Investigation of leukocyte transcriptomes using serial analysis of gene expression

Lawrence Hene

Thesis submitted for the degree of Doctor of Philosophy

Green College

University of Oxford

Trinity Term 2005

Abstract

Investigation of leukocyte transcriptomes using serial analysis of gene expression

Two tag-based gene expression technologies (serial analysis of gene expression, SAGE, and massively parallel signature sequencing, MPSS) were used to profile a variety of lymphocyte transcriptomes. By combining these libraries with publicly available genome and transcriptome data, both immunological and general aspects of gene expression could be considered. Unexpectedly, analysis of the expression of currently known cell surface components and the proteins corresponding to "immune specific" tags in a cytotoxic T-cell (CTL) library suggested that the current knowledge of the immune specific composition of the cell surface components of a resting CTL is largely complete. An analysis of the "immune specific" tags in a natural killer cell library revealed that a small number of tags could not be matched to any previously sequenced transcripts, suggesting the presence of functionally important but previously uncharacterised transcripts or exons in these cells. To examine the entire transcriptome large libraries are required, implying that MPSS would be the most appropriate technology. A comparison of libraries produced by the two tag-based technologies to characterise $CD4^+$ T-cells revealed a relatively poor correlation, suggesting bias in the two techniques. Investigations into this bias led to the conclusion that despite its great depth, the random sampling events involved in the production of a library limit the breadth of MPSS sampling, making it inappropriate for characterising entire transcriptomes. Finally, with the availability of large LongSAGE libraries it is now possible to examine weakly expressed transcript classes. A panel of LongSAGE libraries was used to conduct the first large scale quantitative study of the expression of *cis*-natural antisense transcripts (*cis*-NATs). *cis*-NATs were found to be expressed at approximately one tenth of the level of sense transcripts and across the panel of libraries *cis*-NATs were found for approximately two thirds of all observed sense transcripts. This suggests antisense transcription is more widespread than previously thought.

Acknowledgements

Whilst studying for this thesis many people have given me support, encouragement and, in a few cases, insight. I owe a great deal of thanks to my friends, family and colleagues for this; some of whom I would like to highlight below. Obviously, first and foremost I must thank my supervisor, Simon Davis. Simon was kind enough to leave me to it when things were going well and supportive enough to come up with ideas when they were not. He was prepared to put in time and effort whenever it was required; this has been hugely appreciated.

Thanks must go to all the members of the Davis Group, who provided me with help and buckets of encouragement. On the experimental side, I must thank Raquel, Mai and Lisa for producing and sequencing the SAGE libraries, and Jan for the GLGI and RACE experiments. On the theoretical side, this thesis would not have been possible with the support and ideas of Ed.

Final thanks must go to all those who proof-read some of this thesis: Ed, John, Nick, Noj, Pete and Sara.

This work has been financially supported by the Medical Research Council and the Wellcome Trust.

Abbreviations

AC	Audic Claverie statistical test
APC	Antigen presenting cell
CCP	complement control protein
CD	cluster of differentiation antigens
cDNA	complementary DNA
c-SMAC	central supramolecular activation cluster
CTL	cytotoxic T lymphocyte
DNA	deoxyribonucleic acid
dsRNA	double stranded RNA
EGF	epidermal growth factor
FACS	fluorescence activated cell sorter
GLGI	generation of longer cDNA fragments from serial analysis of gene expression tags for gene identification
GO	gene ontology
HLDA	human leucocyte differentiation antigen
IgSF	immunoglobulin superfamily
LDLR	low-density lipoprotein receptor
LRR	leucine rich repeats
mAb	monoclonal antibody
MHC	major histocompatibility complex
miRNA	microRNA
MPSS	massively parallel signature sequencing
mRNA	messenger RNA
NAT	natural antisense transcript
NCBI	national center for biotechnology information
NK	natural killer
PBMC	peripheral blood mononuclear cell
pMHC	peptide MHC
p-SMAC	peripheral supramolecular activation cluster
RACE	rapid amplification of cDNA ends
ref	reference
RISC	RNA-induced silencing complex
RNA	ribonucleic acid
rRNA	ribosomal RNA
SAGE	serial analysis of gene expression
ScavengerRCR	scavenger receptor cysteine rich
siRNA	small interfering RNA
TCR	T-cell receptor
TMH	transmembrane helix
TNFRSF	tumour necrosis factor receptor superfamily
TNFSF	tumour necrosis factor superfamily
tpm	tags per million
tRNA	transfer RNA
UTBS	Unique Transcripts Both Sites
UTLST	Unique Transcripts LongSAGE Tags
UTR	untranslated region

Table of Contents

Abstract....i

Acknowledgements....ii

Abbreviations.... iii

Table of Contents....iv

I. Introduction....1

A. Immune Cells....1

B. T-cell activation....3

1. Key cell surface components....3

2. Models of activation....14

3. Synapse formation....18

C. Molecular discovery....21

1. Molecule by molecule discovery....21

2. Global technologies....24

D. Thesis Overview....46

II. Materials and Methods....49

A. Bench-based....49

1. Cellular samples....49

2. SAGE library production....52

3. MPSS library production....52

4. Extension of “no match” SAGE tags....53

B. In silico analysis....53

1. SAGE....53

2. LongSAGE/MPSS....59

III. $CD8^+$ T-cell surface....67

A. Introduction 67
B. Results and Discussion 69
1. Immune cell surface 69
2. Non-immune cell surface 98
IV. Natural Killer cell transcriptome 103
A. Introduction 103
B. Results and Discussion 104
1. Cell surface molecules 105
2. “Immune specific” molecules 109
V. $CD4^+$ T-cell transcriptome 133
A. Introduction 133
1. Unique transcript identification 133
2. Depth of sampling 134
3. Breadth of sampling 136
B. Results and Discussion 137
1. Technique comparison 137
2. Depth of sampling 138
3. Breadth 145
4. Most abundant tags 157
5. Identification of potential novel transcriptional loci 158
6. Future Directions 160
VI. Antisense transcription 161
A. Introduction 161
B. Results and Discussion 163
1. Characteristics of *cis*-NAT expression 163

2. Positional effects 168
3. Novel transcriptional loci 173
4. Sensitivity 174
VII. General Discussion 182
VIII. Appendices 190
Appendix A 190
Appendix B 209
Appendix C.1 248
Appendix C.2 279
Appendix D 294
Appendix E.1 304
Appendix E.2 319
Appendix F.1 327
Appendix F.2 333
IX. Bibliography 337

I. Introduction

The initial aim of the work presented in this thesis was to characterise the cell surface of a resting cytotoxic lymphocyte (CTL). In this introduction, I intend to review the roles of the key cell surface components involved in T-cell activation. Then I will discuss the methods that have been used to identify the cell surface molecules found on the CTL. Finally, I will discuss the appropriate global analyses that could be undertaken to answer the initial aim and the technical issues behind these techniques.

A. Immune Cells

Figure 1.1 shows the different lineages of leukocytes. These cells perform a wide range of functions in both the innate and adaptive immune responses. The main focus of this thesis is on T-cells, and they are classified in a variety of ways, for example by function: memory, effector or naïve T-cells; or by the presence of cell surface markers, i.e. $CD4^+$ and $CD8^+$ T-cells. The common factor is that they all express either the αβ or γδ T cell receptor (TCR). T-cells have two major roles in immune defence; they either control the response of the other immune cells (regulatory and helper T-cells) or they are involved in initiating cell death (killer T-cells). For example, $CD4^+$ T-cells can trigger B-cells to produce antibodies and they can also activate $CD8^+$ T-cells. $CD8^+$ T-cells, once activated, initiate cell death for cells that are presenting the correct antigen, for example this may be cells that are virally infected. Natural killer cells (NK cells) are closely related to T-cells, but form part of the innate immune system. NK cells also have a killer function and may kill virus infected cells, by releasing perforins and inducing apoptosis.

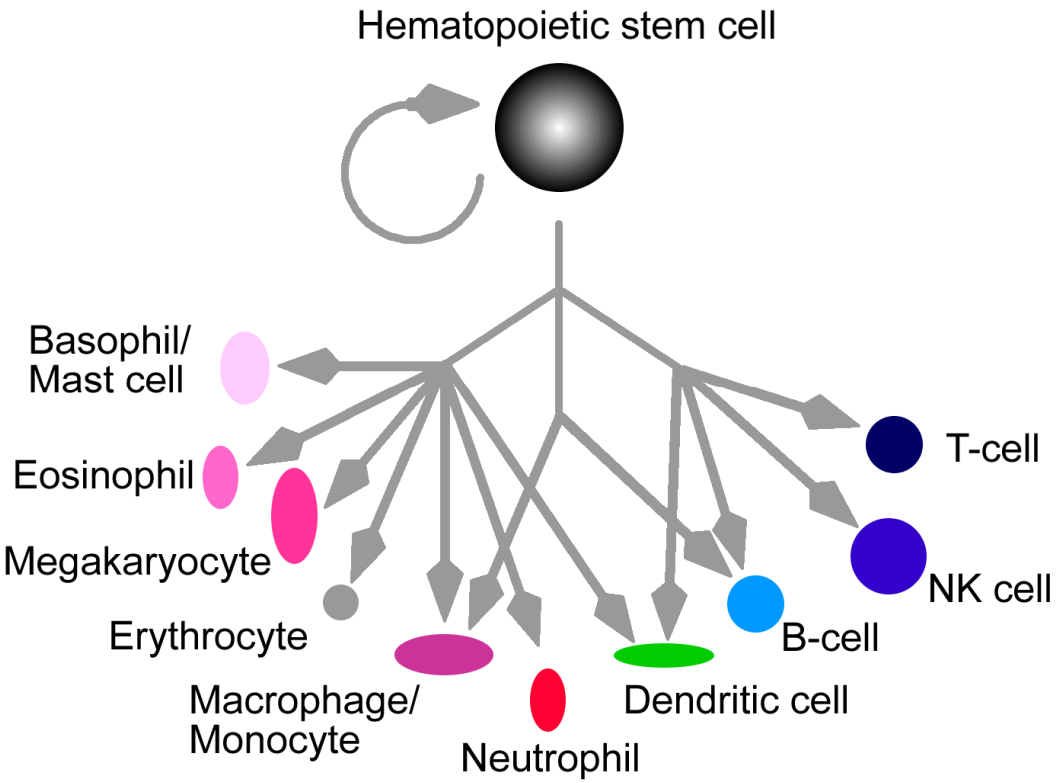


Figure 1.1: Leukocyte lineages.

Leukocytes are derived from hematopoietic stem cells. The nodes represent intermediates, but not all intermediates are shown. The curved arrow represents the option for self renewal of hematopoietic stem cells. Cells coloured blue are lymphoid cells and those coloured pink/red are myeloid cells. This figure is based on Figure 1A by Rothenberg *et al.*[1].

The activation of these killer cells has to be tightly regulated, otherwise inappropriate cells may be killed or diseased cells may not be killed. This has been well studied in T-cells and whilst there are various theories about the exact method of activation it is thought to be relatively well understood, whereas for the NK cells it is still unclear exactly how they are activated. The process of activation differs in the different classes of T-cell but the basic principle is the same. Specific foreign antigen is presented by an antigen presenting cell (APC), and if this antigen is recognised by the receptors on the T-cell, this leads to initial T-cell activation events, followed by immunological synapse formation and subsequent rounds of intracellular signalling events, leading to full-scale T-cell activation. To understand the mechanism of how CTLs are activated it is important to identify all the key components on the cell surface and how they interact. These molecules are responsible for recognising the external signals and translating them into internal signals.

Below I will describe the currently identified key components for T-cell activation. I will then briefly describe how they are thought to interact during both the initial triggering events and in the immunological synapse.

B. T-cell activation

1. Key cell surface components

The molecules listed below are all believed to be expressed on T-cells and are thought to have a key role in the immune response[2, 3]. Their functions, as related to T-cell activation and key structural information, are described. These molecules appear to have at least one of three functional roles: molecular recognition, adhesion or signalling.

The nomenclature of cell surface molecules found on immunological cells is defined by the Human Leukocyte Differentiation Antigens (HLDA) workshops. As described in further detail below, many cell surface proteins on immunological cells were initially identified using antibodies. It was often the case that different antibodies would bind to the same molecule. This lead to a molecule having multiple names but this was unclear to the field[4]. The HLDA workshops undertook to classify these antibodies. Antibodies that were found to bind to the same molecule were grouped into clusters of differentiation (CD). The corresponding antigen (i.e. the cell surface molecule) was given the name of the cluster, e.g. CD124 refers to the IL-4 receptor. There are now over 300 CD molecules[5].

The T-cell receptor complex consists of two classes of molecule; four CD3 proteins (γ, δ, ε and ζ) and two T-cell receptor components (in the large majority of cells, α and β, but in some cells, γ and δ). The exact stoichiometry is unknown but the complex should be considered as one functional molecule[2]. The TCR components are responsible for recognition of the specific peptide antigen and consist primarily of an extracellular domain with a transmembrane domain and a very small, presumably non-functional cytoplasmic tail[6, 7]. The CD3 components are involved in intracellular signalling and consist of a fairly small, or no extracellular domain, a transmembrane domain and a large intracellular domain containing tyrosine phosphorylation motifs[8].

CD2 binds to CD48 and CD58 (found on the APCs) and is thought to enhance T-cell antigen recognition[9]. Whether this enhancement is due to signalling by CD2 or merely due to the increased adhesion between T-cells and APCs is unclear.

CD4 is a coreceptor for T-cell antigen recognition that shows weak affinity for MHC class II molecules. Blockade of CD4 using monoclonal antibodies (mAbs) has been shown to inhibit T-cell functions both *in vivo* and *in vitro*[10]. It is thought to enhance triggering by delivering Lck to the TCR complex[11].

CD5 is believed to downregulate the immune response involved in antigen recognition[12]. It is thought to do this by affecting downstream signalling events, possibly via the pseudo-immunoreceptor tyrosine-based activation motif in its cytoplasmic domain, rather than interfering with the immunological synapse. It is thought to play a similar role in thymocyte development, since T-cells from mice lacking CD5 show an increased response to TCR-mediated stimulation[13].

The exact role for CD6 is unknown. However, it contains potential cytoplasmic binding sites for signalling domains and various studies have suggested a role in T-cell activation[14]. A recent study showed that CD6 binds to the TCR/CD3 complex and is involved in synapse maturation[15]. Inhibition of binding to its normal binding partner (CD166) was found to inhibit the antigen specific T-cell response[16].

CD7 has been shown to be a costimulatory molecule inducing intracellular signalling[17], however, its exact role is unknown. It has recently been shown that this signalling is through a pathway involving a type II phosphatidylinositol 4-kinase[18]. It

has been found to be expressed at different levels in the different functional classes of $CD8^+$ T-cells[19].

CD8 is found as a dimer, either as a CD8α homodimer or a CD8αβ heterodimer. This dimer binds to MHC class I molecules and acts as a coreceptor for antigen recognition. The cytoplasmic domain has been shown to bind the tyrosine kinase Lck, in effect delivering it to the TCR complex and potentiating the activity of the $CD8^+$ T-cell[20].

CD11a and CD18 are found as a heterodimer (LFA-1). LFA-1 is thought to have two roles. Firstly, it acts as an adhesion molecule leading to intercellular adhesion by binding to its ligands (intercellular adhesion molecules: CD50, CD54 and CD102)[21], which are found on a wide variety of cell types. The avidity for its ligands is tightly controlled to prevent inappropriate adhesion[22]. Secondly, it has also been shown to function in costimulatory signalling[23].

CD43 is a large molecule of approximately 45 nm in length and it has been suggested that it functions as an anti-adhesion molecule, inhibiting T-cell interactions[24]. However, studies using antibodies have shown that signalling can occur via CD43 and that it acts as a costimulatory molecule, but the functional relevance of this is unclear[25, 26]. It may well be that CD43 simply forms part of the glycocalyx of the leukocyte.

CD45 is another large cell surface molecule and belongs to the protein tyrosine phosphatase family. There are multiple isoforms, which vary in the large

extracellular domain due to alternative splicing and glycosylation. All isoforms share a common intracellular phosphatase domain[27]. CD45 has been shown to be functionally important in T-cells, with experiments showing that its expression is required for TCR signalling[28]. However, it has been shown to have both an activating and inhibitory role in T-cell activation[29, 30]. The physiological relevance of some of these experiments may be questioned, as CD45 appears to be excluded from the immunological synapse[31, 32].

CD53 is a member of the tetraspanin family (these proteins contain 4 transmembrane domains). CD53 has been shown to non-covalently associate with CD2 in rat T-cells[33]. Its function is unclear but antibody studies suggest a signalling role. Binding of mAbs to CD53 has been shown to deliver a costimulatory signal for the CD3 mediated activation of Jurkats (a T-cell line)[34].

One of the key aspects of the known leukocyte cell surface proteins and GPI-anchored molecules, such as those described above, is that the majority (~80%) have characteristic, modular extracellular protein domains belonging to 19 different superfamilies (CCP, Cytokine receptor, EGF, Fibronectin type II, Fibronectin type III, Galectin, IgSF, Integrin, Letin C-type, LRR, Link, LDLR, Ly-6, MHC, ScavengerRCR, Somatomedin B, TM4SF, TNFSF and TNFRSF)[2]. A further 3% of leukocyte antigens lack these domains but are characterised by the presence of either four (e.g., Fc receptor subunits) or seven transmembrane domains.

This modular architecture is of key importance if one is trying to identify novel cell surface proteins. Obviously, if one can examine the proteins *in situ* then one can

identify the cellular location of a novel protein. However, if one only has sequence information (either nucleotide or amino acid sequences) then one has to predict the location of the protein. To identify cell surface transmembrane proteins one needs firstly to predict the transmembrane helix (TMH) and secondly in which cell membrane it is found. Predicting transmembrane domains is not trivial, as they are characterised by short (~20 amino acids) hydrophobic regions and there are tens of different algorithms publicly available[35]. Generally, most methods identify the majority of TMH (though the simple hydrophobicity based methods falsely predict TMH in globular proteins). However, most methods also incorrectly predict signal peptides to be TMH[36]. This suggests that the majority of transmembrane proteins will be correctly identified as such, but that there will also be an overestimate in the prediction of proteins with a TMH. The next challenge is to identify which membrane these proteins are found in. As stated above, the large majority of cell surface transmembrane proteins contain domains from a limited number of protein superfamilies. These domains are predominantly found at the cell surface (though there are notable exceptions, such as the muscle structural protein titin, which contains both immunoglobulin superfamily domains and fibronectin type 3 domains[37]). Therefore, if one identifies a TMH and one of these domains in a protein sequence then one can be relatively confident that this protein is found at the cell surface.

All these domains belong to superfamilies that contain regions that are relatively well conserved at the sequence level. This conservation at the sequence level means it is possible to identify them by either alignment based approaches or more complex hidden Markov model (HMM) methods such as those used by SMART[38] or Pfam[39].

Hidden Markov models are widely used in domain prediction as they provide a probabilistic framework for profile methods (reviewed by Eddy[40, 41]). HMMs are composed of a number of states and the transition between states is defined by a transition probability[42] (Figure 1.2). Each of these states emits a symbol (e.g. an amino acid residue) according to a characteristic emission probability. It is easiest to explain if one imagines the HMM generating a sequence. Starting from an initial state, a sequence is generated from the HMM by emitting a symbol and moving from one state to another state (possibly itself) until an end state is reached. The probability of the sequence is given by the product of the probabilities of the selected states emitting the symbol observed and the probabilities of each observed state transition. The key point about HMMs is that the probability of moving from one state is defined only by the current state and not those previously observed.

In practise an HMM is not used to generate a sequence but to score a given sequence, (for example to predict domain structure on a protein sequence). The sequence is scanned and each emission is placed in one of the states. The highest scoring state path is chosen, leading to a predicted annotation for the sequence. Computationally it is inefficient to score every possible state sequence and therefore dynamic programming (Virterbi algorithm) is used to avoid scoring sub-optimal solutions.

