GJ, Mager DL. (2002) Insertional polymorphisms of ETn retrotransposons include a disruption of the wiz gene in C57BL/6 mice. Mamm Genome; 13: 423-428.

Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, Kent WJ, Haussler D. (2006) A distal enhancer and an ultraconserved exon are derived from a novel retroposon.

Ben Fredj N, Grange J, Sadoul R, Richard S, Goldberg Y, Boyer V. (2004) Depolarization-induced translocation of the RNA-binding protein Sam68 to the dendrites of hippocampal neurons. J Cell Sci.

Benjamini Y, Hochberg Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Statist. Soc. Ser. B 57 289-300.

Bennett EA, Keller H, Mills RE, Schmidt S, Moran JV, Weichenrieder O, Devine SE. (2008) Active Alu retrotransposons in the human genome. Genome Res; 18: 1875-1883.

Bennetzen JL, Swanson J, Taylor WC, Freeling M. (1984) DNA insertion in the first intron of maize Adh1 affects message levels: cloning of progenitor and mutant Adh1 alleles. Proc Natl Acad Sci U S A; 81: 4125-4128.

Berleth T, Burri M, Thoma G, Bopp D, Richstein S, Frigerio G, Noll M, Nüsslein-Volhard C. (1988) The role of localization of bicoid RNA in organising the anterior pattern of the Drosophila embryo. EMBO J; 7:1749-1756.

Bernassola F, Federici M, Corazzari M, Terrinoni A, Hribal ML, De Laurenzi V, Ranalli M, Massa O, Sesti G, McLean WH, Citro G, Barbetti F, Melino G. (2002). Role of transglutaminase 2 in glucose tolerance: knockout mice studies and a putative mutation in a MODY patient. FASEB J; 16: 1371-1378.

Bernstein E, Caudy AA, Hammond SM, Hannon GJ. (2001) Role for a bidentate ribonuclease in the initiation step of RNA interference. Nature; 409: 363-366.

Bessant DA, Holder GE, Fitzke FW, Payne AM, Bhattacharya SS, Bird AC. (2003) Phenotype of retinitis pigmentosa associated with the Ser50Thr mutation in the NRL gene. Arch Ophthalmol; 121:793-802.

Blichenberg A, Rehbein M, Muller R, Garner CC, Richter D, Kindler S. (2001) Identification of a cis-acting dendritic targeting element in the mRNA encoding the alpha subunit of Ca2+/calmodulin-dependent protein kinase II. Eur J Neurosci; 13: 1881-1888.

Boissinot S, Furano AV. (2001) Adaptive evolution in LINE-1 retrotransposons. Mol Biol Evol; 18: 2186-2194.

Borchert GM, Lanier W, Davidson BL. (2006) RNA polymerase III transcribes human microRNAs. Nat Struct Mol Biol; 13: 1097-1101.

Bowen NJ, Jordan IK. (2002) Transposable elements and the evolution of eukaryotic complexity. Curr Issues Mol Biol; 4: 65-76. Review.

Bradford M, Law MH, Stewart AD, Shaw DJ, Megson IL, Wei J. (2009) The TGM2 gene is associated with schizophrenia in a British population. Am J Med Genet B Neuropsychiatr Genet; 150B: 335-340.

Brenda P, Lightbody J, Sato G, Levine L, Sweet W. (1968) Differentiated rat glial cell strain in tissue culture. Science; 161: 370-371.

Brenner S. (1990).The human genome: the nature of the enterprise. Ciba Found Symp; 149:6-12; discussion 12-7.

Britten R. (2006) Transposable elements have contributed to thousands of human proteins. Proc Natl Acad Sci U S A; 103:1798-1803.

Britten RJ, Davidson EH. (1969) Gene regulation for higher cells: a theory. Science; 165: 349-357.

Britten RJ, Davidson EH. (1971) Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. Q Rev Biol; 46: 111-138.

Britten RJ, Kohne DE. (1968) Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. Science; 161: 529-540.

Britten RJ, Kohne DE. (1970) Repeated segments of DNA. Sci Am; 222: 24-31.

Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran JV, Kazazian HH Jr. (2003) Hot L1s account for the bulk of retrotransposition in the human population. Proc Natl Acad Sci U S A; 100:5280-5285.

Buchon N, Vaury C. (2006) RNAi: a defensive RNA-silencing against viruses and transposable elements. Heredity; 96: 195–202.

Burki F, Kaessmann H. (2004) Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux. Nat Genet; 36: 1061-1063.

Butler M, Goodwin T, Simpson M, Singh M, Poulter R. (2001) Vertebrate LTR retrotransposons of the Tf1/sushi group. J Mol Evol; 52: 260-274.

Caldwell EF, von Cramon-Taubadel N, Weale ME, Thomas MG. (2004) Salivary amylase gene copy number: Have humans adapted to high starch diets? Am J Phys Anthropol; 123: 72.

Callinan PA, Batzer MA.(2006) Retrotransposable elements and human disease. Genome Dyn; 1: 104-115. Review.

Carter-Dawson LD, LaVail MM. (1979) Rods and cones in the mouse retina. II. Autoradiographic analysis of cell generation using tritiated thymidine. J Comp Neurol; 188:263-272.

Casavant NC, Scott L, Cantrell MA, Wiggins LE, Baker RJ, Wichman HA. (2000) The end of the LINE?: lack of recent L1 activity in a group of South American rodents. Genetics; 154: 1809-1817.

Chaley MB, Korotkov EV. (2001) Evolution of MIR elements located in the coding regions of human genome. Mol Biol; 35: 1023-1031.

Chan SW, Henderson IR, Jacobsen SE. (2005) Gardening the genome: DNA methylation in Arabidopsis thaliana. Nat. Rev. Genet; 6: 351–360

Chang YF, Imam JS, Wilkinson MF. (2007) The nonsense-mediated decay RNA surveillance pathway. Annu Rev Biochem; 76: 51–74.

Chen JM, Stenson PD, Cooper DN, Ferec C. (2005) A systematic analysis of LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease. Hum Genet; 117: 411-427.

Chen JS, Mehta K. (1999) Tissue transglutaminase: an enzyme with a split personality. Int J Biochem Cell Biol; 31: 817-836. Review

Citores MJ, Rua-Figueroa I, Rodriguez-Gallego C, Durántez A, García-Laorden MI, Rodríguez-Lozano C, Rodríguez-Pérez JC, Vargas JA, Pérez-Aciego P. (2004) The dinucleotide repeat polymorphism in the 3'UTR of the CD154 gene has a functional role on protein expression and is associated with systemic lupus erythematosus. Ann Rheum Dis; 63:310-317.

Citron BA, Suo Z, SantaCruz K, Davies PJ, Qin F, Festoff BW. (2002) Protein crosslinking, tissue transglutaminase, alternative splicing and neurodegeneration. Neurochem Int; 40:69-78.

Claverie J. (2005) Fewer Genes, More Noncoding RNA. Science; 309: 152-1530.

Colllighan RJ, Griffin M. (2009) Transglutaminase 2 cross-linking of matrix proteins: biological significance and medical applications. Amino Acids; 36: 659-670. Review.

Conne B, Stutz A, Vassalli JD. (2000) The 3' untranslated region of messenger RNA: A molecular 'hotspot' for pathology? Nat Med; 6: 637-41.

Corvelo A, Eyras E. (2008) Exon creation and establishment in human genes. Genome Biol. 2008; 9(9):R141.

Cost GJ, Feng Q, Jacquier A, Boeke JD. (2002) Human L1 element target-primed reverse transcription in vitro. EMBO J; 21: 5899-5910.

Costa-Mattioli M, Sossin WS, Klann E, Sonenberg N. (2009) Translational control of long-lasting synaptic plasticity and memory. Neuron; 61: 10-26. Review.

Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. Genome Res; 14: 1188-1190.

Dahl HH. (1995) Pyruvate dehydrogenase E1 alpha deficiency: males and females differ yet again. Am J Hum Genet; 56:553-557. Review.

Dahm R, Kiebler M, Macchi P. (2007) RNA localisation in the nervous system. Semin Cell Dev Biol; 18: 216-223. Review.

Damiani D, Alexander JJ, O'Rourke JR, McManus M, Jadhav AP, Cepko CL, Hauswirth WW, Harfe BD, Strettoi E. (2008) Dicer inactivation leads to progressive functional and structural degeneration of the mouse retina. J Neurosci; 28: 4878-4887.

Daniels GR, Deininger PL. (1985) Repeat sequence families derived from mammalian tRNA genes. Nature; 317:819-822.

Datta S, Antonyak MA, Cerione RA. (2006) Importance of Ca(2+)-dependent transamidation activity in the protection afforded by tissue transglutaminase against doxorubicin-induced apoptosis. Biochemistry; 45: 13163-13174.

David AS. (1994) Schizophrenia and the corpus callosum: developmental, structural and functional relationships. Behav Brain Res; 64: 203-211.

Day NE, Ugai H, Yokoyama KK, Ichiki AT. (2003) K-562 cells lack MHC class II expression due to an alternatively spliced CIITA transcript with a truncated coding region. Leuk Res; 27: 1027-1038.

de la Puente A, Hall J, Wu YZ, Leone G, Peters J, Yoon BJ, Soloway P, Plass C. (2002) Structural characterization of Rasgrf1 and a novel linked imprinted locus. Gene; 291: 287-297.

Deardorff MA, Kaur M, Yaeger D, Rampuria A, Korolev S, Pie J, Gil-Rodríguez C, Arnedo M, Loeys B, Kline AD, Wilson M, Lillquist K, Siu V, Ramos FJ, Musio A, Jackson LS, Dorsett D, Krantz ID. (2007) Mutations in cohesin complex members SMC3 and SMC1A cause a mild variant of cornelia de Lange syndrome with predominant mental retardation. Am J Hum Genet; 80: 485-494.

Degen SJ, Davie EW. (1987) Nucleotide sequence of the gene for human prothrombin. Biochemistry; 26: 6165-6177.

Deininger PL, Batzer MA, Hutchison CA 3rd, Edgell MH. (1992) Master genes in mammalian repetitive DNA amplification. Trends Genet; 8: 307-311.

Deininger PL, Batzer MA. (1999) Alu repeats and human disease. Mol Genet Metab; 67: 183-193.

Deininger PL, Batzer MA. (2002) Mammalian retroelements. Genome Res; 12: 1455-1465.

Deininger PL, Moran JV, Batzer MA, Kazazian HH Jr. (2003) Mobile elements and mammalian genome evolution. Curr Opin Genet Dev; 13: 651-658.

Devor EJ, Peek AS, Lanier W, Samollow PB. (2009) Marsupial-specific microRNAs evolved from marsupial-specific transposable elements. Gene;9: Jul 3. [Epub ahead of print]

Devor EJ. 2006. Primate microRNAs miR-220 and miR-492 lie within processed pseudogenes. J. Hered; 97:186–90

Dewannieux M, Esnault C, Heidmann T. (2003) LINE-mediated retrotransposition of marked Alu sequences. Nat Genet; 35: 41-48.

Diederich NJ, Fénelon G, Stebbins G, Goetz CG. (2009) Hallucinations in Parkinson disease. Nat Rev Neurol; 5:331-342.

Dieterich W, Ehnis T, Bauer M, Donner P, Volta U, Riecken EO, Schuppan D. (1997) Identification of tissue transglutaminase as the autoantigen of celiac disease. Nat Med; 3:797-801.

Donehower LA, Slagle BL, Wilde M, Darlington G, Butel JS. (1989) Identification of a conserved sequence in the non-coding regions of many human genes. Nucleic Acids Res; 17: 699-710.

Dupuis M, Lévy A, Mhaouty-Kodja S. (2004) Functional coupling of rat myometrial alpha 1-adrenergic receptors to Gh alpha/tissue transglutaminase 2 during pregnancy. J Biol Chem; 279:19257-19263.

Dziembowska M, Fondaneche MC, Vedrenne J, Barbieri G, Wiszniewski W, Picard C, Cant AJ, Steimle V, Charron D, Alca-Loridan C, Fischer A, Lisowska-Grospierre B. (2002) Three novel mutations of the CIITA gene in MHC class II-deficient patients with a severe immunodeficiency. Immunogenetics; 53: 821-829.

Eickbush TH. (1992) Transposing without ends: the non-LTR retrotransposable elements. New Biol; 4:430-440.

Emerich DF, Winn SR, Hantraye PM, Peschanski M, Chen EY, Chu Y, McDermott P, Baetge EE, Kordower JH. (1997) Protective effect of encapsulated cells producing neurotrophic factor CNTF in a monkey model of Huntington's disease. Nature; 386: 395-399.

Fedoroff N, Masson P, Banks JA. (1989) Mutations, epimutations, and the developmental programming of the maize Suppressor-mutator transposable element. Bioessays; 10:139-44. Review.

Feng Q, Moran JV, Kazazian HH Jr, Boeke JD. (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell; 87: 905-916.

Festoff BW, SantaCruz K, Arnold PM, Sebastian CT, Davies PJ, Citron BA. (2002) Injury-induced "switch" from GTP-regulated to novel GTP-independent isoform of tissue transglutaminase in the rat spinal cord. J Neurochem; 81: 708-718.

Festoff BW, Suo Z, Citron BA. (2001) Plasticity and stabilization of neuromuscular and CNS synapses: interactions between thrombin protease signaling pathways and tissue transglutaminase. Int Rev Cytol; 211:153-177. Review.

Fesus L, Piacentini M. (2002) Transglutaminase 2: an enigmatic enzyme with diverse functions. Trends Biochem Sci; 27: 534-539.

Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A. (2008) The Pfam protein families database. Nucleic Acids Res; 36 (Database issue): D281-288.

Fraij BM, Gonzales RA. (1996) A third human tissue transglutaminase homologue as a result of alternative gene transcripts.Biochim Biophys Acta; 1306:63-74.

French NS, Norton JD. (1997) Structure and functional properties of mouse VL30 retrotransposons. Biochim Biophys Acta; 1352: 33-47. Review.

Friedman RC, Farh KK, Burge CB, Bartel DP. (2009) Most mammalian mRNAs are conserved targets of microRNAs. Genome Res; 19:92-105.

Gentles AJ, Karlin S. (1999) Why are human G-protein-coupled receptors predominantly intronless? Trends Genet; 15: 47-49.

Giess R, Mäurer M, Linker R, Gold R, Warmuth-Metz M, Toyka KV, Sendtner M, Rieckmann P. (2002) Association of a null mutation in the CNTF gene with early onset of multiple sclerosis. Arch Neurol; 59: 407-409.

Gilbert N, Labuda D. (1999) CORE-SINEs: eukaryotic short interspersed retroposing elements with common sequence motifs. Proc Natl Acad Sci U S A; 96: 2869-2874.

Gnaccarini C, Ben-Tahar W, Lubell WD, Pelletier JN, Keillor JW. (2009) Fluorometric assay for tissue transglutaminase-mediated transamidation activity. Bioorg Med Chem;17: 6354-6359.

Gomi F, Imaizumi K, Yoneda T, Taniguchi M, Mori Y, Miyoshi K, Hitomi J, Fujikado T, Tano Y, Tohyama M. (2000)Molecular cloning of a novel membrane glycoprotein, pal, specifically expressed in photoreceptor cells of the retina and containing leucine-rich repeat. J Neurosci; 20: 3206-3213.

Goodier JL, Ostertag EM, Du K, Kazazian HH Jr. (2001) A novel active L1 retrotransposon subfamily in the mouse. Genome Res; 11: 1677-1685.

Graur D, Shuali Y, Li WH. (1989) Deletions in processed pseudogenes accumulate faster in rodents than in humans. J Mol Evol; 28:279-285.

Greene B, Walko R, Hake S. (1994) Mutator insertions in an intron of the maize knotted1 gene result in dominant suppressible mutations. Genetics; 138:1275-1285.

Gregory TR. (2001) Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. Biol Rev Camb Philos Soc; 76: 65-101. Review.

Grover D, Mukerji M, Bhatnagar P, Kannan K, Brahmachari SK. (2004) Alu repeat analysis in the complete human genome: trends and variations with respect to genomic composition. Bioinformatics; 20: 813-817.

Gu W, Castoe TA, Hedges DJ, Batzer DA, Pollock DD. (2008) Identification of repeat structure in large genomes using repeat probability clouds . Anal Biochem; 380: 77-83.

Guhaniyogi J, Brewer G. (2001) Regulation of mRNA stability in mammalian cells. Gene; 265: 11-23.

Gwynn B, Lueders K, Sands MS, Birkenmeier EH. (1998) Intracisternal A-particle element transposition into the murine beta-glucuronidase gene correlates with loss of enzyme activity: a new model for beta-glucuronidase deficiency in the C3H mouse. Mol Cell Biol; 18: 6474-6481.

Han K, Sen SK, Wang J, Callinan PA, Lee J, Cordaux R, Liang P, Batzer MA. (2005) Genomic rearrangements by LINE-1 insertion-mediated deletion in the human and chimpanzee lineages. Nucleic Acids Res; 33: 4040-4052.

Häsler J, Strub K. (2006) Alu elements as regulators of gene expression. Nucleic Acids Res; 34: 5491-5497.

Hassoun H, Coetzer TL, Vassiliadis JN, Sahr KE, Maalouf GJ, Saad ST, Catanzariti L, Palek J. (1994) A novel mobile element inserted in the alpha spectrin gene: spectrin dayton. A truncated alpha spectrin associated with hereditary elliptocytosis. J Clin Invest; 94:643-648.

Havecker ER, Gao X, Voytas DF. (2004) The diversity of LTR retrotransposons. Genome Biol; 5:225.

Hendrickson A, Bumsted-O'Brien K, Natoli R, Ramamurthy V, Possin D, Provis J. (2008) Rod photoreceptor differentiation in fetal and infant human retina. Exp Eye Res; 87: 415-426.

Hendriksen PJ, Hoogerbrugge JW, Baarends WM, de Boer P, Vreeburg JT, Vos EA, van der LendeT, Grootegoed JA. (1997) Testis-specific expression of a functional retroposon encoding glucose-6-phosphate dehydrogenase in the mouse. Genomics; 41, 350–359.

Hepler JE, Van Wyk JJ, Lund PK. (1990) Different half-lives of insulin-like growth factor I mRNAs that differ in length of 3' untranslated sequence. Endocrinology; 127:1550-1552.

Hosoki K, Ogata T, Kagami M, Tanaka T, Saitoh S. (2008) Epimutation (hypomethylation) affecting the chromosome 14q32.2 imprinted region in a girl with upd(14)mat-like phenotype. Eur J Hum Genet; 16: 1019-1023.

Houck CM, Rinehart FP, Schmid CW. (1979) A ubiquitous family of repeated DNA sequences in the human genome. J Mol Biol; 132:289-306.

Huang DW, Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, Stephens R, Baseler MW, Lane HC, Lempicki RA. (2007) DAVID Gene Functional Classification Tool: A novel biological module-centric algorithm to functionally analyze large gene list. Genome Biol; 8: R183.

Huang HY, Chien CH, Jen KH, Huang HD. (2006) RegRNA: an integrated web server for identifying regulatory RNA motifs and elements. Nucleic Acids Res; 34: 429-34.

Hughes DC. (2000) MIRs as agents of mammalian gene evolution. Trends Genet; 16: 60-62.

Hughes JF, Coffin JM. (2002). A novel endogenous retrovirus-related element in the human genome resembles a DNA transposon: evidence for an evolutionary link? Genomics; 80:453–35.

Hummel S, Hummel M, Banholzer J, Hanak D, Mollenhauer U, Bonifacio E, Ziegler AG. (2007) Development of autoimmunity to transglutaminase C in children of patients with type 1 diabetes: relationship to islet autoantibodies and infant feeding. Diabetologia; 50: 390-394.

Ikura K, Takahata K, Sasaki R. (1993) Cross-linking of a synthetic partial-length (1-28) peptide of the Alzheimer beta/A4 amyloid protein by transglutaminase. FEBS Lett; 326: 109-111.

Ingason A, Sigmundsson T, Steinberg S, Sigurdsson E, Haraldsson M, Magnusdottir BB, Frigge ML, Kong A, Gulcher J, Thorsteinsdottir U, Stefansson K, Petursson H, Stefansson H. (2007) Support for involvement of the AHI1 locus in schizophrenia. Eur J Hum Genet; 15: 988-991.

Jeon CJ, Strettoi E, Masland RH. (1998) The major cell populations of the mouse retina. J Neurosci; 18: 8936-8946.

Jiang X, Zhao Y, Chan WY, Vercauteren S, Pang E, Kennedy S, Nicolini F, Eaves A, Eaves C. (2004) Deregulated expression in Ph+ human leukemias of AHI-1, a gene activated by insertional mutagenesis in mouse models of leukemia. Blood; 103: 3897-3904.

Jordan IK, Rogozin IB, Glazko GV, Koonin EV. (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet; 19: 68-72.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res; 110: 462-467.

Jurka J, Klonowski P. (1996) Integration of retroposable elements in mammals: selection of target sites. J. Mol. Evo; 43: 685–689.

Jurka J, Zietkiewicz E, Labuda D. (1995) Ubiquitous mammalian-wide interspersed repeats (MIRs) are molecular fossils from the mesozoic era. Nucleic Acids Research; 23: 170-175

Jurka J. (2004) Evolutionary impact of human Alu repetitive elements. Curr Opin Genet Dev; 14: 603-608.

Jurka J. (2008) Conserved eukaryotic transposable elements and the evolution of gene regulation. Cell Mol Life Sci; 65: 201-204.

Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y. (2007) KEGG for linking genomes to life and the environment. Nucleic Acids Res; 36 (Database issue):D480-4.

Kang SK, Kim DK, Damron DS, Baek KJ, Im MJ. (2002) Modulation of intracellular Ca (2+) via alpha (1B)-adrenoreceptor signaling molecules, G alpha(h) (transglutaminase II) and phospholipase C-delta 1. Biochem Biophys Res Commun; 293:383-390.

Kapitonov VV, Jurka J. (2003) A novel class of SINE elements derived from 5S rRNA. Mol Biol Evol; 20: 694-702.

Kapitonov VV, Jurka J. (2005) RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. PLOS Biol. 3:e181

Kazazian HH Jr, Wong C, Youssoufian H, Scott AF, Phillips DG, Antonarakis SE. (1988) Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. Nature; 332: 164-166.

Khanna H, Hurd TW, Lillo C, Shu X, Parapuram SK, He S, Akimoto M, Wright AF, Margolis B, Williams DS, Swaroop A. (2005) RPGR-ORF15, which is mutated in retinitis pigmentosa, associates with SMC1, SMC3, and microtubule transport proteins. J Biol Chem; 280: 33580-33587.

Kidwell MG, Lisch DR. (2001) Perspective: transposable elements, parasitic DNA, and genome evolution. Evolution; 55: 1-24.

Kidwell MG. (2002) Transposable elements and the evolution of genome size in eukaryotes. Genetica; 115: 49-63. Review.

Kim DD, Kim TT, Walsh T, Kobayashi Y, Matise TC, Buyske S, Gabriel A. (2004) Widespread RNA editing of embedded alu elements in the human transcriptome. Genome Res; 14: 1719-1725.

Kim E, Goren A, Ast G. (2008) Alternative splicing: current perspectives. Bioessays; 30: 38-47. Review.

Kim ST, Xu B, Kastan MB. (2002) Involvement of the cohesin protein, Smc1, in Atm-dependent and independent responses to DNA damage. Genes Dev; 16: 560-570.

Kleefstra T, Brunner HG, Amiel J, Oudakker AR, Nillesen WM, Magee A, Geneviève D, Cormier-Daire V, van Esch H, Fryns JP, Hamel BC, Sistermans EA, de Vries BB, van Bokhoven H. (2006) Loss-of-function mutations in euchromatin histone methyl transferase 1 (EHMT1) cause the 9q34 subtelomeric deletion syndrome. Am J Hum Genet; 2: 370-377.

Kohany O, Gentles AJ, Hankus L, Jurka J. (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics; 7: 474.

Kolosha VO, Martin SL. (1997) In vitro properties of the first ORF protein from mouse LINE-1 support its role in ribonucleoprotein particle formation during retrotransposition. Proc Natl Acad Sci U S A; 94: 10155-10160.

Kondrashov FA, Koonin EV. (2001) Origin of alternative splicing by tandem exon duplication. Hum Mol Genet; 10:2661-2669.

König H, Matter N, Bader R, Thiele W, Müller F. (2007) Splicing segregation: the minor spliceosome acts outside the nucleus and controls cell proliferation. Cell; 131: 718-729.

Korotkov EV. (1991) A new family of widely propagated MB1-repeats in the human genome. Mol Biol; 25: 250-263.

Krasnov AN, Kurshakova MM, Ramensky VE, Mardanov PV, Nabirochkina EN, Georgieva SG. (2005) A retrocopy of a gene can functionally displace the source gene in evolution. Nucleic Acids Res; 33:6654-6661.

Kremerskothen J, Kindler S, Finger I, Veltel S, Barnekow A. (2006) Postsynaptic recruitment of Dendrin depends on both dendritic mRNA transport and synaptic anchoring. J Neurochem; 96: 1659-1666.

Kriegs JO, Churakov G, Jurka J, Brosius J, Schmitz J. (2007) Evolutionary history of 7SL RNA-derived SINEs in Supraprimates. Trends Genet; 23:158-161.

Krull M, Brosius J, Schmitz J. (2005) Alu-SINE exonization: en route to protein-coding function. Mol Biol Evol; 22: 1702-1711.

Labuda D, Sinnett D, Richer C, Deragon JM, Striker G. (1991) Evolution of mouse B1 repeats: 7SL RNA folding pattern conserved. J Mol Evol; 32: 405-414.

Lafreniere D, Mann N. (2009) Anosmia: loss of smell in the elderly. Otolaryngol Clin North Am; 42: 123-131.

Lai TS, Liu Y, Li W, Greenberg CS. (2007) Identification of two GTP-independent alternatively spliced forms of tissue transglutaminase in human leukocytes, vascular smooth muscle, and endothelial cells. FASEB J; 21:4131-4143.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody M. C et al. (2001) Initial sequencing and analysis of the human genome. Nature; 409: 860-921.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. (2007) Clustal W and Clustal X version 2.0. Bioinformatics; 23:2947-2948.

Larraín J, Bachiller D, Lu B, Agius E, Piccolo S, De Robertis EM. (2000) BMP-binding modules in chordin: a model for signalling regulation in the extracellular space. Development; 127:821-830.

Le SY, Maizel JV Jr. (2007) Data mining of imperfect double-stranded RNA in 3' untranslated regions of eukaryotic mRNAs. Biomol Eng; 24: 351-359.

Lee HJ, Macbeth AH, Pagani JH, Young WS 3rd. (2009) Oxytocin: the great facilitator of life. Prog Neurobiol; 88:127-151.

Lee JH, Barral S, Cheng R, Chacon I, Santana V, Williamson J, Lantigua R, Medrano M, Jimenez-Velazquez IZ, Stern Y, Tycko B, Rogaeva E, Wakutani Y, Kawarai T, St George-Hyslop P, Mayeux R. (2008) Age-at-onset linkage analysis in Caribbean Hispanics with familial late-onset Alzheimer's disease. Neurogenetics; 9: 51-60.

Lee KN, Birckbichler PJ, Patterson MK Jr. (1989) GTP hydrolysis by guinea pig liver transglutaminase. Biochem Biophys Res Commun; 162: 1370-1375.

Lei H, Day IN, Vorechovsky I. (2003) Exonization of AluYa5 in the human ACE gene requires mutations in both 3' and 5' splice sites and is facilitated by a conserved splicing enhancer. Nucleic Acids Res; 33: 3897-3906.

Lesort M, Chun W, Johnson GV, Ferrante RJ. (1999) Tissue transglutaminase is increased in Huntington's disease brain. J Neurochem; 73: 2018-2027.

Levanon EY, Eisenberg E, Yelin R, Nemzer S, Hallegger M, Shemesh R, Fligelman ZY, Shoshan A, Pollock SR, Sztybel D, Olshansky M, Rechavi G, Jantsch MF. (2004) Systematic identification of abundant A-to-I editing sites in the human transcriptome. Nat Biotechnol; 22:1001-1005.

Lev-Maor G, Sorek R, Shomron N, Ast G. (2003) The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons. Science; 300: 1288-1291.

Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB. (2003) Prediction of mammalian microRNA targets. Cell; 115: 787-798.

Lewis SM, Wu GE. (1997) The origins of V(D)J recombination. Cell; 88: 159-162. Review

Li J, Bench AJ, Vassiliou GS, Fourouclas N, Ferguson-Smith AC, Green AR. (2004) Imprinting of the human L3MBTL gene, a polycomb family member located in a region of chromosome 20 deleted in human myeloid malignancies. Proc Natl Acad Sci U S A; 101: 7341-7346.

Li X, Scaringe WA, Hill KA, Roberts S, Mengos A, Careri D, Pinto MT, Kasper CK, Sommer SS. (2001) Frequency of recent retrotransposition events in the human factor IX gene. Hum Mutat; 17: 511-5119.

Limon-Boulez I, Mhaouty-Kodja S, Coudouel N, Benoit de Coignac A, Legrand C, Maltier JP. (1997) The alpha1B-adrenergic receptor subtype activates the phospholipase C signaling pathway in rat myometrium at parturition. Biol Reprod; 57:1175-1182.

Lin CW, Ting AY. (2006) Transglutaminase-catalyzed site-specific conjugation of small-molecule probes to proteins in vitro and on the surface of living cells. J Am Chem Soc; 128: 4542-4543

Lin L, Jiang P, Shen S, Sato S, Davidson BL, Xing Y. (2009) Large-scale analysis of exonized mammalian-wide interspersed repeats in primate genomes. Hum Mol Genet;18: 2204-2214.

Lin PY, Tsai G. (2004) Meta-analyses of the association between genetic polymorphisms of neurotrophic factors and schizophrenia. Schizophr Res; 71:353-360.

Lindblad-Toh K, Wade CM, Mikkelsen TS *et al.* (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature; 438: 803-819.

Lippman Z, Gendrel AV, Black M, Vaughn MW, Dedhia N, McCombie WR, Lavine K, Mittal V, May B, et al. (2004) Role of transposable elements in heterochromatin and epigenetic control. Nature; 430: 471–476.

Liu H, Wang M, Xia CH, Du X, Flannery JG, Ridge KD, Beutler B, Gong X. (2009) A Novel Rhodopsin Mutation Causes Severe Retinal Degeneration. Invest Ophthalmol Vis Sci. 2009 Sep 9. [Epub ahead of print].

Liu Q, Tan G, Levenkova N, Li T, Pugh EN Jr, Rux JJ, Speicher DW, Pierce EA. (2007) The proteome of the mouse photoreceptor sensory cilium complex. Mol Cell Proteomics; 6: 1299-1317.

Loscher CJ, Hokamp K, Kenna PF, Ivens AC, Humphries P, Palfi A, Farrar GJ. (2007) Altered retinal microRNA expression profile in a mouse model of retinitis pigmentosa. Genome Biol; 8: R248.

Loscher CJ, Hokamp K, Wilson JH, Li T, Humphries P, Farrar GJ, Palfi A. (2008) A common microRNA signature in mouse models of retinal degeneration. Exp Eye Res; 87: 529-534.

Luan DD, Korman MH, Jakubczak JL, Eickbush TH. (1993) Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell; 72: 595-605.

Luedi PP, Dietrich FS, Weidman JR, Bosko JM, Jirtle RL, Hartemink AJ. (2007) Computational and experimental identification of novel human imprinted genes. Genome Res; 17: 1723-1730.

Luehrsen KR, Walbot V. (1990) Insertion of Mu1 elements in the first intron of the Adh1-S gene of maize results in novel RNA processing events. Plant Cell; 2:1225-1238.

Lugli G, Larson J, Martone ME, Jones Y, Smalheiser NR. (2005) Dicer and eIF2c are enriched at postsynaptic densities in adult mouse brain and are modified by neuronal activity in a calpain-dependent manner. J Neurochem; 94: 896-905.

Mager DL, Freeman JD. (2000) Novel mouse type D endogenous proviruses and ETn elements share long terminal repeat and internal sequences. J Virol; 74: 7221-7229.

Mah N, Stoehr H, Schulz HL, White K, Weber BH. (2001) Identification of a novel retina-specific gene located in a subtelomeric region with polymorphic distribution among multiple human chromosomes. Biochim Biophys Acta; 1522: 167-174.

Maiuri L, Ciacci C, Ricciardelli I, Vacca L, Raia V, Rispo A, Griffin M, Issekutz T, Quaratino S, Londei M. (2005) Unexpected role of surface transglutaminase type II in celiac disease. Gastroenterology; 129: 1400-1413.

Maiuri L, Luciani A, Giardino I, Raia V, Villella VR, D'Apolito M, Pettoello-Mantovani M, Guido S, Ciacci C, Cimmino M, Cexus ON, Londei M, Quaratino S. (2008) Tissue transglutaminase activation modulates inflammation in cystic fibrosis via PPARgamma down-regulation. J Immunol; 180: 7697-7705.

Makałowski W. (2000) Genomic scrap yard: how genomes utilize all that junk. Gene; 259: 61-67.

Marmur J, Doty P. (1961) Thermal renaturation of deoxyribonucleic acids. J Mol Biol; 3: 585–594.

Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H. (2005) Emergence of young human genes after a burst of retroposition in primates. PLoS Biol; 3: e357

Mastroberardino PG, Iannicola C, Nardacci R, Bernassola F, De Laurenzi V, Melino G, Moreno S, Pavone F, Oliverio S, Fesus L, Piacentini M. (2002) 'Tissue' transglutaminase ablation reduces neuronal death and prolongs survival in a mouse model of Huntington's disease. Cell Death Differ; 9: 873-880.

Matsui M, Ichihara H, Kobayashi S, Tanaka H, Tsuchida J, Nozaki M, Yoshimura Y, Nojima H, Rochelle JM, Nishimune Y, Taketo MM, Seldin MF. (1997) Mapping of six germ-cell-specific genes to mouse chromosomes. Mamm Genome; 8: 873-874.

Mattick JS, Makunin IV. (2006) Non-coding RNA. Hum Mol Genet; 15: R17-29. Review.

Maubaret C, Griffoin JM, Arnaud B, Hamel C. (2005) Novel mutations in MYO7A and USH2A in Usher syndrome. Ophthalmic Genet; 26: 25-29.

Mauget-Faÿsse M, Vuillaume M, Quaranta M, Moullan N, Angèle S, Friesen MD, Hall J. (2003) Idiopathic and radiation-induced ocular telangiectasia: the involvement of the ATM gene. Invest Ophthalmol Vis Sci; 44: 3257-3262.

Mavrou A, Tsangaris GT, Roma E, Kolialexi A. (2008) The ATM gene and ataxia telangiectasia. Anticancer Res; 28: 401-405.

Mayford M, Baranes D, Podsypanina K, Kandel ER. (1996) The 3'-untranslated region of CaMKII alpha is a cis-acting signal for the localization and translation of mRNA in dendrites. Proc Natl Acad Sci USA; 93: 13250-13255.

McCarrey JR, Kumari M, Aivaliotis MJ, Wang Z, Zhang P, Marshall F, Vandeberg JL. (1996) Analysis of the cDNA and encoded protein of the human testis-specific PGK-2 gene. Dev Genet; 19: 321-332.

McCart AE, Mahony D, Rothnagel JA. (2003) Alternatively spliced products of the human kinesin light chain 1 (KNS2) gene. Traffic; 4: 576-580.

McCarthy BI, Bolton CT. (1963) An approach to the measurement of genetic relatedness among organisms. Proc. Not. Acod. Sri. U S A: 90: 156-64.

McClintock B (1961) Some parallels between gene control systems in maize and in bacteria. Am Naturalist; 95:265–77

McClintock B. (1944) The Relation of Homozygous Deficiencies to Mutations and Allelic Series in Maize. Genetics; 29:478-502.

McClintock B. (1950) The origin and behaviour of mutable loci in maize. Proc Natl Acad Sci U S A; 36: 344–355.

McClure MA. (1991) Evolution of retroposons by acquisition or deletion of retrovirus-like genes. Mol Biol Evol; 8: 835-856.

McEwen DP, Gee KR, Kang HC, Neubig RR. (2002) Fluorescence approaches to study G protein mechanisms. Methods Enzymol; 344: 403-420.

McGuire JR, Rong J, Li SH, Li XJ. (2006) Interaction of Huntingtin-associated protein-1 with kinesin light chain: implications in intracellular trafficking in neurons. J Biol Chem; 281: 3552-3559.

Mears AJ, Kondo M, Swain PK, Takada Y, Bush RA, Saunders TL, Sieving PA, Swaroop A. (2001) Nrl is required for rod photoreceptor development. Nat. Genet; 29: 447–452.

Medstrand P, Mager DL. (1998) Human-specific integrations of the HERV-K endogenous retrovirus family. J Virol; 72: 9782-9787.

Medstrand P, van de Lagemaat LN, Dunn CA, Landry JR, Svenback D, Mager DL. (2005) Impact of transposable elements on the evolution of mammalian gene regulation. Cytogenet Genome Res; 110:342-352.

Medstrand P, van de Lagemaat LN, Mager DL. (2002) Retroelement distributions in the human genome: variations associated with age and proximity to genes. Genome Res; 12: 1483-1495.

Melino G, Piacentini M. (1998) 'Tissue' transglutaminase in cell death: a downstream or a multifunctional upstream effector? FEBS Lett; 430:59-63. Review.

Mercer TR, Dinger ME, Mariani J, Kosik KS, Mehler MF, Mattick JS.(2009) Noncoding RNAs in Long-Term Memory Formation. Neuroscientist; 14: 434-45. Review.

Mersch B, Sela N, Ast G, Suhai S, Hotz-Wagenblatt A. (2007) SERpredict: detection of tissue- or tumor-specific isoforms generated through exonization of transposable elements. BMC Genet; 8:78.

Mhaouty-Kodja S.(2004) Ghα/tissue transglutaminase 2: an emerging G protein in signal transduction. Biol Cell; 96: 363-367.

Mignone F, Grillo G, Licciulli F, Iacono M, Liuni S, Kersey PJ, Duarte J, Saccone C, Pesole G. (2005) UTRdb and UTRsite: a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. Nucleic Acids Res; 33: 141-146.

Mikkelsen TS, Wakefield MJ, Aken B, et al. (2007) Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. Nature; 447: 167-177.

Miller WJ, McDonald JF, Nouaud D, Anxolabéhère D. (1999) Molecular domestication – more than a sporadic episode in evolution. Genetica; 107: 197–207.

Millet C, Lemaire P, Orsetti B, Guglielmi P, François V.(2001)The human chordin gene encodes several differentially expressed spliced variants with distinct BMP opposing activities. Mech Dev; 106:85-96.

Minakami R, Kurose K, Etoh K, Furuhata Y, Hattori M, Sakaki Y. (1992) Identification of an internal cis-element essential for the human L1 transcription and a nuclear factor(s) binding to the element. Nucleic Acids Res; 20: 3139-3145.

Mishra S, Melino G, Murphy LJ. (2007) Transglutaminase 2 kinase activity facilitates protein kinase A-induced phosphorylation of retinoblastoma protein. J Biol Chem; 282: 18108-18115.

Mitchell GA, Labuda D, Fontaine G, Saudubray JM, Bonnefont JP, Lyonnet S, Brody LC, Steel G, Obie C, Valle D. (1991) Splice-mediated insertion of an Alu sequence inactivates ornithine delta-aminotransferase: a role of Alu elements in human mutation. Proc Natl Acad Sci USA; 88: 815-819.

Mitchell JD, Borasio JD. (2007) Amyotrophic lateral sclerosis. Lancet; 369: 2031–2041

Mitton KP, Swain PK, Chen S, Xu S, Zack DJ, Swaroop A. (2000) The leucine zipper of NRL interacts with the CRX homeodomain. A possible mechanism of transcriptional synergy in rhodopsin regulation. J Biol Chem; 275: 29794-29799.

Miyoshi N, Wagatsuma H, Wakana S, Shiroishi T, Nomura M, Aisaka K, Kohda T, Surani MA, Kaneko-Ishino T, Ishino F. (2000) Identification of an imprinted gene, Meg3/Gtl2 and its human homologue MEG3, first mapped on mouse distal chromosome 12 and human chromosome 14q. Genes Cells; 5: 211-220.

Mohr E, Richter D. (2001) Messenger RNA on the move: implications for cell polarity. Int J Biochem Cell Biol; 33: 669-679.

Möller CG, Kimberling WJ, Davenport SL, Priluck I, White V, Biscone-Halterman K, Odkvist LM, Brookhouser PE, Lund G, Grissom TJ. (1989) Usher syndrome: an otoneurologic study. Laryngoscope; 99:73-79.

Monsonego A, Friedmann I, Shani Y, Eisenstein M, Schwartz M. (1998) GTP-dependent conformational changes associated with the functional switch between Galpha and cross-linking activities in brain-derived tissue transglutaminase. J Mol Biol; 282: 713-720.

Monsonego A, Shani Y, Friedmann I, Paas Y, Eizenberg O, Schwartz M. (1997) Expression of GTP-dependent and GTP-independent tissue-type transglutaminase in cytokine-treated rat brain astrocytes. J Biol Chem; 272: 3724-3732.

Moran JV, DeBerardinis RJ, Kazazian HH Jr. (1999) Exon shuffling by L1 retrotransposition. Science; 283: 1530-1534.

Morgan HD, Sutherland HG, Martin DI, Whitelaw E. (1999) Epigenetic inheritance at the agouti locus in the mouse. Nat. Genet; 23: 314–318.

Mori Y, Imaizumi K, Katayama T, Yoneda T, Tohyama M. (2000) Two cis-acting elements in the 3' untranslated region of α-CaMKII regulate its dendritic targeting. Nature Neurosci; 3: 1079-1084.

Morison IM, Ramsay JP, Spencer HG. (2005) A census of mammalian imprinting. Trends Genet; 21: 457-465.

Muotri AR, Marchetto MC, Coufal NG, Gage FH. (2007). The necessary junk: new functions for transposable elements. Hum Mol Genet. 2007 Oct 15; 16 Spec No. 2:R159-67. Review.

Murthy SN, Iismaa S, Begg G, Freymann DM, Graham RM, Lorand L. (2002) Conserved tryptophan in the core domain of transglutaminase is essential for catalytic activity. Proc Natl Acad Sci U S A; 99:2738-2742.

Musio A, Selicorni A, Focarelli ML, Gervasini C, Milani D, Russo S, Vezzoni P, Larizza L. (2006) X-linked Cornelia de Lange syndrome owing to SMC1L1 mutations. Nat Genet; 38: 528-530.

Nakamura H, Yoshida M, Tsuiki H, Ito K, Ueno M, Nakao M, Oka K, Tada M, Kochi M, Kuratsu J, Ushio Y, Saya H. (1998) Identification of a human homolog of the Drosophila neuralized gene within the 10q25.1 malignant astrocytoma deletion region. Oncogene; 16: 1009-1019.

Narita N, Nishio H, Kitoh Y, Ishikawa Y, Ishikawa Y, Minami R, Nakamura H, Matsuo M. (1993) Insertion of a 5' truncated L1 element into the 3' end of exon 44 of the dystrophin gene resulted in skipping of the exon during splicing in a case of Duchenne muscular dystrophy. J Clin Invest; 91: 1862-1867.

Nekrutenko A, Li WH. (2001) Transposable elements are found in a large number of human protein-coding genes. Trends Genet; 17:619-621

Nikaido M, Matsuno F, Abe H, Shimamura M, Hamilton H, Matsubayashi H, Okada N. (2001) Evolution of CHR-2 SINEs in cetartiodactyl genomes: possible evidence for the monophyletic origin of toothed whales. Mamm Genome; 12: 909-915.

Nurtdinov RN, Mironov AA, Gelfand MS. (2009) Rodent-specific alternative exons are more frequent in rapidly evolving genes and in paralogs. BMC Evol Biol; 9: 142.

Oh EC, Cheng H, Hao H, Jia L, Khan NW, Swaroop A. (2008) Rod differentiation factor NRL activates the expression of nuclear receptor NR2E3 to suppress the development of cone photoreceptors. Brain Res; 1236: 16-29.

Ohno S. (1972) So much “junk” DNA in our genome. Brookhaven Symp Biol; 23: 366-370

Okada N, Hamada M. (1997) The 3' ends of tRNA-derived SINEs originated from the 3' ends of LINEs: a new example from the bovine genome. J Mol Evol; 44 Suppl 1:S52-56.

Ono M, Kawakami M, Takezawa T. (1987) A novel human nonviral retroposon derived from an endogenous retrovirus. Nucleic Acids Res; 15: 8725-8737.

Ostertag EM, Goodier JL, Zhang Y, Kazazian HH Jr. (2003) SVA elements are non-autonomous retrotransposons that cause disease in humans. Am J Hum Genet; 73: 1444-1451.

Ostertag EM, Kazazian HH Jr. (2001) Biology of mammalian L1 retrotransposons. Annu Rev Genet; 35: 501-538.

Ovcharenko I, Nobrega MA, Loots GG, Stubbs L. (2004) ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes. Nucleic Acids Res; 32(Web Server issue):W280-6.

Pan D, Zhang L. (2009) Burst of young retrogenes and independent retrogene formation in mammals. PLoS One; 4:e5040. Epub 2009 Mar 27.

Pant PV, Tao H, Beilharz EJ, Ballinger DG, Cox DR, Frazer KA.(2006) Analysis of allelic differential expression in human white blood cells. Genome Res; 16: 331-339.

Parisi MA, Doherty D, Chance PF, Glass IA. (2007) Joubert syndrome (and related disorders) (OMIM 213300). Eur J Hum Genet; 15: 511-521.

