Regulated T cell pre-mRNA splicing as genetic marker of T cell suppression

by
Boitumelo Mofolo

Thesis presented for the degree of Master of Science (Bioinformatics)
Institute of Infectious Disease and Molecular Medicine
University of Cape Town (UCT)

Supervisor:
Assoc. Prof. Nicola Mulder

August 2012

Declaration

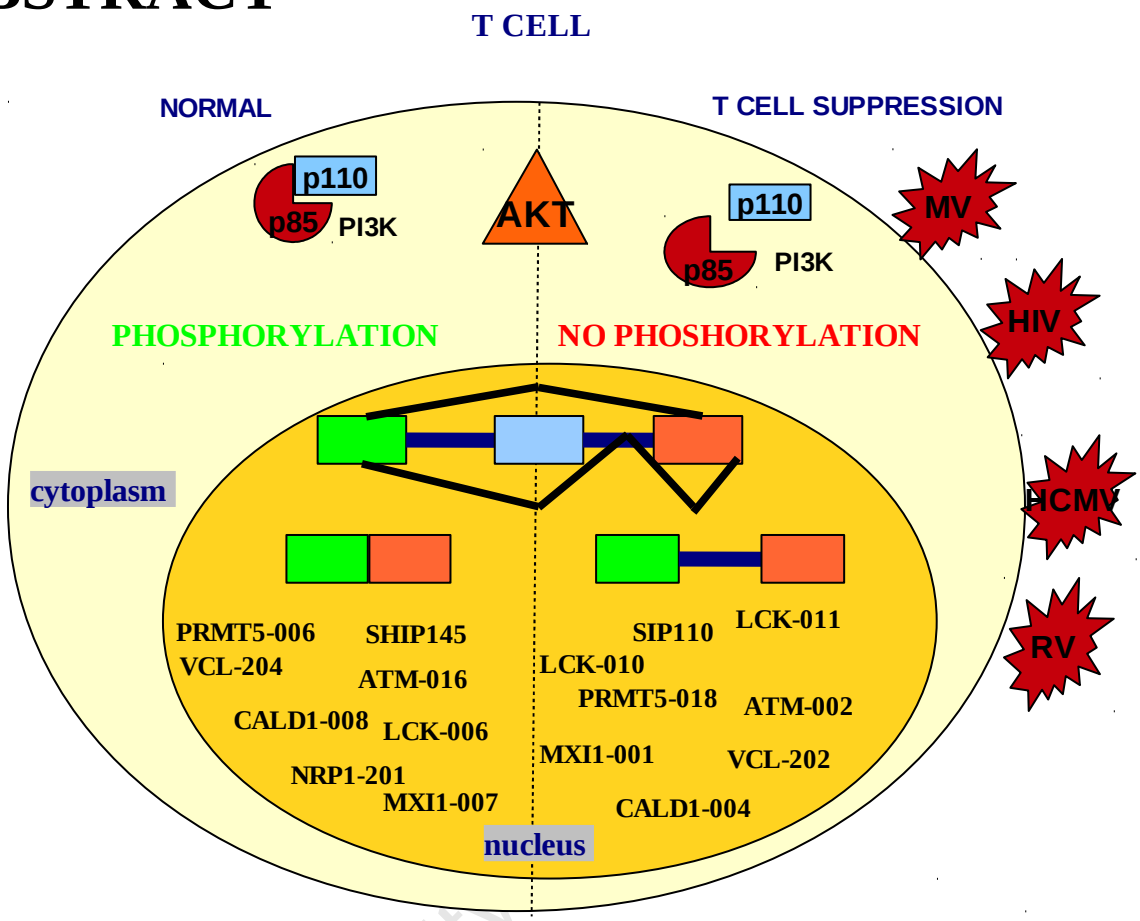
I, Boitumelo Mofolo, declare that all the work in this thesis, excluding that has been cited and referenced, is my own.

Signature Signed by candidate

Signature Removed

Boitumelo Mofolo

ABSTRACT



Background: Measles is a highly contagious disease that mainly affects children and according to the World Health Organisation (WHO), was responsible for over 164000 deaths in 2008, despite the availability of a safe and cost-effective vaccine [56]. The Measles virus (MV) inactivates T-cells, rendering them dysfunctional, and results in virally induced immunosuppression which shares certain features with that induced by HIV. Targets as well as genetic markers distinctive to the MV and common to both (Measles and HIV) have not yet been defined. The MV targets the PI3K/AKT pathway which is involved in a number of cellular processes such as cell growth, proliferation and survival. The MV interferes with the signalling of the pathway and results in the production of alternatively spliced protein isoforms such as the *SIP110*, an alternatively spliced isoform of the phosphatase *SHIP145*, which was shown to directly inhibit T cells [50]. The aim of this project was to determine whether more of these types of alternatively spliced isoforms exist, and could thus be potential genetic markers of T cell suppression.

Methods: A GeneChip Exon array analysis was performed on RNAs isolated from human T cells. In parallel, splicing targets that may be affected by PI3K were identified through the use of publicly available databases such as Genecards and Entrez. Microarray datasets from GEO and ArrayExpress as well as transcript data from dbEST and alternative splicing data from Ensembl and

Genecards were used to support the identified splicing targets. Functional pathway analysis was performed using DAVID for functional clustering and the KEGG Mapper for pathway mapping. InterProScan was used for the protein functionality of the identified isoforms [65], and protein-protein interaction were determined using the STRING database [62].

Results: The genes *ATM*, *CALD1*, *LCK*, *VCL*, *PRMT5*, *NRP1* and *MXI1* were found to show different transcripts under normal versus T cell suppression conditions. *ATM*, *PRMT5* and *VCL* overlapped with the differentially regulated list from the GeneChip exon array whilst *CALD1*, *LCK*, *NRP1* and *MXI1* overlapped with the alternatively spliced list. Functional pathway analysis showed *MXI1* is involved in DNA binding and has transcription repressor activity as well as implications in neurofibrosarcoma and prostate cancer. *CALD1* plays a role in vascular muscle contraction, *LCK* is involved in natural killer cell mediated cytotoxicity and the T cell receptor signalling pathway, *ATM* is involved in apoptosis and plays a role in the p53 signalling pathway, *VCL* is involved in focal adhesion and regulates the actin cytoskeleton, *PRMT5* forms part of the RNA transport pathway and *NRP1* is involved in axon guidance and has been shown to regulate T cell activation at the immune synapse. Low levels of *LCK* have been implicated in severe combined immunodeficiency (SCID) whilst *ATM* is part of a range of cancers including breast cancer, lymphoma as well as ataxia telangiectasia. For some of the genes whose transcripts were found to be expressed under T cell suppression conditions, the biotype was retained introns (*ATM* and *CALD1*) and nonsense-mediated decay (*CALD1 and PRMT5)*, which is the same biotype as the SIP110. Using InterProScan it was determined that protein-coding transcripts found under T cell suppression conditions, seem to lack the protein kinase-like domains and suffered from a loss of conserved motifs that are present in their counterparts expressed in normal conditions. The protein-protein interaction analysis, using STRING, showed that the T cell suppressed isoforms of *LCK* have relationships with genes involved in T cell activation, T cell antigen receptor mediated signalling, T cell mediated killing and assembly and expression of the T cell receptor (TCR). Isoforms identified for *NRP1* under T cell suppression conditions had relationships with genes involved in the cell cycle.

Conclusion: Although challenges to the study included lack of proper annotated EST/cDNAs, weaknesses to the GEO2R and limitations of the interaction tool STRING, potential biomarkers for T cell suppression were identified for *ATM*, *CALD1*, *LCK*, *VCL*, *PRMT5*, *NRP1* and *MXI1*. It was discovered that isoforms found expressed under T cell suppression conditions lacked the protein kinase-like domain and some conserved motifs. Interestingly the proline-rich regions seem to be the 'housekeepers' for the isoforms, as the regions were found conserved in the isoforms expressed under both conditions, the normal and T cell suppression. Further studies would need to be performed in order to better understand how the T cell suppressed isoforms affect the relationships

and an analysis of alternative splicing on the protein interactions.

ACKNOWLEDGEMENTS

To the Deutsche Forschungsgemeinschaft (DFG) and the National Research Foundation (NRF) for funding this study;

To Professor Nicola Mulder for her supervision, support, understanding and patience throughout this study;

To Dr Susanne Kneitz and Professor Sybille Schneider-Schaulies for their support during this study;

To the staff and post-graduate students at the Computational Biology group (CBIO), especially Ayton Meintjes, Jean-Michel Safari Serufuri and Kris Wolfenden for their assistance and friendship;

To the students in the IRTG/NRF programme, especially Alice Riedel and Nigel Makoah Aminake for their friendship and endless support;

To my family, especially my son, for their never-ending belief in me;

I thank you!

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

LITERATURE REVIEW

When the first draft of the Human Genome Project was published in February 2001, researchers were amazed by the findings. The number of human genes appeared to be significantly fewer than the previously thought range of 50,000 - 140,000 genes [6]. The project revealed that the human genome consists of about 30,000 – 40,000 protein-coding genes, only twice as much as the fruit fly [6] but was much more complex with alternative splicing thought to be responsible for generating a lot more protein products. Aberrations in alternative splicing have been implicated in a variety of diseases with large scale studies having been done into biomarkers, using the alternatively spliced products, to distinguish between normal and diseased cells [59].

1.1 Splicing

Splicing is a process whereby introns are removed from the pre-mRNA and exons joined to form a mature mRNA (Figure 1.1). The reaction occurs in the nucleus co- or post-transcriptionally resulting in the mRNA being exported to the cytoplasm for translation into protein products [1].

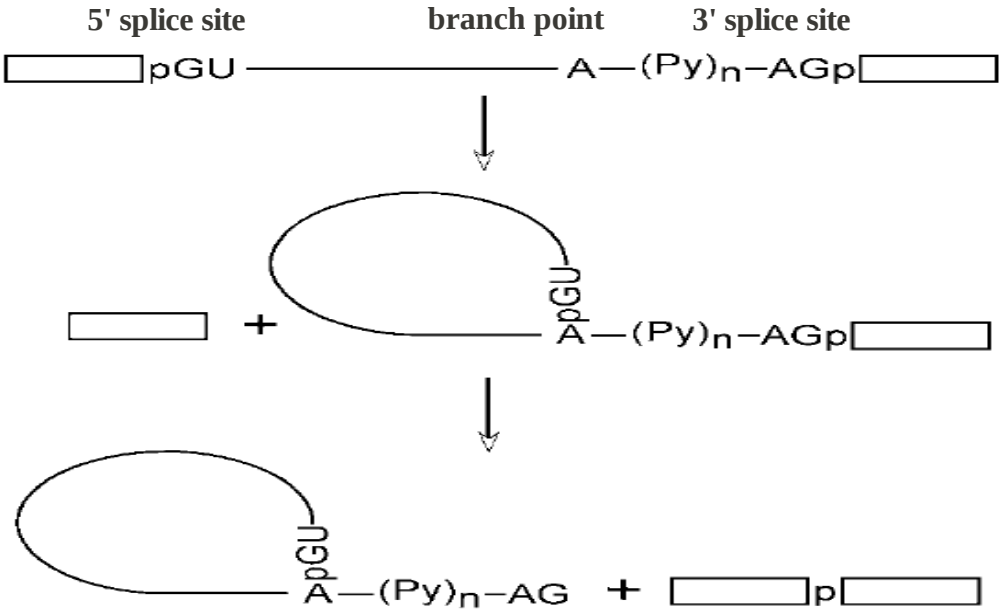


Figure 1.1: The catalytic steps of the nuclear pre-mRNA splicing. Exons are shown as boxes and introns as lines. Step 1 sees the detached 5' exon and an intron/3' exon fragment forming a lariat structure. Step 2 is the ligation of the two exons and the release of the intron lariat to form a mature mRNA [1,3].

Exons make up one tenth of an average pre-mRNA and recognition is tantamount to looking for a needle (exons) in a haystack (introns) [2]. Precision during splicing is very important in order to avoid costly mistakes that can affect the protein-coding ability of a gene [2].

The recognition of the 5' and 3' splice sites of the pre-mRNA is possible due to the presence of the AGGURAGU and the YAG sequences respectively and are limited to the exon/intron borders allowing for the recognition and the excision of introns from the pre-mRNA [1]. A third sequence YNCURAY, called the branch point, is found 18-30 nucleotides upstream of the 3' splice site and allows for the release of the intron post excision (Figure 1.1). The splicing reaction is made possible by the presence of a spliceosome which is a large macromolecular complex that binds onto the aforementioned sequences [3]. The spliceosome is made up of five small ribonucleoproteins (snRNP) that recognise the splice sites and are responsible for the removal of the introns (Figure 1.2).

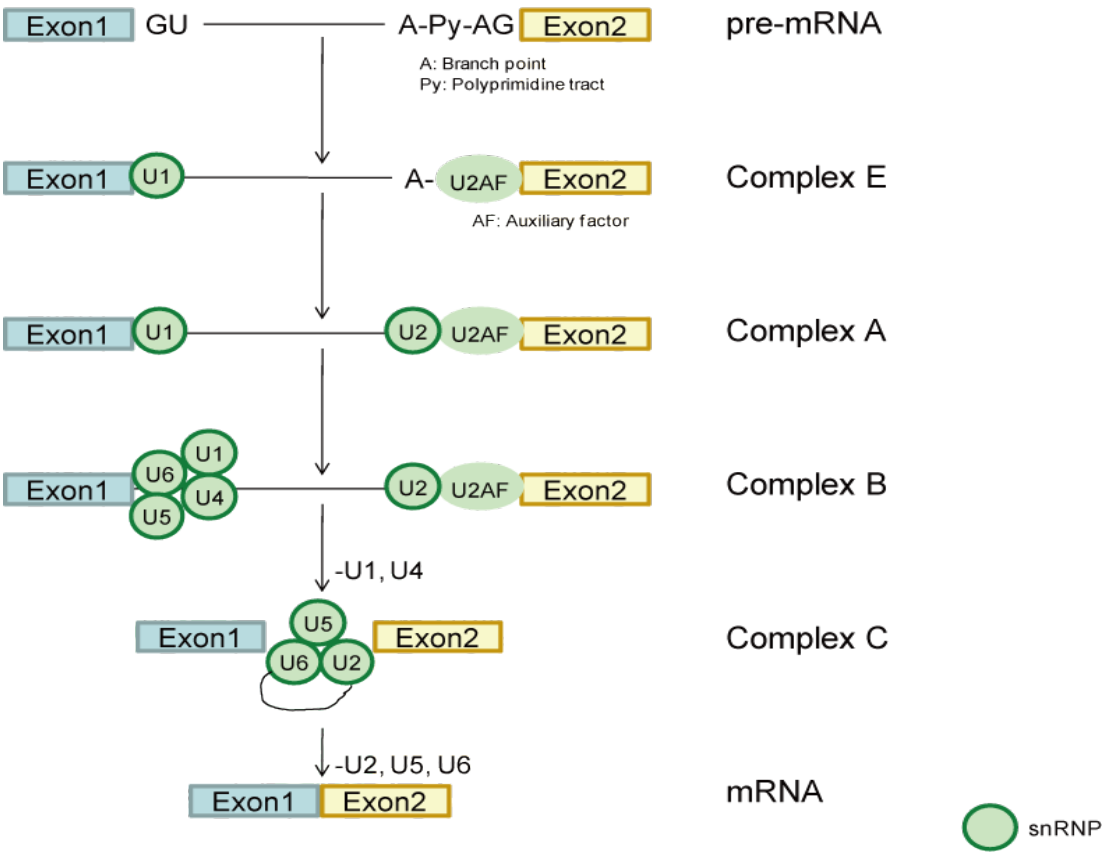


Figure 1.2: The spliceosome complex contains five small nuclear ribonucleoproteins (snRNP) that assemble onto the intron. The complex E contains the U1 snRNP bound to the 5' splice site. The branch point is bound by SF1, the polypyrimidine tract by U2AF 65 and the AG dinucleotide by U2AF 35 as well as the unbound U2 snRNP. When U2 attaches to the branch point via the RNA/RNA base-pairing, complex A is formed. Complex B is formed when complex A is joined by the U4/5/6 tri-snRNP. The interactions of the U1 and U4 snRNPs are lost during the rearrangement of complex B to form the catalytic complex C. The formation of complex C results in splicing [51].

Although the three sequences are necessary for the splicing process, they are not sufficient. In addition to the splice site sequences, other splicing regulatory elements (SREs) aid the splicing process. The SREs are classified as exonic splicing enhancers (ESEs) or silencers (ESSs) if they promote or inhibit the inclusion of the exon where they reside, and as intronic splicing enhancers (ISEs) or silencers (ISSs) if they enhance or inhibit the inclusion of the exon adjacent to the intron where they reside (Figure 1.3) [4].

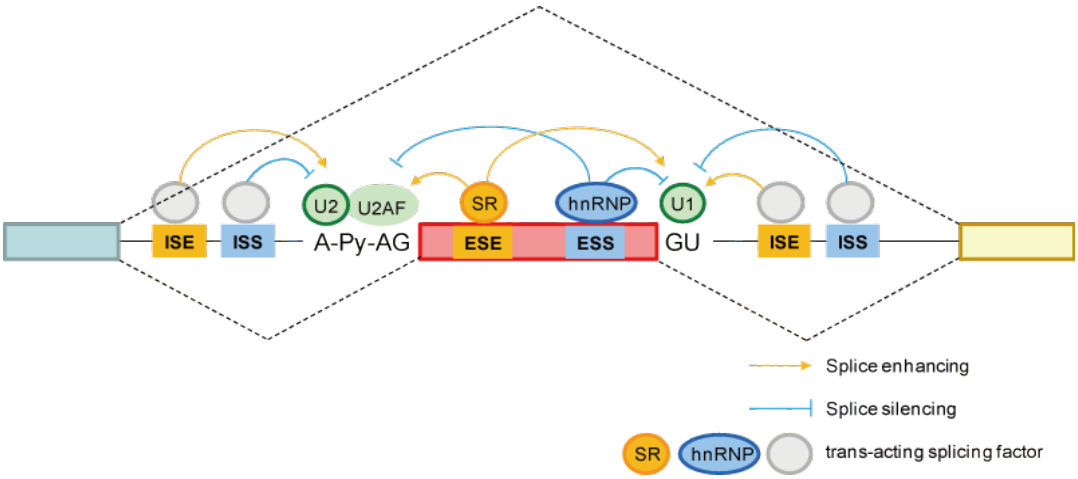


Figure 1.3: Splicing regulatory elements (SRE) . The exonic splicing enhancer (ESE) and intronic splicing enhancer (ISE) are bound by the positive regulator, the SR proteins whilst the exonic splicing silencers and intronic splicing silencers are bound by the negative regulator, the hnRNP proteins [51].

The enhancer elements are bound by positive regulators (SR) whilst silencer elements are bound by negative regulators (hnRNP) as illustrated in Figure 1.3 above. An example of a positive regulator is the SR protein family that contains more than one RNA binding domain as well as a serine/arginine-rich domain whilst the hnRNP proteins are an example of a negative regulator [2]. A fine balance between the regulators is necessary for controlling the level of exon inclusion in the mRNA transcript [2,5].

As mentioned previously, the first draft of the Human Genome Project revealed a much lower number of protein-coding genes than expected, and alternative splicing is thought to be responsible for generating a larger diversity of protein products.

1.2 Alternative splicing

Alternative splicing is the differential removal of exons from mature mRNA or the inclusion of intronic sequences. One gene can produce many protein products partially explaining the genome complexity of eukaryotes (Figure 1.4). It is now believed that up to 94% of human genes are alternatively spliced and that splicing is affected by an estimated 50% of disease causing mutations [2].

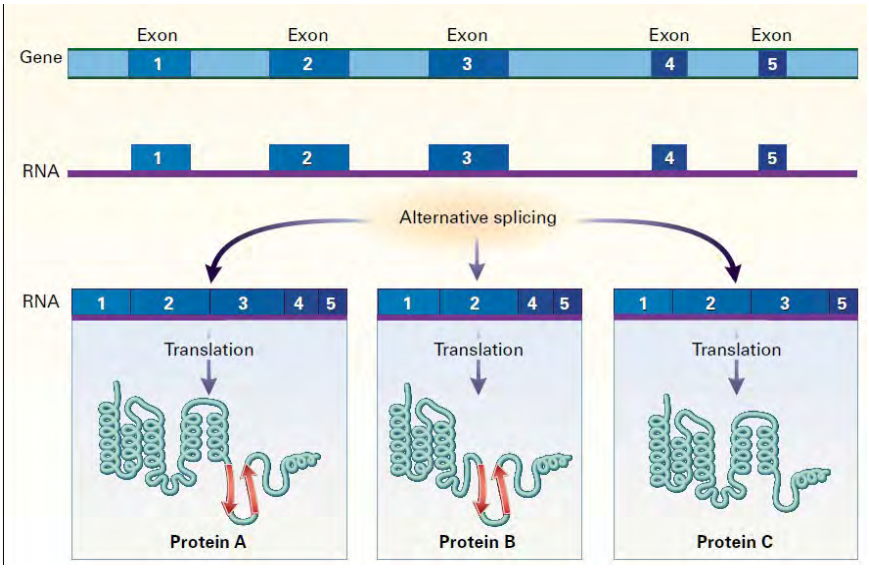


Figure 1.4: Alternative splicing. The process can produce many protein products that are not only structurally different but performs various functions too, all derived from a single mRNA [46].

As illustrated in Figure 1.5 below, different combinations of exons can be spliced together to produce different mRNA isoforms of a gene that encode not only structurally different but also functionally different protein products [7].

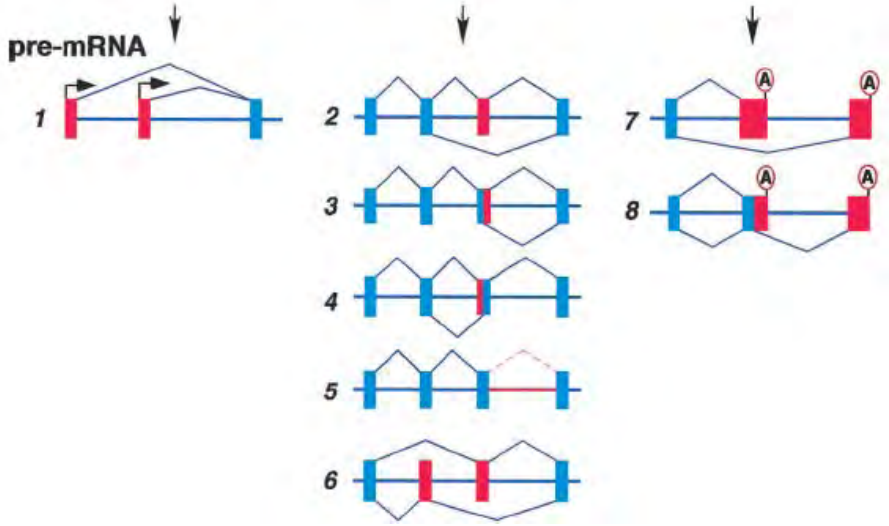


Figure 1.5: Different types of alternative splicing events [48]. 1. Alternative promoter 2. Cassette exon 3. Alternative 5' splice sites 4. Alternative 3' splice sites 5. Intron retention 6. Mutually exclusive 7. and 8. alternative terminal exons.

Most exons are constitutive as they are always spliced or included in the final mRNA product, there is no alternative splicing [3]. The most common type of alternative splicing event is the cassette exon whereby the regulated exon is sometimes included or excluded from the mRNA (Figure 1.5: 2). Multiple cassette exons are what are termed mutually exclusive exons (Figure 1.5: 6), whereby mRNAs consisting of one of several possible exon choices are produced [3]. Alternative 5' and 3' splice sites (Figure 1.5: 3 and 4) results in the lengthening or shortening of a particular exon whilst alternative promoters (Figure 1.5: 1) and alternative polyadenylation sites switch the 5' and 3' exons of a transcript (Figure 1.5: 7 and 8).

A normal protein-coding human gene has on average 8.8 exons which are usually short in length (10 to 400 nt), whilst introns are usually 10 times or more the length of the exons (200,000 nucleotides or more) [1,8]. Exons have always been and are known to have coding ability whilst introns used to be referred to as 'junk DNA' (non-coding DNA)[9]. Some introns are now known to have coding ability, but at a much smaller scale than exons [1]. As an example, introns have been shown to play roles in gene regulation via alternative splicing as well as in nonsense-mediated decay (NMD) [10-12]. When introns are retained within an mRNA (Figure 1.5: 5), in frame stop-codons are usually introduced and prematurely terminated proteins are translated resulting in NMD [1]. NMD thus functions to down-regulate gene expression. Barash et al (2010) noticed that a certain class of alternative exons that introduce a premature termination codon (PTC) and activate

NMD, are found in adult tissues and suppress mRNA expression but are skipped in embryonic tissues resulting in mRNA expression [10]. Some introns have been shown to enhance gene expression, known as intron-mediated enhancement (IME) and are found to occur amongst humans, plants, insects and mice [13].

1.2.1 Regulation of alternative splicing

Studies have demonstrated that cells frequently regulate alternative splicing events in response to external stimuli, cell type, developmental stage and gender [8]. The SREs (see the splicing section) function to regulate alternative splicing by enhancing (ESEs and ISEs) or silencing (ESSs and ISSs) splicing (Figure 1.3). In addition to these SREs, RNA binding proteins belonging to the SR protein and the hnRNPs families are known to promote and inhibit splicing respectively (Figure 1.6) [14]. Changes in splice site selection are influenced by phosphorylation events that are regulated by known signal transduction pathways and the SR-proteins are known to be phosphorylated by several kinases such as the cdc2 like kinases and Clk1-4 [8]. Phosphorylation of the SR-protein SF2/ASF increases its binding to the U1 70K snRNP promoting exon inclusion [8]. The phosphorylation is reversed by phosphatases such as protein phosphatase 1 (PP1), PP2A and protein phosphatase 2Cgamma and the complete blocking of these inhibits splicing as dephosphorylation is necessary for the transesterification step (Figure 1.2) [8]. Both in vitro and in vivo studies have shown that changing levels of the phosphatase has an influence on alternative exon usage suggesting that phosphatases can be used by cells to regulate alternative splicing [8].

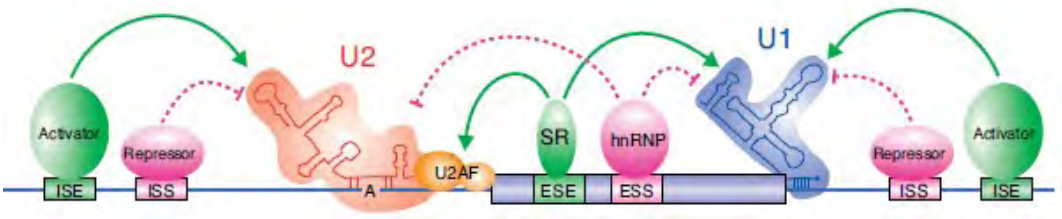


Figure 1.6: The sequences and proteins involved in alternative splicing. In addition to the SREs, SR proteins and hnRNPs are known to promote and inhibit splicing respectively [14].

A lesser known topic of the splicing regulatory mechanisms are the impact of the pre-mRNA structures on alternative splicing. This very topic has been reviewed by McManus, C.J. et al (2011) and they demonstrate that these pre-mRNA structures (both local and long-range structures) regulate alternative splicing preventing the 5' and 3' splice site recognition by the spliceosome [14].

Over the past few years, literature has provided a link between aberrant alternative splicing and different aspects of human diseases with the association mostly between splicing variants and cancer [15].

1.2.2 Alternative splicing and disease

The regulatory elements mentioned above can be altered and cause aberrant regulation of alternative splicing with the changes manifesting as disease. Mutations that affect splicing can cause disease directly or contribute to the susceptibility or severity of disease [2]. See Table 1.1 for some examples. Changes in splicing patterns associated with many cancers are important in the transformation, motility and metastasis of the tumour tissue [16]. In the case of cancers, alternative splicing is known to affect hormonal signalling, apoptosis as well as cell -cell or matrix interactions [8,16].

Examples of abnormally spliced genes detected in cancers include *CD44*, *MUC2*, *SRF*, *NCAM*, *MLH*, *MSH* as well as members of the Wnt pathway [16]. These genes may be used as markers of the diseases or as drug targets but limited larger clinical trials as well as factors such as individual patient differences, tissue complexity and lack of tools pose some challenges [16].

Table 1.1: Examples of disease causing mutations related to splicing

Mutation	**Disease**	**Gene Involved**	**Type of Mutation**	**Mutation effect**	**Symptoms**
Intron mutations that alter splice-site recognition	Familial Dysautonomia (FD)	IKBKAP gene, coding for IKAP	A silent T to C transition in the 6th base of the intron leads to skipping of exon 20	Introducing a PTC resulting in NMD of the mRNA and decreased expression of IKAP	It's a disorder of the autonomic nervous system. Patients are insensitive to pain, unable to produce tears, poor growth
Exon mutations that disrupt splicing	Spinal muscular atrophy (SMA)	Survival motor neuron (SMN) protein encodes for SMN1 and SMN2	SMN2 can't compensate due to a silent C to T substitution in the 6th nucleotide of exon 7 resulting in exon skipping	Production of a truncated, inactive protein	Infant mortality and motor neuron degeneration resulting in progressive paralysis

Mutation	Disease	Gene Involved	Type of Mutation	Mutation effect	Symptoms
Splicing as a genetic modifier of disease	Menkes disease (MD) and occipital horn syndrome (OHS)	ATP7A gene, encodes a copper-transporting ATPase	Mutations in the invariant dinucleotides at the 5 and 3 splice sites eliminate exon recognition cause MD whilst OHS is due to mutations occuring in the less conserved regions	With MD, the expression of the correct splice variant is drastically reduced or even eliminated. With OHS, the splice variants are still produced.	neurological degeneration, connective tissue defects, distinctive kinky and brittle hair as well as early-childhood death
Mutations that alter the ratio of alternatively spliced isoforms	Frontotemporal dementia and Parkinsonism linked to chromosome 17 (FDTP-17)	MAPT gene encodes for tau, a protein involved in microtubule assembly and stability	Mutations in exon10 alter the ratio of tau isoform containing either 3R or 4R repeat microtubule binding sequences	Mutations in exon10 alter the ratio of tau isoform containing either 3R or 4R repeat microtubule binding sequences	Difficulty walking, rigidity, tremor or muscle weakness, difficulty naming familiar objects
Mutations in regulators of alternative splicing	Amyotrophic lateral sclerosis (ALS)	TDP-43		Protein abnormally included in ubiquitinated protein aggregates	
Unstable nucleotide repeat expansions	Myotonic dystrophy (DM)	DMPK and ZNF9 genes	DMI is due to a CTG trinucleotide expansion in the 3 UTR of the DMPK gene. DM2 is due to a CCTG tetranucleotide expansion within the first intron of the ZNF9 gene		Skeletal muscle wasting, cataracts, myotonia, insulin resistance

The abnormally spliced genes are usually not mutated but the defects are as a result of a change in the nuclear environment that regulates splice site choice [2].

1.2.3 Approaches to detection of alternative splicing

Several approaches have been developed or used over the years to study alternative splicing. Techniques such as microarrays, high-throughput sequencing analysis, and ESTs have all been used with varying degrees of success to understand alternative splicing.

1.2.3.1 Expressed sequence tags (ESTs)

Several genome-wide studies for computational detection of alternative splicing, using expressed sequence tags (EST) have been carried out over the past decade [17-22]. ESTs are 'single-pass' cDNA sequences (~360bp) and are a rich source of information for understanding gene expression and structure (Figure 1.7).

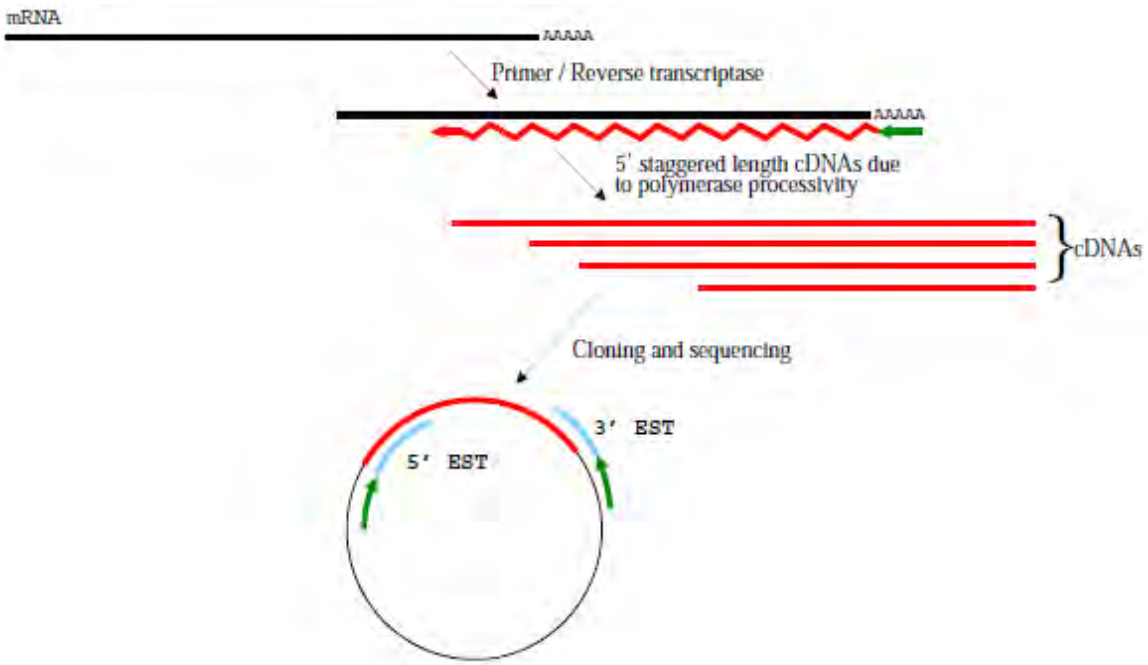


Figure 1.7: Expressed Sequence Tags (EST) are short (200-800 nt) and single-pass sequences derived from cDNA libraries. RNA is reverse transcribed to a double-stranded cDNA using reverse transcriptase. The cDNA is then cloned and randomly sequenced from both directions in a single-pass run to obtain 5' and 3' ESTs [23].