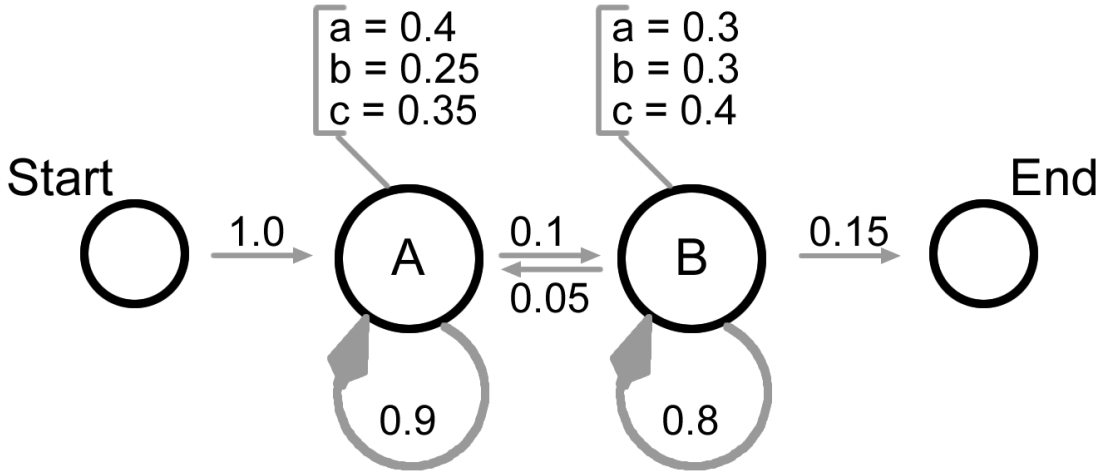


Figure 1.2: Example HMM.

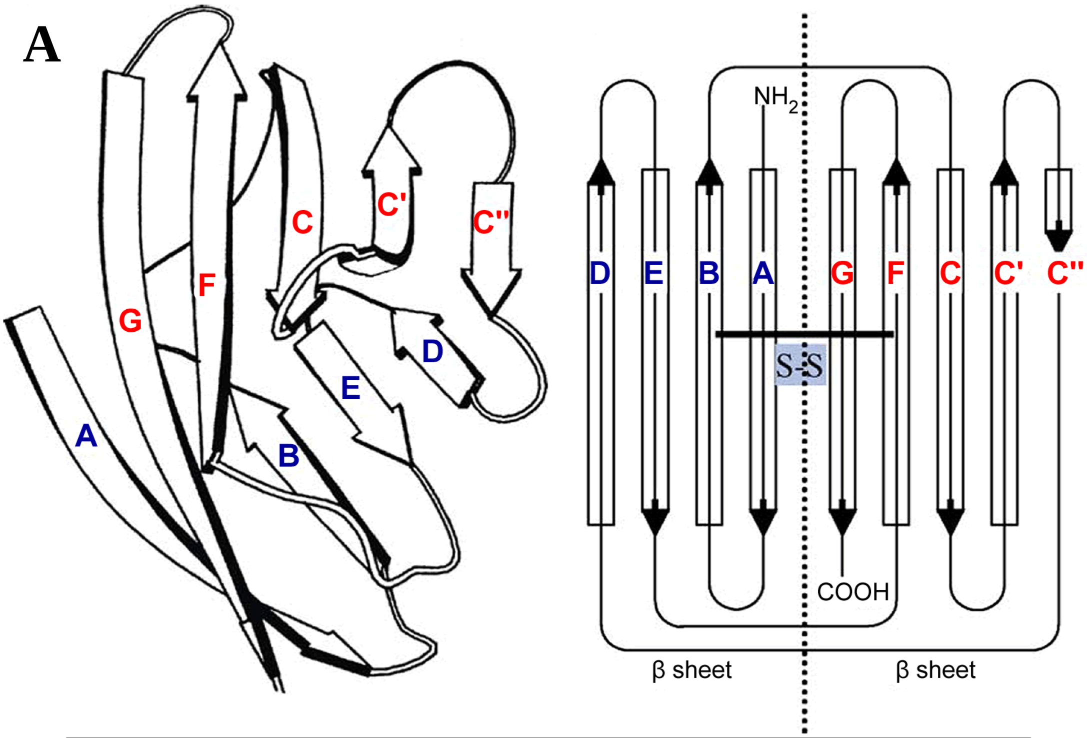
The HMM shown above emits a, b and c at different probabilities (the emission probability) depending upon which state the HMM is in (A or B). The transition probabilities are the probabilities of the HMM changing state. Adapted from Figure 1 by S.R. Eddy[42].

An example of one of these characteristic conserved extracellular domains is the immunoglobulin superfamily (IgSF) domain. It is found in more than a third of known leukocyte cell surface proteins[2] and is predicted by InterPro to be the thirteenth most common domain in the Ensembl Genes dataset (when ranked by number of genes the domain is detected in)[43]. It consists of about 100 amino acids organised in two beta sheets (Figure 1.3 A)[44]. The number and relative length of the strands in the beta sheets depends on the class of fold, and there are four classes: V, C1, C2 and I-set[45]. There are three key residues, which were identified initially by studies of the structure of immunoglobulins and are conserved in the majority of IgSF domains[46]. These residues are two cysteines, which form a disulfide bridge and are located on strands B and F, and a tryptophan found on strand C. These three residues together form a structural motif (the "pin"). The alignment in Figure 1.3 B shows the conserved residues in the V-set domain and these conserved residues can be used to identify other members of the superfamily. Though the fact that IgSF domains contain relatively few highly conserved residues mean that it is likely that current methods under predict the presence of IgSF domains[44]. However, work in my current laboratory suggests that if one uses HMMs specific to sub-families of the IgSF superfamily then it is possible to identify previously undetected IgSF domains[47, 48]. The key point is that if one identifies novel proteins or transcripts when characterising T-cells, then it is possible to predict, with a reasonable level of confidence, whether they are (or encode for) cell surface proteins.

Figure 1.3: Immunoglobulin superfamily domain.

Figure A shows the structure of the Ig fold. The left panel shows the Ig fold structure, with each beta strand represented by an arrow, whereas the right hand panel more clearly shows the topology. Figure B shows an alignment of the V-set domain; the conserved residues are boxed. The approximate positions of strands shown in Figure A are marked. The three key conserved residues that form the pin and are described in the text are boxed in red. Figures A and B are adapted from figures by A.N. Barclay[44].

A



B

```
                A                      B                    C
Ig κ     QMTQSPSSLSASVGDRVTITCQASQD----ISIFLNWYQQKPG-
TCR β    GVIQSPRHEVTEMGQEVTLRCKPISGH-----NSLFWYRQTMM-
CD8 α    QLQLSPKKVDAEIGQEVKLTCEVLRDTS----QGCSWLFRNSSS
CD4 d1   AATQGKKVVLGKKGDTVELTCTASQKK----SIQFHWKNSN---
Thy-1    RGQRVISLTACLVNQNLRLDCRHENNTNLPIQHEFSLTRE----
CD2 d1   ---RDSGTVWGALGHGINLNIPNFQMTD--DIDEVRWERGS---
CD80     -----VIHVTKEVKEVATLSCGHNVSVEELAQTRIYWQKEKK--
CD152    MHVAQPAVVLASSRGIASFVCEY SPGKA-TEVRVTVLRQADSQ

               C'           C"          D           E
Ig κ     KAPKLLIYDA---SKLEAGV---PSRFSGTGS----GTDFTFTI
TCR β    RGLELLIYFN--NNVPIDDSGMPEDRFSAKMPNA---SFSTLKI
CD8 α    ELLQPTFIIYVSSSRSKLNDILDPNLFSARKE----NNKYILTL
CD4 d1   -QIKILGNQG---SFLTKGPSKLNDRADSRRSLWD-QGNFPLII
Thy-1    -KKKHVLSGTL---GVPEHTY--RSRVNLFSDR----FIKVLTL
CD2 d1   --TLVAEFKRKMKPFLKSG------AFEILA--------NGDLKI
CD80     ---MVLTMMSG--DMNIWPEYK--NRTIFDT-------NNLSIVI
CD152    -VTEVCAATYMMGNELTFLDD---SICTGTSSGN----QVNLTI

                   F                  G
Ig κ     SSLQPEDIATYYCQQFDNL-----PLTFGGGTKVDFK
TCR β    QPSEPRDSAVYFCASSFSTCSANYGYTFGSGTRLTVV
CD8 α    SKFSTKNQGYYFCSITSNS-----VMYFSPLVPVFQK
CD4 d1   KNLKIEDSDTYICEVE----------DQKEEVQLLVF
Thy-1    ANFTTKDEGDYMCELRVSGQNPTSSNKTINVIRDKLV
CD2 d1   KNLTRDDSGTYNVTVYSTNG----TRILDKALDSRIL
CD80     LALRPSDEGTYECVVLKYEKDAFK-REHLAEVLLTVK
CD152    QGLR MDTGLYICKVELMYPPPYY-LGIGNGTQIYVI
```

2. Models of activation

The molecules highlighted above, have all been found at the T-cell surface and have been shown to be involved in T-cell activation. However, the definitive functions for some of the molecules remain to be elucidated and various models for the mechanism of activation have been devised. There are three main classes of model for the initial stage of activation (TCR triggering). These models attempt to explain how interactions of the molecules at the cell surface lead to intracellular signalling. The three classes of model are: aggregation, conformational change and segregation[49]. Whilst more than one model has been proposed for each class, here I will just discuss one representative model. These models are based on a wide variety of data obtained by both structural (e.g., NMR and X-ray crystallography) and interaction (e.g., BRET and Biacore) methods.

Aggregation of the TCR complex has been shown to be sufficient to initiate TCR triggering. Using antibodies to the TCR complex led to artificial aggregation which initiated TCR triggering[50]. In the coreceptor independent aggregation model, binding of multiple TCR complexes, on the T-cell, to multiple peptide MHCs (pMHC), on the antigen presenting cell (APC), leads to an increase in concentration or aggregation of the TCR complexes[51]. This aggregation is thought to lead to an increase in concentration of the tyrosine kinases (e.g., Lck) which then leads to TCR triggering.

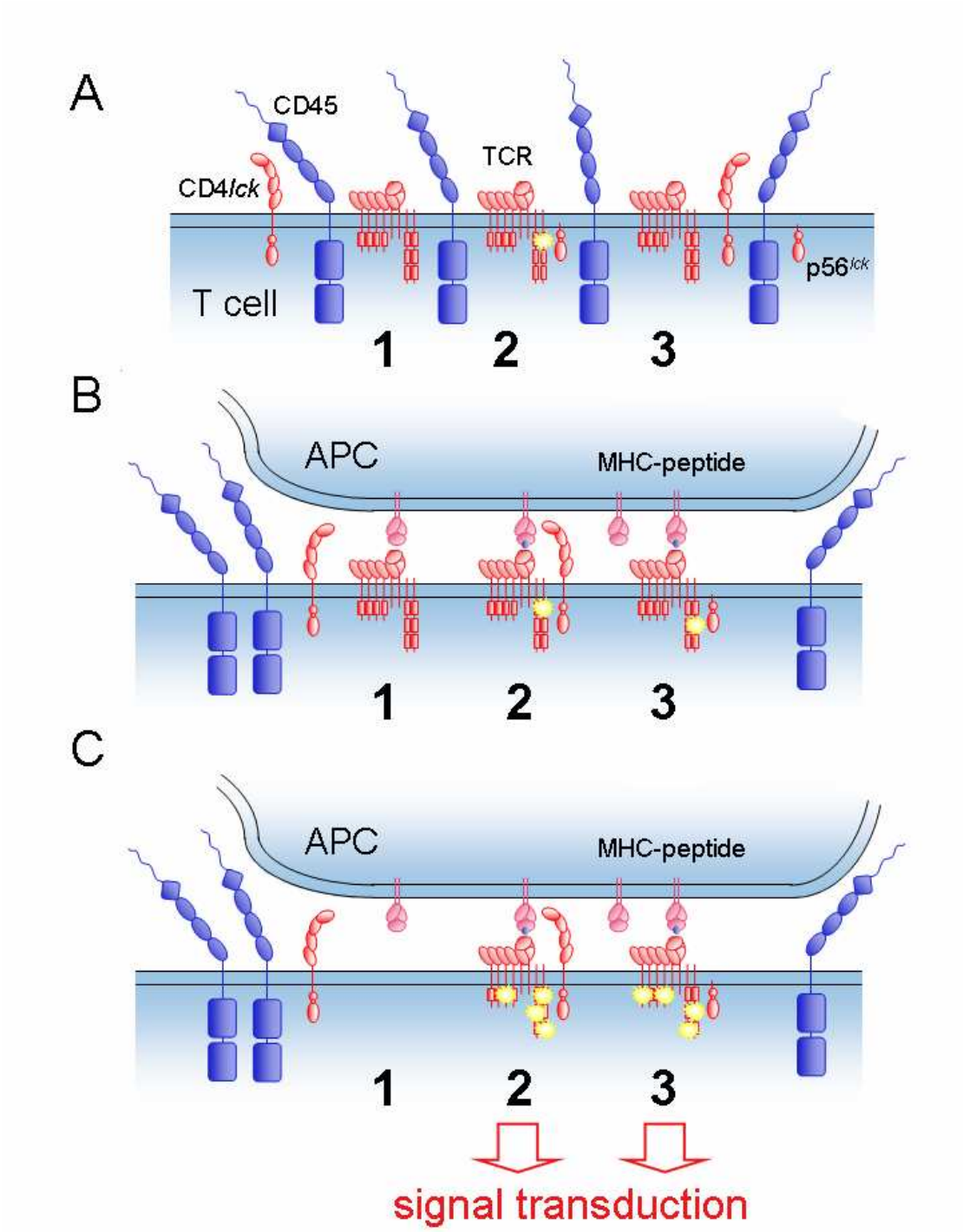
Conformational change models rely on binding of the TCR complex to the pMHC leading to some form of conformational change in TCR complex which results in internal signal transduction. This has been observed in many other signalling systems, for example the nicotinic acetylcholine receptor[52]. One such model

suggests that the force of binding of the pMHC to the TCR complex leads to a rearrangement of the TCR complex and its associated molecules[53]. This rearrangement may lead to exposure of a proline rich region in the cytoplasmic tail of CD3ε, which can then bind the SH3 domain of Nck, leading to intracellular signalling[53].

The final class of models are the segregation based models, where inhibitory molecules are physically separated from the smaller signalling proteins. The first such model, which emerged from my present laboratory, is the "kinetic-segregation model" (Figure 1.4)[9]. This model proposes that upon the TCR complex binding the pMHC on the APC there is a passive, size dependent segregation of large and small molecules on the cell surface. The inhibitory molecules, such as the CD45 and CD148 phosphatases, which have large extracellular domains, are excluded from the contact zone which then leads to net phosphorylation of CD3 by kinases attached to the inner leaflet of the membrane, such as Lck, initiating signalling. Evidence for this model includes the observation that the truncation of CD45 leads to inhibition of triggering[54].

Figure 1.4: Kinetic-segregation (K-S) model of TCR triggering.

(A) In resting T-cells, random interactions of membrane inner leaflet- or co-receptor associated p56lck and CD45 with the TCR results in constitutive TCR phosphorylation/dephosphorylation cycles. The net phosphorylation level is very low but, at any given time, a population of TCRs will have phosphorylated immunoreceptor tyrosine-associated activation motifs (e.g. yellow spot in TCR 2 in A). (B) When the T cell encounters an antigen-presenting cell (APC), the proteins segregate according to size, thereby excluding the phosphatase CD45 and extending the half-life of phosphorylated species within close-contact zones. Note that the K-S model postulates that these close-contact zones are small, as shown in B, and that there may be many such zones within the cell-cell interface. Specific TCR/pMHC interactions hold the TCR in the close-contact zone long enough either for co-receptor recruitment to pre-existing phosphorylated TCRs (e.g. TCR 2 in B), or the phosphorylation of new TCRs (e.g. TCR 3 in B). (C) TCRs that fail to engage MHC bearing appropriate peptide (e.g. TCR 1 in B) will diffuse rapidly from the close-contact zone and be dephosphorylated by CD45 prior to initiating signal transduction. This is the basis of signalling specificity. In contrast, the incipient phosphorylation of engaged TCRs will be amplified by coreceptor recruitment (e.g. TCR 2 in C), leading to full signal transduction. If the highly dynamic contact zone persists for long enough, and the TCR/pMHC interaction is sufficiently stable, engaged TCRs might also be hyper-phosphorylated by free p56lck (e.g. TCR 3 in C), accounting for co-receptor independent signalling. Crucially, the model is not dependent on any explicit ligand (i.e. pMHC)-induced effects beyond simple binding. Other monovalent T-cell signalling molecules dependent on extrinsic kinases, such as CD28, may induce signalling in a similar manner given that, for these molecules, there is no obvious structural basis for receptor dimerization or conformational changes either. This figure is a preprint version of Figure 1 by Davis and van der Merwe[55].


A
CD45
TCR
CD4/ck
p56lck
T cell
1
2
3
B
APC
MHC-peptide
1
2
3
C
APC
MHC-peptide
1
2
3
signal transduction

The key point to take from these models, and the starting point for the current investigation, is that it is not currently possible to determine precisely how T-cells are activated and it may be that key components remain to be identified. If one can get an idea of how many key molecules remain to be discovered, then one can decide where to focus future research. If it appears that many molecules remain to be discovered then future work should be directed at molecular discovery, whereas if the large majority of molecules appear to be known then further functional data should be sought for these key molecules.

3. Synapse formation

The models described above attempt to explain the first stage of T-cell activation, TCR triggering. However, T-cells require at least two signals for their activation and these may be temporally separated. The initial signal is an antigen specific signal provided by interaction between the T-cell receptor complex and the peptide–MHC complex. The effects of triggering can be observed seconds after stimulation of T-cells[56, 57]. The second signal is a non-specific signal provided by costimulatory surface molecules (e.g., CD8) interacting with their ligands (e.g., CD80) on the surface of the APC. Both these signals are required for activation, as it has been shown that TCR triggering without secondary signals leads to anergy (reviewed by Lenschow *et al.*[58]).

The secondary signal is thought to be enhanced by the formation of the immunological synapse, which happens minutes after triggering[59, 60]. Synapse formation occurs in two stages leading initially to formation of the immature synapse followed by the mature synapse [61]. The mature synapse consists of two zones: the central c-SMAC and the peripheral p-SMAC (Figure 1.5)[62]. At the T-cell surface the

c-SMAC contains the TCR and costimulatory molecules such as CD4 whereas the surrounding p-SMAC contains integrin adhesion molecules[63]. It is postulated that this spatial restriction of molecules in the synapse enables the costimulatory molecules to bind to their ligands leading to full T-cell activation. Without this spatial restriction, the low affinity of the costimulatory molecules for their ligands may prevent sufficient binding for costimulation[64].

The synapse is also thought to have other functions in the immune response (Figure 1.5). For example, it has been observed that contents of secretory granules in $CD8^+$ T-cells are released at the synapse[65]. This may be to reduce the widespread diffusion of these products, limiting their effect to only the target cell[66].