Park ES, Won JH, Han KJ, Suh PG, Ryu SH, Lee HS, Yun HY, Kwon NS, Baek KJ. (1998) Phospholipase C-delta1 and oxytocin receptor signalling: evidence of its role as an effector. Biochem J; 331: 283-289.

Patrushev LI, Minkevich IG. (2008) The problem of the eukaryotic genome size. Biochemistry (Mosc); 73: 1519-1552.

Pedersen ED, Aass HC, Rootwelt T, Fung M, Lambris JD, Mollnes TE. (2007) CD59 efficiently protects human NT2-N neurons against complement-mediated damage.Scand J Immunol; 66: 345-351.

Pennings RJ, Fields RR, Huygen PL, Deutman AF, Kimberling WJ, Cremers CW. (2003) Usher syndrome type III can mimic other types of Usher syndrome. Ann Otol Rhinol Laryngol; 11: 525-530.

Pepe IM. (2001) Recent advances in our understanding of rhodopsin and phototransduction. Prog Retin Eye Res; 20: 733-759. Review.

Peterson PA. (1981) Instability among the components of a regulatory element transposon in maize. Cold Spring Harb Symp Quant Biol; 2:447-55.

Petrov DA. (2001) Evolution of genome size: new approaches to an old problem. Trends Genet; 17: 23-28.

Pickard BS, Knight HM, Hamilton RS, Soares DC, Walker R, Boyd JK, Machell J, Maclean A, McGhee KA, Condie A, Porteous DJ, St Clair D, Davis I, Blackwood DH, Muir WJ. (2008) A common variant in the 3'UTR of the GRIK4 glutamate receptor gene affects transcript abundance and protects against bipolar disorder.Proc Natl Acad Sci U S A; 105: 14940-14945.

Pinkstaff JK, Chappell SA, Mauro VP, Edelman GM, Krushel LA. (2001) Internal initiation of translation of five dendritically localized neuronal mRNAs. Proc Natl Acad Sci U S A; 98: 2770-2775.

Piriyapongsa J, Jordan IK. (2008) Dual coding of siRNAs and miRNAs by plant transposable elements. RNA; 14: 814-821.

Piriyapongsa J, Marino-Ramirez L, Jordan IK. (2007) Origin and evolution of human microRNAs from transposable elements. Genetics; 176: 1323-1337.

Piriyapongsa J, Polavarapu N, Borodovsky M, McDonald J. (2007) Exonization of the LTR transposable elements in human genome. BMC Genomics; 8: 291.

Piskurek O, Austin CC, Okada N. (2006) Sauria SINEs: Novel short interspersed retroposable elements that are widespread in reptile genomes. J Mol Evol; 62: 630-644.

Piskurek O, Nishihara H, Okada N. (2008) The evolution of two partner LINE/SINE families and a full-length chromodomain-containing Ty3/Gypsy LTR element in the first reptilian whole-genome of Anolis carolinensis. Gene; 441: 111-118.

Polavarapu N, Mariño-Ramírez L, Landsman D, McDonald JF, Jordan IK. (2008) Evolutionary rates and patterns for human transcription factor binding sites derived from repetitive DNA. BMC Genomics; 9: 226.

Pontius JU, Mullikin JC, Smith DR et al. (2007) Initial sequence and comparative analysis of the cat genome. Genome Res; 17:1675-1689.

Porzio O, Massa O, Cunsolo V et al. (2007) Missense mutations in the TGM2 gene encoding transglutaminase 2 are found in patients with early-onset type 2 diabetes. Hum Mutat; 28: 1150.

Rahner C, Fukuhara M, Peng S, Kojima S, Rizzolo LJ. (2004) The apical and basal environments of the retinal pigment epithelium regulate the maturation of tight junctions during development. J Cell Sci; 117: 3307-3318.

Rasschaert J, Flatt PR, Barnett CR, McClenaghan NH, Malaisse WJ. (1996) D-glucose metabolism in BRIN-BD11 islet cells. Biochem Mol Med; 57:97-105.

Reese MG, Eeckman FH, Kulp D, Haussler D. (1997) Improved splice site detection in Genie. J Comput Biol; 4: 311-323.

Riley JL, Westerheide SD, Price JA, Brown JA, Boss JM. (1995) Activation of class II MHC genes requires both the X box region and the class II transactivator (CIITA). Immunity; 2: 533-543.

Rosskopf D, Manthey I, Habich C, Kielbik M, Eisenhardt A, Nikula C, Urban M, Kohnen S, Graf E, Ravens U, Siffert W. (2003) Identification and characterization of G beta 3s2, a novel splice variant of the G-protein beta 3 subunit. Biochem J; 371: 223-232.

Roy AM, West NC, Rao A, Adhikari P, Aleman C, Barnes AP, Deininger PL. (2000) Upstream flanking sequences and transcription of SINEs. J Mol Biol; 302: 17-25.

Roy-Engel AM, Carroll ML, El-Sawy M, Salem AH, Garber RK, Nguyen SV, Deininger PL, Batzer MA. (2002) Non-traditional Alu evolution and primate genomic diversity. J Mol Biol; 316: 1033-1040.

Royo H, Cavaillé J. (2008) Non-coding RNAs in imprinted gene clusters. Biol Cell; 100: 149-66.

Ruan Q, Johnson GV. (2007) Transglutaminase 2 in neurodegenerative disorders. Front Biosci; 12:891-904.

Rump A, Rosen-Wolff A, Gahr M, Seidenberg J, Roos C, Walter L, Gunther V, Roesler J. (2006) A splice-supporting intronic mutation in the last bp position of a cryptic exon within intron 6 of the CYBB gene induces its incorporation into the mRNA causing chronic granulomatous disease (CGD). Gene; 371: 174-181.

Salem AH, Ray DA, Batzer MA. (2005) Identity by descent and DNA sequence variation of human SINE and LINE elements. Cytogenet Genome Res; 108: 63-72.

Salem AH, Ray DA, Xing J, Callinan PA, Myers JS, Hedges DJ, Garber RK, Witherspoon DJ, Jorde LB, Batzer MA. (2003) Alu elements and hominid phylogenetics. Proc Natl Acad Sci U S A; 100: 12787-12791.

Santangelo AM, de Souza FS, Franchini LF, Bumaschny VF, Low MJ, Rubinstein M. (2007) Ancient exaptation of a CORE-SINE retroposon into a highly conserved mammalian neuronal enhancer of the proopiomelanocortin gene. PLoS Genet; 3:1813-1826.

Schatz DG (1999) Transposition mediated by RAG1 and RAG2 and the evolution of the adaptive immune system. Immunol Res; 19: 169-182.

Schatz DG, Oettinger MA, Baltimore D. (1989) The V(D)J recombination activating gene, RAG-1. Cell; 59: 1035–1048.

Schratt GM, Tuebing F, Nigh EA, Kane CG, Sabatini ME, Kiebler M, Greenberg ME. (2006) A brain-specific microRNA regulates dendritic spine development. Nature; 439: 283-289.

Schulz R, Menheniott TR, Woodfine K, Wood AJ, Choi JD, Oakey RJ. (2006) Chromosome-wide identification of novel imprinted genes using microarrays and uniparental disomies. Nucleic Acids Res; 34: e88.

Sedlacek Z, Munstermann E, Dhorne-Pollet S, Otto C, Bock D, Schutz G, Poustka A. (1999) Human and mouse XAP-5 and XAP-5-like (X5L) genes: identification of an ancient functional retroposon differentially expressed in testis. Genomics; 61: 125-132.

Sela N, Mersch B, Gal-Mark N, Lev-Maor G, Hotz-Wagenblatt A, Ast G. (2007) Comparative analysis of transposed element insertion within human and mouse genomes reveals Alu's unique role in shaping the human transcriptome. Genome Biol; 8: R127.

Serra L, Cercignani M, Lenzi D, Perri R, Fadda L, Caltagirone C, Macaluso E, Bozzali M. (2009) Grey and White Matter Changes at Different Stages of Alzheimer's Disease. J Alzheimers Dis; Sept 11. [Epub ahead of print].

Shedlock AM, Okada N. (2000) SINE insertions: powerful tools for molecular systematics. Bioessays; 22: 148-160.

Shedlock AM, Takahashi K, Okada N. (2004) SINEs of speciation: tracking lineages with retroposons. Trends Ecol Evol; 19: 545-553.

Shen L, Wu LC, Sanlioglu S, Chen R, Mendoza AR, Dangel AW, Carroll MC, Zipf WB, Yu CY. (1994) Structure and genetics of the partially duplicated gene RP located immediately upstream of the complement C4A and the C4B genes in the HLA class III region. Molecular cloning, exon-intron structure, composite retroposon, and breakpoint of gene duplication. J Biol Chem; 269: 8466-8476.

Shen MR, Batzer MA, Deininger PL. (1991) Evolution of the master Alu gene(s). J Mol Evol; 33: 311-320.

Sheng G, Xu X, Lin YF, Wang CE, Rong J, Cheng D, Peng J, Jiang X, Li SH, Li XJ. (2008) Huntingtin-associated protein 1 interacts with Ahi1 to regulate cerebellar and brainstem development in mice. J Clin Invest; 118:2785-2795.

Shiao MS, Khil P, Camerini-Otero RD, Shiroishi T, Moriwaki K, Yu HT, Long M. (2007) Origins of new male germ-line functions from X-derived autosomal retrogenes in the mouse. Mol Biol Evol; 24:2242-2253.

Shimamura M, Yasue H, Ohshima K, Abe H, Kato H, Kishiro T, Goto M, Munechika I, Okada N. (1997) Molecular evidence from retroposons that whales form a clade within even-toed ungulates. Nature; 388: 666-70.

Simell O, Takki K. (1973) Raised plasma ornithine and gyrate atrophy of the choroid and retina. Lancet; I: 1031-1033.

Simms RJ, Eley L, Sayer JA. (2009) Nephronophthisis. Eur J Hum Genet; 17:406-416.

Singer CF, Hudelist G, Walter I, Rueckliniger E, Czerwenka K, Kubista E, Huber AV. (2006) Tissue array-based expression of transglutaminase-2 in human breast and ovarian cancer. Clin Exp Metastasi; 23: 33-39.

Sinzelle L, Izsvák Z, Ivics Z. (2009) Molecular domestication of transposable elements: from detrimental parasites to useful host genes. Cell Mol Life Sci; 66: 1073-1093.

Sisk TJ, Roys S, Chang CH. (2001) Self-association of CIITA and its transactivation potential. Mol Cell Biol; 21: 4919-4928.

Slaughter TF, Achyuthan KE, Lai TS, Greenberg CS.(1992) A microtitre plate transglutaminase assay using (5-biotinamido)pentylamine as substrate. Anal. Biochem; 205: 166–171.

Smalheiser NR, Torvik VI. (2005) Mammalian microRNAs derived from genomic repeats. Trends Genet; 21: 322-326.

Smalheiser NR, Torvik VI. (2006) Alu elements within human mRNAs are probable microRNA targets. Trends Genet; 22: 532-6.

Smit AF, Riggs AD. (1995) MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. Nucleic Acids Res; 23: 98-102.

Smith AM, Sanchez MJ, Follows GA, Kinston S, Donaldson IJ, Green AR, Göttgens B. (2008) A novel mode of enhancer evolution: the Tal1 stem cell enhancer recruited a MIR element to specifically boost its activity. Genome Res; 18: 1422-1432.

Smith TJ, Peterson PE, Schmidt T, Fang J, Stanley CA. (2001) Structures of bovine glutamate dehydrogenase complexes elucidate the mechanism of purine regulation. J Mol Biol; 307: 707-20.

Sorek R, Ast G, Graur D. (2002) Alu-containing exons are alternatively spliced. Genome Res; 12: 1060-7.

Sossin WS, DesGroseillers L. (2006) Intracellular trafficking of RNA in neurons. Traffic; 7: 1581-1589. Review.

Spanopoulou E, Zaitseva F, Wang FH, Santagata S, Baltimore D, Panayotou G. (1996) The homeodomain region of Rag-1 reveals the parallel mechanisms of bacterial and V(D)J recombination. Cell. 87: 263-276.

St Johnston.D. (1995) The intracellular localization of messenger RNAs. Cell; 81: 161-170.

Steimle V, Otten LA, Zufferey M, Mach B. (1993) Complementation cloning of an MHC class II transactivator mutated in hereditary MHC class II deficiency (or bare lymphocyte syndrome). Cell; 75: 135-146.

Storz G, Opdyke JA, Zhang A. (2004) Controlling mRNA stability and translation with small, non-coding RNAs. Curr Opin Microbiol; 7: 140–144.

Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A; 101: 6062-6067.

Suzuki S, Ono R, Narita T, Pask AJ, Shaw G, Wang C, Kohda T, Alsop AE, Marshall Graves JA, Kohara Y, Ishino F, Renfree MB, Kaneko-Ishino T. (2007) Retrotransposon Silencing by DNA Methylation Can Drive Mammalian Genomic Imprinting. PLoS Genet; 3 :e55

Svoboda P, Di Cara A. (2006) Hairpin RNA: a secondary structure of primary importance. Cell Mol Life Sci; 63:901-908.

Swanberg M, Lidman O, Padyukov L, Eriksson P, Akesson E, Jagodic M, Lobell A, Khademi M, Börjesson O, Lindgren CM, Lundman P, Brookes AJ, Kere J, Luthman H, Alfredsson L, Hillert J, Klareskog L, Hamsten A, Piehl F, Olsson T. (2005) MHC2TA is associated with differential MHC molecule expression and susceptibility to rheumatoid arthritis, multiple sclerosis and myocardial infarction. Nat Genet; 37: 486-494.

Szmulewicz MN, Novick GE, Herrera RJ. (1998) Effects of Alu insertions on gene function. Electrophoresis; 19: 1260-1264.

Takahashi K, Okada N. (2002) Mosaic structure and retropositional dynamics during evolution of subfamilies of short interspersed elements in African cichlids. Mol Biol Evol; 19: 1303-1312.

Telci D, Griffin M. (2006) Tissue transglutaminase (TG2)--a wound response enzyme. Front Biosci; 11:867-882. Review.

Terai Y, Takahashi K, Nishida M, Sato T, Okada N. (2003) Using SINEs to probe ancient explosive speciation: "hidden" radiation of African cichlids? Mol Biol Evol; 20: 924-930.

Terai Y, Takahashi K, Okada N. (1998) SINE cousins: the 3'-end tails of the two oldest and distantly related families of SINEs are descended from the 3' ends of LINEs with the same genealogical origin. Mol Biol Evol; 15: 1460-1471.

Terai Y, Takezaki N, Mayer WE, Tichy H, Takahata N, Klein J, Okada N. (2004) Phylogenetic relationships among East African haplochromine fish as revealed by short interspersed elements (SINEs). J Mol Evol; 58: 64-78.

Thakker MM, Huang J, Possin DE, Ahmadi AJ, Mudumbai R, Orcutt JC, Tarbet KJ, Sires BS. (2008) Human orbital sympathetic nerve pathways. Ophthal Plast Reconstr Surg; 24: 360-366.

Thomson SJ, Goh FG, Banks H, Krausgruber T, Kotenko SV, Foxwell BM, Udalova IA. (2009) The role of transposable elements in the regulation of IFN-{lambda}1 gene expression. Proc Natl Acad Sci U S A. 2009 Jul 1. [Epub ahead of print]

Timmusk T, Palm K, Belluardo N, Mudò G, Neuman T. (2002) Dendritic localization of mammalian neuralized mRNA encoding a protein with transcription repression activities. Cell Neurosci; 20: 649-668.

Ting CN, Rosenberg MP, Snow CM, Samuelson LC, Meisler MH. (1992) Endogenous retroviral sequences are required for tissue-specific expression of a human salivary amylase gene. Genes Dev; 6: 1457-1465

Tobimatsu T, Fujisawa H. (1989) Tissue-specific expression of four types of rat calmodulin-dependent protein kinase II mRNAs. J Biol Chem; 264: 17907-17912.

Tolentino PJ, Waghray A, Wang KK, Hayes RL. (2004) Increased expression of tissue-type transglutaminase following middle cerebral artery occlusion in rats. J Neurochem; 89:1301-1307.

Tory K, Lacoste T, Burglen L, Moriniere V, Boddaert N, Macher MA, Llanas B, Nivet H, Bensman A, Niaudet P, Antignac C, Salomon R, Saunier S. (2007) High NPHP1 and NPHP6 Mutation Rate in Patients with Joubert Syndrome and Nephronophthisis: Potential Epistatic Effect of NPHP6 and AHI1 Mutations in Patients with NPHP1 Mutations. J Am Soc Nephrol; 18: 1566-1575.

Treharne KJ, Giles Best O, Mehta A. (2009) Transglutaminase 2 and nucleoside diphosphate kinase activity are correlated in epithelial membranes and are abnormal in cystic fibrosis. FEBS Lett; 583:2789-2792.

Tribe RM. (2001) Regulation of human myometrial contractility during pregnancy and labour: are calcium homeostatic pathways important? Exp Physiol; 86: 247-254. Review.

Tulko JS, Korotkov EV, Pheonix DA. (1997) MIRs are present in coding regions of human genes. DNA Seq; 8: 31-38.

Uechi T, Maeda N, Tanaka T, Kenmochi N. (2002) Functional second genes generated by retrotransposition of the X-linked ribosomal protein genes. Nucleic Acids Res; 30: 5369-5375.

Utton MA, Noble WJ, Hill JE, Anderton BH, Hanger DP. (2005) Molecular motors implicated in the axonal transport of tau and alpha-synuclein. J Cell Sci; 118: 4645-4654.

Valente EM, Brancati F, Dallapiccola B. (2008) Genotypes and phenotypes of Joubert syndrome and related disorders. Eur J Med Genet; 51:1-23.

Valente EM, Brancati F, Silhavy JL, Castori M, Marsh SE, Barrano G, Bertini E, Boltshauser E, Zaki MS, Abdel-Aleem A, Abdel-Salam GM, Bellacchio E, Battini R, Cruse RP, Dobyns WB, Krishnamoorthy KS, Lagier-Tourenne C, Magee A, Pascual-Castroviejo I, Salpietro CD, Sarco D, Dallapiccola B, Gleeson JG. (2006) International JSRD Study Group. AHI1 gene mutations cause specific forms of Joubert syndrome-related disorders. Ann Neurol; 59: 527-534.

Valente EM, Marsh SE, Castori M, Dixon-Salazar T, Bertini E, Al-Gazali L, Messer J, Barbot C, Woods CG, Boltshauser E, Al-Tawari AA, Salpietro CD, Kayserili H, Sztriha L, Gribaa M, Koenig M, Dallapiccola B, Gleeson JG. (2005) Distinguishing the four genetic causes of Jouberts syndrome-related disorders. Ann Neurol; 57: 513-519.

Vaughn MW, Martienssen R. (2005) It's a small RNA world, after all. Science; 309: 1525–1526.

Verardo MR, Viczian A, Piri N, Akhmedov NB, Knox BE, Farber DB.(2009) Regulatory sequences in the 3' untranslated region of the human cGMP-phosphodiesterase beta-subunit gene. Invest Ophthalmol Vis Sci; 50:2591-2598.

Verderio EA, Johnson TS, Griffin M. (2005) Transglutaminases in wound healing and inflammation. Prog Exp Tumor Res; 38:89-114. Review.

Vervoort R, Gitzelmann R, Lissens W, Liebaers I. (1998) A mutation (IVS8+0.6kbdelTC) creating a new donor splice site activates a cryptic exon in an Alu-element in intron 8 of the human beta-glucuronidase gene. Hum Genet; 103: 686-693.

Vervoort R, Lennon A, Bird AC, Tulloch B, Axton R, Miano MG, Meindl A, Meitinger T, Ciccodicola A, Wright AF.(2000) Mutational hot spot within a new RPGR exon in X-linked retinitis pigmentosa. Nature Genet; 25: 462-466.

Vervoort R, Wright AF. (2002) Mutations of RPGR in X-linked retinitis pigmentosa (RP3). Hum. Mutat; 19: 486-500.

Vidal CN, Nicolson R, DeVito TJ, Hayashi KM, Geaga JA, Drost DJ, Williamson PC, Rajakumar N, Sui Y, Dutton RA, Toga AW, Thompson PM. (2006) Mapping corpus callosum deficits in autism: an index of aberrant cortical connectivity. Biol Psychiatry; 60: 218-225.

Volff JN.(2006)Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. Bioessays; 28: 913-922.

Wallace MR, Andersen LB, Saulino AM, Gregory PE, Glover TW, Collins FS. (1991) A de novo Alu insertion results in neurofibromatosis type 1. Nature; 353: 864-866.

Walter J, Hutter B, Khare T, Paulsen M. (2006) Repetitive elements in imprinted genes. Cytogenet Genome Res; 113: 109-115.

Walter J, Paulsen M. (2003) The potential role of gene duplications in the evolution of imprinting mechanisms. Hum Mol Genet; 12 Spec No 2: R215-20.

Walterfang M, Malhi GS, Wood AG, Reutens DC, Chen J, Barton S, Yücel M, Velakoulis D, Pantelis C. (2009) Corpus callosum size and shape in established bipolar affective disorder. Aust N Z J Psychiatry; 43: 838-845.

Wang DS, Dickson DW, Malter JS. (2008) Tissue Transglutaminase, Protein Cross-linking and Alzheimer's Disease: Review and Views. Int J Clin Exp Pathol; 1: 5-18.

Wang H, Xing J, Grover D, Hedges DJ, Kyudong H, Walker JA, Batzer MA. (2005) SVA elements: a hominid-specific retroposon family. J Mol Biol; 354: 994-1007.

Waterston RH, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature; 420: 520–562.

Wen R, Song Y, Kjellstrom S, Tanikawa A, Liu Y, Li Y, Zhao L, Bush RA, Laties AM, Sieving PA. (2006) Regulation of rod phototransduction machinery by ciliary neurotrophic factor. J Neurosci; 26:13523-135230.

Wilhelmus MM, Grunberg SC, Bol JG, van Dam AM, Hoozemans JJ, Rozemuller AJ, Drukarch B. (2009) Transglutaminases and transglutaminase-catalyzed cross-links colocalize with the pathological lesions in Alzheimer's disease brain. Brain Pathol; 19: 612-622.

Will CL, Lührmann R. (2005) Splicing of a rare class of introns by the U12-dependent spliceosome. Biol Chem; 386: 713-724.

Williams AC, Brophy PJ. (2002) The function of the Periaxin gene during nerve repair in a model of CMT4F. J Anat; 200: 323-330.

Wilson RS, Arnold SE, Schneider JA, Boyle PA, Buchman AS, Bennett DA. (2009) Olfactory impairment in presymptomatic Alzheimer's disease. Ann N Y Acad Sci; 1170: 730-735.

Wood AJ, Roberts RG, Monk D, Moore GE, Schulz R, Oakey RJ. (2007) A screen for retrotransposed imprinted genes reveals an association between X chromosome homology and maternal germ-line methylation. PLoS Genet; 3:e20.

Wszolek ZK, Markopoulou K. Olfactory dysfunction in Parkinson's disease. Clin Neurosci. 1998;5(2):94-101.

Wu LL, Zhou XF. (2009) Huntingtin associated protein 1 and its functions. Cell Adh Migr; 3:71-76.

Xiong Y, Eickbush TH. (1990) Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J; 9: 3353-3362.

Xu L, Hynes RO. (2007) GPR56 and TG2: possible roles in suppression of tumor growth by the microenvironment. Cell Cycle; 6: 160-165.

Yang LB, Li R, Meri S, Rogers J, Shen Y. (2000) Deficiency of complement defense protein CD59 may contribute to neurodegeneration in Alzheimer's disease. J Neurosci; 20:7505-7509.

Ying SY, Chang DC, Lin SL. (2008). The microRNA (miRNA): overview of the RNA genes that modulate gene function. Mol Biotechnol; 38: 257-268. Review.

Zechner U, Kohlschmidt N, Rittner G, Damatova N, Beyer V, Haaf T, Bartsch O. (2009) Epimutation at human chromosome 14q32.2 in a boy with a upd(14)mat-like clinical phenotype. Clin Genet; 75: 251-258.

Zeng Y, Cullen BR. (2005) Efficient processing of primary microRNA hairpins by Drosha requires flanking nonstructured RNA sequences. J Biol Chem; 280: 27595-27603.

Zhang MQ. (1998) Identification of human gene core promoters in silico. Genome Res; 8: 319-326.

Zhang MQ. (1998) Statistical features of human exons and their flanking regions. Hum Mol Genet; 7: 919-932.

Zhang SR, Li SH, Abler A, Fu J, Tso MO, Lam TT. (1996) Tissue transglutaminase in apoptosis of photoreceptor cells in rat retina. Invest Ophthalmol Vis Sci; 37:1793-1799.

Zhao F, Qi J, Schuster SC. (2009) Tracking the past: interspersed repeats in an extinct Afrotherian mammal, Mammuthus primigenius. Genome Res;19: 1384-1392.

Zhong J, Zhang T, Bloch LM. (2006) Dendritic mRNAs encode diversified functionalities in hippocampal pyramidal neurons. BMC Neurosci; 7: 17.

Zhou D, Li S, Wen J, Gong X, Xu L, Luo Y. (2008) Genome-wide computational analyses of microRNAs and their targets from Canis familiaris. Comput Biol Chem; 32:60-65.

Zhou X, Duan X, Qian J, Li F. (2009) Abundant conserved microRNA target sites in the 5'-untranslated region and coding sequence. Genetica. 2009 Jul 4. [Epub ahead of print]

Zhu ZB, Hsieh SL, Bentley DR, Campbell RD, Volanakis JE. (1992) A variable number of tandem repeats locus within the human complement C2 gene is associated with a retroposon derived from a human endogenous retrovirus. J Exp Med; 175: 1783-1787.

Zilberman D, Henikoff S. (2005) Epigenetic inheritance in Arabidopsis: selective silence. Curr. Opin. Genet. Dev; 15: 557–562.

Zuckerkandl E. (1976) Gene control in eukaryotes and the c-value paradox "excess" DNA as an impediment to transcription of coding sequences. J Mol Evol; 9: 73-104. Review.

# 9. APPENDICES

## 9.1. MIR-containing genes which are known and validated

All genes are human with the exception of Arhgef17, Arrb1, Cacna1d and Plxnb2 which are noted in rodent genes only. Abbreviations: 5’, 5’-untranslated region; 3’, 3’-untranslated region; CDS, coding sequence; ATG, initiating methionine codon; TAG, stop codon.

| Gene Name | Accession Number | Genomic Location | MIR Location | MIR Type | Loc. |
|---|---|---|---|---|---|
| A4GALT | NM_017436 | - | - | - | - |
| | BC055286 | 7 121 | 112 228 | MIR | 5’ |
| | | 186 317 | 154 17 | MIR3 | 5’ |
| AAA1 | NM_207285 | 43 195 | 261 112 | MIR | ATG |
| | NM_207286 | 43 195 | 261 112 | MIR | ATG |
| | NM_207287 | 43 195 | 261 112 | MIR | - |
| | NM_207288 | - | - | - | - |
| | NM_207289 | 43 195 | 261 112 | MIR | ATG |
| | NM_207290 | - | - | - | - |
| | NM_207283 | - | - | - | - |
| AADACL1 | NM_020792 | 3310 3415 | 74 180 | MIR3 | 3’ |
| ABCC13 | NM_138726 | 1444 1640 | 201 6 | MIR | 3’ |
| | NM_172024 | - | - | - | - |
| | NM_172026 | - | - | - | - |
| | NM_172025 | - | - | - | - |
| ABCC3 | NM_003786 | 4758 4926 | 198 14 | MIR3 | 3’ |
| ABCG2 | NM_004827 | 2755 2852 | 95 207 | MIR3 | 3’ |
| ABHD2 | NM_152924 | 2447 2604 | 45 197 | MIR3 | 3’ |
| | NM_007011 | 171 259 | 143 230 | MIR | 5’ |
| | | 2845 3002 | 45 197 | MIR3 | 3’ |
| ABHD4 | NM_022060 | 1610 1783 | 188 21 | MIR3 | 3’ |
| ACADSB | NM_001609 | 5246 5384 | 202 42 | MIR3 | 3’ |
| ACAN | NM_001135 | - | - | - | - |
| | NM_013227 | - | - | - | - |
| | BC036445 | 2624 2790 | 47 265 | MIRb | 3’ |
| | | 3984 4167 | 13 206 | MIR | 3’ |
| ACAT1 | NM_000019 | - | - | - | - |
| | BC063853 | 1672 1799 | 1 126 | MIRb | 3’ |
| ACBD5 | NM_145698 | 4192 4298 | 52 176 | MIRm | 3’ |
| | | 6459 6581 | 207 91 | THER1_MD | 3’ |
| | | 8746 8847 | 112 216 | MIR | 3’ |
| | | 12866 12956 | 223 127 | MIR | 3’ |
| ACCN4 | NM_018674 | - | - | - | - |
| | NM_182847 | - | - | - | - |
| | BC031812 | 2425 2544 | 4 124 | MIR3 | 3’ |
| ACE2 | NM_021804 | 3334 3381 | 206 254 | MIR | 3’ |
| ACSS1 | NM_032501 | 4 54 | 129 79 | MIR | 5’ |
| ACVR1C | NM_145259 | 2364 2465 | 148 40 | MIR | 3’ |
| ADAM19 | NM_023038 | 2989 3127 | 49 193 | MIR | 3’ |
| | NM_033274 | - | - | - | - |
| ADAMTS4 | NM_005099 | 4224 4324 | 109 222 | MIRb | 3’ |
| ADAMTS7 | NM_014272 | - | - | - | - |
| | AF140675 | 2502 2562 | 93 150 | MIR_Mars | CDS |
| | | 2843 3013 | 30 207 | MIRb | TAG |
| ADAMTS9 | NM_182920 | - | - | - | - |
| | AB037733.1 | 4447 4637 | 2 200 | MIR3 | 3’ |

| | | | | | |
|---|---|---|---|---|---|
| ADAMTSL1 | NM_139238 | 2116 2271 | 35 201 | MIR3 | TAG |
| | NM_052866 | - | - | - | - |
| | NM_139264 | - | - | - | - |
| ADAMTSL5 | NM_213604 | 129 314 | 186 5 | MIRb | 5’ |
| ADC | NM_052998 | 360 481 | 172 29 | MIR3 | 5’ |
| ADCK4 | NM_024876 | 13 112 | 36 138 | MIRb | 5’ |
| ADCY1 | NM_021116 | 10178 10428 | 258 1 | MIRb | 3’ |
| ADCY2 | NM_020546 | 4766 4881 | 22 188 | MIR | 3’ |
| | | 5765 5876 | 153 46 | MIR_Mars | 3’ |
| ADCY6 | NM_015270 | 469 594 | 91 214 | MIR | 5’ |
| | NM_020983 | - | - | - | - |
| ADD2 | NM_001617 | 2766 2930 | 179 3 | MIR3 | 3’ |
| | NM_017482 | 2193 2251 | 140 206 | MIRb | 3’ |
| | NM_017483 | 1848 2012 | 179 3 | MIR3 | 3’ |
| | NM_017484 | 1275 1333 | 140 206 | MIRb | 3’ |
| | NM_017488 | 2852 3016 | 179 3 | MIR3 | 3’ |
| ADIPOQ | NM_004797 | 1075 1258 | 202 21 | MIRb | 3’ |
| ADRA1A | NM_000680 | - | - | - | - |
| | NM_033302 | - | - | - | - |
| | NM_033303 | - | - | - | - |
| | NM_033304 | - | - | - | - |
| | AY491781 | 1302 1404 | 159 55 | MIR | CDS |
| AGGF1 | NM_018046 | 4289 4417 | 119 258 | MIR_Mars | 3’ |
| AGPAT6 | NM_178819 | - | - | - | - |
| | AY358670 | 2033 2215 | 2 195 | MIR | 3’ |
| AGXT2L2 | NM_153373 | 1592 1737 | 227 78 | MIR | 3’ |
| AHCYL1 | NM_006621 | - | - | - | - |
| | AF090905 | 910 1095 | 24 225 | MIRb | 3’ |
| AHI1 | NM_01765 | - | - | - | - |
| | NM_001134832 | 3436 3548 | 8 118 | MIRc | TAG |
| AIF1L | NM_031426 | 1686 1805 | 205 87 | MIR3 | 3’ |
| | | 2175 2215 | 276 234 | MIRm | 3’ |
| | NM_001002260 | 1708 1827 | 205 87 | MIR3 | 3’ |
| | | 2197 2237 | 276 234 | MIRm | 3’ |
| AIPL1 | NM_001033054 | - | - | - | - |
| | NM_001033055 | - | - | - | - |
| | NM_014336 | - | - | - | - |
| | BC007994 | 1403 1466 | 157 103 | MIR | 3’ |
| AK2 | NM_001625 | - | - | - | - |
| | NM_013411 | 1870 2087 | 3 263 | THER1_MD | 3’ |
| | | 2727 2778 | 115 161 | MIRm | 3’ |
| | | 2783 3004 | 258 26 | MIRb | 3’ |
| | | 3250 3324 | 19 112 | MIR | 3’ |
| AKAP13 | NM_006738 | - | - | - | - |
| | NM_007200 | - | - | - | - |
| | NM_144767 | - | - | - | - |
| | BC050312 | 2869 2919 | 103 153 | MIR | 3’ |
| AKAP14 | NM_178813 | - | - | - | - |
| | NM_001008534 | - | - | - | - |
| | NM_001008535 | - | - | - | - |
| | BC066357 | 1 76 | 174 249 | MIR | 5’ |
| AKAP5 | NM_004857 | 109 282 | 89 272 | MIRb | 5’ |
| | | 496 657 | 35 195 | MIR3 | 5’ |
| ALAD | NM_001003945 | 2820 2907 | 143 53 | MIR | 5’ |
| | | 2913 2987 | 78 4 | MIR | 5’ |
| | NM_000031 | 2654 2741 | 143 53 | MIR | 5’ |
| | | 2747 2821 | 78 4 | MIR | 5’ |

| | | | | | |
|---|---|---|---|---|---|
| ALDH3B1 | NM_000694 | 1800 1936 | 30 164 | MIR | 3’ |
| | | 2239 2268 | 165 191 | MIR | 3’ |
| | NM_001030010 | 1689 1825 | 30 164 | MIR | 3’ |
| | | 2128 2157 | 165 191 | MIR | 3’ |
| ALDH3B2 | NM_000695 | 2148 2236 | 48 138 | MIRb | 3’ |
| | NM_001031615 | 1992 2080 | 48 138 | MIRb | 3’ |
| ALG2 | NM_033087 | 2416 2588 | 74 272 | THER1_MD | 3’ |
| AMACR | NM_014324 | 2274 2363 | 203 115 | MIRb | 3’ |
| | NM_203382 | 2113 2202 | 203 115 | MIRb | 3’ |
| AMIGO2 | NM_181847 | 3212 3301 | 90 179 | MIR3 | 3’ |
| | | 3415 3525 | 196 82 | MIR3 | 3’ |
| AMOT | NM_133265 | 4331 4445 | 206 81 | MIR3 | 3’ |
| AMOTL1 | NM_130847 | 4535 4688 | 214 35 | MIRb | 3’ |
| AMPD3 | NM_000480 | 3262 3349 | 99 201 | MIR3 | 3’ |
| | | 3487 3534 | 152 202 | MIR3 | 3’ |
| | | 3540 3666 | 157 14 | MIRb | 3’ |
| | NM_001025389 | 3225 3312 | 99 201 | MIR3 | 3’ |
| | | 3450 3497 | 152 202 | MIR3 | 3’ |
| | | 3503 3629 | 157 14 | MIRb | 3’ |
| | NM_001025390 | 3364 3451 | 99 201 | MIR3 | 3’ |
| | | 3589 3636 | 152 202 | MIR3 | 3’ |
| | | 3642 3768 | 157 14 | MIRb | 3’ |
| ANKRD34C | XM_930512 | - | - | - | - |
| | AK127037 | 3024 3274 | 6 260 | MIRb | 5’ |
| ANKRD42 | NM_182603 | 1460 1628 | 5 193 | MIR | 3’ |
| | | 1659 1669 | 194 205 | MIR | 3’ |
| ANKRD43 | NM_175873 | 2833 2957 | 81 207 | MIRm | 3’ |
| ANKRD49 | NM_017704 | - | - | - | - |
| | AL833977 | 467 551 | 164 258 | MIR_Mars | TAG |
| | | 655 862 | 260 17 | MIRb | 3’ |
| ANKRD5 | NM_022096 | 281 345 | 176 112 | MIR | 5’ |
| | NM_198798 | - | - | - | - |
| ANKS6 | NM_173551 | 3419 3512 | 93 188 | MIR3 | 3’ |
| | | 4362 4586 | 251 5 | MIR | 3’ |
| | | 6376 6510 | 74 235 | MIRb | 3’ |
| | BC012981 | 845 954 | 5 115 | MIRb | 5’ |
| | | 979 1188 | 262 44 | MIR | 5’ |
| ANRIL | NR_003529 | 3480 3581 | 137 38 | MIR | ncRNA |
| | | 3820 4033 | 241 5 | MIR | |
| AP2B1 | NM_001030006 | 3953 4093 | 187 27 | MIR3 | 3’ |
| | NM_001282 | 3911 4051 | 187 27 | MIR3 | 3’ |
| APBA2BP | NM_031231 | 1433 1518 | 40 124 | MIR | 3’ |
| | NM_031232 | 1535 1620 | 40 124 | MIR | 3’ |
| APLN | NM_017413 | 1342 1505 | 201 19 | MIR3 | 3’ |
| APOF | NM_001638 | 1162 1359 | 34 228 | MIRb | 3’ |
| APOLD1 | NM_030817 | 1513 1681 | 23 231 | MIRb | 3’ |
| | | 1868 1879 | 232 245 | MIRb | 3’ |
| APTX | NM_175073 | - | - | - | - |
| | NM_175069 | - | - | - | - |
| | NM_175072 | - | - | - | - |
| | NM_017692 | 1 110 | 184 68 | MIR | 5’ |
| | NM_175071 | - | - | - | |
| AQP1 | NM_198098 | 2424 2484 | 151 97 | MIRm | 3’ |
| AQP2 | NM_000486 | 2201 2294 | 124 15 | MIR3 | 3’ |
| | | 3340 3410 | 254 183 | MIRb | 3’ |
| | | 3434 3680 | 261 9 | MIRb | 3’ |
| AQP9 | NM_020980 | 2136 2285 | 240 90 | MIRb | 3’ |

| | | | | | |
|---|---|---|---|---|---|
| ARHGAP24 | NM_001025616 | - | - | - | - |
| | NM_031305 | 4465 4574 | 41 168 | MIR | 3’ |
| ARHGAP27 | NM_199282 | 2793 2863 | 150 77 | MIR | 3’ |
| | | 3316 3394 | 271 181 | THER1_MD | 3’ |
| Arhgef17 (mus) | NM_001081116 | 8002 8091 | 82 184 | MIR3 | 3’ |
| | | 8447 8506 | 6 65 | MIR | 3’ |
| | | 9072 9168 | 15 120 | MIRb | 3’ |
| ARHGEF6 | NM_004840 | - | - | - | - |
| | BC043505 | 1 86 | 59 158 | MIR3 | 5’ |
| ARL1 | NM_001177 | 1267 500 | 246 5 | MIR | 3’ |
| | | 2907 2980 | 123 208 | MIR3 | 3’ |
| ARL10 | NM_173664 | 1277 1366 | 55 151 | MIR | 3’ |
| | | 2087 2206 | 73 201 | MIR3 | 3’ |
| | BC059361 | 872 904 | 251 217 | MIRb | CDS |
| | | 1208 1344 | 216 73 | MIRb | 3’ |
| ARL11 | NM_138450 | 2855 3091 | 255 11 | MIRb | 3’ |
| ARMCX3 | NM_016607 | 3145 3261 | 122 248 | MIR | 3’ |
| | NM_177947 | 3117 3233 | 122 248 | MIR | 3’ |
| | NM_177948 | 3068 3184 | 122 248 | MIR | 3’ |
| ARNT2 | NM_014862 | 3490 3577 | 3 84 | MIRm | 3’ |
| | | 4707 4779 | 248 172 | MIRm | 3’ |
| ARNTL2 | NM_020183 | - | - | - | - |
| | AF256215 | 5144 5312 | 202 6 | MIR3 | 3’ |
| Arrb1 (mus) | NM_177231 | 2740 2841 | 74 177 | MIR3 | 3’ |
| | NM_178220 | 2716 2817 | 74 177 | MIR3 | 3’ |
| ARSG | NM_014960 | 1820 2011 | 7 268 | MIRb | 3’ |
| ART1 | NM_004314 | 48 114 | 141 74 | MIR | ATG |
| ARTN | NM_003976 | 3 100 | 110 222 | MIRb | 5’ |
| | NM_057091 | - | - | - | - |
| | NM_057160 | - | - | - | - |
| | NM_057090 | - | - | - | - |
| ASL | NM_001024943 | - | - | - | - |
| | NM_000048 | - | - | - | - |
| | NM_001024944 | - | - | - | - |
| | NM_001024946 | - | - | - | - |
| | AY203938 | 635 765 | 262 136 | MIR | 5’ |
| ASTN1 | NM_004319 | 5568 5812 | 252 4 | MIR | 3’ |
| | NM_207108 | - | - | - | - |
| ASTN2 | NM_014010 | - | - | - | - |
| | NM_198186 | - | - | - | - |
| | NM_198187 | - | - | - | - |
| | NM_198188 | 74 185 | 175 61 | MIR3 | 5’ |
| ATM | NM_000051 | 12439 12598 | 257 79 | MIRb | 3’ |
| | NM_138292 | 8269 8428 | 257 79 | MIRb | 3’ |
| | BC061584 | 6 61 | 8 63 | MIR | 5’ |
| ATP2B1 | NM_001001323 | 5933 6089 | 2 187 | MIR3 | 3’ |
| | NM_001682 | 5779 5935 | 2 187 | MIR3 | 3’ |
| ATP6V0E2 | NM_001100592 | 169 256 | 224 130 | MIRb | 5’ |
| | NM_145230 | 169 256 | 224 130 | MIRb | 5’ |
| ATP6V1B1 | NM_001692 | - | - | - | - |
| | BC035978 | 965 1003 | 147 109 | MIR | 3’ |
| ATP6V1E2 | NM_080653 | 194 421 | 267 30 | MIRb | 5’ |
| ATP6V1G2 | NM_130463 | 918 1162 | 21 267 | MIRb | 3’ |
| | NM_138282 | 764 1008 | 21 267 | MIRb | 3’ |
| ATP7B | NM_000053 | 5867 6048 | 272 83 | MIRb | 3’ |
| | NM_001005918 | 5246 5427 | 272 83 | MIRb | 3’ |
| ATPAF1 | NM_001042546 | 3116 3301 | 10 193 | MIRb | 3’ |

| | | | | | |
|---|---|---|---|---|---|
| | NM_022745 | 3320 3505 | 10 193 | MIRb | 3’ |
| | AF111705 | 700 885 | 10 193 | MIRb | 5’ |
| ATPIF1 | NM_016311 | - | - | - | - |
| | NM_178190 | - | - | - | - |
| | NM_178191 | 1068 1125 | 169 113 | MIR3 | 3’ |
| ATXN7L1 | NM_152749 | 1398 1470 | 51 123 | MIRb | 3’ |
| | BC003517 | 1748 1896 | 255 99 | MIR | 3’ |
| AVPR1A | NM_000706 | 128 227 | 237 142 | MIRc | 5’ |
| AVPR2 | NM_000054 | 1482 1545 | 82 148 | MIR | 3’ |
| AXL | NM_021913 | 3309 3409 | 99 202 | MIR3 | 3’ |
| | NM_001699 | 3282 3382 | 99 202 | MIR3 | 3’ |
| B3GNT1 | NM_006057 | 1575 1806 | 220 2 | MIRb | 3’ |
| | NM_033170 | 1490 1673 | 176 2 | MIRb | 3’ |
| | NM_033171 | 1684 1915 | 220 2 | MIRb | 3’ |
| | NM_033172 | 1524 1755 | 220 2 | MIRb | 3’ |
| | NM_033173 | 2270 2453 | 176 2 | MIRb | 3’ |
| B3GNT6 | NM_006876 | 1421 1664 | 2 263 | MIRb | 3’ |
| B4GALT2 | NM_003780 | 1751 1839 | 43 135 | MIR | 3’ |
| | NM_001005417 | 1537 1625 | 43 135 | MIR | 3’ |
| B9D2 | NM_030578 | 759 928 | 3 181 | MIR | 3’ |
| BATF2 | NM_138456 | 1100 1250 | 238 88 | MIRb | 3’ |
| BBS7 | NM_176824 | 2663 2872 | 252 2 | MIRb | 3’ |
| | NM_018190 | - | - | - | - |
| BCAS2 | NM_005872 | 1200 1265 | 238 191 | THER1_MD | 3’ |
| BDKRB2 | NM_000623 | 2100 2261 | 43 259 | MIR_Mars | 3’ |
| | | 2358 2394 | 205 247 | MIRb | 3’ |
| BEND6 | NM_152731 | - | - | - | - |
| | BC022988 | 91 229 | 23 160 | MIR | 5’ |
| BEST2 | NM_017682 | 1716 1839 | 124 262 | MIRb | 3’ |
| BEST3 | NM_152439 | 1885 1947 | 208 144 | MIR3 | 3’ |
| | NM_032735 | - | - | - | - |
| BHLHB9 | NM_030639 | 2842 3083 | 247 4 | MIRb | 3’ |
| BIRC4 | NM_001167 | 5397 5487 | 128 225 | MIRb | 3’ |
| | | 6290 6439 | 206 49 | MIR3 | 3’ |
| BMP1 | NM_006132 | 1374 1455 | 22 97 | MIR | 3’ |
| | NM_001199 | - | - | - | - |
| | NM_006128 | - | - | - | - |
| | NM_006129 | - | - | - | - |
| | NM_006130 | - | - | - | - |
| BMPR2 | NM_001204 | 6577 6796 | 228 7 | MIRb | 3’ |
| BNIPL | NM_138278 | 2019 2112 | 44 147 | MIR | 3’ |
| | NM_138279 | - | - | - | - |
| BPA-1 | XM_001726020 | - | - | - | - |
| | AB088847 | 787 959 | 267 81 | MIRb | 3’ |
| BPESC1 | NM_021812 | 49 206 | 195 21 | THER1_MD | 5’ |
| | | 1323 1490 | 44 229 | MIRb | 3’ |
| | | 1854 1898 | 233 276 | MIRm | 3’ |
| | | 2845 3081 | 1 261 | MIR | 3’ |
| BRF1 | NM_001519 | - | - | - | - |
| | NM_145696 | 304 442 | 165 22 | MIR | 5’ |
| | | 532 673 | 180 19 | MIR_Mars | 5’ |
| | NM_145685 | 19 127 | 146 19 | MIR_Mars | 5’ |
| BRWD1 | NM_018963 | - | - | - | - |
| | NM_033656 | - | - | - | - |
| | NM_001007246 | 2143 2360 | 35 258 | MIR_Mars | 3’ |
| BSDC1 | NM_018045 | 1800 2010 | 51 258 | MIR | 3’ |
| | | 2108 2203 | 113 19 | MIR | 3’ |