Earlier studies have relied heavily on ESTs to study alternative splicing as they represent a wide variety of different tissue types both diseased and normal as well as immortalised cell lines whilst providing a 'snap-shot' of human development (from embryos to late adulthood) [19]. ESTs are also important in gene discovery, genome annotation, gene structure identification, characterization of

splice variants, SNP data mining as well as proteome analysis [23]. Accessibility is also one of the major reasons why ESTs are still popular for examining alternative splicing. They are accessible from public databases such as the GenBank dbEST. As of the 1st January 2012, the organism with the highest number of EST entries was *Homo sapien*s at 8,315,294 followed by *Mus musculus* at 4,853,562 [52].

Although ESTs provide an ideal tool for alternative splicing detection, this is not without challenges or problems. Studies of alternative splicing using ESTs are hampered by protocol differences, transcript end bias, and library coverage limitations (Table 1.2) [24].

Table 1.2: Issues with high-throughput alternative splice detection [22]

Experimental factors	*Type of errors caused*: (+) false positive (–) false negative
EST coverage limitations, bias	(–). Most genes have very few ESTs, from even fewer tissues. The main barrier to alternative splice detection.
RT / PCR artifacts	(+) in methods that don't screen for fully valid splice sites (which requires genomic mapping, intronic sequence).
Genomic coverage, assembly errors	(–) in methods that map ESTs on the genome. Short contigs may cause >25% false negatives.
Chimeric ESTs	(+) in methods that simply compare ESTs.
Genomic contamination	(+) in methods that don't screen for pairs of mutually exclusive splices.
EST orientation error, uncertainty	(+/–) in methods that don't correct misreported orientation, or don't distinguish overlapping genes on opposite strands.
Sequencing error	(+/–). Single-pass EST sequencing error can be very high locally (e.g. >10% at the ends). Need chromatograms.
EST fragmentation	Where ESTs end cannot be treated as significant.
Bioinformatics factors	
Mapping ESTs to the genome	(–) in methods that map genomic location for each EST.
Paralogous genes	(+) in all current methods, but mostly in those that don't map genomic location or don't check all possible locations.
Rigorous measures of evidence	(+/–). How can the strength of experimental evidence for a specific splice form be measured rigorously?
Arbitrary cutoff thresholds	(+/–) in methods that use cutoffs (such as '99% identity').
Alignment size limitations	(–) in methods that can't align $>10^2$, $>10^3$ sequences.
'Pathological' assemblies	(+/–). What should assembly programs do when the assembled reads disagree in regions (such as alt. splicing)? Programs vary.
Nonstandard splice sites?	(+) in methods that don't fully check splice sites; (–) in methods that do restrict to standard splice sites.
Alignment degeneracy	(+/–). Alignment of ESTs to genomic is frequently degenerate around splice sites.
Biological interpretation factors	**Comments**
Spliceosome errors?	Is splicing perfect? That is, does it only make correct forms?
What is truly functional?	Just because a splice form is real (i.e. present in the cell) doesn't mean it's biologically functional. Conversely, even an mRNA isoform that makes a truncated, inactive protein might be a biologically valid form of functional regulation.
Defining the coding region	Predicting ORF in newly discovered genes; splicing may change ORF.
Predicting impact in protein	Motif, signal, domain prediction, and functional effects.
Predicting impact in UTR	Knowledge about effects of mRNA stability, localization, and other possible UTR sequences is incomplete.
Assessing and correcting for bias	Our genome-wide view of function is under construction. Until then, we have unknown selection bias.

When trying to identify alternatively spliced isoforms, ESTs from the same gene are searched for differences that are consistent with alternative splicing, for example a large insertion or a deletion (Figure 1.8). Each candidate splice product can be further assessed by mapping the ESTs exactly to their gene sequence in the draft genome [22].

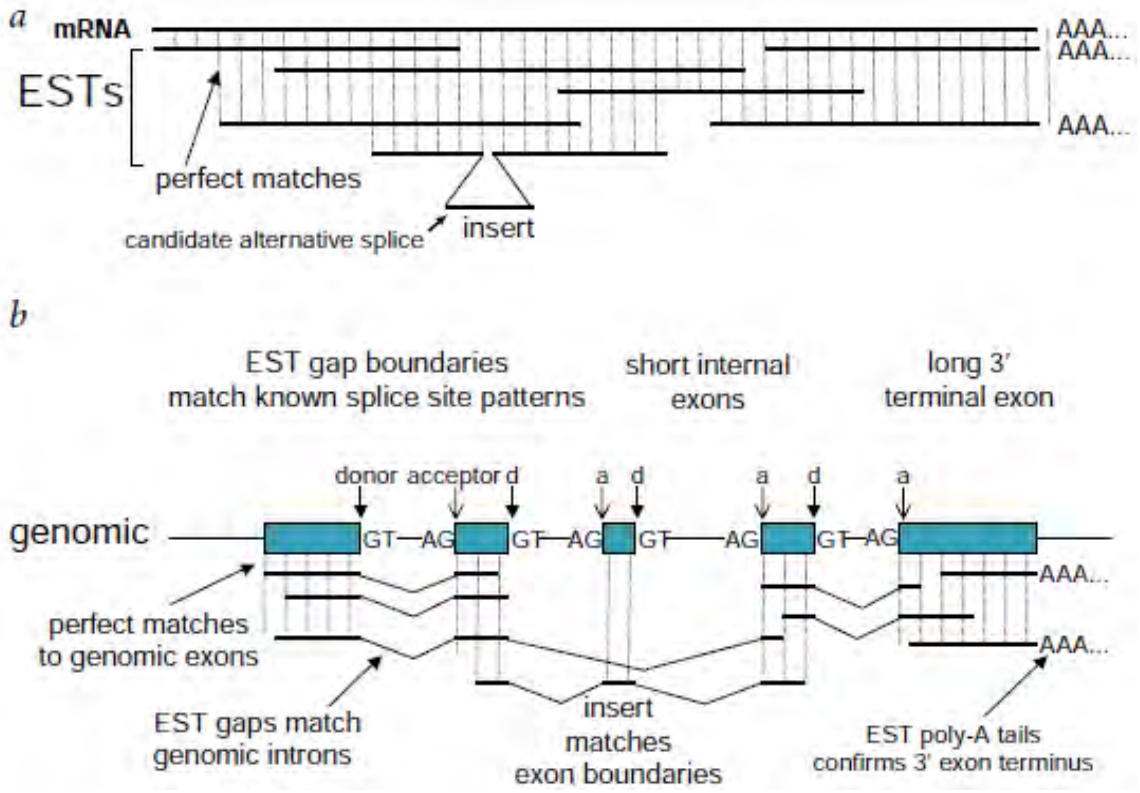


Figure 1.8: Computational identification of alternative splicing. a, insertion and deletion in ESTs relative to mRNA are identified as potential alternative splices. b, splice products are identified and intronic splice junction donor and acceptor sites are checked. Alternative splice products are detected when two exons are mutually exclusive [22].

1.2.3.2 Microarrays

The role that microarrays play in detecting splicing events was suggested as early as 2000 [24,25]. Exon microarray technology enables genome-wide quantification of expression levels of the majority of exons and helps in detecting alternatively spliced isoforms [26]. The most popular method to detect alternative splicing is the Splicing Index (SI), which identifies exons that have different inclusion rates relative to the gene level, between two groups [26,53]. An SI of 0 indicates equal inclusion rates of the exon in both samples, positive values indicate enrichment of that exon in Sample 1, and negative values indicate repression or exon skipping in Sample 1 relative to Sample 2 [53].

There are different types of microarrays. The 3' expression microarrays are the earliest forms of microarrays and are used to approximate the expression of the entire gene. This approach assumes that the 3' end of a gene is clearly defined, has an intact poly-A tail and that the entire length of the gene is expressed as a single unit [54]. This type of microarray doesn't discriminate between alternatively spliced transcripts that have identical 3' ends; transcripts lacking a 3' exon because of alternative splicing, or non-polyadenylation, and genomic deletions etc., are not detected in the 3' based expression experiments [54].

The exon microarrays on the other hand, such as the Affymetrix Human Exon 1.0 ST Array, are used to detect alternative splicing. The microarray platform contains probes for 80% of human exons and detects differential use of specific exons and allows one to distinguish between different splice variants, to quantify expression levels for exons and genes [2,26].

1.2.3.4 Next generation sequencing

RNA Sequencing (RNA-Seq), like ESTs, are generated from cDNAs but are different as they are sequenced using high-throughput Illumina or other next generation sequencing technologies and are then used to define exons, 5' and 3' boundaries and introns as well as gene expression levels [27]. An advantage of using RNA-Seq is that RNAs expressed at very low levels can be detected and quantified.

1.3 T cell silencers

The immune system of a healthy organism maintains immune balance between tolerance and active response. Under normal physiological conditions, the immune balance is a tightly regulated network of several types of immune cells. If this is disturbed, the response can be either inefficient (as in cancer) or, conversely, over-reactive, resulting in conditions such as autoimmunity [28]. Maintaining immune tolerance is vital for organ transplantation and T cells play an important role in determining whether the organ gets rejected or not [29]. Various studies on organ transplantation [29] target T cells and have demonstrated that suppression of the T cells is important in maintaining immune tolerance. *VSIG4*, a member of the B7 family-related proteins, was identified as a T cell silencer [30] whilst dendritic cells transfected with hVSIG4 recombinant adenovirus and shown to cause T cell suppression, was suggested as a way of maintaining immune tolerance for organ transplantation or autoimmune diseases [31]. Another T cell silencer that has been identified is the *B7-H1* or *PD-L1* (a B7 member), a PD-1 ligand [32,33].

A multitude of pathogens that cause infections often result in the immune system being suppressed. The Measles Virus (MV) and the Human Immunodeficiency Virus (HIV) are such pathogens [34,35]. In Vienna, von Pirquet first described measles virus-induced immunosuppression as represented by the disappearance of delayed-type hypersensitivity (DTH) skin test responses to tuberculin in 1908. His work with measles virus provided the first evidence that viruses have the ability to suppress the immune system [55]. In this section, the MV is used to illustrate immunosuppression but do note that MV resembles HIV in many aspects except for being less severe and its spontaneous reversibility [36].

MV belongs to the family *Paramyxoviridae*, subfamily *Paramyxovirinae*, genus *Morbillivirus.* Other members of the *Morbillivirus* genus, specifically canine distemper virus and rinderpest virus, have also been shown to cause immunosuppression [37,38]. Even though the highly contagious measles virus is contracted once in a lifetime and is vaccine-preventable, more than 350 000 children succumb to the disease annually [35,56]. Symptoms of the infection include rash, conjuctivitis, fever and lethal complications due to its suppressive nature and results in opportunistic infections such as pneumonia and encephalitis [35,36].

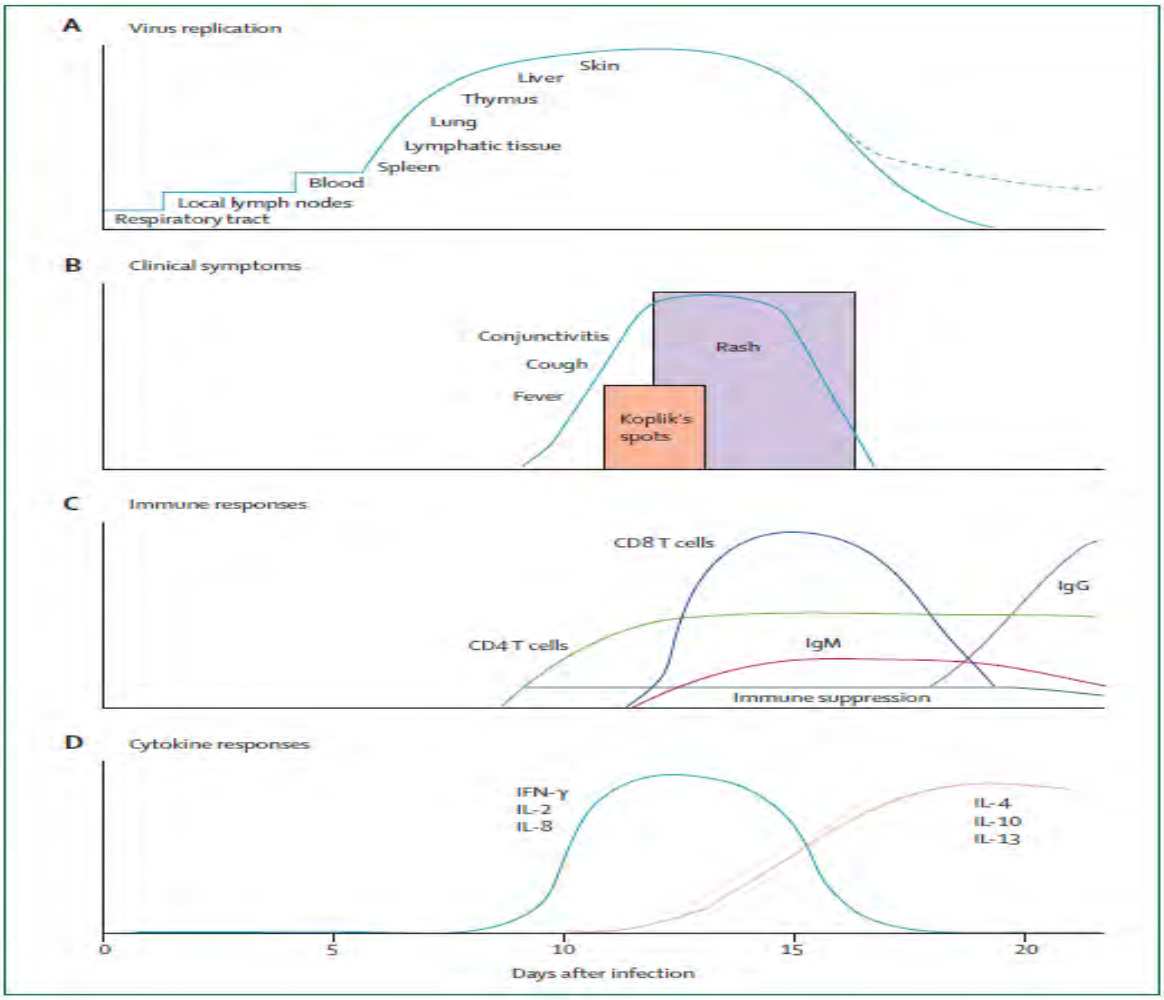


Figure 1.9: Schematic diagram of the pathogenesis of measles from virus infection to recovery. ***A****:The measles-virus infects the respiratory tract and then spreads to the other organs.* ***B****: Symptoms include fever, cough and rash that begins when the virus starts clearing.* ***C****: The CD8+ and CD4+ T cells appear at the same with the CD4+ T cell activation being prolonged. The measles-virus specific IgM is used as a marker for diagnosing measles. During the acute disease and post recovery, the immune system remains suppressed.* ***D****: The production of cytokines aid in viral clearance (IFN-γ) as well as developing antibodies (IL-4 and IL-10) [39].*

The virus causes infection by entering the respiratory tract mucosa of the host by droplets or aerosol dispersion and replicates in both the upper and lower respiratory tracts. The virus then spreads to the lymph nodes entering the bloodstream and then spreads to multiple organs such as the skin, gastrointestinal tract, liver, central nervous system and the thymus (Figure 1.9A) [39]. The virus replicates mainly in the epithelial and endothelial cells of blood vessels, in organs, in macrophages, and monocytes [36,39].

During the acute phase of infection, high IFN-γ concentrations aid in viral clearance (Figure 1.9D). Cytokines IL-4 and IL-10's concentrations (secreted by CD4+ T cells) that are crucial in the

development of antibodies, increase once the infection has been brought under control (Figure 1.9D) [36,39]. After the measles virus infection, MV-induced immunosuppression occurs due to a delayed-type hypersensitivity response to antigens such as tuberculin (initially described by von Pirquet) [39].

The MV glycoprotein complex made up of the fusion and hemagglutinin proteins actively silence T cells by interfering with signalling pathways essential for T cell activation [35]. The signalling pathway targeted by MV is the phosphatidyl-inositol-3-kinase (PI3K)/AKT kinase pathway in T cells. The pathway is important in a number of cellular processes such as cell growth, proliferation and survival and interference of the signalling by MV results in the arrest of the cell cycle in T lymphocytes leading to T cell suppression.

PI3Ks are lipid kinases comprising 8 members that are further divided into 3 classes, based on sequence homology, substrate preference and tissue distribution (Table 1.3) [40,41].

Table 1.3: PI3K family members [41]

Class	Catalitic subunit	Regulatory subunit	Activation	Products
Ia	p110α p110β p110δ	p85	RTK, RAS	PtdIns-3,4,5-P_3 PtdIns-3,4-P_2 PtdIns-3-P
Ib	P110γ	p101	Heterotrimeric G proteins	PtdIns-3,4,5-P_3 PtdIns-3,4-P_2 PtdIns-3-P
II	PI3KC2α PI3KC2β PI3KC2γ		RTK, integrins	PtdIns-3,4,-P_2 PtdIns-3-P
III	VSP34p			PtdIns-3-P

The class I PI3K kinases are the most frequently implicated in cancer [41]. The p110α and it's regulatory subunit p85 (it binds and inhibits the p110α) are mostly involved in cell division regulation and tumorigenesis whilst the p110 β, δ, and γ have no known oncogenic mutations [40]. The p110α and the p85 both consist of 5 domains: the p110α, which is encoded by the PIK3CA gene, has a kinase with a C-terminal domain, an adaptor binding domain that joins the kinase, a

Ras-binding domain (RBD), a C2 domain that joins the p85 subunit and a helical domain. The p85's domains include the N-terminal Src homology-2 (SH2) and the inter-SH2 domains [40].

AKT is a serine-threonine kinase consisting of 3 family members namely, Akt1, 2 and 3 which are encoded by *AKT1*, *AKT2* and *AKT3* respectively and are a crucial downstream effectors of PI3K.

Under normal physiological conditions, when the PI3K is activated, AKT phosphorylates and regulates the activity of a number of targets that includes kinases, transcription factors and other regulatory molecules [41]. Through this phosphorylation, AKT regulates a host of cell functions such as glucose metabolism, cell proliferation as well as cell survival [41].
The signalling cascade is usually activated by tyrosine-kinase receptors such as the insulin-like growth factor-1 (IGF-1) on the T-cells and is responsible for activating the PI3K. The PI3K, i.e. the p110α, phosphorylates the phosphatidylinositol-4,5-biphosphate (PIP2) to produce phosphatidylinositol 3,4,5-triphosphate (PIP3). The tumor suppressor phosphatase and tensin homolog deleted on chromosome 10 (PTEN) and p110β acts as an antagonist and dephosphoryates PIP3 to PIP2 thus terminating PI3K-dependent signalling [41]. PIP3 functions to bring the phosphoinositide-dependent kinase 1 (PDK1) and AKT, into close proximity which results in PDK1 phosphorylating and activating AKT, which in turn phosphorylates and activates downstream target proteins that are important in different cellular functions (Figure 1.10).

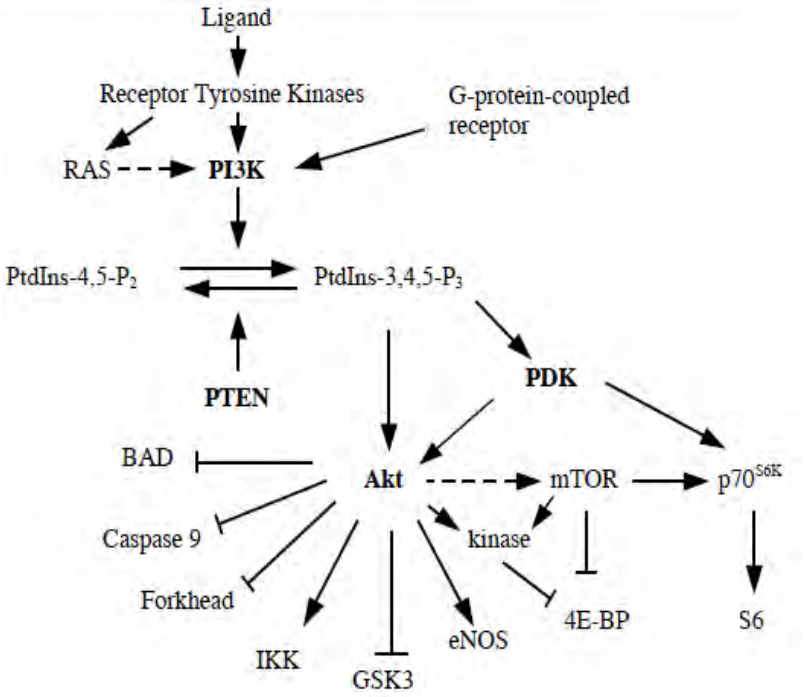


Figure 1.10: PI3K/AKT pathway under normal conditions [41]. Post activation, the PI3K phosphorylates AKT which in turn activates downstream targets that play an important role in cellular functions.

1.4 Alternative splicing and virally-induced immunosuppression

According to [35], the MV is the only pathogen known to interfere with the activation of the pathway in a contact-dependent manner. Figure 1.11 illustrates what happens when the MV interferes with the PI3K/AKT pathway.

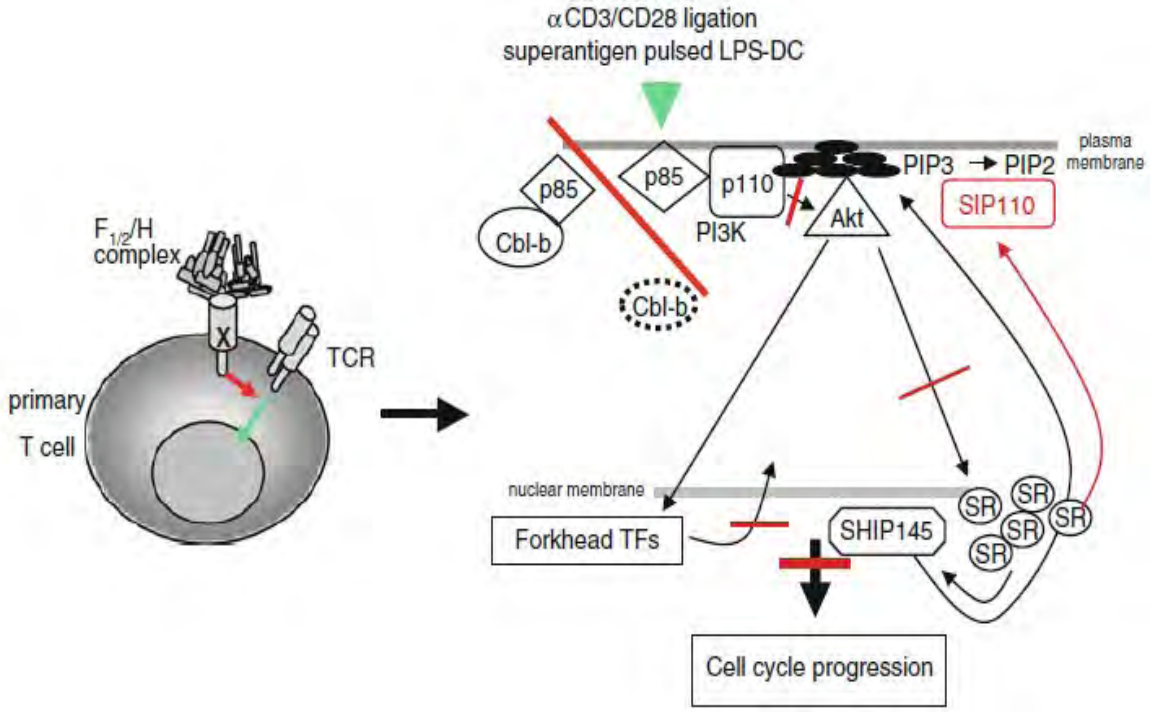


Figure 1.11: MV interferes with the PI3K/AKT signalling pathway [35]. The interference is as a result of the MV preventing degradation of the Cbl-b protein which in turn stops the PIP3 from activating AKT. As a result downstream targets end up not getting phosphorylated.

The virus prevents the degradation of the Cbl-b proteins which results in the accumulation of PIP3 thus preventing the activation of Akt kinase. This in turn results in downstream targets such as forkhead transcription factors as well as splice regulatory proteins (SR) not being phosphorylated. Studies have shown that splice isoforms get produced due to the unphosphorylated splice regulatory proteins. One such isoform is the *SIP110* which is a product of the lipid phosphatase *SHIP145* [35]. PIP3 formation is reduced by the presence of the *SIP110* resulting in the threshold level of T cell activation being increased, leading to T cell suppression [35].

1.5 Problem Identification

1.5.1 Developing biomarkers for T cell suppression

At present, serology is the most common method of confirming measles infection in the laboratory [39]. The measles specific antibody, IgM, is usually used to diagnose measles but is detectable only 4 days or more after a rash manifests and falls to undetectable concentrations within 4-8 weeks of rash onset [39]. A point of care diagnostic test for measles is needed similar to the rapid diagnostic test for malaria which ideally could be done with oral fluid samples [39].

As an estimated 94% of human genes undergo alternative splicing it is important to know which spliced variants are expressed, their relative abundance as well as their biological function in order to understand a gene's function under normal physiological conditions as well as under disease [2]. Spliced variants have been shown to play an important role in the progression and detection of cancer [84, 85] and include the spliced variants of the genes *CD44*, *WT1*, *BRCA1*, *MDM2*, *FGFR* etc. As an example, isoform-specific antibodies for the gene *CD44* (specifically the *CD44v10*) have been developed and used in the differentiation of metastatic and nonmetastatic pancreatic cancer cells [84]. Detection of the CD44v10 using immunohistochemical methods could potentially be used clinically, to diagnose gastric carcinoma progression in patients [84]. Other standard techniques such as RT-PCR and high-throughput methods such as microarrays can be used to detect the splice variants (biomarkers) in a clinical setting [84].

Distinctive genetic markers for MV-induced T cell suppression haven't as yet been defined. The production of alternatively spliced *SIP110* by the lipid phosphatase *SHIP145* due to PI3K interference, might be one of many undiscovered isoforms that can be used as genetic markers for T cell suppression. The main aim of this study was to carry out an analysis of cDNA/EST libraries as well as exon array data to identify potential alternatively spliced isoforms that are T cell suppression-specific. The project aims to address whether more of these types of alternatively spliced isoforms exists, and could thus be potential genetic markers for identifying T cell suppression.

1.6 Specific objectives

- Analysis of publicly available microarray data and EST/cDNA databases for the alternatively spliced isoforms of the PI3K targets
- Pathway and co-expression analysis of the PI3K targets
- SNP analysis of the alternatively spliced isoforms

CHAPTER 2

MATERIALS & METHODS

The methods used to identify potential markers of T cell suppression in this project include the collection or generation of relevant data, the analysis of the data, merging of data from different sources, and then biological interpretation through SNP and functional analysis. A workflow diagram outlining the steps taken, is provided in Figure 2.1.

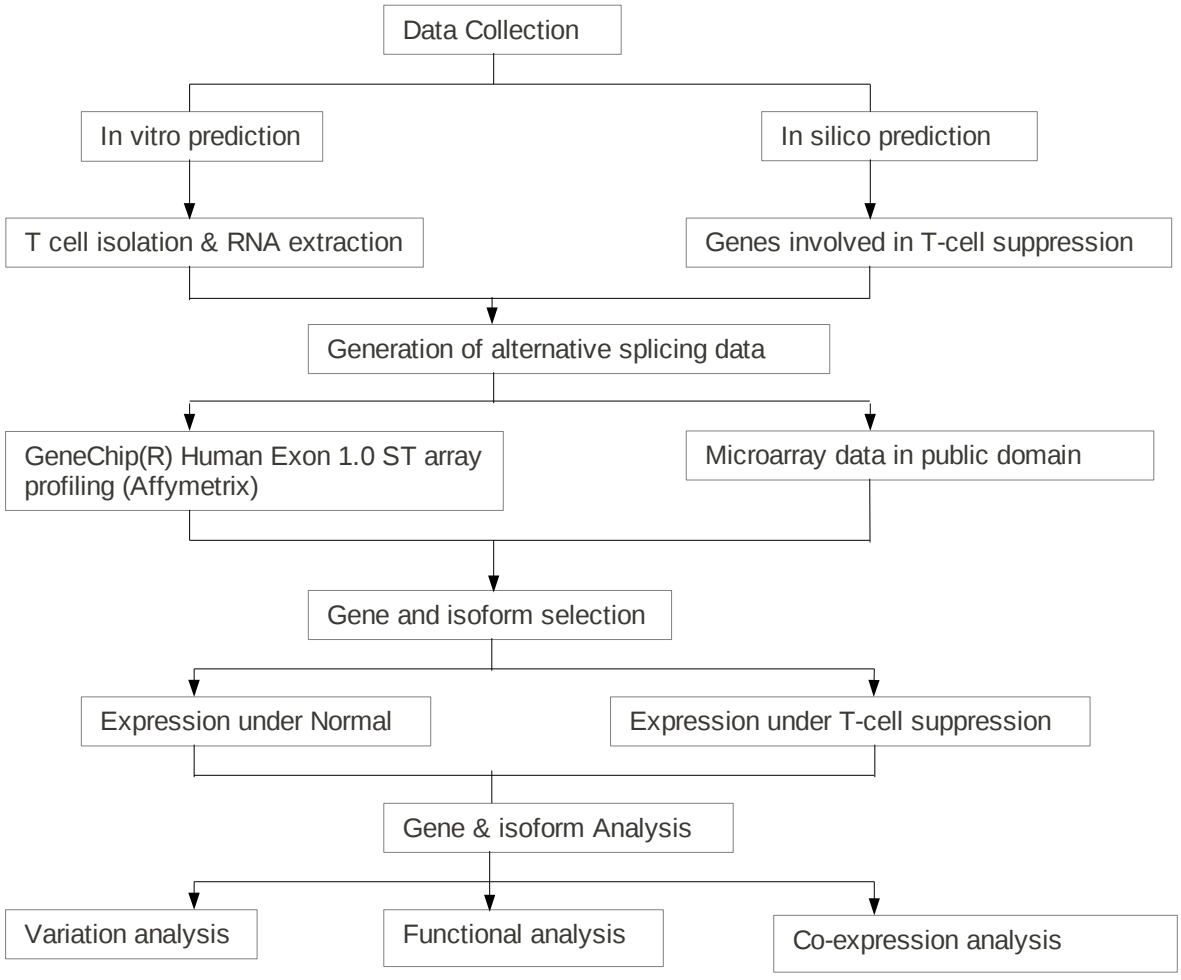


Figure 2.1: Workflow diagram.

2.1 Data collection and generation

In addition to data generated by collaborators within the project, public data was retrieved from a number of different sources listed in Table 2.1.

Table 2.1: A summary of the databases used throughout the study.

Database	**URL**
GENECARDS	http://www.genecards.org/
ARRAYEXPRESS	http://www.ebi.ac.uk/arrayexpress/
NCBI	http://www.ncbi.nlm.nih.gov
GEO	http://www.ncbi.nlm.nih.gov/geo
GEO2R	http://www.ncbi.nlm.nih.gov/geo/geo2r
dbEST	http://www.ncbi.nlm.nih.gov/biosample
BLAST	http://blast.ncbi.nlm.nih.gov/Blast.cgi
DAVID	http://david.abcc.ncifcrf.gov
ENSEMBL	http://www.ensembl.org/
STRING	http://string-db.org/
INTERPROSCAN	http://www.ebi.ac.uk/Tools/pfa/iprscan/
KEGG MAPPER	http://www.genome.jp/kegg/mapper.html

2.1.1 Identification of genes predicted to be involved in T cell suppression

Genes predicted to be involved in T cell suppression were identified through the use of the gene databases from Genecards [69] and NCBI [70]. Querying the databases with the search phrase "T cell suppression" resulted in a combined list of 2309 genes.

2.1.2 Retrieval and generation of alternative splicing data through the use of microarrays

In silico as well as *in vitro* experiments were carried out to investigate the PI3K targets and the role of alternative splicing in T cell suppression.

2.1.2.1 Generation of alternative splicing data through the use of a GeneChip Exon array

A GeneChip Exon array analysis was performed on RNAs isolated from human T cells pre-exposed

to wortmannin, or not, prior to a 24 hour phorbolester/ionomycin exposure. Wortmannin is a fungal metabolite known to inhibit PI3K by binding to the p110 catalytic subunit thus preventing phosphorylation of the kinase [83]. The wortmannin was used in this study to mimic the MV infection. The array analysis was performed through a collaboration with our German partners from the University of Wuerzburg. The human T cells were enriched on nylon wool columns and samples were divided into 2 with half the sample stimulated with 40ng/ml PMA and 0.5µM ionomycin for 24hrs whilst the other half was exposed to 50µM of LY200492 (New England Biolabs, Frankfurt, Germany) for 2hrs and then stimulated. RNA from the two samples was extracted using RNeasy Mini Kit (QIAGEN). Quality control was maintained by gel analysis of the RNAs. cDNA and cRNA synthesis of 100ng of total RNA of both samples was processed according to Affymetrix and then hybridized onto the GeneChip Human Exon 1.0 ST array. The resulting data has been deposited in GEO.

2.1.2.2 Extraction of publicly available data

The Gene Expression Omnibus (GEO) [71] and ArrayExpress [72] databases were queried to identify microarray based experiments that were done under conditions relating to "T cell suppression". A total of 9 experiments were found, 8 in GEO and 1 in ArrayExpress. The list of relevant experiments is provided in Table 2.2.