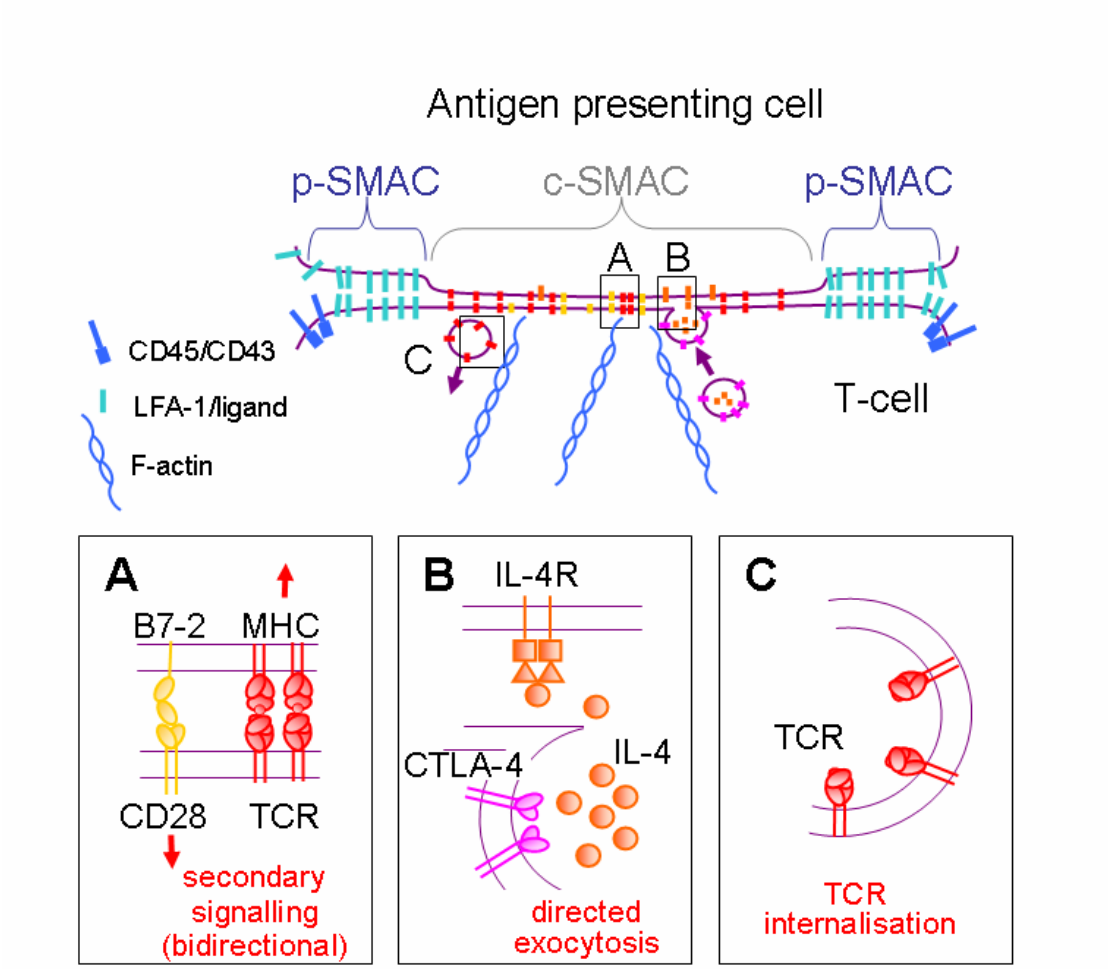


Figure 1.5: The structure and function of the immunological synapse.

The main figure shows the basic layout of the synapse. The panels show the different proposed functions of the synapse. It is proposed that secondary signalling, by e.g., costimulatory molecules and MHC proteins, is enhanced within the central region of the synapse (A), and that the targeted delivery of key effector molecules, such as the cytokine, IL-4, and the inhibitory protein, CTLA-4, to this region via microfilament-dependent exocytosis (B), is necessary to limit bystander effects. Lee *et al*. propose that the synapse is required for TCR internalization (C)[61]. This figure is based on a preprint version of the figure from van der Merwe and Davis[66].

C. Molecular discovery

The molecules highlighted above were characterised individually. To meet the initial aim of this thesis, which is to catalogue all the "immune specific" proteins found on the T-cell surface, global methods will need to be used.

1. Molecule by molecule discovery

To characterise leukocyte cell surface antigens it was necessary to be able to isolate the individual molecules. The initial work involved isolating the proteins and then sequencing the amino acids; this was successful for a limited number of proteins including MHC Class I and II[67, 68]. However, a large increase in the discovery rate was enabled by the combination of two technologies: monoclonal antibody production and genetic techniques. These DNA based methods led to the discovery of most known cell surface antigens[2].

The aim of monoclonal antibody production is to produce a cell line that produces a single antibody that is specific to a single protein. Kohler and Milstein achieved this in their Nobel Prize winning work in the 1970s[69]. The basic principles of mAb production are outlined in Figure 1.6.

Initially, these mAbs were used on affinity columns to purify antigen. Using this antigen, one could sequence enough protein to design primers to either screen cDNA libraries or use PCR to amplify the cDNA. Many of the cell surface proteins, listed above, were identified in this manner, including CD2, CD3δ, CD3ζ, CD5, CD8β, CD11a, CD18, CD43 and CD45.

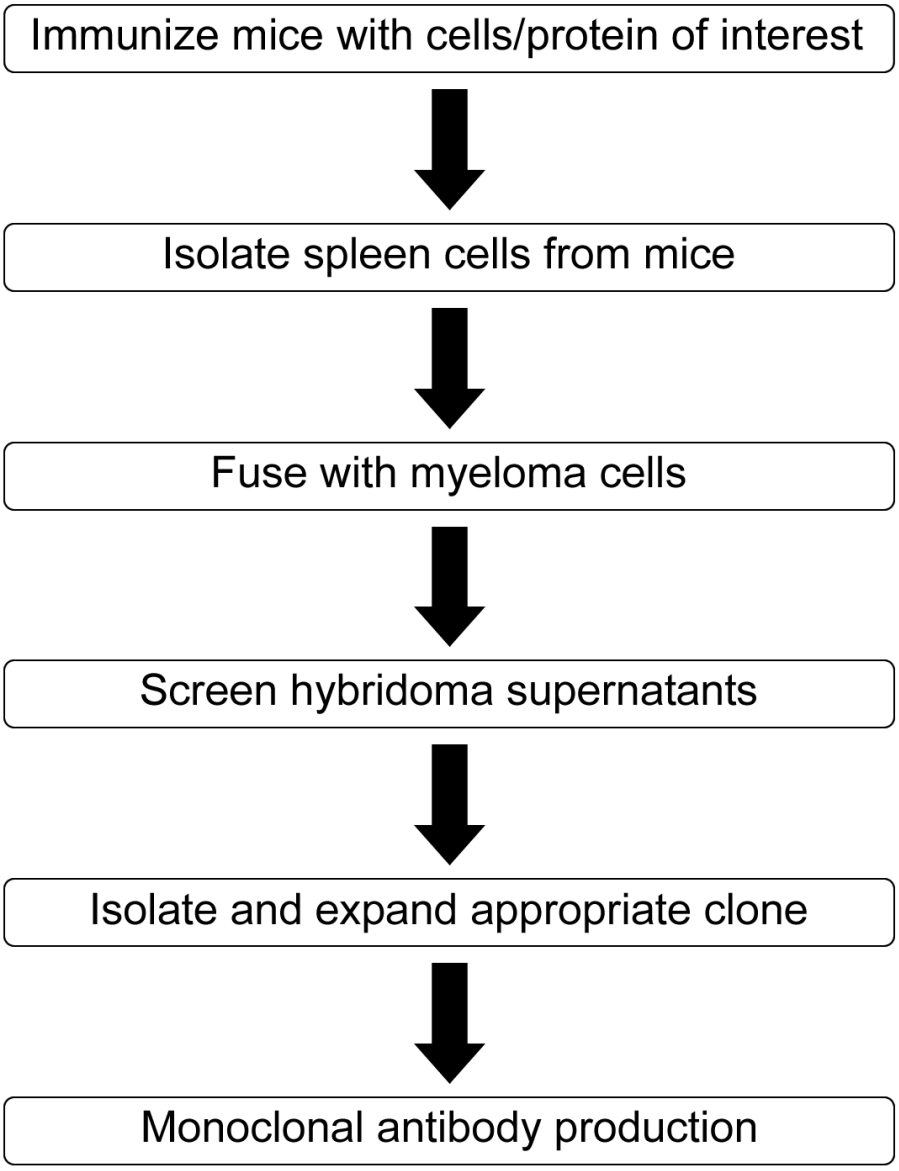


Figure 1.6: Monoclonal antibody production.

This figure highlights the key stages involved in monoclonal antibody production as devised by Kohler and Milstein[69]. This figure is based on Figure 1 from Monoclonal Antibody Production[70].

The next technological shift was to combine the use of mAbs with genetic engineering. DNA was transfected into cells and these cells were screened with the mAbs. The transfected DNA was isolated and then used to screen cDNA or genomic libraries. In the proteins listed above this led to identification of CD4 and CD8α. This method was improved upon by Seed and Aruffo and led to a large increase in the rate of leukocyte cell surface molecule discovery[71]. The basic principles of their method are as follows. They transfected a cDNA library into COS cells for transient expression. These cells can then be screened (or "panned") with mAbs for the antigen of interest (or in a modified version of the protocol, the cells can be screened with known ligands for the molecules of interest[72]). The plasmids from the selected COS cells are then extracted and the process repeated until a single plasmid is identified. The authors identify the key advantages of their method, with the principle advantage being the recovery of a complete cDNA that encodes for a protein that reaches the cell surface and is correctly folded. The use of a COS expression system also enabled high expression levels of the desired protein, increasing the chance of it being detected at the cell surface. The authors used this system to identify CD2 and CD7, and CD6 was also identified by this method[71, 73].

Whilst the mAb approach developed by Kohler and Milstein revolutionised antibody production and use, it is not perfect and novel approaches have been developed that are both faster and avoid the use of animals[74]. The most advanced of these techniques is phage display[75]. In this technique, a bacteriophage library is produced that contains bacteriophage that synthesize a fragment of an antibody which are "displayed" on their surface. The DNA required for these antibody fragments is either artificially produced (e.g., taken from human germline antibody genes and

synthetically rearranged *in vitro*) or naturally derived (e.g., taken from naturally rearranged genes in human B cells)[76]. The library is then screened with the antigen and those phage that bind to the antigen are kept, yielding a polyclonal mixture of phage antibodies[77]. Bacteria are then infected with the phage and single phage species are kept that produce high affinity antibodies. Currently this technology is used to produce antibodies for a wide range of purposes, including therapeutic uses. However antibodies for cell surface markers, e.g. CD54, have been identified by this technique[74]. Therefore, it could equally be used as the antibody production stage in Seed's method to produce antibodies for unknown cell surface proteins. A limitation, however, is that the antibodies are generally of lower affinity than conventional monoclonal antibodies.

2. Global technologies

To examine the proteome of any cell it is possible to use either direct or indirect methods. Direct methods examine the levels of various proteins (proteome) found in the cell while the indirect methods can be used to examine the DNA (genome) or mRNA (transcriptome).

Before considering the merits of these different technology classes it is worth considering the range required by the techniques. One should consider both the total number of proteins currently known and the number different proteins one would expect in a single cell type. Due to the limited power of current proteomic methods, to examine the number of possible proteins it is necessary to examine both the genome and the transcriptome, and to examine the number found in any one cell, one should examine the transcriptome. Whilst there is believed to be a similar order of magnitude in the number of protein and mRNA species, the range is thought to be

very different. It has been estimated that there is perhaps a four orders of magnitude difference between common and rare transcripts, whereas in proteins the difference may be eight orders of magnitude[78].

The number of genes in the human genome has traditionally been estimated at around 80,000-100,000 genes[79]. However, as the sequencing of the human genome progressed it became clear the number was far less. A competition held from 2000 to 2003 to predict the total number of genes showed that many people still expected a large number, with a mean value 61,710 (see www.ensembl.org/Genesweep)[80]. In the "finished" human genome there were 19,438 known protein coding genes[81]. However, this is still believed to be an underestimate due to both the incomplete nature of the genome and the fact many genes have been predicted but not identified in the transcriptome (whether these are false positives remains to be seen). Recent studies using whole genome arrays have suggested that there is large scale transcription outside the annotated genes[82, 83]. Of course, whether this transcription is due to random events or is functionally important is unclear.

Few estimates have been made about the number of different proteins/transcripts found in a cell. Bishop *et al*. used R_0t curves to estimate that there are approximately 35,000 different mRNA species in HeLa cells (these are cells from an "established line of human epithelial cells derived from a cervical carcinoma")[84]. From more recent work, using global gene expression analyses it is difficult to calculate a number of definite species. If one looks at cDNA or standard oligonucleotide microarray data, then one is limited by both the detection rate of the technique and what is on the array, and therefore any gene estimates are likely to be underestimates. However, if

one examines genome tiling array data or tag-based methods, then the criteria chosen for grouping the observed fragments has a large impact on the number of genes identified (this is shown in Chapter 5). What is clear from all these experiments is that any global technique used to catalogue all proteins/transcripts found in a cell must be able to discriminate in the order of tens of thousands of different species.

a) Proteomics

Ideally, one would use direct methods to catalogue all the proteins found in a cell. However, proteomic technologies are currently limited both in the breadth and depth of coverage. The latest technologies can only deal with a limited number of proteins (e.g. a recent study using "shotgun proteomics" managed to characterise 5,130 proteins[85]), have problems with integral membrane proteins and generating quantitative results is not routine[86]. It is clear that with current technological limitations proteomic methods are not appropriate for cataloguing a large number of cell proteins.

b) Genomics

Having ruled out a proteomic based approach one can use either a genome or transcriptome based approach. Theoretically it should be possible to examine the genome and predict which proteins will be found in a given cell. However, there are problems with this approach. The first and most major problem for this approach is that cells respond to external stimuli, and therefore, not only genomic information will be required to predict the contents of a cell. However, even if one could predict the environment of the cell there are still problems. Firstly, even though the genome is now "finished" there remain regions of euchromatin which cannot be sequenced and a large proportion of the heterochromatin has not been sequenced[81]. This means

that some genes (and thus their corresponding proteins) may not have been sequenced. Secondly, current gene identification approaches are not perfect. Genes can be identified using a combination of transcriptome information (e.g., ESTs, SAGE, microarray, full length mRNAs etc.) and gene prediction software. However, transcriptome information is limited by what has currently been sampled, therefore cell specific genes may not have previously been identified if the transcriptome of the cell type of interest has not previously been probed deeply. Whilst gene prediction software is constantly improving, it also not perfect, often having both a high false positive error rate and a low false negative rate[87]. Even bearing these limitations in mind, genome wide studies have been successfully used to identify novel molecules of immune interest. For instance, one study involved the use of hidden Markov models designed on members of the β defensins subfamily to successfully identify novel members of this family[88].

Even if it was possible to identify all genes found in the genome, our understanding of the control of when genes are expressed is still limited. Whilst eukaryotic gene expression is thought to be stochastic[89], various factors have been identified that affect gene expression. For example, there has been shown to be a correlation between the putative peptide content of a gene and its expression level[90]. For many genes, the key components for controlling whether the gene is expressed are still not known. Various transcription factors have been shown to have a major role in the control of gene expression. However, it appears likely that the exact roles of many transcription factors remain to be discovered. Discovery of these roles can enable major progress in understanding immunological processes. For example, recently it was discovered that the transcription factor cKrox is necessary and sufficient for a

thymocyte to become a $CD4^+$ T-cell[91]. In the future, one may hope to use the genome to predict the contents of a given cell, however, it is not currently feasible.

c) Transcriptomics

Having ruled out the use of proteomic and genomic technologies this leaves transcriptome (gene expression) based technologies. There is a wide variety of gene expression technologies, ranging from those that examine a single gene (e.g., Northern blot[92]) to those that aim to profile every transcript class in the cell. As the analyses required in this thesis are intended to be exhaustive, here I will review the different classes of global gene expression technologies focussing on their main strengths and weaknesses. I will then present a summary of how the suitable techniques work and discuss their limitations.

(i) Technologies

There are two main classes of gene expression technologies: "open" and "closed" technologies. With closed technologies, one decides before the experiment which genes one wants to examine, whereas with open technologies there is no such requirement and in theory at least, one can profile all genes expressed in the cell. There are a large number of global gene expression technologies, some of which use proprietary technology (e.g. massively parallel signature sequencing, MPSS[93]) whilst others use readily available technology or approaches (cDNA microarrays). Here I will review the technologies that were considered for the experiments in this thesis: serial analysis of gene expression (SAGE[94] and LongSAGE[95]), MPSS, cDNA microarrays[96] and oligonucleotide microarrays[97]. Table 1.1 compares the relevant characteristics of the different technologies.

	Open technology		Closed technology	
	SAGE/LongSAGE	MPSS	cDNA microarry	Oligonucleotide microarray
Breadth	>99% of genes	>99% genes	Limited by array design	Limited by array design
Depth (tags or equivalent)	Hundreds of thousands	>1 million	?	Estimated to be 120,000
Timescale	Weeks	Weeks	<1 day	<1 day
Quantitative	Yes	Yes	No	Yes*
Data comparable	Yes	Yes	Yes‡	Yes*

Table 1.1: Comparison of "open" and "closed" technologies.

The table above compares some of the key factors of the different classes of gene expression technology. In this thesis, the breadth, depth and comparable nature of the data were the most important factors when choosing which technique to use.

‡ cDNA microarrays work by comparing the relative expression in two samples, so by definition the data is comparable. However, whilst it is possible to compare data between laboratories by the use of reference mRNA samples this increases the error.

* In theory, oligonucleotide microarrays are both quantitative and comparable between labs, however it has been shown that which laboratory the experiments are carried out in has a large effect on the results obtained[98].

The deciding factor in choosing the technology was whether it was open or closed. To meet the initial aims an open technology was required. Microarrays are the most commonly used global expression technology, primarily for their speed and relatively low cost. However, a microarray is only as powerful as the set of probes found on it. Whilst the current generation of microarrays, that contain probes based on both known transcripts and genome data, are almost equivalent to open technologies those available in 1997 (when the strategy was being decided) did not have this breadth of sampling. This lack of transcript coverage on microarrays can be significant, one relatively recent study on platelets found that almost half the genes identified in a small SAGE library were not tested by the oligonucleotide array probes[99]. Therefore microarrays were not used.