| | | | | | |
|---|---|---|---|---|---|
| | | 2680 2758 | 120 41 | THER1_MD | 3’ |
| BSND | NM_057176 | 1257 1302 | 112 158 | MIRb | 3’ |
| BTC | NM_001729 | 1164 1271 | 268 149 | MIRb | 3’ |
| BTN3A1 | NM_007048 | 2085 2252 | 225 52 | MIR | 3’ |
| | | 2250 2481 | 272 21 | MIR | 3’ |
| | NM_194441 | - | - | - | - |
| BTN3A2 | NM_007047 | 2778 2949 | 225 52 | MIR | 3’ |
| | | 2950 3177 | 268 22 | MIR | 3’ |
| BTN3A3 | NM_006994 | 1913 2084 | 225 52 | MIR | TAG |
| | | 2088 2312 | 260 22 | MIR | 3 |
| | NM_197974 | 1823 1994 | 225 52 | MIR | TAG |
| | | 1998 2222 | 260 22 | MIR | 3’ |
| BTNL3 | NM_197975 | 1822 1980 | 84 263 | MIRb | 3’ |
| | NM_006707 | 1720 1878 | 84 263 | MIRb | 3’ |
| BUD31 | NM_003910 | 256 334 | 132 45 | MIR | 5’ |
| BZRAP1 | NM_004758 | 6112 6324 | 208 2 | MIR3 | 3’ |
| C10orf10 | NM_007021 | 1041 1169 | 42 170 | MIR | 3’ |
| C10orf105 | XM_001715769 | 4253 4335 | 229 61 | MIR | 3’ |
| C10orf111 | NM_153244 | 1478 1647 | 48 249 | MIR | 3’ |
| C10orf25 | NM_001039380 | 259 337 | 191 115 | MIRb | 3’ |
| | | 1131 1344 | 245 25 | MIR_Mars | 3’ |
| | | 1436 1527 | 173 273 | MIRb | 3’ |
| | | 2041 2098 | 262 205 | MIR | 3’ |
| | | 2527 2577 | 196 144 | MIR | 3’ |
| C10orf54 | NM_022153 | 4616 4771 | 229 61 | MIRc | 3’ |
| C11orf44 | NM_173580 | 933 1156 | 10 262 | MIRb | ncRNA |
| | | 2487 2667 | 37 216 | MIRb | ncRNA |
| C11orf51 | NM_014042 | 131 216 | 140 55 | MIRb | 5’ |
| C11orf57 | NM_018195 | 2717 2826 | 43 156 | MIR | 3’ |
| C11orf68 | NM_031450 | - | - | - | - |
| | BC010512 | 2 75 | 160 84 | MIR | 5’ |
| C11orf86 | NM_001136485 | 1000 1095 | 198 101 | MIRc | 3’ |
| C11orf92 | NM_207429 | 3032 3171 | 160 27 | MIRb | 3’ |
| | | 3464 3494 | 32 1 | MIRb | 3’ |
| C12orf47 | NM_016534 | 1116 1331 | 263 50 | MIRb | ncRNA |
| C12orf52 | NM_032848 | 281 400 | 90 220 | MIRb | 5’ |
| C12orf76 | NM_207435 | 446 556 | 29 145 | MIR | CDS |
| C13orf1 | NM_020456 | 1737 1933 | 4 232 | MIRb | 3’ |
| | | 2249 2273 | 233 260 | MIRb | 3’ |
| C14orf126 | NM_080664 | 691 784 | 38 136 | MIRb | 3’ |
| C14orf132 | NM_020215 | 1177 1348 | 175 1 | MIR3 | ncRNA |
| | | 2955 3173 | 236 1 | MIRb | ncRNA |
| | | 5012 5088 | 193 117 | MIR3 | ncRNA |
| C14orf139 | NM_024633 | 826 936 | 127 22 | MIR | ncRNA |
| C14orf178 | NM_174943 | 721 789 | 37 108 | MIR_Mars | 3’ |
| | | 793 890 | 163 272 | MIRb | 3’ |
| C14orf33 | XM_375081 | 1757 1949 | 234 3 | MIRb | 3’ |
| C14orf91 | AF113687 | 1241 1410 | 22 200 | MIRb | ncRNA |
| C15orf33 | NM_152647 | - | - | - | - |
| | BC048125 | 1208 1359 | 33 186 | MIR3 | TAG |
| C15orf39 | NM_015492 | 3567 3725 | 24 188 | MIRb | 3’ |
| C15orf5 | NM_030944 | 157 285 | 1 141 | MIR | ncRNA |
| C15orf52 | NM_207380 | 4926 5116 | 3 200 | MIRb | 3’ |
| C16orf47 | NM_207385 | 252 310 | 82 24 | MIR | CDS |
| C16orf88 | NM_001012991 | 1782 1950 | 252 74 | MIR | 3’ |
| C17orf102 | NM_207454 | 1614 1702 | 206 117 | MIR3 | 3’ |
| | | 1875 2119 | 260 16 | MIRb | 3’ |

| | | | | | |
|---|---|---|---|---|---|
| | | 3250 3454 | 65 268 | MIRb | 3’ |
| C17orf65 | NM_178542 | 1616 1770 | 208 50 | MIR3 | 3’ |
| C17orf66 | NM_152781<br>BC033734 | -<br>973 1206<br>1224 1321 | -<br>258 6<br>93 206 | -<br>MIRb<br>MIR3 | -<br>3’<br>3’ |
| C17orf67 | XM_378687 | 727 812 | 107 26 | MIRb | 5’ |
| C18orf12 | AB027121 | 38 205 | 3 171 | MIR | ncRNA |
| C18orf21 | NM_031446 | 929 961 | 156 188 | MIR | 3’ |
| C18orf49 | AK000229<br>BC047606 | 1292 1411<br>1455 1633 | 126 253<br>203 22 | MIRb<br>MIR | ncRNA<br>ncRNA |
| C19orf20 | NM_033513 | 1319 1359 | 109 149 | MIR | 3’ |
| C19orf54 | NM_198476 | 1051 1156 | 174 68 | MIR | 3’ |
| C1orf116 | NM_023938 | 5278 5490 | 40 262 | MIRb | 3’ |
| C1orf167 | XM_209234 | 2864 2957<br>3643 3720 | 136 48<br>58 145 | MIR<br>MIR | CDS<br>3’ |
| C1orf187 | NM_198545 | 1577 1738 | 250 91 | MIRb | 3’ |
| C1orf2 | NM_006589<br>NM_198264<br>BC008854 | 2991 3094<br>2703 2806<br>1192 1296 | 7 127<br>7 127<br>7 127 | MIR<br>MIR<br>MIR | 3’<br>3’<br>TAG |
| C1orf213 | NM_001008896<br>NM_138479 | -<br>1427 1614 | -<br>4 226 | -<br>MIRb | -<br>3’ |
| C1orf220 | NM_207467 | 342 446 | 196 91 | MIRb | 5’ |
| C1QTNF3 | NM_030945<br>NM_181435 | -<br>2024 2079 | -<br>95 38 | -<br>MIR | -<br>3’ |
| C1RL | NM_016546 | 2009 2230 | 222 3 | MIRb | 3’ |
| C20orf160 | NM_080625 | 2358 2513 | 94 259 | MIR | 3’ |
| C20orf54 | NM_033409 | 87 169<br>1967 2094 | 88 168<br>74 197 | MIR3<br>MIRb | 5’<br>3’ |
| C21orf24 | NM_001001789 | 2905 2998 | 123 29 | MIR | 3’ |
| C21orf87 | XR_040389 | 22 196 | 194 27 | MIR | ncRNA |
| C21orf88 | NM_153754 | 34 156 | 182 56 | THER1_MD | ncRNA |
| C21orf96 | XR_040896 | 291 369 | 17 36 | MIR | ncRNA |
| C22orf24 | NM_015372 | 1061 1231 | 19 197 | MIRb | 3’ |
| C22orf26 | NM_018280 | 973 1084 | 100 212 | MIRb | 3’ |
| C2orf18 | NM_017877<br>BC016389 | -<br>326 364<br>549 730 | -<br>150 112<br>258 31 | -<br>MIR<br>MIRb | -<br>3’<br>3’ |
| C2orf73 | NM_001100396 | 1218 1368 | 36 188 | MIRb | 3’ |
| C3orf19 | NM_016474 | 2632 2769 | 154 10 | MIRb | 3’ |
| C3orf27 | NM_007354 | 1622 1726<br>2073 2307 | 178 74<br>260 6 | MIRb<br>MIRb | 3’<br>3’ |
| C3orf35 | NM_178339<br>NM_178341<br>NM_178342 | 1205 1317<br>833 925<br>952 1069 | 141 21<br>144 43<br>160 40 | MIR<br>MIR<br>MIRb | CDS<br>CDS<br>3’ |
| C3orf51 | U88965 | 1476 1625<br>1930 1960 | 255 103<br>102 72 | MIRb<br>MIRb | ncRNA<br>ncRNA |
| C3orf57 | NM_145035<br>BC071875 | 1445 1536<br>61 152 | 135 42<br>135 42 | MIR<br>MIR | 3’<br>5’ |
| C4orf6 | NM_005750 | 702 870 | 221 43 | MIRb | 3’ |
| C5orf20 | NM_130848 | 1067 1293 | 31 260 | MIR | 3’ |
| C5orf3 | NM_018691 | 378 563 | 232 2 | MIR | 5’ |
| C6orf105 | NM_032744 | 1074 1246 | 197 30 | MIR3 | 3’ |
| C6orf114 | ***NM_033069*** | 883 1034<br>1186 1277 | 260 109<br>100 8 | MIRb<br>MIRb | TAG<br>3’ |
| C6orf12 | XM_379403 | 1131 1182<br>1846 2052<br>3544 3684 | 270 213<br>254 27<br>43 187 | MIRm<br>MIRb<br>MIR | ncRNA<br>ncRNA<br>ncRNA |

| | | | | | |
|---|---|---|---|---|---|
| | XM_945157 | 9056 9185 | 22 171 | MIR | ncRNA |
| C6orf155 | NM_024882 | 1356 1448 | 236 147 | MIR | ncRNA |
| | | 1950 2074 | 70 193 | MIR3 | ncRNA |
| C6orf195 | NM_152554 | 41 184 | 190 40 | MIRb | 5’ |
| C6orf227 | NM_207497 | 956 1128 | 201 31 | MIR | 3’ |
| | | 1875 2067 | 18 251 | MIR3 | 3’ |
| C6orf99 | XM_374188 | 148 215 | 123 199 | MIR3 | CDS |
| C7 | NM_000587 | 2866 2995 | 66 195 | MIR3 | 3’ |
| C7orf13 | NM_032625 | 1233 1350 | 35 158 | MIRb | ncRNA |
| C7orf65 | NM_001123065 | 105 157 | 140 85 | MIRb | CDS |
| C8orf14 | NM_054029 | 476 609 | 50 184 | MIR | 5’ |
| C8orf17 | NM_020237 | 90 157 | 190 123 | MIRb | 5’ |
| | | 328 383 | 173 112 | MIR3 | 5’ |
| | | 2628 2680 | 56 109 | MIR | 3’ |
| C8orf43 | NM_052958 | 1348 1539 | 57 262 | MIR | ncRNA |
| C8orf44 | NM_019607 | 897 1119 | 23 248 | MIRb | 3’ |
| C8orf46 | NM_152765 | 1396 1538 | 258 112 | MIR_Mars | 3’ |
| | | 1934 2110 | 258 54 | MIR | 3’ |
| C8orf55 | NM_016647 | 1863 1968 | 158 22 | MIR | 3’ |
| C9orf100 | NM_001031728 | - | - | - | - |
| | NM_032818 | 2669 2755 | 23 115 | MIR | 3’ |
| C9orf109 | XM_379665 | 5005 5221 | 245 1 | MIR | 3’ |
| | | 5303 5432 | 179 59 | MIRm | 3’ |
| C9orf110 | XM_379664 | 5005 5221 | 245 1 | MIR | 3’ |
| | | 5363 5429 | 135 74 | MIRb | 3’ |
| C9orf130 | NR_023389 | - | - | - | - |
| | AF170307 | 502 543 | 139 97 | MIRb | ncRNA |
| C9orf156 | NM_016481 | - | - | - | - |
| | AY189280 | 1245 1403 | 252 84 | MIRb | 5’ |
| C9orf25 | NM_147202 | 1927 2100 | 3 177 | MIR3 | 3’ |
| | | 3100 3169 | 60 130 | MIRb | 3’ |
| C9orf27 | NM_021208 | 616 762 | 39 193 | MIRb | 3’ |
| C9orf31 | XR_040686 | 154 266 | 144 29 | MIRc | ncRNA |
| C9orf5 | NM_032012 | 3658 3856 | 260 41 | MIR_Mars | 3’ |
| | | 6554 6729 | 202 23 | MIR3 | 3’ |
| | | 7351 7509 | 180 25 | MIRb | 3’ |
| C9orf72 | NM_018325 | - | - | - | - |
| | NM_145005 | 1157 1382 | 2 247 | MIR | 3’ |
| | | 1423 1491 | 259 188 | MIR | 3’ |
| | | 1730 1805 | 87 162 | MIRb | 3’ |
| C9orf84 | NM_173521 | 1614 1727 | 59 195 | MIR3 | 3’ |
| C9orf91 | NM_153045 | 1687 1758 | 181 109 | MIR3 | 3’ |
| | | 2687 2796 | 46 154 | MIR3 | 3’ |
| | | 3097 3145 | 155 202 | MIR3 | 3’ |
| | | 3250 3389 | 22 159 | MIRb | 3’ |
| C9orf97 | NM_139246 | - | - | - | - |
| | AF258575 | 1109 1181 | 16 90 | MIRb | 3’ |
| C9orf98 | NM_152572 | 135 184 | 161 109 | MIR | 5’ |
| CA8 | NM_004056 | 2096 2165 | 191 268 | MIRb | 3’ |
| CACHD1 | NM_020925 | 396 479 | 266 172 | MIRb | 5’ |
| | | 491 626 | 250 107 | MIRb | 5’ |
| Cacna1d (rat) | NM_017298 | - | - | - | - |
| | M57682 | 6343 6522 | 2 193 | MIRc | |
| CACNA1G | NM_018896 | - | - | - | - |
| | NM_198376 | - | - | - | - |
| | NM_198377 | - | - | - | - |
| | NM_198378 | - | - | - | - |

| | | | | | |
|---|---|---|---|---|---|
| | NM_198379 | - | - | - | - |
| | NM_198380 | - | - | - | - |
| | NM_198382 | - | - | - | - |
| | NM_198383 | - | - | - | - |
| | NM_198384 | - | - | - | - |
| | NM_198385 | - | - | - | - |
| | NM_198386 | - | - | - | - |
| | NM_198387 | - | - | - | - |
| | NM_198388 | - | - | - | - |
| | NM_198396 | - | - | - | - |
| | NM_198397 | 5006 5081 | 145 66 | MIRb | TAG |
| | | 5061 5236 | 6 174 | MIRb | 3’ |
| CACNA2D4 | NM_172364 | 4984 5090 | 212 106 | MIRb | 3’ |
| | | 5093 5217 | 272 113 | MIR | 3’ |
| | | 5250 5343 | 176 273 | MIR | 3’ |
| | NM_001005737 | 4913 5019 | 212 106 | MIRb | 3’ |
| | | 5022 5146 | 272 113 | MIR | 3’ |
| | | 5179 5272 | 176 273 | MIR | 3’ |
| | NM_001005766 | 4939 5045 | 212 106 | MIRb | 3’ |
| | | 5048 5172 | 272 113 | MIR | 3’ |
| | | 5205 5298 | 176 273 | MIR | 3’ |
| CACNG3 | NM_006539 | - | - | - | - |
| | AF114832 | 1751 1899 | 60 210 | MIRb | 3’ |
| CALML4 | NM_001031733 | - | - | - | - |
| | NM_033429 | 224 364 | 149 5 | MIR | 5’ |
| | | 1206 1375 | 20 196 | MIRb | 3’ |
| | | 3020 3099 | 102 18 | MIR | 3’ |
| CAMK2A | | 3768 3833 | 85 151 | MIRm | 3’ |
| | NM_015981 | 4020 4183 | 8 180 | MIR3 | 3’ |
| | NM_171825 | 3735 3800 | 85 151 | MIRm | 3’ |
| | | 3987 4150 | 8 180 | MIR3 | 3’ |
| CAPN5 | NM_004055 | 2171 2274 | 24 129 | MIR | 3’ |
| | | 3692 3765 | 81 153 | MIR | 3’ |
| CASC2 | NM_178816 | - | - | - | - |
| | NM_201377 | 542 616 | 20 106 | MIRm | ncRNA |
| CASP10 | NM_001230 | 3266 3375 | 200 88 | MIR3 | 3’ |
| | NM_032974 | - | - | - | - |
| | NM_032976 | 3303 3412 | 200 88 | MIR3 | 3’ |
| | NM_032977 | 3395 3504 | 200 88 | MIR3 | 3’ |
| CASR | NM_000388 | 4669 4913 | 33 267 | MIRb | 3’ |
| CATSPER2 | NM_054020 | - | - | - | - |
| | NM_172095 | - | - | - | - |
| | NM_172096 | - | - | - | - |
| | NM_172097 | 1000 1149 | 247 97 | MIR | 3’ |
| CATSPER2P1 | NR_002318 | 971 1120 | 247 97 | MIR | ncRNA |
| CCBL2 | NM_001008661 | 1870 2009 | 95 243 | MIR | 3’ |
| | NM_001008662 | 1770 1909 | 95 243 | MIR | 3’ |
| CCDC113 | NM_014157 | 2701 2856 | 63 267 | MIRb | 3’ |
| CCDC26 | NM_145050 | 853 1095 | 37 267 | MIRb | 3’ |
| CCDC52 | NM_144718 | 4302 4451 | 182 23 | MIR | 3’ |
| | | 4512 4740 | 268 60 | MIRb | 3’ |
| | | 4785 5000 | 3 247 | MIR | 3’ |
| CCDC64B | NM_001103175 | - | - | - | - |
| | XM_085463 | 1086 1225 | 72 229 | MIRb | 5’ |
| CCDC68 | NM_025214 | 2552 2728 | 252 51 | MIR | 3’ |
| CCDC76 | NM_019083 | 2616 2842 | 38 260 | MIR | 3’ |
| CCDC82 | NM_024725 | 78 152 | 26 99 | MIR | 5’ |
| CCDC95 | NM_173618 | - | - | - | - |

| | | | | | |
|---|---|---|---|---|---|
| | AY189289 | 1025 1193 | 1 161 | MIR3 | 3’ |
| | | 2139 2226 | 36 124 | MIR | 3’ |
| CCNF | NM_001761 | 4128 4222 | 88 197 | MIRb | 3’ |
| CCNO | NM_021147 | 29 159 | 94 230 | MIRb | ATG |
| | NM_001024592 | 613 730 | 152 35 | MIRb | 3’ |
| | | 735 789 | 94 150 | MIRb | 3’ |
| CCR5 | NM_000579 | 2781 2881 | 153 259 | MIR | 3’ |
| CCRL2 | NM_003965 | 1439 1660 | 261 23 | MIRb | 3’ |
| CD28 | NM_006139 | 2645 2767 | 21 157 | MIRb | 3’ |
| CD300A | NM_007261 | 1692 1818 | 175 51 | MIR_Mars | 3’ |
| CD300LG | NM_145273 | - | - | - | - |
| | AF427619 | 691 761 | 200 138 | MIR3 | CDS |
| | | 895 987 | 141 49 | MIR | CDS |
| CD33 | NM_001772 | 1255 1433 | 29 210 | MIR | 3’ |
| CD4 | NM_000616 | 1901 2028 | 147 6 | MIRb | 3’ |
| | | 2126 2332 | 212 17 | MIRb | 3’ |
| CD48 | NM_001778 | - | - | - | - |
| | BC030224 | 915 1009 | 30 126 | MIRb | 3’ |
| | | 1483 1576 | 122 216 | MIRb | 3’ |
| CD59 | NM_000611 | 3585 3709 | 152 20 | MIRb | 3’ |
| | | 3967 4153 | 2 211 | MIRb | 3’ |
| | | 5456 5560 | 99 1 | MIR3 | 3’ |
| | | 5900 6098 | 28 261 | MIR | 3’ |
| | | 6252 6429 | 268 67 | MIRb | 3’ |
| | NM_203329 | 3634 3758 | 152 20 | MIRb | 3’ |
| | | 4016 4202 | 2 211 | MIRb | 3’ |
| | | 5505 5609 | 99 1 | MIR3 | 3’ |
| | | 5949 6147 | 28 261 | MIR | 3’ |
| | | 6301 6478 | 268 67 | MIRb | 3’ |
| | NM_203330 | 3746 3870 | 152 20 | MIRb | 3’ |
| | | 4128 4314 | 2 211 | MIRb | 3’ |
| | | 5617 5721 | 99 1 | MIR3 | 3’ |
| | | 6061 6259 | 28 261 | MIR | 3’ |
| | | 6413 6590 | 268 67 | MIRb | 3’ |
| | NM_203331 | 3630 3754 | 152 20 | MIRb | 3’ |
| | | 4012 4198 | 2 211 | MIRb | 3’ |
| | | 5501 5605 | 99 1 | MIR3 | 3’ |
| | | 5945 6143 | 28 261 | MIR | 3’ |
| | | 6297 6474 | 268 67 | MIRb | 3’ |
| CD80 | NM_005191 | 1457 1591 | 193 20 | MIR3 | 3’ |
| CD84 | NM_003874 | - | - | - | - |
| | U96627 | 1057 1193 | 47 195 | MIR3 | 3’ |
| CD8A | NM_001768 | 1108 1290 | 201 9 | MIR3 | 3’ |
| | NM_171827 | 997 1179 | 201 9 | MIR3 | 3’ |
| CDC14B | NM_003671 | - | - | - | - |
| | NM_033331 | - | - | - | - |
| | NM_033332 | 3922 4028 | 1 116 | MIR | 3’ |
| CDC20B | NM_152623 | 1625 1829 | 1 199 | MIR | TAG |
| | | 2174 2234 | 259 197 | THER1_MD | 3’ |
| | | 2559 2644 | 29 119 | MIR | 3’ |
| CDC25B | NM_021873 | 33 141 | 44 152 | MIRm | 5’ |
| | NM_004358 | 33 141 | 44 152 | MIRm | 5’ |
| | NM_021872 | 33 141 | 44 152 | MIRm | 5’ |
| CDH22 | NM_021248 | 3285 3457 | 265 70 | MIRb | 3’ |
| | | 3457 3651 | 65 262 | MIR | 3’ |
| CDK5RAP3 | NM_176095 | 838 1023 | 268 66 | MIRb | 3’ |

| | | | | | |
|---|---|---|---|---|---|
| | NM_025197<br>NM_176096 | 838 1023<br>- | 268 66<br>- | MIRb<br>- | 3’<br>- |
| CDKN2A | NM_000077<br>NM_058197<br>NM_058195 | 1045 1128<br>1319 1402<br>1036 1119 | 149 60<br>149 60<br>149 60 | MIRb<br>MIRb<br>MIRb | 3’<br>3’<br>3’ |
| CDKN2B | NM_004936<br>NM_078487 | 2606 2743<br>2729 2866 | 252 91<br>252 91 | MIRb<br>MIRb | 3’<br>3’ |
| CECR6 | NM_031890 | 3350 3513 | 41 208 | MIR3 | 3’ |
| CENPC1 | NM_001812<br>BC030695 | -<br>1918 2005 | -<br>180 94 | -<br>MIR3 | -<br>3’ |
| CENTB2 | NM_012287 | 6167 6219 | 82 31 | MIR | 3’ |
| CENTD3 | NM_022481 | 4862 4967 | 81 201 | MIR3 | 3’ |
| CES4 | NM_016280 | 958 1158 | 208 1 | MIR3 | 3’ |
| CFP | NM_002621 | 12 148 | 193 69 | MIRb | 5’ |
| CGNL1 | NM_032866 | 6680 6860 | 3 207 | MIR3 | 3’ |
| CGREF1 | NM_006569 | 1653 1852 | 274 60 | MIRb | 3’ |
| CHL1 | NM_006614 | 6582 6726 | 26 171 | THER1_MD | 3’ |
| CHRD | NM_003741<br>NM_177978<br>NM_177979 | -<br>498 551<br>- | -<br>134 81<br>- | -<br>MIR<br>- | -<br>TAG<br>- |
| CHRDL2 | NM_015424 | 1372 1429 | 135 78 | MIR | CDS |
| CHST5 | NM_012126<br>NM_024533 | -<br>849 1020 | -<br>207 39 | -<br>MIR | -<br>5’ |
| CHURC1 | NM_145165 | 1612 1804<br>2019 2114<br>2204 2247<br>2566 2604 | 28 216<br>261 151<br>140 97<br>114 76 | MIRb<br>MIR<br>MIR<br>MIR | 3’<br>3’<br>3’<br>3’ |
| CIITA | NM_000246<br>AF410154 | -<br>769 862 | -<br>174 76 | -<br>MIR | -<br>TAG |
| CLDN1 | NM_021101 | 1894 2048<br>2432 2669 | 22 191<br>6 240 | MIRb<br>MIRb | 3’<br>3’ |
| CLDN10 | NM_182848<br>NM_006984 | 2267 2435<br>2091 2259 | 221 39<br>221 39 | MIR<br>MIR | 3’<br>3’ |
| CLEC12B | NM_205852 | 1081 1252 | 173 7 | MIR | 3’ |
| CLEC4M | NM_014257<br>NM_214675<br>NM_214676<br>NM_214677<br>NM_214678<br>NM_214679 | -<br>-<br>-<br>885 972<br>885 972<br>822 909 | -<br>-<br>-<br>69 155<br>69 155<br>69 155 | -<br>-<br>-<br>MIRb<br>MIRb<br>MIRb | -<br>-<br>-<br>3’<br>3’<br>3’ |
| CLEC5A | NM_013252 | 2780 2972<br>3083 3133 | 8 198<br>93 145 | MIRb<br>MIR | 3’<br>3’ |
| CLIP4 | NM_024692 | 2901 3066 | 191 16 | MIR3 | 3’ |
| CLN5 | NM_006493 | 3705 3881 | 225 43 | MIRb | 3’ |
| CLN8 | NM_018941 | 4233 4440 | 234 14 | MIR | 3’ |
| CLYBL | NM_138280<br>NM_206808 | 2392 2546<br>- | 79 251<br>- | MIR<br>- | 3’<br>- |
| CNOT7 | NM_013354<br>NM_054026 | -<br>1234 1399 | -<br>206 17 | -<br>MIR3 | -<br>3’ |
| CNTF | NM_000614 | 1139 1367 | 268 20 | MIRb | 3’ |
| CNTN2 | NM_005076 | 4602 4792 | 199 13 | MIR3 | 3’ |
| CNTN3 | NM_020872 | 3675 3814 | 19 164 | MIRc | 3’ |
| COBLL1 | NM_014900 | 2188 2347 | 27 198 | MIRb | 5’ |
| COL28A1 | NM_001037763<br>XM_209824 | -<br>1438 1652 | -<br>225 8 | -<br>MIRb | -<br>3’ |
| COL4A6 | NM_001847<br>NM_033641 | -<br>- | -<br>- | -<br>- | -<br>- |

| | | | | | |
|---|---|---|---|---|---|
| | BC005305 | 876 1105 | 7 268 | MIRb | 3’ |
| COL8A1 | NM_001850<br>NM_020351<br>AF170702 | -<br>-<br>1966 2090 | -<br>-<br>66 194 | -<br>-<br>MIR3 | -<br>-<br>3’ |
| CORO2B | NM_006091 | 2095 2305<br>2651 2788 | 223 7<br>261 86 | MIRb<br>MIR_Mars | 3’<br>3’ |
| COX15 | NM_078470<br>NM_004376 | 3618 3672<br>2379 2433 | 151 205<br>151 205 | MIR3<br>MIR3 | 3’<br>3’ |
| CPLX3 | NM_001030005 | 1020 1177 | 19 188 | MIR3 | 3’ |
| CPM | NM_001874<br>NM_198320<br>NM_001005502 | 5342 5368<br>5328 5354<br>5310 5336 | 127 153<br>127 153<br>127 153 | THER1_MD<br>THER1_MD<br>THER1_MD | 3’<br>3’<br>3’ |
| CPNE5 | NM_020939 | 2720 2880 | 31 193 | MIR3 | 3’ |
| CRB1 | NM_201253<br>BX640729 | -<br>3555 3702 | -<br>257 94 | -<br>MIRb | -<br>3’ |
| CRB2 | NM_173689 | 4729 4837 | 136 243 | MIR_Mars | 3’ |
| CREBL2 | NM_001310 | 3588 3693 | 81 197 | MIRb | 3’ |
| CREG2 | NM_153836 | 1867 2008 | 255 109 | MIRb | 3’ |
| CRELD1 | NM_001031717<br>NM_015513<br>BC008720 | -<br>-<br>1127 1305 | -<br>-<br>216 38 | -<br>-<br>MIR | -<br>-<br>CDS |
| CRHR2 | NM_001883 | 1530 1587 | 125 67 | MIR | 3’ |
| CRNKL1 | NM_016652 | 2931 3103 | 37 256 | MIRb | 3’ |
| CRSP2 | NM_004229 | 7622 7855 | 262 13 | MIR | 3’ |
| CRTAC1 | NM_018058 | 2616 2854 | 1 252 | MIRb | 3’ |
| CRYZL1 | NM_145858<br>AK057604 | 1029 1201<br>1608 1780 | 218 35<br>218 35 | MIR<br>MIR | 3’<br>3’ |
| CSF3R | NM_000760<br>NM_156038<br>NM_156039<br>NM_172313 | 99 179<br>99 179<br>99 179<br>99 179 | 133 57<br>133 57<br>133 57<br>133 57 | MIR<br>MIR<br>MIR<br>MIR | TAG<br>TAG<br>TAG<br>TAG |
| CSMD2 | NM_052896 | 11451 11656 | 235 7 | MIRb | 3’ |
| CUL1 | NM_003592<br>BC034318 | -<br>61 170 | -<br>49 179 | -<br>THER1_MD | -<br>5’ |
| CUL3 | NM_003590 | 5849 5970<br>5971 6049 | 268 143<br>111 29 | MIR<br>MIRb | 3’<br>3’ |
| CX3CR1 | NM_001337 | 2146 2228<br>2198 2248 | 43 154<br>117 171 | MIRm<br>MIR | 3’<br>3’ |
| CXorf52 | NM_173168<br>AY168775 | 266 307<br>266 307 | 110 151<br>110 151 | MIRb<br>MIRb | TAG<br>TAG |
| CXorf9 | NM_018990 | 1816 1938 | 65 202 | MIR3 | 3’ |
| CYB561D1 | NM_182580 | 1090 1257 | 2 190 | MIR3 | 3’ |
| CYB5D1 | NM_144607 | 1639 1792 | 182 24 | MIRb | 3’ |
| CYB5RL | NM_001031672 | 1319 1464<br>2025 2082<br>2301 2482<br>2938 3059 | 20 170<br>28 84<br>190 6<br>1 122 | MIRc<br>MIR3<br>MIR3<br>MIRc | 3’<br>3’<br>3’<br>3’ |
| CYBB | NM_000397 | 1877 2123 | 262 1 | MIR_Mars | 3’ |
| CYBRD1 | NM_024843 | 1773 2012 | 241 1 | MIR | 3’ |
| CYLD | NM_015247 | 4808 4883 | 118 194 | MIR3 | 3’ |
| CYorf15B | NM_032576 | 2336 2517 | 262 77 | MIR | 3’ |
| CYP2U1 | NM_183075 | 2000 2104<br>3050 3248 | 99 197<br>261 37 | THER1_MD<br>MIRb | 3’<br>3’ |
| CYP4F3 | NM_000896 | 2758 2844 | 111 20 | MIR | 3’ |
| CYP4V2 | NM_207352 | 3257 3378 | 184 57 | MIR3 | 3’ |
| CYSLTR2 | NM_020377 | 2135 2298 | 253 73 | MIRb | 3’ |
| DAAM2 | NM_015345 | - | - | - | - |

| | | | | | |
|---|---|---|---|---|---|
| | BC047575 | 1409 1570 | 190 41 | MIR3 | 3’ |
| DAB2IP | NM_032552 | - | - | - | - |
| | NM_138709 | 41 137 | 128 226 | THER1_MD | 5’ |
| DACT3 | NM_145056 | 1832 1938 | 30 145 | MIR | 3’ |
| DARS2 | NM_018122 | 2904 3100 | 89 263 | MIR_Mars | 3’ |
| DBF4B | NM_145663 | 2936 3068 | 64 199 | MIR3 | 3’ |
| | NM_025104 | 765 887 | 184 61 | MIRb | 3’ |
| DBH | NM_000787 | 2574 2718 | 12 175 | MIRb | 3’ |
| DCDC1 | NM_198462 | 4281 4423 | 147 4 | MIRb | 3’ |
| | NM_181807 | - | - | - | - |
| DCLRE1B | NM_022836 | 3106 3260 | 89 263 | MIR_Mars | 3’ |
| DCLRE1C | NM_001033855 | - | - | - | - |
| | NM_022487 | - | - | - | - |
| | NM_001033858 | 242 329 | 137 51 | MIR | 5’ |
| | NM_001033857 | 242 329 | 137 51 | MIR | 5’ |
| DCX | NM_000555 | 3674 3737 | 172 110 | MIR3 | 3’ |
| | | 4465 4680 | 203 3 | MIR3 | 3’ |
| | NM_178152 | 3337 3400 | 172 110 | MIR3 | 3’ |
| | | 4128 4343 | 203 3 | MIR3 | 3’ |
| | NM_178153 | 3322 3385 | 172 110 | MIR3 | 3’ |
| | | 4113 4328 | 203 3 | MIR3 | 3’ |
| | NM_178151 | 3390 3453 | 172 110 | MIR3 | 3’ |
| | | 4181 4396 | 203 3 | MIR3 | 3’ |
| DDN | NM_015086 | 2616 2792 | 16 201 | MIR3 | 3’ |
| | | 3548 3730 | 73 253 | MIRb | 3’ |
| DDR1 | NM_013993 | - | - | - | - |
| | NM_001954 | - | - | - | - |
| | NM_013994 | 3 165 | 90 258 | MIR | 5’ |
| DDX52 | NM_007010 | 3265 3357 | 77 170 | MIRb | 3’ |
| DDX58 | NM_014314 | 4164 4304 | 95 260 | MIRb | 3’ |
| DEC1 | NM_017418 | 91 161 | 42 111 | MIR | 5’ |
| DEF8 | NM_017702 | - | - | - | - |
| | NM_207514 | 3082 3180 | 70 166 | MIR | 3’ |
| DENND1B | NM_144977 | 2116 2174 | 253 193 | MIR | 3’ |
| DENND2C | NM_198459 | 3845 3975 | 261 126 | MIRb | 3’ |
| DENND5B | NM_144973 | 6214 6328 | 161 45 | MIR3 | 3’ |
| DGKA | NM_201444 | - | - | - | - |
| | NM_201445 | 87 208 | 124 1 | MIRb | 5’ |
| | NM_001345 | - | - | - | - |
| | NM_201554 | - | - | - | - |
| DHCR24 | NM_014762 | 1820 2045 | 5 262 | MIRb | 3’ |
| | | 2980 3108 | 34 168 | MIR | 3’ |
| DHCR7 | NM_001360 | 1699 1742 | 104 153 | MIRb | 3’ |
| DHRS12 | NM_001031719 | 912 1040 | 229 103 | MIR | 3’ |
| | NM_024705 | - | - | - | - |
| DHX33 | NM_020162 | 2332 2495 | 64 235 | MIRb | 3’ |
| DIS3 | NM_014953 | 5683 5792 | 233 121 | MIRb | 3’ |
| DISC1 | NM_001012957 | 6120 6332 | 252 15 | MIRb | 3’ |
| | NM_001012958 | - | - | - | - |
| | NM_001012959 | 2049 2165 | 133 11 | MIR | TAG |
| | NM_018662 | 6186 6398 | 252 15 | MIRb | 3’ |
| CTXN2 | NM_001145668 | 1181 1293 | 9 131 | MIRm | 3’ |
| | | 1408 1635 | 29 261 | MIR | 3’ |
| | | 1687 1807 | 87 200 | MIR3 | 3’ |
| | | 1935 2134 | 196 8 | MIR | 3’ |
| DKK3 | NM_015881 | 1831 2060 | 251 4 | MIRb | 3’ |
| | NM_013253 | 1817 2046 | 251 4 | MIRb | 3’ |

| | | | | | |
|---|---|---|---|---|---|
| | NM_001018057 | 1649 1878 | 251 4 | MIRb | 3’ |
| DLEU1 | NR_002605 | 223 280 | 219 167 | MIRb | ncRNA |
| | AF490255 | 64 121 | 219 167 | MIRb | |
| DLGAP4 | NM_014902 | 4835 5007 | 44 204 | MIR | 3’ |
| | NM_183006 | 3622 3794 | 44 204 | MIR | 3’ |
| DMBT1 | NM_004406 | 5693 5799 | 109 215 | MIRb | 3’ |
| | NM_007329 | 7577 7683 | 109 215 | MIRb | 3’ |
| | NM_017579 | 7547 7653 | 109 215 | MIRb | 3’ |
| DMRTC2 | NM_033052 | 1449 1558 | 136 33 | MIRb | 3’ |
| DMWD | NM_004943 | 1949 2056 | 140 30 | MIR | CDS |
| | | 2700 2770 | 154 77 | MIR3 | 3’ |
| DNAJB13 | NM_153614 | - | - | - | - |
| | AF516185 | 1233 1299 | 144 73 | THER1_MD | TAG |
| DNAJB14 | NM_001031723 | - | - | - | - |
| | NM_024920 | 1507 1640 | 73 224 | MIRb | 3’ |
| DNAJB8 | NM_153330 | 845 911 | 48 121 | MIR | 5’ |
| DNAJC18 | NM_152686 | 2080 2246 | 254 68 | MIRb | 3’ |
| DNASE2 | NM_001375 | 1836 1995 | 34 195 | MIRb | 3’ |
| DOCK2 | NM_004946 | - | - | - | - |
| | BC016996 | 1498 1589 | 91 190 | MIRb | 3’ |
| DOCK5 | NM_024940 | 6830 6911 | 204 118 | MIR3 | 3’ |
| DOCK8 | NM_203447 | - | - | - | - |
| | BC045629 | 1592 1722 | 192 31 | MIR | TAG |
| DPM2 | NM_152690 | 183 341 | 246 91 | MIRb | TAG |
| | NM_003863 | - | - | - | - |
| | BC015374 | 153 311 | 246 91 | MIRb | TAG |
| DPP4 | NM_001935 | 19 155 | 43 183 | MIR3 | 5’ |
| DPP6 | NM_130797 | - | - | - | - |
| | NM_001936 | - | - | - | - |
| | NM_001039350 | - | - | - | - |
| | BC035912 | 1039 1123 | 133 49 | MIRb | ATG |
| DPP8 | NM_130434 | - | - | - | - |
| | NM_017743 | - | - | - | - |
| | NM_197960 | - | - | - | - |
| | NM_197961 | - | - | - | - |
| | BC040203 | 13 252 | 260 10 | MIR3 | 5’ |
| DPYSL2 | NM_001386 | 3290 3378 | 109 22 | MIRc | 3’ |
| DSC2 | NM_024422 | 4987 5123 | 87 215 | THER1_MD | 3’ |
| | NM_004949 | 5033 5169 | 87 215 | THER1_MD | 3’ |
| DUOX1 | NM_017434 | 109 239 | 134 4 | MIR3 | 5’ |
| | NM_175940 | - | - | - | - |
| DUSP18 | NM_152511 | 2128 2327 | 35 267 | MIRb | 3’ |
| DUSP3 | NM_004090 | 2563 2748 | 198 24 | MIRb | 3’ |
| | | 3486 3535 | 139 187 | MIR3 | 3’ |
| | BC008286 | 1199 1367 | 198 24 | MIRb | ATG |
| | | 2122 2171 | 139 187 | MIR3 | 3’ |
| DVL3 | NM_004423 | 4906 5060 | 28 189 | MIRb | 3’ |
| DYNC2LI1 | NM_016008 | - | - | - | - |
| | NM_015522 | - | - | - | - |
| | NM_001012665 | - | - | - | - |
| | BC040558 | 1671 1743 | 119 191 | MIRb | 3’ |
| DYNLRB1 | NM_014183 | - | - | - | - |
| | NM_177954 | 557 718 | 6 171 | MIR | 3’ |
| | | 813 963 | 207 61 | MIRb | 3’ |
| DZIP1 | NM_014934 | 4958 5103 | 203 47 | MIR | 3’ |
| | | 6076 6244 | 39 221 | MIR | 3’ |
| | NM_198968 | 5015 5160 | 203 47 | MIR | 3’ |

| | | | | | |
|---|---|---|---|---|---|
| | | 6133 6301 | 39 221 | MIR | 3’ |
| E2F6 | NM_001952 | 2120 2312 | 53 264 | MIRm | 3’ |
| | NM_198256 | 2191 2383 | 59 264 | MIRm | 3’ |
| | NM_198325 | 2253 2445 | 53 264 | MIRm | 3’ |
| | NM_198258 | 2324 2516 | 53 264 | MIRm | 3’ |
| | NM_198257 | 2065 2257 | 59 264 | MIRm | 3’ |
| | NM_212540 | 2044 2236 | 59 264 | MIRm | 3’ |
| EBF2 | NM_02265 | - | - | - | - |
| | AK001144 | 1714 1837 | 1 131 | MIRb | 3’ |
| | | 2059 2218 | 204 34 | MIR3 | 3’ |
| EBI3 | NM_005755 | 776 892 | 40 154 | MIRb | 3’ |
| EDA | NM_001399 | 3344 3445 | 129 13 | MIRb | 3’ |
| | | 4568 4688 | 203 89 | MIR3 | 3’ |
| | NM_001005609 | 3338 3439 | 129 13 | MIRb | 3’ |
| | | 4562 4682 | 203 89 | MIR3 | 3’ |
| | NM_001005610 | - | - | - | - |
| | NM_001005611 | - | - | - | - |
| | NM_001005612 | - | - | - | - |
| | NM_001005613 | - | - | - | - |
| | NM_001005614 | - | - | - | - |
| | NM_001005615 | - | - | - | - |
| EDG6 | NM_003775 | 1452 1495 | 82 125 | MIR | 3’ |
| EEPD1 | NM_030636 | 3477 3597 | 182 51 | MIRb | 3’ |
| EFEMP2 | NM_016938 | 1829 1935 | 39 156 | MIR3 | 3’ |
| EGLN2 | NM_053046 | - | - | - | - |
| | NM_017555 | 1325 1560 | 159 37 | MIR | 5’ |
| | | 1775 1887 | 159 37 | THER1_MD | 5’ |
| | NM_080732 | - | - | - | - |
| EHD3 | NM_014600 | 2678 2881 | 254 54 | MIRb | 3’ |
| EHD4 | NM_139265 | 1689 1869 | 25 205 | MIR3 | 3’ |
| EHMT1 | NM_024757 | - | - | - | - |
| | AF461894 | 760 928 | 214 35 | MIR | 5’ |
| EIF1AX | NM_001412 | 3733 3967 | 5 229 | MIRb | 3’ |
| EIF4G3 | NM_003760 | 18 88 | 140 66 | MIRb | 5’ |
| ELF1 | NM_172373 | 2687 2816 | 65 202 | MIR3 | 3’ |
| ELF3 | NM_004433 | 849 866 | 204 185 | MIR3 | 5’ |
| | | 1162 1299 | 184 37 | MIR3 | 5’ |
| | | 1480 1703 | 16 220 | MIRb | 5’ |
| ELK4 | NM_001973 | 140 374 | 253 7 | MIRb | 5’ |
| | NM_021795 | 132 374 | 262 7 | MIRb | 5’ |
| EMR2 | NM_013447 | 4582 4635 | 209 155 | MIRb | 3’ |
| | | 4933 5078 | 154 7 | MIRb | 3’ |
| | NM_152916 | 4435 4488 | 209 155 | MIRb | 3’ |
| | | 4786 4931 | 154 7 | MIRb | 3’ |
| | NM_152917 | 4303 4356 | 209 155 | MIRb | 3’ |
| | | 4654 4799 | 154 7 | MIRb | 3’ |
| | NM_152918 | 4156 4209 | 209 155 | MIRb | 3’ |
| | | 4507 4652 | 154 7 | MIRb | 3’ |
| | NM_152919 | 4549 4602 | 209 155 | MIRb | 3’ |
| | | 4900 5045 | 154 7 | MIRb | 3’ |
| | NM_152920 | 4402 4455 | 209 155 | MIRb | 3’ |
| | | 4753 4898 | 154 7 | MIRb | 3’ |
| | NM_152921 | 4270 4323 | 209 155 | MIRb | 3’ |
| | | 4621 4766 | 154 7 | MIRb | 3’ |
| ENAM | NM_031889 | 5241 5368 | 137 270 | MIRm | 3’ |
| | | 5491 5558 | 198 135 | MIR3 | 3’ |
| ENSA | NM_207042 | - | - | - | - |
| | NM_207043 | 1216 1395 | 6 187 | MIRm | 3’ |