Table 2.2: Microarray experiments on T cell suppression from GEO and ArrayExpress.

Experiment	**Description**	**PMID***
GSE6263	Analysis of HCT116 colon cancer cells in which tumor suppressor gene PTEN had been deleted by gene targeting	17060456
GSE980	Analysis of dendritic cells for up to 24 hrs after infection with measles virus	16492729
GSE6260	Analysis of CD34+ erythroid progenitors stimulated with erythropoietin (Epo) with or without LY294002, a PI3K inhibitor	16965383
GSE9601	Analysis of monocytes treated with an NF-kappaB or PI3K inhibitor and then infected with the human cytomegalovirus (HCMV). CMV tends to induce immunosuppression followed by lasting immunity	18003728
GSE17493	Molecular and functional characterization of alloantigen-specific anergic T cell suitable for cell therapy IL-10 anergized	20713457
GSE2729	Expression profiling of peripheral blood mononuclear cells (PBMCs) from children with acute rotavirus diarrhea / rotavirus activates B but impairs T lymphocytes	17267507
GSE5220	Analysis of CD14+ monocytes from 8 HIV patients at the aviremic state during highly active antiretroviral therapy (HAART) and at the viremic state after the cessation of HAART. Results provide insight into the impact of HIV infection and high-level HIV viremia on the function of monocytes	17005663

Experiment	Description	PMID*
E-MTAB-62	Human gene expression atlas of 5372 samples representing 369 different cell and tissue types, disease states and cell lines.	20379172

***PMID - Pubmed ID**

2.2 Identification of potential isoforms through the use of microarray data

2.2.1 Identification of isoforms from analysis of the GeneChip exon array data

Post processing of the exon array, transcripts that were found to be expressed in PI3K inhibited cells were detected using an in-house developed algorithm [64]. The gene lists were generated in collaboration with our colleagues at the University of Wuerzburg. For the workflow diagram outlining the detection of the transcripts, please see Figure 2.2.

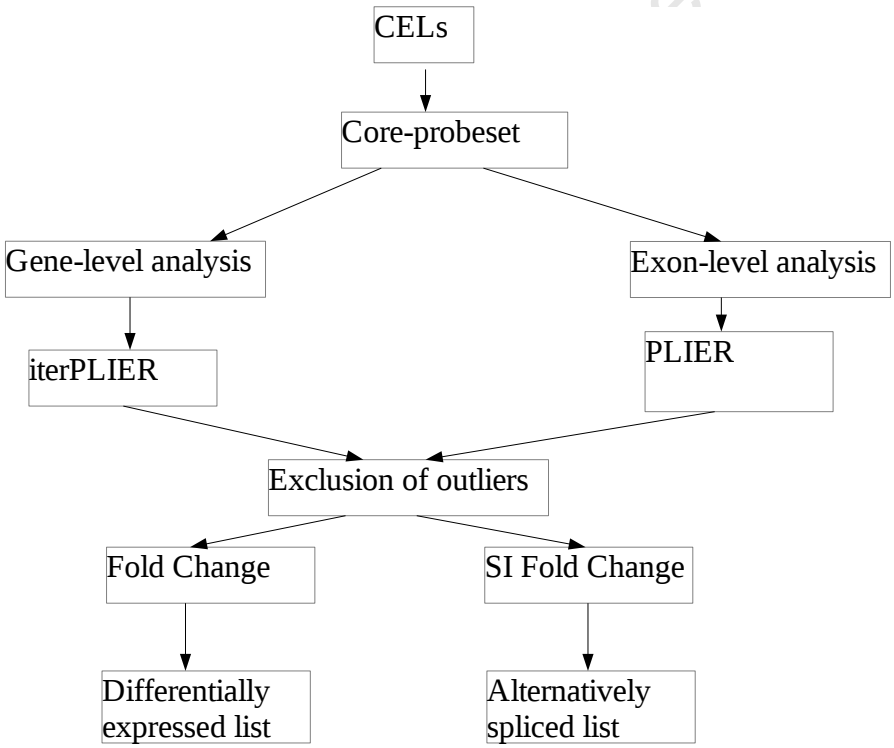


Figure 2.2: Workflow diagram for the gene- and exon-level analysis.

The Expression Console software (Affymetrix) was used for signal estimation, filtering was done using the statistical language R and data normalised by quantile-quantile normalisation [73]. Probe sets annotated as 'core' were used for the gene and exon level analysis. Signal estimation algorithms

PLIER and iterPLIER were used to calculate the expression signal values for the exons and genes respectively [53]. The alternatively spliced list was obtained by applying the following thresholds: a) A sliding window approach of size 3 was applied to genes with more than 6 exons, b) followed by second filtering using a range-cutoff of 0.75 and, c) Splicing Index (SI) of >0.5.

For differentially expressed genes and exons, a modified t-test was calculated and the following thresholds had to be met: a) Genes had a fold change <1.5 and p-value < 0.01, b) exons had to have a fold change above 1.35 and c) 90% of all exons had to be differentially expressed.

2.2.2 Identification of isoforms from analysis of public array data

GEO2R [57] was used to identify differentially expressed genes from the GEO experiments identified in section 2.1.2.2. The Graphical User Interface (GUI) or the tool uses the GEO to query as well as the Limma R packages from the Bioconductor project and doesn't require the user to have command line expertise [58]. The output of the results is a table of the top 250 genes ranked by P-value (the smaller, the more significant) as well as t, B, F statistics and the logFC. The Benjamini & Hochberg false discovery rate method was selected to calculate the P-values to adjust between the discovery of statistically significant genes and false positives [57]. The t statistic is used to test if two samples are different from each other, the B statistic is the log-odds that a gene is differentially expressed and a value of 0 means that a gene has a 50-50 chance that it is differentially expressed [57]. A fold change (logFC) is a ratio of an experimental sample over the control sample and fold change above 2 is deemed significant.

The complete results set was saved into an excel spreadsheet by clicking on "save all results". The output table is populated with identifiers such as Accession (probe sets), Title, Source name and the fields can be amended by choosing other identifiers using the Columns box [57].

In order to identify differentially expressed genes, GEO2R default parameters were used. For each experiment the probe sets, depending on the level of expression, were either assigned as occurring under 'Normal' or 'T cell suppression' conditions. For our research purposes, the level of identification was not sufficient as we needed to identify the isoforms expressed. As the publicly identified microarray data are not exon arrays, Ensembl was utilised in identifying the microarray probes.

The mapping of the probes via Ensembl is a 2 step process. Firstly the individual probes are mapped to the genome as well as the cDNA sequences and also the alignments capture probes which cover the length of the introns [59]. A 1bp mismatch between the probe and the genome sequence

assembly is allowed. The second step in the mapping process involves associating the alignments identified in step 1, with the Ensembl transcript predictions [59]. For arrays with probe sets, Ensembl is set so that 50% of the probes within the set match a transcript. For arrays that do not have probe sets, individual probes are matched to a transcript. A probe or probeset is said to match a transcript if it overlaps with an exon or the untranslated region (UTR) with at least a 1bp mismatch [59]. The individual probe alignments can be accessed via the Ensembl web browser (http://www.ensembl.org) in the 'Region in detail' view whilst probes that match a transcript can be found in the 'Oligo probes' view in the transcript page [59]. Other ways of accessing the data involve programmatic access via the ensembl-functgenomics API and via Biomart. We thus queried Ensembl using Biomart [59] to extract these details. The Ensembl Genes 67 database was queried, with the *Homo sapiens* genes (GRCh37.p7) as the dataset of choice. To restrict/filter the query, the accessions (probe sets) from the GEO2R results of the experiments (with the exception of GEO6260) were used with the appropriate identifier. As the aim of using Biomart was to identify what isoform the probe sets represented, under the attribute node the features selected were Associated Transcript Name as well as the default Ensembl Gene and Transcript IDs. After previewing the results, the output was saved.

The identifiers used for the GEO6260's accessions were not available in Biomart. The NCBI BLAST program with the *Homo sapiens* database [59] was used to align the accessions against the RefSeq DNA database. The results were then used to query Biomart as done for the other microarray experiments.

Post isoform identification, genes that have different isoforms under the normal compared to the T cell suppression conditions were kept for further analysis. To understand the nature of the isoform changes on a gene level, Ensembl was again queried for cDNA/EST evidence to support the transcripts, specifically the non-coding transcripts. The transcripts identified as retained introns, processed transcript and nonsense-mediated decay are all non-coding and the EST/cDNA evidence was used to identify the sources and/or conditions under which these transcripts were found. The supporting evidence shows all mRNA and protein entries in public databases (UniProt/SwissProt, UniProt/TrEMBL and RefSeq) that were used to make an Ensembl transcript prediction [59].

For genes that had different protein-coding transcripts expressed under normal and T cell suppression conditions, the functionality of the proteins under the two conditions was examined using InterProScan [65]. InterProScan is a powerful tool that is used to classify the protein sequences at different levels such as superfamily, family and subfamily. It compares the sequences against the InterPro member databases which include PROSITE, PRINTS, Pfam, SMART etc.,

resulting in the characterisation of the protein's domains and/or functional sites [65]. The different InterPro member databases have different strengths, for example, PROSITE is used to find short motifs whilst PRINTS is good for sub-family membership and by combining multiple databases, InterProScan ensures that reliable results are obtained for a given protein sequence [65]. The amino acid sequences of the isoforms that are protein-coding were retrieved in FASTA format from Ensembl, and run through InterProScan [65].

2.3 Variation analysis

The Ensembl Variation 67 database was queried, with *Homo sapiens* Variation (dbSNP 135;ENSEMBL) as the dataset of choice. SNPs were selected based on the following criteria: 1. should have phenotype data, 2. the SNP should be located in the splice site, i.e. can influence alternative splicing and thus the function of the gene or 3. the SNP be located in a splicing regulatory element (SRE) for example an ESE. Genes that satisfied the criteria were analysed further.

2.4 Functional analysis

2.4.1 Enrichment and pathway analysis

Functional analysis was performed on gene lists using the Database for Annotation, Visualisation and Integrated Discovery (DAVID) web-based tool [60]. The default parameter of count 2 was used together with a lowered Ease score (p-value) of 0.01 from 0.1 (the lower the Ease score the more enriched the terms are). KEGG Mapper [74] was also used as a database of choice and used to map the genes onto pathways. This tool takes a gene list and searches for KEGG pathways the genes are involved in, providing the KEGG maps with the relevant genes in highlighted the colour chosen by the user. If no colour is defined, the default is pink.

2.4.2 Interaction analysis with STRING (Search Tool for the Retrieval of Interacting Genes/Proteins)

STRING [62], a database of known and predicted protein interactions, was also used to find, if any, interactions between PIK3CA and the other genes resulting from the analysis. PIK3CA is the catalytic subunit of the phosphatidyl-inositol-3-kinase (PI3K). The PI3K/AKT signaling pathway is

targeted by the measles virus (MV) in T cells. The pathway is important in a number of cellular processes such as cell growth, proliferation and survival and interference of the signaling by MV results in the arrest of the cell cycle in T lymphocytes leading to T cell suppression. The direct and indirect interactions obtained from STRING, are derived from seven sources: neighbourhood, gene fusion, cooccurence, coexpression, experiments, databases and textmining [63].

CHAPTER 3

RESULTS

The aim of the project was to identify markers of T cell suppression. This was achieved by using publicly available as well as internally generated microarray data in order to predict the alternatively spliced isoforms that could potentially be used as genetic markers for T cell suppression.

3.1 Identification of genes predicted to be involved in T cell suppression

In order to identify genes involved in T cell suppression, we searched Genecards [69] and NCBI [70]. This resulted in a list of 2309 genes, which is provided in the Appendix A.

3.2 Identification of potential isoforms through the use of microarray data

3.2.1 Identification of alternatively spliced genes from the GeneChip exon array

A GeneChip exon array experiment was performed by our collaborators from the University of Wuerzburg. Preprocessing and analysis of the exon array was done as described in the methods section to detect transcripts specifically expressed in PI3K inhibited cells. These were assigned to categories defining differentially regulated (DR) (654 candidates) and alternatively spliced (AS) species (1985 candidates) [63]. To be considered alternatively spliced, candidate genes were obtained by applying the following thresholds: a) A sliding window approach of size 3 was applied to the genes with more than 6 exons, b) followed by second filtering using a range-cutoff of 0.75 and, c) Splicing Index (SI) of >0.5. For the differentially expressed genes and exons, a modified t-test was calculated and genes had to have a fold change <1.5 and p-value < 0.01 whilst exons had to have a fold change above 1.35 and that 90% of all exons had to be differentially expressed. Validation of a selection of transcripts, based on signal intensities and splice indices, was carried out through the use of RT-PCR. Functional annotation and analysis of representation of these genes in molecular networks and pathways was carried out.

Applying filters described in the methods section, 1985 genes were assigned as being alternatively

spliced whilst 654 genes were differentially regulated. Nine and seven candidate genes from the differentially regulated and alternatively spliced list respectively, were chosen for validation using RT-PCR. It was found that gene assignment to both categories seemed valid. Please see work done by Riedel et. al for more detail [63].

DAVID [60] was used for functional analysis. Only pathways that showed a p-value < 0.01 and at least 5 genes were considered to be significantly enriched. The analysis revealed that alternatively spliced genes were involved in extra cellular matrix (ECM)-receptor interaction and focal adhesion, purine metabolism, and natural killer cell mediated cytotoxicity. The differentially regulated list was associated with cytokine-receptor interaction, the Jak-STAT and p53 pathways, as well as DNA replication. Cell cycle regulation was enriched in both the alternatively spliced and the differentially regulated lists.

3.2.2 Intersection of publicly identified genes and GeneChip Exon array genes

We then determined whether there was an overlap between the genes from the exon array and those retrieved from gene databases using T cell suppression and related search terms. An overlap of the differentially regulated (654) as well as the alternatively spliced (1985) genes was found with the genes identified in section 3.1. Figure 3.1 illustrates the intersection between the 3 gene lists.

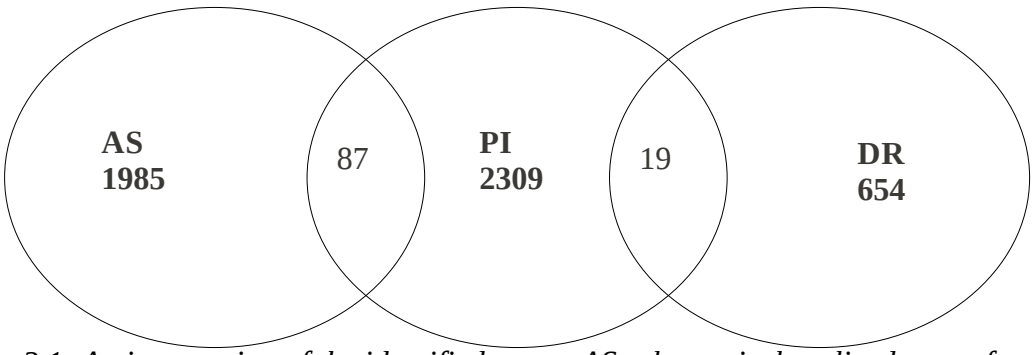


Figure 3.1: An intersection of the identified genes. AS: alternatively spliced genes from the GeneChip exon array; PI: publicly identified genes and DR: differentially regulated genes from the GeneChip exon array. Genes that intersected were kept for further analysis.

The genes that intersected were kept for further analysis and are listed in the next page in Table 3.1.

Table 3.1: An intersection of publicly identified genes and genes from the GeneChip exon array.

Gene	Description	Assignment*
AGAP2	ArfGAP with GTPase domain, ankyrin repeat and PH domain 2	AS
ALCAM	activated leukocyte cell adhesion molecule	DR
AMPD3	adenosine monophosphate deaminase 3	AS
AR	androgen receptor	AS
ATM	ataxia telangiectasia mutated	DR
BACE1	beta-site APP-cleaving enzyme 1	AS
BIN1	bridging integrator 1	AS
BMP4	bone morphogenetic protein 4	AS
BRCA1	breast cancer 1, early onset	AS
CALD1	caldesmon 1	AS
CCDC50	coiled-coil domain containing 50	AS
CD244	CD244 molecule, natural killer cell receptor 2B4	AS
CD44	CD44 molecule (Indian blood group)	AS
CD79B	CD79b molecule, immunoglobulin-associated beta	AS
CDC25A	cell division cycle 25 homolog A (S. pombe)	AS
CDH11	cadherin 11, type 2, OB-cadherin (osteoblast)	AS
CDKN1A	cyclin-dependent kinase inhibitor 1A (p21, Cip1)	DR
CFB	complement factor B	AS
CHAF1A	chromatin assembly factor 1, subunit A (p150)	AS
CHN1	chimerin (chimaerin) 1	AS
CTLA4	cytotoxic T-lymphocyte-associated protein 4	DR
DDX58	DEAD (Asp-Glu-Ala-Asp) box polypeptide 58	AS
DFFB	DNA fragmentation factor, 40kDa, beta polypeptide (caspase-activated DNase)	AS
DUT	deoxyuridine triphosphatase	AS
ELAC2	elaC homolog 2 (E. coli)	DR
EPB49	erythrocyte membrane protein band 4.9 (dematin)	AS
ERCC1	excision repair cross-complementing rodent repair deficiency, complementation group 1 (includes overlapping antisense sequence)	AS
ETV7	ets variant 7	AS
EXO1	exonuclease 1	AS
FANCD2	Fanconi anemia, complementation group D2	AS
FCGR2B	Fc fragment of IgG, low affinity IIb, receptor (CD32)	AS
FCGR2C	Fc fragment of IgG, low affinity IIc, receptor for (CD32) (gene/pseudogene)	AS
FLT3LG	fms-related tyrosine kinase 3 ligand	AS
GAD1	glutamate decarboxylase 1 (brain, 67kDa)	AS
GCK	glucokinase (hexokinase 4)	AS
GRIA3	glutamate receptor, ionotropic, AMPA 3	AS
HDAC5	histone deacetylase 5	AS

HDAC9	histone deacetylase 9	AS
HPR	haptoglobin-related protein	AS
ICOS	inducible T cell co-stimulator	DR
IKZF1	IKAROS family zinc finger 1 (Ikaros)	DR
IL12RB1	interleukin 12 receptor, beta 1	AS
IMPDH1	IMP (inosine 5'-monophosphate) dehydrogenase 1	AS
INPPL1	inositol polyphosphate phosphatase-like 1	AS
IRAK1	interleukin-1 receptor-associated kinase 1	AS
IRF7	interferon regulatory factor 7	DR
ITGAM	integrin, alpha M (complement component 3 receptor 3 subunit)	AS
ITGAV	integrin, alpha V (vitronectin receptor, alpha polypeptide, antigen CD51)	DR
KCNQ1	potassium voltage-gated channel, KQT-like subfamily, member 1	AS
KCNQ5	potassium voltage-gated channel, KQT-like subfamily, member 5	DR
KHK	ketohexokinase (fructokinase)	AS
KIAA0913	KIAA0913	DR
KIT	v-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog	AS
KLRC3	killer cell lectin-like receptor subfamily C, member 3	DR
LAMA3	laminin, alpha 3	AS
LCK	lymphocyte-specific protein tyrosine kinase	AS
MAP3K12	mitogen-activated protein kinase kinase kinase 12	AS
MCM7	minichromosome maintenance complex component 7	DR
MEN1	multiple endocrine neoplasia I	AS
MEST	mesoderm specific transcript homolog (mouse)	AS
MKI67	antigen identified by monoclonal antibody Ki-67	AS
MUC1	mucin 1, cell surface associated	AS
MXI1	MAX interactor 1	AS
MYB	v-myb myeloblastosis viral oncogene homolog (avian)	AS
MYH10	myosin, heavy chain 10, non-muscle	AS
NAB2	NGFI-A binding protein 2 (EGR1 binding protein 2)	AS
NRP1	neuropilin 1	AS
OGDH	oxoglutarate (alpha-ketoglutarate) dehydrogenase (lipoamide)	DR
OPN4	opsin 4	DR
P4HTM	prolyl 4-hydroxylase, transmembrane (endoplasmic reticulum)	AS
PCGF6	polycomb group ring finger 6	AS
PDE2A	phosphodiesterase 2A, cGMP-stimulated	AS
PLAT	plasminogen activator, tissue	AS
PNPLA8	patatin-like phospholipase domain containing 8	DR
PPARGC1B	peroxisome proliferator-activated receptor gamma, coactivator 1 beta	AS
PRDM10	PR domain containing 10	AS
PRMT5	protein arginine methyltransferase 5	DR

PTK7	PTK7 protein tyrosine kinase 7	AS
PVRL2	poliovirus receptor-related 2 (herpesvirus entry mediator B)	AS
RAD1	RAD1 homolog (S. pombe)	AS
RARA	retinoic acid receptor, alpha	AS
RARG	retinoic acid receptor, gamma	AS
RASGRF1	Ras protein-specific guanine nucleotide-releasing factor 1	AS
RASSF5	Ras association (RalGDS/AF-6) domain family member 5	AS
RBBP6	retinoblastoma binding protein 6	AS
RBL1	retinoblastoma-like 1 (p107)	AS
RUNX3	runt-related transcription factor 3	AS
SEMA3B	sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3B	AS
SGOL1	shugoshin-like 1 (S. pombe)	DR
SIRT2	sirtuin 2	AS
SLC25A14	solute carrier family 25 (mitochondrial carrier, brain), member 14	AS
SLCO4A1	solute carrier organic anion transporter family, member 4A1	AS
SMAD6	SMAD family member 6	AS
SMG1	smg-1 homolog, phosphatidylinositol 3-kinase-related kinase (C. elegans)	AS
SPHK1	sphingosine kinase 1	AS
SPP1	secreted phosphoprotein 1	AS
STK36	serine/threonine kinase 36	AS
SYK	spleen tyrosine kinase	AS
TLR4	toll-like receptor 4	AS
TNC	tenascin C	AS
TRPV4	transient receptor potential cation channel, subfamily V, member 4	AS
UBD	ubiquitin D	AS
USF2	upstream transcription factor 2, c-fos interacting	AS
VCL	vinculin	DR
VWCE	von Willebrand factor C and EGF domains	AS
WWC1	WW and C2 domain containing 1	AS

***Assignment: AS - Alternatively Spliced; DR - Differentially Regulated**

3.2.3 Identification of isoforms from the public microarray data

Publicly available microarray data were used to identify potential isoforms that might be formed during normal and T cell suppression conditions as well as to provide further support for the genes that overlapped with those identified through the use of the GeneChip Exon array. GEO2R [57] was used to identify the alternatively spliced isoforms from the different microarray data. As the publicly identified microarray data are not exon arrays, Ensembl was used to identify the microarray probes. Ensembl starts by mapping the individual probes to the genome as well as the cDNA sequences with room for a 1bp mismatch between the probe and the genome sequence assembly. The second step involves associating the alignments identified in step 1, with the Ensembl transcript predictions [59]. A probe or probeset is said to match a transcript if it overlaps with an exon or the untranslated region (UTR) with at least a 1bp mismatch. We thus queried Ensembl using Biomart [59] to extract the required details.

Two sets of gene lists were created as a result of the Biomart search, namely a list of genes that had different isoforms expressed under the different conditions whilst another list was of genes that had different isoforms all expressed under a single condition. Genes that exhibited different isoforms under different conditions were kept for further analysis. The following genes that overlapped with the GeneChip exon array, had different isoforms formed under normal versus T cell suppression conditions: *ATM*, *CALD1*, *LCK*, *VCL*, *MXI1*, *NRP1* and *PRMT5*. See Table 3.2 for summary of the identified isoforms. *ATM*, *VCL* and *PRMT5* were found to be differentially regulated on the whole transcript level using the GeneChip exon array, while *CALD1*, *LCK*, *MXI1* and *NRP1* were shown to be alternatively spliced.

Table 3.2: Genes that have different isoforms under Normal vs. T cell suppression conditions.

Ensembl Gene ID	Ensembl Transcript ID	Associated Transcript Name	Condition	RefSeq / EST evidence	Experiment
ENSG00000149311	ENST00000527805	ATM-001	Normal	YES	GSE980
ENSG00000149311	ENST00000530958	ATM-002	T-cell suppression	YES	GSE980
ENSG00000149311	ENST00000526567	ATM-004	T-cell suppression	YES	GSE980
ENSG00000149311	ENST00000532931	ATM-016	Normal	YES	GSE980
ENSG00000122786	ENST00000430085	CALD1-004	T-cell suppression	YES	GSE9601, GSE6263, GSE5220
ENSG00000122786	ENST00000482470	CALD1-008	Normal	YES	GSE6263, GSE5220
ENSG00000122786	ENST00000543443	CALD1-204	T-cell suppression	YES	GSE5220, GSE6263
ENSG00000182866	ENST00000495610	LCK-006	Normal	YES	GSE6263
ENSG00000182866	ENST00000373557	LCK-010	T-cell suppression	YES	GSE980
ENSG00000182866	ENST00000477031	LCK-011	T-cell suppression	YES	GSE980
ENSG00000182866	ENST00000398345	LCK-202	Normal	YES	GSE9601
ENSG00000119950	ENST00000369613	MXI1-001	T-cell suppression	YES	GSE9601
ENSG00000119950	ENST00000369614	MXI1-002	T-cell suppression	YES	GSE9601
ENSG00000119950	ENST00000460667	MXI1-003	T-cell suppression	YES	GSE9601
ENSG00000119950	ENST00000485566	MXI1-007	Normal	YES	GSE980
ENSG00000119950	ENST00000393134	MXI1-009	T-cell suppression	YES	GSE9601
ENSG00000119950	ENST00000369612	MXI1-011	T-cell suppression	YES	GSE9601
ENSG00000119950	ENST00000484030	MXI1-012	T-cell suppression	YES	GSE9601
ENSG00000119950	ENST00000369619	MXI1-201	T-cell suppression	YES	GSE9601
ENSG00000035403	ENST00000372755	VCL-001	Normal	YES	E-MTAB62, GSE980
ENSG00000035403	ENST00000415462	VCL-202	T-cell suppression	YES	E-MTAB62
ENSG00000035403	ENST00000537043	VCL-204	Normal	YES	E-MTAB62, GSE980
ENSG00000099250	ENST00000265371	NRP1-001	T-cell suppression	YES	GSE6263, GSE26050
ENSG00000099250	ENST00000374823	NRP1-005	T-cell suppression	YES	GSE6263, GSE26050
ENSG00000099250	ENST00000374814	NRP1-201	Normal	YES	GSE6263, GSE26050
ENSG00000100462	ENST00000557443	PRMT5-006	Normal	YES	GSE26050
ENSG00000100462	ENST00000553550	PRMT5-014	T-cell suppression	YES	GSE17493
ENSG00000100462	ENST00000553502	PRMT5-017	T-cell suppression	YES	GSE17493
ENSG00000100462	ENST00000556043	PRMT5-018	T-cell suppression	YES	GSE17493
ENSG00000100462	ENST00000555530	PRMT5-020	T-cell suppression	YES	GSE17493
ENSG00000100462	ENST00000553787	PRMT5-025	T-cell suppression	YES	GSE17493

3.3 Gene expression levels of isoforms identified from microarray data

3.3.1 GeneChip exon array: probe set intensity plots

Log ratio of the probe set intensities, which show the abundance of a transcript, for the genes identified in section 3.2.3 were plotted. Figure 3.2 illustrates the gene views of the probe set intensity plots for the genes *LCK* and *PRMT5* as well as the different transcript profiles, of the genes, below the graphs. The plots are for the stimulated samples (red line) and both stimulated and inhibited samples (blue line). *LCK* was found to be alternatively spliced as there were regions in the gene where the probe set expressions were found to be different between the stimulated samples and the inhibited/stimulated samples whilst other regions the expression was the same between the different samples as the SI<0.5. *PRMT5* on the other hand was differentially regulated as the different samples showed different expression levels throughout the gene.

Please refer to Appendix B to view plots for *ATM*, *VCL*, *CALD1*, *MXI1* and *NRP1*.

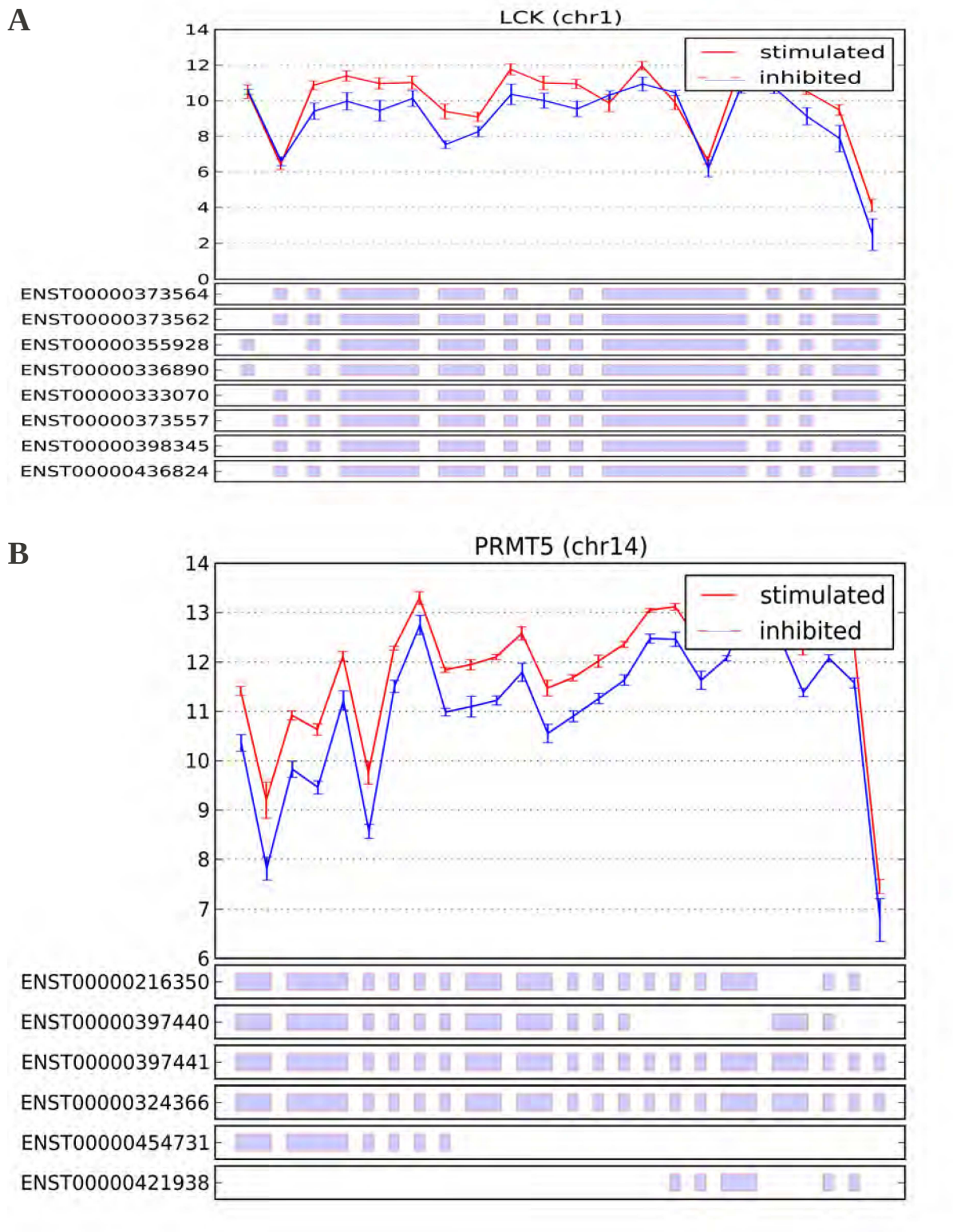


Figure 3.2: Probe set intensity plots for the alternatively spliced gene LCK and differentially regulated gene PRMT5. Accession numbers of the alternatively spliced forms are provided below each panel. ***A****: LCK is alternatively spliced due to regions in the gene where the probe set expressions are different between the stimulated and the inhibited/stimulated samples whilst other regions have the same expression between the different samples.* ***B****: PRMT5 is differentially regulated as the different samples showed different expression levels throughout.*

3.3.2 Publicly identified microarray data

3.3.2.1 Different transcripts, different conditions

The log fold changes and p-values (were calculated using the Benjamini & Hochberg false discovery rate method) for the different experiments retrieved from GEO and ArrayExpress were calculated using GEO2R, see section 2.2.2. For each experiment, using GEO2R, samples within an experiment were assigned to two groups, i.e. a control and a test group and the log fold change was calculated between the two experimental conditions [57]. See Table 3.3 for the condition under which the log fold changes were calculated. The p-value was log transformed and the higher the value, the more significant the probe's expression, whilst a fold change of 2 or above is desired. Log transformed p-values for the probes belonging to the same gene, were plotted against the fold change expressions for the seven genes that overlapped with the GeneChip exon array gene lists (Figure 3.3). The graphs were generated using OpenOffice.org Calc [64].