Whilst microarrays have been discounted for use in the study, as they are the most commonly used global expression technique it is worth reviewing the correlations observed when array and tag-based experiments have been compared. Interestingly, the two differing microarray technologies have been compared, and there was found to be little correlation between the two methods[100]. However, this used publicly available data and the comparisons were between the same originating cell lines but from samples that were grown separately. In another report, it was shown that the effect of samples being produced in a different laboratory was greater than that of using differing array technologies[98]. A comparison of arrays and MPSS data on arabidopsis found a high correlation if only the abundantly expressed genes were analysed[101]. Equally, when the authors examined the differentially expressed genes, if they excluded the lowly expressed genes, they obtained a high correlation. A similar result was observed when comparing oligonucleotide arrays with SAGE data

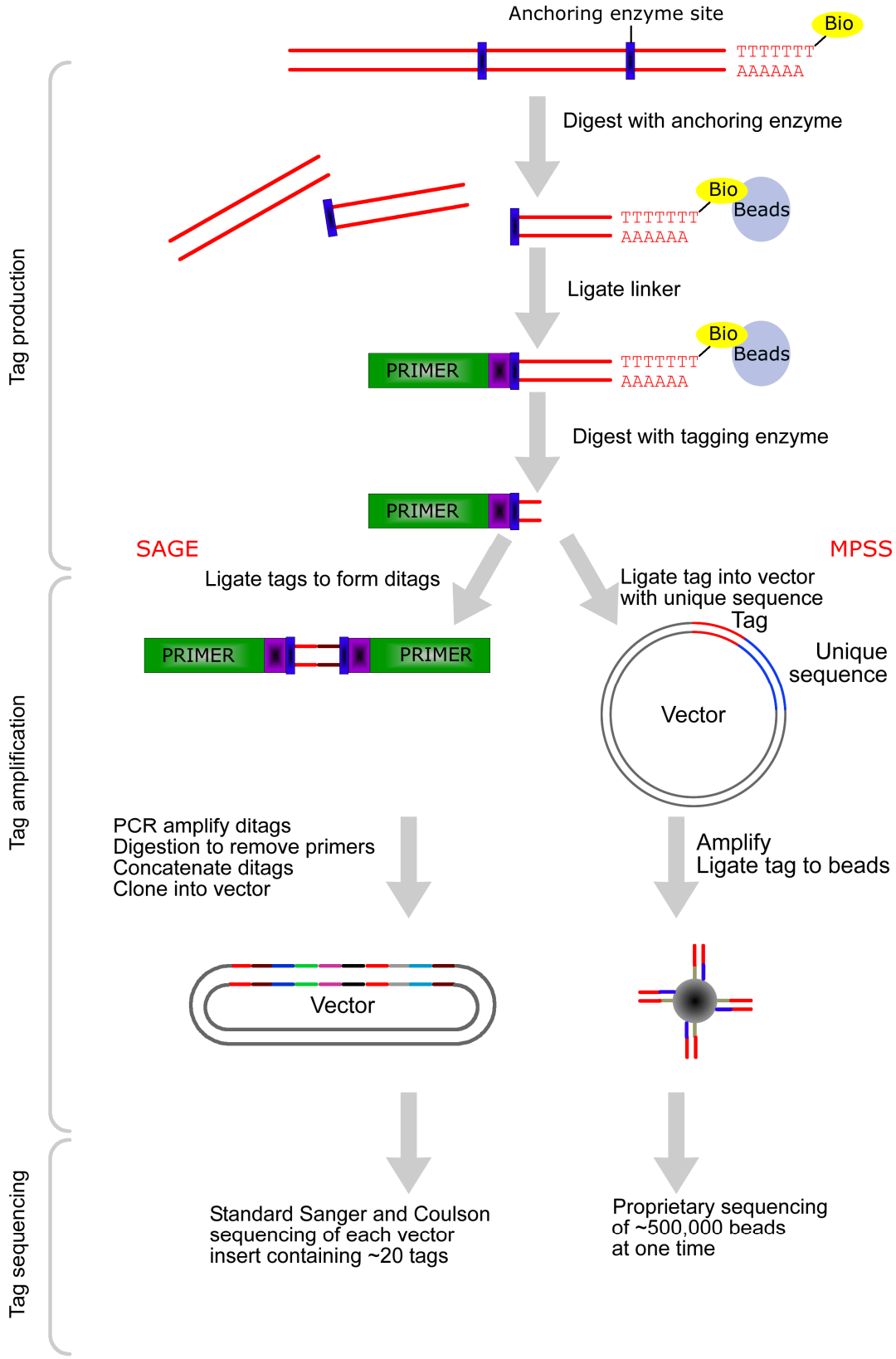
for barley[102]. The authors "found better agreement when comparing genes expressed at high levels than for those expressed at low levels". The general pattern observed is that when comparing only the highly expressed genes, one finds a reasonable level of correlation but if one compares the data for all genes then the correlation is low (and that when one examines the highly expressed differentially expressed genes there is also a high correlation). This is not surprising, as neither the arrays nor the tag-based technologies profile the entire transcriptome. Therefore one would expect that by chance, both experiments would sample the higher expressed transcripts but not the less abundant ones. Thus one would expect that excluding the less abundant transcripts from the comparisons would lead to an improved correlation. However, there are problems with interpreting array data, for example it has been reported that there are errors in the probe design, and this may be contributing to these poor correlations[103]. A recent study suggested that on the Affymetrix Gene Chip U95A "approximately 11% of the probes are non-specific and 9% of the probes are mistargeted"[104]. The authors found an increased agreement between SAGE/EST and microarray data once these probes were ignored. To address whether these poor correlations are due to array specific issues or random sampling bias that affects all techniques, one would need to compare two tag-based experiments. However, there are currently no published examples of this; therefore this will be investigated in this thesis. Other published studies are summarised in Table 1 of the paper by van Ruissen *et al.*[105].

Both SAGE and MPSS are tag-based methods and can be thought to consist of three main stages: tag production, amplification and sequencing. Both are similar for the tag production stage but differ greatly in their amplification and sequencing steps

(Figure 1.7). (LongSAGE is almost identical to SAGE except for the use of a different restriction enzyme in the tag production stage and so will not be considered separately here.) The basic principle behind both technologies is similar to that of sequencing EST libraries. By reducing the length transcript that is required (i.e., in EST libraries the length is reduced from a full length mRNA down to a few hundred bases) one can sequence more transcripts for the same cost, in effect increasing the depth of transcriptome coverage. Both of these technologies take this to its extreme, by cleaving the sequence at its 3′ most restriction site for a given enzyme, generating a short sequence (i.e. a 14-21 base tag) that can, theoretically at least, be uniquely assigned to its corresponding transcript. These tags are then sequenced in parallel (either by concatenation and standard sequencing in SAGE or by using proprietary sequencing technology in MPSS). This significantly reduces the cost and time required to produce a library with deep coverage.

Figure 1.7: Key stages in the tag-based gene expression methods, SAGE and MPSS.

There are three key stages in tag-based gene expression techniques, tag production, unbiased tag amplification and simultaneous tag sequencing. Both SAGE and MPSS have a similar tag production stage but differ in both the amplification and sequencing stages. At the level of detail described here, LongSAGE and SAGE are equivalent. The key principle behind the tag production stage is that by the use appropriate restriction enzymes it is possible to generate a tag of fixed length and relative position from each transcript. In SAGE, unbiased tag amplification is achieved by ligating two tags to form ditags, these ditags are then amplified by PCR. The lack of bias is achieved in the data processing stage, as the sequence of any one ditag will only be kept and processed into tags *once*. This avoids bias by preventing the analysis of multiple copies of the same sequence. MPSS uses a different approach to avoid bias. Each tag is ligated into a vector that contains a unique sequence. A library of beads is produced, in which each bead contains a unique sequence which is complementary to a unique sequence in a vector. Tags are then ligated to beads, via the unique sequence. As there is only one bead with each unique sequence only one copy of each tag will be sequenced. The power of both techniques is that more than one tag is sequenced per sequencing run, i.e. simultaneous sequencing. In SAGE, the tags are concatenated, so that ~20 tags are sequenced per sequencing run. In MPSS, proprietary technology is used, to sequence the tags attached to the beads. Approximately half a million beads are sequenced simultaneously. This figure is adapted from figures produced by Harbers and Carninci[106], and Reinartz *et al*.[107].


Tag production
Tag amplification
Tag sequencing
Anchoring enzyme site
Bio
TTTTTTT
AAAAAA
Digest with anchoring enzyme
Bio
Beads
TTTTTTT
AAAAAA
Ligate linker
PRIMER
Bio
Beads
TTTTTTT
AAAAAA
Digest with tagging enzyme
PRIMER
SAGE
MPSS
Ligate tags to form ditags
Ligate tag into vector
with unique sequence
Tag
Unique
sequence
Vector
PRIMER
PRIMER
PCR amplify ditags
Digestion to remove primers
Concatenate ditags
Clone into vector
Amplify
Ligate tag to beads
Vector
Standard Sanger and Coulson
sequencing of each vector
insert containing ~20 tags
Proprietary sequencing
of ~500,000 beads
at one time

(ii) Limitations of tag-based technologies

Having established that tag-based transcriptome approaches are the only currently suitable methods to solve the initial aim, it is worth examining their limitations in further detail. Firstly, I will highlight the largest problem in tag-based technology data interpretation, then I will review the different species of RNA found in the cell and finally I will examine the level of correlation between protein and mRNA abundance.

(a)Transcript identification

The biggest weakness of tag-based approaches is the transcript identification or "tag-to-gene" mapping step. This involves converting tag information into transcript information. In an ideal world, each tag would uniquely map to one known transcript and one location on the genome. There are two key factors which affect the tag-to-gene mapping. The first is tag length, the longer the tag the more likely it is to be a unique match to both the transcriptome and genome. However, increasing tag length increases the cost and time involved in sequencing and also increases the chances of sequencing errors. Unneberg *et al.*[108] recommend an optimal tag length of 16-17 bases when looking at human data and Pleasance *et al.*[109] find that 16 bases is ideal for drosophila data; even with tags of this length one would not expect all tags to be unique in both transcriptome and genome. SAGE tags are 14 bases, LongSAGE tags are 21 bases and MPSS tags are 20 bases long. The effect of the differing tag lengths on the ability to uniquely map tags to their corresponding transcript is investigated in Chapter 5.

The second factor which affects tag-to-gene mapping is the quality of the database used. The ideal tag-to-gene mapping database has three properties. Firstly, tags would be extracted from the genome, secondly tags would be extracted from the transcriptome and thirdly this data would be combined and tag-to-gene annotations would be ranked by the quality of the mappings (based on both transcriptome and genome data).

At the time of the initial analysis in this thesis there was only one publicly available mapping database, which is one component of the ever improving SAGEmap utility at NCBI. The SAGEmap database consists of tags extracted from the 3′ most restriction site of sequences found in Unigene[110]. The tags are classed depending on both the sequence type (i.e., EST or cDNA), direction and presence of a polyA signal or tail. However, there are problems with this set-up. Firstly, tags are not extracted from the genome, this means that tags that are found in novel transcripts or exons will not be classified. Secondly, using only the 3′ most restriction site limits the mapping, as some tags will be derived from sites that are not the 3′ most site (due to various factors, such as alternative splicing, internal priming, incomplete cleavage by restriction enzyme, and polymorphisms). Thirdly, basing the tag extractions on Unigene data provides two limitations. Unigene focuses on protein coding transcripts[111], excluding non-coding RNAs, whereas one of the advantages of SAGE is that it can be used to catalogue a wide range of RNA species (discussed further below). Also, Unigene cluster IDs are not stable, this causes problems because SAGEmap is not updated simultaneously with every release of Unigene and therefore Unigene clusters identified in SAGEmap often no longer exist. Finally, and most importantly, the classification data is not made available in the publicly accessible

database. This means that many tags falsely appear to match multiple genes (the extra false matches are due to poor quality or un-oriented sequences) and this significantly limits the power of the technique.

To prevent these problems limiting the technique and to enable the aims of the thesis to be met, two approaches were decided upon. Firstly, for the SAGE experiments, the SAGEmap database was used, however, each tag match was manually curated. This led to the removal of tag matches that were solely due to poor quality or un-oriented sequences, reducing the number of tags that appear to artificially map to multiple genes. However, this is a field where improvements are being made and recently new tag-to-gene mapping databases have become available (e.g., SAGE Genie[112], Human Transcriptome Map[113] and DiscoverySpace, see www.bcgsc.ca/discoveryspace/). In particular, SAGE Genie addresses many of the problems highlighted above and was used as an additional database in some of the analysis. Secondly, a custom tag-to-gene database was designed using the human genome sequence as its framework. This was used to compare LongSAGE to MPSS and also, to examine *cis*-NAT transcription.

(b) RNA classes

RNA has been shown to have a variety of cellular roles; these include: catalytic (ribosymes), information storage (mRNA), information transfer (tRNA), regulatory (e.g., microRNA) and structural roles (rRNA and histone RNA). Only RNA with an information storage or regulatory role will be considered here. For the tag-based expression techniques used here, the direction, function or cellular location have no effect on whether it can be studied, the only requirements are the presence of a polyA tail and the relevant restriction site. Historically, RNA has been classified by

direction, i.e. whether it is transcribed from the putative sense or antisense strand; though this distinction is now becoming blurred, for example, microRNAs can be found to be transcribed in the sense direction from one gene but bind in an antisense manner to another.

Traditionally, the bulk of sense RNA has been presumed to be mRNA. This is RNA that is transcribed from the genome, producing a transcript in the sense direction that is capped at the 5′ end, has a polyA tail at 3′ end, is spliced, exported to the cytosol and then translated. This class of mRNA is presumed to account for the majority of mRNA that leads to protein production. However, studies in the seventies suggested the existence of translatable transcripts that lack polyA tails, and recent global analyses have made it clear that there are a large number of RNA molecules found in a cell that are different to this stereotypical mRNA. Sonenshein *et al*. isolated an mRNA molecule from mouse sarcoma 180 ascites cells that lacked a polyA tail but could be translated *in vitro*[114]. Cheng *et al.* used high resolution arrays to examine the sites of transcription of polyadenylated and non-polyadenylated RNA for 10 human chromosomes in a variety of cell lines[83]. Of the transcribed sequences they found that 19.4% of sequences were polyadenylated, 43.7% were non-polyadenylated and the remaining 36.9% were found in both states. However, this work was not quantitative and they are merely examining whether transcription was found at a site in any of ten tissues. Therefore, one might speculate that the large majority of this non-polyadenylated transcription occurs at a low level (possibly due to random transcription events) and is functionally unimportant. Indeed, the authors point out that some of these non-polyadenylated transcripts are likely to be excised introns.

The situation for antisense transcripts is more complicated, primarily because the widespread nature of antisense transcription and variety of roles for antisense transcripts have only recently been discovered. The first distinction to make for antisense transcripts, is between naturally occurring antisense (NATs) and "artificial" antisense. Artificial antisense transcripts are those antisense transcripts that are not produced by the cell under normal conditions, but are found due to exogenous events, such as viral infections or RNAi experiments.

The natural antisense transcripts (NATs) contain a wide range of classes but can be classified into two groups, *cis*- and *trans*-NATs. *Trans*-NATs are RNA molecules produced at different locations in the genome to the mRNA that they are complementary to and may only exhibit partial complentarity, whereas *cis*-NATs are produced on the opposite strand at the same location in genome as the mRNA they are complementary to and show perfect complentarity[115]. By definition *cis*-NATs can only be the antisense partner of one other species of mRNA, whereas *trans*-NATs may be complementary to many different mRNA molecules. Some of these molecules appear to be processed by similar pathways to sense mRNA, whereas others use distinct pathways.

microRNAs (miRNA) are *trans*-NATs. The majority are found to be transcribed from genomic loci separate to other transcripts, although a quarter of human miRNAs were found to be introns of pre-mRNAs[116]. They were found initially in *C. elegans* and the first example *lin-4* was shown to be involved in translational repression[117]. They have since been found in many other higher organisms, including humans. Their biogenesis is thought to involve a pri-miRNA which is processed in the nucleus

to become a pre-miRNA. This is a stem loop structure that is approximately 60-70 bases long, which is cleaved (by Dicer homologs) in the cytoplasm to produce an approximately 22 base microRNA that is found in a RISC (RNA-induced silencing complex)[116]. Once in the RISC, the microRNA is thought to lead to either mRNA cleavage or translational repression of its target(s). As these microRNAs do not contain polyA tails, they cannot be profiled by the tag-based methods. However, human pri-miRNAs have recently been shown to be capped and polyadenylated, and also function as mRNAs[118, 119]. Therefore they can be detected by the tag-based methods.

Small interfering RNAs (siRNA) are similar to microRNAs in both size and function, however, they have a different biogenesis[120]. The initial stages of the formation of the double stranded RNA (dsRNA) required for endogenous siRNA production are not fully understood. In the case of transposon silencing in *C. elegans* it was found that the dsRNA was produced by "snap back" in the Tc*1* transcript (a transposable element)[121]. This RNA was found to be spliced but not polyadenylated. This suggests that the precursors of natural siRNAs would not be detected by SAGE or MPSS. However, whether this is true for all siRNAs or whether Tc*1* is a special case will require further research.

Cis-NATs are perhaps the most widespread of all the antisense RNA species. Recent work has identified thousands of *cis*-NATs in both humans and mice[122-124]. However, their function and structure are perhaps the least well characterised of the antisense transcripts. Structurally these antisense molecules may be one of two types of transcript, in that the antisense transcript may be coding or non-coding (a recent

report suggested the majority of these sequences are non-coding[122]). If the antisense transcript is coding, then defining which transcript is the antisense transcript is arbitrary and the overlap appears to be primarily at the 3′ UTRs of the transcripts (e.g., TESK1 and CD72)[125, 126]. For this type of *cis*-NAT one would expect the structure of the antisense transcript to be that of a typical mRNA molecule. However, if the antisense transcript is non-coding then the structure may be different and the overlap may appear at any position along the sense transcript. Large-scale analyses have revealed interesting structural insights about *cis*-NATs. One report showed that some of the antisense transcripts were complementary to the sense transcripts even at the splice junctions[83]. This led the authors to suggest that these transcripts may be cRNA copies synthesised by an RNA-dependent RNA polymerase, as has been observed in both *A. thaliana* and *C. elegans*. If this is correct, then one would expect that the transcripts would by polyA negative. Indeed, Kiyosawa *et al*. use custom oligonucleotide microarrays and find that a large percentage of mouse *cis*-NATs appear to be polyA negative[127]. This suggests that SAGE and MPSS will only identify a proportion of *cis*-NAT transcripts.

Functionally, the role of these antisense molecules is less clear. It has been suggested they form dsRNA with their complementary sense transcript, and are then deaminated leading to nuclear retention[128, 129]. This in effect reduces the net production of the sense transcript, and suggests these transcripts may have a regulatory role. However, recent work on arabidopsis transcripts, found those gene pairs that had overlapping transcripts (where both transcripts contained an ORF), had increased alternative splicing and alternative polyadenylation[130]. This suggests that perhaps these antisense transcripts have a role in the structure of the protein product

rather than regulation of translation, or this may be artefactual and the increased splicing and polyadenylation is the cause of the detection of the antisense overlap in the first place.

There have been a variety of large scale studies examining antisense transcription. The techniques used range from database mining to high density arrays. A clear pattern seems to be emerging, in that the deeper the sampling, the more widespread the detection of antisense transcription is. In one of the earliest large scale human studies Lehner *et al.* examined transcripts that code for proteins (from RefSeq and those annotated as "complete cds" from EMBL), and found approximately ~2% of transcripts were NATs[131]. Whereas in a larger study, which clustered ESTs and cDNAs, regardless of whether they coded for proteins, antisense transcription was found for approximately 20% of clusters[122].

(c) Correlation

Transcription and translation are controlled by both gene and cell specific factors. For example, if the cell is proliferating then general protein synthesis will be increased, whereas if a memory $CD4^+$ T-cell is co-stimulated by endothelial cells then the half lives of certain cytokine mRNAs are increased, leading to increased cytokine synthesis[132]. At the transcriptional level, gene expression will be influenced by transcription factors, whereas protein abundance will be influenced by factors such as mRNA stability. The correlation between the mRNA and protein abundance may depend on where the key controlling step is in protein synthesis. Is the dominant control step translation or transcription or a mixture of the two? If translation is the dominant control step then one would expect to find little correlation between protein level and gene expression, in that it is only necessary for transcripts to reach the

ribosomes and then other factors control the rate of translation. If transcription is the dominant step, one would expect all transcripts to be translated and there to be a high correlation between protein level and gene expression. In fact, it has been shown that at different times, either one of these control points may be dominant for the same protein (for example, 20 minutes after starvation there is an increase in production of the yeast transcription factor Gcn4p without an increase in transcript level, i.e. translation control, whereas 3-4 hours after starvation there is an increase in transcript level[133]).

One other factor could have a major influence on whether a correlation is observed and this is the relative half-lives of mRNA and proteins. Protein half-lives have been shown to vary by many orders of magnitude, from seconds to days[134, 135]. Eukaryotic mRNA half-lives have also been shown to vary, with observed half-lives being from a few minutes to more than 24 hours[136]. The half-life of a given transcript is determined by both gene specific factors (e.g., ARE elements in the 3'-UTR) and general cell factors (such as environmental stresses)[137]. For a given gene, if the half-lives of the transcript and protein are very different, then it is unlikely that a correlation between protein and mRNA abundance will be observed.

It is likely to be the case that for different genes/proteins different factors control their abundance and thus different correlations are observed. This means that when analysing global gene expression data the observed transcript level of certain genes will better reflect the level of the corresponding protein found in the cell than other genes.

There have been many studies attempting to compare mRNA and protein abundance on a large scale. Both the number of genes studied and number of samples used are now greater than 100. A variety of techniques have been used in the comparisons, from microarrays and flow cytometry[138] to SAGE and mass spectrometry[139] (others reviewed by Greenbaum *et al.*[140]). The overall picture appears to be the same in all the experiments, that is there appears to be little correlation between protein and mRNA abundance. However, when the data is broken down into sub-groups interesting patterns appear. Both Gygi *et al.*[139] and Ørntoft *et al.*[141] find that there is a high correlation for the high abundance genes and proteins. Griffin *et al*. find a high correlation in *Saccharomyces cerevisiae* when looking at the change in abundance of proteins involved in glycolysis upon carbon perturbation[142]. Interestingly, Greenbaum *et al*. find a high level of correlation for those proteins where the transcription of the gene appears to be tightly controlled over the cell cycle. This suggests that when the cell is "putting effort" into controlling the expression of a gene, this relates to a change in protein abundance. Interestingly, Kern *et al.*, as well as examining correlations, also gave data comparing merely the detection of the gene and protein[138]. At this resolution, they found, in over 100 samples and 39 genes, that in approximately 70% of cases, the gene expression data (microarray) and proteomic data (flow cytometry) were congruent. In the majority of cases, where the data were not congruent, this was due to a transcript being detected when the corresponding protein was not observed.