| | | | | | |
|---|---|---|---|---|---|
| | NM_004436 | - | - | - | - |
| | NM_207044 | 1168 1347 | 6 187 | MIRm | 3’ |
| | NM_207045 | - | - | - | - |
| | NM_207046 | - | - | - | - |
| | NM_207047 | 1032 1211 | 6 187 | MIRm | 3’ |
| | NM_207168 | - | - | - | - |
| EPB41L1 | NM_012156 | 5754 5953 | 3 201 | MIR3 | 3’ |
| | NM_177996 | 5347 5540 | 9 201 | MIR3 | 3’ |
| EPB41L4B | NM_018424 | 2251 2364 | 84 199 | MIR3 | 3’ |
| | NM_019114 | - | - | - | - |
| EPHA10 | NM_001004338 | - | - | - | - |
| | NM_173641 | 1137 1198 | 218 156 | MIR | 3’ |
| | | 1505 1626 | 168 42 | MIRb | 3’ |
| EPHA8 | NM_020526 | 4135 4294 | 181 13 | MIR3 | 3’ |
| | NM_001006943 | 1699 1839 | 109 254 | MIRb | 3’ |
| EPHB2 | NM_017449 | - | - | - | - |
| | NM_004442 | - | - | - | - |
| | BC067861 | 1565 1690 | 132 5 | MIR | TAG |
| EPM2A | NM_005670 | 2058 2179 | 133 5 | MIRb | 3’ |
| | | 2645 2868 | 252 8 | MIRb | 3’ |
| | NM_001018041 | - | - | - | - |
| ERBB3 | NM_001982 | 4724 4854 | 258 113 | MIRb | 3’ |
| | | 5152 5230 | 112 26 | MIRb | 3’ |
| | NM_001005915 | - | - | - | - |
| ERGIC1 | NM_001031711 | - | - | - | - |
| | NM_020462 | 708 881 | 22 204 | MIRb | 3’ |
| | | 1254 1362 | 203 85 | MIRb | 3’ |
| ERLIN2 | NM_007175 | - | - | - | - |
| | NM_001003790 | - | - | - | - |
| | BC067765 | 1095 1230 | 181 37 | MIR3 | 3’ |
| ERMAP | NM_001017922 | - | - | - | - |
| | NM_018538 | - | - | - | - |
| | BX537371 | 236 433 | 254 36 | MIRb | 5’ |
| ESR1 | NM_000125 | - | - | - | - |
| | BX640939 | 1247 1306 | 86 145 | MIRb | 3’ |
| | | 1695 1804 | 207 93 | MIR3 | 3’ |
| | | 3029 3227 | 8 254 | MIRb | 3’ |
| ETV6 | NM_001987 | 5309 5409 | 151 49 | MIR | 3’ |
| ETV7 | NM_016135 | 1507 1544 | 115 151 | MIRb | 3’ |
| EVI5 | NM_005665 | 4367 4416 | 116 72 | MIRb | 3’ |
| | | 4547 4592 | 71 31 | MIRb | 3’ |
| EXOC5 | NM_006544 | 5800 6040 | 267 10 | MIRb | 3’ |
| EXOSC1 | NM_016046 | 988 1049 | 251 189 | MIRb | 3’ |
| EXTL1 | NM_004455 | 613 786 | 3 186 | MIR3 | 5’ |
| F2RL1 | NM_005242 | 1713 1845 | 198 55 | MIR3 | 3’ |
| FABP2 | NM_000134 | 1846 1907 | 268 208 | THER1_MD | 3’ |
| FADD | NM_003824 | 1309 1371 | 75 10 | MIR | 3’ |
| FAIM2 | NM_012306 | 2814 2968 | 197 47 | MIRb | 3’ |
| | | 3159 3275 | 27 143 | MIR | 3’ |
| FAM100A | NM_145253 | - | - | - | - |
| | AF447881 | 299 434 | 208 61 | MIR3 | 5’ |
| | | 483 574 | 147 40 | MIR | 5’ |
| FAM109A | NM_144671 | 110 195 | 134 48 | MIR | 5’ |
| FAM110A | NM_207121 | 1 49 | 207 156 | MIR | 5’ |
| | NM_031424 | - | - | - | - |
| FAM111A | NM_022074 | 2157 2247 | 262 152 | MIR | 3’ |
| | NM_198847 | 2118 2218 | 262 141 | MIR | 3’ |

| | | | | | |
|---|---|---|---|---|---|
| FAM139A | NM_173678 | 2829 2916 | 3 94 | MIR | 3’ |
| FAM163A | NM_173509 | 2602 2713 | 25 152 | MIRb | 3’ |
| FAM167A | NM_053279 | 2388 2486 | 173 74 | MIR3 | 3’ |
| FAM168A | NM_015159<br>D87470 | -<br>2031 2152<br>3206 3335<br>3507 3663<br>4532 4637 | -<br>197 74<br>259 112<br>238 43<br>180 72 | -<br>MIR3<br>MIRc<br>MIRc<br>MIRc | -<br>3’<br>3’<br>3’<br>3’ |
| FAM179B | NM_015091<br>BC057255 | -<br>4604 4849 | -<br>240 4 | -<br>MIRc | -<br>3’ |
| FAM184A | NM_024581<br>NM_001100411<br>BX640728 | -<br>-<br>2285 2493 | -<br>-<br>27 253 | -<br>-<br>MIRb | -<br>-<br>3’ |
| FAM26C | NM_001001412 | 2073 2292<br>2385 2422<br>2726 2879 | 2 223<br>143 106<br>2 171 | MIRb<br>MIR<br>MIRb | 3’<br>3’<br>3’ |
| FAM26D | NM_153036 | 831 1017 | 40 257 | MIRb | 3’ |
| FAM40A | NM_033088 | 2530 2645<br>4323 4360 | 151 34<br>113 153 | MIR<br>MIRb | 3’<br>3’ |
| FAM49B | NM_016623<br>BC016345 | -<br>34 154 | -<br>136 16 | -<br>MIR | -<br>5’ |
| FAM50B | NM_012135 | 1390 1618 | 12 262 | MIR | 3’ |
| FAM53B | NM_014661<br>BC031654 | -<br>1333 1397 | -<br>179 111 | -<br>MIR | -<br>TAG |
| FAM62C | NM_031913 | 2140 2231 | 157 65 | MIRm | 3’ |
| FAM79B | NM_198485 | 3505 3615 | 87 191 | MIR | 3’ |
| FAM92B | NM_198491 | 1339 1532<br>1706 1884 | 70 261<br>43 254 | MIR<br>MIRb | 3’<br>3’ |
| FAM98B | NM_173611 | 1250 1382 | 14 142 | MIRb | 3’ |
| FANCF | NM_022725 | 2147 226 | 42 177 | MIR3 | 3’ |
| FAT3 | XM_926199<br>XM_936538 | 16562 16728<br>16561 16727 | 185 6<br>185 6 | MIR3<br>MIR3 | 3’<br>3’ |
| FBLIM1 | NM_017556<br>NM_001024215<br>NM_001024216 | 77 212<br>-<br>- | 178 33<br>-<br>- | MIRb<br>-<br>- | 5’<br>-<br>- |
| FBXL12 | NM_017703 | 1337 1475 | 21 173 | MIRb | 3’ |
| FBXL7 | NM_012304 | 2543 2624 | 48 135 | MIR | 3’ |
| FBXO18 | NM_032807<br>NM_178150 | 31 156<br>- | 6 125<br>- | MIR3<br>- | ATG<br>- |
| FBXO31 | NM_024735 | 2177 2309 | 13 152 | MIR | 3’ |
| FBXO44 | NM_033182<br>NM_183412<br>NM_183413<br>NM_001014765 | 2270 2305<br>2256 2291<br>2145 2180<br>277 373<br>2642 2677 | 205 240<br>205 240<br>205 240<br>14 119<br>205 240 | MIR<br>MIR<br>MIR<br>MIRm<br>MIR | 3’<br>3’<br>3’<br>5’<br>3’ |
| FCMD | NM_006731 | 4375 4414<br>5634 5764 | 67 30<br>116 254 | MIR<br>MIRb | 3’<br>3’ |
| FEZ2 | NM_005102<br>L17328 | -<br>114 276 | -<br>72 235 | -<br>MIR | -<br>5’ |
| FGD1 | NM_004463 | 3960 4111 | 37 194 | MIR3 | 3’ |
| FGD2 | NM_173558<br>BC062363 | -<br>699 768<br>955 1084 | -<br>147 75<br>30 157 | -<br>MIR<br>MIR | -<br>3’<br>3’ |
| FGD3 | NM_033086 | 4453 4551 | 129 31 | MIR | 3’ |
| FGF7 | NM_002009 | 3385 3517 | 202 46 | MIR3 | 3’ |
| FGR | NM_005248 | 1832 1924 | 204 112 | MIR3 | 3’ |
| FHAD1 | XM_934878 | - | - | - | - |

| | | | | | |
|---|---|---|---|---|---|
| | XM_934883 | - | - | - | - |
| | XM_934889 | 3227 3343 | 141 22 | MIRb | 3’ |
| | XM_934892 | - | - | - | - |
| | XM_934885 | - | - | - | - |
| | XM_057107 | 3786 3994 | 231 22 | MIRb | 3’ |
| | XM_934887 | 1026 1205 | 78 256 | MIR | 3’ |
| | XM_934881 | - | - | - | - |
| | XM_934876 | - | - | - | - |
| | XM_934890 | 20 130 | 140 37 | MIRb | CDS |
| FIBIN | NM_203371 | 2635 2706 | 170 93 | MIRc | 3’ |
| FIGNL1 | NM_022116 | 48 160 | 135 17 | MIR | 5' |
| | | 225 3178 | 4 98 | MIR | 5’ |
| FILIP1 | NM_015687 | 4224 4369 | 33 199 | MIR3 | 3’ |
| FKRP | NM_001039885 | 2376 2503 | 4 137 | MIR | 3’ |
| | | 3088 3344 | 262 9 | MIR | 3’ |
| | NM_024301 | 2157 2284 | 4 137 | MIR | 3’ |
| | | 2869 3125 | 262 9 | MIR | 3’ |
| FLT3LG | NM_001459 | - | - | - | - |
| | U03858 | 818 915 | 186 87 | MIRb | 3’ |
| FMNL3 | NM_175736 | 7401 7646 | 255 11 | MIRb | 3’ |
| | | 7772 7866 | 127 26 | MIR | 3’ |
| | | 8187 8376 | 256 15 | MIRb | 3’ |
| | | 10124 10264 | 146 1 | MIR | 3’ |
| | NM_198900 | 7248 7493 | 255 11 | MIRb | 3’ |
| | | 7619 7713 | 127 26 | MIR | 3’ |
| | | 8034 8223 | 256 15 | MIRb | 3’ |
| | | 9971 10111 | 146 1 | MIR | 3’ |
| FNDC3B | NM_022763 | 4804 4965 | 3 165 | MIR | 3’ |
| FOXD2 | NM_004474 | 135 241 | 51 157 | MIR3 | 5’ |
| | | 351 496 | 166 9 | MIR | 5’ |
| | | 472 622 | 138 58 | MIRb | 5’ |
| FOXI1 | NM_012188 | 1411 1609 | 245 37 | MIRb | 3’ |
| | NM_144769 | 1126 1326 | 245 37 | MIRb | 3’ |
| FRAG1 | NM_014489 | 1558 1690 | 53 181 | MIR3 | 3’ |
| FREQ | NM_014286 | 2112 2271 | 30 202 | MIR3 | 3’ |
| FRMPD2L1 | NM_001042524 | 1512 1675 | 254 94 | MIRb | 3’ |
| | NM_001042525 | 1498 1661 | 254 94 | MIRb | 3’ |
| FRMPD2L2 | NM_001042515 | 1512 1675 | 254 94 | MIRb | 3’ |
| | NM_001042516 | 1498 1661 | 254 94 | MIRb | 3’ |
| FRRS1 | NM_001013660 | 509 571 | 118 53 | MIRb | 5’ |
| FSTL1 | NM_007085 | 2113 2171 | 73 132 | MIR | 3’ |
| FTO | XM_051200 | 2499 2623 | 143 25 | MIR3 | 3’ |
| | | 2561 2778 | 42 261 | MIRm | 3’ |
| | | 3155 3278 | 87 219 | MIRb | 3’ |
| | XM_931743 | 1253 1470 | 42 261 | MIRm | 3’ |
| | | 1847 1970 | 87 219 | MIRb | 3’ |
| | XM_931747 | 1448 1572 | 143 25 | MIR3 | 3’ |
| | | 1510 1727 | 42 261 | MIRm | 3’ |
| | | 2104 2227 | 87 219 | MIRb | 3’ |
| | XM_931752 | 1424 1548 | 143 25 | MIR3 | 3’ |
| | | 1486 1703 | 42 261 | MIRm | 3’ |
| | | 2080 2203 | 87 219 | MIRb | 3’ |
| FUT10 | NM_032664 | 3146 3273 | 49 179 | MIR3 | 3’ |
| FUT8 | NM_178155 | 101 193 | 64 161 | MIRb | 5’ |
| | NM_178154 | - | - | - | - |
| | NM_178156 | 101 193 | 64 161 | MIRb | 5’ |
| | NM_004480 | - | - | - | - |
| | NM_178157 | - | - | - | - |

| | | | | | |
|---|---|---|---|---|---|
| FXC1 | NM_012192 | 746 817 | 188 116 | MIRb | 3’ |
| FXN | NM_000144 | 1838 2008 | 44 208 | MIR3 | 3’ |
| | NM_181425 | 1846 2016 | 44 208 | MIR3 | 3’ |
| FYB | NM_001465 | 3803 3916 | 57 185 | MIRb | 3’ |
| | NM_199335 | 3665 3778 | 57 185 | MIRb | 3’ |
| GAB2 | NM_080491 | 5118 5186 | 110 180 | MIR3 | 3’ |
| | NM_012296 | 5164 5232 | 110 180 | MIR3 | 3’ |
| GAB3 | NM_080612 | 2370 2630 | 249 4 | MIRb | 3’ |
| GABBR2 | NM_005458 | 4457 4675 | 19 221 | MIRb | 3’ |
| GABRA4 | NM_000809 | 3608 3703 | 204 114 | MIR3 | 3’ |
| | | 5071 5220 | 51 201 | MIR3 | 3’ |
| GABRE | NM_004961 | 2637 2770 | 245 117 | MIR_Mars | 3’ |
| GAFA2 | XM_001725248 | 27 78 | 91 147 | MIR | CDS |
| GALNT10 | NM_198321 | - | - | - | - |
| | NM_017540 | 1 157 | 96 262 | MIRb | 3’ |
| GALNT11 | NM_022087 | 1740 1953 | 242 8 | MIRb | 3’ |
| GALNT4 | NM_003774 | 3003 3042 | 29 71 | THER1_MD | 3’ |
| | | 3073 3199 | 72 202 | MIR3 | 3’ |
| GAS2L3 | NM_174942 | 75 153 | 140 58 | MIRb | 5’ |
| GAS7 | NM_003644 | 5945 6028 | 90 173 | MIR3 | 3’ |
| | | 7079 7167 | 61 147 | MIR | 3’ |
| | NM_201432 | 6353 6436 | 90 173 | MIR3 | 3’ |
| | | 7487 7575 | 61 147 | MIR | 3’ |
| | NM_201433 | 6383 6466 | 90 173 | MIR3 | 3’ |
| | | 7517 7605 | 61 147 | MIR | 3’ |
| GCET2 | NM_152785 | 657 704 | 109 158 | MIR3 | TAG |
| | | 1335 1501 | 207 20 | MIR3 | 3’ |
| | | 2551 2694 | 4 156 | MIRm | 3’ |
| | NM_001008756 | 824 871 | 109 158 | MIR3 | TAG |
| | | 1502 1668 | 207 20 | MIR3 | 3’ |
| | | 2718 2861 | 4 156 | MIRm | 3’ |
| GCLC | NM_001498 | - | - | - | - |
| | BC022487 | 1738 1907 | 5 194 | MIR3 | 3’ |
| GCNT2 | NM_145649 | 2932 3068 | 246 111 | MIRb | 3’ |
| | NM_001491 | 3200 3336 | 246 111 | MIRb | 3’ |
| | NM_145655 | 2464 2509 | 256 210 | MIR | 3’ |
| | | 2728 2864 | 246 111 | MIRb | 3’ |
| GDAP2 | NM_017686 | 2659 2762 | 108 221 | MIRb | 3’ |
| GFI1B | NM_004188 | 1277 1354 | 72 149 | MIR | 3’ |
| GGCX | NM_000821 | 2913 2982 | 260 191 | MIRb | 3’ |
| | | 2954 3088 | 166 20 | MIR_Mars | 3’ |
| GGTA1 | NR_003191 | 1258 1377 | 127 251 | MIR | ncRNA |
| GHDC | NM_032484 | 295 438 | 179 46 | MIRb | 5’ |
| GIPC1 | NM_005716 | 94 153 | 136 78 | MIR | 5’ |
| | NM_202467 | 94 153 | 136 78 | MIR | 5’ |
| | NM_202468 | - | - | - | - |
| | NM_202469 | - | - | - | - |
| | NM_202470 | - | - | - | - |
| | NM_202494 | - | - | - | - |
| GJB3 | NM_024009 | 268 384 | 236 118 | MIR_Mars | 5’ |
| | NM_001005752 | - | - | - | - |
| GLB1L3 | NM_001080407 | 1687 1879 | 48 250 | MIRb | CDS |
| GLDN | NM_181789 | 1474 1608 | 12 163 | MIR3 | 3’ |
| GNA12 | NM_007353 | 3597 3715 | 90 199 | MIR3 | 3’ |
| GOLGA8A | NM_181077 | - | - | - | - |
| | AF163441 | 11 199 | 48 246 | MIRb | 5’ |
| GOLGA8B | NM_001023567 | 54 144 | 159 235 | MIR | 5’ |

| | | | | | |
|---|---|---|---|---|---|
| | | 385 560 | 48 233 | MIRb | 5’ |
| GOLPH2 | NM_016548 | - | - | - | - |
| | NM_177937 | 2590 2674 | 201 113 | MIR3 | 3’ |
| GP5 | NM_004488 | 3276 3527 | 18 268 | MIRb | 3’ |
| GPI | NM_000175 | - | - | MIRb | - |
| | AB209575 | 1757 1869 | 163 47 | | 3’ |
| GPR114 | NM_153837 | 247 421 | 46 182 | MIRm | 5’ |
| | | 2362 2504 | 23 147 | MIRb | 3’ |
| | | 3561 3761 | 24 248 | MIR | 3’ |
| GPR132 | NM_013345 | 2902 3024 | 55 191 | MIRb | 3’ |
| GPR135 | NM_022571 | 1846 1908 | 83 23 | MIR | 3’ |
| GPR155 | NM_001033045 | 3038 3198 | 60 216 | MIRb | 3’ |
| | NM_152529 | 2889 3049 | 60 216 | MIRb | 3’ |
| GPR171 | NM_013308 | 98 178 | 13 94 | MIR | 5’ |
| GPR26 | NM_153442 | 5610 5702 | 187 106 | MIR | 3’ |
| GPR44 | NM_004778 | 1832 1987 | 213 57 | MIRb | 3’ |
| GPR78 | NM_080819 | 1722 1775 | 81 134 | MIR | 3’ |
| GPR81 | NM_032554 | 3179 3276 | 100 201 | MIR | 3’ |
| | | 3365 3425 | 182 246 | MIRb | 3’ |
| GPR97 | NM_170776 | 1731 1927 | 27 208 | MIR3 | 3’ |
| GPX7 | NM_015696 | 912 1024 | 132 252 | MIR | 3’ |
| GRAMD1B | XM_370660 | 5729 5896 | 23 190 | MIR3 | 3’ |
| GREB1 | NM_014668 | 7521 7732 | 223 4 | MIRb | 3’ |
| | NM_033090 | - | - | - | - |
| | NM_148903 | 1896 1985 | 140 50 | MIRb | 3’ |
| GRIP2 | XM_042936 | - | - | - | - |
| | XM_940982 | - | - | - | - |
| | XM_944910 | 1176 1348 | 181 8 | MIRb | 3’ |
| | | 1927 2064 | 61 201 | MIR3 | 3’ |
| | | 2594 2626 | 91 125 | MIRm | 3’ |
| GRPEL2 | NM_152407 | 3363 3518 | 51 224 | MIRb | 3’ |
| GSG1L | NM_144675 | 1582 1715 | 151 19 | MIRc | 3’ |
| | | 3151 3334 | 46 241 | MIR | 3’ |
| | NM_001109763 | 1966 2099 | 151 19 | MIRc | 3’ |
| | | 3535 3718 | 46 241 | MIR | 3’ |
| | AY302134 | 371 420 | 126 77 | MIR | CDS |
| GTPBP1 | NM_004286 | 3794 3931 | 5 147 | MIRb | 3’ |
| | | 4091 4148 | 220 158 | MIRb | 3’ |
| | AF077204 | 4454 4495 | 157 114 | MIRb | 3’ |
| | | 92 125 | 147 114 | MIRb | 5’ |
| GUCA1B | NM_002098 | 1839 1974 | 26 162 | MIR3 | 3’ |
| H2AFJ | NM_018267 | 1952 2070 | 143 268 | MIRb | 3’ |
| | NM_177925 | - | - | - | - |
| HAP1 | NM_003949 | 2729 2910 | 207 2 | MIRb | 3’ |
| | NM_177977 | 2573 2754 | 207 2 | MIRb | 3’ |
| HAPLN3 | NM_178232 | - | - | - | - |
| | BC062320 | 139 274 | 140 18 | MIR | 5’ |
| HAPLN4 | NM_023002 | 2790 2848 | 179 123 | MIRb | 3’ |
| | | 3220 3347 | 34 175 | MIRb | 3’ |
| HAS1 | NM_001523 | 1861 1908 | 111 157 | MIR3 | 3’ |
| HBS1L | NM_006620 | 6683 6892 | 236 3 | MIRb | 3’ |
| HCK | NM_002110 | 1849 1896 | 105 151 | MIRb | 3’ |
| | | 1869 1951 | 118 193 | MIR3 | 3’ |
| HDAC6 | NM_006044 | - | - | - | - |
| | BC011498 | 653 805 | 96 274 | MIRb | 3’ |
| HDAC8 | NM_018486 | - | - | - | - |
| | AF212246 | 575 695 | 16 144 | MIR | 3’ |

| | | | | | |
|---|---|---|---|---|---|
| HDHD3 | NM_031219 | 8 142 | 19 150 | MIR | 5’ |
| | | 366 443 | 191 115 | MIRb | 5’ |
| | | 533 610 | 18 96 | MIR | 5’ |
| HEMK1 | NM_016173 | 2507 2611 | 159 46 | MIRm | 3’ |
| | | 3288 3482 | 28 232 | MIRb | 3’ |
| | | 4031 4071 | 157 117 | MIR3 | 3’ |
| HFE | NM_000410 | 2122 2215 | 241 149 | MIR | 3’ |
| | NM_139002 | - | | - | - |
| | NM_139003 | - | - | - | - |
| | NM_139004 | 1785 1910 | 245 125 | MIRb | 3’ |
| | NM_139005 | - | - | - | - |
| | NM_139006 | - | - | - | - |
| | NM_139007 | - | - | - | - |
| | NM_139008 | - | - | - | - |
| | NM_139009 | - | - | - | - |
| | NM_139010 | - | - | - | - |
| | NM_139011 | - | - | - | - |
| HHAT | NM_018194 | 1861 2009 | 175 1 | MIR3 | 3’ |
| HIF3A | NM_152794 | - | - | - | - |
| | NM_022462 | - | - | - | - |
| | NM_152795 | 1903 2008 | 151 36 | MIR3 | TAG |
| | NM_152796 | - | - | - | - |
| HIST1H2AC | NM_003512 | - | - | - | - |
| | U90551 | 1350 1466 | 80 210 | MIRb | 3’ |
| HK1 | NM_000188 | - | - | - | - |
| | NM_033496 | 3 128 | 20 153 | MIR | 5’ |
| | NM_033497 | - | - | - | - |
| | NM_033498 | - | - | - | - |
| | NM_033500 | - | - | - | - |
| HK2 | NM_000189 | 476 581 | 2 112 | MIRb | 5’ |
| | | 869 942 | 113 188 | MIRb | 5’ |
| HKR1 | NM_181786 | - | - | - | - |
| | BC053845 | 1 108 | 125 12 | MIR | 5’ |
| HM13 | NM_030789 | - | - | - | - |
| | NM_178580 | - | - | - | - |
| | NM_178581 | - | - | - | - |
| | NM_178582 | 540 691 | 22 154 | MIRb | 3’ |
| | | 877 1107 | 8 262 | MIR | 3’ |
| | | 1110 1293 | 202 18 | MIRb | 3’ |
| HNF4A | NM_178849 | 2692 2819 | 265 136 | MIRb | 3’ |
| | | 2876 2997 | 145 26 | MIRb | 3’ |
| | NM_000457.3 | 2722 2849 | 265 136 | MIRb | 3’ |
| | | 2906 3027 | 145 26 | MIRb | 3’ |
| | NM_178850 | 1285 1385 | 49 149 | MIRb | TAG |
| | NM_175914 | - | - | - | - |
| | NM_001030003 | - | - | - | - |
| | NM_001030004 | - | - | - | - |
| HNF4G | NM_004133 | 2285 2442 | 196 39 | MIRb | 3’ |
| HOP | NM_032495 | - | - | - | - |
| | NM_139211 | - | - | - | - |
| | NM_139212 | - | - | - | - |
| | AF492678 | 83 197 | 77 196 | MIR3 | 5’ |
| HPCAL4 | NM_016257 | 2796 2992 | 253 54 | MIRb | 3’ |
| HPS1 | NM_000195 | 2731 2883 | 45 262 | MIR | 3’ |
| | NM_182637 | 2684 2836 | 45 262 | MIR | 3’ |
| | NM_182639 | - | - | - | - |
| | NM_182638 | 2337 2489 | 45 262 | MIR | 3’ |
| HRH1 | NM_000861 | 2477 2637 | 217 57 | MIRb | 3’ |

| | | | | | |
|---|---|---|---|---|---|
| HRH2 | NM_022304 | 1669 1842 | 266 84 | MIRb | 3’ |
| HS1BP3 | NM_022460 | - | - | - | - |
| | BC027947 | 1086 1211 | 69 206 | MIR | 3’ |
| HS2ST1 | NM_012262 | 2585 2839 | 1 257 | MIR_Mars | 3’ |
| | | 6464 6510 | 8 55 | MIR3 | 3’ |
| HS6ST3 | NM_153456 | 3163 3361 | 251 13 | MIRb | 3’ |
| | | 5549 5749 | 210 6 | ***MIRb*** | 3’ |
| HSD11B1L | NM_198705 | 1145 1180 | 150 115 | MIRb | 3’ |
| | NM_198706 | 1348 1383 | 150 115 | MIRb | 3’ |
| | NM_198707 | 1308 1343 | 150 115 | MIRb | 3’ |
| | NM_198708 | 1236 1271 | 150 115 | MIRb | 3’ |
| | NM_198533 | 1213 1248 | 150 115 | MIRb | 3’ |
| | NM_198704 | 1050 1085 | 150 115 | MIRb | 3’ |
| HSDL2 | NM_032303 | 2725 2810 | 169 80 | MIRm | 3’ |
| HSH2D | NM_032855 | 1806 1988 | 246 53 | MIRb | 3’ |
| HSPA12B | NM_052970 | 2887 3022 | 10 143 | MIR3 | 3’ |
| HSPA14 | NM_016299 | - | - | - | - |
| | NM_001037538 | 1831 2027 | 39 256 | MIR | 3’ |
| | | 3193 3287 | 5 113 | MIR | 3’ |
| | | 3567 3646 | 103 187 | MIR3 | 3’ |
| | | 4265 4439 | 63 239 | MIR | 3’ |
| HSPB8 | NM_014365 | 434 515 | 56 145 | MIR | 5’ |
| HTR3E | NM_182589 | 5 209 | 57 262 | MIR | ATG |
| HTR4 | NM_000870 | 2371 2519 | 32 183 | MIR3 | 3’ |
| | NM_199453 | - | - | - | - |
| HUNK | NM_014586 | 6257 6394 | 201 62 | MIR3 | 3’ |
| HVCN1 | NM_001040107 | 87 174 | 140 52 | MIRb | 5’ |
| | NM_032369 | 60 149 | 142 52 | MIRb | 5’ |
| HYAL1 | NM_007312 | 2214 2436 | 268 19 | MIRb | 3’ |
| | NM_153282 | 1649 1861 | 259 19 | MIRb | 3’ |
| | NM_153283 | 1382 1594 | 259 19 | MIRb | 3’ |
| | NM_153284 | 1402 1614 | 259 19 | MIRb | 3’ |
| | NM_153285 | 964 1176 | 259 19 | MIRb | 3’ |
| | NM_153286 | 815 1027 | 259 19 | MIRb | 3’ |
| | NM_033159 | 1729 1951 | 268 19 | MIRb | 3’ |
| | NM_153281 | 2024 2246 | 268 19 | MIRb | 3’ |
| HYDIN | NM_017558 | - | - | - | - |
| | BC028351 | 2684 2754 | 38 110 | MIRb | 3’ |
| IBRDC1 | NM_152553 | 1044 1120 | 80 2 | MIR | CDS |
| ICMT | NM_012405 | 2258 2371 | 98 218 | MIRb | 3’ |
| IFIT5 | NM_012420 | 1656 1726 | 267 197 | MIRb | 3’ |
| IFT122 | NM_052985 | - | - | - | - |
| | NM_018262 | - | - | - | - |
| | NM_052989 | - | - | - | - |
| | NM_052990 | - | - | - | - |
| | AK024435 | 3427 3491 | 27 93 | MIRb | 3’ |
| | | 3519 3614 | 158 259 | MIR | 3’ |
| | AK096891 | 100 138 | 86 125 | MIR | 5’ |
| | AK124140 | 291 412 | 135 7 | MIR | 5’ |
| IFT81 | NM_014055 | - | - | - | - |
| | NM_031473 | 1863 1991 | 153 29 | MIR | 3’ |
| IGF1 | NM_000618 | 1745 1974 | 1 200 | MIRb | 3’ |
| IGF2BP2 | NM_006548 | - | - | - | - |
| | NM_001007225 | - | - | - | - |
| | AF057352 | 284 372 | 26 115 | MIR | 5’ |
| IGSF1 | NM_001555 | - | - | - | - |

| | | | | | |
|---|---|---|---|---|---|
| | NM_205833 | 1496 1659 | 5 168 | MIRb | 3’ |
| | | 1651 1813 | 78 259 | MIR | 3’ |
| IKBKG | NM_003639 | - | - | - | - |
| | AY114157 | 1 68 | 171 104 | MIR3 | ATG |
| IL10RA | NM_001558 | 2757 2884 | 6 153 | MIR | 3’ |
| | | 3379 3627 | 3 268 | MIRb | 3’ |
| IL12RB1 | NM_005535 | 2046 2100 | 141 87 | MIRb | 3’ |
| | NM_153701 | 1777 1842 | 191 255 | MIRb | 3’ |
| IL12RB2 | NM_001559 | 135 267 | 196 63 | MIR | 5’ |
| | | 3582 3626 | 275 231 | MIRm | 3’ |
| IL16 | NM_004513 | 3151 3255 | 268 152 | MIRb | 3’ |
| | | 4677 4890 | 225 22 | MIR_Mars | 3’ |
| | NM_172217 | 5573 5677 | 268 152 | MIRb | 3’ |
| | | 7099 7312 | 225 22 | MIR_Mars | 3’ |
| IL17RE | NM_153480 | - | - | - | - |
| | NM_153481 | - | - | - | - |
| | NM_144640 | - | - | - | - |
| | NM_153482 | - | - | - | - |
| | NM_153483 | - | - | - | - |
| | BC063110 | 1227 1362 | 108 259 | MIRb | 3’ |
| IL18BP | NM_173042 | 579 671 | 72 162 | MIR3 | 5’ |
| | NM_001039659 | - | - | - | - |
| | NM_001039660 | - | - | - | - |
| IL18R1 | NM_003855 | 2529 2768 | 2 253 | MIR | 3’ |
| IL1A | NM_000575 | 1944 2153 | 268 57 | MIR | 3’ |
| | | 2121 2334 | 18 219 | MIRb | 3’ |
| IL1RAP | NM_002182 | 2790 2957 | 21 197 | MIR3 | 3’ |
| | NM_134470 | - | - | - | - |
| IL20RB | NM_144717 | 1338 1485 | 3 151 | MIR3 | 3’ |
| IL21R | NM_021798 | 2343 2526 | 58 265 | MIRb | 3’ |
| | | 2735 2780 | 276 231 | MIRm | 3’ |
| | NM_181078 | 2456 2639 | 58 265 | MIRb | 3’ |
| | | 2848 2893 | 276 231 | MIRm | 3’ |
| | NM_181079 | 2358 2541 | 58 265 | MIRb | 3’ |
| | | 2750 2795 | 276 231 | MIRm | 3’ |
| IL22RA1 | NM_021258 | 2573 2794 | 36 260 | MIRb | 3’ |
| IL22RA2 | NM_052962 | 1 191 | 68 261 | MIRb | 5’ |
| | NM_181309 | 1 191 | 68 261 | MIRb | 5’ |
| | NM_181310 | 1 191 | 68 261 | MIRb | 5’ |
| IL23R | NM_144701 | 2281 2342 | 146 83 | MIR | 3’ |
| IL27RA | NM_004843 | 2347 2501 | 237 57 | MIRb | 3’ |
| IL28RA | NM_170743 | 2160 2357 | 35 259 | MIRb | 3’ |
| | | 3252 3322 | 152 81 | MIR | 3’ |
| | NM_173064 | 2073 2270 | 35 259 | MIRb | 3’ |
| | | 3165 3235 | 152 81 | MIR | 3’ |
| | NM_173065 | 2029 2226 | 35 259 | MIRb | 3’ |
| | | 3121 3191 | 152 81 | MIR | 3’ |
| IL2RB | NM_000878 | 3683 3782 | 63 166 | MIR3 | 3’ |
| IL6ST | NM_002184 | 148 240 | 9 101 | MIRb | 5’ |
| | NM_175767 | - | - | - | - |
| IL8 | NM_000584 | 907 984 | 113 183 | MIR | 3’ |
| IMP3 | NM_018285 | 668 692 | 141 117 | MIR | 3’ |
| INOC1 | NM_017553 | - | - | - | - |
| | BX640651 | 5252 5334 | 114 212 | MIRb | 3’ |
| INPP5D | NM_001017915 | 4089 4144 | 254 199 | MIRb | 3’ |
| | NM_005541 | 4086 4141 | 254 199 | MIRb | 3’ |
| INTS2 | NM_020748 | 5264 5329 | 119 194 | MIR | 3’ |

| | | | | | |
|---|---|---|---|---|---|
| INTU | NM_015693<br>BC051698 | -<br>3231 3466 | -<br>251 7 | -<br>MIRb | -<br>3' |
| IREB2 | NM_004136<br>BC017880 | -<br>1278 1481 | -<br>206 3 | -<br>MIR | -<br>3' |
| IRF6 | NM_006147 | 1894 2051 | 16 185 | MIR3 | 3 |
| ISCA1 | NM_030940 | 1166 1292 | 183 58 | MIR3 | 3' |
| ISG20 | NM_002201<br>BC016341 | -<br>566 645<br>1365 1444<br>1924 2020<br>2084 2180 | -<br>9 96<br>85 165<br>166 262<br>168 75 | -<br>MIRb<br>MIR<br>MIR<br>MIR3 | -<br>3'<br>3'<br>3'<br>3' |
| ISG20L1 | NM_022767 | 2183 2422 | 14 262 | MIRb | 3' |
| ITGA2 | NM_002203 | 5153 5302 | 47 198 | MIR3 | 3' |
| ITGAL | NM_002209 | 3612 3793 | 16 196 | MIRb | 3' |
| ITGB3 | NM_000212 | 4000 4163 | 41 207 | MIR3 | 3' |
| ITGB7 | NM_000889 | 26 135 | 133 10 | MIR | 5' |
| ITIH4 | NM_002218 | 3075 3141 | 186 120 | MIR3 | 3' |
| ITIH5 | NM_030569<br>NM_032817<br>NM_001001851 | 5054 5132<br>4394 4472<br>- | 140 64<br>140 64<br>- | MIR<br>MIR<br>- | 3'<br>3'<br>- |
| ITK | NM_005546 | 3130 3360 | 248 25 | MIRb | 3' |
| ITPK1 | NM_014216 | 2297 2400 | 91 195 | MIRm | 3' |
| JAK3 | NM_000215 | 3444 3527 | 49 139 | MIR | 3' |
| JARID1B | NM_006618 | 418 597 | 65 262 | MIR | 5' |
| JHDM1D | NM_030647 | 6188 6302 | 25 140 | MIR | 3' |
| KCNA6 | NM_002235 | 4039 4170 | 42 171 | MIR | 3' |
| KCNA7 | NM_031886 | 4037 4105<br>4161 4284 | 152 83<br>191 56 | MIRb<br>MIRm | 3'<br>3' |
| KCNE3 | NM_005472 | 1725 1860 | 58 200 | MIR3 | 3' |
| KCNG3 | NM_133329<br>NM_172344 | 3333 3398<br>3300 3365 | 177 110<br>177 110 | MIR3<br>MIR3 | 3'<br>3' |
| KCNG4 | NM_172347<br>NM_133490 | -<br>1174 1276<br>1287 1422 | -<br>60 180<br>270 113 | -<br>MIRb<br>MIRm | -<br>3'<br>3' |
| KCNH5 | NM_139318<br>NM_172376<br>NM_172375 | -<br>2152 2288<br>- | -<br>2 151<br>- | -<br>MIR<br>- | -<br>3'<br>- |
| KCNJ13 | NM_002242 | 2262 2330 | 124 197 | MIR3 | 3' |
| KCNK3 | NM_002246 | 2102 2267 | 261 99 | MIRb | 3' |
| KCNK6 | NM_004823 | 1421 1553<br>2066 2206<br>2525 2555 | 29 165<br>31 176<br>177 207 | MIR<br>MIR<br>MIR | 3'<br>3'<br>3' |
| KCNS1 | NM_005552<br>NM_002251 | -<br>232 394<br>2315 2417 | -<br>261 85<br>157 15 | -<br>MIRb<br>MIRb | -<br>5'<br>3' |
| KCTD17 | NM_024681 | 1398 1425 | 161 188 | MIRb | 3' |
| KHK | NM_000221<br>NM_006488 | 1727 1880<br>1727 1880 | 34 192<br>34 192 | MIRb<br>MIRb | 3'<br>3' |
| KIF1A | NM_004321 | 7821 7982 | 5 166 | MIR3 | 3' |
| KLC1 | NM_005552<br>NM_182923 | 1958 2011<br>- | 137 84<br>- | MIRb<br>- | TAG<br>- |
| KLF12 | NM_007249<br>NM_016285 | 7815 8049<br>- | 267 8<br>- | MIRb<br>- | 3'<br>- |
| KLF17 | NM_173484 | 2541 2627 | 37 121 | MIRb | 3' |
| KLF3 | NM_016531<br>BC051687 | -<br>1639 1707 | -<br>63 134 | -<br>MIRb | -<br>3' |

| | | | | | |
|---|---|---|---|---|---|
| KLHDC6 | NM_207335 | 2614 2832 | 36 261 | MIRb | 3’ |
| KLHL23 | NM_144711 | - | - | - | - |
| | BC016950 | 1358 1481 | 4 130 | MIR3 | 3’ |
| KLHL32 | NM_052904 | 2778 2916 | 52 186 | MIR | 3’ |
| KLHL6 | NM_130446 | 2056 2305 | 272 2 | MIRb | 3’ |
| KLK10 | NM_002776 | 1328 1543 | 22 255 | MIRb | 3’ |
| | NM_145888 | 1191 1406 | 22 255 | MIRb | 3’ |
| KLK15 | NM_023006 | 1036 1188 | 76 229 | MIR | 3’ |
| | NM_138563 | 1017 1169 | 76 229 | MIR | 3’ |
| | NM_138564 | 899 1051 | 76 229 | MIR | 3’ |
| | NM_017509 | 1154 1306 | 76 229 | MIR | 3’ |
| KLK6 | NM_002774 | 1306 1508 | 56 262 | MIR | 3’ |
| | NM_001012964 | 1322 1524 | 56 262 | MIR | 3’ |
| | NM_001012965 | 1274 1476 | 56 262 | MIR | 3’ |
| | NM_001012966 | 1165 1367 | 56 262 | MIR | 3’ |
| KRI1 | NM_023008 | 2551 2579 | 148 120 | MIRc | 3’ |
| KRT4 | NM_002272 | 2 77 | 100 25 | MIR | 5’ |
| KRT74 | NM_175053 | 2257 2490 | 268 2 | MIRb | 3’ |
| L3MBTL | NM_015478 | 2546 2760 | 4 224 | MIRb | 3’ |
| | NM_032107 | - | - | - | - |
| LAMP3 | NM_014398 | 2417 2640 | 274 7 | MIRb | 3’ |
| LAS1L | NM_031206 | 1116 1231 | 135 4 | MIR3 | CDS |
| LDLRAP1 | NM_015627 | 1903 2010 | 164 18 | MIRb | 3’ |
| LENG4 | NM_024298 | - | - | - | - |
| | BC006309 | 1544 1591 | 98 145 | MIR | 3’ |
| | | 1626 1799 | 28 216 | MIR | 3’ |
| LEP | NM_000230 | 2458 2595 | 177 33 | MIR3 | 3’ |
| LHFPL5 | NM_182548 | 1341 1429 | 96 186 | MIR3 | 3’ |
| LIG4 | NM_002312 | 109 244 | 230 93 | MIRb | 5’ |
| | NM_206937 | 66 123 | 145 87 | MIR | 5’ |
| LILRB2 | NM_005874 | - | - | - | - |
| | BC025766 | 1586 1728 | 35 188 | MIR3 | 3’ |
| | | 1905 2122 | 14 247 | MIRb | 3’ |
| LINGO1 | NM_032808 | 2681 2931 | 3 258 | MIRb | 3’ |
| LINS1 | NM_018148 | 2384 2466 | 69 154 | MIRb | CDS |
| | NM_001040614 | - | - | MIRb - | - |
| | NM_001040615 | - | - | - | - |
| | NM_001040616 | - | - | - | - |
| | AK001445 | 2397 2466 | 85 154 | MIRb | |
| LIPG | NM_006033 | 3014 3144 | 167 27 | MIRb | 3’ |
| LMBR1L | NM_018113 | 273 401 | 120 249 | MIR_Mars | CDS |
| LMOD1 | NM_012134 | 3685 3850 | 244 66 | MIRb | 3’ |
| LPPR2 | NM_022737 | 190 274 | 104 188 | MIRb | 5’ |
| LRAT | NM_004744 | 4040 4084 | 188 233 | MIR | 3’ |
| LRFN1 | XM_290842 | 98 156 | 141 83 | MIRb | 5’ |
| LRRC15 | NM_130830 | 4913 5009 | 263 162 | MIRb | 3’ |
| LRRC21 | NM_015613 | 2045 2100 | 262 207 | MIR | 3’ |
| LRRC25 | NM_145256 | 430 477 | 161 116 | MIR | 3’ |
| | | 1552 1754 | 28 236 | MIR3 | 3’ |
| LRRC32 | NM_005512 | 7 90 | 14 101 | MIRb | 5’ |
| | | 2107 2232 | 142 1 | MIR | 3’ |
| | | 2940 3158 | 232 13 | MIRb | 3’ |
| | | 3388 3488 | 186 82 | MIR3 | 3’ |
| | | 3511 3580 | 128 198 | MIR3 | 3’ |
| | | 3841 3924 | 152 47 | MIRm | 3’ |
| LRRC48 | NM_031294 | - | - | - | - |
| | AL136926 | 1001 1084 | 219 129 | MIRb | 3’ |