Table 3.3: Conditions under which the log fold changes were calculated

Experiment	Control	Test
GSE6263	Human colon cancer cells with intact PTEN	Human colon cancer cells with deleted PTEN
GSE980	Monocyte-derived dendritic cells	Measles virus-infected dendritic cells
GSE6260	Erythroid progenitors cultured in medium	Erythroid progenitors stimulated with EPO and LY
GSE9601	HCMV-infected monocytes	HCMV-infected monocytes pretreated with LY
GSE17493	CD3-depleted cells	CD3-depleted cells stimulated with IL-10
GSE2729	Peripheral blood mononuclear cells (PBMC)	Rotavirus-infected PBMC
GSE5220	HIV-infected monocytes on HAART	HIV-infected monocytes off HAART
E-MTAB-62	Healthy PBMC	PBMC with: trauma; anergy; HIV

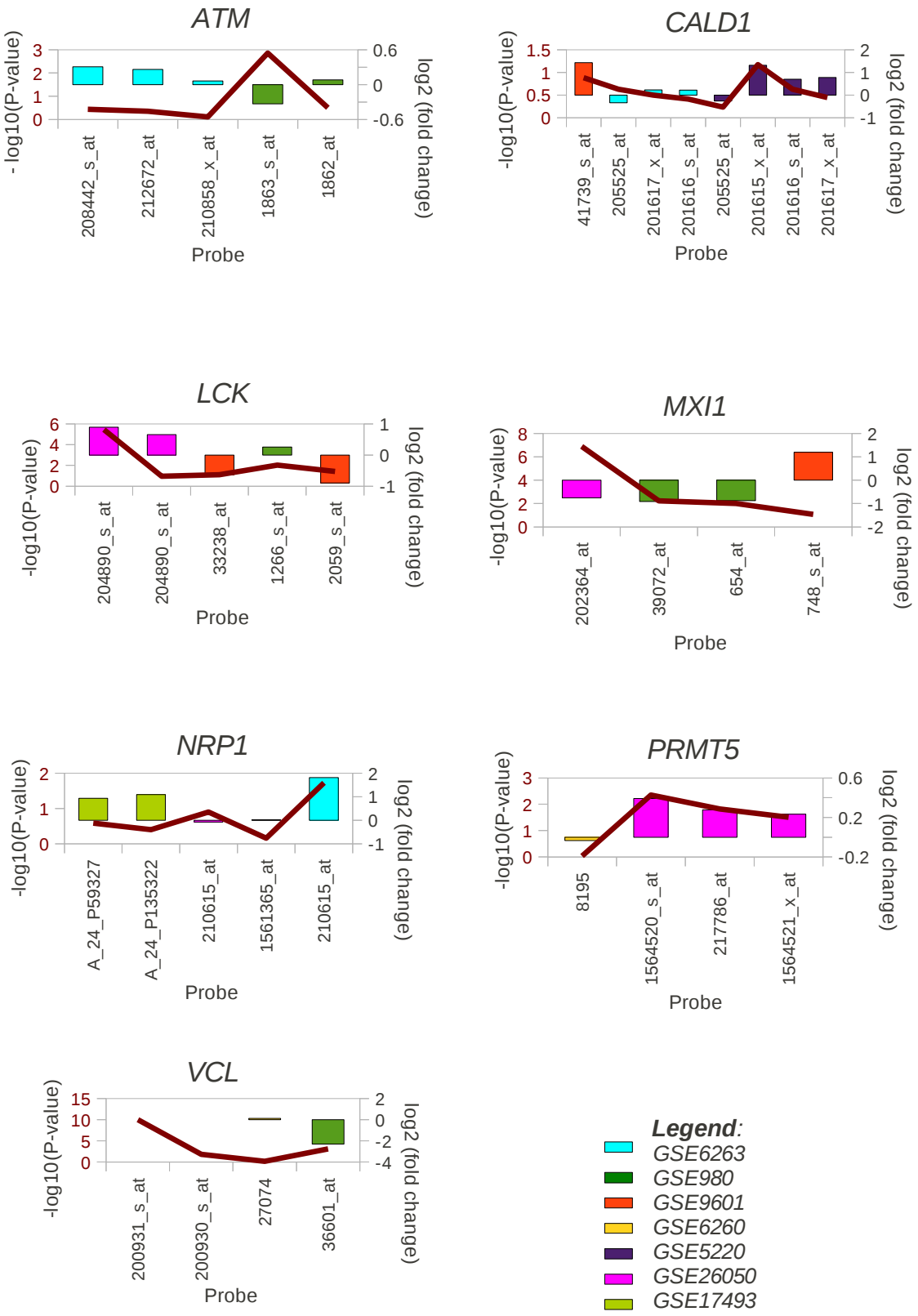


Figure 3.3: Expression levels of the gene's probes from public array data for the seven genes of interest. The P-values for the probes of the different genes are graphed (red line), along with the fold change expression levels (coloured bars, see legend).

ATM

For the gene *ATM*, probe 1863_s_at, representing isoforms *ATM-001* and *ATM-016* was found to be expressed under normal conditions in the experiment GSE980 whilst *ATM-002* and *ATM-004* were expressed under T cell suppression conditions and represented by the probe 1862_at in the same experiment. The probe 1863_s_at was expressed under normal conditions at a significant p-value of 0.0014 whilst the probe 1862_at had an insignificant p-value of 0.313. Probes 208442_s_at (p-value=0.3614), 212672_at (p-value=0.4407) and 210858_x_at (p-value=0.7823) were all expressed under T cell suppression conditions in the experiment GSE6263.

LCK

Transcripts LCK-006, represented by probe 204890_s_at (p value = 0.111 and 0.0000037, respectively) in experiments GSE26050 and GSE6263, LCK-010 and LCK-011, both represented by probe 1266_s_at (p-value =0.00948) in the experiment GSE980, were all expressed under T-cell suppression conditions. *LCK-202* represented by probes 2059_s_at (p-value=0.0379) and 33238_at (p-value= 0.0782) had significant expression under normal conditions in the experiment GSE9601.

CALD1

CALD1-008 represented by probe 205525_at (p-value of 0.2357 and 0.5832) was expressed under normal conditions in the experiments GSE6263 and GSE5220 whilst under T cell suppression conditions *CALD1*-004 represented by probe 201617_x_at (p-value of 0.319 and 0.3575) and probe 201616_s_at (p-value of 0.389) was expressed in both the experiments GSE6263 and GSE5220. *CALD1-204* (probe 201616_s_at with p-value of 0.389) was expressed in the experiment GSE5220 under T cell suppression conditions.

MXI1

Seven *MXI1* transcripts, represented by probe 748_s_at with p-value of 0.08, were expressed under T cell suppression conditions, namely *MXI1-009*, *MXI1-011*, *MXI1-201*, *MXI1-001*, *MXI1-002*, *MXI1-003* and *MXI1-012*. *MXI1-007* represented by probe 39072_at (p-value of 0.006) in experiment GSE980, probe 654_at (p-value of 0.009) in experiment GSE980 and probe 202364_at (p-value of 0.00000013) in experiment GSE26050 was the only isoform found to be expressed under normal conditions.

VCL

VCL-001 and *VCL-204*, represented by the probes 200931_s_at (p-value of 0.1 e-9) and probe 36601_at (p-value of 0.000828) were expressed under normal conditions in the experiments

GSE980 and E-MTAB-62. *VCL-202* represented by probe 200930_s_at (p-value of 0.016) was expressed under T cell suppression conditions in the experiment E-MTAB-62.

PRMT5

Probe 8195 was found expressed in the experiment GSE6260 under normal conditions with a p-value of 0.9242. Probes 1564520_s_at (p-value of 0.0044), 217786_at (p-value of 0.0151) and 1564521_x_at (p-value of 0.0313) were all expressed in the experiment GSE26050 under T cell suppression conditions. Isoform PRMT5-006 is represented by probe 8195 and the probes expressed under T cell suppression conditions represent isoforms PRMT5-014, PRMT5-017, PRMT5-018, PRMT5-020 and PRMT5-025.

NRP1

The probe 1561365_at (p-value of 0.689), representing isoform NRP1-201, was found expressed under normal conditions in the experiment GSE26050. Probes A_24_P59327(p-value=0.2583) and 210615_at (p-value=0.1270 and 0.0189) both code for NRP1-005 and were found expressed under T cell suppression conditions in the experiments GSE17493, GSE26050 and GSE6263. Isoform NRP1-001, represented by probe A_24_P135322 (p-value=0.3981), was found expressed under T cell suppression conditions as well in the experiment GSE17493.

3.3.2.2 Different transcripts, same condition

The gene *CD44*, which is known to produce spliced isoforms, appeared under the list of genes that have isoforms produced under only one condition, either normal or T cell suppression but not both. The list, not shown, also included the gene *INPPL1* which is part of the SHIP family of genes. Genes such as *BRCA1*, *FGF1* and *PSEN1* were also on the list. This is interesting as these genes' alternative splicing has been well studied [49].

3.4 Gene summaries, EST/cDNA evidence and protein functionality

The following section highlights the gene summaries, from Ensembl [59], of the genes identified in section 3.2.3. The graphs of the gene summaries show the distribution of the exons of the different isoforms along the chromosomes. EST/cDNA supporting evidence for the transcripts are also included for the transcripts identified as non-coding, i.e. retained introns, processed transcripts and nonsense-mediated decay. For genes that had different protein-coding transcripts expressed under both normal compared to T cell suppression conditions, the functionality of the proteins under the two conditions was examined using InterProscan [65] by looking at the protein's domains. The different isoforms identified in section 3.2.3, are shown with arrows in the graphs. The green arrow identifies isoforms expressed under normal conditions and the purple arrow is for isoforms expressed under T cell suppression conditions.

3.4.1 *ATM*

3.4.1.1 *ATM* gene summary

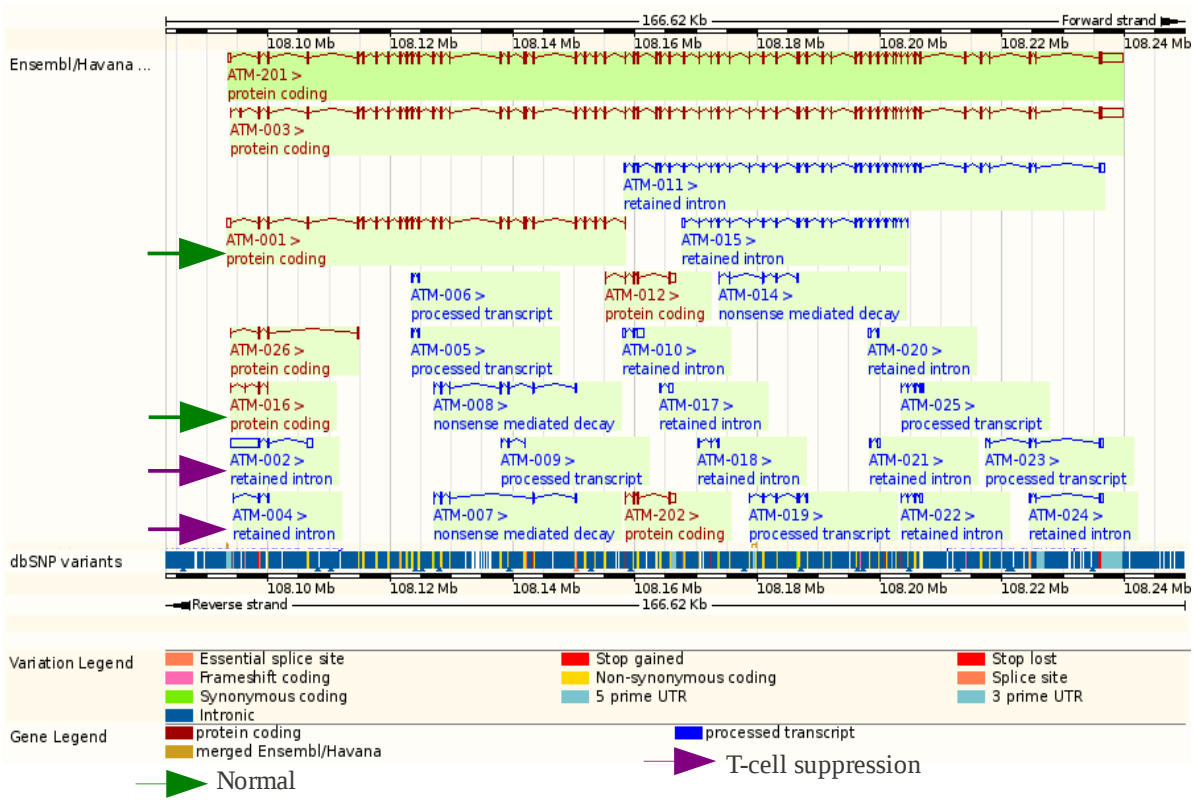
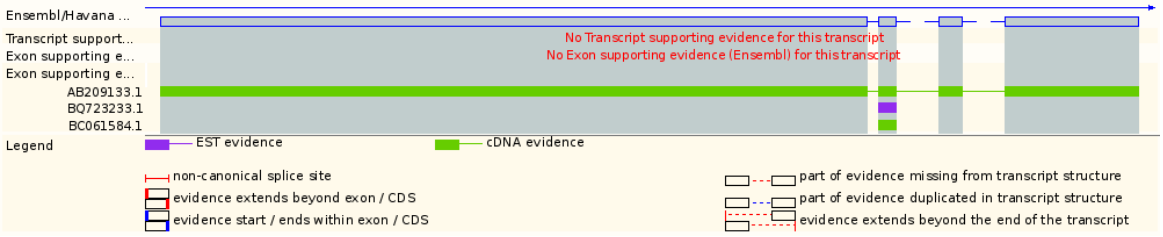


Figure 3.4: ATM gene summary modified from Ensembl. The longer ATM-001 and the short ATM-016 were found expressed under normal conditions. ATM-002 and ATM-004, both retained introns with no protein products, were found expressed under T cell suppression conditions.

Four transcripts were identified from probes that were found to be expressed under the different conditions. *ATM-001* and *ATM-016* were expressed under normal conditions and are protein-coding with *ATM-001* being the longer isoform. The transcripts found under T cell suppression conditions were *ATM-002* and *ATM-004*, both retained introns and are predicted to have no protein product. Supporting evidence for the transcripts *ATM-002* and *ATM-004* is provided in Figure 3.5.

3.4.1.2 ATM supporting evidence

ATM-002



ATM-004

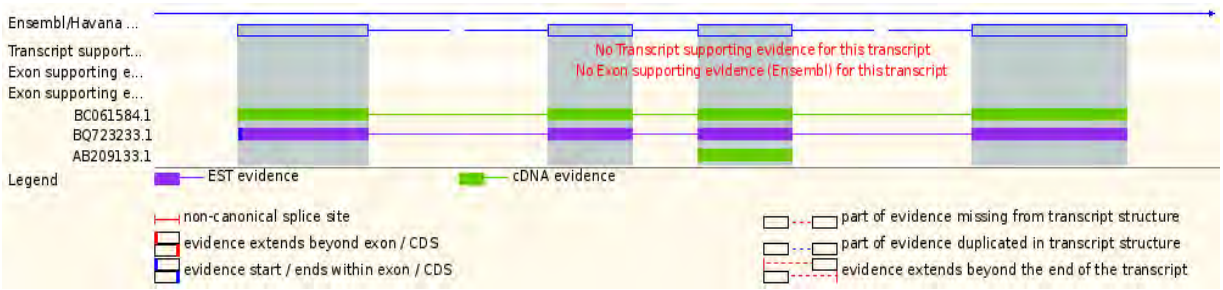


Figure 3.5: Ensembl EST/cDNA supporting evidence for the ATM-002 and ATM-004 transcripts.

Not much supporting evidence exists for these non-coding transcripts. *ATM-004* has both EST and cDNA supporting all exons whilst for the transcript *ATM-002* only cDNA evidence supports all the exons. For *ATM-002*, AB209133.1 was found expressed in brain tissue, BQ723233.1 and BC061584.1 were found expressed in the sympathetic trunk (nerve fibres that run between the base of the skull and the tailbone) of a 16 year old male. *ATM-004* had the same supporting evidence as *ATM-002* but supporting different exons.

3.4.2 *CALD1*

3.4.2.1 *CALD1* gene summary

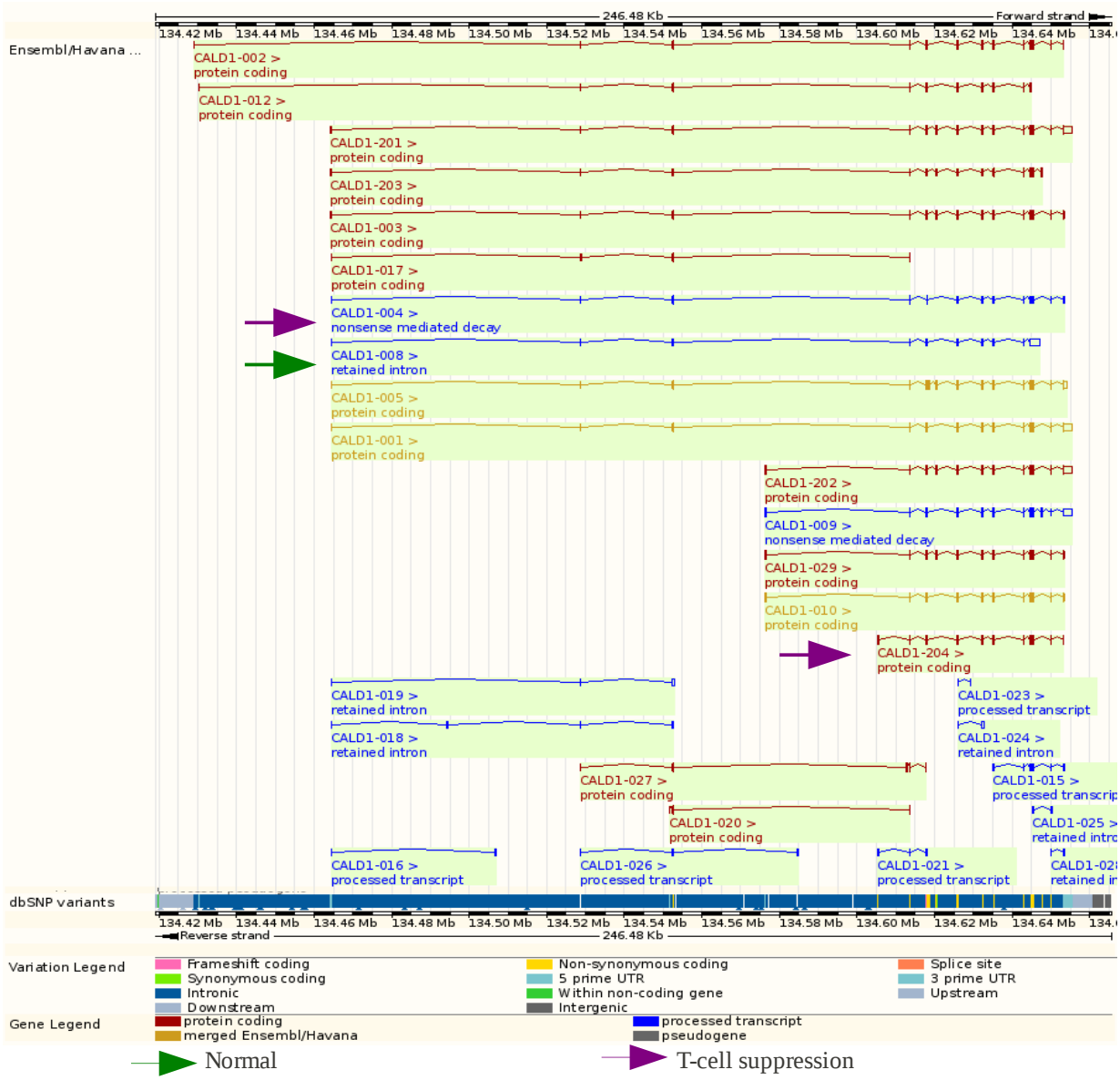


Figure 3.6: CALD1 gene summary modified from Ensembl. The protein-coding CALD1-204 was found expressed under T cell suppression conditions together with the non-coding CALD1-004. CALD1-008, also non-coding, was expressed under normal conditions.

CALD1-204 and *CALD1-004* are expressed under T cell suppression whilst *CALD1-008* was found expressed under normal conditions. *CALD1-004* and *CALD1-008* are both non-coding and *CALD1-204* is protein-coding. The difference between the non-coding transcripts, *CALD1-004* and *CALD1-008*, is not only the number of exons but *CALD1-008* has an untranslated region (UTR).

3.4.2.2 *CALD1* supporting evidence

CALD1-008

CALD1-004

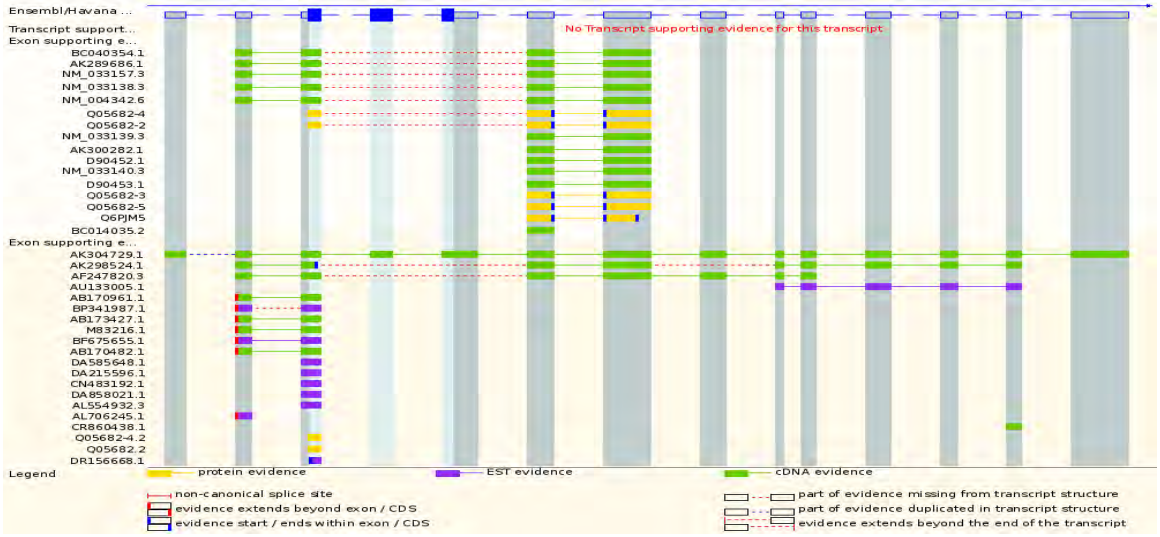


Figure 3.7: Ensembl EST/cDNA supporting evidence for the CALD1 transcripts. CALD1-008 was found expressed in both non-muscle and smooth muscle tissues. CALD1-004 was found in the uterus.

For both *CALD1-008* and *CALD1-004* while there is no transcript evidence, there is plenty of exon supporting evidence that interestingly includes evidence at the protein level given the supposed non-coding ability of the transcripts. According to this evidence, *CALD1-008* is mainly expressed in non-muscle tissues or cells such as amygdala and is also found in smooth muscles. The cDNA AK304729.1 supports all of the *CALD1-004* exons and was found expressed in the library LIBEST_018544 UTERU3 in the uterus tissue.

3.4.3 *MXI1*

3.4.3.1 *MXI1* gene summary

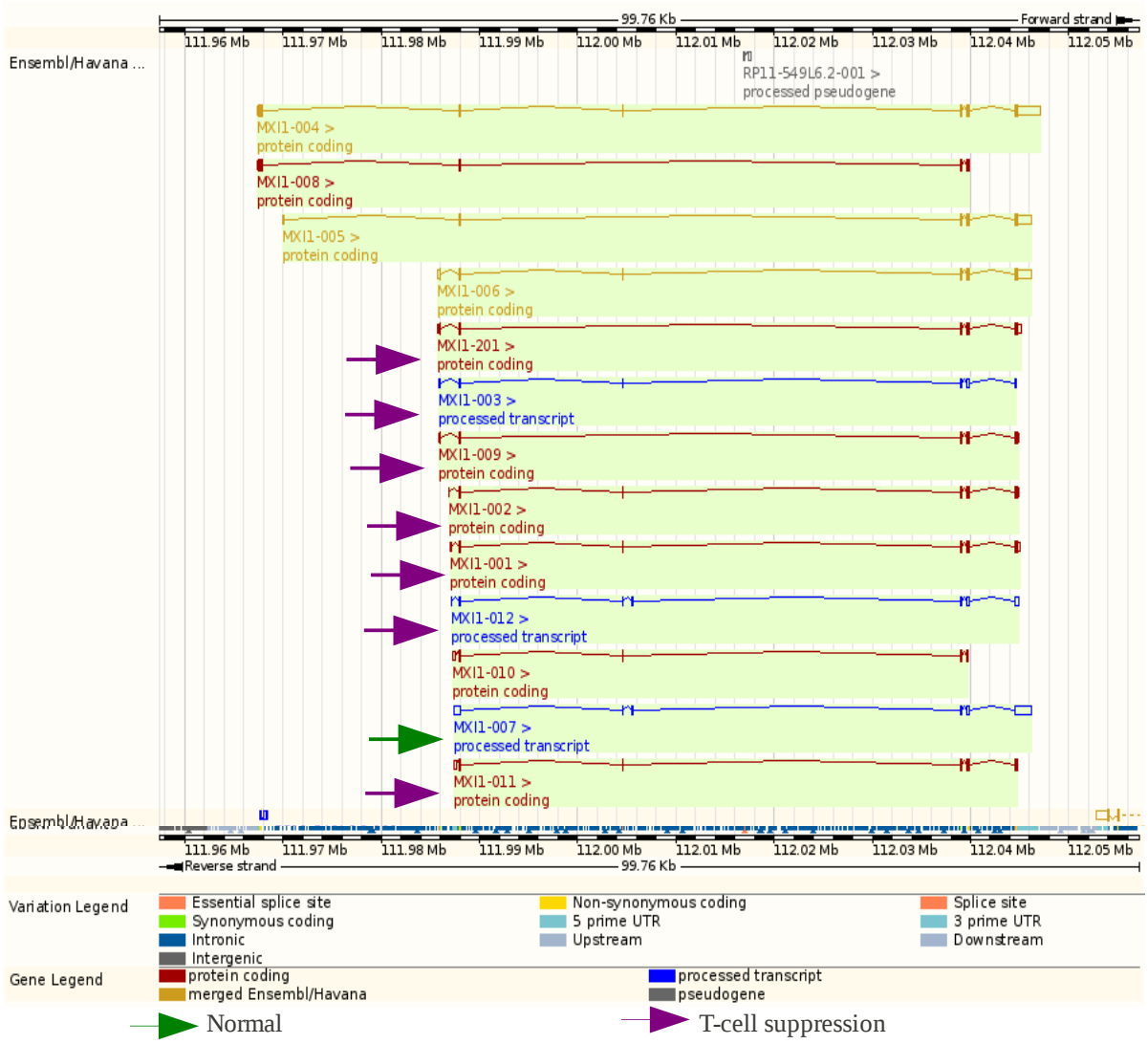
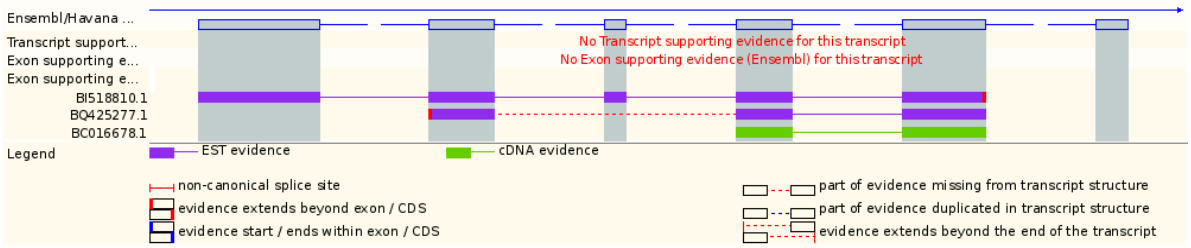


Figure 3.8: MXI1 gene summary modified from Ensembl. MXI1-007, a non-coding transcript without an ORF, was found expressed under normal conditions whilst a mixture of protein-coding and non-coding transcripts was found expressed under T cell suppression conditions.

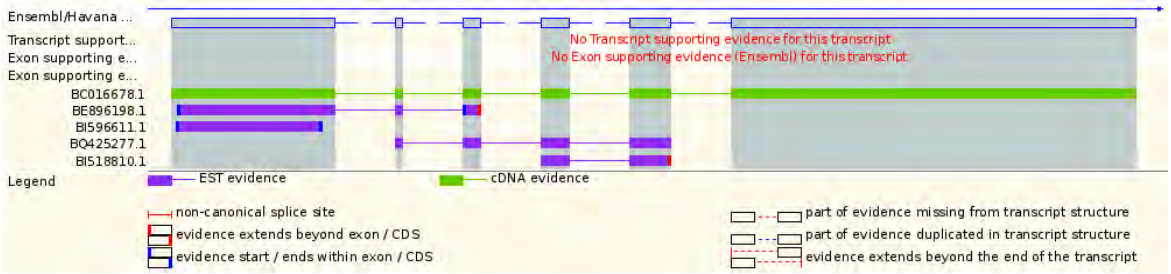
The *MXI1-007* which is a processed transcript (non-coding transcript without an ORF), was found to be expressed under normal conditions. Transcripts expressed under T cell suppression conditions were a mixture of protein-coding and non-coding. Supporting evidence for the non-coding transcripts is shown in section 3.4.3.2.

3.4.3.2 *MXI1* supporting evidence

MXI1-003



MXI1-007



MXI1-012

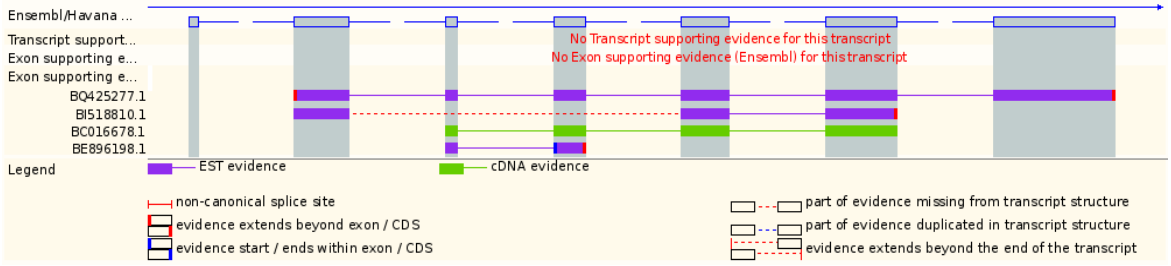


Figure 3.9: Ensembl EST/cDNA supporting evidence for the MXI1 transcripts. The MXI1 transcripts were found expressed on skin tissue.

Just like the *ATM* non-coding transcripts, there seem to be very little evidence for these *MXI1* transcripts. The EST BQ425277.1 evidence was found in skin tissue and the cDNA BC016678.1 was found expressed in skin tissue with melanotic melanoma generated from the library NIH_MGC_72.

3.4.4 *LCK*

3.4.4.1 *LCK* gene summary

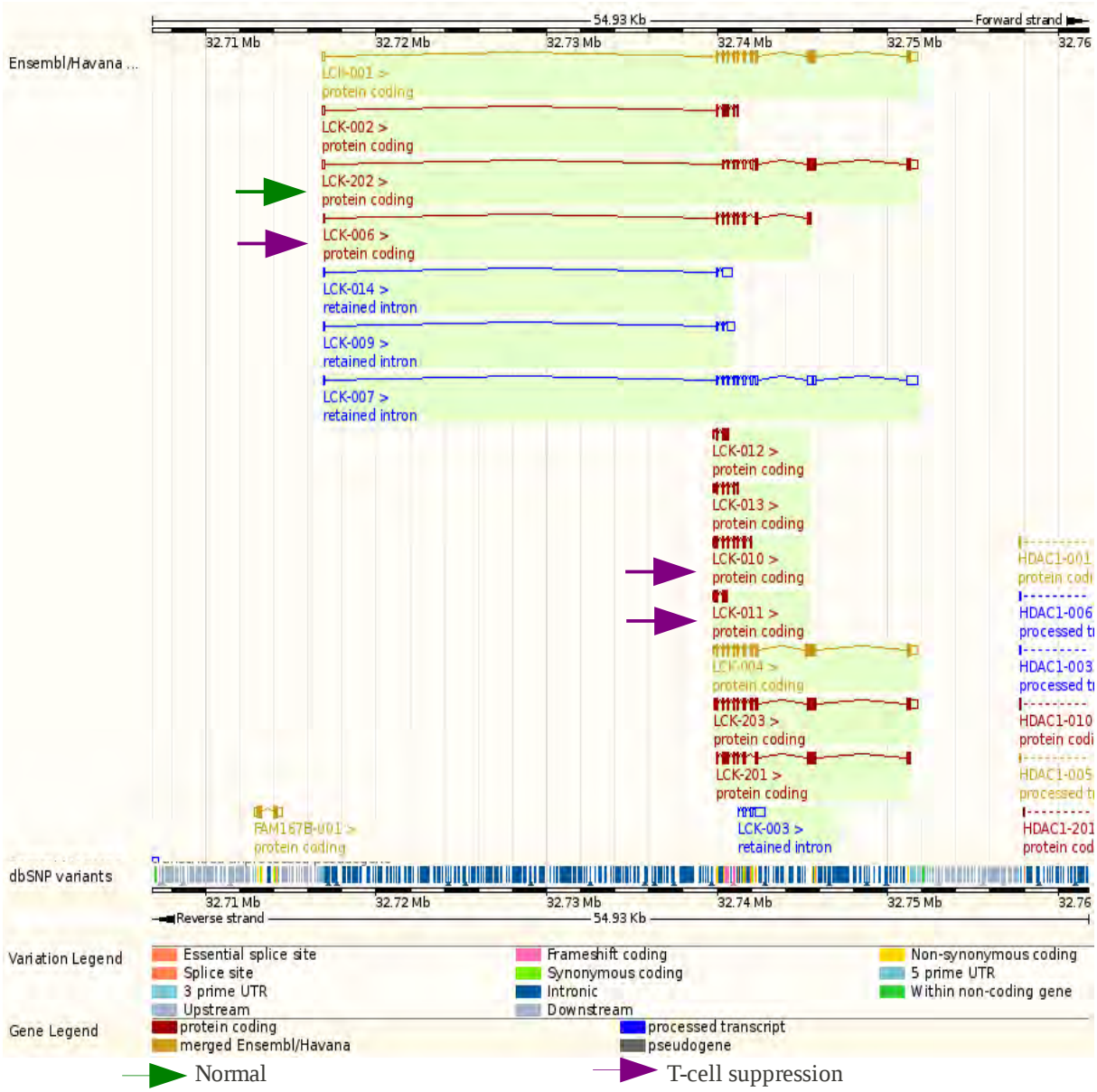


Figure 3.10: LCK gene summary modified from Ensembl. All the identified transcripts are protein-coding. LCK-010, LCK-011 and LCK-006 were expressed under T cell suppression conditions whilst LCK-202 was expressed under normal conditions. InterProScan was used to examine the domains of the proteins in section 3.4.4.2.