Overall, it appears that on a global scale there is a low level of correlation between protein and mRNA abundance. However, one must remember that current studies are limited by both the technologies used and samples studied. In all papers, only a

limited number of proteins/genes have been studied and so it is possible that the correlations (or lack of) are biased in the subset studied. Gene/protein specific factors may also affect the detection of certain genes/proteins and this may affect the correlations. Experimental factors may also be affecting the observed correlations. There is noise in both gene expression and protein abundance experiments, and when comparing the results from two experiments this noise is likely to interfere with any true correlations. Indeed, Beyer *et al.* find that averaging the results from a range of experiments increases the significance of correlations; this suggests that noise is reducing the correlations[135]. Also in some of the experiments the techniques used were not entirely quantitative (for example, multidimensional protein identification technology as used by Greenbaum *et al*.).

(iii) Previous tag-based global analyses

Despite the potential limitations imposed on tag-based global gene expression techniques by the relatively poor correlation between protein and mRNA abundance, and the inability to profile all forms of cellular RNA, they have been successfully used in a wide variety of discovery driven analyses of immune cells.

One group has used SAGE to characterise a large number of human immune cells, ranging from dendritic cells to sub-types of $CD4^+$ T-cells[143-146]. Others have examined the change in expression in $CD4^+$ T-cells upon HIV infection[147]. In mice, one group has used SAGE to characterise a large number of immune cells to identify important transcripts involved in immune regulation and tolerance[148-151]. Others have examined intraepithelial lymphocytes[152].

Tag based-analyses have also been revealing in the field of gene expression. Caron *et al.* identified "ridges" of transcription by mapping SAGE tags to the genome[153]. These ridges suggested that certain euchromatic regions of the genome are more commonly transcribed than others. Patankar *et al.* first demonstrated that SAGE was suitable for detecting large scale antisense transcription, when they produced a library from the malarial parasite *Plasmodium falciparum*[154]. Large scale antisense transcription has since been detected by tag-based methods in both mouse and arabidopsis libraries[155, 156].

SAGE has also been modified to examine the 5′ end of transcripts. Two groups have published techniques to examine the sites of initiation of transcription (CAGE[157] and 5′ SAGE[158]). Using CAGE on the mouse transcriptome they identified different transcription start sites for up to 27% of the transcriptional units profiled. Using 5′ SAGE it was shown there is a preference for adenosine as the first base of transcription (41%) and that similar to the alternative polyadenylation found at the 3′ end of transcripts there is heterogeneity at transcription start sites. Wei *et al.* modified LongSAGE to produce tags for both the transcription start and end sites[159]. They mapped these tags to the genome and predicted transcriptional units. They then used these tags as primers for RT-PCR and confirmed transcription of genes that had previously only been predicted.

D. Thesis Overview

As stated at the beginning of this introduction, the initial aim of this thesis was to characterise the cell surface of a resting CTL. As described above, an open tag-based technology is the only current technology class that may meet this need. Due to their increased tag length (and thus increased specificity) both LongSAGE and MPSS are

preferable to SAGE. However, when the strategy was being decided upon SAGE was the only technology that was ready to use and therefore was used to characterise a CTL clone (Chapter 3).

Having successfully used SAGE to characterise a CTL, it was decided to use the same methodology to analyse the NK cell surface (Chapter 4). In this chapter, tags that matched nothing ("no matches") in the "tag-to-gene" mapping were examined to discover if they were false negatives due to the technique or genuine.

For the remaining work, all three technologies were now available and a choice had to be made between them to fulfil each aim. Having examined the cell surface, the next aim was to examine the entire transcriptome of another lymphocyte, a $CD4^+$ T-cell (Chapter 5). To examine the entire transcriptome one needs to sample very deep (hundreds of thousands of tags), this is routine with MPSS and now feasible with SAGE due to reduced sequencing costs, so either technology could be used. However, as both LongSAGE and MPSS are limited by the fact that a small percentage of genes lack the required restriction site[108] both MPSS and LongSAGE were used to characterise a $CD4^+$ T-cell transcriptome. This also enabled a direct comparison of the two tag-based technologies.

To carry out this technique comparison one needs a reliable tag-to-gene mapping database. As no database is currently available that can match both SAGE and MPSS tags, one had to be developed. One of the key advantages of using LongSAGE or MPSS, which have longer tags than SAGE, is that the large majority of the tags only appear once in the genome. When taking advantage of the depth of

sampling provided by these techniques (effectively attempting to sample the entire transcriptome), one would expect to identify transcripts that have not been previously identified. Therefore, the ideal approach to integrate SAGE and MPSS data would be to use a tag-to-gene mapping database based on the genome. The key decision when producing a tag-to-gene database is what dataset to use for the source of the transcripts. If one uses all publicly deposited sequences there is a large amount of redundancy, therefore one should use an approach that groups the data. Unigene has been used in the publicly available mappings, however, I prefer the Ensembl Genes dataset[160]. There are two main reasons for this. Firstly, Ensembl genes are, in effect, defined by a consensus sequence, in that they are defined by genome co-ordinates. Secondly, Ensembl Gene Ids are stable, meaning that they stay the same between releases. Therefore a tag-to-gene mapping database was developed based on Ensembl and the principles established, in Chapter 4, by studying the tags for which it was not possible to map to transcripts using the publicly available tag-to-gene mapping databases.

As described above, there are many different classes of mRNA; using the tag-gene-mapping database developed in Chapter 5 it was possible to examine *cis*-NAT expression. In Chapter 6, the aim was to examine the extent of *cis*-NAT production in a range of cells. As there are many publicly available LongSAGE libraries (and no publicly available MPSS libraries), LongSAGE was used to examine antisense transcription.

II. Materials and Methods

A. Bench-based

1. Cellular samples

a) $CD8^+$ T-cell clone (clone 32)

This clone was also used by Edward Evans in his D.Phil thesis and the details are repeated here for completeness purposes[47]. The CTL clone used was grown by T. Dong. Briefly, the clone was generated from an HIV virus-specific CTL line grown from the peripheral blood mononuclear cells (PBMC) of an HIV-seronegative, highly exposed but apparently HIV-resistant woman from Nairobi, Kenya. The CTL line was an early antigen-specific line, stimulated directly with epitope peptide (the HLA A68 restricted peptide ETAYFILKL) as described[161]. The clone was generated by limiting dilution at 0.3 cells/well in the presence of 100µl "cloning mix" (10^5/ml peptide-pulsed autologous irradiated B lymphoblastoid cell-lines and 10^6/ml mixed allogeneic irradiated (4000rads) PBMCs from three donors, in RPMI 1640 (GibcoBRL, Paisley, UK) supplemented with 10% FCS and 5µg/ml phytohaemaglutinin). Clones were plated out after two weeks in a further 1ml of "cloning mix" and were subsequently maintained by weekly restimulation with mixed, heterologous, irradiated, PHA-activated PBMCs, and grown in medium supplemented with IL-2 as 10% Lymphocult-T (Biotest, Solihull, UK).

RNA was extracted from $2x10^8$ clone 32 CTL four weeks after their last stimulation with PBMC (during which time they were cultured in IL-2 containing medium as above but in the absence of antigen). Total cellular RNA was extracted from CTL by

suspension of cells in TRIzol. 1ml of TRIzol was added per $7.5x10^6$ cells. The mixture was vortexed to ensure cell lysis and incubated at room temperature for 5-10 minutes. 0.2ml chloroform was added per 1ml of TRIzol; the tubes were shaken, incubated for 2-3 minutes at room temperature and centrifuged at 3000rpm (1500g) for 30 minutes at 4°C. RNA was extracted from the aqueous (upper) phase by addition of 0.5ml propan-2-ol per 1ml TRIzol initially used. The RNA was pelleted by centrifugation at 3000rpm (1500g) for 30 minutes at 4°C, washed in 75% ethanol and dried. FACS analysis of the CTL at the time of RNA extraction was performed by T. Dong on a FACScan machine (Becton Dickinson Biosciences, Erembodegem, Belgium).

b) $CD4^+$ T-cell clone (clone 29)

Clone 29 was grown by Julian Sutton for his D.Phil thesis[162]. PBMC from a vaccinee were established in culture with IL-7 and two gag 15mer peptides (A and C)[163]. The clone was generated by limiting dilution at 0.5, 1 and 2 cells/well in the presence of 100µl "cloning mix" (10^6/ml mixed irradiated (300rads) PBMCs from three healthy donors with 10% human AB serum and 90 µg/ml phytohaemagglutinin (Murex Biotech Limited, UK)). At day 3, 100µl of IL-2 (200U/ml) in RPMI with 10% heat inactivated human AB+ serum was added and was maintained at this concentration thereafter. The cells were split and medium replaced from day 4 onwards as required. Clones were restimulated with cloning mix every 20-25 days with the addition of il-2 (200U/ml) at day 3 in the cycle. Cells were split into new wells of 24-well plates or 20ml flasks and medium replaced as required. The activated cells were stimulated for 16 hrs, in the absence of APCs, by PMA (50ng/ml) and Ionomycin (1µg/ml), brefeldin A was added after 2 hours.

16ug of total RNA was extracted from $9.4x10^6$ cells of the resting $CD4^+$ T-cell clone 29 and 120ug of total RNA was extracted from $10.8x10^6$ cells of the activated $CD4^+$ T-cell clone 29. The total RNA was prepared using TRIzol reagent method (GibcoBRL). The same method was followed for both the resting and activated $CD4^+$ T-cell clones. The cells were counted and centrifuged at 1,300rpm for 5 minutes at room temperature. The cell pellet was resuspended in 2ml of TRIzol reagent. The mixture was pipetted up and down 5 times until the cells were lysed and then stored at -80^{o}C for ~10 weeks. The cells-TRIzol mix was thawed at room temperature for 5 minutes and then split into 2 aliquots of 1ml. To each aliquot 0.2ml of chloroform was added and the mixture was shaken vigorously by hand for 15 seconds, then incubated at room temperature for 2 minutes. After this incubation, the mixture was transferred to a phase-lock tube (Eppendorf, Cambridge, UK) and centrifuged at 13,000rpm for 1 minute at room temperature. The upper aqueous layer was kept and mixed with 0.5ml of isopropanol. The mixture was shaken vigorously by hand for 15 seconds and incubated at room temperature for 10 minutes before being centrifuged at 13,000rpm for 30 minutes at 4^{o}C. The resulting total RNA pellet was washed twice with 500µl of 75% ethanol then stored in 75% ethanol at -20^{o}C for 5 months. The ethanol was discarded and the total RNA pellet air dried for 5 minutes and resuspended in a total volume of 10µl of double processed tissue culture water (Sigma, Dorset, UK). The sample was diluted 100-fold in water and an absorbance reading was taken at 260 nm and 260/280 nm to estimate the concentration and quality of the RNA. A sample of the total RNA was also run on a 1% agarose analytical gel to show that there was no degradation of the RNA. The total RNA was aliquoted appropriately and stored at -20^{o}C.

c) NK cell line

The NK cell line was grown by Veronique Braud and Christelle Retiere. Monocyte depleted peripheral blood lymphocytes were depleted of $CD3^+$ cells using MACS beads, stimulated with irradiated fresh PBMC and allogeneic B cells and cultured in the presence of IL-2 with three further rounds of CD3 depletion. Four weeks after stimulation of these cells, RNA was extracted from $2.1x10^8$ NK, as described for clone 32.

2. SAGE library production

a) SAGE libraries – NK and CTL

The CTL clone 32 SAGE library was produced by Lisa Sparks using the original SAGE protocol[94]. The NK cell SAGE library was also generated using the same methods by Lisa Sparks and sequenced by Raquel Manso-Sancho and myself.

b) LongSAGE libraries – $CD4^+$ activated and resting

The $CD4^+$ T-cell activated and resting LongSAGE libraries were produced by Mai Vuong. The Long SAGE library was produced using the I-SAGE long kit version B (Invitrogen, UK). Two activated libraries were made: the first from 16µg total RNA and the second using 30µg total RNA as starting material. The vast majority of the sequenced tags were from the 16µg library. The resting $CD4^+$ T-cell clone 29 library was made using 16µg of total RNA.

3. MPSS library production

The MPSS library was produced by Lynx Therapeutics Inc. from 30µg of total RNA. The protocol used is essentially that as described by Brenner *et al.*[93].

4. Extension of "no match" SAGE tags

Two techniques were used to extend the NK SAGE tags for which no corresponding transcripts were found in SAGEmap. GLGI was performed for all 61 tags using the original protocol[164]. The only major difference being that the GLGI library was not made at the same time as the SAGE library, therefore random sampling events may mean that certain transcripts that were sampled in the initial SAGE library were not present in the GLGI library.

For the majority of tags where GLGI produced no PCR products 3′ RACE was used to extend the tag length[165]. The Invitrogen Gene Racer kit was used (Invitrogen). The 5′ RACE primers were designed by matching all tags to the genome. The tags were matched to the genome using BLAST at NCBI accessed with Perl. All exact matches for each tag were kept, and 21 base pair primers were designed by extending the match 3′ of the tag by 7 bases.

B. In silico analysis

1. SAGE

Unless otherwise stated all data was processed using a combination of Microsoft Access, Excel and Visual Basic for Applications.

a) Tag extraction

The SAGE 2000 program (obtained from K. Kinzler) was used to identify ditags from the concatamer sequences, exclude duplicate ditags (which are likely to have arisen from PCR bias) and produce a list of unique 10 bp tag sequences together with their frequency of observation (tag count). Tags that may be derived from linker sequence

were excluded - a list of such tags is provided with the SAGE program group software. For some SAGE tags it was possible to extract an 11th base. For comparisons to other libraries, for which only 10 bases are available for each tag, only the first ten bases of tags in our library were used.

Public SAGE databases were downloaded from SAGEmap and from the Matsushima group at the University of Tokyo (Blood SAGE[143] at www.prevent.m.u-tokyo.ac.jp/sage) as text files. These were imported into Microsoft Access and treated in the same way as the libraries generated in house.

b) Tag-to-gene mapping

Tag files were linked to the SAGEmap Tag to UniGene Mapper[166]. For all tags of interest this data was manually curated. Apparent Unigene matches were checked to ensure that the tag did not match only a single EST, only 5′ ESTs (i.e. tag was not truly at a 3′ most *Nla*III site), ESTs which had been incorrectly clustered or those which appeared to contain single base errors compared to the true transcript (i.e. that represented by the mRNAs and the majority of ESTs). Where possible, matches were also checked to ensure that they had the correct 11th base.

Once transcripts had been matched to tags, they were categorised using two systems - one based on current knowledge of the transcript function in the literature and online databases and a second based on the probable molecular function of the transcript.

The categories used for these two systems were as follows:

Knowledge:

Code	Classification
K	Has a **K**nown immune function
O	Has **O**ther known function
H	**H**ypothetical protein / function unknown
E	Cluster contains **E**STs only
N	**N**o true matches to this tag
xs	Multiple true matches to this tag

Function:

Code	Classification
E	Secreted **E**ffector molecules
CS	**C**ell **s**urface molecules (not MHC molecules)
S	**S**ignalling/adaptor molecules
AP	**A**ntigen **P**resentation & MHC molecules
T	**T**ranscriptional regulation
Cy	**Cy**toskeleton & vesicle transport related
CC	**C**ell **C**ycle / viability / apoptosis related
P	**P**rotein and mRNA synthesis related
O	**O**ther (incl. Housekeeping)
U	**U**nclassified transcript of known sequence
Est	Tag matches one or more **E**STs only
N	**N**o true matches to this tag
xs	Multiple true matches to this tag

Transcripts were assigned to these categories based on their Unigene[110] descriptions, Locuslink and RefSeq[167] entries (where available) and information from publications.

Uncharacterised but sequenced transcripts were analysed using BLAST[168] at www.ncbi.nlm.nih.gov/blast. BLASTp was used if a predicted amino acid sequence was given and BLASTx (with the DNA sequence) was used if not. If no amino acid sequence was given, the longest forward-reading ORF was also identified, using the NCBI ORF Finder tool (www.ncbi.nlm.nih.gov/gorf). The given amino acid sequence or translated ORF was used to search databases of known domains and profiles (Interpro[43] and SMART[38], including prediction of signal peptides and transmembrane domains with SignalP[169] and TMHMM2 [ref:[170]] respectively). These results were used to screen the uncharacterised transcripts for possible cell surface or signalling molecules, transcription factors or regulators of the cytoskeleton.

For tags matching Unigene clusters containing only ESTs, the clusters were assembled using CAP3 (ref:[171]) hosted at http://bio.ifom-firc.it/ASSEMBLY/assemble.html. The resulting contigs were used in BLASTn and

BLASTx searches at NCBI to identify possible homologues. Matching transcripts were then analysed as for uncharacterised but sequenced transcripts.

Tags that matched no known transcripts using the above methods were characterised further to examine three aspects: internally derived tags, tags caused by sequencing errors and tags due to antisense transcription. To examine internally derived tags (due to either internal priming or incomplete cleavage), the SAGE Genie flat files were downloaded and the database reassembled in Access. The "no match" tags were then matched to the best-gene match in SAGE Genie. This is the highest quality match for each tag found in the SAGE Genie database. Information was kept for those tags that were either classed as "internal" or "internally primed".

Potential sequencing errors were highlighted using the following algorithm developed in collaboration with Peter Collingridge (a summer student in the laboratory). All tags in a library were sorted by descending abundance. Working down the list, each tag was compared to each tag beneath it in the list, if a tag was a daughter (a possible 1 base sequencing error, due to insertion, deletion or mismatch) then the tag and count of the parent was marked next to the daughter. "No match" tags were examined and any tag found to have a parent with a count of greater than ten fold that of the "no match" tag was deemed to be a possible sequencing error.

Cis-NAT antisense transcription was examined by searching RefSeq. All transcripts were obtained from RefSeq (October 2004) and tags were extracted from both the forward and reverse strands. The "no match" tags were compared to the tags

extracted from RefSeq. Those tags that matched tags only found in RefSeq in the antisense direction were classed as antisense tags.

c) Gene-to-tag mapping, for genes encoding cell surface proteins

Two sets of Gene-to-tag mapping were undertaken: a manually curated set and a partially automated set. The manually curated set was produced for a list of genes encoding cell surface proteins as chosen at theHLDA Workshops. The list contained 247 CD antigen clusters defined at the 7th HLDA Workshop (except those containing carbohydrate or uncharacterised antigens) and genes that were to be considered for CD status at the 8th HLDA workshop (www.hlda8.org/PotentialCDs.htm). SAGEmap Gene to Tag Mapper was used to provide the initial tags and then matches were manually curated as for tag-to-gene mapping. Genes were excluded if one of their tags was found to match several, unrelated transcripts. Combining these lists gave a total of 374 transcripts that could be assigned SAGE tags (see Appendix A for the full list of transcripts). One should note that this process may produce more than one tag for each transcript.