| | | | | | |
|---|---|---|---|---|---|
| | | 1385 1468 | 128 40 | MIRb | 3’ |
| | | 3823 4028 | 246 36 | MIR | 3’ |
| LRRC59 | NM_018509 | 1600 1748 | 196 35 | MIR3 | 3’ |
| LTBP2 | NM_000428 | 7283 7435 | 249 87 | MIR | 3’ |
| | AF113211 | 1 149 | 245 87 | MIR | 5’ |
| LUZP1 | NM_033631 | 4932 5110 | 20 201 | MIR3 | 3’ |
| | | 5263 5330 | 110 177 | MIR3 | 3’ |
| LVRN | NM_173800 | - | - | - | - |
| | BC070028 | 134 186 | 142 90 | MIR | 5’ |
| LYPD5 | NM_001031749 | 1219 1323 | 2 107 | MIR | 3’ |
| | | 1326 1405 | 27 107 | MIR | 3’ |
| | | 1408 1613 | 27 218 | MIR | 3’ |
| | NM_182573 | 1183 1287 | 2 107 | MIR | 3’ |
| | | 1290 1369 | 27 107 | MIR | 3’ |
| | | 1372 1600 | 27 255 | MIR | 3’ |
| LYPD6 | NM_194317 | 2512 2645 | 254 119 | MIRb | 3’ |
| LZIC | NM_032368 | 1148 1292 | 197 32 | MIR3 | 3’ |
| MACF1 | NM_012090 | 98 147 | 134 85 | MIR | 5’ |
| | NM_033044 | - | - | - | - |
| MAFF | NM_012323 | 1248 1349 | 134 38 | MIR | 3’ |
| | NM_152878 | 1198 1299 | 134 38 | MIR | 3’ |
| MAGIX | NM_001099680 | - | - | - | - |
| | NM_001099681 | - | - | - | - |
| | NM_001099682 | - | - | - | - |
| | NM_024859 | 2409 2531 | 155 17 | MIRb | TAG |
| MALT1 | NM_006785 | 4201 4347 | 152 4 | MIRb | 3’ |
| | NM_173844 | 4168 4314 | 152 4 | MIRb | 3’ |
| MAN2A2 | NM_006122 | 6056 6130 | 193 268 | THER1_MD | 3’ |
| MAOB | NM_000898 | 2103 2257 | 202 46 | MIR3 | 3’ |
| MAP2K3 | NM_002756 | - | - | - | - |
| | NM_145109 | - | - | - | - |
| | NM_145110 | 105 140 | 143 108 | MIR | 5’ |
| MAP3K14 | NM_003954 | 3861 3984 | 109 229 | MIR | 3’ |
| MAP3K8 | NM_005204 | 578 673 | 263 163 | MIR | 5’ |
| MAP4K5 | NM_006575 | 4164 4332 | 199 11 | MIRb | 3’ |
| | NM_198794 | 4212 4380 | 199 11 | MIRb | 3’ |
| MAPK4 | NM_002747 | 3097 3256 | 176 17 | MIR3 | 3’ |
| MAPKAPK2 | NM_004759 | 1472 1629 | 175 16 | MIR3 | 3’ |
| | NM_032960 | - | - | - | - |
| MARCH5 | NM_017824 | 3261 3438 | 25 197 | MIRb | 3’ |
| MARCH8 | NM_001002265 | 511 622 | 74 190 | MIR3 | 5’ |
| | NM_145021 | 511 622 | 74 190 | MIR3 | 5’ |
| | NM_001002266 | - | - | - | - |
| MARCH9 | NM_138396 | - | - | - | - |
| | BC009489 | 48 224 | 200 22 | MIR | 5’ |
| MARK4 | NM_031417 | 3060 3132 | 84 153 | MIRb | 3’ |
| MASP1 | NM_001879 | 2599 2825 | 14 254 | MIRb | 3’ |
| | | 3111 3244 | 2 138 | MIR3 | 3’ |
| | | 3970 4042 | 92 165 | MIR | 3’ |
| | NM_139125 | 2721 2791 | 87 157 | MIR3 | 3’ |
| | | 3657 3821 | 37 207 | MIR3 | 3’ |
| | NM_001031849 | - | - | - | - |
| MAST3 | XM_038150 | 5331 5414 | 46 140 | MIR | 3’ |
| | | 5429 5603 | 256 71 | MIRb | 3’ |
| | XM_938007 | 5328 5411 | 46 140 | MIR | 3’ |
| | | 5426 5600 | 256 71 | MIRb | 3’ |
| MAT1A | NM_000429 | 1944 2187 | 255 1 | MIR | 3’ |

| | | | | | |
|---|---|---|---|---|---|
| MATN1 | NM_002379 | - | - | - | - |
| | M55683 | 1579 1780 | 258 38 | MIRb | 3’ |
| MB | NM_005368 | - | - | - | - |
| | NM_203377 | 93 166 | 23 96 | MIR | 5’ |
| | NM_203378 | - | - | - | - |
| MBL2 | NM_000242 | 3241 3423 | 37 226 | MIR | 3’ |
| MC1R | NM_002386 | 1661 1737 | 82 172 | MIR3 | 3’ |
| MCCC2 | NM_022132 | 1987 2034 | 106 153 | MIR | 3’ |
| MCM10 | NM_182751 | 3130 3235 | 89 194 | MIR3 | 3’ |
| | NM_018518 | 3127 3232 | 89 194 | MIR3 | 3’ |
| MDGA1 | NM_153487 | 7447 7567 | 193 62 | MIR3 | 3’ |
| | | 8274 8356 | 187 93 | MIR3 | 3’ |
| MED25 | NM_030973 | - | - | - | - |
| | AF447873 | 586 676 | 240 158 | MIRb | 5’ |
| | | 1436 1479 | 136 93 | MIRm | 5’ |
| MED8 | NM_201542 | 982 1208 | 15 260 | MIRb | 3’ |
| | NM_052877 | - | - | - | - |
| | NM_001001651 | - | - | - | - |
| | NM_001001653 | 982 1208 | 15 260 | MIRb | 3’ |
| | NM_001001654 | 939 1165 | 15 260 | MIRb | 3’ |
| MEG3 | AF447875 | 4756 4882 | 121 254 | MIRb | 3’ |
| MEGF11 | NM_032445 | 5149 5276 | 84 205 | MIR3 | 3’ |
| MEGF9 | XM_929502 | - | - | - | - |
| | XM_933704 | 1597 1806 | 4 256 | MIR | 3’ |
| | XM_941600 | - | - | - | - |
| | XM_945294 | 1597 1806 | 4 256 | MIR | 3’ |
| MESDC2 | NM_015154 | - | - | - | - |
| | BC009210 | 1813 1912 | 96 202 | MIR3 | 3’ |
| | | 1854 2012 | 187 22 | MIR3 | 3’ |
| | | 2389 2558 | 93 272 | MIRb | 3’ |
| | | 2678 2795 | 38 178 | MIRm | 3’ |
| METTL13 | NM_001007239 | 2746 2830 | 262 182 | MIRb | 3’ |
| | NM_014955 | 2889 2973 | 262 182 | MIRb | 3’ |
| | NM_015935 | 3214 3298 | 262 182 | MIRb | 3’ |
| MFAP3 | NM_005927 | 1900 1968 | 111 185 | MIR3 | 3’ |
| MFAP5 | NM_003480 | 1744 1880 | 20 156 | MIR3 | 3’ |
| MFSD2 | NM_032793 | - | - | - | - |
| | AF370364 | 145 273 | 134 8 | MIR | 5’ |
| MFSD4 | NM_181644 | - | - | - | - |
| | AY358107 | 2853 3106 | 7 262 | MIR | 3’ |
| MGAT5 | NM_002410 | 1245 1458 | 9 233 | MIR | 5’ |
| MICA | NM_000247 | 1186 1360 | 24 223 | MIR | TAG |
| MICAL2 | NM_014632 | - | - | - | - |
| | BX538021 | 1 187 | 36 262 | MIRb | 5’ |
| MIF4GD | NM_020679 | 1136 1205 | 70 145 | MIR | 3’ |
| MIPOL1 | NM_138731 | 4121 4203 | 261 177 | MIR_Mars | 3’ |
| | | 4277 4387 | 165 26 | MIRb | 3’ |
| MIR16 | NM_016641 | 1854 1966 | 198 85 | MIR3 | 3’ |
| | | 2257 2506 | 258 4 | MIRb | 3’ |
| MLLT6 | NM_005937 | - | - | - | - |
| | BC064612 | 2072 2269 | 197 1 | MIR | 3’ |
| MLXIPL | NM_032951 | 2767 2938 | 33 191 | MIRm | 3’ |
| | NM_032952 | 2710 2881 | 33 191 | MIRm | 3’ |
| | NM_032953 | 2761 2932 | 33 191 | MIRm | 3’ |
| | NM_032954 | 2704 2875 | 33 191 | MIRm | 3’ |
| MMP19 | NM_002429 | 2019 2250 | 1 264 | MIRb | 3’ |

| | | | | | |
|---|---|---|---|---|---|
| | NM_001032360 | 1672 1903 | 1 264 | MIRb | 3’ |
| MMRN1 | NM_007351 | 4538 4776 | 266 3 | MIRb | 3’ |
| MOBKL2C | NM_145279 | 2171 2322 | 193 39 | MIR3 | 3’ |
| | NM_201403 | 2123 2283 | 202 39 | MIR3 | 3’ |
| MON1B | NM_014940 | 2649 2740 | 108 197 | MIRb | 3’ |
| MORF4L2 | NM_012286 | - | - | - | - |
| | BC056899 | 110 164 | 164 231 | MIR | 5’ |
| MPL | NM_005373 | 3082 3305 | 12 249 | MIR | 3’ |
| MPO | NM_000250 | 2693 2901 | 2 228 | MIR | 3’ |
| MPP3 | NM_001932 | 10 147 | 124 265 | MIRb | 5’ |
| | | 2099 2203 | 198 93 | MIR3 | 3’ |
| MRAP2 | NM_138409 | 1348 1485 | 53 207 | MIR3 | 3’ |
| MRI1 | NM_001031727 | 1310 1441 | 21 151 | MIR | 3’ |
| | | 1770 1804 | 138 172 | MIR | 3’ |
| | NM_032285 | 1169 1300 | 21 151 | MIR | 3’ |
| | | 1629 1663 | 138 172 | MIR | 3’ |
| MRM1 | NM_024864 | 1776 1816 | 162 202 | THER1_MD | 3’ |
| MRO | NM_031939 | 1629 1877 | 262 2 | MIR | 3’ |
| MRPL10 | NM_145255 | 1218 1406 | 247 27 | MIRb | 3’ |
| | NM_148887 | 1316 1504 | 247 27 | MIRb | 3’ |
| MRPL27 | NM_016504 | - | - | - | - |
| | NM_148571 | 340 445 | 266 161 | MIRb | 3’ |
| | | 1059 1287 | 262 15 | MIR | 3’ |
| | | 1401 1503 | 104 213 | MIR | 3’ |
| | NM_148570 | - | - | - | - |
| MRPL42 | NM_014050 | 1279 1438 | 220 9 | MIRb | 3’ |
| | NM_172177 | 1282 1441 | 220 9 | MIRb | 3’ |
| | NM_172178 | 1302 1461 | 220 9 | MIRb | 3’ |
| MRPL49 | NM_004927 | 1727 1841 | 30 155 | MIRb | 3’ |
| MRRF | NM_138777 | - | - | - | - |
| | NM_199177 | - | - | - | - |
| | BC002814 | 677 798 | 26 145 | MIR | 3’ |
| MRS2L | NM_020662 | - | - | - | - |
| | BC069009 | 1337 1545 | 2 230 | MIR | 3’ |
| MS4A10 | NM_206893 | 817 894 | 128 49 | MIRb | TAG |
| | | 2097 2241 | 6 145 | MIR3 | 3’ |
| MS4A3 | NM_006138 | 1440 1653 | 1 219 | MIRb | 3’ |
| | NM_001031809 | 1302 1515 | 1 219 | MIRb | 3’ |
| | NM_001031666 | 1131 1344 | 1 219 | MIRb | 3’ |
| MS4A6A | NM_152851 | - | - | - | - |
| | NM_022349 | 1109 1240 | 167 35 | MIR3 | 3’ |
| | NM_152852 | - | - | - | - |
| MSR1 | NM_138715 | 2910 2973 | 2 66 | MIR | 3’ |
| | NM_002445 | - | - | - | - |
| | NM_138716 | 2721 2784 | 2 66 | MIR | 3’ |
| MST150 | NM_032947 | 5 142 | 191 38 | THER1_MD | 5’ |
| | | 180 340 | 183 22 | MIR3 | ATG |
| | | 946 1127 | 34 212 | MIRb | 3’ |
| MSX1 | NM_002448 | 1319 1466 | 182 41 | MIR3 | 3’ |
| MTIF2 | NM_001005369 | 375 529 | 3 160 | MIR3 | 5’ |
| | NM_002453 | 92 246 | 3 160 | MIR3 | 5’ |
| MUC15 | NM_145650 | 199 283 | 99 178 | MIRb | 3’ |
| MUC3B | XM_168578 | 2233 2367 | 43 193 | MIRb | 3’ |
| | XM_374502 | 1114 1242 | 43 187 | MIRb | 3’ |
| | XM_940796 | - | - | - | - |
| MYCBP | NM_012333 | 1387 1551 | 78 241 | MIRb | 3’ |
| MYEOV | NM_138768 | 1372 1532 | 261 87 | MIR | CDS |

| | | | | | |
|---|---|---|---|---|---|
| | | 1926 2022 | 230 110 | MIRb | 3’ |
| MYO15A | NM_016239 | 7364 7465<br>10947 11025 | 188 82<br>55 139 | MIR<br>MIR | CDS<br>3’ |
| MYO5A | NM_000259 | 7666 7742<br>9821 9950 | 115 193<br>221 84 | MIR3<br>MIRc | 3’<br>3’ |
| MYO7A | NM_000260<br>HSU55209 | -<br>3829 3958 | -<br>29 159 | -<br>MIRb | -<br>CDS |
| MYOZ3 | NM_133371 | 2394 2564<br>2661 2866 | 250 73<br>28 262 | MIRb<br>MIRb | 3’<br>3’ |
| MYRIP | NM_015460 | 3996 4165 | 197 2 | MIRb | 3’ |
| MYSM1 | XM_055481<br>XM_941796<br>BX537912 | -<br>-<br>2208 2361<br>2517 2713 | -<br>-<br>20 200<br>7 205 | -<br>-<br>MIR3<br>MIR3 | -<br>-<br>3’<br>3’ |
| MYT1L | NM_015025 | 666 798 | 144 5 | MIRm | 5’ |
| NAALAD2 | NM_005467<br>BC038840 | -<br>1333 1503 | -<br>70 251 | -<br>MIR | -<br>3’ |
| NAIF1 | NM_197956<br>BC021580 | -<br>1077 1329 | -<br>2 261 | -<br>MIRb | -<br>3’ |
| NAPE-PLD | NM_198990<br>BC071604 | -<br>2881 3003 | -<br>82 216 | -<br>MIR | -<br>3’ |
| NAT9 | NM_015654 | 1725 1786 | 149 86 | MIRm | 3’ |
| NCAN | NM_004386 | 5366 5620 | 262 3 | MIRb | 3’ |
| NCOA6 | NM_014071<br>AF208227 | -<br>1740 1909 | -<br>64 258 | -<br>MIR | -<br>5’ |
| NCR2 | NM_004828 | 798 914 | 95 219 | MIRb | TAG |
| NDST1 | NM_001543 | 4712 4858<br>7175 7293 | 144 3<br>23 153 | MIR3<br>MIR | 3’<br>3’ |
| NDST2 | NM_003635 | 121 219 | 5 102 | MIR | 5’ |
| NDUFAF1 | NM_016013<br>BC000780 | -<br>166 313 | -<br>66 235 | -<br>MIR | -<br>5’ |
| NDUFB2 | NM_004546<br>BC063026 | -<br>135 181 | -<br>126 83 | -<br>MIR | -<br>5’ |
| NDUFS1 | NM_005006.5 | 2781 3023 | 262 5 | MIRb | 3’ |
| NEDD4 | NM_006154<br>NM_198400 | 4603 4810<br>5864 6008 | 32 243<br>32 186 | MIRb<br>MIRb | 3’<br>3’ |
| NEK3 | NM_002498<br>NM_152720 | 2040 2211<br>1963 2134 | 15 181<br>15 181 | MIRb<br>MIRb | 3’<br>3’ |
| NEK6 | NM_014397<br>BC012761 | -<br>1874 2071 | -<br>62 259 | -<br>MIRb | -<br>3’ |
| NETO1 | NM_138999<br>NM_153181<br>NM_138966<br>BC050329 | -<br>-<br>-<br>1901 1977<br>2103 2172 | -<br>-<br>-<br>141 56<br>130 195 | -<br>-<br>-<br>MIRm<br>MIRb | -<br>-<br>-<br>3’<br>3’ |
| NEU3 | NM_006656 | 2 2478 2627 | 16 167 | MIR3 | 3’ |
| NEURL | NM_004210 | 3520 3639 | 18 150 | MIRb | 3’ |
| NEUROD2 | NM_006160<br>AB021742 | -<br>1 117 | -<br>51 184 | -<br>MIR3 | -<br>5’ |
| NEUROD4 | NM_021191<br>BC040961 | -<br>1109 1295 | -<br>41 199 | -<br>MIR3 | -<br>3’ |
| NFATC4 | NM_004554 | 3702 3841 | 156 5 | MIR | 3' |
| NFXL1 | **NM_152995**<br>**BC051193** | -<br>2106 2220 | -<br>135 14 | -<br>MIR3 | -<br>CDS |
| NGFR | NM_002507 | 1915 2061 | 201 52 | MIR3 | 3’ |

| | | | | | |
|---|---|---|---|---|---|
| NHLRC1 | NM_198586 | 1202 1285 | 28 118 | MIRb | 3’ |
| | | 1352 1452 | 264 168 | MIRb | 3’ |
| | | 1721 1856 | 167 38 | MIRb | 3’ |
| NHP2L1 | NM_005008 | 860 934 | 17 85 | MIR | 3’ |
| | NM_001003796 | 779 853 | 17 85 | MIR | 3’ |
| NIP30 | NM_024946 | - | - | - | - |
| | BC063409 | 1927 2073 | 201 52 | MIR3 | 5’ |
| NIP7 | NM_016101 | 1210 1390 | 47 206 | MIRb | 3’ |
| | | 1673 1746 | 207 271 | THER1_MD | 3’ |
| NIPA1 | NM_144599 | - | - | - | - |
| | BX537997 | 5479 5616 | 4 152 | MIR | 3’ |
| NIPSNAP1 | NM_003634 | 1 98 | 154 44 | MIR3 | 5’ |
| NIT1 | NM_005600 | - | - | - | - |
| | BC046149 | 1742 1887 | 59 201 | MIRb | 3’ |
| NKIRAS2 | NM_001001349 | 2189 2301 | 70 185 | MIR3 | 3’ |
| | NM_017595 | 1713 1825 | 70 185 | MIR3 | 3’ |
| NMNAT2 | NM_015039 | 3198 3395 | 200 6 | MIR3 | 3’ |
| | NM_170706 | 2983 3180 | 200 6 | MIR3 | 3’ |
| NMUR1 | NM_006056 | 1817 1894 | 35 112 | MIR | 3’ |
| | | 2192 2287 | 113 207 | MIR3 | 3’ |
| NOD2 | NM_022162 | 4277 4336 | 191 251 | MIRc | 3’ |
| NOTCH2NL | NM_203458 | - | - | - | - |
| | BX537434 | 2854 3100 | 262 8 | MIRb | 3’ |
| | | 3672 3773 | 202 97 | MIRb | 3’ |
| NOV | NM_002514 | 1947 1990 | 151 112 | MIR | 3’ |
| NPAL3 | NM_020448 | - | - | - | - |
| | BX640883 | 444 561 | 183 54 | MIR | 3’ |
| NPCDR1 | XR_001027 | 1416 1566 | 199 36 | MIR | ATG |
| | AF134979 | 1644 1794 | 199 36 | MIR | ATG |
| NPHP1 | NM_000272 | 2962 3144 | 262 56 | MIR | 3’ |
| | NM_207181 | - | - | - | - |
| NPTXR | NM_014293 | 2581 2720 | 3 152 | MIR3 | 3’ |
| | | 3427 3550 | 141 6 | MIRm | 3’ |
| | | 4550 4764 | 232 14 | MIRb | 3’ |
| | | 5220 5359 | 179 19 | MIR3 | 3’ |
| | NM_058178 | 2578 2717 | 3 152 | MIR3 | 3’ |
| | | 3424 3547 | 141 6 | MIRm | 3’ |
| | | 4547 4761 | 232 14 | MIRb | 3’ |
| | | 5217 5356 | 179 19 | MIR3 | 3’ |
| NR2E3 | NM_016346 | - | - | - | - |
| | NM_014249 | - | - | - | - |
| | BC041421 | 273 373 | 141 25 | MIRb | 5’ |
| NR4A3 | NM_006981 | 4596 4774 | 208 19 | MIR3 | 3’ |
| | NM_173198 | 5343 5521 | 208 19 | MIR3 | 3’ |
| | NM_173200 | 3944 4122 | 208 19 | MIR3 | 3’ |
| | NM_173199 | 2437 2527 | 14 98 | MIRb | 3’ |
| NRIP3 | NM_020645 | 1300 1443 | 176 5 | MIR3 | 3‘ |
| NRL | NM_006177 | 1629 1791 | 20 190 | MIR3 | 3’ |
| | BC012395 | - | - | - | - |
| NRSN1 | NM_080723 | 1770 1924 | 37 195 | MIR | 3’ |
| NRSN2 | NM_024958 | 11 148 | 141 8 | MIR | 5’ |
| | | 1019 1096 | 82 160 | MIR3 | 3’ |
| | | 1244 1417 | 210 29 | MIR_Mars | 3’ |
| NSAP11 | XM_001722999 | - | - | - | - |
| | AY176665 | 1799 2009 | 17 242 | MIRc | 3’ |
| NT5E | NM_002526 | 2093 2329 | 255 8 | MIRb | 3’ |

| | | | | | |
|---|---|---|---|---|---|
| NTRK3 | NM_001012338 | - | - | - | - |
| | NM_002530 | - | - | - | - |
| | NM_001007156 | 1758 1883 | 207 82 | MIR3 | CDS |
| | BT007291 | 1596 1721 | 207 82 | MIR3 | CDS |
| NTSR1 | NM_002531 | 3061 3147 | 86 162 | MIR | 3’ |
| | | 3528 3660 | 24 173 | MIRb | 3’ |
| NTSR2 | NM_012344 | 1550 1694 | 101 263 | THER1_MD | 3’ |
| NUBPL | NM_025152 | 1579 1745 | 162 8 | MIR | 3’ |
| NUCB2 | NM_005013 | 114 211 | 58 168 | MIRb | 5’ |
| NUDT12 | NM_031438 | 2657 2780 | 190 75 | MIR3 | 3’ |
| NUMA1 | NM_006185 | 184 329 | 2 172 | MIR | CDS |
| | | 1294 1467 | 175 2 | MIRb | 3’ |
| NUMB | NM_001005743 | 206 306 | 146 48 | MIRb | 5’ |
| | NM_001005744 | 206 306 | 146 48 | MIRb | 5’ |
| | NM_003744 | 206 306 | 146 48 | MIRb | 5’ |
| | NM_001005745 | 206 306 | 146 48 | MIRb | 5’ |
| NUP133 | NM_018230 | 3831 3941 | 49 176 | MIR3 | 3’ |
| NXF4 | NR_002216 | 2140 2326 | 17 222 | MIR | ncRNA |
| OAS1 | NM_016816 | 1328 1471 | 179 21 | MIR3 | 3’ |
| | NM_002534 | 1388 1470 | 256 176 | MIRb | 3’ |
| | NM_001032409 | 1230 1373 | 179 21 | MIR3 | 3’ |
| OCLN | NM_002538 | 2340 2430 | 210 122 | MIRb | 3’ |
| OLFML2A | NM_182487 | 1655 1857 | 249 38 | MIRb | 3’ |
| | | 2888 2999 | 187 63 | MIR3 | 3’ |
| | | 3998 4019 | 168 146 | MIR | 3’ |
| | | 4333 4426 | 153 53 | MIRb | 3’ |
| | | 4441 4581 | 246 87 | THER1_MD | 3’ |
| | | 5348 5460 | 2 118 | MIR3 | 3’ |
| | | 5526 5676 | 203 47 | MIRb | 3’ |
| OPHN1 | NM_002547 | 4206 4352 | 180 10 | MIR3 | 3’ |
| | | 5827 5962 | 145 5 | MIR3 | 3’ |
| OR4D2 | NM_001004707 | 1550 1665 | 22 137 | MIRb | 3’ |
| OR51E2 | NM_030774 | 2189 2434 | 262 2 | MIR | 3’ |
| ORAOV1 | NM_153451 | 999 1194 | 257 43 | MIRb | 3’ |
| ORC2L | NM_006190 | - | - | - | - |
| | AF315716 | 115 233 | 189 58 | MIR3 | ATG |
| OSCAR | NM_206818 | 1580 1646 | 12 84 | MIR | 3’ |
| | NM_206817 | 1544 1610 | 12 84 | MIR | 3’ |
| | NM_130771 | 1066 1132 | 12 84 | MIR | 3’ |
| | NM_133169 | 1054 1120 | 12 84 | MIR | 3’ |
| | NM_133168 | 1021 1087 | 12 84 | MIR | 3’ |
| OSM | NM_020530 | 1321 1464 | 3 154 | MIR3 | 3’ |
| OTUB2 | NM_023112 | 1090 1172 | 61 150 | MIRb | 3’ |
| | | 3361 3557 | 34 235 | MIR | 3’ |
| | AF318378 | 2025 2221 | 34 235 | MIRb | TAG |
| P2RX7 | NM_002562 | 2360 2564 | 27 263 | MIR | 3’ |
| | NM_177427 | 2263 2467 | 27 263 | MIR | 3’ |
| P2RY2 | NM_176072 | 1918 2005 | 156 64 | MIR3 | 3’ |
| | NM_002564 | 1784 1871 | 156 64 | MIR3 | 3’ |
| | NM_176071 | 1850 1937 | 156 64 | MIR3 | 3’ |
| P2RY6 | NM_176797 | - | - | - | - |
| | NM_176798 | - | - | - | - |
| | NM_176796 | - | - | - | - |
| | NM_004154 | 21 126 | 33 142 | MIRb | 5’ |
| P53AIP1 | NM_022112 | 1833 2077 | 9 262 | MIR | 3’ |
| PALM2 | NM_053016 | 6011 6071 | 184 124 | MIR3 | 3’ |
| | NM_001037293 | - | - | - | - |

| | | | | | |
|---|---|---|---|---|---|
| PAPLN | NM_173462 | 5376 5473 | 104 201 | MIR3 | 3’ |
| PAPPA | NM_002581 | 9499 9598 | 83 169 | MIR | 3’ |
| | AY189937 | 74 173 | 83 169 | MIR | TAG |
| PAPSS2 | NM_004670 | 2988 3051 | 185 122 | MIR3 | 3’ |
| | | 3126 3184 | 78 17 | MIR | 3’ |
| | NM_001015880 | 3003 3066 | 185 122 | MIR3 | 3’ |
| | | 3141 3199 | 78 17 | MIR | 3’ |
| PAQR3 | NM_001040202 | - | - | - | - |
| | BC031256 | 1040 1126 | 15 112 | MIR | TAG |
| PAQR5 | NM_017705 | - | - | - | - |
| | AY424283 | 10 94 | 45 136 | MIR | 5’ |
| PARD3B | NM_205863 | - | - | - | - |
| | NM_152526 | - | - | - | - |
| | NM_057177 | - | - | - | - |
| | AB053321 | 936 1189 | 271 1 | MIR | 3’ |
| PARS2 | NM_152268 | 1773 1893 | 31 169 | MIR | 3’ |
| | | 1981 2127 | 7 173 | MIR3 | 3’ |
| PARVA | NM_018222 | 1541 1707 | 260 95 | MIRb | 3’ |
| PAX5 | NM_016734 | 2470 2515 | 192 147 | MIR3 | 3’ |
| | | 3233 3366 | 2 130 | MIR3 | 3’ |
| PAX7 | NM_002584 | 2105 2214 | 155 29 | MIR | TAG |
| | NM_013945 | 2099 2208 | 155 29 | MIR | TAG |
| PBOV1 | NM_021635 | 123 245 | 168 5 | MIR | CDS |
| PCBP4 | NM_020418 | 32 171 | 219 80 | MIRb | 5’ |
| | NM_033009 | 32 171 | 219 80 | MIRb | 5’ |
| | NM_033008 | - | - | - | - |
| | NM_033010 | - | - | - | - |
| PCDH12 | NM_016580 | 197 421 | 234 14 | MIR | 5’ |
| PCDH21 | NM_033100 | 3404 3657 | 2 258 | MIRb | 3’ |
| | | 3867 3955 | 154 66 | MIR3 | 3’ |
| PCDHB12 | NM_018932 | 2994 3188 | 206 4 | MIRb | 3’ |
| PCDHB7 | NM_018940 | 2949 3150 | 216 1 | MIR | 3’ |
| PCGF3 | NM_006315 | - | - | - | - |
| | BX640894 | 646 866 | 254 7 | MIRb | 5’ |
| | | 889 1042 | 21 177 | MIR | 5’ |
| PCTK3 | NM_212503 | - | - | - | - |
| | NM_212502 | - | - | - | - |
| | NM_002596 | - | - | - | - |
| | BC040529 | 1046 1252 | 31 258 | MIR | TAG |
| PDCD2 | NM_002598 | - | - | - | - |
| | NM_144781 | 1204 1367 | 261 72 | MIR | 3’ |
| PDE11A | NM_016953 | 4250 4374 | 25 158 | MIRb | 3’ |
| PDE2A | NM_002599 | 3071 3201 | 51 204 | MIRb | 3’ |
| | | 3601 3802 | 3 208 | MIR3 | 3’ |
| PDE4A | NM_006202 | - | - | - | - |
| | U18088 | 50 251 | 7 262 | MIR3 | 5’ |
| PDE4C | NM_000923 | - | - | - | - |
| | U66347 | 1091 1216 | 182 41 | MIR | 5’ |
| PDGFRB | NM_002609 | 3955 4120 | 30 189 | MIR3 | 3’ |
| | | 4657 4792 | 57 194 | MIR3 | 3’ |
| PDLIM2 | NM_176871 | 3234 3327 | 187 73 | MIR3 | 3’ |
| | | 3329 3533 | 261 43 | MIR | 3’ |
| | NM_021630 | - | - | - | - |
| | NM_198042 | - | - | - | - |
| PDSS2 | NM_020381 | 3400 3476 | 80 159 | MIRm | 3’ |
| PELI3 | NM_145065 | 270 335 | 141 77 | MIRb | CDS |
| | AF487456 | 168 233 | 141 77 | MIRb | CDS |

| | | | | | |
|---|---|---|---|---|---|
| PGGT1B | NM_005023 | - | - | - | - |
| | L25441 | 145 207 | 195 261 | MIR | 5' |
| | | 211 328 | 57 171 | MIR3 | ATG |
| PHF12 | NM_001033561 | - | - | - | - |
| | NM_020889 | 2728 2928 | 211 3 | MIR3 | 3' |
| PHF19 | NM_015651 | - | - | - | - |
| | NM_001009936 | 825 922 | 34 154 | MIRb | TAG |
| | BC022374 | 685 782 | 34 154 | MIRb | TAG |
| PHLDB3 | NM_198850 | 1304 1512 | 262 59 | MIR | 3' |
| PIGL | NM_004278 | 979 1074 | 151 233 | THER1_MD | 3' |
| PIGS | NM_033198 | 2277 2499 | 25 260 | MIRb | 3' |
| PIGZ | NM_025163 | 2339 2562 | 5 230 | MIR3 | 3' |
| PIK3AP1 | NM_152309 | 3065 3205 | 183 31 | MIR3 | 3' |
| PIK3IP1 | NM_052880 | 1986 2100 | 132 7 | MIR | 3' |
| PIP5KL1 | NM_173492 | 84 167 | 100 17 | MIRb | 3' |
| PKD1L1 | NM_138295 | 8524 8760 | 261 11 | MIRb | 3' |
| PLA2G3 | NM_015715 | 2240 2404 | 51 231 | MIR | 3' |
| | | 2421 2571 | 62 209 | MIR | 3' |
| PLA2G4B | NM_005090 | - | - | - | - |
| | BC025290 | 1203 1374 | 77 260 | MIRb | 3' |
| PLAGL1 | NM_002656 | - | - | - | - |
| | NM_006718 | - | - | - | - |
| | BX537397 | 1228 1410 | 186 2 | MIRb | 5' |
| PLAUR | NM_002659 | 1295 1417 | 61 203 | MIRb | 3' |
| | NM_001005376 | - | - | - | - |
| | NM_001005377 | 1160 1282 | 61 203 | MIRb | 3' |
| PLCB2 | NM_004573 | 3984 4105 | 15 168 | MIR3 | 3' |
| PLCB4 | NM_000933 | 5404 5459 | 111 167 | MIR3 | 3' |
| | NM_182797 | - | - | - | - |
| PLD1 | NM_002662 | 5088 5204 | 171 56 | MIR3 | 3' |
| PLEK | NM_002664 | 1338 1442 | 75 183 | MIR3 | 3' |
| PLSCR1 | NM_021105 | - | - | - | - |
| | BC017901 | 670 823 | 259 93 | MIR | 3' |
| PLXNA2 | NM_025179 | - | - | - | - |
| | AY358496 | 361 465 | 13 118 | MIR3 | 5' |
| Plxnb2 (mouse) | XM_484491 | 125 228 | 161 59 | MIR | |
| PML | NM_033238 | 3800 3945 | 58 205 | MIR3 | 3' |
| | | 5318 5542 | 33 261 | MIR_Mars | 3' |
| | NM_033240 | - | - | - | - |
| | NM_033242 | - | - | - | - |
| | NM_033244 | - | - | - | - |
| | NM_002675 | - | - | - | - |
| | NM_033246 | - | - | - | - |
| | NM_033247 | 1729 1775 | 221 268 | MIRm | 3' |
| | NM_033239 | - | - | - | - |
| | NM_033249 | - | - | - | - |
| | NM_033250 | - | - | - | - |
| | NM_033245 | - | - | - | - |
| PNKD | NM_015488 | - | - | - | - |
| | NM_022572 | - | - | - | - |
| | BC036457 | 2947 2993 | 20 253 | MIR | 3' |
| PNMA2 | NM_007257 | 3206 3440 | 261 3 | MIRb | 3' |
| | | 4162 4332 | 85 262 | MIRb | 3' |
| PNPLA1 | NM_173676 | 1578 1810 | 5 255 | MIRb | 3' |
| | NM_001039725 | - | - | - | - |
| PODXL | NM_001018111 | 2187 2354 | 177 15 | MIR3 | 3' |
| | | 5212 5372 | 195 38 | MIR3 | 3' |

| | | | | | |
|---|---|---|---|---|---|
| | NM_005397 | 2091 2258<br>5116 5276 | 177 15<br>195 38 | MIR3<br>MIR3 | 3'<br>3' |
| POFUT1 | NM_015352<br><br>NM_172236 | 3124 3260<br>3323 3535<br>1321 1465 | 86 233<br>232 27<br>209 67 | MIR_Mars<br>MIRb<br>MIRb | 3'<br>3'<br>3' |
| POLG | NM_002693 | 133 299 | 172 1 | MIR3 | 5' |
| POLH | NM_006502 | 2617 2796 | 2 210 | MIRb | 3' |
| POLQ | NM_199420<br>AF090919 | -<br>1961 2125 | -<br>5 174 | -<br>MIR | -<br>3' |
| POLR3D | NM_001722 | 1682 1910 | 13 254 | MIRb | 3' |
| POPDC2 | NM_022135 | 1459 1634 | 70 253 | MIRb | 3' |
| PPA2 | NM_176869<br>NM_006903<br>NM_176866<br>NM_176867<br>NM_001034191<br>NM_001034192<br>NM_001034193<br>BC008246 | -<br>-<br>-<br>-<br>-<br>-<br>-<br>993 1172 | -<br>-<br>-<br>-<br>-<br>-<br>-<br>28 236 | -<br>-<br>-<br>-<br>-<br>-<br>-<br>MIRb | -<br>-<br>-<br>-<br>-<br>-<br>-<br>3' |
| PPBPL2 | XM_926381 | 823 942 | 246 126 | MIRb | TAG |
| PPP1R12B | NM_002481<br>NM_032105<br>NM_032103<br>NM_032104 | 7940 8086<br>8121 8267<br>-<br>- | 33 207<br>33 207<br>-<br>- | MIR3<br>MIR3<br>-<br>- | 3'<br>3'<br>-<br>- |
| PPP1R3E | AK024489<br><br>BX248758 | 1908 2082<br>3579 3746<br>432 556<br>1385 1559 | 184 3<br>197 30<br>70 198<br>184 3 | MIR<br>MIR3<br>MIR3<br>MIR | 3'<br>3'<br>5'<br>3' |
| PRCD | -<br>AK054729<br>AK125617 | -<br>560 638<br>1078 1314<br>2261 2477 | -<br>20 96<br>6 256<br>266 31 | -<br>MIR MIR<br>MIRb | -<br>U<br>U<br>U |
| PRELID2 | NM_182960<br><br>NM_138492<br><br>NM_205846 | 1309 1437<br>1441 1679<br>1508 1636<br>1640 1878<br>1273 1401<br>1405 1643 | 72 201<br>30 267<br>72 201<br>30 267<br>72 201<br>30 267 | MIRb<br>MIRb<br>MIRb<br>MIRb<br>MIRb<br>MIRb | 3'<br>3'<br>3'<br>3'<br>3'<br>3' |
| PRELP | NM_002725<br><br>NM_201348 | 2656 2701<br>3005 3039<br>2646 2691<br>2995 3029 | 120 161<br>162 192<br>120 161<br>162 192 | MIR3<br>MIR3<br>MIR3<br>MIR3 | 3'<br>3'<br>3'<br>3' |
| PRKACB | NM_182948<br><br>NM_002731<br><br>NM_207578 | 3846 3930<br>4266 4319<br>3876 3960<br>4296 4349<br>- | 113 197<br>179 231<br>113 197<br>179 231<br>- | MIR3<br>MIR_Mars<br>MIR3<br>MIR_Mars<br>- | 3'<br>3'<br>3'<br>3'<br>- |
| PRRX1 | NM_006902<br>NM_022716 | 2391 2503<br>2319 2431 | 66 182<br>66 182 | MIR<br>MIR | 3'<br>3' |
| PRX | NM_020956<br>NM_181882 | 42 167<br>- | 37 153<br>- | MIR<br>- | 3'<br>- |
| PSD4 | NM_012455 | 1267 1374<br>4781 4841 | 7 124<br>202 260 | MIRm<br>MIRm | CDS<br>3' |
| PSEN2 | NM_000447<br>NM_012486 | 55 165<br>- | 189 78<br>- | MIR3<br>- | 5'<br>- |
| PSMD12 | NM_002816<br>NM_174871 | -<br>- | -<br>- | -<br>- | -<br>- |

| | | | | | |
|---|---|---|---|---|---|
| | BC065826 | 143 292 | 207 47 | MIR | 5’ |
| PSMD6 | NM_014814 | - | - | - | - |
| | AY359879 | 8 126 | 50 169 | MIR | 5’ |
| PSMF1 | NM_006814 | 1717 1831 | 6 127 | MIR | 3’ |
| | NM_178578 | 1680 1794 | 6 127 | MIR | 3’ |
| | NM_178579 | 1634 1748 | 6 127 | MIR | 3’ |
| PSTPIP2 | NM_024430 | 1432 1566 | 182 46 | MIR3 | 3’ |
| PTCD2 | NM_024754 | 1719 1788 | 80 2 | MIRm | 3’ |
| | | 1844 1985 | 88 252 | MIRb | 3’ |
| PTGIS | NM_000961 | 3168 3341 | 264 71 | MIRb | 3’ |
| | | 5362 5454 | 184 93 | MIRm | 3’ |
| PTHLH | NM_198965 | - | - | - | - |
| | NM_002820 | 1659 1867 | 35 259 | MIRb | 3’ |
| | NM_198964 | 1640 1848 | 35 259 | MIRb | 3’ |
| | NM_198966 | - | - | - | - |
| PTHR1 | NM_000316 | 94 156 | 138 78 | MIR | 5’ |
| PTK7 | NM_002821 | - | - | - | - |
| | NM_152880 | - | - | - | - |
| | NM_152881 | - | - | - | - |
| | NM_152882 | - | - | - | - |
| | NM_152883 | - | - | - | - |
| | BC046109 | 1623 1813 | 15 210 | MIRb | 3’ |
| PTPN3 | NM_002829 | - | - | - | - |
| | BC063287 | 2433 2634 | 37 258 | MIR_Mars | 3’ |
| PTPRT | NM_133170 | 5811 5964 | 256 115 | MIR | 3’ |
| | | 7701 7791 | 150 61 | MIRm | 3’ |
| | | 8241 8445 | 229 21 | MIRb | 3’ |
| | NM_007050 | 5868 6021 | 256 115 | MIR | 3’ |
| | | 7758 7848 | 150 61 | MIRm | 3’ |
| | | 8298 8502 | 229 21 | MIRb | 3’ |
| PTPRU | NM_133178 | 5440 5586 | 100 249 | MIR_Mars | 3’ |
| | NM_133177 | 5451 5597 | 100 249 | MIR_Mars | 3’ |
| | NM_005704 | 5470 5616 | 100 249 | MIR_Mars | 3’ |
| PUS7L | NM_031292 | 2233 2401 | 254 25 | MIRb | 3’ |
| | | 2832 2916 | 32 122 | MIRb | 3’ |
| | | 3563 3806 | 7 251 | MIR | 3’ |
| PVRL1 | NM_002855 | - | - | - | - |
| | NM_203285 | - | - | - | - |
| | NM_203286 | 1311 1394 | 137 223 | MIRb | 3’ |
| PVT1 | XM_928984 | 716 850 | 155 21 | MIRb | 3’ |
| PXMP4 | NM_007238 | 928 1006 | 42 124 | MIR | 3’ |
| | NM_183397 | 729 807 | 42 124 | MIR | 3’ |
| PXN | NM_002859 | - | - | - | - |
| | BC052611 | 1166 1411 | 13 268 | MIRb | 3’ |
| QPRT | NM_014298 | 1053 1222 | 13 192 | MIR | 3’ |
| RAB39B | NM_171998 | 1812 1946 | 231 93 | MIRb | 3’ |
| RAB4B | NM_016154 | 161 375 | 10 226 | MIR | 5’ |
| | | 778 835 | 221 273 | THER1_MD | 5’ |
| RAB7B | NM_177403 | 1462 1620 | 24 192 | MIRb | 3’ |
| RAE1 | NM_003610 | - | - | - | - |
| | NM_001015885 | 10 90 | 188 109 | MIR3 | 5’ |
| RAG1 | NM_000448 | 4389 4598 | 262 17 | MIRb | 3’ |
| RALGPS1 | NM_014636 | - | - | - | - |
| | BC032372 | 1616 1662 | 120 169 | MIRb | 3’ |
| RANBP10 | NM_020850 | 2751 2799 | 126 72 | MIRm | 3’ |
| | | 3969 4111 | 202 57 | MIR3 | 3’ |
| RAP1GAP | NM_002885 | 202 57 | 19 200 | MIR3 | 3’ |

| | | | | | |
|---|---|---|---|---|---|
| RAPGEF3 | NM_006105<br>U78169 | -<br>3955 4092 | -<br>24 166 | -<br>MIR | -<br>3’ |
| RASD2 | NM_014310 | 1953 2099<br>2084 2140<br>2426 2542 | 154 8<br>272 220<br>194 67 | MIR3<br>MIRm<br>MIRb | 3’<br>3’<br>3’ |
| RASGRF1 | NM_002891<br>NM_153815 | -<br>676 821 | -<br>68 195 | -<br>MIRb | -<br>5’ |
| RASL10B | NM_033315 | 2563 2679 | 3 128 | MIRm | 5’ |
| RASL12 | NM_016563 | 2333 2460 | 162 20 | MIR | 3’ |
| RBBP9 | NM_006606 | 905 943 | 68 30 | MIR | 3’ |
| RBM19 | NM_016196 | 3343 3523<br>3933 4124 | 37 231<br>21 216 | MIRb<br>MIR | 3’<br>3’ |
| RBM43 | NM_198557 | 1250 1458 | 251 37 | MIRb | 3’ |
| RCHY1 | NM_015436<br>NM_001008925<br>NM_001009922 | 2863 3042<br>2861 3040<br>2836 3015 | 223 21<br>223 21<br>223 21 | MIRb<br>MIRb<br>MIRb | 3’<br>3’<br>3’ |
| RD3 | NM_183059 | 2241 2336 | 204 109 | MIR3 | 3’ |
| RECQL5 | NM_004259<br>NM_001003715<br>NM_001003716 | -<br>-<br>2053 2144 | -<br>-<br>139 32 | -<br>-<br>MIR | -<br>-<br>3’ |
| RELB | NM_006509 | 2033 2080 | 145 97 | MIR | 3’ |
| REXO1L1 | NM_172239 | 2003 2143<br>5984 6137 | 4 149<br>40 208 | MIRb<br>MIRb | 3’<br>3’ |
| REXO1L6P | XM_376784 | 2447 2587<br>6431 6584 | 4 149<br>40 208 | MIRb<br>MIR | 3’<br>3’ |
| RFESD | NM_173362 | 453 631<br>919 1006 | 243 36<br>144 236 | MIR<br>MIR | 5’<br>5’ |
| RFX3 | NM_002919<br>NM_134428 | 2724 2970<br>- | 261 12<br>- | MIR<br>- | 3’<br>- |
| RGPD4 | BX537861 | 1877 1958 | 187 274 | THER1_MD | 3’ |
| RGS6 | NM_004296 | 2849 2927 | 193 267 | MIRb | 3’ |
| RGS9BP | NM_207391 | 163 304 | 27 175 | MIR3 | 5’ |
| RHBDD2 | NM_001040456<br>NM_001040457 | 1244 1386<br>1366 1508 | 262 124<br>262 124 | MIRb<br>MIRb | TAG<br>TAG |
| RHBDD3 | NM_012265 | 181 232 | 137 84 | MIRm | 5’ |
| RHO | NM_000539 | 1469 1529<br>1601 1797 | 216 276<br>79 260 | MIRm<br>MIRb | 3’<br>3’ |
| RIC3 | NM_024557<br>AF040723 | -<br>2572 2753 | -<br>207 2 | -<br>MIRb | -<br>3’ |
| RIG | XM_932493<br>U32331 | -<br>1804 1949 | -<br>111 268 | -<br>MIRb | -<br>3’ |
| RILP | NM_031430 | 1677 1774 | 115 215 | MIRb | 3’ |
| RIMS1 | NM_014989<br>AB045726 | -<br>5520 5691<br>6009 6158 | -<br>245 45<br>179 23 | -<br>MIRb<br>MIRb | -<br>3’<br>3’ |
| RIMS3 | NM_014747 | 2524 2620<br>6170 6352 | 28 128<br>8 206 | MIRb<br>MIR3 | 3’<br>3’ |
| RIMS4 | NM_182970 | 3499 3620<br>3766 3862<br>4243 4490 | 190 69<br>145 43<br>258 15 | MIR3<br>MIR<br>MIR_Mars | 3’<br>3’<br>3’ |
| RMND5B | NM_022762<br>AY359092 | -<br>1006 1159 | -<br>148 6 | -<br>MIRb | -<br>3’ |
| RNASE4 | NM_194430<br>NM_002937<br>NM_194431 | 1485 1592<br>-<br>- | 122 9<br>-<br>- | MIR3<br>-<br>- | 3’<br>-<br>- |
| RNASE7 | NM_032572 | 1186 1341 | 167 8 | MIRb | 3’ |
| RNASEL | NM_021133 | 2683 2814 | 209 75 | MIRb | 3’ |