All the transcripts identified for *LCK* are protein-coding. Three of the transcripts, *LCK-010, LCK-011* and *LCK-006* were found to be expressed under T cell suppression conditions whilst *LCK-202* was found under normal conditions. As for both conditions all the transcripts are protein-coding, domains of the proteins were examined using InterProScan [65] to assess how the proteins differ functionally between the normal and suppressed conditions. Please see section 3.4.4.2 for more details.

3.4.4.2 *LCK* protein functional analysis - InterProScan

Figure 3.11: LCK protein characterisation using InterProScan. The InterProScan result showed that LCK-202 is the most complex of the transcripts whilst LCK-011 has low complexity.

Running the protein sequence of *LCK-202* through InterProScan results in 5 domain matches as well as 1 active site match (Figure 3.11). The domains belong to the protein kinase family and are:

- **Protein kinase, catalytic domain (IPR000719)**
  - This domain is found in serine/threonine-protein kinases, tyrosine-protein kinases and dual specificity protein kinases. Protein kinases play an important role in cellular processes such as proliferation, differentiation and apoptosis [65]. The domain has conserved regions such as the N-terminal of the domain has glycine rich residues that are involved in ATP binding whilst the middle of the domain has an aspartic acid residue which is crucial in the functionality of the kinase [65].
  - The 3-D fold of this domain is similar to the PI3K catalytic domain [65].
- **Serine-threonine/tyrosine-protein kinase catalytic domain (IPR001245)**
- **Tyrosine-protein kinase, active site (IPR008266)**
  - The kinase can transfer a phosphate group from ATP to a tyrosine residue in a protein and are divided into two groups, the receptor and cytoplasmic/non-receptor tyrosine kinases.
- **Protein kinase-like domain (IPR011009)** and
- **Tyrosine-protein kinase, catalytic domain (IPR020635)**.

The InterProScan result for the protein sequence of *LCK-006* matched 5 domains, namely:

- **Protein kinase, catalytic domain (IPR000719)**
- **SH2 domain (IPR000980)**
  - The SH2 domain is found within the Src oncoprotein as well as other signal-transducing proteins. The domain functions to recognise the phophorylated tyrosine residues on proteins and thus aids other proteins to bind the tyrosine phosphorylated sites.
- **Serine-threonine/tyrosine-protein kinase catalytic domain (IPR001245)**
- **Src homology-3 domain (IPR001452)**
  - The 55aa proteins are found in membrane-associated proteins such as proteins with enzymatic activity, in adaptor proteins without catalytic sequences and in cytoskeletal proteins.
- **Protein kinase-like domain (IPR011009)**

LCK-010 InterProScan results returned only two hits, namely the SH2 domain (IPR000980) and the Src homology-3 domain (IPR001452). *LCK-011* protein search resulted in no matches.

The InterProScan result show that *LCK-202* is the most complex of the transcripts whilst *LCK-011* has low complexity. Both *LCK-202* and *LCK-006* were expressed under normal conditions and seem to have similar domains except for the lack of the SH2 domain and the Src homology-3 domain in *LCK-202*. *LCK-010* lacks the protein kinase domains whilst *LCK-011* seem to be a very simple protein; both transcripts were found expressed under T cell suppression conditions. In the case of *LCK*, it seems that the loss of certain domains (e.g. the protein kinase-like domain) is what may contribute to the immunosuppression of the T cells.

3.4.5 *VCL*

3.4.5.1 VCL gene summary

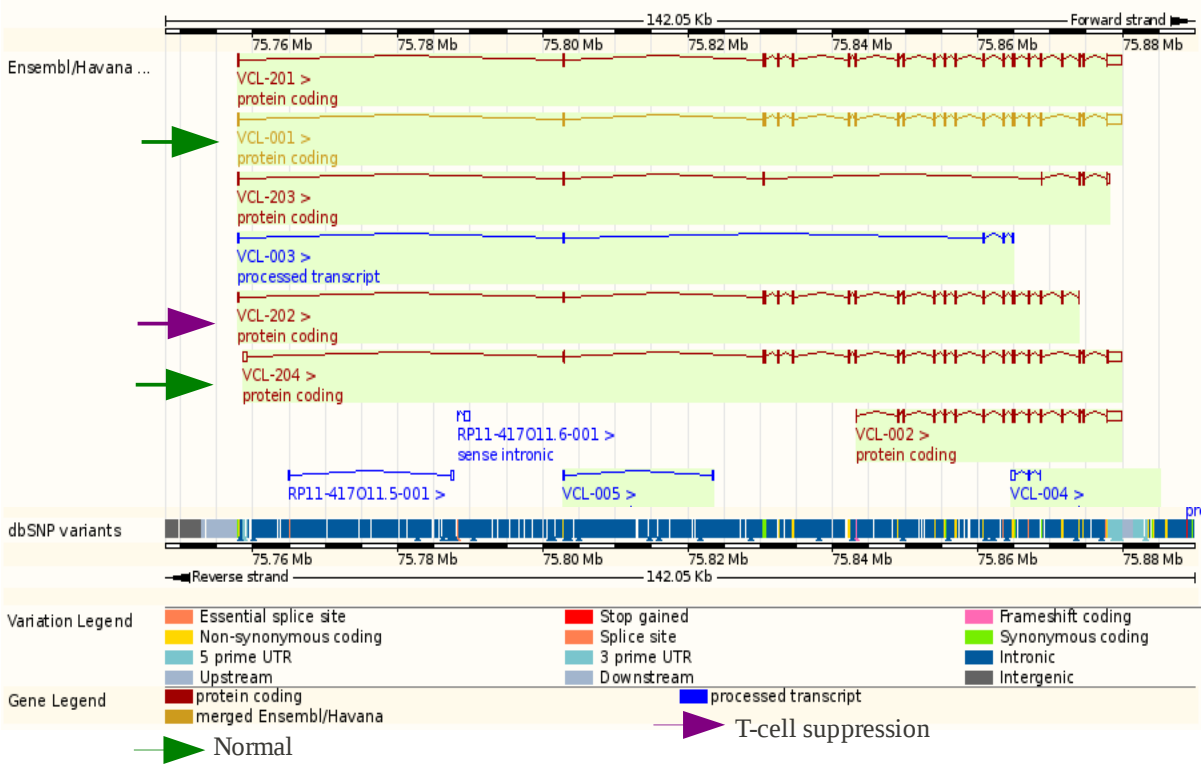


Figure 3.12: VCL gene summary modified from Ensembl. All the identified VCL transcripts are protein-coding. VCL-001 and VCL-204 are expressed under normal conditions whilst the truncated transcript, VCL-202, was found expressed under T cell suppression conditions. InterProScan was used to examine the domains of the proteins in section 3.4.5.2.

Transcripts *VCL-001* and *VCL-204* were found to be expressed under normal conditions with *VCL-202* expressed under T cell suppression conditions. The transcripts are very similar except that *VCL-202* is truncated. As all the transcripts are protein-coding, protein functionality was examined using InterProScan. Please see section 3.4.5.2, Figure 3.13.

3.4.5.2 VCL protein functional analysis - InterProScan

VCL-001 and VCL-204

VCL-202

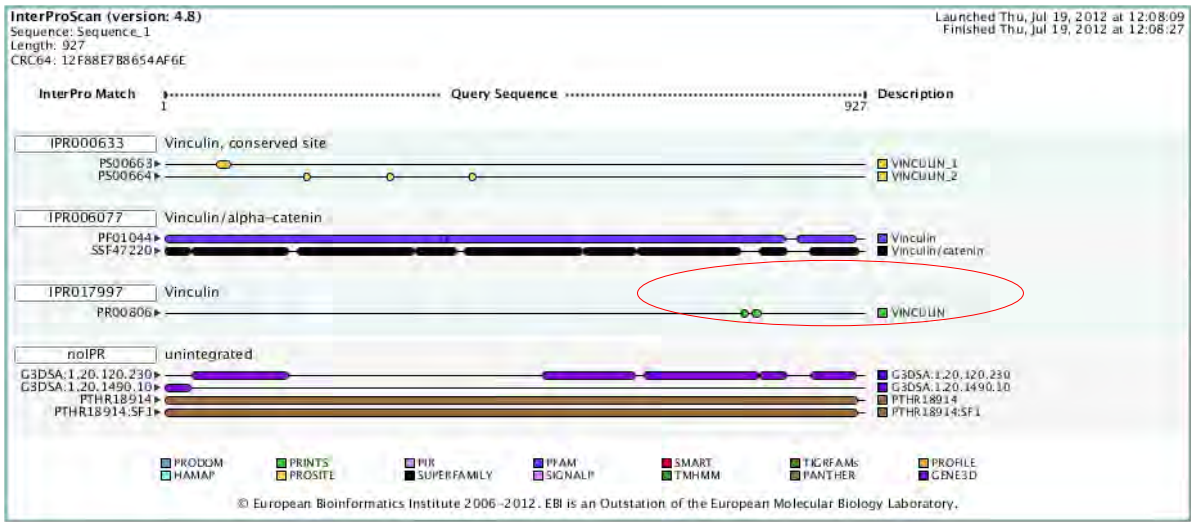


Figure 3.13: VCL protein characterisation using InterProScan. VCL-001 and VCL-204 both have all of the 7 motifs present on the transcripts whilst VCL-202 has lost some of the conserved motifs and only has the proline-rich motifs 1 and 2 present on the transcript.

Three InterProScan entries matched the protein sequence of *VCL-001* and *VCL-204*, namely:

- **Vinculin, conserved site (IPR000633)**
- **Vinculin/alpha-catenin (IPR006077)**
  - Alpha-catenins are related to vinculin and associate with the cytoplasmic domain of a variety of cadherins resulting in a complex that is linked to the actin filament network, and is important in the cadherins cell-adhesion properties. There are three types: alpha,

beta, and gamma. Alpha-catenins are related to vinculin but lack the repeated domain as well as the proline-rich segment.

- **Vinculin (IPR017997)**
  - The protein has 7 motifs that were drawn from conserved regions of the C-terminal; motifs 1 and 2 are in the proline-rich region and the rest of the motifs 3-7 are in the C-terminal domain [65].

VCL-001 and *VCL-204* both expressed under normal conditions, and have all of the 7 motifs present on the transcripts. *VCL-202*, which is found under T cell suppression conditions, has lost some of the conserved motifs and only has the proline-rich motifs 1 and 2 present on the transcript. Motif 1 has the following sequence (DELAPPKPPLP) and is located at position 763-773 on the *VCL-202* protein sequence whilst motif 2 is located at position 778-791 on the transcript and its amino acid sequence is (PPPRPPPPEEKDEE). The poly-proline regions are known to form a helical conformation that aids in intermolecular interactions such as signal transduction, antigen recognition, cell-cell communication and cytoskeletal organization [75].

3.4.6 *NRP1*

3.4.6.1 NRP1 gene summary

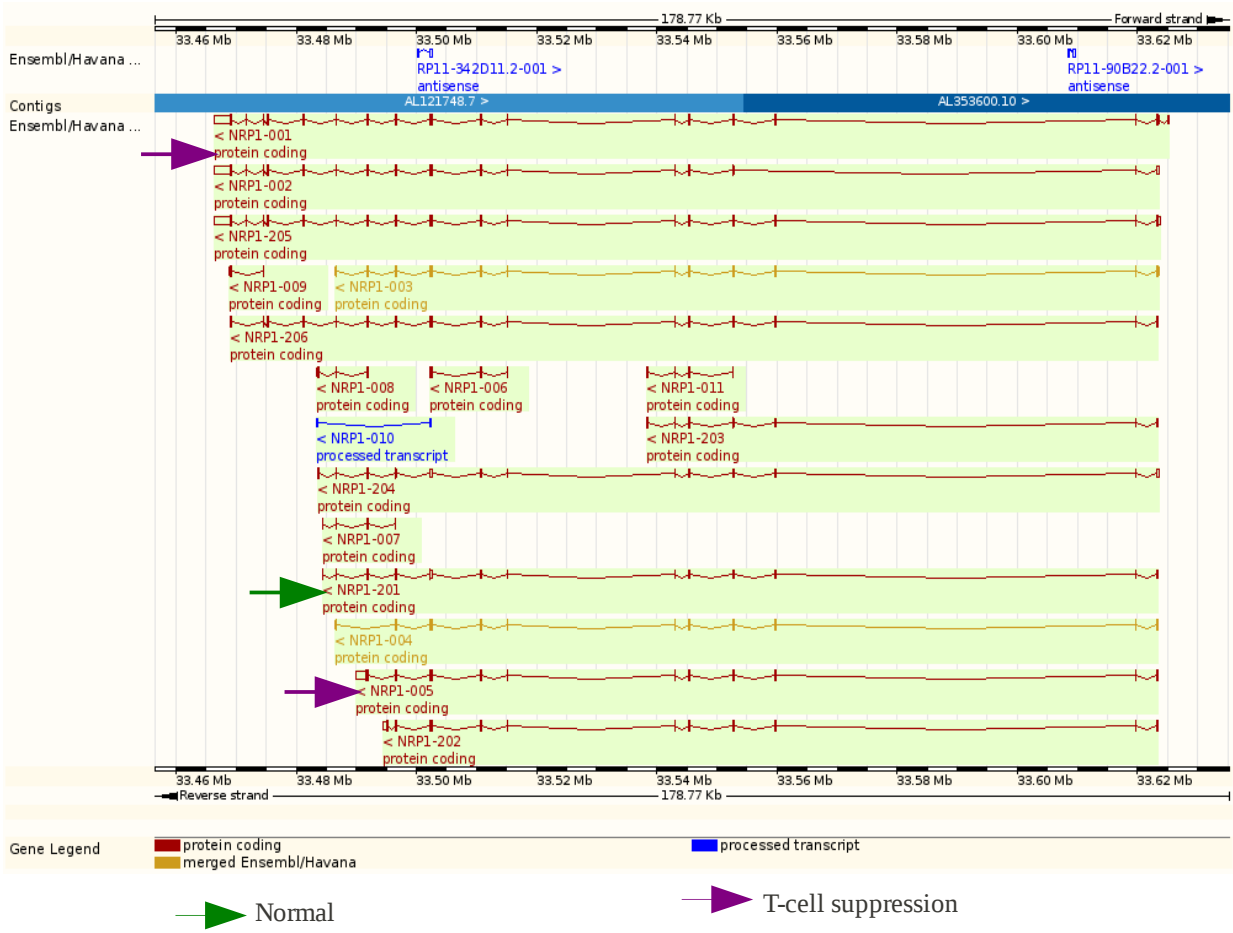


Figure 3.14: NRP1 gene summary modified from Ensembl. All the identified transcripts are protein-coding. NRP1-001 and NRP1-005 was expressed under T cell suppression conditions whilst NRP1-201 was expressed under normal conditions. InterProScan was used to examine the domains of the proteins in section 3.4.6.2.

Transcripts *NRP1-001* and *NRP1-005* were expressed under T cell suppression conditions whilst *NRP1-201* was expressed under normal conditions. All transcripts are protein-coding and were further analysed using InterProScan in section 3.4.6.2.

3.4.6.2 NRP1 protein functional analysis - InterProScan

NRP1-001

Figure 3.15: InterProScan result search for NRP1 protein-coding transcripts.

Seven hits matched the *NRP1-001* protein sequence, namely:

- **Coagulation factor 5/8 C-terminal type domain (IPR000421)**
  - This domain forms part of a larger functional domain which promotes binding to anionic phospholipids on the surface of platelets and endothelial cells [65].
- **CUB (IPR000859)**
  - The domain is found in extracellular and plasma membrane-associated proteins and is involved in a variety of functions including tissue repair, axon guidance and angiogenesis, cell signalling, fertilisation, haemostasis, inflammation, receptor-mediated endocytosis, and tumour suppression [65]. The domain contains four cysteine residues that form two disulphide bridges [78]
- **MAM domain (IPR000998)**
  - The MAM domain is found in cell surface proteins and it acts as an adhesion domain.
- **Galactose-binding domain-like (IPR008979)**
  - The domain binds ligands such as cell-surface-attached carbohydrate substrates for

galactose oxidase, phospholipids on the outer side of the mammalian cell membrane for coagulation factor Va and membrane-anchored ephrin for the Eph family of receptor tyrosine kinases [65].

- **Concanavalin A-like lectin/glucanase (IPR008985)**
  - Con A-like domains are important in cell recognition and examples of proteins that have the domain include the sex hormone-binding globulins which transport sex steroids in blood and regulate their access to target tissues, neurexins which are expressed in hundreds of isoforms on the neuronal cell surface where they may function as cell recognition molecules and sialidases that are found in both microorganisms and animals and function in cell adhesion and signal transduction [65]. Other proteins include pentraxins PTX3 which is a TNFalpha-induced protein produced during inflammation by adipose cells [65].
- **Neuropilin (IPR014648)**
  - This domain is the parent of the other domains identified for the transcript *NRP1-001*.
- **Neuropilin-1, C-terminal (IPR022579)**

NRP1-201 only had two InterProScan matches that are also found in *NRP1-001*, namely the Coagulation factor 5/8 C-terminal type domain (IPR000421) and galactose-binding domain-like (IPR008979). *NRP1-005* had four matches, Coagulation factor 5/8 C-terminal type domain (IPR000421), CUB (IPR000859), Galactose-binding domain-like (IPR008979) and Neuropilin (IPR014648). The lack of the CUB domain in the *NRP1-201* transcript could potentially result in cancer as the transcript cannot bind semaphorins which are proteins with anti-tumour properties [78], whilst the loss of axon guidance and cell signalling functions could lead to T cell silencing as *NRP1* has been shown to play a role in the stimulation of resting T cells by dendritic cells (DCs) [79].

3.4.7 *PRMT5*

3.4.7.1 PRMT5 gene summary

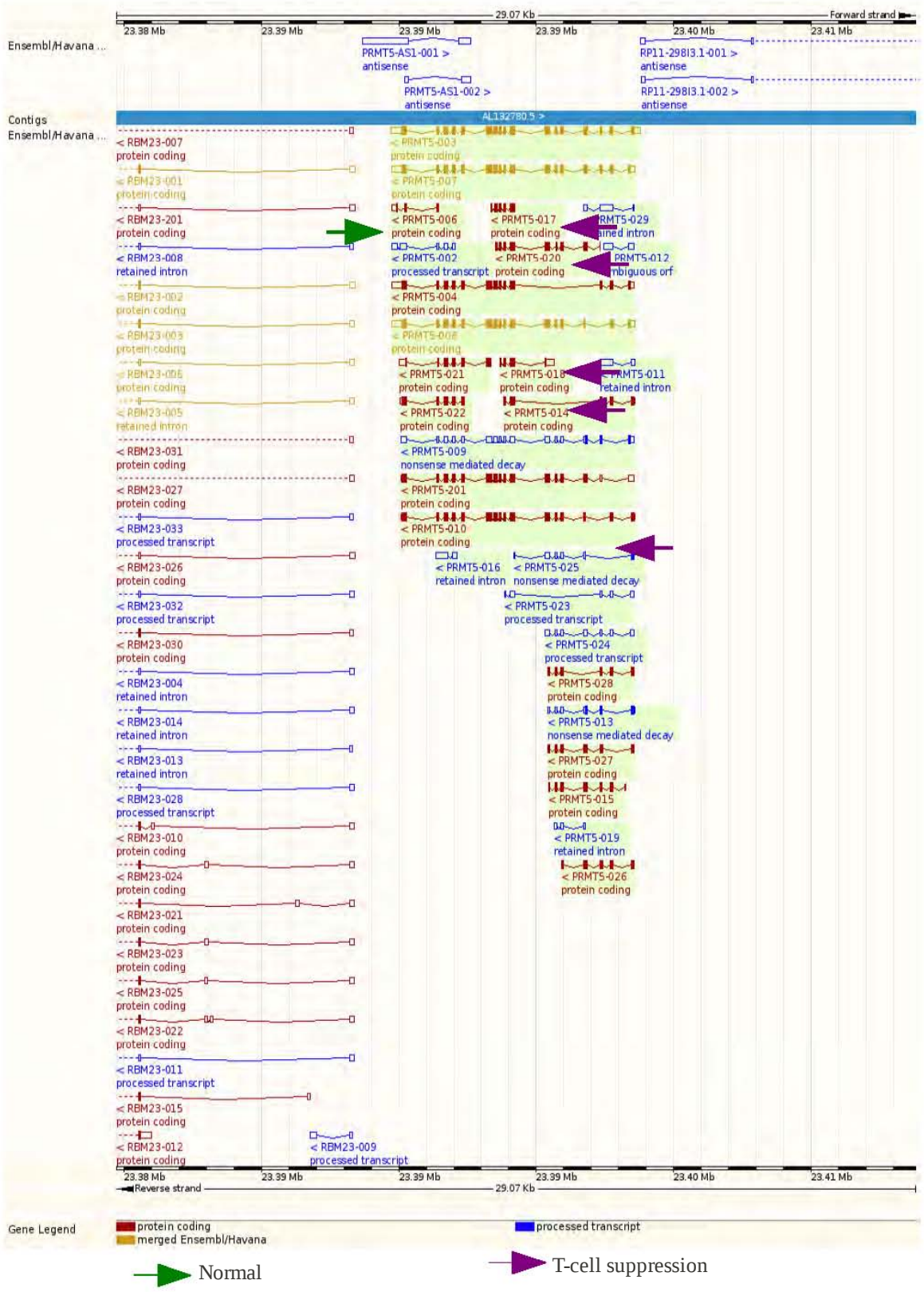


Figure 3.16: PRMT5 gene summary modified from Ensembl. All the identified transcripts are protein-coding. Further analysis was performed using InterProScan, see section 3.4.7.4.

Almost all the transcripts were found expressed under T cell suppression conditions and are protein-coding. Exceptions are *PRMT5-006*, a protein coding transcript, which was expressed under normal conditions, and *PRMT5-025*, a T cell suppression transcript, which undergoes nonsense-mediated decay. Supporting evidence for the non-coding transcript is provided in section 3.4.7.3 whilst for the protein-coding transcripts further analysis was performed using InterProScan as shown in section 3.4.7.4.

3.4.7.3 PRMT5 supporting evidence

PRMT5-025

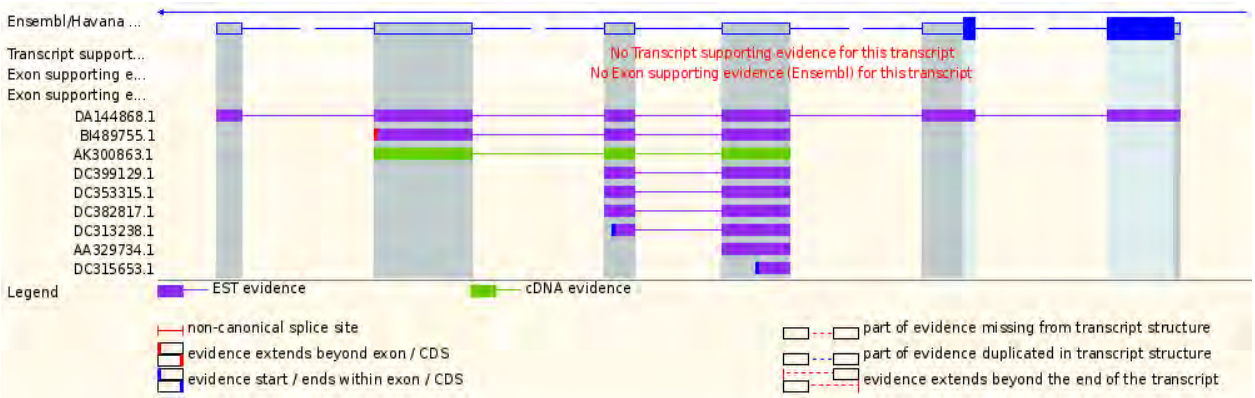


Figure 3.17: Ensembl EST/cDNA supporting evidence for the PRMT5 transcripts. PRMT5-025 was found expressed in brain tissue (DA144868.1 , DC353315.1 and DC313238.1), lung and testis (BI489755.1, DC399129.1 and DC382817.1) and the small intestine (AK300863.1).

The transcript, according to EST evidence, was found expressed in brain tissue (DA144868.1 , DC353315.1 and DC313238.1), lung and testis (BI489755.1, DC399129.1 and DC382817.1) and the small intestine (AK300863.1). Unfortunately the state under which the ESTs were found, could not be ascertained.

3.4.7.4 PRMT5 protein functional analysis - InterProScan

The protein sequences for *PRMT5-006*, *PRMT5-017*, *PRMT5-018* and *PRMT5-020* matched 2 InterPro entries, namely:

- **Protein arginine N-methyltransferase PRMT5 (IPR007857)**
  - The protein is a key mitotic regulator and is involved in Jak signalling. Methyltransferases are also involved in biosynthesis, signal transduction, protein repair, chromatin regulation as well as gene silencing [65].

- **Protein arginine N-methyltransferase (IPR025799)**

Unlike the previous cases discussed, there does not appear to be any difference in domain composition of PRMT5 proteins produced under normal versus T cell suppression conditions.

3.5 Transcript variations

We then looked for SNPs within the isoforms identified in section 3.4, to investigate the contribution, if any, to T cell suppression. SNPs were selected, using dbSNP, based on the following criteria: (1) should have phenotype data, (2) the SNP can be located in the splice site, i.e. can influence the alternative splicing and thus the function of the gene, or (3) the SNP can be located in an splicing regulatory element (SRE) for example an exon splice enhancer (ESE) site. Genes that satisfied at least two of the criteria were further analysed.

The gene *ATM* had SNPs that satisfied the first two criteria. The gene, located on 11q22-q23, encodes a protein that belongs to the PI3/PI4-kinase family of proteins and mutations are located all over the gene and are associated with ataxia telangiectasia as well as cancer [42]. One SNP that satisfied criteria 1. and 2. is CM063853, which, according to the public Human Gene Mutation Database (HGMD) (www.hgmd.org), is an A/T splice site variant in the *ATM* gene on chromosome 11:108100050 (Figure 3.18).

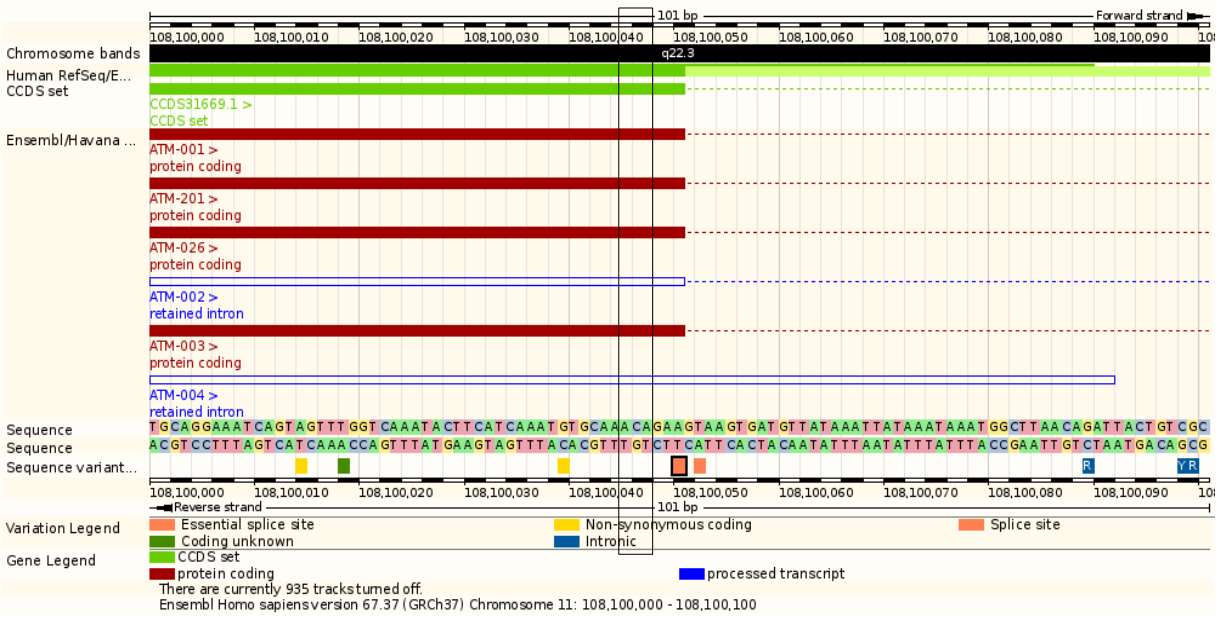


Figure 3.18: The chromosome location of the A/T splice variant, CM063853. The variant is known to contribute to ataxia telangiectasia.

The SNP is present in the T cell suppression-related transcripts *ATM-002* and *ATM-004* (retained introns) as well as other transcripts, refer to Figure 3.18. The T allele of CM063853 is associated with the increased risk of ataxia telangiectasia, an autosomal recessive disorder.

Another SNP found on the *ATM-002* and the *ATM-004* (retained intron) as well as the other transcripts is the rs1800054, which is a C/G variation on chromosome 11:108098576 (Figure 3.19). The variant results in missense mutations causing a ser49 to cys switch which has been linked to breast cancer susceptibility [43-45]. Missense mutations are known to change the sequence and structure of a protein leading to disease as a result of the altered splicing machinery process [76]. This non-synonymous SNP occurs in an exon splice enhancer (ESE) site recognised by the SR-protein sc35 [58]. Since the SNPs, rs1800054 and CM063853, are found in both normal and T cell suppression transcripts they could have an effect under all conditions, although the SNP in the splice site may determine whether the longer T cell suppression *ATM-004* transcript is produced preferentially.

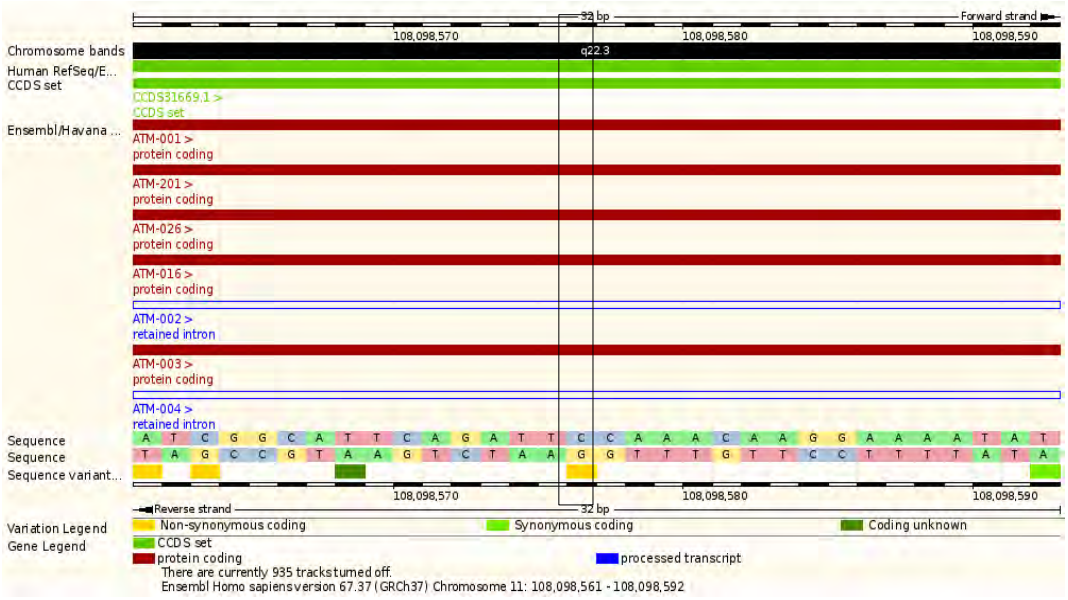


Figure 3.19: The chromosome location of the variant rs1800054.

Like the SNPs identified for *ATM*, the SNP for *MXI1* is present in transcripts found under both normal and T cell suppression conditions. The SNP rs14401, which is a C/T substitution, is found downstream of the gene on chromosome 10 at position 112045171, which is a 3' UTR (Figure 3.20). This SNP occurs in an exon splice enhancer (ESE) site recognised by the SR-protein srp40 and the SNP causes a 'lose' effect on the site [15]. SNPs located in the 3' UTR could affect

MXI1 mRNA by destabilising the *MXI1* transcripts or polyadenylation site selection [77]. The variant results in cytotoxicity of lymphoblastoid cell lines (BLCLs) to CTL-1B9.