The semi-automated set was produced using SAGE Genie mappings. UniGene IDs were downloaded from the Genome Ontology[172] (GO) database at CGAP for potential plasma membrane proteins (from GO categories: 0016021 integral to membrane and 0005886 plasma membrane). Each gene in the list was manually checked (using Locuslink entries, where available, and information from publications) and those corresponding to confirmed cell surface proteins were retained. The UniGene IDs were then matched to those in SAGE Genie and, where possible, the “best tag” was assigned to the UniGene. If the tag was found to match more than one UniGene cluster in the dataset, then only one cluster with the tag was used for further analysis.

d) Inter library comparisons – identifying "immune specific" tags

To identify a subset of tags that were thought to correspond to proteins with an important immunological function it was necessary to identify tags that were up-regulated in our libraries. There are a variety of statistical tests available for identifying differential expression in tag-based experiments. Romualdi *et al*. compared various statistical tests for detecting differential gene expression and showed that a test developed by Audic and Claverie was the most efficient for pairwise comparisons[173]. This test developed by Audic and Claverie (hereafter referred to as the AC test) was initially used for detecting differential expression in EST libraries, but can equally be applied to SAGE data[174]. For the pairwise comparisons this statistic was calculated in Excel using Visual Basic. For the highly observed tags it was not possible to calculate this statistic (combined count of >140), as the factorials involved were too large. Therefore for these tags a different test was required. The 2x2 χ^2 test was used and is suitable because the high expression level of the tags mean the counts in each library will be well above the minimum of 5 required for accurate calculation of probabilities (it will be at least 70). This statistic was also calculated in Excel using Visual Basic, the p-value was then calculated using the χ^2 distribution function built into Excel. The abundances of those tags that were found to be significantly more abundant against both cerebellum and ovary epithelium were then compared to those in a panel of 12 cancer libraries. The panel of cancer libraries consisted of libraries derived from five brain tumours (astrocytoma, glioblastoma, medulloblastoma, oligodenroglioma and recurrent ependyoma), a breast ductal carcinoma, a colon adenocarcinoma, a gastroesophageal junction adenocarcinoma, a pancreas adenocarcinoma, a prostate tumour (including stroma and epithelium) and two ovarian cancers (carcinoma and cystadenoma). If

the tag was found at more than a third of the level in the immune library in less than two libraries in the panel, then the tag was deemed to be specific.

2. LongSAGE/MPSS

The computational aim of the LongSAGE/MPSS analysis was to make it platform independent and remotely accessible. To meet this aim the tag extractions were performed using Perl and the Perl API at Ensembl. The data was stored locally in a MySQL database on a computer set-up as a remotely accessible MySQL server. Data was primarily processed using Perl and the DBI module for database interactions. Most of the data was stored as flat tables to simplify error checking and optimise the speed of access.

a) Tag extractions

(i) Libraries

Sequencing runs from the LongSAGE library were initially processed with Phred to remove sequencing errors[175]. When choosing a Phred setting one has to balance the need for high quality sequence against that of losing genuine sequences due to false positives. Analysis of sequences of a similar length to SAGE ditags led Prosdocimi *et al*. to conclude that low Phred settings allowed the optimal ratio of genuine sequences retained to errors removed[176]. Therefore for this analysis, sequences with Phred scores of 10 and above were kept. The Phred screening and tag extraction from ditags was done by Perl scripts written by A.G. McArthur (Marine Biological Laboratory, Woods Hole, USA). Tags were imported into a MySQL database and tags possibly derived from linker sequence were removed (the list of linker sequence

derived tags was courtesy of NCBI). Tag counts were then normalised to tags per million (tpm).

Public LongSAGE libraries were downloaded from GEO[177] and imported into the MySQL database using the same normalisation and linker removal script as above.

(ii) Ensembl data

There were two classes of extractions necessary for the tag-to-gene mapping. Tags had to be extracted from both the genome and transcriptome. Ensembl (version 29, human genome sequence 35b) was used as the data source[178].

Tags were extracted from all chromosomes and mitochondrial DNA of the unmasked genome at all restriction sites (of *Nla*III and *Dpn*II). Further information was extracted for each tag. Three windows were examined, both up and downstream of the tag, for the presence of gene annotation (the three windows were: at the restriction site, end of restriction site-1000 bases and 1001-5000 bases). For these windows all Ensembl genes and predicted genes were recorded. The tag matches to genes were classed as: outside of a gene (default), exonic, intronic, boundary (i.e. the tag overlaps the exon intron boundary) and multiple genes (this was only assigned if the tag was found to match multiple genes and the gene matches were different classes).

Tags were extracted from all transcripts in the Ensembl Genes dataset[160] at all restriction sites (of *Nla*III and *Dpn*II) in both the sense and antisense direction. If tags overlapped the 3′ end of a transcript the tag was extended along the genome unless the transcript was predicted to contain a polyA site (as defined in the supplementary material by Caron *et al.*[153]), then adenosines were added to the 3′ end

of the tag to complete the length of the tag. This was repeated for the ESTGenes dataset[179].

b) Tag-to-gene mapping

Information from the extractions described above was combined for the tag-to-gene-mapping. The first stage was to calculate frequencies for each tag, both in the genome and transcriptome. Each tag was matched to the genome and depending on the number of hits to the genome classed as one of the following: a single match, multiple match, no match or excess matches (more than 20 hits to the genome). No further analysis was undertaken for the excess matches. For the single and multiple matches the annotations for each tag found in the genome were examined and if genes were annotated to the tag, the tags were classed as one of the following: single antisense, single sense, downstream of single antisense, downstream of single sense and multiple genes. For all tags, apart from those that were found to match multiple times to the genome and have multiple gene annotations, the Ensembl Gene and Ensembl ESTGene extractions were then examined. The resulting matches were classed as single antisense, single sense or multiple matches, and then compared to the transcripts identified in the genome matches. It was noted if the gene annotation for the genome matches were different to the transcriptome extractions. The result of this mapping process is a list of tags and for each tag all the matches identified are returned along with a four part code. This code can then be used to assess the quality of the tag-to-gene match.

c) Gene-to-tag mapping

To examine either *cis*-NAT expression or the correlation between the different tag-based technologies it is necessary to compare transcripts for which all the potential

tags can be uniquely assigned to both the transcriptome and genome. The extractions and frequency information described above were used to reduce the Ensembl transcripts to a set with unique tags. To compare MPSS and LongSAGE, transcripts were kept if they met the following criteria. The transcripts had to have at least one restriction site for both enzymes (*Dpn*II and *Nla*III). All forward and reverse tags (from both types of restriction site) had to be found only once in the transcriptome and at most once in the genome (some tags may absent from the genome due to splicing, polyadenylation and the fact the genome is not complete). This set was termed UTBS dataset (Unique Transcripts Both Sites) and consisted of 5857 transcripts. For the antisense analysis only LongSAGE data was used, so the transcripts only needed to have an *Nla*III restriction site and the corresponding forward and reverse tags (extracted from the restriction site) had to be found only once in the transcriptome and at most once in the genome. This produced a dataset known as the UTLST dataset (Unique Transcripts LongSAGE Tags) which consisted of 8954 transcripts.

d) Altering library size

Many of the experiments required libraries to be of a different size to the sequenced library. These libraries of different size were produced using a script in Perl. All tags from a library were placed into an array, the array was randomised and then desired number of tags was taken from this array. The process was repeated for the number of data points required.

e) Identification of potentially novel regions of transcription

Potentially novel regions of transcription were identified in two ways, firstly using a combination of LongSAGE and MPSS data and secondly using only LongSAGE data

where nearby transcription was observed on opposite strands. The basic principle was the same for both experiments. Tags were selected that matched the genome only once and were not found in the transcriptome or downstream of any known transcripts (<5000 bases). Tags were then kept if there was another tag within a certain distance on the genome. For the LongSAGE/MPSS experiment the criteria was that tag was within either 5000 bases or five recognition site sequences of a tag from the other technique. These pairs were then screened to remove tags that were found within the ESTGene dataset or <5000 bases downstream of an ESTGene. For the antisense experiments, tags were kept if both the forward and reverse tag at any one loci were observed.

f) Number of transcriptional loci

In tag-based expression experiments one gene may be represented by multiple tags. This may be due to be biological reasons (e.g., alternative splicing and alternative polyadenylation) or technical reasons (e.g., incomplete cleavage or internally primed cDNA synthesis). To establish the number of genes (or transcriptional loci) one must group the tags found in similar locations. Tags from the expression library were matched to the genome and those that were found only once in the genome were grouped. Two criteria were used to group the tags. The first criterion is the maximum distance between the first and last observed tag in the group (this was set at between 50,000 - 200,000 bases in 50,000 base steps). The second criterion is the maximum distance between any two consecutive tags (this was set at between 5,000 - 20,000 bases in 5,000 base steps). The strand that the tags matched to was ignored to prevent an artificial increase in loci due to *cis*-NATs. The resulting figure for the number of loci then needed to be corrected to account for the tags excluded because they did not match uniquely to the genome. The correction factor was calculated by

dividing the number of unique tags in the library by the number of unique tags in the library that uniquely matched the genome. This will be a slight underestimate, as some of the observed tags that match the genome in more than one location will be derived from more than one transcript.

g) Inter-library comparisons

The UTBS dataset was used to compare LongSAGE and MPSS produced from different tissues and activation states. Spearman correlation coefficients were calculated for natural log transformed data using the statistical analysis program, R (available from http://www.r-project.org)[180]. The data was log transformed to reduce the effects of heteroscadicity on the correlations. Spearman correlation coefficients are suitable for examining large scale gene expression experiments because the calculation uses rank data rather than absolute values and is therefore not influenced by outliers[181].

h) Basal level of transcription

Before examining *cis*-NAT expression it was necessary to check that antisense transcription occurs at a different rate to the basal level of transcription. To examine the basal level of transcription expression in the $CD4^+$ activated library was examined on the positive strand of Chromosome 22. Only those tags that appear once in the genome were examined. A sliding window based approach was used, the size of the window was set at 10 *Nla*III restriction sites. The windows were classed as being either sense, antisense or outside known genes depending on the class of transcripts which aligned to the windows. Priority was given to sense matches, then antisense and finally outside known genes. The average expression for each class of window was then calculated to give a basal transcription level.

i) Antisense – external data comparisons

To evaluate the quality of the *cis*-NAT detection method proposed here it was necessary to compare the analysis to other published analyses. For the chosen comparative publications the lists of sense and antisense transcripts ("original transcripts") were downloaded. Only "original transcript" pairs which fulfilled the following criteria were kept. Both "original transcripts" had to match the genome, at most for one Ensembl transcript, and that match had to be >99% identity. If at least one "original transcript" in the pair matched an Ensembl transcript, the pair was kept. If the "original transcripts" matched different Ensembl transcripts, the pair was discarded. Finally, only those "original transcript" pairs which matched to an Ensembl transcript contained in the UTLST dataset were kept.

j) Antisense – positional information

Positional information was required to examine two facets of *cis*-NAT expression. Firstly it was used to see if there was a correlation between sense and antisense tags at the same restriction site, a high level of correlation would indicate that this antisense expression may be due to a technical artefact. Secondly, it was used to examine the relative frequency of the use of different restriction sites. One script was written to provide the data for both analyses. The script analysed the count of tags in the $CD4^+$ activated LongSAGE library that corresponded to the tags in the UTLST dataset. For each restriction site, the counts for both the forward and reverse tags were outputted to enable an analysis of the correlation between forward and reverse tags at the same site. In the final analysis, tags from the 3' most end were removed, to eliminate bias due to transcripts with only one restriction site.

To examine the location of the approximate 3′ end of the *cis*-NATs only transcripts that contained a polyA site (as defined by Caron *et al*.[153]) were used, as one can have a higher confidence that these transcripts are full length. To combine the results of multiple transcripts found over a wide range of expression levels it was necessary to normalise the data. Transcripts were grouped by the number of *Nla*III restriction sites. Then for each transcript, the counts for all tags observed in one direction were summed and the counts converted into a fraction of this sum. For each position the normalised counts were summed for all transcripts with a set number of restriction sites and then divided by the number of transcripts observed with the set number of restriction sites. For each position this gave a percentage occupancy that was independent of the level of expression of the individual transcripts.

III. CD8+ T-cell surface

A. Introduction

When trying to understand how any system works it is important to appreciate the scope of current knowledge. If all the key components of a system are known, then sensible models of the how system works can be designed. Conversely, if the key components are not known, then it is unlikely that functional models will be correct. In the immune system, many models have been generated to explain T-cell activation (e.g. conformational change[182], aggregation[183] and kinetic segregation[9]). One of these contrasting models may well be correct. However, until all the cell surface proteins that are involved in this process are known, it is not possible to know that a model is correct. Therefore, in the work presented here I am going to focus on identifying the proteins found at the T-cell surface using a transcriptome based approach, with the aim of gauging the scope of current knowledge.

The two key requirements for any technique to be able to catalogue a list of proteins expressed in a cell are: firstly that no prior knowledge is necessary about the protein structure and secondly, that the technique can discriminate between thousands of proteins. Ideally, one would use a proteomic based approach. However, even the latest technologies can only deal with a limited number of proteins (e.g., a recent study using "shotgun proteomics" managed to characterise only 5,130 proteins[85]), have problems with integral membrane proteins and generating quantitative results is not routine[86]. If it is not possible to examine the proteins directly then one should examine gene expression. The various types of global gene expression methods were compared in Chapter 1. In 1997 when the technique was being selected, it was clear

that SAGE, due its potential for deep sampling and that it is a truly open expression technology, was the most appropriate method.

Having chosen an appropriate technique the next important step is to choose an appropriate sample. A human $CD8^+$ cytotoxic T-cell clone (clone 32, a gift from T. Dong) was chosen to study. This particular clone was chosen as it is a simple, effective killer, acting in a CD8-independent manner. Using a clone in global gene expression studies is desirable, as it provides a pure sample. This means that the results are not biased due any un-quantified heterogeneity in the starting sample. This also, in effect, means that one can sample deeper as there will be a smaller pool of transcripts to identify.

Having chosen the technology and the sample to be used, there were three aims of this work. Firstly, it was necessary to confirm that the combination of SAGE and a clonal sample was appropriate for identifying the functionally important "immune specific" cell surface molecules. The second aim was to catalogue all the "immune specific" cell surface proteins on the T-cell surface. Finally, the aim was to use a partially automated approach to examine the expression of transcripts encoding all the known cell surface proteins in a variety of libraries. The expression of the immune specific cell surface proteins can then be analysed in the context of all proteins on the cell surface.

The work presented below was carried out in partnership with Edward Evans and is a continuation of work that was presented in his doctorate thesis[47].

B. Results and Discussion

1. Immune cell surface

As described in the materials and methods chapter, a SAGE library was constructed from a sample of RNA from a resting CD8$^+$ T-cell clone (clone 32). After the exclusion of linker tags, a total of 63,270 tags were sequenced, representing 20,204 unique tags.

a) Library purity and depth

To examine whether one has sampled enough tags, one can look at the rate of addition of novel tags to the library. Once all available tags have been sequenced, the rate of addition of novel tags should be zero. One would expect a curve of total tags sequenced versus unique tags to become asymptotic if all transcripts have been sampled. As expected for a library of this size, Figure 3.1 shows the library has not reached saturation. This suggests that the lower expressed transcripts may not be detected in this library.

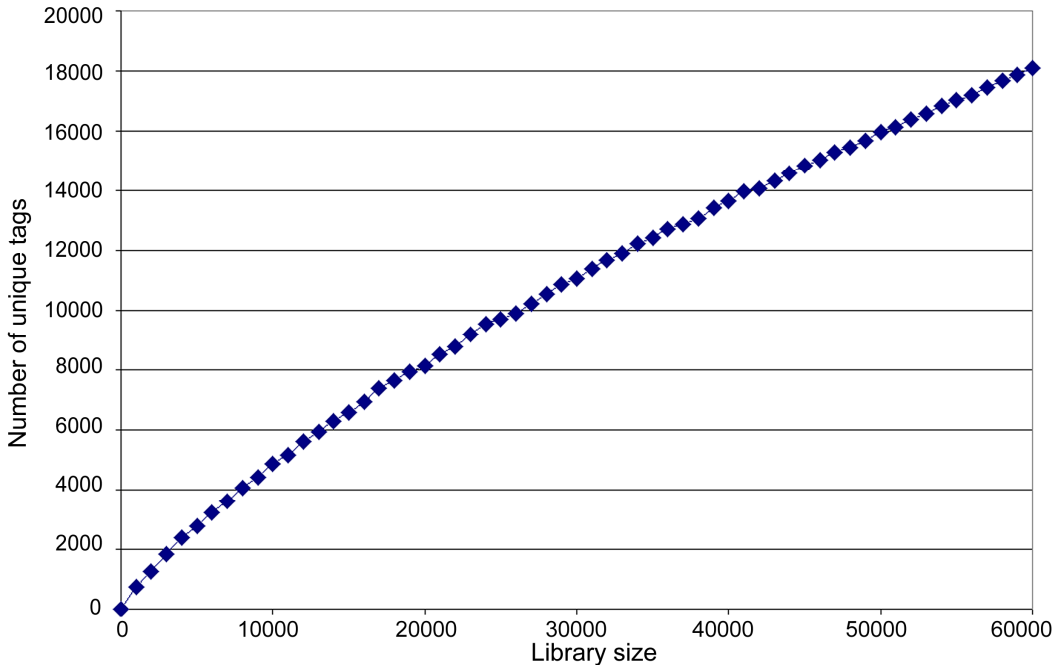


Figure 3.1: Effect of total number of tags sequenced on number of unique tags identified.

The clone 32 SAGE library was sampled at various sizes to examine the effect of library size on the number of unique tags identified. If the library is large enough to sample all available tags, then increasing the library size will not increase the number of transcripts detected.

To examine both the purity of the library and the impact of the lack of saturation of the library it is informative to search the library for tags corresponding to known cell surface proteins. To use a systematic approach for identifying known cell surface proteins, rather than one biased by personal preferences, the list of proteins searched for was based on those defined at the HLDA Workshops. SAGE tags for all 247 CD antigen clusters defined at the 7th HLDA Workshop[184], except those containing carbohydrate or uncharacterised antigens, were obtained using the SAGEmap Unigene-to-Tag mapping database[166], with manual curation of the matches where necessary. This method may produce more than one tag for each transcript. CD antigens for which a tag matched several, unrelated transcripts were excluded from further analysis, as these tags are likely to be found in many libraries and may represent a different transcript in each case. Tags specific for T-cell receptor transcripts and for the genes that were to be considered for CD status at the 8th HLDA workshop (www.hlda8.org/PotentialCDs.htm) were also included. Combining these lists gave a total of 374 transcripts that could be assigned SAGE tags (see Appendix A for the full list of transcripts).

Of 374 transcripts encoding cell surface molecules that were examined in this way, 111 were present in the clone 32 library, at levels of ~1.5 to almost 200 tags per 100,000. This set of transcripts included all of the principal T-cell markers, i.e. each of the TCR/CD3 components, CD2, CD5, CD6, CD8, CD11a (LFA-1α), CD43, CD45 and CD53. CD18 tags were found at high levels in the library, but one of these tags matched several other genes, preventing abundance determination. CD28 was absent and this is generally characteristic of antigen-experienced human

cytotoxic T-cells, and particularly of CTL clones. Transcripts encoding CD150 (SLAM), CD152 (CTLA-4), CD154 (CD40L) and ICOS were absent or only weakly expressed, consistent with the resting phenotype of the clone. Using this list of 374 genes, tags corresponding to transcripts whose proteins were detected by FACS analysis were also examined. (FACS analysis was performed by Tao Dong.) Tags corresponding to all proteins detected by FACS were identified by SAGE, with the exception of CD32. This may be due to random sampling effects or that an alternatively spliced/polyadenylated form of CD32 was expressed in clone 32. The fact that all key T-cell markers and the large majority of markers detected by FACS were also found in the SAGE library suggests that this library is deep enough to be used to identify the expression of transcripts encoding "immune specific" cell surface proteins.

The purity of the library was checked by examining the expression of tags corresponding to lineage marker for B cells (e.g. CD19, CD20, CD21, CD22) and myeloid cells (e.g. CD14, CD32, CD33). All these tags were absent, confirming that the library was free from significant numbers of feeder cell-derived transcripts.

To check that clone 32 is representative of CTLs, the library produced here was compared to a previously published library generated for purified CD8$^+$ T-cells[146]. Forty-nine of the 50 most abundant tags in the library generated from *ex vivo* purified CD8$^+$ T-cells are among the 2.5% most abundant tags in the clone 32 library; the only tag that is absent corresponds to an MHC class I allele, which is replaced by a tag from an alternative allele with a single base substitution. This, along with the

presence of the expected T-cell markers, suggests that the clonal library presented here is representative of CTLs.