| | | | | | |
|---|---|---|---|---|---|
| | | 3634 3867 | 263 22 | MIRb | 3’ |
| RNF213 | NM_020914 | 10721 10818 | 91 199 | MIR3 | 3’ |
| RNF7 | NM_014245 | - | - | - | - |
| | NM_183237 | 175 289 | 140 13 | MIR | CDS |
| | AF312226 | - | - | - | - |
| RNF8 | NM_003958 | 4015 4122 | 37 138 | MIR | 3’ |
| | NM_183078 | 3810 3917 | 37 138 | MIR | 3’ |
| RNMT | NM_003799 | 5208 5360 | 260 96 | MIRb | 3’ |
| RNPS1 | NM_006711 | - | - | - | - |
| | NM_080594 | - | - | - | - |
| | L37368 | 2298 2400 | 172 72 | MIR | 3’ |
| RP1-21O18.1 | NM_001017999 | 30 235 | 246 17 | MIR | 5’ |
| | NM_001018000 | - | - | - | - |
| | NM_001018001 | - | - | - | - |
| | NM_015209 | - | - | - | - |
| | NM_201628 | 3674 3824 | 63 231 | MIRc | 3’ |
| RPE | NM_199229 | 1397 1587 | 49 246 | MIR | 3’ |
| | | 1694 1748 | 168 114 | MIRb | 3’ |
| | NM_006916 | 1400 1590 | 49 246 | MIR | 3’ |
| | | 1697 1751 | 168 114 | MIRb | 3’ |
| RPGR | NM_001034853 | 4351 4435 | 110 193 | MIR3 | 3’ |
| | NM_000328 | - | - | - | - |
| RPL28 | NM_000991 | 3337 3561 | 12 249 | MIR | 3’ |
| | | 3699 3824 | 63 190 | MIR | 3’ |
| RPL32P3 | XM_379215 | 278 375 | 182 90 | MIRc | 5’ |
| RPL7AL2 | XM_066102 | 104 168 | 165 100 | MIR | CDS |
| RPRD1B | NM_021215 | 3918 4158 | 266 11 | MIR_Mars | 3’ |
| RPUSD3 | NM_173659 | 1075 1201 | 19 166 | MIRb | 3’ |
| RRAS | NM_006270 | 762 809 | 78 136 | MIRb | 3’ |
| RRM2B | NM_015713 | 1516 1710 | 11 209 | MIR | 3’ |
| | | 3005 3178 | 260 74 | MIRb | 3’ |
| RTP2 | NM_001004312 | 1127 1269 | 20 163 | MIR3 | 3’ |
| RUFY4 | NM_198483 | 200 282 | 187 102 | MIR3 | 5’ |
| | | 1232 1399 | 205 31 | MIR3 | 5’ |
| | | 3209 3344 | 57 199 | MIR3 | 3’ |
| RUNX1 | NM_001754 | - | - | - | - |
| | NM_001001890 | - | - | - | - |
| | D43967 | 2471 2603 | 105 241 | MIR | 3’ |
| | D10570 | 1661 1793 | 105 241 | MIR | 3’ |
| SAG | NM_000541 | - | - | - | - |
| | BX647827 | 5307 5419 | 133 244 | MIRb | 3’ |
| SAP18 | NM_005870 | 1449 1530 | 163 259 | MIR | 3’ |
| SAPS3 | NM_018312 | 95 151 | 142 86 | MIR | 5’ |
| SARM1 | NM_015077 | 3237 3399 | 218 41 | MIRb | 3’ |
| | | 4669 4709 | 102 62 | MIR | 3’ |
| | | 6603 6801 | 37 266 | MIRb | 3’ |
| SART3 | NM_014706 | 704 903 | 248 26 | MIRb | 3’ |
| | BC041638 | 1026 1179 | 20 184 | MIRb | 3’ |
| SCAND2 | NM_022050 | - | - | - | - |
| | NM_033633 | - | - | - | - |
| | NM_033634 | - | - | - | - |
| | NM_033635 | 2821 2976 | 67 230 | MIR_Mars | 3’ |
| | NM_033636 | 2863 3018 | 67 230 | MIR_Mars | 3’ |
| | NM_033640 | - | - | - | - |
| SCARB2 | NM_005506 | 429 2 4357 | 163 96 | MIRb | 3’ |
| SCD | NM_005063 | 4937 5093 | 65 229 | MIR | 3’ |
| SCD5 | NM_001037582 | - | - | - | - |

| | | | | | |
|---|---|---|---|---|---|
| | NM_024906 | 1231 1323 | 110 206 | MIR | 3’ |
| SCEL | NM_003843 | - | - | - | - |
| | NM_144777 | - | - | - | - |
| | BC047536 | 2756 2860 | 98 202 | MIR3 | 3’ |
| SCMH1 | NM_001031694 | - | - | - | - |
| | NM_012236 | 56 184 | 142 6 | MIRm | 5’ |
| SCML4 | NM_198081 | 1822 2038 | 5 251 | MIR | 3’ |
| SCN1B | NM_001037 | - | - | - | - |
| | NM_199037 | 860 936 | 50 126 | MIRm | TAG |
| SCN4B | NM_174934 | 2928 3029 | 93 189 | MIRb | 3’ |
| SCN8A | NM_014191 | - | - | - | - |
| | XM_001141985 | 10744 10948 | 2 216 | MIRb | 3’ |
| SCNN1D | NM_002978 | 8 89 | 99 7 | MIRb | 5’ |
| SCNN1G | NM_001039 | 2748 2914 | 177 10 | MIR | 3’ |
| SCYL1BP1 | NM_152281 | - | - | - | - |
| | BC064945 | 792 892 | 249 137 | MIR | 3’ |
| SEC23B | NM_006363 | - | - | - | - |
| | NM_032985 | - | - | - | - |
| | NM_032986 | - | - | - | - |
| | BC005404 | 152 243 | 153 42 | MIR | 5’ |
| SEC61A1 | NM_013336 | 3103 3232 | 136 4 | MIR | 3’ |
| SEMA4F | NM_004263 | 4095 4249 | 54 210 | MIRb | 3’ |
| SEPN1 | NM_020451 | 3662 3820 | 209 33 | MIRb | 3’ |
| | NM_206926 | 3560 3718 | 209 33 | MIRb | 3’ |
| SEPT6 | NM_145799 | 3857 3960 | 131 28 | MIR | 3’ |
| | NM_015129 | 2548 2651 | 131 28 | MIR | 3’ |
| | NM_145800 | - | - | - | - |
| | NM_145802 | - | - | - | - |
| SERINC1 | NM_020755 | 2553 2601 | 160 113 | MIR3 | 3’ |
| SERPINA13 | NM_207378 | 23 161 | 237 95 | MIRb | 5’ |
| SERPINA5 | NM_000624 | 1543 1688 | 168 26 | MIR3 | 3’ |
| | | 1798 1933 | 22 165 | MIR | 3’ |
| SERTAD3 | NM_013368 | 14 149 | 136 3 | MIR3 | 5’ |
| | | 219 281 | 154 91 | MIRb | 5’ |
| | NM_203344 | - | - | - | - |
| SFTPA2 | NM_006926 | 1095 1193 | 184 82 | MIR3 | 3’ |
| | | 1739 1827 | 92 190 | MIRb | 3’ |
| | | 1883 2127 | 255 4 | MIRb | 3’ |
| SFXN5 | NM_144579 | 2212 2332 | 59 182 | MIR3 | 3’ |
| SGCB | NM_000232 | 1631 1904 | 268 2 | MIRb | 3’ |
| SGCD | NM_000337 | 7560 7785 | 262 5 | MIR | 3’ |
| | NM_172244 | - | - | - | - |
| SGIP1 | NM_032291 | - | - | - | - |
| | BC040516 | 1465 1548 | 140 53 | MIRb | CDS |
| SH2B | NM_015503 | 683 853 | 2 188 | MIR | 5’ |
| | | 847 1098 | 262 2 | MIRb | 5’ |
| | | 1571 1601 | 113 143 | MIRm | 5’ |
| SH2D1B | NM_053282 | 846 992 | 1 163 | MIRb | 3’ |
| SH2D4B | NM_207372 | 2480 2580 | 1 100 | MIRb | 3’ |
| | | 3380 3552 | 37 215 | MIRm | 3’ |
| SH3BP2 | NM_003023 | 2539 2635 | 106 2 | MIR | 3’ |
| | | 4489 4610 | 263 115 | MIRb | 3’ |
| | | 4607 4680 | 44 121 | MIR | 3’ |
| | | 4918 4982 | 156 99 | MIRb | 3’ |
| SH3TC2 | NM_024577 | - | - | - | - |
| | AF370410 | 2728 2862 | 180 38 | MIR3 | 3’ |
| SHC3 | NM_016848 | - | - | - | - |

| | | | | | |
|---|---|---|---|---|---|
| | BX641139 | 2431 2469<br>3140 3300 | 9 48<br>49 212 | MIR<br>MIR | 5’<br>5’ |
| SHC4 | NM_203349 | 3465 3595 | 73 205 | MIR3 | 3’ |
| SHF | NM_138356 | 1247 1402 | 84 268 | MIRb | 3’ |
| SHOX2 | NM_006884<br>NM_003030 | 2383 2450<br>2029 2096 | 190 261<br>190 261 | MIRb<br>MIRb | 3’<br>3’ |
| SHPRH | NM_173082 | 6530 6748 | 252 20 | MIRb | 3’ |
| SIAE | NM_170601<br>BC040966 | -<br>617 710 | -<br>249 147 | -<br>MIR | -<br>TAG |
| SIAH3 | NM_198849 | 3422 3548 | 157 23 | MIRc | 3’ |
| SIGLEC15 | NM_213602<br>AK095432 | -<br>1192 1404 | -<br>16 237 | -<br>MIRc | -<br>CDS |
| SIGLEC5 | NM_003830 | 1938 2113 | 20 210 | MIRb | 3’ |
| SIL1 | NM_001037633<br>NM_022464 | 1838 1953<br>1750 1865 | 32 143<br>32 143 | MIRb<br>MIRb | 3’<br>3’ |
| SIM1 | NM_005068 | 3690 3816 | 11 145 | MIRb | 3’ |
| SIM2 | NM_005069<br>NM_009586 | -<br>2113 2210 | -<br>75 178 | -<br>MIRm | -<br>3’ |
| SIRPA | NM_001040022<br>NM_001040023<br>NM_080792 | 3088 3140<br>2996 3048<br>2755 2807 | 123 174<br>123 174<br>123 174 | THER1_MD<br>THER1_MD<br>THER1_MD | 3’<br>3’<br>3’ |
| SLAMF1 | NM_003037<br>BC067847 | -<br>1641 1814 | -<br>24 202 | -<br>MIRb | -<br>3’ |
| SLC11A1 | NM_000578 | 1903 2087 | 7 222 | MIR | 3’ |
| SLC16A12 | NM_213606 | 2499 2554 | 208 154 | MIR3 | 3’ |
| SLC16A14 | NM_152527 | 1930 2009 | 153 233 | MIRb | 3’ |
| SLC16A4 | NM_004696<br>BC021664 | -<br>1276 1637 | -<br>131 193 | -<br>MIR3 | -<br>3’ |
| SLC16A9 | NM_194298 | 3177 3224 | 212 156 | MIR | 3’ |
| SLC17A8 | NM_139319 | 2369 2525<br>3621 3699 | 228 57<br>184 261 | THER1_MD<br>MIR | 3’<br>3’ |
| SLC18A1 | NM_003053<br>BC006317 | -<br>120 263 | -<br>5 151 | -<br>MIR3 | -<br>5’ |
| SLC1A2 | NM_004171 | 3330 3404<br>6674 6845 | 178 253<br>2 185 | MIRb<br>MIR3 | 3’<br>3’ |
| SLC1A3 | NM_004172 | 2431 2552 | 82 205 | MIR3 | 3’ |
| SLC1A4 | NM_003038 | 3037 3165 | 117 241 | MIRb | 3’ |
| SLC1A7 | NM_006671 | 2082 2211<br>2407 2491 | 168 26<br>189 102 | MIRb<br>MIR3 | 3’<br>3’ |
| SLC22A16 | NM_033125<br>AB055798 | -<br>169 266 | -<br>259 162 | -<br>MIRb | -<br>5’ |
| SLC22A7 | NM_153320<br>NM_006672 | 2004 2230<br>1998 2224 | 5 216<br>5 216 | MIR<br>MIR | 3’<br>3’ |
| SLC23A1 | NM_005847<br>NM_152685 | 1883 2046<br>1871 2034 | 3 153<br>3 153 | MIR<br>MIR | 3’<br>3’ |
| SLC25A29 | NM_152333<br><br>NM_001039355 | 1256 1430<br>1888 1952<br>1300 1426 | 33 231<br>94 162<br>33 176 | MIRb<br>MIRb<br>MIRb | 3’<br>3’<br>3’ |
| SLC25A34 | NM_207348 | 1379 1578 | 235 26 | MIRb | 3’ |
| SLC26A5 | NM_198999<br>NM_206883<br>NM_206884<br>NM_206885 | -<br>2313 2474<br>1786 1947<br>1243 1404 | -<br>4 166<br>4 166<br>4 166 | -<br>MIRb<br>MIRb<br>MIRb | -<br>3’<br>3’<br>3’ |
| SLC29A2 | NM_001532 | 2174 2412 | 260 2 | MIR_Mars | 3’ |
| SLC2A10 | NM_030777 | 2532 2676<br>2706 2925 | 41 202<br>5 241 | MIR3<br>MIRb | 3’<br>3’ |
| SLC2A5 | NM_003039 | - | - | - | - |

| | | | | | |
|---|---|---|---|---|---|
| | BC035878 | 968 1105<br>1356 1572 | 237 100<br>245 13 | MIRb<br>MIR | 3’<br>3’ |
| SLC30A6 | NM_017964<br>BC032525 | -<br>3421 3591 | -<br>5 172 | -<br>MIR3 | -<br>3’ |
| SLC30A7 | NM_133496 | 4859 5052 | 196 3 | MIR3 | 3’ |
| SLC35A5 | NM_017945 | 2644 2750 | 143 37 | THER1_MD | 3’ |
| SLC35C2 | NM_173179<br>NM_015945<br>NM_173073 | 2294 2427<br>-<br>2036 2169 | 43 189<br>-<br>43 189 | MIR<br>-<br>MIR | 3’<br>-<br>3’ |
| SLC36A3 | NM_181774 | 2259 2418 | 19 193 | MIR3 | 3’ |
| SLC36A4 | NM_152313 | 2272 2432 | 177 12 | MIR3 | 3’ |
| SLC43A3 | NM_014096<br>NM_199329<br>NM_017611 | 2369 2484<br>2264 2379<br>2242 2357 | 41 157<br>41 157<br>41 157 | MIRm<br>MIRm<br>MIRm | 3’<br>3’<br>3’ |
| SLC45A2 | NM_016180<br>NM_001012509 | -<br>1689 1797<br>2461 2578 | -<br>173 51<br>12 132 | -<br>MIRb<br>MIRb | -<br>3’<br>3’ |
| SLC4A1 | NM_000342 | 3206 3383<br>3393 3488 | 82 258<br>107 19 | MIRb<br>MIRb | 3’<br>3’ |
| SLC6A12 | NM_003044 | 72 189<br>390 535<br>3221 3382 | 144 20<br>12 164<br>159 3 | MIR<br>MIRm<br>MIR3 | 5’<br>3’<br>3’ |
| SLC6A13 | NM_016615<br>BC020867 | -<br>343 443<br>457 665 | -<br>204 90<br>10 220 | -<br>MIR3<br>MIRb | -<br>TAG<br>3’ |
| SLC6A15 | NM_182767<br>NM_018057 | -<br>2217 2464<br>3994 4153 | -<br>256 13<br>224 59 | -<br>MIRb<br>MIRb | -<br>3’<br>3’ |
| SLC6A6 | NM_003043<br>BC006252 | -<br>999 1128 | -<br>37 163 | -<br>MIRb | -<br>3’ |
| SLC6A7 | NM_014228 | 2513 2649 | 157 8 | MIR3 | 3’ |
| SLC9A1 | NM_003047<br>BC012121 | -<br>1840 1988 | -<br>52 203 | -<br>MIR | -<br>TAG |
| SLCO2B1 | NM_007256 | 2464 2581 | 31 152 | MIR | 3’ |
| SLIC1 | NM_182854<br>BC027944 | -<br>404 510 | -<br>140 27 | -<br>MIR | -<br>TAG |
| SMAD6 | NM_005585<br>BC052569 | -<br>2675 2721<br>3085 3152<br>3813 3927 | -<br>273 108<br>46 114<br>33 154 | -<br>MIRm<br>MIR<br>MIR3 | -<br>3’<br>3’<br>3’ |
| SMC1A | NM_006306 | 4571 4723<br>5971 6160<br>6154 6248<br>6221 6283<br>6525 6599<br>6636 6739<br>8429 8550 | 22 173<br>14 237<br>197 97<br>76 16<br>152 240<br>124 20<br>193 42 | MIR<br>MIRb<br>MIR3<br>MIR_Mars<br>MIRb<br>MIR<br>MIR3 | 3’<br>3’<br>3’<br>3’<br>3’<br>3’<br>3’ |
| SMCR5 | NM_144774<br>AF467442 | 948 1168<br>1727 1817<br>1966 2172 | 22 259<br>81 166<br>270 31 | MIRb<br>MIR3<br>MIR | 3’<br>3’<br>3’ |
| SNHG3 | AJ006834<br>AJ006835 | -<br>1878 1907<br>1934 1999 | -<br>240 210<br>142 68 | -<br>MIR<br>MIR | -<br>ncRNA<br>ncRNA |
| SNPH | NM_014723<br>AF187733 | -<br>333 528 | -<br>4 202 | -<br>MIR3 | -<br>5’ |
| SNX21 | NM_033421 | 2535 2688<br>3086 3164 | 211 37<br>157 77 | MIR<br>MIRm | 3’<br>3’ |

| | | | | | |
|---|---|---|---|---|---|
| | NM_152897 | 1298 1436 | 196 37 | MIR | 3’ |
| | | 1834 1906 | 157 83 | MIR3 | 3’ |
| SNX27 | NM_030918 | 2050 2232 | 16 202 | MIR3 | 3’ |
| SNX6 | NM_021249 | 2898 3057 | 19 194 | MIRb | 3’ |
| | NM_152233 | 2793 2952 | 19 194 | MIRb | 3’ |
| SORBS2 | NM_003603 | - | - | - | - |
| | NM_021069 | - | - | - | - |
| | AF090937 | 866 929 | 178 118 | MIR3 | 3’ |
| SORCS2 | NM_020777 | 4356 4509 | 193 17 | MIR3 | 3’ |
| SOST | NM_025237 | 1102 1273 | 5 170 | MIR3 | 3’ |
| SOX12 | NM_006943 | 3224 3271 | 81 128 | MIR | 3’ |
| | | 3594 3759 | 11 230 | MIR | 3’ |
| | | 3765 3869 | 158 46 | MIR_Mars | 3’ |
| SP2 | NM_003110 | - | - | - | - |
| | BC005914 | 863 1009 | 174 24 | MIRb | 3’ |
| SPAG8 | NM_001039592 | - | - | - | - |
| | NM_172312 | 1865 1981 | 158 37 | MIRb | 3’ |
| SPAR | L10123 | 30 207 | 65 262 | MIR | 5’ |
| SPATA12 | NM_181727 | 1029 1150 | 92 219 | MIR | CDS |
| | | 1377 1481 | 193 89 | MIRb | 3’ |
| SPEF2 | NM_024867 | - | - | - | - |
| | NM_144722 | 2153 2291 | 37 169 | MIRb | 3’ |
| SPRY3 | NM_005840 | 2449 2654 | 204 5 | MIR3 | 3’ |
| | | 4530 4761 | 258 22 | MIR | 3’ |
| | | 5305 5452 | 177 33 | MIR3 | 3’ |
| | | 8033 8180 | 152 4 | MIRb | 3’ |
| | | 8196 8247 | 115 165 | MIRb | 3’ |
| SPRY4 | NM_030964 | 4372 4474 | 129 24 | MIRc | 3’ |
| | NM_001127496 | 4321 4423 | 129 24 | MIRc | 3’ |
| SPTBN4 | NM_020971 | - | - | - | - |
| | NM_025213 | 2160 2229 | 272 195 | MIR | TAG |
| SRC | NM_005417 | 187 279 | 156 53 | MIRb | 5’ |
| | | 275 412 | 142 3 | MIR | 5’ |
| | | 2674 2828 | 187 37 | MIR3 | 3’ |
| | NM_198291 | 100 178 | 142 53 | MIRb | 5’ |
| | | 174 311 | 142 3 | MIR | 5’ |
| | | 2573 2727 | 187 37 | MIR3 | 3’ |
| SRGAP3 | NM_014850 | 5608 5729 | 38 159 | MIRb | 3’ |
| | | 6231 6324 | 155 66 | MIR | 3’ |
| | NM_001033117 | 5536 5657 | 38 159 | MIRb | 3’ |
| | | 6159 6252 | 155 66 | MIR | 3’ |
| | NM_001033116 | 5505 5626 | 38 159 | MIRb | 3’ |
| | | 6128 6221 | 155 66 | MIR | 3’ |
| SRI | NM_003130 | 1650 1811 | 45 215 | MIRb | 3’ |
| | NM_198901 | 1632 1793 | 45 215 | MIRb | 3’ |
| SS18L1 | NM_198935 | 3005 3126 | 216 94 | MIR | 3’ |
| | NM_015558 | 2193 2314 | 216 94 | MIR | 3’ |
| SSPO | NM_198455 | - | - | - | - |
| | XM_376720 | 4470 4588 | 16 143 | MIR | 3’ |
| ST3GAL1 | NM_003033 | 379 433 | 141 87 | MIRb | 5’ |
| | NM_173344 | 379 433 | 141 87 | MIRb | 5’ |
| ST6GAL1 | NM_173216 | 3173 3272 | 120 14 | MIRb | 3’ |
| | | 3361 3571 | 34 260 | MIRb | 3’ |
| | NM_003032 | 3051 3150 | 120 14 | MIRb | 3’ |
| | | 3239 3449 | 34 260 | MIRb | 3’ |
| | NM_173217 | 2516 2615 | 120 14 | MIRb | 3’ |
| | | 2704 2914 | 34 260 | MIRb | 3’ |

| | | | | | |
|---|---|---|---|---|---|
| ST6GALNAC6 | NM_013443 | 89 147 | 141 83 | MIRb | ATG |
| ST7L | NM_017744<br>NM_138727<br>NM_138728<br>NM_138729 | 1982 2190<br>1931 2139<br>1889 2097<br>- | 23 255<br>23 255<br>23 255<br>- | MIRb<br>MIRb<br>MIRb<br>- | TAG<br>TAG<br>TAG<br>- |
| STAB1 | NM_015136<br>AB052957 | -<br>2351 2443 | -<br>70 167 | -<br>MIR3 | -<br>CDS |
| STARD10 | NM_006645 | 12 119 | 153 38 | MIR | 5’ |
| STAT5A | NM_003152 | 111 188 | 188 114 | MIR3 | 5’ |
| STEAP2 | NM_152999 | 2765 3007 | 255 5 | MIR | 3’ |
| STK32B | NM_018401 | 1415 1621<br>2363 2568 | 15 216<br>1 209 | MIR3<br>MIRb | 3’<br>3’ |
| STK4 | NM_006282 | 2633 2793<br>3345 3534<br>5646 5823 | 196 30<br>268 63<br>180 4 | MIR3<br>MIRb<br>MIRb | 3’<br>3’<br>3’ |
| STK40 | NM_032017 | 3099 3184 | 71 152 | MIR | 3’ |
| STRA6 | NM_022369<br>BC015881 | -<br>599 687<br>1468 1516<br>1701 1870 | -<br>141 49<br>30 80<br>96 252 | -<br>MIRm<br>MIRb<br>MIRb | -<br>TAG<br>3’<br>3’ |
| STYK1 | NM_018423 | 2621 2730 | 132 254 | MIRb | 3’ |
| SUHW4 | NM_017661<br>NM_001002843<br>NM_001002844 | -<br>-<br>1072 1284 | -<br>-<br>260 48 | -<br>-<br>MIR | -<br>-<br>3’ |
| SUMF1 | NM_182760 | 1402 1566 | 15 184 | MIR3 | 3’ |
| SUOX | NM_000456<br>NM_001032386<br>NM_001032387 | 138 271<br>-<br>- | 142 16<br>-<br>- | MIR<br>-<br>- | 5’<br>-<br>- |
| SV2C | NM_014979<br>XM_043493 | -<br>5375 5551 | -<br>23 186 | -<br>MIR3 | -<br>3’ |
| SYDE2 | NM_032184<br>AL834286 | -<br>2587 2697 | -<br>32 141 | -<br>MIR | -<br>3’ |
| SYNGR1 | NM_004711<br><br><br>NM_145731<br>NM_145738 | 1293 1442<br>1807 1891<br>2632 2706<br>1116 1323<br>- | 153 2<br>121 33<br>170 103<br>35 268<br>- | MIRb<br>MIR<br>MIRb<br>MIRb<br>- | 3’<br>3’<br>3’<br>3’<br>- |
| SYNPO | NM_007286 | 3323 3488 | 7 164 | MIR3 | 3’ |
| SYT11 | NM_152280 | 3192 3223 | 65 34 | MIR | 3’ |
| SYT7 | NM_004200 | 3699 3859 | 15 182 | MIR | 3’ |
| TACR1 | NM_001058<br>NM_015727 | 1549 1726<br>- | 2 186<br>- | MIRm<br>- | 3’<br>- |
| TADA3L | NM_006354<br>NM_133480<br>NM_133481 | 456 521<br>456 521<br>456 521 | 234 167<br>234 167<br>234 167 | MIRb<br>MIRb<br>MIRb | 5’<br>5’<br>5’ |
| TAF9B | NM_015975 | 2451 2654 | 43 258 | MIRb | 3’ |
| TAP2 | NM_000544<br>NM_018833 | 2822 3041<br>5111 5316 | 260 27<br>254 41 | MIRb<br>MIRb | 3’<br>3’ |
| TAPBP | NM_003190<br>NM_172208<br>NM_172209 | -<br>2144 2256<br>- | -<br>2 117<br>- | -<br>MIRb<br>- | -<br>3’<br>- |
| TBC1D2 | NM_018421 | 2898 3005 | 32 142 | MIR | 5’ |
| TBC1D24 | NM_020705 | 3431 3540 | 81 182 | MIR | 3’ |
| TBC1D30 | XM_037557 | 3830 3926 | 176 86 | MIR | 3’ |
| TBC1D5 | NM_014744 | 497 548 | 253 202 | MIR | 5’ |
| TBN | NM_138572 | 1681 1750 | 262 190 | THER1_MD | 3’ |
| TBRG1 | NM_032811 | - | - | - | - |

| | | | | | |
|---|---|---|---|---|---|
| | BC032312 | 452 611 | 68 262 | MIR | TAG |
| | | 638 767 | 192 41 | MIRb | 3’ |
| | | 1136 1210 | 271 196 | MIRm | 3’ |
| TCF2 | NM_000458 | - | - | - | - |
| | NM_006481 | 2981 3108 | 140 9 | MIR | 3’ |
| | | 3632 3748 | 91 211 | MIRb | 3’ |
| TCL1B | NM_199206 | 672 813 | 198 52 | MIRb | 3’ |
| | NM_004918 | - | - | - | - |
| TCL6 | NM_020553 | 2372 2484 | 138 25 | MIRb | 3‘ |
| | NM_020554 | 1393 1567 | 197 2 | MIR3 | ATG |
| | NM_014418 | 113 287 | 197 2 | MIR3 | 5’ |
| | | 1723 1755 | 69 38 | MIRb | ATG |
| | | 2312 2431 | 124 2 | MIRc | 3’ |
| | NM_020550 | 113 287 | 197 2 | MIR3 | 5’ |
| | | 1723 1755 | 69 38 | MIRb | ATG |
| | | 113 287 | 197 2 | MIR3 | 3’ |
| | NM_020552 | 1723 1755 | 69 38 | MIRb | ATG |
| | | 2728 2840 | 138 25 | MIRb | 3’ |
| | NM_012468 | 113 287 | 197 2 | MIR3 | 5’ |
| | | 1723 1755 | 69 38 | MIRb | ATG |
| | | 2312 2431 | 124 2 | MIRc | 3’ |
| TDP1 | NM_018319 | 3224 3338 | 179 64 | MIR3 | 3’ |
| | NM_001008744 | 3001 3115 | 179 64 | MIR3 | 3’ |
| TFCP2L1 | NM_014553 | 2826 2900 | 145 68 | MIR | 3’ |
| | | 4606 4791 | 183 1 | MIR3 | 3’ |
| TFEB | NM_007162 | 196 273 | 126 206 | MIR | 5’ |
| TFIP11 | NM_001008697 | - | - | - | - |
| | NM_012143 | - | - | - | - |
| | AF070662 | 1 121 | 229 108 | MIRb | 5’ |
| TGFB1 | NM_000660 | - | - | - | - |
| | X02812 | 2324 2462 | 249 103 | MIR | 3’ |
| TGFBR2 | NM_001024847 | 474 519 | 136 91 | MIR | CDS |
| | NM_003242 | - | - | - | - |
| TGM2 | NM_004613 | 2724 2859 | 256 116 | MIRb | 3' |
| | | 2835 2984 | 137 6 | MIRc | 3' |
| | | 3050 3222 | 17 203 | MIR3 | 3' |
| | NM_198951 | 1812 1879 | 251 184 | MIR | 3' |
| | CR604340 | 1203 1333 | 47 180 | MIR | 3' |
| | AK126508 | 391 488 | 137 40 | MIR | 5' |
| | | 614 832 | 28 261 | MIR | ATG |
| | | 948 1087 | 142 1 | MIRb | CDS |
| | | 1099 1191 | 28 121 | MIR | CDS |
| | | 2270 2405 | 256 116 | MIRb | 3' |
| | | 2381 2530 | 137 6 | MIRc | 3' |
| | | 2596 2768 | 17 203 | MIR3 | 3' |
| THAP5 | NM_182529 | 1743 1894 | 21 179 | MIR3 | 3’ |
| | | 2967 3126 | 91 262 | MIRb | 3’ |
| THRSP | NM_003251 | 753 955 | 9 220 | MIRb | 3’ |
| TICAM2 | NM_021649 | 2643 2801 | 6 177 | MIR3 | 3’ |
| TK2 | NM_004614 | 1445 1666 | 246 3 | MIRb | 3’ |
| TLL1 | NM_012464 | 5803 5989 | 222 21 | MIRb | 3’ |
| TLOC1 | NM_003262 | 23 148 | 26 143 | MIR | 5’ |
| TMC1 | NM_138691 | 262 337 | 11 108 | MIRb | 3’ |
| TMEM10 | NM_033207 | 1407 1607 | 18 256 | MIRb | 3’ |
| | NM_020847 | 3225 3378 | 67 209 | MIRb | 3’ |
| | | - | - | - | - |
| TMEM106B | NM_018374 | 4043 4170 | 153 24 | MIRm | 3’ |
| TMEM118 | NM_032814 | 1517 1622 | 4 126 | MIRb | TAG |

| | | | | | |
|---|---|---|---|---|---|
| TMEM126B | NM_018480<br>BC017574 | -<br>660 802<br>831 1054 | -<br>262 123<br>22 249 | -<br>MIR_Mars<br>MIRb | -<br>3’<br>3’ |
| TMEM132B | NM_052907<br>AB058689 | -<br>5285 5377 | -<br>65 160 | -<br>MIRb | -<br>3’ |
| TMEM139 | NM_153345 | 1526 1641 | 141 22 | MIRb | 3’ |
| TMEM144 | NM_018342 | 2553 2783 | 30 262 | MIR | 3’ |
| TMEM16C | NM_031418 | 22 127 | 89 196 | MIRb | 5’ |
| TMEM173 | NM_198282 | 1547 1683 | 93 242 | MIRc | 3’ |
| TMEM186 | NM_015421 | 978 1139 | 2 191 | MIRb | 3’ |
| TMEM215 | NM_212558 | 2665 2791 | 83 204 | MIR3 | 3’ |
| TMEM29 | NM_014138 | 6 96 | 74 167 | MIRb | 5’ |
| TMEM50A | NM_014313 | 1796 1911 | 196 76 | MIR3 | 3’ |
| TMEM68 | NM_152417 | 1363 1419 | 73 129 | MIRm | 3’ |
| TMEM72 | NM_001123376<br>BX538120 | -<br>1717 1941 | -<br>248 13 | -<br>MIRb | -<br>3’ |
| TMPRSS11B | NM_182502 | 2163 2284 | 205 86 | MIR3 | 3’ |
| TMUB2 | NM_024107<br>NM_177441 | -<br>259 482<br>433 539 | -<br>21 229<br>268 166 | -<br>MIR<br>MIRb | -<br>5’<br>5’ |
| TNFAIP8L3 | NM_207381 | 1374 1421 | 160 207 | MIRb | 3’ |
| TNFSF12 | BC071837<br>NM_003809 | 447 522<br>- | 120 44<br>- | MIR<br>- | TAG<br>- |
| TNFSF4 | NM_003326 | 2362 2432<br>2929 3086 | 274 204<br>2 165 | THER1_MD<br>MIR3 | 3’<br>3’ |
| TNFSF8 | NM_001244 | 1097 1193 | 1 100 | MIRm | 3’ |
| TNNI1 | NM_003281 | 3320 3557<br>3556 3758<br>4679 4759 | 262 7<br>260 37<br>121 29 | MIR<br>MIRb<br>MIR | 3’<br>3’<br>3’ |
| TNRC6A | NM_014494<br>NM_020847 | 80 190<br>- | 116 3<br>- | MIRb<br>- | 5’<br>- |
| TOMM20 | NM_014765 | 1936 2009<br>2976 3035 | 42 121<br>200 256 | MIR<br>MIR | 3’<br>3’ |
| TOMM40 | BC047528 | 2230 2313 | 180 85 | MIR3 | 3’ |
| TOR1A | NM_000113<br>BC014484 | -<br>1468 1578 | -<br>29 149 | -<br>MIRb | 3’ |
| TOR2A | NM_130459 | 875 1120<br>1516 1592 | 19 239<br>52 129 | MIRm<br>MIRb | 3’<br>3’ |
| TP53 | NM_000546 | 1801 1903 | 110 217 | MIRb | 3’ |
| TP53I11 | NM_006034<br>AK056379 | -<br>164 263 | -<br>170 61 | -<br>MIRb | -<br>5’ |
| TP53INP1 | NM_033285 | 2200 2296 | 86 185 | MIR | 3’ |
| TP53RK | NM_033550 | 2639 2828 | 249 19 | MIRb | 3’ |
| TP73L | NM_003722<br>AF061512 | -<br>264 331 | -<br>238 160 | -<br>MIR | -<br>5’ |
| TPK1 | NM_022445 | 1751 1842 | 105 194 | MIR3 | 3’ |
| TPP1 | NM_000391 | 3067 3288 | 14 252 | MIRb | 3’ |
| TRAF1 | NM_005658 | 151 343 | 203 17 | MIR3 | 3’ |
| TRAF3 | NM_145725<br>NM_145726<br>NM_003300 | 196 308<br>196 308<br>- | 135 23<br>135 23<br>- | MIR<br>MIR<br>- | 5’<br>5’<br>- |
| TREML1 | NM_178174<br>AY358357 | -<br>1007 1108 | -<br>202 97 | -<br>MIR3 | -<br>3’ |
| TREML2 | NM_024807 | 2881 3071 | 7 216 | THER1_MD | 3’ |
| TRIAD3 | NM_207111<br>NM_207116<br>AY177398 | -<br>-<br>988 1049 | -<br>-<br>132 196 | -<br>-<br>MIR3 | -<br>-<br>5’ |

| | | | | | |
|---|---|---|---|---|---|
| TRIM10 | NM_006778 | 1779 1907 | 2 131 | MIR3 | 3’ |
| | NM_052828 | 1243 1372 | 2 131 | MIR3 | TAG |
| TRIM14 | NM_014788 | 3327 3545 | 241 2 | MIR | 3’ |
| | | 3813 3941 | 68 213 | THER1_MD | 3’ |
| | NM_033219 | - | - | - | - |
| | NM_033220 | - | - | - | - |
| | NM_033221 | 2566 2784 | 241 2 | MIR | 3’ |
| | | 3052 3180 | 68 213 | THER1_MD | 3’ |
| TRIM16 | NM_006470 | - | - | - | - |
| | AK056026 | 557 727 | 91 266 | MIRb | 5’ |
| | | 763 872 | 130 4 | MIR3 | 5’ |
| | AB209899 | 4628 4800 | 183 6 | MIR | 3’ |
| TRIM22 | NM_006074 | 2498 2667 | 40 212 | MIR | 3’ |
| TRIM35 | NM_171982 | - | - | - | - |
| | BC018337 | 512 619 | 139 30 | MIR | CDS |
| TRIM38 | NM_006355 | - | - | - | - |
| | BX640949 | 1585 1743 | 53 231 | MIR | 3’ |
| TRIM4 | NM_033017 | 2394 2580 | 250 51 | MIRb | 3’ |
| | NM_033091 | 2316 2513 | 250 40 | MIRb | 3’ |
| TRIM5 | NM_033034 | 2208 2315 | 18 129 | MIR | |
| | NM_033092 | - | - | - | 3’ |
| | NM_033093 | - | - | - | |
| TRIM62 | NM_018207 | 1710 1866 | 23 201 | MIR3 | 3’ |
| TRIOBP | NM_138632 | 1202 1330 | 160 31 | MIRb | TAG |
| TRPS1 | NM_014112 | 9232 9407 | 91 260 | MIR_Mars | 3’ |
| TRPV3 | NM_145068 | - | - | - | - |
| | AF514998 | 2455 2570 | 82 202 | MIR3 | 3’ |
| TRSPAP1 | NM_017846 | - | - | - | - |
| | BC039879 | 277 386 | 197 90 | MIR3 | 5’ |
| TSGA14 | NM_018718 | - | - | - | - |
| | AY186739 | 2 99 | 124 226 | MIRb | 5’ |
| TSKU | NM_015516 | 1727 1906 | 194 2 | MIR3 | 3’ |
| | | 2010 2070 | 70 128 | MIRb | 3’ |
| TSPAN2 | NM_005725 | 2497 2625 | 85 202 | MIR3 | 3’ |
| TTC12 | NM_017868 | - | - | - | - |
| | BC032355 | 2341 2392 | 138 87 | MIR | 3’ |
| TTC21A | NM_145755 | - | - | - | - |
| | BC062621 | 1107 1318 | 36 260 | MIRb | 3’ |
| | | 2055 2142 | 33 119 | THER1_MD | 3’ |
| TTC3 | NM_003316 | 98 296 | 27 239 | MIRb | 5’ |
| | NM_001001894 | - | - | - | - |
| TTC31 | NM_022492 | 2718 2891 | 73 252 | MIRb | 3’ |
| TTC39A | NM_001080494 | - | - | - | - |
| | AB007921 | 94 241 | 5 140 | MIRb | ATG |
| | | 1491 1612 | 154 33 | MIR3 | 3’ |
| | | 3199 3410 | 229 15 | MIRb | 3’ |
| | | 3795 3962 | 36 192 | MIR3 | 3’ |
| TTC9 | XM_027236 | 2816 3059 | 261 1 | MIRb | 3’ |
| | XM_938197 | 2816 3059 | 261 1 | MIRb | 3’ |
| TTL | NM_153712 | 3356 3496 | 57 191 | MIRm | 3’ |
| TTN | NM_003319 | - | - | - | - |
| | NM_133378 | - | - | - | - |
| | NM_133379 | - | - | - | - |
| | NM_133432 | - | - | - | - |
| | NM_133437 | - | - | - | - |
| | BC013396 | 1790 1953 | 268 91 | MIRb | 3’ |
| | | 1948 2079 | 124 260 | MIR_Mars | 3’ |

| | | | | | |
|---|---|---|---|---|---|
| TTPAL | NM_024331 | 1982 2063 | 115 199 | MIR3 | 3’ |
| TUB | NM_003320 | 3028 3177 | 202 36 | MIR3 | 3’ |
| | | 5861 5935 | 78 154 | MIRb | 3’ |
| | NM_177972 | 2771 2920 | 202 36 | MIR3 | 3’ |
| | | 5604 5678 | 78 154 | MIRb | 3’ |
| TUBGCP3 | NM_006322 | - | - | - | - |
| | BC007763 | 1854 1976 | 66 201 | THER1_MD | 3’ |
| TUFT1 | NM_020127 | 2980 3047 | 123 191 | MIR3 | 3’ |
| TULP4 | NM_020245 | - | - | - | - |
| | NM_001007466 | 7122 7269 | 38 185 | MIR3 | 3’ |
| TXNDC2 | NM_032243 | 18 107 | 23 117 | MIR | 5’ |
| TYK2 | NM_003331 | 151 298 | 228 81 | MIR | 5’ |
| U2AF1 | NM_006758 | - | - | - | - |
| | NM_001025203 | - | - | - | - |
| | NM_001025204 | - | - | - | - |
| | AF370386 | 306 450 | 226 57 | MIRb | 5’ |
| UBE1L2 | NM_018227 | - | - | - | - |
| | BC031637 | 1218 1423 | 60 274 | MIR | 3’ |
| UBE2V1 | U39360 | 8 160 | 194 31 | MIRb | ATG |
| | NM_021988 | 165 260 | 111 214 | MIR | 5’ |
| | | 321 473 | 194 31 | MIRb | ATG |
| | NM_199144 | 165 258 | 111 212 | MIR | 5’ |
| | | 260 396 | 178 31 | MIRb | ATG |
| | NM_022442 | 124 262 | 180 31 | MIRb | ATG |
| UBE3B | NM_130466 | 397 498 | 14 123 | MIRb | 5’ |
| | NM_183415 | - | - | - | - |
| UBL7 | NM_032907 | - | - | - | - |
| | NM_201265 | - | - | - | - |
| | BC007913 | 92 226 | 74 213 | MIRb | 5’ |
| UBXD5 | NM_145345 | - | - | - | - |
| | NM_183008 | - | - | - | - |
| | NM_145346 | - | - | - | - |
| | BC038106 | 1768 1861 | 110 207 | MIR3 | 3’ |
| UCN2 | NM_033199 | 1276 1376 | 40 139 | MIR | 3’ |
| ULK2 | NM_014683 | 5511 5602 | 154 62 | MIRb | 3’ |
| ULK3 | NM_001099436 | 2265 2500 | 10 229 | MIR | 3’ |
| UNC13A | XM_038604 | 5820 5916 | 187 85 | MIR3 | 3’ |
| | XM_937931 | - | - | - | - |
| UNC50 | NM_014044 | 815 907 | 8 119 | MIR | 5’ |
| UNQ9370 | NM_207447 | 1215 1417 | 1 205 | MIRb | 3’ |
| USH2A | NM_007123 | - | - | - | - |
| | NM_206933 | 13178 13219 | 168 208 | MIR3 | CDS |
| USP12 | NM_182488 | - | - | - | - |
| | AF022789 | 4091 4330 | 25 262 | MIRb | 3’ |
| USP45 | XM_371838 | 366 517 | 8 170 | MIR | 3’ |
| USP49 | NM_018561 | 2572 2654 | 141 217 | MIR | 3’ |
| VAPA | NM_003574 | 3046 3206 | 179 21 | MIR3 | 3’ |
| | | 5963 6113 | 204 33 | MIRb | 3’ |
| | NM_194434 | 2911 3071 | 179 21 | MIR3 | 3’ |
| | | 5828 5978 | 204 33 | MIRb | 3’ |
| VASH1 | NM_014909 | - | - | - | - |
| | BC009031 | 1058 1129 | 109 180 | MIR3 | TAG |
| VASN | NM_138440 | 2295 2405 | 81 197 | MIR3 | 3’ |
| VDR | NM_000376 | 3337 3417 | 152 59 | MIR | 3’ |
| | | 4576 4643 | 273 210 | THER1_MD | 3’ |
| | NM_001017535 | 3459 3539 | 152 59 | MIR | 3’ |
| | | 4698 4765 | 273 210 | THER1_MD | 3’ |