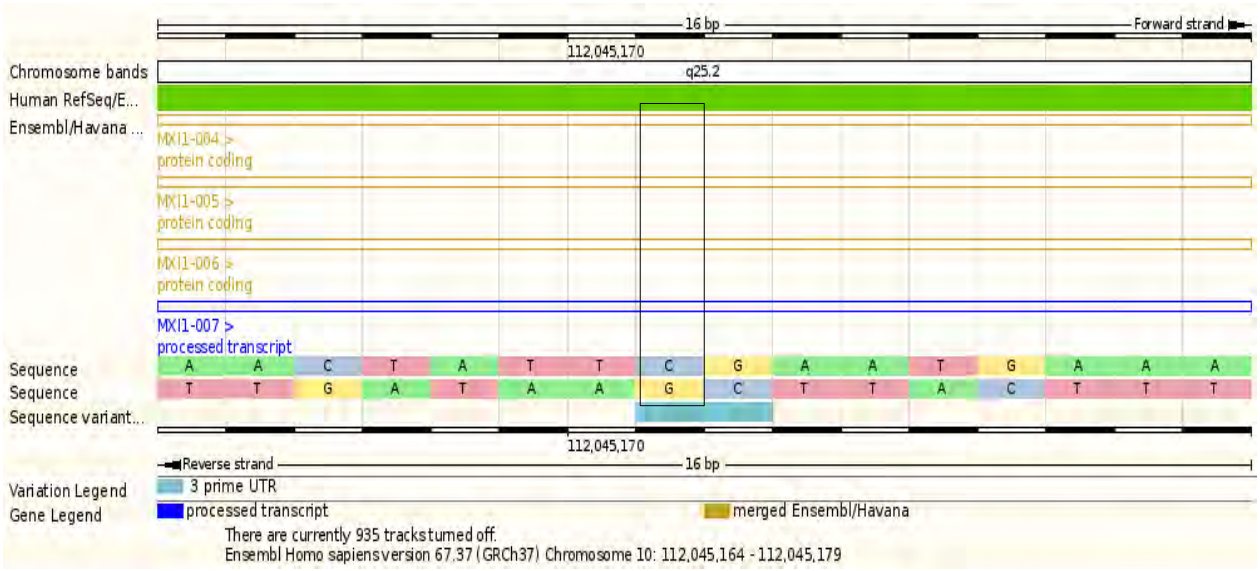


Figure 3.20: The chromosome location of the variant rs14401, which is known to cause cytotoxicity of the lymphoblastoid cell lines.

It seems that SNPs found within the splice sites and/or an SRE (splicing enhancers or silencers), do have an effect on alternative splicing. This modification can potentially lead to disease, as identified in this study, or may affect drug metabolism [88].

SNPs found in the other genes only satisfied one criteria, i.e. only had phenotype data.

3.6 Functional analysis

In order to interpret the results biologically, we performed functional analysis using DAVID and the KEGG Mapper [74]. Enrichment analysis was performed using DAVID which allows for a large set of genes to be grouped into functionally related genes. The KEGG Mapper was used to map the genes onto relevant pathways. To understand the gene to gene interactions, interaction network analysis was performed using the STRING database.

3.6.1. Enrichment analysis - DAVID

Using DAVID for enrichment analysis, using default parameters of count threshold of 2 and Ease score (p-value) of 0.1, results in 31 chart records for the set of 7 genes overlapping all the datasets. However, if we leave the count threshold at 2 (terms are meaningful if more than one gene is associated with it) but lower the Ease score from 0.1 to 0.01 (the lower the score, the more enriched the terms are) we end up with 3 enrichment terms associated with the 7 genes passing the score, namely: alternative splicing (7 genes), the GO term cytoskeleton (4 genes) and disease mutation (4 genes). The genes annotated to the term cytoskeleton are *CALD1*, *LCK*, *ATM* and *VCL*, whilst the disease mutation term applied to *MXI1*, *LCK*, *ATM* and *VCL*.

To further analyse the genes, we used DAVID's functional annotation table which allows one to view a lot more detail about the individual genes. *MXI1* is involved in DNA binding, has transcription repressor activity, acts as a T cell silencer as well as implication in neurofibrosarcoma and prostate cancer. Low levels of *LCK* have been implicated in severe combined immunodeficiency (SCID) according to OMIM whilst *ATM* is part of a range of cancers including breast cancer, lymphoma as well as ataxia telangiectasia.

3.6.2 Pathway analysis - KEGG Mapper

The KEGG Mapper was used for pathway mapping of the 7 genes. The pathway analysis showed that *ATM*, *LCK* and *NRP1* are involved in HTLV-1 infection (hsa05166). ATM has been shown to regulate T cell survival during HTLV-1 infection [66]. *ATM* and *LCK* are represented in the NF-kappa B signalling pathways (hsa04064). *CALD1* plays a role in vascular muscle contraction (hsa04270), while *LCK* is involved in natural killer cell mediated cytotoxicity (hsa04650) and T cell receptor signalling pathway (hsa04660). *ATM* is involved in apoptosis (hsa04210), it plays a role in the p53 signalling pathway (hsa04115) and is involved in transcriptional misregulation in cancer (hsa05202), whilst *VCL* is involved in focal adhesion (hsa04510) and regulates the actin cytoskeleton (hsa04810). *NRP1* is involved in axon guidance (hsa04360) and has been shown to regulate T cell activation at the immune synapse [67,68]. *PRMT5* forms part of the RNA transport pathway (hsa03013).

3.6.3 Interaction analysis with STRING

In order to determine the level of interaction between the isoforms identified as occurring under T cell suppression conditions, coexpression analysis was performed using the web-based tool STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) [62].

STRING was used to determine the level of interactions between the identified isoforms. Only T cell suppressed isoforms that have protein products were analysed. We not only examined the level of interaction between the isoforms but also between PIK3CA. PIK3CA is the catalytic subunit of the phosphatidyl-inositol-3-kinase (PI3K) and the signalling pathway is targeted by the measles virus (MV) in T cells. The pathway is important in a number of cellular processes such as cell growth, proliferation and survival, and interference of the signalling by MV results in the arrest of the cell cycle in T lymphocytes leading to T cell suppression.

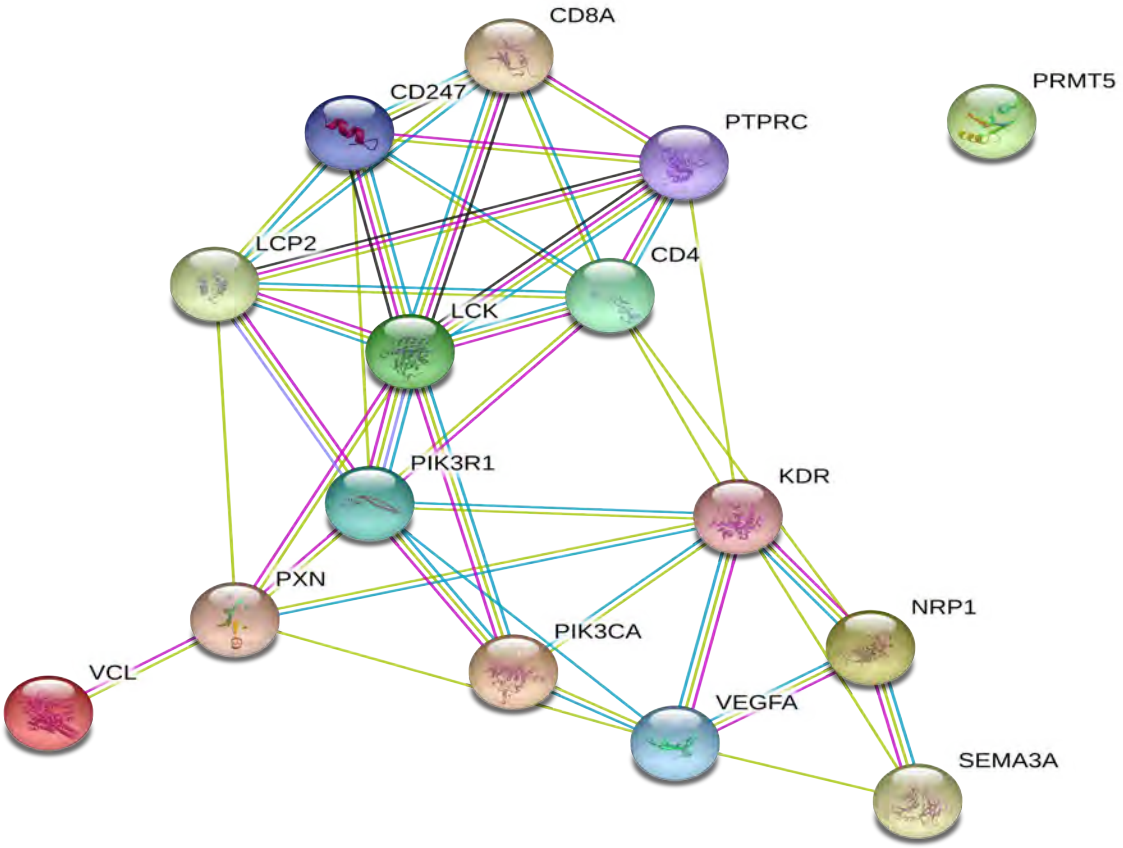


Figure 3.21: STRING predicted gene interactions. The colour of each of the edges represents the type of evidence that exists for that interaction: a red line indicates the presence of fusion evidence, a green line indicates neighbourhood evidence, a blue line indicates co-occurrence evidence, a purple line indicates experimental evidence, a yellow line indicates text-mining evidence, a light blue line indicates database evidence, and a black line indicates co-expression evidence [62]. The genes of interest were ran through STRING and are shown to have interaction with PI3K with the exception of PRMT5. The rest of the genes are predicted functional partners of the kinase and the isoforms.

Using the multiple sequences tab to find the interactions, Figure 3.21 illustrates the output from the search. As already mentioned, only isoforms that have protein products were analysed using the tool. Protein sequences, in FASTA format, of the following transcripts were run through the program: PIK3CA, *LCK-010*, *LCK-011*, *VCL-202*, *NRP1-001*, *NRP1-005*, *PRMT5-017*, *PRMT5-018*, *PRMT5-020* and *PRMT5-014*. As seen in Figure 3.20, the isoforms of the genes had some level of interaction with the PIK3CA except for *PRMT5*. None of the *PRMT5* isoforms had any interaction with the kinase or the genes. The other genes shown in Figure 3.21 are genes predicted by the program to be functional partners of the kinase as well as the isoforms.

Table 3.4: STRING predicted functional partners of the isoforms.

Gene	Description	Interaction	
		Isoform	Score
CD4	CD4 molecule. Accessory protein for MHC class-II antigen/T cell receptor interaction. Regulates T cell activation. Induces the aggregation of lipid rafts.	LCK-006 LCK-010 LCK-011 LCK-202	0.999
PIK3R1	Phosphoinositide-3-kinase. It's the regulatory subunit 1 (alpha) and binds to activated (phosphorylated) protein-Tyr kinases, through its SH2 domain, and acts as an adapter, mediating the association of the p110 catalytic unit to the plasma membrane. Necessary for the insulin-stimulated increase in glucose uptake and glycogen synthesis in insulin-sensitive tissues.	LCK-006 LCK-010 LCK-011 LCK-202	0.994
VEGFA	Vascular endothelial growth factor A. Growth factor active in angiogenesis, vasculogenesis and endothelial cell growth. Induces endothelial cell proliferation, promotes cell migration, inhibits apoptosis, and induces permeabilization of blood vessels. Binds to the VEGFR1/Flt-1 and VEGFR2/Kdr receptors, heparan sulfate and heparin. Neuropilin-1 binds isoforms VEGF-165 and VEGF-145. Isoform VEGF165B binds to VEGFR2/Kdr but doesn't activate downstream signaling pathways, doesn't activate angiogenesis and inhibits tumor growth.	NRP1-001 NRP1-005 NRP1-201	0.999
		PIK3CA	0.990
CD247	CD247 molecule. Probable role in assembly and expression of the TCR complex as well as signal transduction upon antigen triggering.	LCK-006 LCK-010 LCK-011 LCK-202	0.999
PTPRC	Protein tyrosine phosphatase. Receptor type, C and is required for T cell activation through the antigen receptor. Upon T cell activation, recruits and dephosphorylates SKAP1 and FYN.	LCK-006 LCK-010 LCK-011 LCK-202	0.999
KDR	Kinase insert domain receptor (a type III receptor tyrosine kinase). Receptor for VEGF or VEGFC. Has a tyrosine-protein kinase activity. The VEGF-kinase ligand/receptor signaling system plays a key role in vascular development and regulation of vascular permeability. In case of HIV-1 infection, the interaction with extracellular viral Tat protein seems to enhance angiogenesis in Kaposi's sarcoma lesions.	NRP1-001 NRP1-005 NRP1-201	0.999
		PIK3CA	0.952
PXN	Paxillin. Cytoskeletal protein involved in actin-membrane attachment at sites of cell adhesion to the extracellular matrix (focal adhesion).	LCK-006 LCK-010 LCK-011 LCK-202	0.657
		VCL-001 VCL-202 VCL-204	0.999
CD8A	CD8a molecule. Identifies cytotoxic/suppressor T-cells that interact with MHC class I bearing targets. CD8 is thought to play a role in the process of T cell mediated killing. CD8 alpha chains binds to class I MHC molecules alpha-3 domains.	LCK-006 LCK-010 LCK-011 LCK-202	0.999
SEMA3A	Sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3A. Induces the collapse and paralysis of neuronal growth cones. Binds to the complex neuropilin-1/plexin-1.	NRP1-001 NRP1-005 NRP1-201	0.999
LCP2	Lymphocyte cytosolic protein 2. Involved in T cell antigen receptor mediated signalling.	LCK-006 LCK-010 LCK-011 LCK-202	0.999

Table 3.4 highlights the predicted partners together with their descriptions as per the STRING database. Also included in the table is a column that shows which of the identified isoforms have a

direct interaction with the predicted partners as well as a score value for the interactions. The score is a probabilistic confidence score, which is an estimate of how likely a given association describes a functional linkage between two proteins that is at least as specific as that between an average pair of proteins annotated on the same 'map' or 'pathway' in KEGG [62]. The various major sources of interaction/association data in STRING are benchmarked independently and a combined score is calculated which gives a higher score when more than one type of information supports a given interaction [62]. From the analysis, it is clear that *LCK* seems to have direct interactions (scores < 0.99) with many predicted functional partners and the *PIK3CA*. The predicted functional partners are involved in a range of T cell functions such as T cell activation (*CD4*, *PTPRC*), T cell antigen receptor mediated signalling (*LCP2*), T cell mediated killing (*CD8A*) as well as assembly and expression of the T cell receptor TCR (*CD247*). *NRP1* and *PIK3CA* have direct relationship with *VEGFA*, *KDR* and *SEMA3A*. *VEGFA* plays a role in cell proliferation, cell migration and inhibition of apoptosis; *KDR* is involved in vascular regulation via its tyrosine-protein kinase activity and *SEMA3A* has a negative effect on neuronal growth cones. *VCL* is associated with *PXN*, which is involved in focal adhesion just like *VCL*.

The results confirm that there is a functional connection between the final gene list we identified as being involved in T cell suppression through alternative splicing. Compared to the controls, i.e. the isoforms found under the normal conditions, the isoforms found under the T cell suppression conditions are polymorphic in nature due to alternative splicing. This led to loss of domains important in the phosphorylation of downstream targets, which we think might be one of the causes of gene silencing.

CHAPTER 4

4.1 DISCUSSION

MV induced immunosuppression occurs as a result of the interference of PI3K/AKT signalling leading to T cell suppression, and the interference has been shown to promote the production of alternatively spliced protein isoforms such as the SIP100, an alternatively spliced isoform of *SHIP145* [35]. These alternatively spliced protein products are thought to cause T cell suppression by interfering with the signalling pathways that are important in T cell activation [35]. This project addressed the question of whether more of these alternatively spliced isoforms exist, and could thus be potential genetic markers to be used in identifying T cell suppression.

A total of 9 microarray experiments were identified and further analysed using GEO2R, a gene expression analysis tool. GEO2R is a web-based application that uses Limma to analyse the data and allows one to download the R script used to run the analysis. Although the tool is useful in that a large number of GEO data can be analysed, there are some weaknesses to the system one must be aware of:

- *GEO allows data that is processed and normalised by different methodologies*
    - Submitters to the GEO repository employ different methodologies when processing and normalising their data. GEO2R takes the values as they are and identifies differentially expressed genes, unless if no log transformation has been applied to the values, then it will log transform the values.
- *Not all experiments are analysed by GEO2R*
    - GEO2R works only on data tables (i.e have values that can be used to identify the differentially expressed genes). Data from high-throughput sequencing and genome-array tiling cannot be analysed by the system.
    - Data should only use one series.
    - Works on sample size limit of 255.
    - A processing time limit of 10 minutes is allowed on jobs.

Nevertheless, the tool was used, and the following genes were found to have different isoforms

produced under normal vs. T cell suppression conditions: *ATM*, *CALD1*, *LCK*, *VCL*, *MXI1*, *NRP1* and *PRMT5*. *ATM*, *PRMT5* and *VCL* overlapped with the differentially regulated list of the GeneChip exon array whilst *CALD1*, *MXI1*, *NRP1* and *LCK* overlapped with the alternatively spliced list. Genes such as *CD44*, *BRCA1*, *FGF1* and *PSEN1* only had isoforms found under one condition only, either normal or T cell suppression but not both. This is interesting as these genes' alternative splicing has been well studied [49]. The list that showed genes that had isoforms produced under one condition only, also included the gene *INPPL1* which is part of the *SHIP* family of genes.

The probes representing the different isoforms for genes *LCK*, *MXI1*, *VCL* and *PRMT5*, were mostly significantly expressed with p-values <0.05. Probes representing *ATM*, *CALD1* and *NRP1* had p-values >0.05. The Benjamini & Hochberg false discovery rate method was selected in GEO2R to calculate the p-values as it provides a good balance between discovery of statistically significant genes and limitation of false positives [57]. Ensembl was used to identify the gene transcripts represented by the probes. It was interesting to note that for three of the genes whose transcripts were found to be expressed under either T cell suppression or normal conditions, the biotype was retained introns (*ATM* and *CALD1*), nonsense-mediated decay (*CALD1*) and processed transcripts (*MXI1*). Both the nonsense-mediated and processed transcripts are non-coding with the former preventing expression of truncated proteins and the latter does not have an open reading frame [59]. Retained introns are known to cause in frame stop-codons and result in translation of prematurely terminated proteins causing nonsense-mediated mRNA decay (NMD) [1]. Therefore for these three genes, under at least one of the conditions (either normal or T cell suppression), no actual protein product is made. Interestingly the *SIP110*, found to be expressed under T cell suppression conditions, also has a retained intron. Please see Figure 4.1, the retained intron is circled in red.

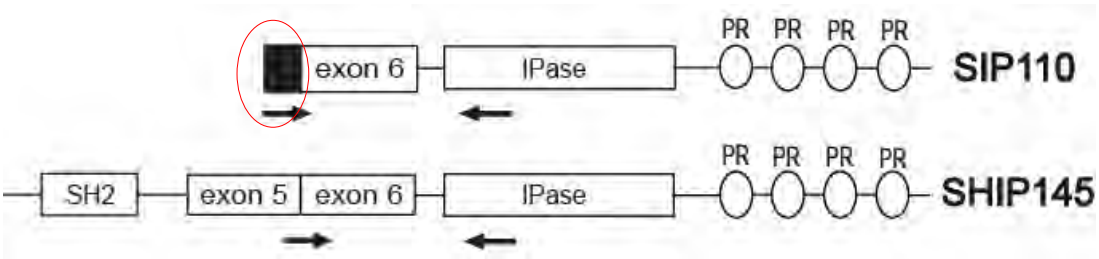


Figure 4.1: SIP110 is a retained intron [50]. Exon 5 + 6 are constitutively spliced in SHIP145. SIP110 is an alternatively spliced form of SHIP145 which has intronic sequences.

It seems that intron retention by the alternative splice sites of the identified isoforms, forms the basis of pathogen-mediated T cell suppression and could potentially be used as biomarkers. Biomarkers are biologic molecules that are used to indicate the state and progression of a disease [86]. A good biomarker should be easy to detect at an early stage and should be measurable across different populations [86]. In this study, supporting evidence in the form of EST/cDNAs for the non-coding transcripts was examined and as mentioned in chapter 1, one of the challenges with using ESTs is poor annotation. The source of the EST/cDNA was easily identified for this study but the condition under which it was found was difficult to ascertain. The reader must bear in mind these challenges. The EST/cDNA supporting evidence for the identified non-coding isoforms *ATM-002*, *ATM-004*, *CALD1-004* and *PRMT5-025* could potentially be used as biomarkers as they can be detected clinically using RT-PCR but do not meet all the criteria that constitute a good biomarker as we do not know under which condition the EST/cDNAs were found. On the other hand, the EST/cDNA supporting evidence for the identified non-coding isoforms *MXI1-003* and *MXI1-012* had better annotation, as the cDNA BC016678.1 was found expressed in skin tissue with melanotic melanoma and could thus make a better candidate for a biomarker for T cell suppression. Further studies would need to be performed to determine not only the conditions but the stage of the disease under which the EST/cDNAs are found.

For genes that had protein-coding transcripts expressed under both normal and T cell suppression conditions, protein functionality was examined using InterProscan [65]. The isoform *LCK-010*, found expressed under T cell suppression conditions, was found to lack the protein kinase-like domain which was found present in the *LCK-202* and *LCK-006* both expressed under normal conditions. The protein kinase-like domain plays an important role in the phosphorylation of other protein targets and it seems that the loss of the domain contributes to the immunosuppression of the T cells. Further evidence in support of the importance of the protein kinase-like domain's role in immunosuppression can be seen in the *LCK-202* transcript. *LCK-202*, found expressed in normal conditions, lacks the SH2 and the SH3 domains but has the protein kinase-like domain. The SH2 and SH3 domains are both adaptors that play a role in signalling protein-protein interactions [87].

Another gene that had interesting results from the InterProScan was *VCL*. *VCL-001* and *VCL-204* both expressed under normal conditions, were found to have all 7 motifs present on the transcripts. In the case of the *VCL-202* found expressed under T cell suppression conditions, a lack or loss of the conserved motifs seems to contribute to the pathogen-mediated immunosuppression. The proline-rich motifs 1 and 2 seem to be crucial in both the normal and diseased conditions, as they were found present on all the *VCL* transcripts. The proline-rich regions play a role in signal

transduction, antigen recognition, cell-cell communication as well as cytoskeletal organization [75]. The identified protein-coding isoforms could potentially be used as biomarkers as antibodies that bind specifically to the protein isoforms could be manufactured and used to detect the isoforms using immunohistochemical techniques such as enzyme-linked immunosorbance assay (ELISA) and RT-PCR [86].

We performed further analysis on the identified isoforms to investigate the variations. SNPs were selected using dbSNP based on phenotype data, location within a splice site, or the mutation results in an SRE. Two SNPs for *ATM* which are SNP CM063853 (A/T) and rs1800054 (C/G) passed the criteria and are known to cause ataxia telangiectasia as well as breast cancer respectively [42-45]. The SNP is found in an exon splice enhancer (ESE) site for the SR-protein sc35 [59]. For *MXI1*, the SNP which passed the criteria is the SNP rs14401 (C/T) which is found downstream of the gene at the 3' UTR.
There seems to be evidence, albeit very little in the form of only 3 SNPs belonging to the 2 genes, that suggests that the isoforms identified in this study to regulate T cell suppression are polymorphic [86]. Techniques such as PCR and ELISA could be used as well to detect these polymorphic isoforms/biomarkers.

As the aim of the study was to identify the genes and their alternatively spliced products that occur as a result of PI3K interference, an interaction analysis with PI3K was performed using the STRING database. The interactions between the protein-coding isoforms and PIK3CA were examined. PIK3CA is the catalytic subunit of the phosphatidyl-inositol-3-kinase (PI3K) and the signalling pathway is targeted by the measles virus (MV) in T cells, which results in the arrest of the cell cycle leading to T cell suppression. The program predicted functional partners for the isoforms and *LCK* had direct interactions (scores < 0.99) with many predicted functional partners as well as the *PIK3CA*. The predicted functional partners that *LCK* had relationships with, are involved in T cell activation (*CD4*, *PTPRC*), T cell antigen receptor mediated signalling (*LCP2*), T cell mediated killing (*CD8A*) and assembly and expression of the T cell receptor (TCR) (*CD247*). *NRP1* and *PIK3CA* have a direct relationship with *VEGFA*, *KDR* and *SEMA3A*. *VEGFA* plays a role in cell proliferation, cell migration and inhibition of apoptosis; *KDR* is involved in vascular regulation through its tyrosine-protein kinase activity and *SEMA3A* has a negative effect on neuronal growth cones. *VCL* is associated with *PXN*, which is involved in focal adhesion just like *VCL*.

It seems that the identified transcripts do not act alone, but rather in cooperation with other co-receptors. Perhaps a biomarker could be designed in such a way as to detect not only *LCK* and *VCL*

isoforms in this case, but also their functional partners as predicted by STRING. This could potentially increase the accuracy with which to detect pathogen-mediated T cell suppression. As an example, the *LCK* transcript was predicted to interact with the the co-receptors *CD8A*, *CD247* and *CD4*. This was not surprising as these co-receptors are known to play a role in the TCR [80]. It appears that co-receptor signalling, kinase signalling, conservation of motifs as well as assembly of the TCR play an important role in T cell suppression.

4.2. CONCLUSION

The microarray experiments that were used to support the genes identified as causing T cell suppression, involved an investigation into the effect of pathogens (MV, HCMV, HIV and RV) on monocytes or PBMCs. According to Avota et. al (2006) the retained intron isoform, SIP110, was found expressed under T cell suppression conditions [50]. In this study, we also identified alternatively spliced isoforms expressed under T cell suppression conditions that had intronic sequences and could potentially be used as biomarkers. The lack of properly annotated EST/cDNAs for the identified non-coding isoforms, with the exception of *MXI1*, will require further studies to be performed to better ascertain under what conditions and what stage in the disease the EST/cDNAs were found.

The study also discovered that protein-coding transcripts found under T cell suppression conditions, seem to lack the protein kinase-like domain and suffer from a loss of conserved motifs. It would seem that the presence of the proline-rich regions in both conditions, normal and T cell suppression, is crucial for the functioning of the identified isoforms.

The protein interactions analysis revealed that cooperation of the identified isoforms together with the co-receptors is crucial in T cell suppression. How the T cell suppressed isoforms interact with the co-receptors, has to be studied further. The protein interaction database, STRING, does not work on splicing isoforms but rather reduces the protein isoforms to a single protein per gene by choosing the longest known isoform [82]. This has unfortunately resulted in only protein interactions at the gene level being analysed in this study and not at the isoform level. A possible future study would require an analysis of the impact of alternative splicing on the protein interactions as well as protein scaffolding.

REFERENCES:

[1] Kramer A. The structure and function of proteins involved in mammalian pre-mRNA splicing. Annu Rev Biochem 1996;65(1):367-409

[2] Ward AJ, Cooper TA. The pathobiology of splicing. J Pathol 2010;220(2):152-163.

(3) Black DL. Mechanisms of alternative pre-messenger RNA splicing. Annu Rev Biochem 2003;72(1):291-336.

[4] Wen J, Chiba A, Cai X. Computational identification of tissue-specific alternative splicing elements in mouse genes from RNA-Seq. Nucleic Acids Res 2010;38(22):7895-7907.

[5] Smith CWJ, Valcárcel J. Alternative pre-mRNA splicing: the logic of combinatorial control. Trends Biochem Sci 2000;25(8):381-388.

[6] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature 2001;409(6822):860-921.

[7] Gilbert W. Why genes in pieces? Nature 1978 Feb 9;271(5645):501.

[8] Tazi J, Bakkour N, Stamm S. Alternative splicing and disease. Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease 2009;1792(1):14-26.

[9] Orgel LE, Crick FH. Selfish DNA: the ultimate parasite. Nature 1980;284(5757):604.

[10] Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, et al. Deciphering the splicing code. Nature 2010;465(7294):53-59.

[11] Kriventseva EV, Koch I, Apweiler R, Vingron M, Bork P, Gelfand MS, et al. Increase of functional diversity by alternative splicing. Trends in Genetics 2003;19(3):124-128.

[12] Stetefeld J, Ruegg MA. Structural and functional diversity generated by alternative mRNA splicing. Trends Biochem Sci 2005;30(9):515-521.

[13] Parra G, Bradnam K, Rose AB, Korf I. Comparative and functional analysis of intron-mediated enhancement signals reveals conserved features among plants. Nucleic Acids Res 2011;39(13):5328-5337.

[14] McManus CJ, Graveley BR. RNA structure and the mechanisms of alternative splicing. Curr Opin Genet Dev 2011.

[15] Ferreira EN, Galante PAF, Carraro DM, de Souza SJ. Alternative splicing: a bioinformatics perspective. Molecular BioSystems 2007;3(7):473-477.

[16] Gardina P, Clark T, Shimada B, Staples M, Yang Q, Veitch J, et al. Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. BMC Genomics 2006;7(1):325.

[17] Modrek B, Resch A, Grasso C, Lee C. Genome-wide detection of alternative splicing in expressed sequences of human genes. Nucleic Acids Res 2001;29(13):2850-2859.

[18] Mironov AA, Fickett JW, Gelfand MS. Frequent alternative splicing of human genes. Genome Res 1999;9(12):1288-1293.

[19] Brett D, Hanke J, Lehmann G, Haase S, Delbrück S, Krueger S, et al. EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. FEBS Lett 2000;474(1):83-86.

[20] Kirsch IR, Green ED, Yonescu R, Strausberg R, Carter N, Bentley D, et al. ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. Nat Genet 2000;24:340-341.

[21] Kan Z, Rouchka EC, Gish WR. Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. Genome Res 2001;11(5):889-900.

[22] Modrek B, Lee C. A genomic view of alternative splicing. Nat Genet 2002;30(1):13-19.

[23] Nagaraj SH, Gasser RB, Ranganathan S. A hitchhiker's guide to expressed sequence tag (EST) analysis. Briefings in Bioinformatics 2007;8(1):6-21.

[24] Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, et al. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science 2003;302(5653):2141-2144.

[25] Black DL. Protein Diversity from Alternative Minireview Splicing: A Challenge for Bioinformatics and Post-Genome Biology. Cell 2000;103:367-370.

[26] Chen P, Lepikhova T, Hu Y, Monni O, Hautaniemi S. Comprehensive exon array data processing method for quantitative analysis of alternative spliced variants. Nucleic Acids Res 2011;39(18):e123-e123.

[27] Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 2008;320(5881):1344-1349.

[28] Jana S, Campbell H, Woodliff J, Waukau J, Jailwala P, Ghorai J, et al. The type of responder T cell has a significant impact in a human in vitro suppression assay. PloS one 2010;5(12):e15154.

[29] Cobbold SP. T cell tolerance in transplantation: possibilities for therapeutic intervention. Expert opinion on therapeutic targets 2002;6(5):583-599.

[30] Vogt L, Schmitz N, Kurrer MO, Bauer M, Hinton HI, Behnke S, et al. VSIG4, a B7 family-related protein, is a negative regulator of T cell activation. J Clin Invest 2006;116(10):2817.

[31] Xu S, Sun Z, Li L, Liu J, He J, Song D, et al. Induction of T cells suppression by dendritic cells transfected with VSIG4 recombinant adenovirus. Immunol Lett 2010;128(1):46-50.

[32] Selenko-Gebauer N, Majdic O, Szekeres A, Höfler G, Guthann E, Korthäuer U, et al. B7-H1 (programmed death-1 ligand) on dendritic cells is involved in the induction and maintenance of T cell anergy. The Journal of Immunology 2003;170(7):3637.

[33] Rodríguez-García M, Porichis F, de Jong OG, Levi K, Diefenbach TJ, Lifson JD, et al. Expression of PD-L1 and PD-L2 on human macrophages is up-regulated by HIV-1 and differentially modulated by IL-10. J Leukoc Biol 2011;89(4):507-515.

[34] Engelking O, Fedorov LM, Lilischkis R, ter Meulen V, Schneider-Schaulies S. Measles virus-

induced immunosuppression in vitro is associated with deregulation of G1 cell cycle control proteins. J Gen Virol 1999;80(7):1599.

[35] Avota E, Gassert E, Schneider-Schaulies S. Measles virus-induced immunosuppression: from effectors to mechanisms. Med Microbiol Immunol (Berl) 2010;199(3):227-237.

[36] Hilleman MR. Current overview of the pathogenesis and prophylaxis of measles with focus on practical implications. Vaccine 2001;20(5):651-665.

[37] Krakowka S, Higgins R, Koestner A. Canine distemper virus: review of structural and functional modulations in lymphoid tissues. Am J Vet Res 1980;41(2):284-292.

[38] McCl-IESNEY MB, OLDSTONE MBA. Virus-induced immunosuppression: infections with measles virus and human immunodeficiency virus. Adv Immunol 1989;45:335.

[39] Moss WJ. Measles review article, Moss & Griffin.

[40] Castaneda CA, Cortes-Funes H, Gomez HL., and Ciruelos EM. The phosphatidyl inositol 3-kinase/AKT signaling pathway in breast cancer. Cancer Metastasis Rev 2010; 29(4):751-759.

[41] Paez J, and Sellers W. PI3K/PTEN/Akt Pathway. Signal transduction in cancer 2004:145-167.

[42] Cavalieri S, Funaro A, Porcedda P, Turinetto V, Migone N, Gatti RA, et al. ATM mutations in Italian families with ataxia telangiectasia include two distinct large genomic deletions. Human Mutation 2006;27(10):1061-1061.

[43] Stredrick DL, Garcia-Closas M, Pineda MA, Bhatti P, Alexander BH, Doody MM, et al. The ATM missense mutation p. Ser49Cys (c. 146C> G) and the risk of breast cancer. Human Mutation 2006;27(6):538-544.

[44] Maillet P, Bonnefoi H, Vaudan-Vutskits G, Pajk B, Cufer T, Foulkes W, et al. Constitutional alterations of the ATM gene in early onset sporadic breast cancer. J Med Genet 2002; 39(10):751-753.

[45] Buchholz TA, Weil MM, Ashorn CL, Strom EA, Sigurdson A, Bondy M, et al. A Ser49Cys variant in the ataxia telangiectasia, mutated, gene that is more common in patients with breast carcinoma compared with population controls. Cancer 2004;100(7):1345-1351.

[46] Guttmacher AE, Collins FS, Guttmacher AE, Collins FS. Genomic medicine—a primer. N Engl J Med 2002; 347(19):1512-1520.

[47] Feero WG, Guttmacher AE, Feero WG, Guttmacher AE, Collins FS. Genomic medicine—an updated primer. N Engl J Med 2010; 362(21):2001-2011.