Having established that this library was both pure and deep enough to detect known cell surface proteins, and that the clone was representative of CTLs the next aim was identify all genes corresponding to "immune specific" cell surface proteins expressed in the clone. Ideally, each tag would be individually matched to a gene and then this match would be manually checked. This is not feasible for a library of over 20,000 unique tags. However, if a subset of tags could be identified that was enriched in tags corresponding to the immune cell surface proteins, then this set could be searched for tags corresponding to novel cell surface proteins.

b) Identifying an 'immune specific' subset

There are various possible strategies that could be used to identify a subset of tags enriched with those corresponding to transcripts with an immune function. One such strategy is to look at the most highly expressed tags. However, over half of the top 100 tags correspond to transcripts involved in protein synthesis[47]. Therefore, a different strategy was required to isolate tags corresponding to transcripts with an immune function. As described in detail in Chapter 2, a 3-stage filter process was used. This involved using a series of pair-wise comparisons to SAGE libraries unrelated to the clone 32 library.

The initial filter process proposed by Edward Evans compared the clone 32 library to a colon library, an ovary epithelium library and a panel of cancer libraries. It was shown that this filter successfully identified a subset of transcripts enriched for those with an immune function. The logic behind the different stages was as follows. To

identify transcripts that are up-regulated in an immune cell, the library should be compared to a library from a non-immune tissue. Colon was chosen as there was a publicly available library derived from a normal colon sample and the library was a suitable size. Tags that were significantly over expressed (p-value <0.05) in the CTL were selected. The properties of the proteins corresponding to the tags were examined and it was found that the percentage of transcripts with a protein synthesis function had been reduced when compared to the top 100 list. However, ~26% of those tags that matched to transcripts with a known function were found to be involved in protein synthesis, implying that this initial pairwise comparison reflected, to a significant degree, the non-proliferative properties of colon, rather than cell type-specific differences. To reduce the number of tags corresponding to protein synthesis proteins the clone 32 library was then compared to a cell expected to be proliferating (ovary epithelium). Only those tags that were significantly (p-value <0.05) more abundant in clone 32 than ovary epithelium were kept. A final filter involved comparisons with a panel of 12 different tumour-derived SAGE libraries[185]. It was reasoned that transcripts present in two or more of these libraries, at more than one-third their level in the clone 32 library, are more likely to be linked to cell proliferation than any immune-specific function. After the final filter, tags corresponding to transcripts involved in protein synthesis were reduced to ~4% of the tags selected.

Since the 3-stage filter process was designed additional SAGE libraries have become publicly available. One of these libraries was derived from cerebellum. As cerebellum is unlikely to contain a significant number of leukocytes, it was decided to update the filter and replace the comparison with the colon library, which may be

contaminated with blood cells, with a comparison to cerebellum. This change increased the number of tags corresponding to cell surface proteins with a known immune function that were found in the final subset of tags by ~15%.

The modified 3-stage filter process was carried out on the clone 32 library. The effects of the first two stages of the filter (i.e. the comparison to cerebellum and ovary epithelium) are shown in Figure 3.2. The fraction of tags that, at the 0.95 confidence level, are significantly over-represented in the T-cell library versus both the cerebellum and ovary libraries (3.75%), is higher than that versus each non T-cell library alone (1.68% and 1.69%). This implies that transcripts corresponding to tags selected by the filter are likely to constitute a differentially expressed subset.

To verify the robustness of the 3-stage filter process the tags were matched to transcripts which were categorised at each stage of the process. These transcripts were categorised according to the function of the encoded protein, where known, using RefSeq[167], Unigene[110] and LocusLink[167] annotations, or the literature. At each stage of the process, the proteins corresponding to tags upregulated in the clone 32 library were categorised as having (1) a known immune function, (2) some other known function, (3) known sequence but no characterised function (fully sequenced cDNAs or hypothetical proteins), (4) incomplete known sequence (i.e. from ESTs only), (5) no matches to Unigene or (6) multiple matches. Those proteins that had a known function were then classed as belonging to one of the following functional categories: (1) soluble effector molecules, (2) cell surface molecules, (3) signalling molecules, (4) antigen presentation, (5) transcriptional regulation, (6) cytoskeleton

related, (7) cell cycle or viability related, (8) protein synthesis related and (9) other miscellaneous functions, including housekeeping roles.

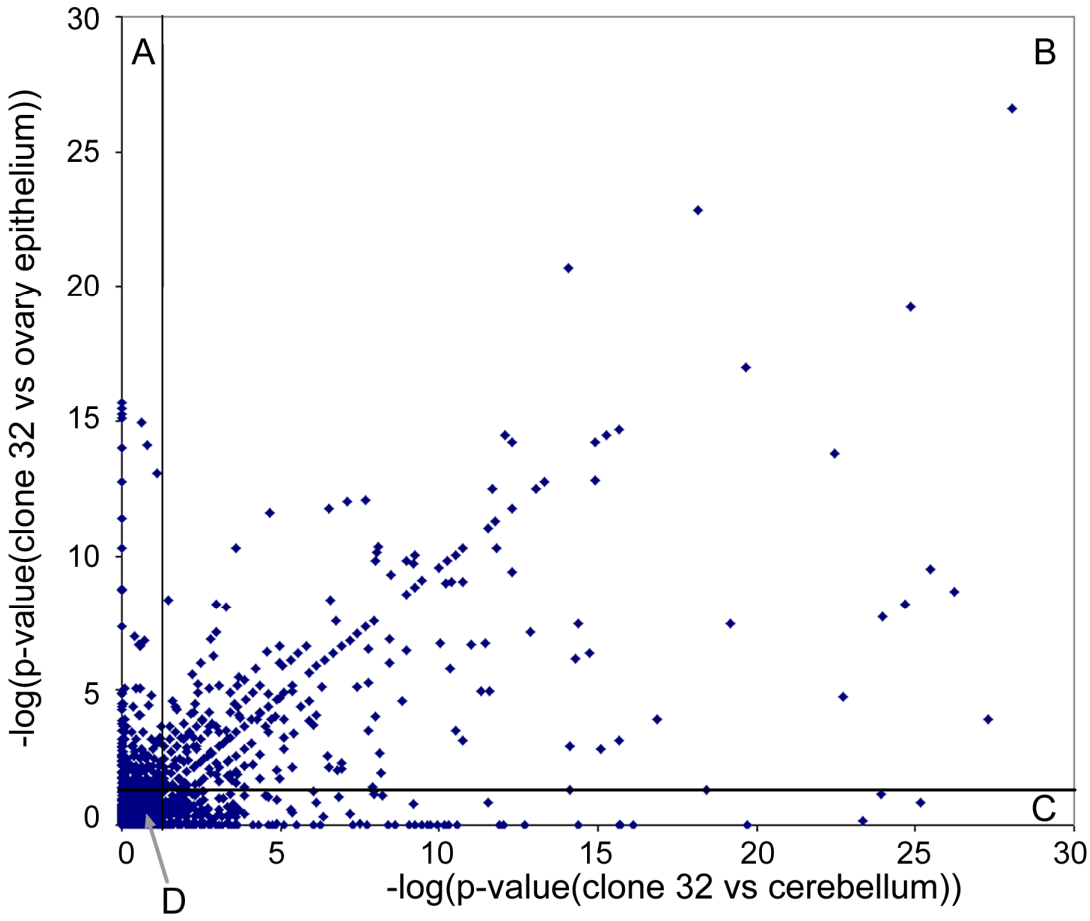


Figure 3.2: Distribution of CTL-specific tags.

As part of the 3-stage filter process the clone 32 library was compared to libraries from cerebellum and ovary epithelium. P-values of 5% were selected and the above figure shows the distribution of the upregulated tags. The lines marked in black represent the negative log of a 5% p-value. Quadrant A contains the tags specific against ovary epithelium only (1.69% of unique tags in the clone 32 library). Quadrant B contains the tags that are specific against both cerebellum and ovary epithelium (3.75%). Quadrant C contains the tags that are specific against cerebellum only (1.68%). Quadrant D contains the large majority of tags which are not specific against either cerebellum or ovary epithelium (92.88%). Thirty one outlying tags are not shown.

An initial comparison with cerebellum indicated that 1098 transcripts are significantly more abundant ($p \leq 0.05$) in clone 32 than in cerebellum. The tag matches and their categorisations are listed in Appendix B. As expected, this list included a large number of transcripts involved in protein/mRNA synthesis and processing (24% of the genes with known function; Figure 3.3, left). The comparison to ovary epithelium reduced the set of transcripts to 758 in total, and the fraction involved in protein synthesis from 24% to 20% (Figure 3.3, centre). The final filter, against a panel of cancer libraries, reduced the list of CTL-specific transcripts to 387, among which only 5% of the 202 known genes are involved in protein/mRNA synthesis (Figure 3.3, right). The final subset of tags consisted of 30% of tags which corresponded to proteins with a known immune function. This suggests that the filter has successfully reduced the number of non-specific transcripts and increased the percentage of functionally important transcripts.

One concern is that whilst the filter process may remove unspecific protein synthesis proteins it may also remove tags corresponding to proteins with an important immune function. Overall, the 3-stage filter process only eliminated fifty-nine transcripts with known immune function from the initial list. Most of these are expected to be expressed in non-immune tissues, e.g. those encoding MHC molecules and proteasome subunits.

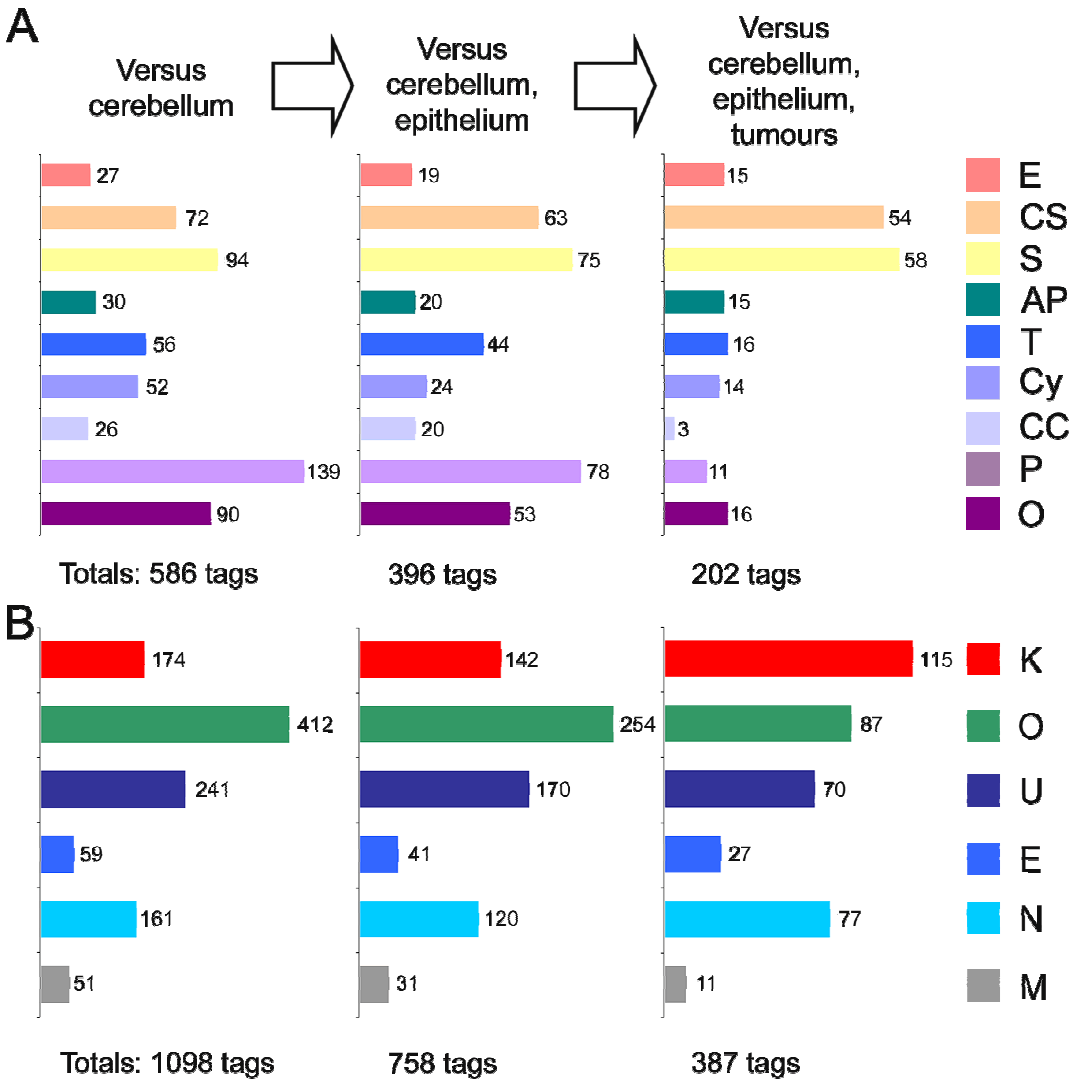


Figure 3.3: Function and extent of characterisation of CTL-specific transcripts.

Three sets of increasingly stringently defined CTL-specific SAGE tags were identified as those significantly more abundant (p-value ≤ 0.05 by AC test) in the clone 32 library compared to libraries derived from normal cerebellum (left) and ovary epithelium (centre), and those that are also not found at similar levels (i.e. at least one third their level in CTL) in more than one of twelve tumour libraries (right). SAGE tags were assigned to transcripts and these transcripts were then categorised according to the broad function of their protein products (A) and level of characterisation (B) using LocusLink, Unigene and RefSeq database entries. Functional classes (A) are: E, soluble Effector molecules; CS, Cell Surface molecules; S, Signalling molecules; AP, Antigen Presentation; T, Transcriptional regulation; Cy, Cytoskeleton related; CC, Cell Cycle or viability related; P, Protein synthesis related; and O, Other. Characterisation classes (B) are: K, Known immune function; O, Other known function; U, Uncharacterised cDNA; E, EST cluster; N, No true Unigene match and M, Multiple true matches. The numbers in the bar chart represent the number of SAGE tags in category.

The above showed that the 3-stage filter process is effective, in that the final subset of tags is enriched for tags corresponding to transcripts with known immune function. However, the aim of this work was to catalogue the tags corresponding to proteins that are found on the cell surface. Therefore, it is important to see if the filter process has succeeded in the initial aim of selecting tags corresponding to immune cell surface proteins.

From the list of 374 transcripts encoding cell surface molecules, tags corresponding to thirty five of the transcripts survived the selection procedure. Surprisingly, this included all of the 18 principal T-cell markers detectable in the library (i.e. TCRα, β1 and β2, CD3γ, δ and ε, CD2, CD5, CD6, CD7, CD8α, β1 and β2, CD11a (LFA-1α), CD43, CD45, CD53 and CD247(ζ)). The selection of all of the transcripts encoding each of the principal T-cell markers by the 3-stage filter process is unlikely to be due to uniformly high expression of these particular genes, as their detected expression varies by more than two orders-of-magnitude. Moreover, the abundance of these known transcripts is likely to be representative for cell surface molecules in general since most genes encoding these molecules were identified on the basis of protein abundance or antigenicity rather than high mRNA expression levels. Looking at the individual tags, rather than for all observed tags for a given transcript, 14 of the 18 key T-cell markers passed through the 3-stage filter process (individual tags for CD8α, CD43, CD45 and TCRβ1 were not selected). Overall this suggests that the filter process is highly effective at producing a subset of tags that correspond to transcripts with important function on the immune cell surface. Therefore, if there are currently unidentified cell surface molecules with an important function in the CTL it is likely that their tags will be found in this subset.

To confirm that the selection of the 18 key T-cell markers was a facet of the T-cell library rather than the libraries it was being filtered against, twenty two non-immune derived libraries and eleven non T-cell immune derived libraries were passed through the filter. For the non-immune libraries in only two (out of 396) instances was a marker selected by the filter (CD45 in both cervix and lung). For the non T-cell immune libraries, in all but the two NK libraries, two or fewer markers were selected by the filter (for both the NK libraries ten markers were passed through the filter). Overall this suggests that the selection of the markers is particular to the T-cell library, rather than due to characteristics of the filter. The expression of these key T-cell markers in the panel of libraries is further investigated in section c below.

Throughout the filter process a large percentage (42-45%) of CTL-specific tags match transcripts that have yet to be assigned any clear function (Figure 3.3 B). The relative invariance of this fraction suggests that, regardless of how tissue specificity is defined, only ~55-60% of the differentially regulated, moderately to highly abundant transcripts in the resting T-cell transcriptome have been characterised. At the highest level of stringency, more than half the known, CTL-specific transcripts encode cell surface molecules (27%) or signalling/adaptor proteins (29%, Figure 3.3 A, right). An additional 29% of the transcripts are evenly distributed between those encoding soluble effector molecules (7%), transcriptional regulators (8%), antigen processing and presentation molecules (7%), and cytoskeletal elements (7%). Of the remaining highly CTL-specific transcripts, eleven (5%) encode proteins involved in protein/mRNA synthesis and nineteen (9%) have other miscellaneous functions. One would expect that the 97 CTL-specific transcripts matching Unigene clusters but lacking any prescribed function should encode a set of proteins with a similarly broad

range of functions. To examine this, the putative open reading frames encoded by these transcripts were characterised using standard bioinformatics tools (see Chapter 2 for further details).

Standard bioinformatics tools should be able to identify novel "immune specific" cell surface proteins because the large majority (~80%) of the known leukocyte cell surface antigens defined thus far, including type I and II membrane proteins and GPI-anchored molecules, have characteristic, modular extracellular protein domains belonging to nineteen different protein superfamilies[2]. A further three percent of leukocyte antigens lack these domains, but are characterised by the presence of either four (e.g., Fc receptor subunits) or seven (i.e., G protein-coupled receptors) transmembrane domains. If when examining the uncharacterised transcripts one finds that they encode for proteins with the modular architecture characteristic of ~80% of leukocyte surface antigens one can be confident that they are cell surface proteins.

Of the 97 stringently defined CTL-specific transcripts matching Unigene clusters but lacking any prescribed function, putative complete open reading frames were available for 70 of these transcripts (Appendix C.1). Fourteen of these were predicted to have one or more transmembrane domains, only two of which also contain a signal peptide (Table 3.1). However, none of predicted proteins possess the modular architecture present in the large majority of known leukocyte cell surface antigens and, therefore, are unlikely to be genuine cell surface proteins. Only one previously uncharacterised transcript encodes a protein that has the 7 transmembrane pass topology characteristic of G-protein coupled receptors and is therefore likely to

be a cell surface molecule. Of the other transcripts containing open reading frames, eleven apparently encode signalling or adaptor proteins (Table 3.1), three encode DNA binding domains, three others appear to encode cell cycling proteins, four encode domains linked to rearrangements of the actin cytoskeleton and another encodes a BAR domain-containing protein (associated with vesicle transport). Finally, four transcripts encode domains associated with mRNA processing and protein synthesis while three have domains associated with other housekeeping functions.

Table 3.1: Uncharacterised CTL-specific proteins containing putative transmembrane helices or signalling domains.

Transcripts whose functions have not been characterised, but for which sequence is available, that match tags classified as CTL-specific using the three stage method described in the text, were analyzed using BLAST, SMART, InterProScan and the conserved domain search tool at NCBI (see Chapter 2). Those predicted to have one or more putative transmembrane helix (TMH) by TMHMM2 (run via SMART) or that contain putative signalling domains are listed. Domains found in each protein are listed in order from the N terminus to the C terminus.

* Indicates matches to incomplete domains that may therefore be false hits.