| | | | | | |
|---|---|---|---|---|---|
| VISA | NM_020746 | 2505 2684 | 45 247 | MIRb | 3’ |
| | | 2745 2936 | 9 204 | MIR | 3’ |
| VPREB3 | NM_013378 | 464 529 | 93 161 | MIRb | 3’ |
| VPS24 | NM_016079 | 1782 1843 | 75 15 | MIR | 3’ |
| | NM_001005753 | 1721 1782 | 75 15 | MIR | 3’ |
| VTA1 | NM_016485 | 2760 2893 | 127 266 | MIR | 3’ |
| VTN | NM_000638 | 53 172 | 220 98 | MIRb | 5’ |
| WARS2 | NM_015836 | 1790 1947 | 217 57 | MIRb | 3’ |
| | NM_201263 | 1819 1976 | 217 57 | MIRb | 3’ |
| WASF2 | NM_006990 | 2547 2696 | 186 21 | MIR3 | 3’ |
| WDR25 | NM_024515 | 1832 1985 | 85 268 | MIRb | 3’ |
| WDR31 | NM_001012361 | - | - | - | - |
| | NM_001006615 | - | - | - | - |
| | NM_145241 | - | - | - | - |
| | BC012352 | 2586 2768 | 204 12 | MIRb | 3’ |
| WDR73 | NM_032856 | 1182 1360 | 7 206 | MIRb | 3’ |
| WDR82 | NM_025222 | 2168 2238 | 138 66 | MIR | 3’ |
| WFDC1 | NM_021197 | 941 1097 | 12 175 | MIR3 | 3’ |
| WISP1 | NM_003882 | 2494 2735 | 263 19 | MIRb | 3’ |
| | NM_080838 | - | - | - | - |
| WIT1 | NM_015855 | 340 387 | 151 104 | MIR | 5’ |
| | | 404 589 | 8 192 | MIR | 5’ |
| | | 643 708 | 98 32 | MIR | 5’ |
| WNT8A | NM_058244 | 1211 1281 | 103 34 | MIR | 3’ |
| XK | NM_021083 | 4372 4522 | 109 263 | MIR_Mars | 3’ |
| XKR5 | NM_207411 | 2282 2388 | 58 174 | MIR3 | 3’ |
| XPNPEP2 | NM_003399 | 2934 3139 | 250 65 | MIR_Mars | 3’ |
| XTP3TPA | NM_024096 | 980 1117 | 19 163 | MIR | 3’ |
| YIPF1 | NM_018982 | 67 183 | 134 2 | MIR | 5’ |
| YWHAZ | NM_003406 | 2750 2789 | 219 258 | MIR | 3’ |
| | NM_145690 | 2792 2831 | 219 258 | MIR | 3’ |
| ZAK | NM_016653 | - | - | - | - |
| | NM_133646 | 3252 3489 | 7 261 | MIR | 3’ |
| ZBTB44 | NM_014155 | 1640 1721 | 23 109 | MIR | TAG |
| ZBTB7B | NM_015872 | 3 137 | 107 257 | MIR | 5’ |
| ZC3H12C | XM_370654 | 496 678 | 3 186 | MIRb | 5’ |
| | XM_931631 | - | - | - | - |
| ZC3H13 | NM_015070 | 5333 5561 | 7 262 | MIR | 3’ |
| | | 6056 6181 | 73 205 | MIRm | 3’ |
| ZC3H7B | NM_017590 | 3660 3763 | 100 202 | MIR3 | 3’ |
| | | 3799 3937 | 195 45 | MIR3 | 3’ |
| ZCCHC24 | NM_153367 | 1736 1933 | 5 200 | MIR3 | 3’ |
| | | 2207 2260 | 123 189 | THER1_MD | 3’ |
| ZCCHC3 | NM_033089 | 1922 2110 | 208 16 | MIR3 | 3’ |
| ZCCHC4 | NM_024936 | 2652 2792 | 66 197 | MIRb | 3’ |
| ZFAND5 | NM_006007 | - | - | - | - |
| | BC027707 | 506 639 | 181 33 | MIR3 | 5’ |
| ZFAT1 | NM_001029939 | 2907 3017 | 143 255 | MIR | 3’ |
| | | 3423 3536 | 11 124 | MIRb | 3’ |
| | | 4061 4169 | 47 156 | MIRb | 3’ |
| | | 4161 4249 | 185 274 | THER1_MD | 3’ |
| | | 4729 4863 | 191 56 | MIRb | 3’ |
| | NM_020863 | - | - | - | - |
| ZFP2 | NM_030613 | 2227 2388 | 32 205 | MIR3 | 3’ |
| ZFP28 | NM_020828 | 2827 2891 | 188 252 | MIRb | 3’ |
| ZFP82 | NM_133466 | 2050 2142 | 65 162 | MIRm | 3’ |
| ZFP90 | NM_133458 | 2682 2831 | 165 19 | MIR3 | 3’ |

| | | | | | |
|---|---|---|---|---|---|
| ZFP91-CNTF | NM_053023<br>NM_170768 | -<br>2726 2954 | -<br>268 20 | -<br>MIRb | -<br>3’ |
| ZNF10 | NM_015394 | 2973 3211<br>3976 4121 | 12 262<br>113 262 | MIR<br>MIR | 3’<br>3’ |
| ZNF142 | NM_001105537<br>NM_005081 | 169 225<br>265 414 | 135 84<br>237 83 | MIR<br>MIR | 5’<br>5’ |
| ZNF184 | NM_007149 | 2667 2782 | 50 172 | MIR3 | 3’ |
| ZNF185 | NM_007150 | 2216 2348<br>2408 2452 | 260 114<br>262 218 | MIRb<br>MIRb | 3’<br>3’ |
| ZNF192 | NM_006298 | 4850 4897<br>5289 5360 | 262 215<br>165 92 | MIR<br>MIRb | 3’<br>3’ |
| ZNF193 | NM_006299<br>AY261373 | -<br>938 1075<br>1450 1541 | -<br>27 171<br>172 267 | -<br>MIRb<br>MIRb | -<br>3’<br>3’ |
| ZNF195 | NM_007152<br>BX537525 | -<br>3514 3646<br>3706 37 | -<br>260 114<br>262 218 | -<br>MIRb<br>MIRb | -<br>3’<br>3’ |
| ZNF197 | NM_006991<br>NM_001024855 | 4120 4301<br>- | 25 216<br>- | MIRb<br>- | 3’<br>- |
| ZNF2 | NM_021088<br>NM_001017396 | 2212 2460<br>2140 2388 | 260 8<br>260 8 | MIR<br>MIR | 3’<br>3’ |
| ZNF248 | NM_021045 | 18 80 | 126 188 | THER1_MD | 5’ |
| ZNF25 | NM_145011 | 2769 2993 | 2 261 | MIR | 3’ |
| ZNF275 | NM_001080485 | 2625 2781<br>2814 2877 | 107 268<br>82 19 | MIR<br>MIR | 3’<br>3’ |
| ZNF300 | NM_052860 | 2713 2827 | 26 145 | MIRb | 3’ |
| ZNF323 | NM_030899<br>NM_145909 | 2421 2585<br>2784 2996 | 16 193<br>16 262 | MIR<br>MIR | 3’<br>3’ |
| ZNF324 | NM_014347 | 2368 2495 | 270 114 | MIR | 3’ |
| ZNF333 | NM_032433 | 3796 3987 | 262 67 | MIR | 3’ |
| ZNF341 | NM_032819 | 3055 3302 | 2 259 | MIRb | 3’ |
| ZNF354C | NM_014594<br>BC063312 | -<br>2870 2905 | -<br>77 43 | -<br>MIR | -<br>3’ |
| ZNF365 | NM_014951<br>NM_199450<br>NM_199451<br><br>NM_199452<br> | -<br>-<br>2168 2377<br>2937 3166<br>1487 1696<br>2256 2485 | -<br>-<br>230 14<br>272 7<br>230 14<br>272 7 | -<br>-<br>MIRb<br>MIRb<br>MIRb<br>MIRb | -<br>-<br>3’<br>3’<br>3’<br>3’ |
| ZNF398 | NM_170686<br>NM_020781 | 3531 3617<br>3210 3296 | 67 154<br>67 154 | MIR3<br>MIR3 | 3’<br>3’ |
| ZNF409 | XM_375065 | 2902 2975 | 133 201 | MIR3 | TAG |
| ZNF420 | NM_144689 | 2548 2744 | 58 262 | MIR | 3’ |
| ZNF445 | NM_181489 | 5462 5650<br>7920 8099<br>8409 8489 | 1 172<br>81 274<br>72 153 | MIRb<br>MIR<br>MIRb | 3’<br>3’<br>3’ |
| ZNF45 | NM_003425 | 165 212<br>262 306 | 110 157<br>120 75 | MIRb<br>MIR | 5’<br>5’ |
| ZNF470 | NM_001001668 | 515 607<br>3044 3125 | 136 29<br>176 259 | MIR<br>MIRb | 5’<br>3’ |
| ZNF471 | NM_020813 | 4308 4381 | 184 258 | MIRb | 3’ |
| ZNF473 | NM_015428<br>NM_001006656 | 3145 3378<br>3023 3256 | 4 251<br>4 251 | MIR<br>MIR | 3’<br>3’ |
| ZNF483 | NM_133464<br>NM_001007169 | 2698 2842<br>1057 1233 | 5 189<br>20 167 | MIRb<br>MIR3 | 3’<br>3’ |
| ZNF488 | NM_153034 | 1457 1672 | 31 252 | MIRb | 3’ |

| | | | | | |
|---|---|---|---|---|---|
| | | 2821 2965 | 271 128 | THER1_MD | 3’ |
| ZNF498 | NM_145115<br>BX640720 | -<br>1998 2074 | -<br>167 91 | -<br>MIRb | -<br>3’ |
| ZNF501 | NM_145044 | 1364 1489<br>2272 2437 | 55 196<br>223 51 | MIRb<br>MIRb | 3’<br>3’ |
| ZNF514 | NM_032788 | 2605 2677 | 163 88 | MIR3 | 3’ |
| ZNF526 | NM_133444 | 3738 3905 | 70 256 | MIR | 3’ |
| ZNF540 | NM_152606 | 2542 2644 | 65 182 | MIR | 3’ |
| ZNF544 | NM_014480 | 79 201 | 180 71 | MIR | 5’ |
| ZNF546 | NM_178544 | 2262 2447<br>3109 3216 | 244 43<br>249 131 | MIRb<br>MIR | TAG<br>3’ |
| ZNF566 | NM_032838 | 2248 2378 | 134 262 | MIR | 3’ |
| ZNF580 | NM_016202<br>NM_207115 | 1190 1280<br>903 993 | 34 136<br>34 136 | MIRm<br>MIRm | 3’<br>3’ |
| ZNF583 | NM_152478 | 2272 2470 | 42 248 | MIRb | 3’ |
| ZNF597 | NM_152457 | 1517 1636 | 245 133 | MIRb | 3’ |
| ZNF606 | NM_025027 | 3930 4135 | 232 2 | MIR | 3’ |
| ZNF607 | NM_032689 | 3159 3272 | 66 195 | MIR | 3’ |
| ZNF639 | NM_016331<br>AB097862 | 613 745<br>613 745 | 141 2<br>141 2 | MIRm<br>MIRm | CDS<br>CDS |
| ZNF660 | NM_173658 | 1874 2004 | 18 148 | MIRb | 3’ |
| ZNF662 | NM_207404 | 1812 2039 | 266 25 | MIRb | 3’ |
| ZNF663 | NM_173643 | 604 738 | 118 253 | MIR | 5’ |
| ZNF664 | NM_152437 | 4545 4684<br>4763 4948 | 47 207<br>19 202 | MIR3<br>MIR3 | 3’<br>3’ |
| ZNF689 | NM_138447 | 1821 1981 | 180 1 | MIR3 | 3’ |
| ZNF70 | NM_021916 | 2684 2915 | 6 221 | MIRb | 3’ |
| ZNF75D | NM_007131 | 1044 1302 | 2 266 | MIRb | 5’ |
| ZNF767 | NM_024910 | 574 657 | 162 83 | MIRc | TAG |
| ZNF781 | NM_152605 | 3040 3074 | 166 200 | MIR | 3’ |
| ZNRF1 | NM_032268 | 3265 3400 | 39 169 | MIRm | 3’ |
| ZSCAN2 | NM_181877<br>NM_017894<br>NM_001007072 | 2263 2382<br>-<br>- | 23 152<br>-<br>- | MIR<br>-<br>- | 3’<br>-<br>- |
| ZSCAN20 | NM_145238 | 3610 3845 | 259 2 | MIR | 3’ |
| ZSCAN22 | NM_181846 | 4061 4212<br>4424 4580 | 9 170<br>253 61 | MIRb<br>MIRb | 3’<br>3’ |
| ZSCAN23 | NM_001012455 | 2393 2567 | 194 8 | MIR3 | 3’ |
| ZXDA | NM_007156 | 2716 2859 | 17 154 | MIR3 | 3’ |
| ZYG11A | NM_001004339 | 3570 3681 | 241 131 | THER1_MD | 3’ |

## 9.2. Hypothetical or unknown genes which have recruited MIR elements.

The gene symbol and accession numbers are listed, 216 hypothetical proteins in total.

| Gene Name | Accession Number | Gene Name | Accession Number |
|---|---|---|---|
| FLJ42957 | NM_207436 | LOC388823 | XM_373931 |
| FLJ90680 | NM_207475 | LOC388890 | XM_373953 |
| KIAA0467 | NM_015284 | LOC388893 | XM_373954 |
| LOC143188 | XM_378251<br>NM_001025447 | LOC389025 | XM_374004 |
| LOC146795 | XM_378701 | LOC389147 | XM_371663 |
| LOC199800 | XM_373810 | LOC389149 | XM_374051 |
| LOC219908 | XM_169057 | LOC389150 | XM_374052 |
| LOC285422 | XM_379280 | LOC389237 | XM_374094 |
| LOC285711 | XM_211988 | LOC389345 | XM_374150 |
| LOC286208 | XM_379668 | LOC389683 | XM_374277 |
| LOC387940 | XM_373571 | LOC390217 | XM_372420 |
| LOC388276 | XM_373685 | LOC399715 | XM_374766 |
| LOC389166 | XM_374060 | LOC399786 | XM_378236 |
| LOC389791 | XM_372142<br>NM_001013652 | LOC399875 | XM_378279 |
| LOC400551 | XM_378621 | LOC399919 | XM_378299 |
| LOC400605 | XM_378685 | LOC399959 | XM_378316 |
| LOC400614 | XM_378693 | LOC400004 | XM_378340 |
| LOC400651 | XM_378747 | LOC400084 | XM_378389 |
| LOC401057 | XM_379181 | LOC400144 | XM_378421 |
| LOC401280 | XM_376549<br>NM_001013682 | LOC400320 | XM_375163 |
| LOC401280 | XM_376560 | LOC400400 | XM_378529 |
| LOC401553 | XM_379671 | LOC400496 | XM_378562 |
| LOC441242 | BC066990 | LOC400552 | XM_378623 |
| MGC15885 | XM_378517 | LOC400553 | XM_378625 |
| PRO0195 | AF090901 | LOC400572 | XM_378648 |
| FLJ10489 | XM_379597 | LOC400577 | XM_378653 |
| FLJ13439 | XM_378280 | LOC400581 | XM_375418 |
| FLJ20674 | BC034471 | LOC400586 | XM_375424 |
| FLJ21408 | XM_378573 | LOC400655 | XM_378753 |
| FLJ25076 | XM_059689 | LOC400743 | XM_378843 |
| FLJ26484 | XM_378703 | LOC400756 | XM_378865 |
| FLJ27365 | NM_207477 | LOC400796 | XM_378917 |
| FLJ31183 | XM_379381 | LOC400813 | XM_378947 |
| FLJ34048 | XM_379510 | LOC400831 | XM_378956 |

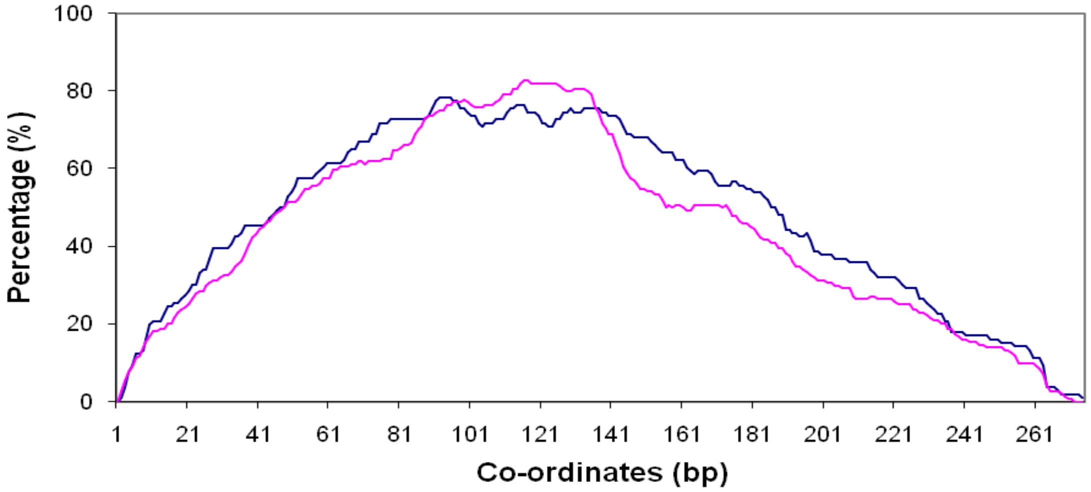
| | | | |
|---|---|---|---|
| FLJ34515 | XM_378620 | LOC400847 | XM_378982 |
| FLJ35700 | NM_205851 | LOC400880 | XM_379030 |
| FLJ39051 | XM_378321 | LOC400942 | XM_379075 |
| FLJ40852 | NM_173677 | LOC400960 | XM_379102 |
| FLJ41046 | NM_207479 | LOC400964 | XM_379106 |
| FLJ41278 | XM_378362 | LOC401005 | XM_379135 |
| FLJ41350 | XM_378247 | LOC401006 | XM_379136 |
| FLJ41481 | XM_379078 | LOC401032 | XM_379156 |
| FLJ41562 | XM_376320 | LOC401062 | XM_379190 |
| FLJ42280 | NM_207503 | LOC401165 | XM_379299 |
| FLJ43903 | XM_379595 | LOC401212 | XM_379363 |
| FLJ43963 | XM_376334 | LOC401234 | XM_379395 |
| FLJ44076 | NM_207486 | LOC401256 | XM_379409 |
| FLJ44255 | XM_378741 | LOC401324 | XM_380100 |
| FLJ44385 | NM_207478 | LOC401456 | XM_379562 |
| FLJ44674 | NM_207449 | LOC401477 | XM_379605 |
| FLJ44817 | XM_375090 | LOC401480 | XM_379608 |
| FLJ45248 | NM_207505 | LOC401490 | XM_379619 |
| FLJ45721 | NM_207490 | LOC401530 | XM_379643 |
| FLJ45872 | XM_376795 | LOC439951 | XM_374769 |
| FLJ46320 | XM_498649 | LOC440791 | XM_498870 |
| FLJ46836 | NM_207509 | LOC440894 | XM_379121 |
| KIAA0040 | NM_014656 | LOC441046 | BC025996 |
| KIAA0087 | XM_376586 | LOC51145 | AF159054 |
| KIAA0125 | NM_014792 | LOC51215 | AF113674 |
| KIAA0141 | BC007855 | LOC51216 | AF113685 |
| KIAA0317 | AB002315 | LOC642863 | AY358216 |
| KIAA0319L | AF275679 | LOC643738 | AY358772 |
| KIAA0492 | XM_378914 | LOC645200 | AB088847 |
| KIAA0748 | XM_374983 | LOC645238 | AY358248 |
| KIAA1199 | NM_018689 | LOC87600 | AF251048 |
| KIAA1239 | XM_049078 | LOC87623 | XM_039762 |
| KIAA1257 | XM_371664 | MGC21881 | NM_203448 |
| KIAA1274 | NM_014431 | PRO0386 | AF116603 |
| KIAA1305 | XM_370756 | PRO0514 | AF090933 |
| KIAA1644 | XM_376018 | PRO0767 | AF113012 |
| KIAA1715 | NM_030650 | PRO0902 | AF305821 |
| KIAA1853 | NM_194286 | PRO0943 | AF116611 |
| KIAA1881 | XM_375558 | PRO1097 | AF119844 |
| LOC114137 | AF305822 | PRO1386 | AF118062 |
| LOC119893 | AF227517 | PRO1483 | AF116635 |

| | | | |
|---|---|---|---|
| LOC142937 | BC008131 | PRO1496 | AF116665 |
| LOC144678 | XM_378390 | PRO1808 | AF118077 |
| LOC145216 | XM_378487 | PRO2832 | AF119902 |
| LOC147977 | AF187554 | PRO2949 | AF119907 |
| LOC149134 | NM_207326 | PRO3073 | AF119912 |
| LOC150568 | XM_379117 | LOC387943 | XM_373574 |
| LOC150577 | XM_379114 | LOC388073 | XM_373616 |
| LOC155036 | XM_376722 | LOC388418 | XM_373748 |
| LOC170425 | XM_378240 | LOC388641 | XM_373848 |
| LOC199899 | XM_378866 | LOC388813 | XM_373925 |
| LOC201229 | XM_375430 | LOC339022 | XM_294775 |
| LOC220799 | AF320070 | LOC339442 | XM_378855 |
| LOC221140 | XM_167908 | LOC339468 | XM_373847 |
| LOC222701 | XM_167152 | LOC339524 | NM_207357 |
| LOC254808 | XM_374069 | LOC339807 | XM_379099 |
| LOC255177 | XM_378516 | LOC339809 | XM_291020 |
| LOC256176 | XM_172889 | LOC339902 | XM_295097 |
| LOC283034 | XM_210860 | LOC347475 | XM_298045 |
| LOC283143 | XM_378312 | LOC387647 | XM_373451 |
| LOC283177 | XM_378327 | LOC387693 | BC071736 |
| LOC283214 | XM_378303 | LOC387824 | XM_373520 |
| LOC283332 | XM_378355 | LOC387826 | XM_373521 |
| LOC283403 | XM_211028 | LOC387888 | XM_373550 |
| LOC283432 | XM_378379 | LOC387927 | XM_370726 |
| LOC283663 | XM_378514 | LOC284688 | XM_378912 |
| LOC283682 | XM_378544 | LOC284798 | XM_378971 |
| LOC283761 | XM_378542 | LOC284898 | XM_379044 |
| LOC283854 | XM_378599 | LOC284931 | XM_211694 |
| LOC284100 | XM_375443 | LOC285045 | XM_379086 |
| LOC284551 | XM_375713 | LOC285418 | AF318325 |
| LOC285547 | XM_379258 | LOC285484 | XM_376303 |
| LOC286149 | XM_379594 | | |

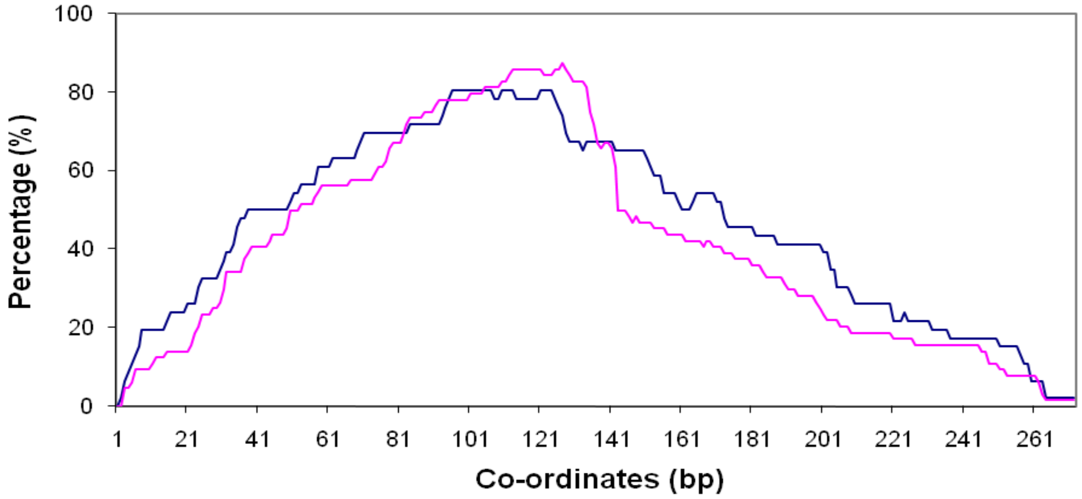
## 9.3. Conservation of human MIR elements

The number of MIR elements which corresponds to each nucleotide of the generic consensus sequence has been plotted for all of the MIR sub-families located within the 3’-UTR (A) coding sequence (B) and 3’-UTR (C). MIR elements in the direct orientation are blue and inverse pink.

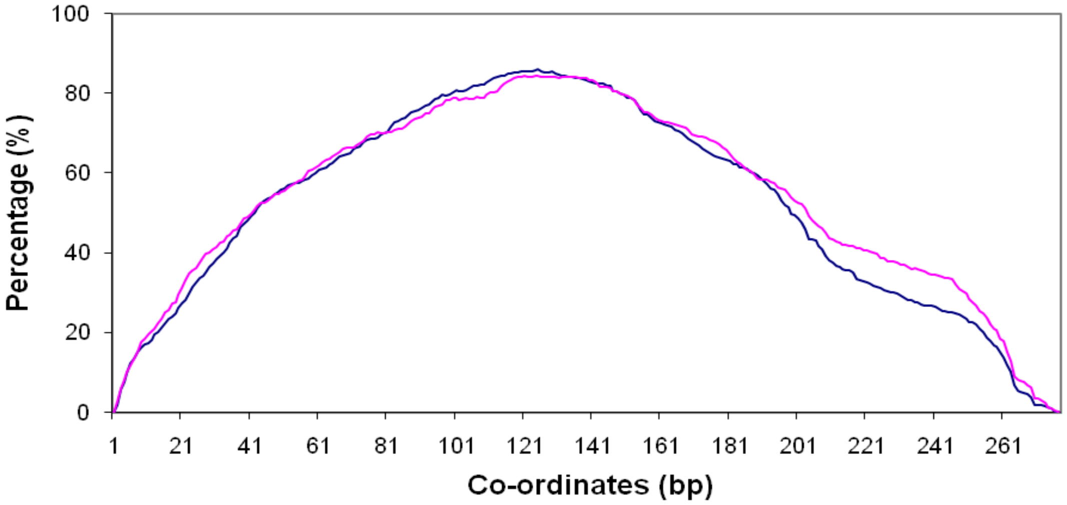
### A) 5’-UTR



### B) CDS



### C) 3’-UTR



## 9.4. Accession numbers for DNA sequences included in multiple sequence alignments

All accession numbers are obtainable for the sequence database at NCBI. Accession numbers for sequence data included in multiple sequence alignments. The sequence types are obtained following BLASTn searches of sequence databases (Db) of either genomic DNA (gDNA; wgs, ggs, refseq_genomic), mRNA (refsq_rna, nr/nt) or expressed sequence tags (est). Query sequences (source) were the generic MIR consensus sequence or from specific MIR-containing genes. Source gene HGNC symbols have been listed.

| Fig. | Page | Query Source | Accession No. | Db | Sequence Type | Species |
|---|---|---|---|---|---|---|
| 3.3 | 50 | MIR consensus | AC013758<br>167249-135967 | wgs | gDNA | *H.sapiens* |
| 3.3 | 50 | MIR consensus | NC_006479<br>37834932-37834867 | wgs | gDNA | *P.troglodytes* |
| 3.3 | 50 | MIR consensus | DAAA02013700<br>17640-17575 | wgs | gDNA | *B.taurus* |
| 3.3 | 50 | MIR consensus | AAEX02022366<br>135713-135776 | wgs | gDNA | *C.familiaris* |
| 3.3 | 50 | MIR consensus | NC_010459<br>34977720-34977655 | wgs | gDNA | *S.scrofa* |
| 3.3 | 50 | MIR consensus | AAHY01125103<br>21760-21695 | wgs | gDNA | *M.musculus* |
| 3.3 | 50 | MIR consensus | ABQO010260434<br>319-254 | wgs | gDNA | *M.eugenii* |
| 3.3 | 50 | MIR consensus | AC234489<br>37870-37935 | ggs | gDNA | *T. aculeatus* |
| 3.3 | 50 | MIR consensus | AAPN01000352<br>15020-14955 | wgs | gDNA | *O.anatinus* |
| 3.3 | 50 | MIR consensus | NM_000618<br>1910-1975 | refseq_ rna | mRNA | *H.sapiens* |
| 3.3 | 50 | MIR consensus | XM_001141985<br>10817-10882 | refseq_ rna | mRNA | *P.troglodytes* |
| 3.3 | 50 | MIR consensus | BC148869<br>3541-3605 | nr/nt | mRNA | *B.taurus* |
| 3.3 | 50 | MIR consensus | DT537928<br>73-138 | est | mRNA | *C.familiaris* |
| 3.3 | 50 | MIR consensus | DN106777<br>155-218 | est | mRNA | *S.scrofa* |
| 3.3 | 50 | MIR consensus | AK160844<br>2516-2581 | nr/nt | mRNA | *M.musculus* |
| 3.3 | 50 | MIR consensus | EC321646<br>319-383 | est | mRNA | *T.vulpecula* |
| 3.3 | 50 | MIR consensus | EH002999<br>160-125 | est | mRNA | *O.anatinus* |
| 5.2 | 108 | KLC1 | NC_000014.8<br>104095525-104167888 | refseq_ genomic | gDNA | *H.sapiens* |
| 5.2 | 108 | CHRD | NC_000003.11<br>184097861-184107617 | refseq_ genomic | gDNA | *H.sapiens* |
| 5.3 | 109 | KLC1 | NT_026437.12<br>85151305-85151379 | refseq_ genomic | gDNA | *H.sapiens* |

| Fig. | Page | Query Source | Accession No. | Db | Sequence Type | Species |
|---|---|---|---|---|---|---|
| 5.3 | 109 | KLC1 | NC_006481.2<br>104177189-104177263 | refseq_genomic | gDNA | *P.troglodytes* |
| 5.3 | 109 | KLC1 | NC_009167.2<br>45145894-45145954 | refseq_genomic | gDNA | *E. caballus* |
| 5.3 | 109 | KLC1 | NC_006590.2<br>74478371-74478445 | refseq_genomic | gDNA | *C.familiaris* |
| 5.3 | 109 | KLC1 | NC_007319.3<br>68615799-68615873 | refseq_genomic | gDNA | *B.taurus* |
| 5.3 | 109 | KLC1 | NC_005105.2<br>136585194-136585266 | refseq_genomic | gDNA | *R.norvegicus* |
| 5.3 | 109 | KLC1 | NT_166318.1<br>23984950-23985010 | refseq_genomic | gDNA | *M.musculus* |
| 5.3 | 109 | KLC1 | NC_008801.1<br>312097480-312097406 | refseq_genomic | gDNA | *M.domestica* |
| 5.3 | 109 | KLC1 | ABQO010977032.1<br>2115-2041 | wgs | gDNA | *M.eugenii* |
| 5.3 | 109 | KLC1 | NC_009094.1<br>22698893-22698965 | refseq_genomic | gDNA | *O.anatinus* |
| 5.3 | 109 | CHRD | NC_000003.11<br>184098853-184098913 | refseq_genomic | gDNA | *H.sapiens* |
| 5.3 | 109 | CHRD | NC_006490.2<br>189910344-189910420 | refseq_genomic | gDNA | *P. troglodytes* |
| 5.3 | 109 | CHRD | NC_007859.1<br>102282539-102282464 | refseq_genomic | gDNA | *M.mulatta* |
| 5.3 | 109 | CHRD | NC_007299.3<br>84687537-84687462 | refseq_genomic | gDNA | *B.taurus* |
| 5.3 | 109 | CHRD | NC_006616.2<br>20263690-20263765 | refseq_genomic | gDNA | *C.familiaris* |
| 5.3 | 109 | CHRD | NC_000082.5<br>20733929-20734004 | refseq_genomic | gDNA | *M.musculus* |
| 5.3 | 109 | CHRD | AC_000079.1<br>79019711-79019636 | refseq_genomic | gDNA | *R.norvegicus* |
| 5.8 | 116 | AHI1 | NT_025741.15<br>39878560-39878488 | refseq_genomic | gDNA | *H.sapiens* |
| 5.8 | 116 | AHI1 | NC_006473.2<br>137724940-137724867 | refseq_genomic | gDNA | *P. troglodytes* |
| 5.8 | 116 | AHI1 | NC_007861.1<br>128447624-128447697 | refseq_genomic | gDNA | *M.mulatta* |
| 5.8 | 116 | AHI1 | NC_009153.2<br>81120331-81120256 | refseq_genomic | gDNA | *E. caballus* |
| 5.8 | 116 | AHI1 | NC_007307.3<br>75649837-75649757 | refseq_genomic | gDNA | *B.taurus* |
| 5.8 | 116 | CIITA | NG_009628.1<br>36039-36122 | refseq_genomic | gDNA | *H.sapiens* |
| 5.8 | 116 | CIITA | NC_006483.2<br>11232769-11232829 | refseq_genomic | gDNA | *P. troglodytes* |
| 5.8 | 116 | CIITA | NC_007877.1<br>10978663-10978747 | refseq_genomic | gDNA | *M.mulatta* |
| 5.8 | 116 | CIITA | NC_009156.2<br>33510196-33510136 | refseq_genomic | gDNA | *E. caballus* |

| Fig. | Page | Query Source | Accession No. | Db | Sequence Type | Species |
|---|---|---|---|---|---|---|
| 5.8 | 116 | CIITA | NC_000082.5<br>10512495-10512557 | refseq_<br>genomic | gDNA | *M.musculus* |
| 5.8 | 116 | CIITA | NC_005109.2<br>5101848-5101788 | refseq_<br>genomic | gDNA | *R.norvegicus* |
| 5.8 | 116 | GSG1L | NC_000016.9<br>27835248-27835188 | refseq_<br>genomic | gDNA | *H.sapiens* |
| 5.8 | 116 | GSG1L | NC_006483.2<br>28256956-28256896 | refseq_<br>genomic | gDNA | *P. troglodytes* |
| 5.8 | 116 | GSG1L | NC_007877.1<br>26274918-26274858 | refseq_<br>genomic | gDNA | *M.mulatta* |
| 5.8 | 116 | GSG1L | NC_006588.2<br>21891470-21891541 | refseq_<br>genomic | gDNA | *C.familiaris* |
| 5.8 | 116 | GSG1L | NC_000073.5<br>133043735-133043664 | refseq_<br>genomic | gDNA | *M.musculus* |
| 5.8 | 116 | GSG1L | AC_000069.1<br>178131185-178131115 | wgs | gDNA | *R.norvegicus* |
| 6.7 | 141 | TGM2 | NC_000020.10<br>36761415-36761327 | refseq_<br>genomic | gDNA | *H.sapiens* |
| 6.7 | 141 | TGM2 | NC_007867.1<br>26327686-26327773 | refseq_<br>genomic | gDNA | *M.mulatta* |
| 6.9 | 143 | TGM2 | NC_000020.10<br>36759606-36758589 | refseq_<br>genomic | gDNA | *H.sapiens* |
| 6.9 | 143 | TGM2 | NC_006487.2<br>35312705-35311170 | refseq_<br>genomic | gDNA | *P. troglodytes* |
| 6.9 | 143 | TGM2 | NC_007867.1<br>26329975-26328958 | refseq_<br>genomic | gDNA | *M.mulatta* |
| 6.9 | 143 | TGM2 | NC_007311.3<br>67599512-67598021 | refseq_<br>genomic | gDNA | *B.taurus* |
| 6.9 | 143 | TGM2 | NC_005102.2<br>148835362-148833769 | refseq_<br>genomic | gDNA | *R.norvegicus* |
| 6.9 | 143 | TGM2 | NC_000068.6<br>157944694-157943047 | refseq_<br>genomic | gDNA | *M.musculus* |

## 9.5. MIR-containing genes clustered on chromosomes 11 and 9.

| Gene symbol | Position | Gene symbol | Position |
|---|---|---|---|
| SLC1A2 | 11p13-p12 | C9orf58 | 9q34.13-q34.3 |
| RAG1 | 11p12 | C9orf90 | 9q34.11 |
| TP53I11 | 11p11.2 | C9orf98 | 9q34.13 |
| SLC43A3 | 11q12.1 | DBH | 9q34.2 |
| ZFP91 | 11q12.1 | DPM2 | 9q34.11-13 |
| CNTF | 11q12.1 | EHMT1 | 9q34.3 |
| FAM111A | 11q12.1 | FREQ | 9q34.11 |
| MS4A3 | 11q12.1 | GFI1B | 9q34.2-13 |
| MS4A6A | 11q12.1 | NEK6 | 9q33.3-q34.11 |
| MS4A10 | 11q12.2 | PIP5KL1 | 9q34.11 |
| GPR44 | 11q12.2 | ST6GALNAC6 | 9q34.11 |
| SYT7 | 11q12.2 | TOR1A | 9q34.11 |
| BATF2 | 11q13.1 | TOR2A | 9q34.11 |
| MRPL49 | 11q13.1 | | |
| DKFZp761E198 | 11q13.1 | | |
| C11orf68 | 11q13.1 | | |
| SLC29A2 | 11q13.2 | | |
| PELI3 | 11q13.2 | | |
| ALDH3B2 | 11q13.2 | | |
| ALDH3B1 | 11q13.2 | | |
| SAPS3 | 11q13.2 | | |
| MYEOV | 11q13.3 | | |
| ORAOV1 | 11q13.3 | | |
| FADD | 11q13.3 | | |

## 9.6. MIR-containing genes involved in regulatory pathways

The pathway name and MIR-containing gene symbol is listed. The total number of genes known to participate in each pathway is noted including the frequency (Freq) of MIR-containing genes involved in each pathway.

| Pathway name | MIR-containing gene | Total | Freq. |
|---|---|---|---|
| Glycolysis / Gluconeogenesis | ACAS2L, ACSS1, ALDH3B1, ALDH3B2, AMPD3, DLD, HK1, HK2, IL21R | 65 | 0.138 |
| Inositol phosphate metabolism | BDKRB2, CASR, CCR5, INPP5D, PIP5KL1, PLCB2, PLCB4 | 52 | 0.135 |
| Propanoate metabolism | ACADSB, ACAS2L, ACAT1, ACSS1, DHCR24 | 39 | 0.128 |
| Stilbene, coumarine and lignin biosynthesis | MPO | 8 | 0.125 |
| Pyruvate metabolism | ACAS2L, ACAT1, ACSS1, DLD, IL21R | 43 | 0.116 |
| Galactose metabolism | B4GALT2, HK1, HK2 | 32 | 0.094 |
| Fructose and mannose metabolism | B4GALT2, HK1, HK2, KHK | 48 | 0.083 |
| Citrate cycle | DLD, CLYBL | 28 | 0.071 |
| Aminosugar metabolism | HK1, HK2 | 30 | 0.067 |
| Starch and sucrose metabolism | DDX52, FBXO18, HK1, HK2 | 78 | 0.051 |
| Pentose and glucuronate interconversions | RPE | 24 | 0.042 |
| Pentose phosphate pathway | RPE | 27 | 0.037 |
| Butanoate metabolism | ACAT1 | 44 | 0.023 |
| Sulfur metabolism | PAPSS2, SUOX | 14 | 0.143 |
| ATP synthesis | ATP6V1G2 ATP6V1B1, ATP6V1E2, ATP6VOE2L | 40 | 0.1 |
| Methane metabolism | MPO | 10 | 0.1 |
| Reductive carboxylate cycle CO2 fixation) | ACSS1 | 11 | 0.091 |
| Oxidative phosphorylation | ATP6V0E2L, ATP6V1B1, ATP6V1E2, ATP6V1G2, ATP7B, COX15, MEGF11, NDUFB2, NDUFS1, PPA2 | 130 | 0.077 |
| Nitrogen metabolism | CA8 | 24 | 0.042 |
| Carbon fixation | RPE | 25 | 0.04 |
| Fatty acid biosynthesis (path 1) | MCCC2 | 8 | 0.125 |
| Synthesis and degradation of ketone bodies | ACAT1 | 8 | 0.125 |
| Glycerolipid metabolism | AGPAT6, B4GALT2, DGKA, LIPG, PLD1, PLA2G3, YWHAZ | 60 | 0.117 |
| Biosynthesis of steroids | DHCR7 | 18 | 0.056 |
| Fatty acid biosynthesis (path 2) | ACAT1 | 18 | 0.056 |
| Glycerophospholipid metabolism | AGPAT6, DGKA, PLA2G3, PLD1 | 77 | 0.052 |
| Fatty acid metabolism | ACADSB, ACAT1 | 51 | 0.039 |
| Linoleic acid metabolism | PLA2G3 | 35 | 0.029 |
| Sphingolipid metabolism | NEU3 | 38 | 0.026 |
| Bile acid biosynthesis | ACADSB | 40 | 0.025 |
| Androgen and estrogen metabolism | HEMK1 | 53 | 0.019 |
| Arachidonic acid metabolism | GPX7 | 57 | 0.018 |
| Purine metabolism | ADCY1, ADCY2, ADCY6, AK2, AMPD3, | 155 | 0.123 |

| | | | |
|---|---|---|---|
| | ATP6V1G2, MPP3, NT5E, PAPSS2, PDE11A, PDE2A, PDE4A, PDE4C, POLG, POLH, POLQ, POLR3D, RECQL5, RRM2B | | |
| Pyrimidine metabolism | EHD4, NT5E, POLG, POLH, POLQ, POLR3D, RRM2B, TK2, TXNDC2 | 93 | 0.097 |
| Phenylalanine metabolism | ALDH3B1, ALDH3B2, MAOB, MPO | 30 | 0.133 |
| Methionine metabolism | AHCYL1, MAT1A | 16 | 0.125 |
| Alanine and aspartate metabolism | ASL, DARS2, IL1A | 25 | 0.12 |
| Histidine metabolism | ALDH3B1, ALDH3B2, HEMK1, MAOB | 41 | 0.098 |
| Tyrosine metabolism | ALDH3B1, ALDH3B2, DBH, HEMK1, MAOB | 60 | 0.083 |
| Arginine and proline metabolism | ADC, ASL, MAOB, PARS2 | 58 | 0.069 |
| Glutamate metabolism | ADC, GCLC | 31 | 0.065 |
| Lysine degradation | ACAT1, EHMT1, OTUB2 | 52 | 0.058 |
| Tryptophan metabolism | ACAT1, DHCR24, HEMK1, MAOB, WARS2 | 86 | 0.058 |
| Valine, leucine and isoleucine degradation | ACADSB, ACAT1, MCCC2 | 54 | 0.056 |
| Glycine, serine and threonine metabolism | DLD, MAOB | 45 | 0.044 |
| Urea cycle and metabolism of amino groups | ASL | 26 | 0.038 |
| Selenoamino acid metabolism | AHCYL1, HEMK1, MAT1A, PAPSS2 | 34 | 0.118 |
| Aminophosphonate metabolism | HEMK1 | 20 | 0.05 |
| Glutathione metabolism | GCLC, GPX7 | 41 | 0.049 |
| beta-alanine metabolism | ACADSB | 27 | 0.037 |
| Keratan sulfate biosynthesis | B3GNT6, B4GALT2, FUT8, ST3GAL1, ST6GAL1 | 15 | 0.333 |
| O-Glycan biosynthesis | B4GALT2, GALNT4, GALNT10, GALNT11, ST3GAL1, ST6GALNAC6 | 25 | 0.24 |
| N-Glycan biosynthesis | ALG2, B4GALT2, FUT8, MAN2A2, MGAT5, NEU3, ST6GAL1 | 37 | 0.189 |
| Glycosylphosphatidylinositol(GPI)-anchor biosynthesis | B4GALT2, PIGS, PIGL | 23 | 0.13 |
| N-Glycan degradation | NEU3 | 16 | 0.063 |
| Glycosaminoglycan degradation | HYAL1 | 17 | 0.059 |
| Glycosphingolipid metabolism | B4GALT2, NEU3 | 43 | 0.047 |
| Glycan structures - biosynthesis 1 | GALNT4 | 107 | 0.009 |
| Thiamine metabolism | TPK1 | 6 | 0.167 |
| Ubiquinne biosynthesis | NDUFB2, NDUFS1 | 15 | 0.133 |
| Boitin metabolism | OTUB2 | 8 | 0.125 |
| Nicotinate and nicotinamide metabolism | NMNAT2, NT5E, NUDT12, QPRT | 43 | 0.093 |
| One carbon pool by folate | EHD4 | 16 | 0.063 |
| Folate biosynthesis | DDX52, FBXO18 | 37 | 0.054 |
| Porphyrin and chlorophyll metabolism | ALAD | 36 | 0.028 |
| 1,1,1-Trichloro-2,2-bis(4-chlorophenyl)ethane | DHCR24 | 3 | 0.333 |
| Nitrobenzene degradation | HEMK1 | 15 | 0.067 |
| 1- and 2-Methylnaphthalene degradation | ACADSB | 27 | 0.037 |
| Benzoate degradation via CoA | ACAT1 | 31 | 0.032 |