[48] Faustino NA, Cooper TA. Pre-mRNA splicing and human disease. Genes Dev 2003;17(4):419-437.

[49] Wang H, Hubbell E, Hu J, Mei G, Cline M, Lu G, et al. Gene structure-based splice variant deconvolution using a microarry platform. Bioinformatics 2003; 19(suppl 1):i315-i322.

[50] Avota E, Harms H, Schneider-Schaulies S. Measles virus induces expression of *SIP110*, a constitutively membrane clustered lipid phosphatase, which inhibits T cell proliferation. Cell Microbiol 2006;8(11):1826–1839

[51] Yamasaki C, Murakami K, Fujii Y, Sato Y, Harada E, Takeda J, et al. The H-Invitational Database (H-InvDB), a comprehensive annotation resource for human genes and transcripts. Nucleic Acids Res 2008;36(Database issue):D793. URL: http://h-invitational.jp/h-dbas/as_mechanism.jsp

[52] http://www.ncbi.nlm.nih.gov/dbEST

[53] Affymetrix technical note. Identifying and validating alternative splicing events: an introduction to managing data provided by GeneChip Exon Arrays. URL: http://media.affymetrix.com/support/technical/technotes/id_altsplicingevents_technote.pdf.

[54] Affymetrix technical note. Whole transcript expression analysis. URL: http://media.affymetrix.com/support/technical/technotes/wt_appnote.pdf

[55] von Pirquet C. Verhalten der kutanen tuberkulin-reaktion wahrend der Masern. Dtsch. Med. Wochenschr. 1908;34:1297–1300

[56] http://www.who.int/en/

[57] http://www.ncbi.nlm.nih.gov/geo/geo2r/

[58] Conde L, Vaquerizas JM, Santoyo J, Al-Shahrour F, Ruiz-Llorente S, Robledo M, et al. PupaSNP Finder: a web tool for finding SNPs with putative effect at transcriptional level. Nucleic Acids Res 2004;32(suppl 2):W242-W248. URL: http://pupasuite.bioinfo.cipf.es/

[59] Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A, Kersey P, Flicek P. Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database (Oxford)*. Vol. 2011. URL: http://www.ensembl.org/biomart/martview

[60] Huang, D.W., Sherman, B.T. and Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. (2009) Nat Protoc. 4(1):44 -57. URL: http://david.abcc.ncifcrf.gov/home.jsp

[61] Obayashi, T and Kinoshita, K. COXPRESdb: a database to compare gene coexpression is seven model animals. Nucleic Acids Research 2011;39:1016-1022

[62] Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen LJ and von Mering C. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Research 2011; 39:561-568. URL: http://string-db.org

[63] Riedel, A, Mofolo, B, Avota, E, Schneider-Schaulies, S, Meintjes, A, Mulder, N and Kneitz, S. Accumulation of splice variants and transcripts in response to PI3K inhibition in T cells: potential role of their gene products in cell silencing (2012) (Accepted for publication in PLoS One)

[64] http://www.openoffice.org

[65] Zdobnov E.M. and Apweiler R. InterProScan - an integration platform for the signature-recognition methods in InterPro. Bioinformatics, 2001, 17(9): 847-8. doi:10.1093/bioinformatics/17.9.847. URL: http://www.ebi.ac.uk/Tools/pfa/iprscan/

[66] Anupam R, Datta A, Kesic M, Green-Church K, Shkriabai N, et al. (2011) Human T-lymphotropic virus type 1 p30 interacts with REGgamma and modulates ATM (ataxia telangiectasia mutated) to promote cell survival. J Biol Chem 286: 7661-7668.

[67] Tordjman R, Lepelletier Y, Lemarchandel V, Cambot M, Gaulard P, et al. (2002) A neuronal receptor, neuropilin-1, is essential for the initiation of the primary immune response. Nat Immunol 3: 477-482.

[68] Tran-Van H, Avota E, Bortlein C, Mueller N, Schneider-Schaulies S (2011) Measles virus modulates dendritic cell/T cell communication at the level of plexinA1/neuropilin-1 recruitment and activity. Eur J Immunol 41: 151-163.

[69] Stelzer, G, Dalah, I, Iny Stein, T, Satanower, Y, Rosen, N, Nativ, N, Oz-Levi, D, Olender, T, Belinky, F, Bahir, I, Krug, H, Perco, P, Mayer, B, Kolker, E, Safran, M and Lancet, D. In-silico Human Genomics with GeneCards, Human Genomics, 2011 Oct;5(6):709-17. URL: http://www.genecards.org/

[70] http://www.ncbi.nlm.nih.gov/entrez

[71] Barrett, T, Troup, D.B, Wilhite, S.E, Ledoux, P, Evangelista, C, Kim, I.F, Tomashevsky, M, Marshall, K.A, Phillippy, K.H, Sherman, P.M, Muertter, R.N, Holko, M, Ayanbule, O, Yefanov, A, Soboleva, A. NCBI GEO: archive for functional genomics data sets—10 years on. Nucleic Acids Res. 2011; 39:1005–1010. URL: http://www.ncbi.nlm.nih.gov/geo

[72] Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, et al. ArrayExpress—a public database of microarray experiments and gene expression profiles. Nucleic Acids Res 2007;35(suppl 1):D747-D750. URL: http://www.ebi.ac.uk/arrayexpress/

[73] Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19: 185-193.

[74] Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res 2010;38(suppl 1):D355-D360. URL: http://www.genome.jp/kegg/tool/map_pathway2.html

[75] Srinivasan , M and Dunker, A.K. Proline-rich motifs as drug targets in immune mediated disorders. Int J Pept. 2012: 634769.

[76] Gargani M, Valentini A, Pariset L. A novel point mutation within the EDA gene causes an exon dropping in mature RNA in Holstein Friesian cattle breed affected by X-linked anhidrotic ectodermal dysplasia. BMC veterinary research 2011;7(1):35.

[77] Boffa MB, Maret D, Hamill JD, Bastajian N, Crainich P, Jenny NS, et al. Effect of single nucleotide polymorphisms on expression of the gene encoding thrombin-activatable fibrinolysis inhibitor: a functional analysis. Blood 2008;111(1):183-189.

[78] Ellis LM. The role of neuropilins in cancer. Molecular cancer therapeutics 2006;5(5):1099.

[79] Tordjman R, Lepelletier Y, Lemarchandel V, Cambot M, Gaulard P, Hermine O, et al. A neuronal receptor, neuropilin-1, is essential for the initiation of the primary immune response. Nat Immunol 2002;3(5):477-482.

[80] Trobridge PA, Forbush KA, Levin SD. Positive and negative selection of thymocytes depends on LCK interaction with the CD4 and CD8 co-receptors. The Journal of Immunology 2001; 166(2):809-818.

[81] Jarvis A, Allerston CK, Jia H, Herzog B, Garza-Garcia A, Winfield N, et al. Small molecule

inhibitors of the neuropilin-1 vascular endothelial growth factor A (VEGF-A) interaction. J Med Chem 2010;53(5):2215-2226.

[82] Von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Krüger B, et al. STRING 7—recent developments in the integration and prediction of protein interactions. Nucleic Acids Res 2007;35(suppl 1):D358-D362.

[83] Ng SSW, Tsao MS, Nicklee T, and Hedley DW. Wortmannin inhibits pkb/akt phosphorylation and promotes gemcitabine antitumor activity in orthotopic human pancreatic cancer xenografts in immunodeficient mice. Clinical Cancer Research 2001; *7:* 3269-3275

[84] Brinkman, BMN. Review: Splice variants as cancer biomarkers. Clinical Biochemistry 2004; 37: 584-594

[85] Stastna, M, and Van Eyk, JE. Review: Analysis of protein isoforms: can we do it better? Proteomics 2012; 12: 2937-2948

[86] Lotze MT, Wang E, Marincola FM, Hanna N, Bugelski PJ, et al. Workshop on Cancer Biometrics: Identifying Biomarkers and Surrogates of Cancer in Patients. Journal of Immunotherapy 2005; 28:79-119

[87] Schlessinger J. SH2/SH3 signaling proteins. Current Opinion in Genetics & Development 1994; 4(1): 25-30

[88] Faber K, Glatting K, Mueller PJ, Risch A and Hotz-Wagenblatt A. Genome-wide prediction of splice-modifying SNPs in human genes using a new analysis pipeline called AASsites. BMC Bioinformatics 2011; 12(Suppl 4):S2

APPENDIX A: Genes predicted to be involved in T cell suppression

ABCB1	ANTXR1	BDP1	CCL3L3	CHD8	CXCL13	E2F1	FAP
ABCC1	ANXA1	BEST1	CCL4	CHEK1	CXCL3	E2F3	FAS
ABCC2	ANXA11	BFSP1	CCL5	CHEK2	CXCL5	E2F4	FASLG
ABCC8	ANXA3	BGLAP	CCL7	CHKA	CXCR3	E2F5	FASN
ABCD3	ANXA5	BID	CCL8	CHN1	CXCR4	E2F6	FBLN1
ABCG1	ANXA7	BIN1	CCNA1	CHRD	CXCR6	E4F1	FBLN5
ABCG2	APAF1	BIRC2	CCNA2	CHRM2	CYB5A	EAF2	FBXO32
ABHD5	APBA2	BIRC3	CCNB1	CHRM3	CYBA	EBAG9	FBXW7
ABI1	APBA3	BIRC7	CCND1	CHUK	CYBB	ECE1	FBXW8
ABL1	APC	BLM	CCND2	CIITA / MHC2TA	CYCS	EDN1	FCER2
AC012652.1	APEX1	BLNK	CCND3	CISH	CYP11A1	EDN3	FCGR1A
ACAA1	API5	BLOC1S2	CCNE1	CKB	CYP11B1	EDNRA	FCGR2A
ACACA	APLN	BLVRB	CCNH	CKM	CYP11B2	EDNRB	FCGR2B
ACAT1	APLNR	BMI1	CCNT1	CLCN1	CYP19A1	EEF2K	FCGR3A
ACAT2	APOA1	BMP2	CCR1	CLCN3	CYP1A1	EFNA1	FDFT1
ACD	APOA2	BMP4	CCR3	CLDN10	CYP1A2	EGF	FDPS
ACE	APOB	BMPR1A	CCR4	CLDN2	CYP1B1	EGFR	FDXR
ACHE	APOBEC3F	BMPR2	CCR6	CLEC4A	CYP21A2	EGLN1	FES
ACLY	APOBEC3G	BNIP1	CCR8	CLIC4	CYP24A1	EGLN2	FGA
ACO1	APOC3	BNIPL	CCR9	CLU	CYP2A6	EGLN3	FGF1
ACP5	APOE	BPI	CCT6A	CMA1	CYP2B6	EGR1	FGF10
ACPP	APPL1	BRAF	CD163	CNBP	CYP2C19	EGR2	FGF19
ACSL1	APRT	BRCA1	CD1A	CNN1	CYP2E1	EHHADH	FGF23
ACTA1	AQP1	BRCA2	CD2	CNP	CYP2J2	EHMT2	FGF4
ACTG2	AQP4	BRD4	CD200	CNR1	CYP3A4	EI24	FGF5
ACTN1	AQP7	BRD7	CD200R1	CNR2	CYP3A5	EIF2AK1	FGF7
ACTN2	AQP8	BRMS1	CD209	CNTF	CYP3A7	EIF2AK2	FGF8
ACTN3	AR	BTG3	CD244	CNTN1	CYP4A11	EIF2AK3	FGFR1
ACTN4	AREG	BTK	CD247	CNTN2	DAB1	EIF2C2	FGFR2
ACTR2	ARF6	BTRC	CD274	COL15A1	DAB2	EIF3A	FGFR3
ACVR1B	ARG1	C11ORF58	CD28	COL18A1	DACH1	EIF4A1	FGFR4
ACVR2A	ARG2	C19ORF2	CD33	COL1A1	DAO	EIF4E	FGG
ACVRL1	ARHGAP24	C1QBP	CD34	COL2A1	DAPK1	EIF4EBP1	FGR
ADA	ARHGAP5	C2	CD36	COL4A1	DARC	EIF4G1	FHIT
ADAM15	ARHGDIA	C3AR1	CD38	COL4A2	DAXX	EIF5A	FHL1
ADAM17	ARHGDIB	C4BPB	CD3D	COL4A3	DBN1	EIF6	FIBP
ADAMTS4	ARHGEF2	C6ORF25	CD4	COL4A3BP	DCN	ELAC2	FKBP1A
ADAMTS5	ARID4A	CA2	CD40	COL4A4	DCT	ELANE	FKBP1B
ADAR	ARMC10	CA3	CD40LG	COL4A6	DCTN1	ELAVL1	FKBP4
ADC	ARNT	CA6	CD44	COPS8	DCX	ELF1	FLI1
ADCY1	ARNTL	CABIN1	CD46	COX4I2	DCXR	ELF3	FLNA
ADCY10	ARRB1	CABP1	CD55	COX5A	DDB2	ELF4	FLT1
ADCY2	ARRB2	CACNA1B	CD58	COX8A	DDC	ELK1	FLT3
ADCYAP1	ARSH	CACNA1I	CD59	CPOX	DDIT3	ELN	FLT3LG
ADCYAP1R1	ASAH1	CACNA2D2	CD68	CPS1	DDOST	EMB	FLT4
ADD1	ASPH	CACYBP	CD7	CPSF4	DDR1	ENG	FMN1
ADIPOQ	ATF1	CAD	CD74	CPT1A	DDX3X	ENO1	FMR1
ADIPOR1	ATF2	CADM1	CD79A	CR1	DDX5	ENO2	FOLH1
ADK	ATF3	CALCA	CD79B	CRABP2	DDX58	ENO3	FOLR1
ADM	ATF4	CALCR	CD80	CRAT	DEFB1	ENPP1	FOLR2

ADORA1	ATF6	CALCRL	CD82	CRB1	DES	ENPP2	FOS
ADORA2A	ATM	CALD1	CD83	CREB1	DFFB	ENTPD1	FOSB
ADORA2B	ATN1	CALM1	CD86	CREB3	DFNA5	EP300	FOSL1
ADORA3	ATP1A1	CALR	CD8A	CREBBP	DGKB	EPAS1	FOXK1
ADRBK1	ATP1A2	CALU	CD8B	CREM	DGKD	EPB41L1	FOXM1
AEBP1	ATP1B2	CAMK2A	CD9	CRH	DGKQ	EPB41L2	FOXN3
AES	ATP1B3	CAMK2D	CDC14A	CRHR1	DGUOK	EPB41L3	FOXO1
AFP	ATP2A2	CAMK2G	CDC25A	CRHR2	DHODH	EPB49	FOXO3
AGAP2	ATP2A3	CAMK2N2	CDC25C	CRK	DICER1	EPCAM	FOXO4
AGER	ATP2B1	CAMK4	CDC27	CRKL	DIO2	EPHA1	FOXP3
AGPAT2	ATP2C1	CAMKK2	CDC42	CRP	DKK1	EPHA2	FPR1
AGRN	ATP5A1	CAMP	CDC6	CRY1	DLC1	EPHA3	FRS2
AGTR1	ATP6V0A2	CANT1	CDC7	CRYAB	DLG4	EPHB2	FSCN1
AGTRAP	ATP6V1A	CANX	CDH1	CS	DLK1	EPHB6	FSHB
AHCY	ATP7A	CAPN1	CDH11	CSDA	DLL1	EPHX2	FSHR
AHR	ATP7B	CAPN10	CDH2	CSF1	DMBT1	EPM2A	FST
AHSG	ATR	CAPN2	CDH5	CSF1R	DMD	EPO	FTL
AICDA	ATRN	CAPRIN2	CDK2	CSF2	DMP1	EPOR	FURIN
AIF1	ATRX	CAPZA1	CDK4	CSF2RA	DMPK	EPRS	FUT3
AIFM1	ATXN1	CARD8	CDK5	CSF3	DMTF1	EPS15	FUT8
AK1	ATXN3	CASP1	CDK5R1	CSF3R	DNAJA1	ERBB2	FYN
AKAP12	AURKA	CASP10	CDK5R2	CSH1	DNAJA3	ERBB2IP	G6PC
AKR1A1	AURKB	CASP2	CDK6	CSK	DNAJB1	ERBB3	G6PD
AKR1C1	AXIN1	CASP3	CDK7	CSN2	DNAJB6	ERBB4	GAB1
AKR1C3	AXL	CASP4	CDK9	CSNK1A1	DNASE1	ERCC1	GAB2
AKT1	B2M	CASP6	CDKN1A	CSNK2A1	DNM3	ERCC5	GABBR1
AKT2	B3GALTL	CASP7	CDKN1B	CSNK2A2	DNMT1	ERP29	GABPB1
AKT3	B3GAT1	CASP8	CDKN1C	CSNK2B	DNMT3A	ESPN	GABRG2
ALCAM	BACE1	CASP9	CDKN2A	CTBP1	DNMT3B	ESR1	GAD1
ALDH2	BACE2	CASR	CDKN2B	CTBP2	DNTT	ESR2	GAD2
ALDH7A1	BACH1	CAST	CDKN3	CTCF	DOCK4	ESRRB	GADD45A
ALDH9A1	BAD	CAT	CEACAM1	CTLA4	DOK1	ETS1	GADD45B
ALK	BAIAP2L1	CAV1	CEACAM5	CTNNA1	DPP4	ETS2	GADD45G
ALOX12	BAK1	cav2	CEBPA	CTNNB1	DPYD	ETV6	GALC
ALOX15	BARD1	CBFB	CEBPB	CTNND1	DRD2	ETV7	GALE
ALOX15B	BAX	CBL	CEBPD	CTSB	DSC1	EWSR1	GALT
ALOX5	BBS9	CBS	CES1	CTSC	DSG1	EXO1	GANAB
ALPL	BCAR1	CBX2	CES2	CTSD	DSG2	EZH2	GAP43
ALPP	BCCIP	CBX5	CETP	CTSE	DSG3	EZR	GAPDH
ALS2	BCHE	CCDC50	CFD	CTSG	DSP	F10	GAS6
AMACR	BCKDHB	CCK	CFDP1	CTSL1	DUSP1	F2	GAST
AMBP	BCL2	CCKBR	CFH	CTTN	DUSP13	F2R	GATA1
AMFR	BCL2A1	CCL1	CFL1	CUL7	DUSP4	F2RL1	GC
AMHR2	BCL2L1	CCL18	CFLAR	CUX1	DUSP5	F8	GCG
AMOT	BCL2L11	CCL2	CFTR	CX3CL1	DUSP6	FABP3	GCH1
AMPD3	BCL3	CCL23	CGA	CX3CR1	DUT	FABP4	GCK
ANAPC7	BCLAF1	CCL25	CGB5	CXADR	DYNC1H1	FADD	GDF15
ANGPT2	BCR	CCL27	CHAF1A	CXCL1	DYNC2H1	FAM49B	GDNF
ANK1	BDKRB2	CCL3	CHAT	CXCL11	DYNLL1	FANCA	GFAP
ANPEP	BDNF	CCL3L1	CHD7	CXCL12	DYRK1A	FANCD2	GFM1

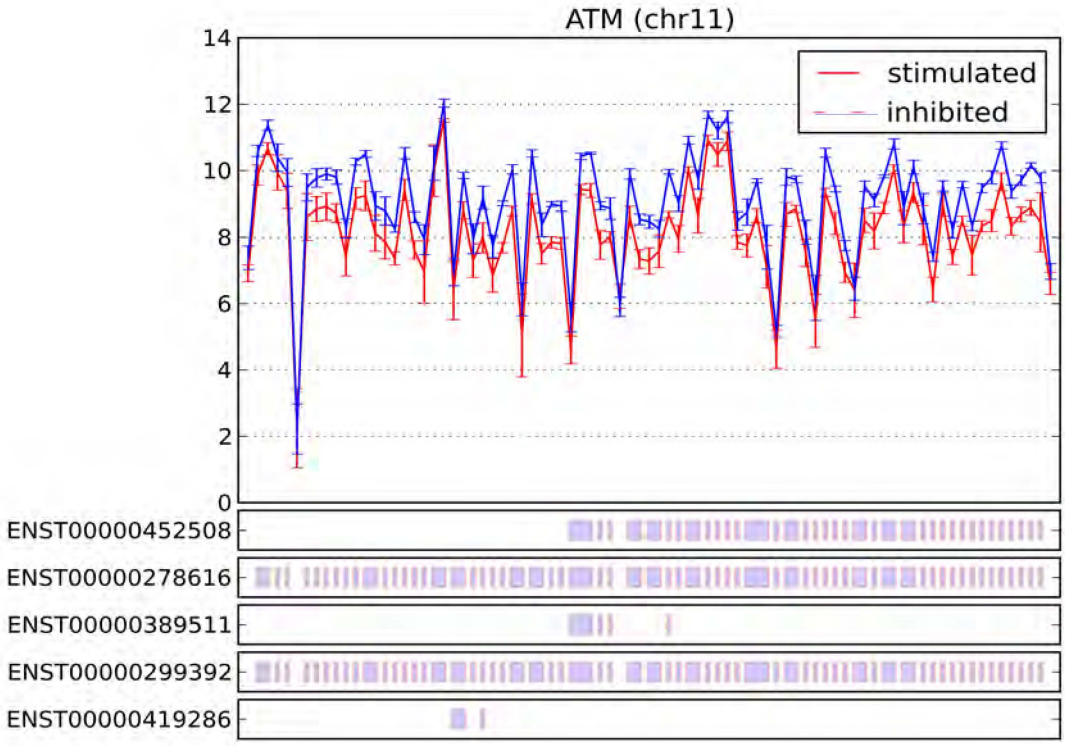
GFPT1	HDAC6	IGFBP5	KCNJ2	LYN	MSR1	NPHS1	PDE4A
GFRA2	HDAC9	IGFBP6	KCNN2	LYST	MSRA	NPM1	PDE4B
GGCX	HDGF	IKBKB	KCNN3	M6PR	MST1R	NPPA	PDE5A
GGT1	HEMGN	IKBKG	KCNN4	MADD	MSTN	NPPB	PDE7A
GH1	HES6	IKZF1	KCNQ1	MAEA	MT1A	NPPC	PDGFA
GHR	HFE2	IKZF2	KCNQ2	MAG	MT2A	NPR1	PDGFB
GHRH	HGF	IL10	KCNQ4	MAGED1	MT3	NPR3	PDGFC
GHRHR	HGFAC	IL11	KCNQ5	MALT1	MTA1	NPTN	PDGFD
GHRL	HGS	IL12A	KDR	MAN2B1	MTDH	NPY	PDGFRA
GJA1	HHIP	IL12RB1	KEAP1	MAOA	MTF1	NPY5R	PDGFRB
GJA5	HIC1	IL12RB2	KHDRBS1	MAP1B	MTHFR	NQO1	PDIA3
GJB2	HIF1A	IL13	KHK	MAP1LC3A	MTMR14	NR0B1	PDK1
GK	HIF3A	IL13RA1	KIAA0101	MAP1S	MTMR3	NR0B2	PDK2
GLI1	HIPK2	IL13RA2	KIF1B	MAP2	MTNR1A	NR1H2	PDK4
GLI2	HK1	IL15	KIF5B	MAP2K1	MUC1	NR1H3	PDLIM2
GLI3	HK2	IL15RA	KISS1	MAP2K2	MUC15	NR1H4	PDLIM3
GLO1	HLA-B	IL17A	KISS1R	MAP2K3	MUC20	NR1I2	PDLIM4
GLRX	HLA-G	IL17F	KIT	MAP2K4	MUC4	NR1I3	PDLIM5
GLRX2	HLCS	IL17RA	KITLG	MAP2K6	MUSK	NR2C1	PDLIM7
GLS	HMBS	IL18	KL	MAP2K7	MUTYH	NR2C2	PDPK1
GLUL	HMGA1	IL19	KLF10	MAP3K1	MVP	NR2E3	PDPN
GNA13	HMGA2	IL1A	KLF2	MAP3K14	MXD3	NR3C1	PEMT
GNAI1	HMGB1	IL1B	KLF6	MAP3K4	MXI1	NR3C2	PENK
GNAO1	HMGCR	IL1R1	KLK3	MAP3K5	MYB	NR4A1	PEPD
GNAS	HMMR	IL1RAP	KLK8	MAP3K7	MYBL2	NR5A1	PER1
GNB2	HMOX1	IL1RL1	KLRC1	MAP3K8	MYCN	NR5A2	PER2
GNB2L1	HMOX2	IL1RN	KLRD1	MAP4	MYD88	NRAS	PF4
GNLY	HNF1A	IL2	KNG1	MAP4K1	MYF5	NRF1	PFKFB3
GNRH1	HNF1B	IL21	KRAS	MAP6	MYH10	NRG1	PFN1
GNRHR	HNF4A	IL24	KRT14	MAPK1	MYH11	NRIP1	PGF
GOT2	HNRNPA1	IL25	KRT18	MAPK10	MYH14	NRP1	PGK1
GP1BA	HNRNPA2B1	IL2RA	KRT19	MAPK11	MYH9	NRP2	PGR
GPD1	HNRNPK	IL3	KRT7	MAPK12	MYL2	NSUN5	PHB
GPI	HNRNPR	IL32	KRT8	MAPK14	MYL4	NTF3	PHKA2
GPR56	HOMER1	IL4	L1CAM	MAPK3	MYLK	NTM	PI3
GPRC6A	HOXA10	IL4R	LAG3	MAPK8	MYOC	NTRK1	PIAS2
GPS1	HOXA3	IL5	LALBA	MAPK9	MYOCD	NTRK2	PIAS3
gpt	HOXA9	IL6	LAMA1	MAPKAP1	MYOD1	NTRK3	PIGP
GPT2	HP	IL6R	LAMA3	MAPKAPK2	MYOG	NTS	PIK3C2A
GPX4	HPD	IL6ST	LAMB1	MAPKAPK5	MYOM1	NUB1	PIK3C3
GRB10	HPN	IL7	LAMC2	MAPRE2	NAB2	NUDT1	PIK3CA
GRB2	HPR	IL7R	LAMP2	MAPT	NAGLU	NUDT10	PIK3CB
GRB7	HPSE	IL8	LAP3	MAT1A	NAIP	NUDT6	PIK3CG
GREM1	HPX	ILF3	LBR	MATK	NAT1	NUMB	PIK3IP1
GRIA1	HRAS	ILK	LCK	MATN1	NBL1	OAT	PIK3R1
GRIA2	HRH2	IMPA1	LCN1	MAX	NBN	OCLN	PIK3R2
GRIA3	HRH3	IMPDH1	LCP1	MBD1	NCAM1	OGDH	PIM1
GRIA4	HRH4	ING1	LDHA	MBD2	NCF1	OGG1	PIP5KL1
GRIN1	HSD11B1	ING2	LDHB	MBD3	NCF2	OLR1	PITPNA
GRIN2A	HSD11B2	INHA	LDLR	MBL2	NCK1	ONECUT1	PITX1

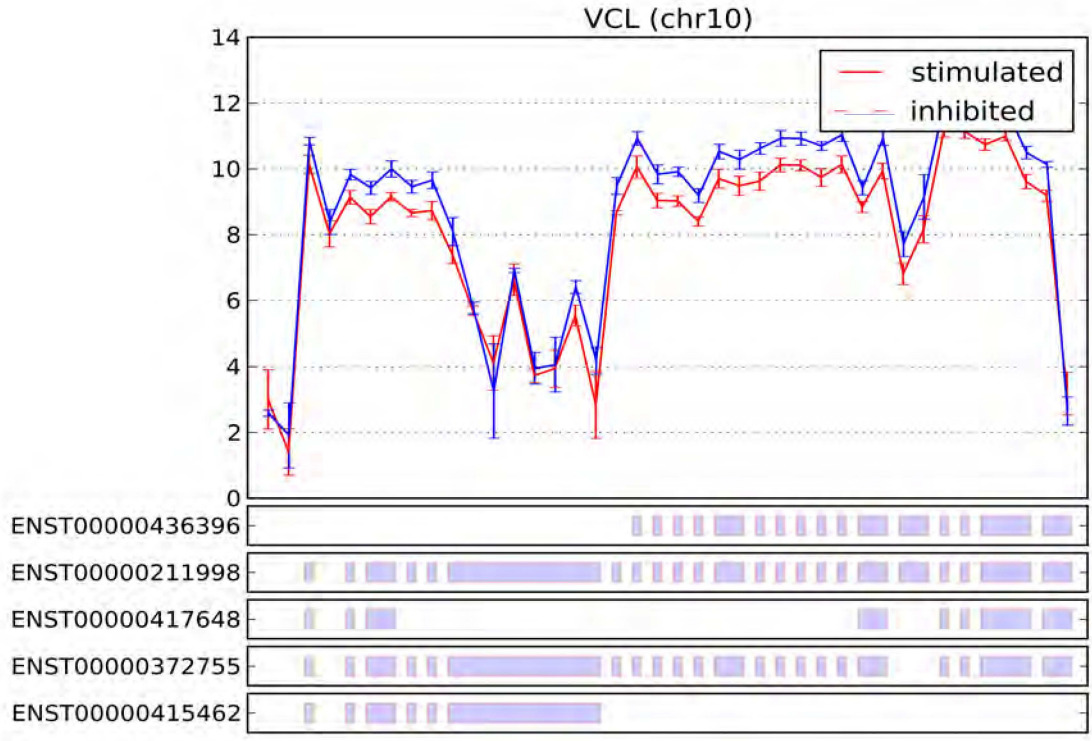
GRIN2B	HSD17B2	INPPL1	LECT1	MC2R	NCL	OPN4	PITX2
GRIP1	HSF1	INS	LEF1	MCC	NCOA1	OPRD1	PKD1
GRK4	HSP90AA1	INSR	LEP	MCF2	NCOA2	OPRM1	PKIB
GRM1	HSP90AB1	INTS6	LEPRE1	MCL1	NCOA3	ORAI1	PKLR
GRM2	HSP90B1	IRAK1	LGALS1	MCM7	NCOA4	ORC1	PKM2
GRM5	HSPA14	IRAK3	LGALS12	MDM2	NCOA6	OSBPL8	PLA2G1B
GRP	HSPA1A	IRF1	LGALS3	MDM4	NCOR1	OSM	PLA2G2A
GSK3B	HSPA4	IRF2	LGALS3BP	MECP2	NCOR2	OTC	PLA2G4A
GSN	HSPA5	IRF3	LGALS4	MED1	NCR3	OTOF	PLA2G5
GSR	HSPA8	IRF7	LGALS9	MED23	NDRG2	OTUD5	PLA2G6
GSTA4	HSPA9	IRS1	LGI1	MEF2A	NEDD4	OTX2	PLA2G7
GSTM1	HSPB1	ITCH	LGMN	MEFV	NEDD4L	OXTR	PLA2R1
GSTP1	HSPE1	ITGA2	LHCGR	MEIS1	NEDD8	P4HTM	PLAGL1
GTF2A1	HSPG2	ITGA2B	LHX3	MEN1	NEK2	PA2G4	PLAT
GTF2B	HTATIP2	ITGA3	LIF	MEST	NES	PABPC4	PLAU
GTF2F1	HTR2A	ITGA4	LIFR	MET	NEU1	PABPN1	PLAUR
GTF2I	HTR2B	ITGA5	LIG4	METAP2	NEUROG3	PACRG	PLCB1
GUCA2A	HTR2C	ITGA6	LILRA3	MGAT3	NF1	PADI1	PLCB3
GUK1	HTR3A	ITGAL	LILRB1	MID1	NF2	PAEP	PLCB4
GUSB	HTR4	ITGAM	LILRB2	MITF	NFATC1	PAFAH1B1	PLCE1
GYS1	HTRA1	ITGAV	LILRB3	MKI67	NFATC2	PAICS	PLCG1
GYS2	HTT	ITGB1	LILRB4	MKKS	NFE2	PAK1	PLCG2
GZMA	HUWE1	ITGB2	LIMK1	MKNK1	NFE2L2	PAK2	PLCL1
GZMB	HYAL1	ITGB3	LIN28B	MLH1	NFIC	PAK4	PLCZ1
H1F0	HYAL2	ITGB3BP	LIPA	MLL	NFKB1	PAM	PLD1
HADHA	HYOU1	ITGB4	LIPC	MME	NFKB2	PAPSS2	PLD2
HADHB	IAPP	ITIH4	LIPE	MMP1	NFKBIA	PARD3	PLEK
HAGH	IBSP	ITM2A	LIPF	MMP11	NGF	PARK2	PLEKHO1
HAMP	ICAM1	ITPA	LIPG	MMP12	NGFR	PARP1	PLG
HAPLN1	ICOS	ITPR1	LMNA	MMP13	NGLY1	PAWR	PLK1
HAS1	ICOSLG	IVL	LMNB1	MMP16	NID1	PAX2	PLK3
HAS2	ID1	JAG1	LONP1	MMP2	NIPBL	PAX3	PLOD2
HAS3	ID2	JAK1	LOX	MMP3	NISCH	PAX5	PLOD3
HBB	ID3	JAK2	LOXL2	MMP7	NKX2-1	PAX6	PLP1
HBEGF	IDH1	JAK3	LPA	MMP9	NKX3-1	PAX7	PLS3
HBG1	IDH2	JMJD6	LPAR1	MOG	NLRC4	PAX8	PLSCR1
HBG2	IDO1	JPH4	LPAR3	MPG	NLRP1	PBK	PLUNC
HCFC1	IDS	JUN	LPL	MPO	NLRP3	PBX1	PMAIP1
HCLS1	IFI16	JUNB	LPO	MPP1	NLRP7	PC	PML
HCN1	IFNAR1	JUND	LPXN	MPZ	NME4	PCBP2	PMS2
HCN2	IFNAR2	JUP	LRP1	MR1	NNAT	PCGF6	PNCK
HCN4	IFNG	KAT2B	LRP8	MRC2	NOD2	PCK2	PNKD
HCRT	IFT81	KAT5	LRPAP1	MRE11A	NODAL	PCMT1	PNLIP
HCRTR1	IGF1	KBTBD10	LRRK2	MSC	NOS1	PCNA	PNPLA8
HCRTR2	IGF1R	KCND1	LSP1	MSH2	NOS2	PCNT	POLB
HDAC1	IGF2BP1	KCND2	LSS	MSH6	NOS3	PCSK1	POMC
HDAC2	IGFALS	KCNE1	LTA	MSI1	NOTCH1	PDCD1LG2	POMT1
HDAC3	IGFBP1	KCNH2	LTB	MSLN	NOTCH3	PDCD4	PON1
HDAC4	IGFBP3	KCNJ1	LTF	MSMB	NOX1	PDE2A	POR
HDAC5	IGFBP4	KCNJ11	LY75	MSN	NOX4	PDE3B	PORCN