| | | | |
|---|---|---|---|
| ligation | | | |
| RNA polymerase | POLR3D | 25 | 0.04 |
| Basal transcription factors | TAF9B | 36 | 0.028 |
| Aminoacyl-tRNA synthetases | DARS2, WARS2 | 32 | 0.063 |
| Ribosome | RPL28 | 131 | 0.008 |
| Proteasome | PSMD12 | 31 | 0.032 |
| Protein export | SEC61A1 | 13 | 0.077 |
| Ubiquitin mediated proteolysis | CUL3, CUL1, NEDD4 | 45 | 0.067 |
| DNA polymerase | POLG, POLH, POLQ | 25 | 0.12 |
| ABC transporters - General | ABCC3, ABCG2, TAP2 | 43 | 0.07 |
| Cytokine-cytokine receptor interaction | CCR5, CNTF, CSF3R, CX3CR1, EDA, FLT3LG, IL1A, IL1RAP, IL10RA, IL12RB1, IL12RB2, IL18R1, IL2RB, IL21R, IL22RA1, IL22RA2, 1L23R, IL28RA, IL6ST, IL8, LEP, MPL, NGFR, OSM, PDGFRB, TGFB1, TNFSF12, TNSSF4, TNFSF8, | 255 | 0.114 |
| Cell adhesion molecules (CAMs) | CD28, CD4, CD8A, CD80, CNTN2, CLDN1, CLDN10, ITGB7, ITGAL, OCLN, PVRL1 | 132 | 0.083 |
| ECM-receptor interaction | COL4A6, GP5, ITGA2, ITGB3, ITGB7, SV2C, VTN | 87 | 0.08 |
| Neuroactive ligand-receptor interaction | APLN, AVPR2, BDKRB2, CRHR2, CYSLTR2, EDG6, F2RL1, GABBR2, GABRE, GABRA4, HRH1, HRH2, HTR4, LEP, MC1R, NMUR1, NTSR1, NTSR2, P2RY2, P2RY6, P2RX7, PTHR1, TACR1, UCN2 | 302 | 0.079 |
| Calcium signaling pathway | ADCY1, ADCY2, ADCY6, AKAP5, ATP2B1, BDKRB2, CACNAID, CACNA1G, CACNA2D4, CACNG3, CAMK2A, CYSLTR2, ERBB3, HDAC6, HDAC8, HRH1, HRH2, HTR4, NFATC4, NTSR1, P2RX7, PDGFRB, PLCB2, PLCB4, PRKACB, TACR1, TNNI1 | 176 | 0.153 |
| Phosphatidylinositol signaling system | CDC25B, DGKA, DUSP3, DUSP18, INPP5D, PLCB2, PLCB4, PTPN3, PTPRU, SIRPA | 79 | 0.127 |
| Jak-STAT signaling pathway | CNTF, CSF3R, JAK3, IL2RB, IL10RA, IL12RB1, IL12RB2, IL21R, IL22RA1, IL22RA2, IL23R, IL28RA, IL6ST, LEP, MPL, OSM, STAT5A, SPRY3, TYK2 | 153 | 0.124 |
| MAPK signaling pathway | ACVR1C, ARRB1, CASP10, CDC25B, DUSP3, ELK4, FGF7, FGR, GNA12, IKBKG, IL1A, MAP2K3, MAP3K8, MAP3K14, MAPK4, MAPKAPK2, MOBKL2C, NFATC4, PDGFRB, PLA2G3, PRKACB, RASGRF1, RRAS, SHC3, STK4, TGFB1, TP53, ZAK | 271 | 0.103 |
| TGF-beta signaling pathway | ACVR1C, BMPR2, CDKN2B, CHRD, CUL1, SMAD6, TGFB1 | 84 | 0.083 |
| Wnt signaling pathway | CUL1, CAMK2A, DAAM2, DVL3, NFATC4, PLCB2, PLCB4, PRKACB, TP53, WNT8A | 147 | 0.068 |
| Notch signaling pathway | DVL3, NUMB, PSEN2 | 46 | 0.065 |
| Hedgehog signaling pathway | PRKACB, WNT8A | 57 | 0.035 |
| VEGF signaling | EIF1AX, PXN | 72 | 0.028 |
| Regulation of actin cytoskeleton | ARHGEF6, BDKRB2, FGF7, FGD1, FGD3, GNA12, IL17RE, ITGA2, ITGAL, ITGB3, | 206 | 0.083 |

| | | | |
|---|---|---|---|
| | ITGB7, PDGFRB, PPP1R12B, PXN, RRAS, SLC9A1, WASF2 | | |
| Apoptosis | ATM, BIRC4, CAPN5, CASP10, FADD, IKBKG, IL1A, IL1RAP, MAP3K14, NGFR, PRKACB, TRAF1, TRAF3, TP53 | 84 | 0.167 |
| Cell cycle | ATM, CDC14B, CDC25B, CDKN2A, CDKN2B, CUL, E2F6, FGR, HDAC6, HDAC8, ORC2L, SMC1L1, TGFB1, TP53, YWHAZ, | 112 | 0.134 |
| Gap junction | ADCY1, ADCY2, ADCY6, MC1R, PDGFRB, PLCB2, PLCB4, PRKACB, RRAS, SRC | 99 | 0.101 |
| Focal adhesion | BIRC4, CAPN5, COL4A6, FGR, HCK, IGF1, ITGA2, ITGB3, ITGB7, PARVA, PDGFRB, PXN, RRAS, SHC3, SHC4, SRC, STYK1, VTN | 194 | 0.093 |
| Tight junction | AMOT, AMOTL1, CLDN10, CLDN1, EPB41L1, OCLN, RRAS, SRC, VAPA, ZAK | 119 | 0.084 |
| Adherens junction | ACVR1C, PVRL1, SRC, WASF2 | 77 | 0.052 |
| Cell communication | COL4A6, DSC2, GJB3, KRT4, VTN | 122 | 0.041 |
| Adipocytokine signaling pathway | ADIPOQ, JAK3, LEP, TYK2 | 69 | 0.058 |
| Insulin signaling pathway | INPP5D, MOBKL2C, PRKACB, RRAS, SHC3, SHC4 | 135 | 0.044 |
| B cell receptor signaling pathway | CAMK2A, GAB2, IKBKG, INPP5D, MALT1, MAP2K3, MAP3K8, MAP3K14, NFATC4, PIK3AP1, RRAS | 63 | 0.175 |
| Hematopoietic cell lineage | CD33, CD4, CD59, CD8A, CSF3R, FLT3LG, GP5, IL1A, ITGA2, ITGB3, ITGB7 | 88 | 0.125 |
| T cell receptor signaling pathway | CD4, CD28, CD8A, IKBKG, ITK, MALT1, MAP3K8, MAP3K14, NFATC4, RELB, RRAS | 93 | 0.118 |
| Natural Killer cell mediated Cytotoxicity | ACVR1C, ARRB1, CACNA1D, CACNA1G, CACNA2D4, CACNG3, CASP10, CDC25B, DUSP3, ELK4, FGF7, GNA12, HSPA1B, IKBKG, IL1A, MAP2K3, MAP3K14, MAP3K8, MAPKAPK2, NFATC4, PDGFRB, PLA2G3, PRKACB, RASGRF1, RRAS, STK4, TGFB1, TP53, ZAK | 271 | 0.107 |
| Complement and coagulation cascades | BDKRB2, C7, CD59, MASP1, MBL2, PLAUR, SERPINA5 | 69 | 0.101 |
| Toll-like receptor signaling pathway | ATM, CD80, FADD, IL8, IKBKG, MAP2K3, MAP3K14, RELB, TICAM2 | 91 | 0.099 |
| Antigen presenting pathway | TAP2, TAPBP | 86 | 0.023 |
| Dorso-ventral axis formation | ETV6, ETV7, RRAS | 28 | 0.107 |
| Axon guidance | EPHA8, NFATC4, PLXNA2, PLXNB2, RRAS, SEMA4F, SRGAP3, UNC50 | 130 | 0.062 |
| Circadian rhythm | ARNTL2 | 18 | 0.056 |
| Amyotrophic lateral sclerosis (ALS) | SLC1A2, TP53 | 17 | 0.118 |
| Huntington's disease | HAP1, TGM2, TP53 | 30 | 0.1 |
| Neurodegenerative Disorders | NGFR, PSEN2 | 35 | 0.057 |
| Alzheimer's disease | PSEN2 | 22 | 0.045 |
| Type I diabetes mellitus | CD28, CD80, IL1A | 44 | 0.068 |
| Type II diabetes mellitus | ADIPOQ, Cacna1d, CACNA1G | 44 | 0.068 |

## 9.7. Gene ontology data for the human MIR-containing genes.

The gene ontology identification number and term are listed. The percentage of MIR-containing genes is included and the p value. The ontologies are categorised as follows; **A)** GO terms relating to protein binding; **B)** growth and development; **C)** neuronal function; **D)** mammalian reproduction; **E)** cell compartment and **F)** immune responses.

**A)** GO terms relating to protein binding.

| GO ID | Gene Ontology Term | MIRs (%) | P value |
|---|---|---|---|
| GO:0046914 | Metal ion binding | 37.83 | 5.32E-05 |
| GO:0043169 | Cation binding | 35.2 | 0.000141201 |
| GO:0042277 | Peptide binding | 1.9 | 0.0300147 |
| GO:0005509 | Calcium ion binding | 8.68 | 0.252313 |
| GO:0030246 | Carbohydrate binding | 2.32 | 0.291063 |
| GO:0005524 | ATP binding | 16.42 | 0.305178 |
| GO:0000166 | Nucleotide binding | 14.56 | 0.422993 |
| GO:0005529 | Sugar binding | 1.55 | 0.446001 |
| GO:0008289 | Lipid binding | 3.06 | 0.495413 |
| GO:0003677 | DNA binding | 17.54 | 0.527381 |
| GO:0005102 | Receptor binding | 5.61 | 0.645644 |
| GO:0005525 | GTP binding | 3.45 | 0.853497 |
| GO:0003723 | RNA binding | 3.82 | 0.994431 |

**B)** GO terms pertaining to growth and development.

| GO ID | Gene Ontology Term | MIRs (%) | P value |
|---|---|---|---|
| GO:0001501 | Skeletal development | 3.63 | 0.00128131 |
| GO:0048856 | Anatomical structure development | 17.89 | 0.00328912 |
| GO:0048513 | Organ development | 11.79 | 0.0253845 |
| GO:0009887 | Organ morphogenesis | 4.41 | 0.0421475 |
| GO:0009888 | Tissue development | 3.89 | 0.0586727 |
| GO:0007568 | Aging | 0.33 | 0.163553 |
| GO:0021675 | Nerve development | 0.13 | 0.282332 |
| GO:0055001 | Cell development | 13.73 | 0.288051 |
| GO:0001701 | Embryonic development | 2.28 | 0.384237 |
| GO:0001558 | Regulation of cell growth | 1.55 | 0.61238 |
| GO:0016049 | Cell growth | 1.69 | 0.643293 |
| GO:0007548 | Sex differentiation | 0.36 | 0.906218 |
| GO:0007389 | Pattern specification | 0.68 | 0.9597 |

**C)** GO terms related to neuronal function.

| GO ID | Gene Ontology Term | MIRs (%) | P value |
|---|---|---:|---:|
| GO:0007611 | Neurotransmitter transport | 1.46 | 0.000341128 |
| GO:0006836 | GPCR receptor activity | 1.84 | 0.0211317 |
| GO:0043005 | Neuron projection | 1.35 | 0.0360011 |
| GO:0046928 | Transmission of nerve impulse | 3.04 | 0.243314 |
| GO:0030901 | Neurite development | 0.17 | 0.250773 |
| GO:0048666 | Synaptic transmission | 2.59 | 0.420828 |
| GO:0048168 | learning and/or memory | 0.46 | 0.430455 |
| GO:0001764 | Regulation of neuronal synaptic plasticity | 0.34 | 0.459277 |
| GO:0007218 | Midbrain development | 0.17 | 0.514246 |
| GO:0030900 | Neuron development | 3.97 | 0.540783 |
| GO:0030182 | Neuron differentiation | 3.97 | 0.540783 |
| GO:0007417 | Regulation of neurotransmitter secretion | 0.99 | 0.591897 |
| GO:0007268 | Neuron migration | 0.52 | 0.645013 |
| GO:0031175 | Central nervous system development | 1.82 | 0.659488 |
| GO:0008528 | Neuropeptide signaling pathway | 0.77 | 0.798758 |
| GO:0019226 | Forebrain development | 0.15 | 0.806132 |

**D)** GO terms pertaining to mammalian reproduction.

| GO ID | Gene Ontology Term | MIRs (%) | P value |
|---|---|---:|---:|
| GO:0007595 | Lactation | 0.61 | 0.00444562 |
| GO:0007567 | Parturition | 0.24 | 0.130172 |
| GO:0009790 | Embryonic development | 2.28 | 0.384237 |
| GO:0030317 | Sperm motility | 0.12 | 0.546723 |
| GO:0007369 | Gastrulation | 0.36 | 0.554173 |
| GO:0001890 | Placenta development | 0.15 | 0.574902 |
| GO:0046661 | Male sex differentiation | 0.26 | 0.634846 |
| GO:0007565 | Pregnancy | 0.49 | 0.63798 |
| GO:0019098 | Reproductive behaviour | 0.11 | 0.697223 |
| GO:0019953 | Sexual reproduction | 1.89 | 0.835177 |
| GO:0007276 | Gametogenesis | 1.6 | 0.846263 |
| GO:0009566 | Fertilisation | 0.24 | 0.867359 |
| GO:0042698 | Menstrual cycle | 0.12 | 0.893952 |
| GO:0046660 | Female sex differentiation | 0.13 | 0.895515 |

**E)** GO terms relating to the cell compartment.

| GO ID | Gene Ontology Term | MIRs (%) | P value |
|---|---|---:|---:|
| GO:0005886 | Plasma membrane | 20 | 0.000723 |
| GO:0005794 | Golgi apparatus | 6.8 | 0.001245 |
| GO:0030425 | Dendrite | 0.75 | 0.031293 |
| GO:0043005 | Neuron projection | 1.35 | 0.036001 |
| GO:0001917 | Photoreceptor | 0.22 | 0.03732 |
| GO:0043203 | Axon hillock | 0.12 | 0.125862 |
| GO:0031982 | Vesicle | 2.41 | 0.131283 |
| GO:0031982 | Endoplasmic reticulum | 9.35 | 0.154074 |
| GO:0005768 | Endosome | 1.7 | 0.210515 |
| GO:0019861 | Flagellum | 0.37 | 0.273242 |
| GO:0005773 | Vacuole | 2.38 | 0.301798 |
| GO:0005929 | Cilium | 0.51 | 0.330358 |
| GO:0030424 | Axon | 0.5 | 0.338993 |
| GO:0043205 | Fibril | 0.12 | 0.375564 |
| GO:0044456 | Synapse part | 0.66 | 0.671878 |
| GO:0005581 | Collagen | 0.25 | 0.70366 |
| GO:0005737 | Cytoplasm | 34 | 0.757031 |
| GO:0005856 | Cytoskeleton | 10.37 | 0.760973 |
| GO:0005739 | Mitochondrion | 8.67 | 0.780529 |
| GO:0005829 | Cytosol | 4.25 | 0.914759 |
| GO:0005634 | Nucleus | 45.75 | 0.957493 |
| GO:0005694 | Ribosome | 1.36 | 0.995429 |

**F)** GO terms pertaining to immune responses.

| GO ID | Gene Ontology Term | MIRs (%) | P value |
|---|---|---|---|
| GO:0048535 | Cytokine receptor activity | 1.58 | 4.61E-05 |
| GO:0019882 | Immune response | 7.78 | 0.00228131 |
| GO:0046649 | Wound healing | 1.94 | 0.00633122 |
| GO:0045321 | Lymphocyte activation | 2.19 | 0.0729672 |
| GO:0006956 | Leukocyte activation | 2.17 | 0.111272 |
| GO:0002706 | Lymphocyte mediated immunity | 0.91 | 0.283123 |
| GO:0042119 | Neutrophil activation | 0.13 | 0.328416 |
| GO:0006954 | Inflammatory response | 2.67 | 0.349137 |
| GO:0006956 | Complement activation | 1.32 | 0.358703 |
| GO:0006959 | Humoral immune response | 0.99 | 0.379501 |
| GO:0006935 | Chemotaxis | 1.3 | 0.480615 |
| GO:0019882 | Antigen presentation | 0.78 | 0.628384 |
| GO:0042379 | Chemokine receptor binding | 0.16 | 0.972179 |

## 9.8. MIR elements which are spliced, and are within coding sequences of human genes

The gene symbol and accession number were obtained from Entrez gene at NCBI. The transcript no has been provided if the exonisation occurs in a reference sequence (-) signifies that there is only one sequence and the exonisation occurs in the reference sequence. Abbreviations: Ex, exon; Int, intron; UTR, untranslated region; CDS, coding sequence; ATG, methionine sequence; TAG, stop codon; Alt, alternative; Acc, acceptor; Don, donor; Con, constitutive; SS, splice site; RT, read through exon. Intronic sequences are in lowercase and exonic nucleotides capitals. Start and stop codons are bold and italicised.

| Gene symbol | Transcript No. | MIR type | mRNA region | RefSeq. region | Splicing event | SS: 3’ = 23-mer<br>5’ = 9-mer |
|---|---|---|---|---|---|---|
| ABCC2 | - | MIR | CDS | Ex 7 | Middle Con Ex | - |
| ABHD14B | 1, 2 | MIRb | 5' UTR | Ex 2 | Con Acc SS | ttattgttcccatatttcagATC |
| ABHD2 | 1 | MIR<br>MIR3<br>MIR3 | 5' UTR<br>3' UTR<br>3’ UTR | Int 1<br>-<br>- | Alt Acc/Don SS | tttcctcacctctaaaacagAAA<br>AGGgtaagc |
| ACACA | nr | MIRb | CDS | Int 1 | Alt Don SS | AAGgtaggc |
| ADAMTS7 | nr | MIR_Mars<br>MIRb | CDS<br>TAG | -<br>Ex 15 | Alt TAG /Ex 15 RT | CATGGA***TGA***GGCAGG |
| ADAMTSL5 | - | MIRb | ATG | Ex 2 | Con ATG | GGCTCT***ATG***GACTCG |
| ADRA1A | nr | MIR | CDS | Int 1 | Alt Acc SS | ctaagtctttacttttacagATG |
| AHI1 | 4 | MIRc | TAG | Int 15 | Alt TAG | AAGCAC***TAG***ACTCAGG |
| ANAPC5 | nr | MIRb | TAG | Int 11 | Alt Acc SS / Alt TAG | ttgttatccctatttcacagAAC<br>***TAA***AGA |
| ANKRD49 | nr | MIR_Mars<br>MIRb | TAG<br>3’ UTR | Int 2 | Alt TAG / Ex 2 RT | TCTGCT***TGG***CCACTT |
| ANKRD9 | - | MIR3<br>MIR3 | 5' UTR<br>5' UTR | Ex 1<br>Ex 2 | Con Acc SS<br>Con Acc/Don SS | gcaggggacctgagttctagGCC<br>acccctcccctttcttacagATA/<br>CAGgtaggt |
| ARL10 | nr | MIRb<br>MIRb | CDS<br>TAG | EX 3 | Alt TAG /Ex 3 RT | TTTTTT***TGA***GACAGA |
| ARNTL | 1 | MIRb | 5’ UTR | Ex 3 | Alt Don SS | GTGgtaagc |
| ART1 | - | MIR | ATG | Ex 2 | Con Acc SS | ttaacactgcaattttccagATG<br>ACCAGC***ATG***CAGATG |
| BCL2L1 | nr | MIRb | CDS | Int 2 | Alt Don SS | TGGgtacta |
| BTN3A3 | a, b | MIR | TAG<br>3’ UTR | Ex 11 | Con TAG | CTTTAC***TGA***TATTCA |
| C10orf118 | nr | MIRc | 5' UTR | Int 1 | Alt Acc SS | cccatcttgattttgtttagAGG |
| C11orf51 | - | MIRb | 5' UTR | Ex 2 | Con Acc/Don SS | ttattatccccattttacagATG/<br>CAGgtctgt |
| C12orf36 | - | MIRc<br>MIR3 | ATG<br>3’ UTR | Ex 2 | Con Don SS | GAGgtgagt<br>- |
| C12orf52 | - | MIRb | 5' UTR | Ex 2 | Con Don SS | tttcacttccaccgttttagCTT |
| C12orf76 | - | MIR | CDS | Ex 3 | Con Don SS | GAGgtaata |
| C14orf178 | - | MIR_Mars<br>MIRb | CDS<br>TAG | Ex 3 | Middle ex<br>Con TAG | -<br>GGG***TAA***TGA |
| C15orf33 | nr | MIR3 | TAG | Ex 10 | Alt TAG | AAC***TAG***TAA |
| C16orf47 | - | MIR | CDS | Ex 3 | Con Acc SS | attattactcccactttacagATA |

| | | | | | | |
|---|---|---|---|---|---|---|
| C19orf39 | - | MIR | TAG | Ex 2 | Con TAG | CCC***TGA***GCT |
| C1orf220 | - | MIRb | ATG | Ex 2 | Con ATG | CCC***ATG***ATC |
| C3orf35 | b | MIR | CDS | Ex 6 | Alt TAG | GGA***TAG***AAA |
| C6orf114 | - | MIRb<br>MIRb | TAG<br>3' UTR | Ex 2 | Con TAG | ACA***TAG***TAC |
| C7orf65 | - | MIRb | CDS | Ex 2 | Con Acc/Don SS | tgcctgctcttattgtgcagGTA/<br>GAGgtaata |
| CACNA1G | 15 | MIRb | TAG<br>3' UTR | Ex 23 | Ex 23 RT /Alt TAG | GGG***TAA***GGG |
| CCDC82 | - | MIRb | 5' UTR | Ex 2 | Con Don SS | TCGgtaagt |
| CCNO | - | MIRb | ATG | Ex 1 | Con ATG | ATC***ATG***GTG |
| CD300LG | nr | MIR3<br>MIR | TAG<br>3’ UTR | Int 6 | Alt Acc SS/Alt TAG | ctttccatccccccttttagAAT<br>GTT***TGA***AGC |
| CD99L2 | nr | MIRb | CDS | Int 2 | Alt Don SS | AAGgtgagt |
| CDC20B | - | MIR<br>THER1_MD<br>MIR | TAG<br>3’ UTR<br>3’ UTR | Ex 12 | Con TAG | TAC***TAG***CAC |
| CDC42BPG | - | MIR3 | CDS | Ex 2 | Middle con ex | - |
| CHRD | 2 | MIR | TAG | Int 2 | Alt Acc/ Don SS | attattatccccattttccaGAC<br>AGGgtaagtcttg |
| CHRDL2 | - | MIR | CDS | Ex 10 | Con Acc / Don SS | attattatccccattttacagATG<br>CCAgtaagt |
| CHRNA1 | 1 | MIR3 | CDS | Ex 4 | Alt Don SS | GAGgtcagt |
| CIITA | nr | MIR | TAG | Ex 11 | Ex 11 RT<br>Alt TAG | AGA***TGA***GGA |
| CPA5 | 2 | MIRc | 5' UTR | Int 1 | Alt Acc/ Don SS | ataatacattgcattttgcagGCT<br>GAGgtaggg |
| CRELD1 | nr | MIR | CDS | Int 10 | Alt Acc SS | tgtgtcagatgctgttctagGTG |
| CSF3R | a, b, c, d | MIR3 | 5’ UTR | Ex 2 | Con Acc/ Don SS | atgttatttctttcccacagATG<br>CTGgtaagt |
| DISC1 | S | MIR | TAG | Int 9 | Alt TAG | GCC***TGA***GGA |
| DMWD | - | MIR<br>MIR3 | CDS<br>3’ UTR | Ex 4 | Con Acc/ Don SS | agacccctctgtctccgtagTTC<br>CAAgtcagt |
| DNAJB13 | nr | MIR3 | ATG | Int 6 | Alt Acc SS/Alt ATG | cataagatttcattgtacagATG<br>GAC***ATG***GAG |
| DPM2 | 2 | MIRb | TAG | Int 2 | Ex 2 RT/Alt TAG | TCT***TGA***ACT |
| DUOX1 | 1 | MIR3 | 5' UTR | Int 1 | Alt Acc SS | aggcatatttgctcccatagCTG |
| EIF4G3 | nr | MIRb | 5’ UTR | Int 1 | Alt Acc/ Don SS | ttattttgctaattctttccagCAA<br>AAGgtaatc |
| ELMO2 | 1 | MIRb | 5' UTR | Int 2 | Alt Don SS | AAGgtagga |
| EPB41L1 | nr | MIR | 5' UTR | Int 1 | Alt Acc SS<br>Alt ATG | tctgtttcttatctgtaaaCAG<br>CTC***ATG***GTT |
| EPHB2 | nr | MIR | TAG | Int 6 | Alt Acc SS<br>Alt TAG | ttactatccccctttacagATG<br>AGT***TAA***GGA |
| FAM109A | - | MIR | 5' UTR | Ex 2 | Con Acc SS | gttacgatcccattttacagATG |
| FAM189B | nr | MIR | TAG | Ex 5 | Alt TAG | TGG***TAA***GGC |
| FAM23B | - | MIR | TAG | Ex 5 | Con TAG | ATT***TGA***GGT |
| FAM53B | nr | MIR | TAG | Int 4 | Alt Acc SS<br>Alt TAG | cctgttgtattttctcctccagATA<br>CCA***TGA***GGA |
| FBXO18 | 1 | MIR3 | ATG | Int 1 | Alt ATG<br>Alt Don SS | CTC***ATG***AGC<br>CAGgcaagt |
| FHAD1 | nr | MIRc | CDS | Int 33 | Alt Acc SS | cgatgatccctattttacagATG |

| | | | | | | |
|---|---|---|---|---|---|---|
| FIGNL1 | 1<br><br>2 | MIR<br>MIR<br>MIR<br>MIR | 5' UTR<br>5' UTR<br>5' UTR<br>5' UTR | Ex 2<br>Ex 2<br>Ex 2<br>Ex 3 | Con Acc SS<br>Alt Don SS<br>Alt Don SS<br>Con Acc SS | tgttgctgcgcttcttttatagGTG<br>TGAgtaggc<br>CAAgtaagt<br>TTGgtactg |
| FXYD4 | - | MIR | 5' UTR | Ex 2 | Con Acc / Don SS | tgatgatcctccttttacagGAC<br>TTGgtaagt |
| GAFA2 | - | MIR | CDS | Ex 1 | Middle Con Ex | - |
| GCET2 | 1,2 | MIR3<br>MIR3<br>MIRm | TAG<br>3' UTR<br>3' UTR | Ex 6 | Con TAG | TTA***TAG***TGA |
| GDPD5 | nr | MIR<br>MIRb | 5' UTR | Int 2 | Alt Acc / Don SS | tgcaaatccttccttcctagGGC<br>CAGgtttgt |
| GGA1 | nr | MIRb | CDS | Int 9 | Alt Acc | atcttatttactacccgcagCTG |
| GIPC1 | 1, 2 | MIR | 5' UTR | Int 1 | Alt Acc / Don SS | ggaattccacccatttttcagATG<br>CTGgtaagt |
| GPR171 | - | MIR | 5' UTR | Ex 2 | Con Don SS | TGGgtaagt |
| GSG1L | nr | MIR | CDS | Int 4 | Alt Acc / Don SS | ccaaagtgatgggattacagGTG<br>CTGgtaagt |
| HHIPL1 | b | MIR3 | TAG | Ex 8 | Alt TAG | AAA***TGA***GCA |
| HIF3A | 3 | MIR3 | TAG | Ex 13 | Ex 13 RT /Alt TAG | TGG***TAG***GCC |
| HM13 | 4 | MIRb<br>MIR<br>MIRb | TAG<br>3' UTR<br>3' UTR | Ex 3 | Ex 3 RT/ Alt TAG | GGA***TGA***GAA |
| HNF4A | 3 | MIR | TAG | Ex 8 | Ex 8 RT/ Alt TAG | GCT***TAA***CCT |
| HTR3E | - | MIR | ATG | Ex 1 | Con ATG | CAA***ATG***TTA |
| IBRDC1 | - | MIR | TAG | Ex 9 | Con TAG | TTGTAAAAA |
| IKBKG | 2 | MIR3 | ATG | Ex 1 | Alt ATG | CCC***ATG***GCC |
| IL6ST | 1, 2 | MIRb | 5’ UTR | Ex 2 | Con Acc / Don SS | agactctcctctctttacagTGA<br>TGGgtaagt |
| INPP5J | nr | MIR | 5’ UTR | Int 1 | Alt Acc SS | cccatcccttcattttacagACA |
| ITGB7 | - | MIR | 5' UTR | Ex 2 | Con Acc SS | tcaacttctccactttgcagATG |
| KIF19 | nr | MIR | TAG | Ex 12 | Ex 12 RT/ Alt TAG | GTT***TGA***GAG |
| KIFC3 | nr | MIRb<br>MIRb<br>MIRb | 5' UTR<br>5' UTR<br>TAG | Int 3 | Alt Don SS<br>Alt TAG | CAAgtaagt<br>CTC***TAG***AGG |
| KLC1 | 1 | MIRb | TAG | Int 13 | Alt Acc / Don SS<br>Alt TAG | gtcctccccacattttaaagATG<br>AAC***TGA***CTT<br>ATGgtgagt |
| LAS1L | - | MIR3 | CDS | Ex 9 | Con Acc / Don SS | ccaacatcctcatgttacagATG<br>ACAgtgagt |
| LIG4 | 1, 2 | MIRb | 5' UTR | Int 1 | Alt Don SS | CAGgtattc |
| LINS1 | 1 | MIRb | CDS | Ex 8 | Middle con Ex | - |
| LMBR1L | nr | MIR_Mars | ATG | Int 2 | Alt ATG | ataacatttctgatggtcagTTT<br>GAG***ATG***GTT |
| MICA | - | MIR | TAG | Ex 6 | Con TAG | GCC***TAG***ACT |
| MORF4L2 | 1, 6, 8 12, 13, 14, 15, 16 | MIR | 5' UTR | Int 2 | Alt Acc / Don SS | taatagttcctacttcatagGAT<br>CATgtaagt |
| MPG | a | MIR | ATG | Int 2 | Alt ATG | AGG***ATG***GTC |

| | | | | | | |
|---|---|---|---|---|---|---|
| MRRF | nr | MIR | CDS | Int 4 | Alt Acc SS | tttttgttattggtccactagGAA |
| MS4A10 | - | MIRb<br>MIR3 | TAG<br>3’ UTR | Ex 8 | Con Acc SS<br>Con TAG | ttaatatctccatttcatgcAGG<br>AAC***TGA***AGA |
| MST150 | - | THER1_MD<br>MIR3<br>MIRb | 5’ UTR<br>ATG<br>3’ UTR | Ex 1 | Con ATG | GAC***ATG***TCC |
| MTIF2 | 1, 2 | MIR3 | 5’ UTR | Ex 2 | Con Don SS | GGAgtaagt |
| MUC15 | a, c | MIRb | 5' UTR | Int 1 | Alt Acc / Don SS | ttgcctcttccattttccagCTT<br>GTTgtaagt |
| MYEOV | - | MIR<br>MIRb | TAG<br>3’ UTR | Ex 3 | Con TAG | TGT***TGA***GGA |
| MYO15A | - | MIR<br>MIR | CDS<br>3’ UTR | Ex 33 | Con Don SS | AAGgtcaac |
| MYO7A | 3 | MIRb | TAG | Int 27 | Alt TAG | AGT***TAG***GAC |
| MYT1L | - | MIRm | 5’UTR | Ex 2 | Con Acc SS | ttgctttttatattttacagATG |
| NCR2 | - | MIRb | CDS | Ex 5 | Middle Exon | - |
| NDST2 | - | MIR | 5' UTR | Ex 2 | Con Don SS | AAGgttcta |
| NDUFB2 | nr | MIR | ATG | Int 1 | Alt ATG | GTG***ATG***GAG |
| NEK11 | 1, 2 | MIR 3 | 5’ UTR | Int 1 | Alt Acc /Don SS | attctgtccttctctaaaagGAA<br>TCTgtaagt |
| NFXL1 | nr | MIR3 | TAG | Int 16 | Alt Acc SS<br>Alt TAG | gtatactctttgttttgcagATA<br>ATT***TAA***ATA |
| NIPSNAP1 | - | MIR3 | ATG | Ex 1 | Con ATG | AAC***ATG***GCT |
| NSD1 | nr | MIRb | CDS | Int 2 | Alt Don SS | GGGgtaagc |
| NTRK3 | 3 | MIR3 | CDS | Int 13 | Alt Don SS | ATGgtggag |
| NUMB | 1,2, 3,4 | MIRb | 5’ UTR | Ex 3 | Con Acc / Don SS | ttattattctcattttacagATG<br>CAGgtctgt |
| OTUB2 | nr | MIRb | TAG | Ex 3 | Alt TAG | TTC***TAA***CTT |
| OTUD6B | nr | MIR3 | ATG | Int 3 | Alt ATG<br>Alt Don SS | AAGgtgctt |
| PAQR3 | nr | MIR | TAG | Int 5 | Alt ATG<br>Alt Don SS | GTT***TGA***GAA<br>TAAgtaagt |
| PAX7 | 1 | MIR | TAG | Int 8 | Alt TAG | TAC***TAG***GGC |
| PBOV1 | - | MIR | CDS | Ex 1 | Middle con Ex | - |
| PCTK3 | nr | MIR | TAG | Int 12 | Alt TAG | ACT***TAG***CTG |
| PELI3 | 2 | MIRb | CDS | Int 2 | All Acc / Don SS | tcatttcaatcttcacaaagATG<br>CTGgtaagt |
| PFKFB4 | nr | MIRc | CDS | Int 3 | Alt Don SS | gtcaccccctacttttacagATG |
| PGGT1B | nr | MIR<br>MIR3 | 5’ UTR<br>ATG | Int 1 | Alt ATG | GGC***ATG***GTG |
| PHF19 | 2 | MIRb | TAG | Int 5 | Ex 5 RT / Alt TAG | TTG***TAG***ACT |
| PSD4 | - | MIRm<br>MIRm | CDS<br>3’ UTR | Ex 3 | Con Don SS | TGGgcaagt |
| PTGS1 | nr | MIRb | CDS | Int 2 | Alt ATG<br>Alt Don SS | CAA***ATG***AGG<br>TCAgtaggt |
| PUS10 | nr | MIR<br>MIRb | CDS<br>3’ UTR | Int 2 | Alt Acc SS | atgattgtctctattttcagATG |
| RGAG1 | - | MIRb | 5' UTR | Ex 1 | Con Don | TGGgtaagt |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | MIRb<br>MIRb | 5' UTR<br>3' UTR | -<br>- | SS | |
| RHBDD2 | 1, 2 | MIRb | TAG | Ex 4/5 | Con TAG | CCC***TGA***GAG |
| RHBDD3 | - | MIRm | 5' UTR | Ex 2 | Con Acc / Don SS | gatatccttggtttctacagAAG<br>AAGgtaggc |
| RNASEL | nr | MIR<br>MIR3<br>MIRb<br>MIRC | TAG<br>3' UTR<br>3' UTR<br>3' UTR | Int 4 | Alt Acc SS<br>Alt TAG | aatcatctatatttttgcagATG<br>CAG***TAA***GGG |
| RNF7 | nr | MIR | CDS | Int 2 | Alt Acc SS | gtattgtcccttttttacagATG |
| RNFT2 | 1 | MIRb | TAG | Ex | Alt TAG | GTG***TAA***GGA |
| RNFT2 | 2 | MIRb | TAG | Int 11 | Alt TAG | GTG***TAA***GGA |
| SAPS3 | - | MIR | 5' UTR | Ex 2 | Con Acc SS | tctaaacctctcatttacagATA |
| SCN1B | b | MIRm | TAG | Ex 3 | Ex 3 RT / Alt TAG | GTT***TGA***GCC |
| SGIP1 | nr | MIRb | CDS | Int 15 | Alt Acc / Don SS | ttatctttcccttttttacagATG<br>CAGgtctgt |
| SIAE | nr | MIR | TAG | Int 2 | Ex 3 RT/Alt TAG | ACT***TAA***CAA |
| SLC22A23 | 2 | MIRb | 5' UTR | Int 2 | Alt Don SS | AAGgtagag |
| SLC6A13 | nr | MIR3<br>MIRb | TAG<br>3’ UTR | Ex 2T | Ex 2 RT /Alt TAG | CCA***TAG***ATG |
| SLC9A1 | nr | MIR | TAG | Ex 5 | Ex 5 RT /Alt TAG | AAG***TGA***GAA |
| SLIC1 | 2 | MIR3 | TAG | Int 3 | Alt Acc SS<br>Alt TAG | tcactcacctctgtttacagATG<br>CTC***TGA***GAG |
| SPATA12 | - | MIR<br>MIRb | CDS<br>3’ UTR | Ex 2 | Middle con Ex | - |
| SPTBN4 | Σ5 | MIR | TAG | Ex 27 | Ex 27 RT /Alt TAG | GAA***TGA***CAA |
| SRRM2 | - | MIRb | CDS | Ex 4 | Middle con Ex | - |
| ST3GAL1 | 1, 2 | MIRb | 5' UTR | Ex 3 | Con Don SS | tattttcttctgtttttcagATG |
| ST6GALNAC6 | - | MIRb | ATG | Ex 2 | Con ATG | CAC***ATG***GCT |
| ST7L | 1, 2, 3 | MIRb | TAG | Ex 15 | Alt TAG | GGC***TGA***GCC |
| STAB1 | nr | MIR3 | CDS | Ex 21 | Ex 21 RT /Alt TAG | AAT***TAG***AGA |
| STRA6 | 5 | MIRm<br>MIRb<br>MIRb | TAG<br>3’ UTR<br>3’ UTR | Ex 6 | Ex 6 RT /Alt TAG | ATT***TGA***ACC |
| TCL6 | a1, a2, a3, b1 | MIR3<br>MIRb<br>MIRc | 5’ UTR<br>ATG<br>3’ UTR | Int 4 | Alt ATG | AGG***ATG***GAG |
| TGFBR2 | 1 | MIRb | CDS | Int 1 | Alt Acc SS | ataattatcctgttttacagATG |
| TGM2 | nr | MIR<br>MIR<br>MIRb<br>MIR<br>MIRb<br>MIRc<br>MIR3 | 5- UTR<br>ATG<br>CDS<br>CDS<br>3- UTR<br>3- UTR<br>3- UTR | Int 10 | Alt ATG | AGG***ATG***AAG |
| TIMM8B | nr | MIRb | CDS | Int 1 | Alt Acc SS | cctccctctgctccttgcagGCC |
| TIPARP | - | MIRb | CDS | Ex 6 | Middle con Ex | - |

| | | | | | | |
|---|---|---|---|---|---|---|
| TMEM14C | nr | MIRc | 5' UTR | Int 1 | All Acc SS | actgagacttaggttgcgtagCTT |
| TMUB2 | 2, 3 | MIR<br>MIRb | 5' UTR<br>5' UTR | Int 1 | Alt Don SS | GAGgtaggt |
| TNFSF12 | nr | MIR | TAG | Int 5 | Alt Acc SS<br>Alt TAG | taattacctccattttacagATG<br>CCA***TGA***GAT |
| TP53I11 | nr | MIRb | 5' UTR | Int 1 | Alt Acc / Don SS | ttttgccccgggctccctagGAA<br>GAGgtaagg |
| TRAF3 | 1, 2 | MIR3 | 5' UTR | Int 1 | Alt Acc SS | ccaatccctttattttacagATG |
| TRIM10 | 2 | MIR3 | TAG | Int 7 | Alt TAG | GCA***TAG***AAA |
| TRIM35 | nr | MIR | CDS | Int 1 | Alt Acc SS | atcctttgcccatttttaagGTG |
| TRIOBP | - | MIRb | TAG | Ex 8 | Con TAG | ATT***TGA***GCG |
| TYK2 | - | MIR | 5’ UTR | Ex 1 | Con Don SS | CCGGTGGGT |
| UBE2V1 | 1, 2, 3 | MIRb | ATG | Int 1 | Alt Acc /Don SS<br>Alt ATG | ggaaagcattttatctccacAGC<br>AAG***ATG***GCA<br>AAGgtgagt |
| USH2A | b | MIR3 | CDS | Ex 45 | Middle Con Ex | - |
| VASH1 | nr | MIR3 | TAG | Ex 4 | Ex 4 RT /Alt TAG | GGT***TGA***ATG |
| WBSCR27 | - | MIRc | TAG | Ex 6 | Con TAG | AAGTGAGAT |
| YIPF1 | - | MIR | 5' UTR | Ex 2 | Con Acc SS | gttatgcacccattttacagATG |
| ZBTB44 | nr | MIR | TAG | Int 5 | Alt TAG | TGT***TGA***TTA |
| ZFAND5 | a | MIR3 | 5’ UTR | Int 1 | Alt Don SS | CAGgtgagt |
| ZFHX2 | nr | MIR3 | TAG | Int 4 | Alt Acc SS<br>Alt TAG | actgctggacttctataaaaGGA<br>AAA***TGA***TCT |
| ZNF211 | nr | MIR | CDS | Int 2 | Alt TAG<br>Alt Don SS | TGG***ATG***AGG<br>CAGgttaga |
| ZNF546 | - | MIRb<br>MIR | TAG<br>3’ UTR | Ex 7 | Con TAG | ATG***TAA***AGA |
| ZNF639 | - | MIRm | CDS | Ex 12 | Con Acc SS | tcttccaaatccctttttacAGA |
| ZNF767 | - | MIRc | TAG | Ex 4 | Con Acc SS<br>Con TAG | tattgtcttccatttgacagATG<br>GAT***TGA***GTA |

## 9.9. Human EST sequences which contained spliced MIR elements

The gene symbol and accession number are listed, 73 EST sequences in total.

| Gene Name | Accession No. | Gene region | Gene Name | Accession No. | Gene region |
|---|---|---|---|---|---|
| ABCA10 | DB226553 | CDS | PCSK2 | BE257514 | CDS |
| AGTRAP | BM772673 | CDS | PGAP3 | BI518652 | CDS |
| ANKK1 | CR741028 | CDS | PIP5K1B | AL705046 | 5' UTR |
| ARPC4 | DB140773 | CDS | PLCB4 | BX414883 | CDS |
| BCL2L14 | DA920483 | 5' UTR | POLG | CN411689 | 5' UTR |
| C10orf57 | DB468324 | CDS | POMT2 | BP331592 | TAG |
| C16orf45 | BF309500 | CDS | PPP2R2A | DA828841 | CDS |
| CARM1 | CN352771 | CDS | RPL18 | BM009191 | CDS |
| CCDC149 | BE710079 | 5' UTR | RUFY4 | BX641664 | 5' UTR |
| CCT7 | DC355319 | CDS | RUNX1T1 | AU117637 | CDS |
| CDH5 | DC423813 | 5' UTR | SAP30BP | DA684650 | CDS |
| CHD6 | BM799444 | 5' UTR | SCYL2 | AA280659 | CDS |
| CLN5 | DA018669 | TAG | SDF2 | CF126565 | CDS |
| COG7 | BM821557 | CDS | SEC13 | AL548842 | 5’ UTR |
| CYP4F12 | CD694538 | CDS | SGSH | BP285106 | CDS |
| DDX27 | BI914073 | CDS | SLC25A32 | BG714806 | CDS |
| FAM35A | CB111653 | CDS | SMCR7L | BP315581 | 5' UTR |
| FREM1 | CN425444 | 3' UTR | SMPDL3B | BE747168 | CDS |
| GALE | BI826045 | 5' UTR | SPC25 | BI912854 | CDS |
| GAS7 | DC418644 | CDS | SSTR2 | DB000266 | 5' UTR |
| GBA2 | AA203694 | CDS | STAT6 | CR979664 | CDS |
| GLTP | BI752839 | CDS | TBL1XR1 | CN256124 | 5' UTR |
| GNPDA1 | DA733468 | 5' UTR | THOC1 | CN288816 | CDS |
| GPR137B | CD521447 | CDS | TIMM17A | BM826653 | CDS |
| IFNAR2 | CA309324 | CDS | TLL2 | AL530454 | CDS |
| KARS | BQ437255 | CDS | TMEM68 | CN429772 | CDS |
| KIF1B | BI829019 | CDS | TRAF4 | BQ056660 | CDS |
| MAGI3 | CR996494 | CDS | TRNAU1AP | DR007034 | CDS |
| MAPKAP1 | BG327258 | 5' UTR | TTC4 | BG180840 | CDS |
| MCM8 | CA495297 | CDS | TTLL3 | AL045121 | CDS |
| METAP1 | DA136874 | 5' UTR | UROS | BM476860 | CDS |
| MFN2 | BG723252 | CDS | WWOX | AW874690 | CDS |
| NDUFA12 | BG179091 | CDS | YES1 | DA764027 | 5' UTR |
| NSUN5B | BI765473 | 5' UTR | YJEFN3 | DA289550 | CDS |
| NSUN5C | BI765473 | 5' UTR | ZFPM2 | BX088883 | CDS |
| NSUN7 | BG718408 | CDS | ZNF470 | CR984692 | 5' UTR |
| NUMA1 | BP277968 | 5' UTR | | | |