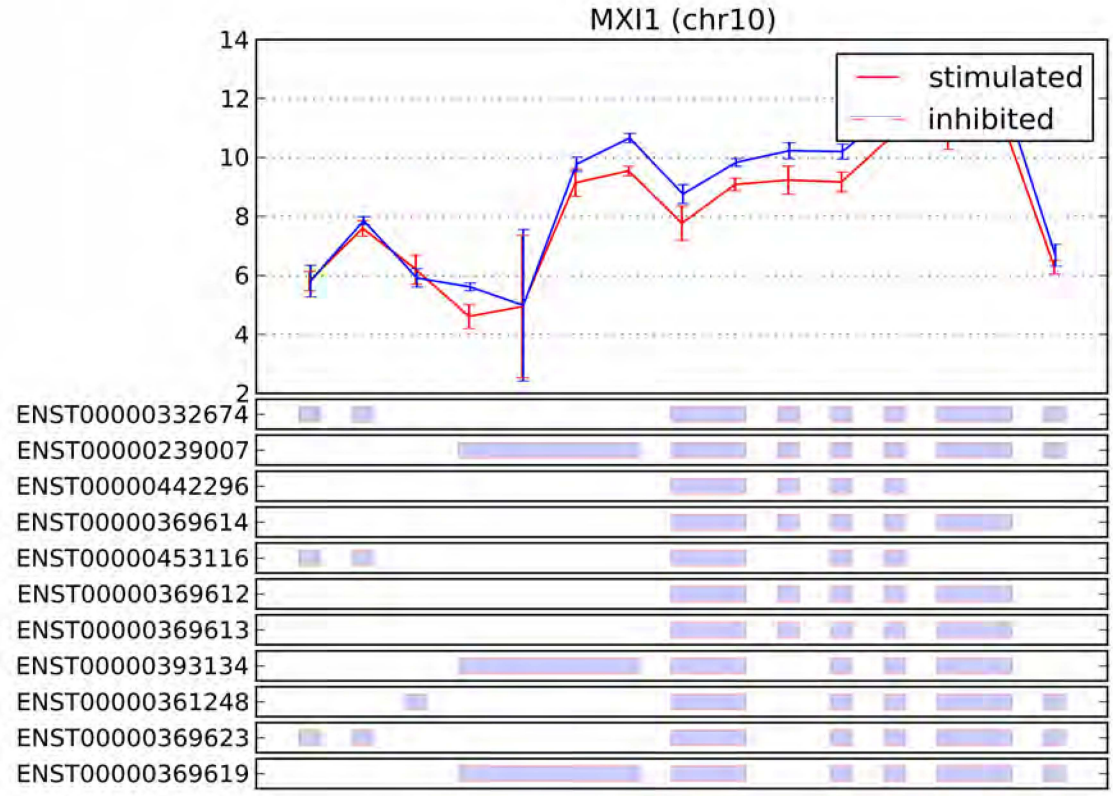
PRKCB	RANBP2	SAFB	SLC6A12	STX2	TNFRSF25	VCAN
PRKCD	RAP1A	SALL1	SLC6A2	STXBP1	TNFRSF4	VCL
PRKCE	RAPGEF3	SAT1	SLC6A3	SUFU	TNFRSF8	VCP
PRKCG	RARA	SCARA3	SLC6A4	SULT1E1	TNFRSF9	VDAC1
PRKCI	RARB	SCARB1	SLC7A1	SULT2A1	TNFSF10	VDR
PRKCZ	RARG	SCARB2	SLC8A1	SUMO1	TNFSF11	VEGFC
PRKD1	RARS	SCARF1	SLC9A1	SUZ12	TNFSF13	VGF
PRKD2	RASA1	SCARF2	SLC9A2	SVIL	TNFSF13B	VHL
PRKD3	RASA2	SCD	SLC9A3	SYK	TNFSF14	VIM
PRKDC	RASA4	SCGB1A1	SLC9A3R1	SYNM	TNFSF15	VIP
PRKG1	RASGRF1	SCN5A	SLCO4A1	SYNPO	TNFSF18	VIPR1
PRKG2	RASSF1	SCN7A	SLPI	SYNPO2	TNNI3	VPS24
PRL	RASSF5	SCNN1A	SMAD2	SYP	TNNT1	VWCE
PRLR	RB1	SCNN1B	SMAD3	SYVN1	TNS1	WAS
PRMT1	RB1CC1	SCP2	SMAD4	T	TOB1	WASF2
PRMT5	RBBP4	SCRIB	SMAD6	TAC1	TOM1	WDR26
PRNP	RBBP6	SCYL1	SMAD7	TAC4	TOP2A	WEE1
PRODH	RBBP8	SDC1	SMARCA2	TACR1	TP53	WHSC1
PROK1	RBL1	SDC2	SMARCA4	TAF1	TP53BP1	WIPF1
PROM1	RBM3	SDCBP	SMARCB1	TAF6	TP53BP2	WNK1
PROS1	RBM5	SDHB	SMARCD3	TAF9	TP53I3	WNT1
PROX1	RBMX	SEC14L2	SMC1A	TANK	TP63	WNT2
PRSS2	RBP1	SELE	SMC3	TAT	TP73	WNT3A
PRSS3	RBP2	SELL	SMG1	TAZ	TPH1	WRN
PSAT1	RBP3	SELP	SMOX	TBP	TPI1	WWC1
PSEN1	RBP4	SEMA3B	SMPD1	TBX21	TPM1	WWOX
PSEN2	RBPJ	SEMA3F	SMPD2	TBX3	TPM3	XAF1
PSMA1	RCAN1	SEPT4	SMTN	TBXA2R	TPO	XBP1
PSMA3	RDX	SEPT9	SNAI1	TBXAS1	TPPP3	XDH
PSMA4	RECQL	SERPINA1	SNAI2	TCEB2	TPSAB1	XIAP
PSMA7	RECQL4	SERPINA3	SNCA	TCF3	TRAF1	XPC
PSMB5	RECQL5	SERPINA6	SNCAIP	TCF4	TRAF2	XRCC4
PSMB7	REL	SERPINA7	SNIP1	TCF7L2	TRAF3	XRCC5
PSMB8	RELB	SERPINB3	SNRNP70	TCL1A	TRAF3IP2	XRCC6
PSMC1	RERE	SERPINB4	SNX6	TCP1	TRAF6	YAP1
PSMC4	RET	SERPINB5	SOAT1	TEAD1	TREH	YARS
PSMC5	REXO2	SERPINC1	SOAT2	TEAD4	TRH	YBX1
PSMC6	RFC2	SERPINE1	SOCS1	TEC	TRIM21	YME1L1
PSMD10	RFC3	SERPINF1	SOCS3	TENC1	TRIM24	YWHAB
PSMD13	RFWD2	SERPINF2	SOD1	TERF1	TRIM63	YWHAE
PSMD4	RFX1	SERPING1	SOD2	TERT	TRPC3	YWHAG
PSMD9	RFXANK	SERPINI2	SOD3	TES	TRPC4	YWHAH
PSME1	RGS2	SET	SORBS3	TF	TRPC6	YWHAQ
PSME3	RGS3	SETDB1	SORT1	TFAM	TRPM1	YWHAZ
PSRC1	RGS4	SFN	SOS1	TFAP2A	TRPM2	YY1
PTBP1	RGS9	SFRP1	SOX5	TFAP2B	TRPM4	ZAP70
PTCH1	RHOA	SFTPA1	SP1	TFAP2C	TRPV4	ZBTB16
PTEN	RHOB	SFTPC	SP100	TFDP1	TSC1	ZEB1
PTGDR	RHOC	SFTPD	SP3	TFR2	TSC2	ZFHX3
PTGDS	RHOD	SGCB	SP7	TFRC	TSC22D3	ZFP36

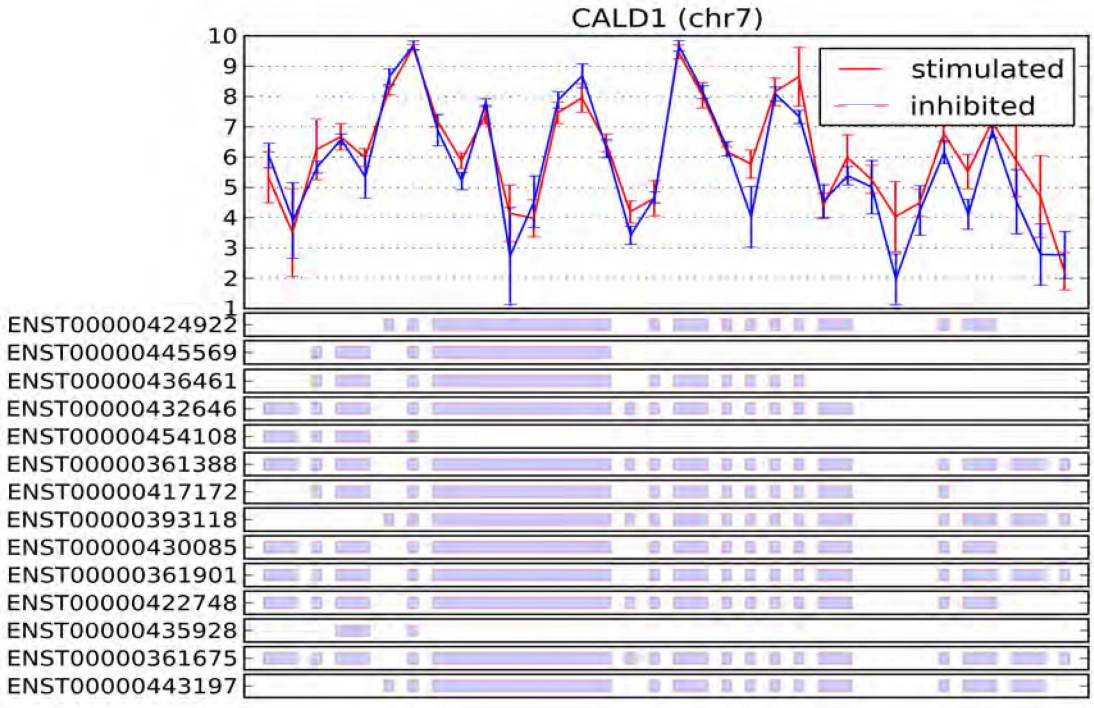
POSTN	PTGER1	RHOH	SGK1	SPAG9	TG	TSG101	ZFPM1
POT1	PTGER2	RIPK1	SGOL1	SPARC	TGFA	TSHB	ZFR
POU1F1	PTGER3	RIPK2	SH2D1A	SPHK1	TGFB1	TSHR	ZFYVE9
POU2AF1	PTGER4	RIPK3	SH3D19	SPI1	TGFB1I1	TSPAN7	ZMYND11
POU2F1	PTGES	RLN1	SH3GLB1	SPINK1	TGFB2	TSPAN8	ZNF148
POU2F2	PTGES2	RNASE1	SHC1	SPINK5	TGFB3	TSPO	ZNF217
POU4F1	PTGFR	RNASEL	SHOX	SPINLW1	TGFBI	TSPYL2	ZNF384
POU5F1	PTGIR	RNF123	SIAH1	SPINT1	TGFBR1	TTC4	
PPA1	PTGIS	RNF128	SIAH2	SPINT2	TGFBR2	TTR	
PPAP2A	PTGS1	RNF135	SIK1	SPP1	TGFBR3	TUBA1B	
PPARA	PTGS2	RNF14	SIN3A	SPTAN1	TGIF1	TUBA4A	
PPARD	PTH	RNF19A	SIRT1	SPTBN1	TGM1	TUBB	
PPARG	PTH1R	RNF34	SIRT2	SPTLC1	TGM2	TUBG1	
PPARGC1A	PTHLH	RNH1	SIT1	SPTLC2	TH	TWIST1	
PPARGC1B	PTK2	ROBO1	SIVA1	SQSTM1	THBS1	TXN	
PPBP	PTK2B	ROCK1	SKI	SRA1	THOC1	TXNRD1	
PPIA	PTK7	ROCK2	SKP2	SRC	THPO	TXNRD2	
PPID	PTP4A3	RORA	SLAMF6	SRD5A1	THRA	TYMS	
PPIG	PTPN1	RPA1	SLC11A2	SREBF1	THRB	TYR	
PPM1A	PTPN11	RPL11	SLC13A2	SREBF2	TIAF1	TYRP1	
PPM1D	PTPN12	RPN2	SLC16A1	SRF	TIMP1	UAP1	
PPP1CA	PTPN2	RPS20	SLC17A5	SRPK1	TIMP3	UBASH3B	
PPP1R13L	PTPN3	RPS24	SLC19A1	SS18	TIRAP	UBC	
PPP1R15B	PTPN6	RPS4X	SLC1A1	SSB	TJP1	UBE2C	
PPP1R2	PTPN7	RPS6KA1	SLC1A2	SST	TJP2	UBE2E1	
PPP1R8	PTPRB	RPS6KA3	SLC1A3	SSTR1	TK1	UBE2K	
PPP1R9B	PTPRC	RPS6KB1	SLC22A1	ST3GAL5	TKT	UBE2N	
PPP2CA	PTPRO	RRAS	SLC22A12	ST5	TLE1	UBE3A	
PPP2R2B	PTPRZ1	RRBP1	SLC22A18	ST7	TLK1	UBQLN1	
PPP2R4	PTS	RRM1	SLC22A2	ST7L	TLN1	UCHL1	
PPP3CA	PUF60	RRM2	SLC22A7	ST8SIA1	TLR1	UCN	
PPP5C	PVR	RSF1	SLC25A13	STAT1	TLR2	UCP1	
PPY	PVRL2	RTN4	SLC25A14	STAT2	TLR3	UCP2	
PQBP1	PXK	RUFY3	SLC25A22	STAT3	TLR4	UCP3	
PRDM1	PXN	RUNX1	SLC25A27	STAT4	TLR9	UHRF1	
PRDM10	PYGB	RUNX1T1	SLC25A4	STAT5A	TMEFF2	UNG	
PRDM2	PYY	RUNX2	SLC25A5	STAT5B	TMPO	UPF3A	
PRDX1	QSOX1	RUNX3	SLC26A4	STAT6	TMPRSS11A	USF1	
PRDX2	RAB11FIP1	RXRA	SLC27A2	STAU1	TMPRSS2	USF2	
PRDX4	RAB1A	RXRB	SLC29A1	STIM1	TNC	USP2	
PRDX5	RABAC1	RXRG	SLC2A1	STK11	TNF	USP4	
PRDX6	RAC1	RYR1	SLC2A2	STK17B	TNFAIP3	USP5	
PRKAA1	RAC2	RYR2	SLC2A4	STK36	TNFRSF10A	USP7	
PRKAA2	RAD1	S100A1	SLC30A9	STMN1	TNFRSF10B	UTRN	
PRKAB1	RAD51	S100A6	SLC35C1	STMN2	TNFRSF11A	UTS2	
PRKACA	RAD52	S100A7	SLC39A4	STS	TNFRSF11B	UXT	
PRKAR1A	RAET1E	S100B	SLC3A2	STT3A	TNFRSF13C	VAV1	
PRKAR2A	RAF1	S1PR1	SLC4A1	STT3B	TNFRSF18	VAV2	
PRKAR2B	RALA	S1PR2	SLC4A2	STX11	TNFRSF1A	VAV3	
PRKCA	RAMP2	SAA4	SLC4A4	STX1A	TNFRSF1B	VCAM1	

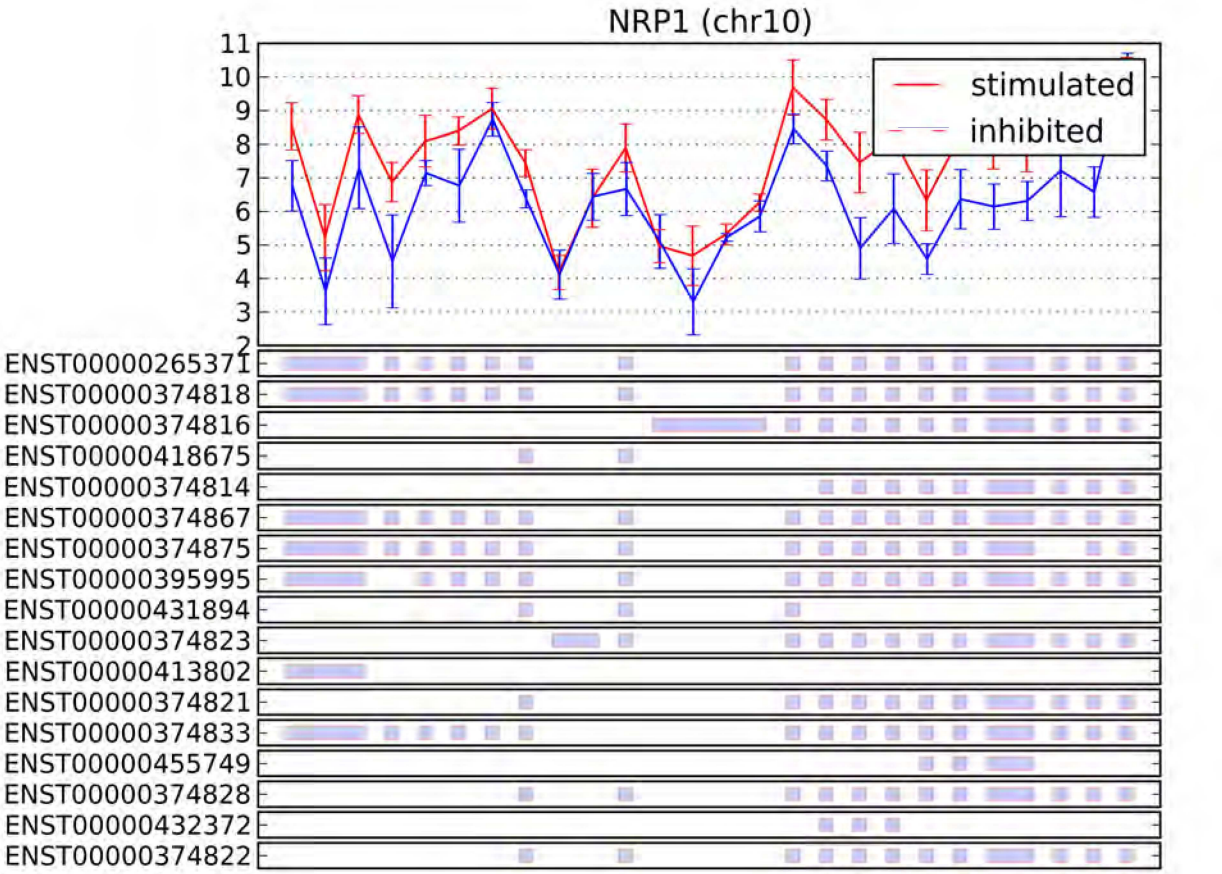
APPENDIX B: GeneChip exon array: probe set intensity plots






MXI1 (chr10)
stimulated
inhibited
14
12
10
8
6
4
2
ENST00000332674
ENST00000239007
ENST00000442296
ENST00000369614
ENST00000453116
ENST00000369612
ENST00000369613
ENST00000393134
ENST00000361248
ENST00000369623
ENST00000369619


CALD1 (chr7)
stimulated
inhibited
10
9
8
7
6
5
4
3
2
1
ENST00000424922
ENST00000445569
ENST00000436461
ENST00000432646
ENST00000454108
ENST00000361388
ENST00000417172
ENST00000393118
ENST00000430085
ENST00000361901
ENST00000422748
ENST00000435928
ENST00000361675
ENST00000443197


NRP1 (chr10)
stimulated
inhibited
11
10
9
8
7
6
5
4
3
2
ENST00000265371
ENST00000374818
ENST00000374816
ENST00000418675
ENST00000374814
ENST00000374867
ENST00000374875
ENST00000395995
ENST00000431894
ENST00000374823
ENST00000413802
ENST00000374821
ENST00000374833
ENST00000455749
ENST00000374828
ENST00000432372
ENST00000374822

APPENDIX C: Ensembl protein sequences

LCK-202

MGYYNGHTKVAVKSLKQGSMSPDAFLAEANLMKQLQHQRLVRLYAVVTQEPIYIITEYMENGSLVDFLKTPSG
IKLTINKLLDMAAQIAEGMAFIEERNYIHRDLRAANILVSDTLSCKIADFGLARLIEDNEYTAREGAKFPIKWTAP
EAINYGTFTIKSDVWSFGILLTEIVTHGRIPYPGMTNPEVIQNLERGYRMVRPDNCPEELYQLMRLCWKERPED
RPTFDYLRSVLEDFFTATEGQYQPQP

LCK-006

MGCGCSSHPEDDWMENIDVCENCHYPIVPLDGKGTLLIRNGSEVRDPLVTYEGSNPPASPLQDNLVIALHSYEP
SHDGDLGFEKGEQLRILEQSGEWWKAQSLTTGQEGFIPFNFVAKANSLEPEPWFFKNLSRKDAERQLLAPGNT
HGSFLIRESESTAGSFSLSVRDFDQNQGEVVKHYKIRNLDNGGFYISPRITFPGLHELVRHYTRYYNGHTKVAVK
SLKQGSMSPDAFLAEANLMKQLQHQRLVRLYAVVTQEPIYIITEYMENGSLVDFLKTPSGIKLTINKLLDMAAQI
AEGMAFIEERNYIHRDLRAANI

LCK-010

MGIPGSHNLRYFWNFPGQGPIISDVGGADLGGAPSAPSSIPSGTMGCGCSSHPEDDWMENIDVCENCHYPIVPLD
GKGTLLIRNGSEVRDPLVTYEGSNPPASPLQDNLVIALHSYEPSHDGDLGFEKGEQLRILEQSGEWWKAQSLTTGQEGFI
PFNFVAKANSLEPEPWFFKNLSRKDAERQLLAPGNTHGSFLIRESESTAGSFSLSVRDFDQNQGEVVKHYKIRNLDNGGF
YISPRITFPGLHELVRHYTNASDGLCTRLSRPCQTQKPQKPWWEDEWEVP

VCL-001

MPVFHTRTIESILEPVAQQISHLVIMHEEGEVDGKAIPDLTAPVAAVQAAVSNLVRVGKETVQTTEDQILKRDMPPA
FIKVENACTKLVQAAQMLQSDPYSVPARDYLIDGSRGILSGTSDLLLTFDEAEVRKIIRVCKGILEYLTVAEVVETMEDL
VTYTKNLGPGMTKMAKMIDERQQELTHQEHRVMLVNSMNTVKELLPVLISAMKIFVTTKNSKNQGIEEALKNRNFTVEKM
SAEINEIIRVLQLTSWDEDAWASKDTEAMKRALASIDSKLNQAKGWLRDPSASPGDAGEQAIRQILDEAGKVGELCAGKE
RREILGTCKMLGQMTDQVADLRARGQGSSPVAMQKAQQVSQGLDVLTAKVENAARKLEAMTNSKQSIAKKIDAAQNWLAD
PNGGPEGEEQIRGALAEARKIAELCDDPKERDDILRSLGEISALTSKLADLRRQGKGDSPEARALAKQVATALQNLQTKT
NRAVANSRPAKAAVHLEGKIEQAQRWIDNPTVDDRGVGQAAIRGLVAEGHRLANVMMGPYRQDLLAKCDRVDQLTAQLAD
LAARGEGESPQARALASQLQDSLKDLKARMQEAMTQEVSDVFSDTTTPIKLLAVAATAPPDAPNREEVFDERAANFENHS
GKLGATAEKAAAVGTANKSTVEGIQASVKTARELTPQVVSAARILLRNPGNQAAYEHFETMKNQWIDNVEKMTGLVDEAI
DTKSLLDASEEAIKKDLDKCKVAMANIQPQMLVAGATSIARRANRILLVAKREVENSEDPKFREAVKAASDELSKTISPM
VMDAKAVAGNISDPGLQKSFLDSGYRILGAVAKVREAFQPQEPDFPPPPPDLEQLRLTDELAPPKPPLPEGEVPPPRPPP
PEEKDEEFPEQKAGEVINQPMMMAARQLHDEARKWSSKGNDIIAAAKRMALLMAEMSRLVRGGSGTKRALIQCAKDIAKA
SDEVTRLAKEVAKQCTDKRIRTNLLQVCERIPTISTQLKILSTVKATMLGRTNISDEESEQATEMLVHNAQNLMQSVKET
VREAEAASIKIRTDAGFTLRWVRKTPWYQ

VCL-202

MLQSDPYSVPARDYLIDGSRGILSGTSDLLLTFDEAEVRKIIRVCKGILEYLTVAEVVETMEDLVTYTKNLGPGMTK
MAKMIDERQQELTHQEHRVMLVNSMNTVKELLPVLISAMKIFVTTKNSKNQGIEEALKNRNFTVEKMSAEINEIIRVLQL
TSWDEDAWASKDTEAMKRALASIDSKLNQAKGWLRDPSASPGDAGEQAIRQILDEAGKVGELCAGKERREILGTCKMLGQ
MTDQVADLRARGQGSSPVAMQKAQQVSQGLDVLTAKVENAARKLEAMTNSKQSIAKKIDAAQNWLADPNGGPEGEEQIRG
ALAEARKIAELCDDPKERDDILRSLGEISALTSKLADLRRQGKGDSPEARALAKQVATALQNLQTKTNRAVANSRPAKAA
VHLEGKIEQAQRWIDNPTVDDRGVGQAAIRGLVAEGHRLANVMMGPYRQDLLAKCDRVDQLTAQLADLAARGEGESPQAR
ALASQLQDSLKDLKARMQEAMTQEVSDVFSDTTTPIKLLAVAATAPPDAPNREEVFDERAANFENHSGKLGATAEKAAAV
GTANKSTVEGIQASVKTARELTPQVVSAARILLRNPGNQAAYEHFETMKNQWIDNVEKMTGLVDEAIDTKSLLDASEEAI
KKDLDKCKVAMANIQPQMLVAGATSIARRANRILLVAKREVENSEDPKFREAVKAASDELSKTISPMVMDAKAVAGNISD
PGLQKSFLDSGYRILGAVAKVREAFQPQEPDFPPPPPDLEQLRLTDELAPPKPPLPEGEVPPPRPPPPEEKDEEFPEQKA
GEVINQPMMMAARQLHDEARKWSSKPGIPAAEVGIGVVAEADAADAAGFPVPPDMEDDYEPELLLMPSNQPVNQPILAAA
QSLHREATKWSSKGNDIIAAAKRMALLMAEMSRLVRGGSVPRTSPRPQMR

VCL-204

MPPAFIKVENACTKLVQAAQMLQSDPYSVPARDYLIDGSRGILSGTSDLLLTFDEAEVRKIIRVCKGILEYLTVA
EVVETMEDLVTYTKNLGPGMTKMAKMIDERQQELTHQEHRVMLVNSMNTVKELLPVLISAMKIFVTTKNSKNQGIEEALK
NRNFTVEKMSAEINEIIRVLQLTSWDEDAWASKDTEAMKRALASIDSKLNQAKGWLRDPSASPGDAGEQAIRQILDEAGK
VGELCAGKERREILGTCKMLGQMTDQVADLRARGQGSSPVAMQKAQQVSQGLDVLTAKVENAARKLEAMTNSKQSIAKKI
DAAQNWLADPNGGPEGEEQIRGALAEARKIAELCDDPKERDDILRSLGEISALTSKLADLRRQGKGDSPEARALAKQVAT
ALQNLQTKTNRAVANSRPAKAAVHLEGKIEQAQRWIDNPTVDDRGVGQAAIRGLVAEGHRLANVMMGPYRQDLLAKCDRV
DQLTAQLADLAARGEGESPQARALASQLQDSLKDLKARMQEAMTQEVSDVFSDTTTPIKLLAVAATAPPDAPNREEVFDE
RAANFENHSGKLGATAEKAAAVGTANKSTVEGIQASVKTARELTPQVVSAARILLRNPGNQAAYEHFETMKNQWIDNVEK
MTGLVDEAIDTKSLLDASEEAIKKDLDKCKVAMANIQPQMLVAGATSIARRANRILLVAKREVENSEDPKFREAVKAASD
ELSKTISPMVMDAKAVAGNISDPGLQKSFLDSGYRILGAVAKVREAFQPQEPDFPPPPPDLEQLRLTDELAPPKPPLPEG
EVPPPRPPPPEEKDEEFPEQKAGEVINQPMMMAARQLHDEARKWSSKGNDIIAAAKRMALLMAEMSRLVRGGSGTKRALI
QCAKDIAKASDEVTRLAKEVAKQCTDKRIRTNLLQVCERIPTISTQLKILSTVKATMLGRTNISDEESEQATEMLVHNAQ
NLMQSVKETVRELKLLQSKFEQMLDLHCAGLERLPGTSRHLAEPGWHRNLY

NRP1-001

MERGLPLLCAVLALVLAPAGAFRNDKCGDTIKIESPGYLTSPGYPHSYHPSEKCEWLIQAPDPYQRIMINFNPHFD
LEDRDCKYDYVEVFDGENENGHFRGKFCGKIAPPPVVSSGPFLFIKFVSDYETHGAGFSIRYEIFKRGPECSQNYTTPSG
VIKSPGFPEKYPNSLECTYIVFVPKMSEIILEFESFDLEPDSNPPGGMFCRYDRLEIWDGFPDVGPHIGRYCGQKTPGRI
RSSSGILSMVFYTDSAIAKEGFSANYSVLQSSVSEDFKCMEALGMESGEIHSDQITASSQYSTNWSAERSRLNYPENGWT
PGEDSYREWIQVDLGLLRFVTAVGTQGAISKETKKKYYVKTYKIDVSSNGEDWITIKEGNKPVLFQGNTNPTDVVVAVFP
KPLITRFVRIKPATWETGISMRFEVYGCKITDYPCSGMLGMVSGLISDSQITSSNQGDRNWMPENIRLVTSRSGWALPPA
PHSYINEWLQIDLGEEKIVRGIIIQGGKHRENKVFMRKFKIGYSNNGSDWKMIMDDSKRKAKSFEGNNNYDTPELRTFPA
LSTRFIRIYPERATHGGLGLRMELLGCEVEAPTAGPTTPNGNLVDECDDDQANCHSGTGDDFQLTGGTTVLATEKPTVID
STIQSEFPTYGFNCEFGWGSHKTFCHWEHDNHVQLKWSVLTSKTGPIQDHTGDGNFIYSQADENQKGKVARLVSPVVYSQ
NSAHCMTFWYHMSGSHVGTLRVKLRYQKPEEYDQLVWMAIGHQGDHWKEGRVLLHKSLKLYQVIFEGEIGKGNLGGIAVD
DISINNHISQEDCAKPADLDKKNPEIKIDETGSTPGYEGEGEGDKNISRKPGNVLKTLDPILITIIAMSALGVLLGAVCG
VVLYCACWHNGMSERNLSALENYNFELVDGVKLKKDKLNTQSTYSEA

NRP1-201

MDDSKRKAKSFEGNNNYDTPELRTFPALSTRFIRIYPERATHGGLGLRMELLGCEVEAPTAGPTTPNGNLVDECDD
DQANCHSGTGDDFQLTGGTTVLATEKPTVIDSTIQSGSRFFKHHHKQQSMRPQNLHHNSILLL

PRMT5-006

XIRPETHSPGMFSWFPILFPIKQPITVREGQTICVRFWRCSNSKKGSSHQSMKTSGQGVRN

PRMT5-017

SQLEVQFIITGTNHHSEKEFCSYLQYLEYLSQNRPPPNAYELFAKGYEDYLQSPLQPLMDNLESQTYEVFEKDPIK
YSQYQQAIYKCLLDRVPEEEKDTNVQVLMVLGAGRGPLVNASLRAAKQADRRIKLLENWQFE

PRMT5-020

RRNSEAAMLQELNFGAYLGLPAFLLPLNQEDNTNLARVLTNHIHTGHHSSMFWMRVPLVAPEDLRDDIIENAPTTHTEEY
SGEEKTWMWWHNFRTLCDYTLEIGADLPSNHVIDRWLGEPIKAAILPTSIFLTNKKGFPVLSKMHQRLIFRLLKLEVQFI
ITGTNHHSEKEFCSYLQYLEYLSQNRPPPNAYELFAKGYEDYLQSPLQPLMDNLESQTYEVFEKDPIKYSQYQQAIYKCL
LDRVPEEEKDTNVQVLMVLGAGRGPLVNASLRAAKQADR

PRMT5-018

MHQRLIFRLLKLEVQFIITGTNHHSEKEFCSYLQYLEYLSQNRPPPNAYELFAKGYEDYLQSPLQPLMDNLESQTYEVFE
KDPIKYSQYQQAIYKCLLDRVPEE

PRMT5-014

MAAMAVGGAGGSRVSSGRDLNCVPEIADTLGAVAKQGFDFLCMPVFHPRFKREFIQEPAKNRPGPQTRSDLLLSGRDWNT
LIVGKLSPWIRPDSKVEKIRRNSEALEVQFIITGTNHHSEKEFCSYLQYLEYLSQNRPPPNAYELFAKGYEDYLQSPLQP
LMDNLESQTYEVFEKDPIKYSQY