	Tag Sequence	Tag count (per 100,000)	Unigene Cluster	Accession no.	Domains
Transcripts encoding transmembrane helices	ACCATTGGAT	103	146360	NM_003641	2 TMH
	GACTTGGCCT	30	16291	AK057590	2 TMH
	CTCCTCCAAG	67	15284	BC028076	3 TMH
	AGCAAGAAAC	19	107393	NM_019895	Signal peptide, 3 TMH, Claudin*
	TACGAGGCCG	17	16165	NM_007267	10 TMH
	TGGGGCCGCA	17	288455	AK026923	7 TMH
	CATTTACTCT	32	17109	NM_004867	1 TMH
	AGGCCACTGG	11	323634	NM_024070	Signal peptide, 1 TMH
	GACAGATGGA	11	83575	BC014077	GRAM, 1 TMH
	CTTCTTTCCA	10	259737	NM_020179	1 TMH
	GATGAAAAGG	16	343473 / 172847	NM_005528	DNAJ, 1TMH
	GGGCGCCTGG	8	159955	NM_130759	MMR GTPase*, 1 TMH
	TTCTCAAGAA	8	37189	NM_007069	NLP/P60*, 1 TMH
	ATAAACAGAT	28	334825	AK027658	Coiled-coil, 1TMH
	TGTTGACTCT	19	8882	CAP3 contig 1	Signal peptide, 7TMH, Rhodopsin-like receptor
Transcripts encoding signalling domains	CACCCAATGG	65	110121	NM_012455	Sec7-like GEF, PH
	TAAGGACGAG	33	238707	NM_024901	DENN (AEX-3)
	TGCAAGAGAG	30	238954	AL832852	RhoGAP
	GGAGCTTGAG	27	61469	NM_018990	SH3, SAM
	GGCGGGGCCA	17	54985	AB002301 / BC003646	Protein kinase (PK), PK extension, PDZ domain
	AGGCTCCGTG	13	16229	AB037794	JAB/MPN
	ACCTGCAGGC	11	147066	BC016615	GTPase (Rab, Ras or Rho like)
	TTTGGGACCC	11	270	NM_004288	PDZ
	CAGGTTAAGC	8	99877	BC028068	SH2
	GAACCGTCCT	47	123164	AW512177	TBC
	TGCCAATTAA	16	165337	AI990569	GTPase (Rab, Rac, Ras, Ran or Arf like)

The remaining twenty-seven tags linked to transcripts with unknown function matched Unigene clusters containing only EST sequences. Possible full length sequences were generated and open reading frames identified for these clusters (Appendix C.2). Only seven of the putative encoded proteins have recognizable domains, i.e. a probable G-protein coupled receptor, two G-protein-related signalling proteins, a probable Zinc finger-containing transcription factor, an RNA-editing domain-containing protein, a metabolic enzyme and the Ly49L pseudogene. Tags not matching any Unigene cluster were not investigated further, though work in the following chapters suggests some of the these tags are likely to correspond to sequencing errors and antisense transcripts.

Overall, none of the 97 stringently defined CTL-specific but uncharacterised transcripts appear to encode proteins with the modular architecture characteristic of ~80% of leukocyte surface antigens and only two appear to encode for genuine cell surface proteins. It might be a coincidence that the two novel proteins have the same architecture (GPCR) or perhaps genes encoding these types of receptors are more difficult to identify via conventional cloning methods (e.g., due to more limited immunogenicity). To put this into context, these 97 uncharacterised transcripts represented ~25% of all the CTL-specific transcripts and contained only two predicted cell surface proteins whereas all 18 of the key T-cell markers, along with another 17 known immune cell surface proteins, were identified amongst the CTL-specific transcripts with a known function. There is no reason to suspect that the 77 additional tags that currently do not give matches to any database correspond to transcripts encoding a higher proportion of cell surface molecules than tags matching uncharacterised sequences, as neither EST nor cDNA sequencing ought to be biased

toward such transcripts. Overall, therefore, this strongly suggests that the "immune specific" protein composition of the human resting CD8$^+$ T-cell surface is largely defined.

Figure 3.4 summarizes the immune cell surface proteins identified by SAGE in the T-cell library. From this it appears that the T-cell surface is dominated by relatively large molecules, including integrins and proteins with mucin-like segments, and by proteins with seven transmembrane domains. A striking correlation exists between abundance and functional importance within this set of transcripts: proteins with the most critical roles in initiating adhesion and T-cell activation, such as CD2, LFA-1, the T-cell receptor, the coreceptor, CD8, and CD45, are all encoded by the 20% most highly expressed transcripts for cell surface molecules. However, it is of considerable interest to see how, on a T-cell, the numbers of cell surface proteins with an immune function compares with that of the cell surface proteins that have a non-immune function. This is investigated in Section 2 below.

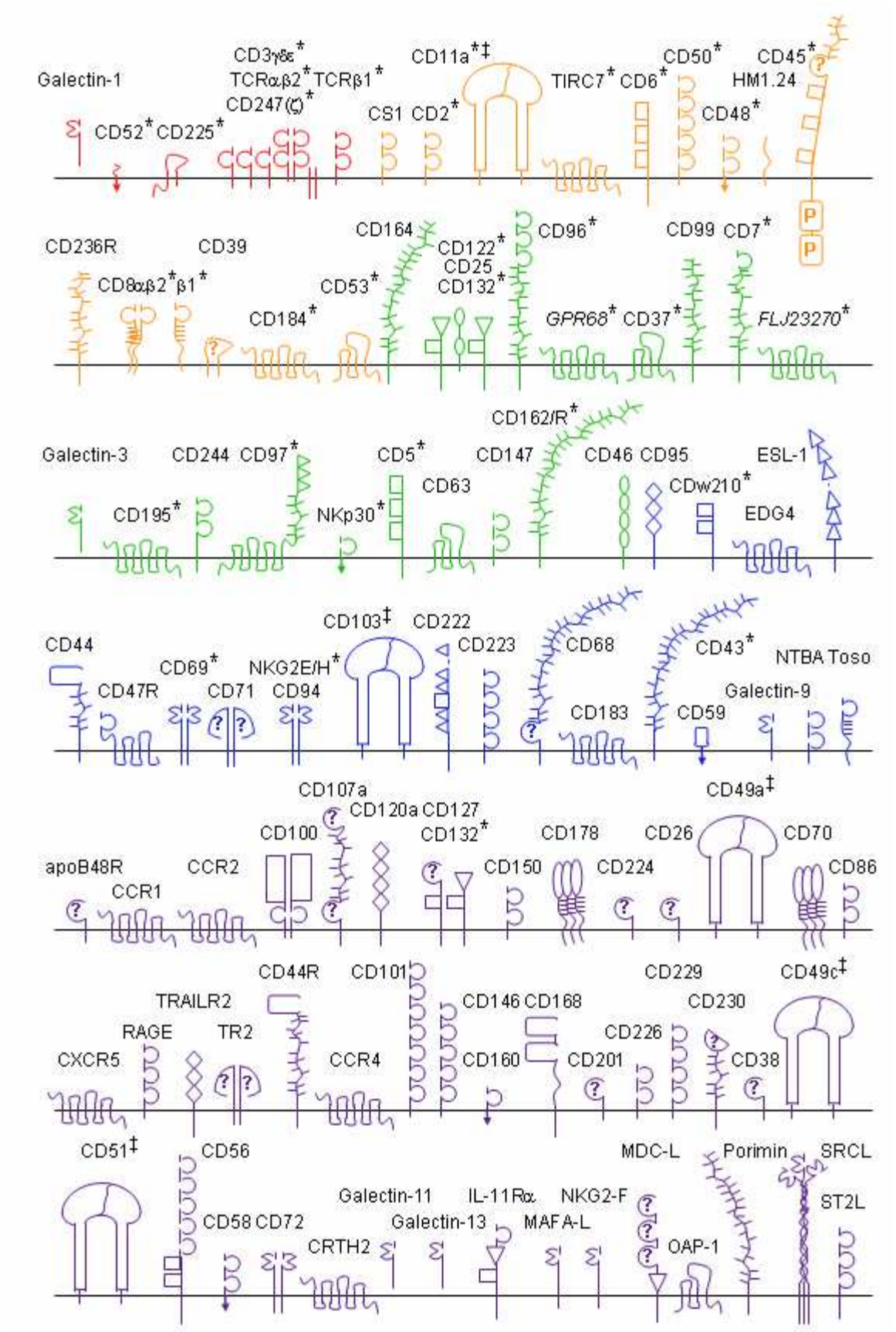
As stated previously, the main strength of using a clonal population for this analysis was that it effectively made the library deeper than a similarly sized non-clonal library. Applying the three-stage filter process to the *ex vivo* purified CD8$^+$ T-cell library[146] demonstrates this clearly. CD2, CD3ε, CD3γ, and CD45 are all found in the *ex vivo* library but do not survive the 3-stage filter process.

Figure 3.4: Cell surface molecules encoded by transcripts identified in the clone 32 SAGE library.

Schematic representations of the defined and proposed CD antigens and TCR components whose expression in clone 32 was detected by SAGE are shown. The two new 7 TM proteins identified are also included (italics). The architecture of these proteins is drawn approximately to scale according to the conventions of Barclay *et al.*[2]. Unconventional domains, or those for which there are no structures, are labelled "?". The molecules are coloured according to transcript abundance per 100,000: purple = ≤ 3, blue = 4 - 9, green = 10 - 27, orange = 28-81, red = > 81. Complexes are represented at the level of the most abundant subunit-encoding transcript.

*Stringently defined, CTL-specific molecules.

‡ Five integrin α-domains were detected in clone 32 by our analysis: CD11a, CD49a, CD49c, CD51 and CD103, which associate with the integrin β-chains CD18, CD29, CD61 and β7. Tags derived from CD18 and CD61 were found at high levels in the library, but matched several other genes preventing abundance determination. Integrin β7 tags were also present but this protein has not been considered for a CD designation. No tags derived from CD29 were observed, presumably due to sampling effects.


Galectin-1
CD3γδε*
TCRαβ2*TCRβ1*
CD247(ζ)*
CD11a*‡
TIRC7*
CD6*
CD50*
CD45*
HM1.24
CD52*
CD225*
CS1
CD2*
CD48*
P
P
CD236R
CD39
CD8αβ2*β1*
CD164
CD53*
CD184*
CD122*
CD25
CD132*
CD96*
CD99
CD7*
GPR68*
CD37*
FLJ23270*
Galectin-3
CD244
CD97*
CD5*
CD162/R*
CD147
CD46
CD95
ESL-1
CD195*
NKp30*
CD63
CDw210*
EDG4
CD44
CD103‡
CD222
CD223
CD68
CD43*
NTBA
Toso
CD47R
CD69*
CD71
CD94
NKG2E/H*
CD183
CD59
Galectin-9
CD107a
CD120a
CD127
CD49a‡
CD100
CD132*
CD178
CD26
CD70
apoB48R
CCR2
CCR1
CD150
CD224
CD86
CD44R
CD101
CD229
CD49c‡
TRAILR2
CD146
CD168
CD230
RAGE
TR2
CD226
CXCR5
CCR4
CD160
CD201
CD38
CD51‡
CD56
MDC-L
Porimin
SRCL
Galectin-11
IL-11Rα
NKG2-F
ST2L
CD58
CD72
Galectin-13
MAFA-L
CRTH2
OAP-1

An obvious and important caveat concerns the extent to which the entire pool of transcripts encoding cell surface molecules was sampled in our 63,270 tag SAGE library. In this library, the key T-cell surface components were identified and almost all proteins found by FACS were also identified. This suggests that the library is large enough to examine the tissue specific components but not all components expressed by the cell. Whilst one cannot formally exclude the possibility that an entirely new class of T-cell specific, weakly differentially expressed transcripts encoding cell surface molecules exists, this seems unlikely since none of the genes encoding surface molecules already known to be critical for the function of T-cells show this pattern. Conversely, had there been large numbers of transcripts encoding new proteins with the modular architecture and expression properties of known cell surface proteins, our level of sampling would have ensured that many of these would have been identified. The effect of increasing library size on detection of both tissue specific and non-specific transcripts is investigated below.

c) Transcript detection sensitivity

The above work showed that by using differential expression patterns one can identify tags corresponding to proteins with a tissue specific function. However, it would be interesting to examine how the expression levels of these purportedly tightly regulated transcripts vary over a wider range of tissues. To this end, a panel of 44 libraries was assembled, of which 20 were derived from immune cells. The non-immune libraries were obtained from NCBI[166] and the immune libraries were either generated in our lab or obtained from the BloodSAGE database[143].

Two lists of purportedly tightly regulated genes were used for the comparisons. The first list is the 18 key T-cell markers that were described above. The second list is a

list of 43 retina specific genes identified by Schultz *et al.*[186]. Retinal genes were chosen for two reasons. Firstly, there were four publicly available libraries[187], of which two were very large (>100,000 tags), so it was possible to confirm that these genes were found in the retina by SAGE. The other advantage of examining retinal genes, as opposed to T-cell markers, is that tissues are unlikely to be contaminated with retina, whereas tissues are likely to contain some leukocytes. Tags were successfully generated for 37 of the retinal genes in exactly the same way as they were for the immune cell surface proteins (four of the genes had no CATG site, whilst the other two had SAGE tags that matched multiple genes).

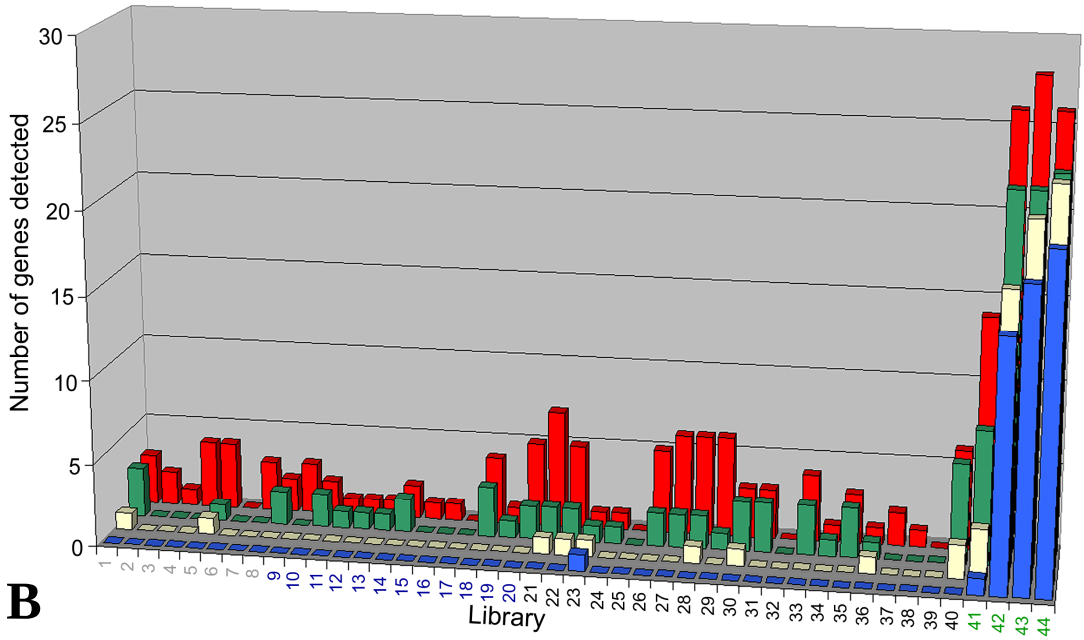
In the large majority of cases genes appear to be "switched off" when one would expect them to be (see Figure 3.5). For the non-retinal tissues, the retinal genes are not found in over 90% of cases. Of the 21 genes that were found in the non-retinal libraries two thirds were found in four libraries or less. It is likely that the remaining seven genes are either not retina specific and as such are misclassified or the assigned tags correspond to other genes which are not retina specific. Unfortunately, Schlutz *et al*. did not publish their criteria and so is not possible to understand why this misclassification may have occurred. It is interesting to note that five of thirty-seven genes analysed are not found in the retina libraries. This may be because the libraries are not large enough, the subset of cells in which the genes are expressed have not been profiled, or a form of the transcripts are expressed which are not profiled by the tags chosen.

Figure 3.5: Expression of "tissue specific" genes in a panel of libraries.

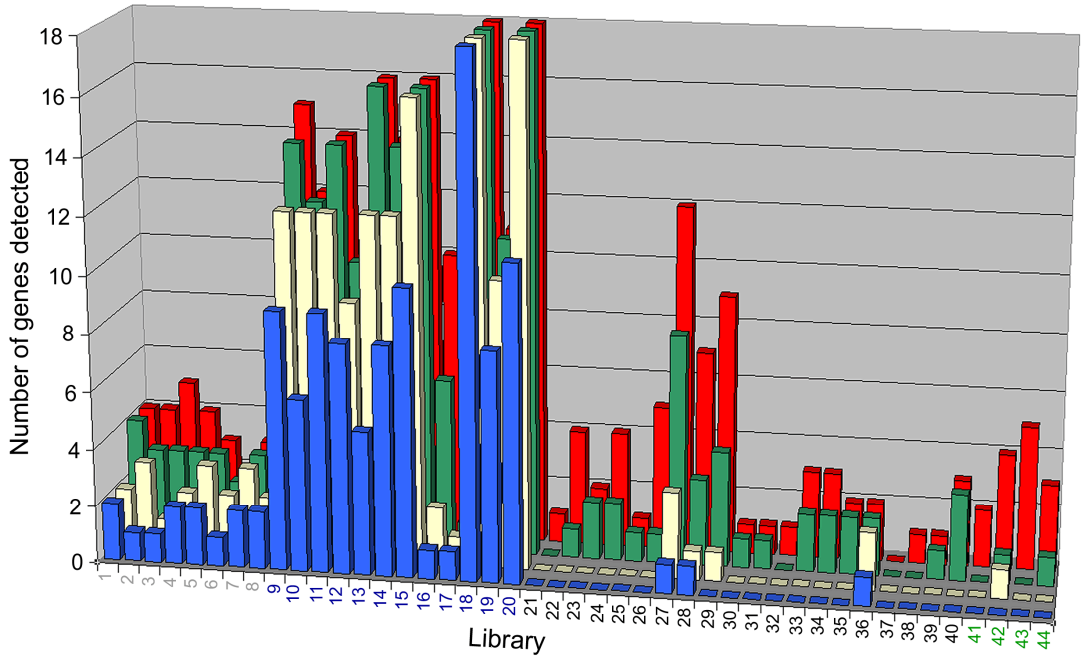
Tags were extracted for "tissue specific" genes and their expression levels were examined in a panel of libraries. The expression pattern of the retinal genes is shown in Figure A, a total of 37 genes were tested. The expression pattern of the 18 key T-cell markers is shown in Figure B. In both graphs the red bars represent genes expressed at any level; green bars represent >2 tags per 100,000; yellow bars represent >5 tags per 100,000 and blue bars represent >10 tags per 100,000.

Grey numbered libraries are myeloid derived libraries, numbers 1-8 represent: Granulocyte, Monocyte, Macrophage-Colony Stimulating Factor Macrophage, Granulocyte Macrophage-Colony Stimulating Factor Macrophage, Immature Dendritic Cells, Mature Dendritic Cells, Langerhans-like Cells and Lipopolysaccharide Monocytes. Blue numbered libraries are lymphoid derived libraries, numbers 9-20 represent: Naive T-cells, Resting Th1, Resting Th2, Activated Th1, Activated Th2, Japanese NK, Japanese CD8 T-cell, B-cell, Activated B-cell, Clone 32, NK, αCD8 activated Clone 32. Black numbered libraries are non-immune cell derived libraries, excluding the retina derived libraries, libraries 21-40 represent: Cerebellum, Cerebellum Duke, Colon Nc1, Heart, Kidney, Liver, Lung, Peritoneum mesothelium, Prostate, Ovary epithelium, Colon Nc2, Astrocytes, White matter, Vascular endothelium, Breast epithelium, Cervix, Muscle old, Muscle young, Stomach epithelium, Thalamus. Green numbered libraries are retina derived libraries, libraries 41-40 represent: Retina epithelium, Retina macula, Retina peripheral and Retina peripheral GSM572.

A


Number of genes detected
0
5
10
15
20
25
30
Library
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44

B


Number of genes detected
0
2
4
6
8
10
12
14
16
18
Library
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44

For the T-cell markers there was a similar if less marked pattern, in that in the non-immune cell derived libraries the T-cell markers are not found in 70% of cases. There is a clear difference in the number of immune specific genes expressed by the immune and non-immune cells. Three non-immune libraries (lung, prostate and peritoneum mesothelium) stand out due to their expression of a large number of the immune specific proteins. It is tempting to suggest that these libraries are contaminated with leukocytes rather than being fundamentally different to the other non-immune libraries.

The level of expression of the retinal genes in the non-retina libraries and the immune genes in the non-immune libraries appears to be low. More than half the T-cell markers present in the non-immune libraries are found at less 5 tags per 100,000 and ~90% of the retinal genes are found at less than 5 tags per 100,000. The fact that these genes are found at a low level suggests that the size of the library may be an important factor in determining whether they are detected or not.

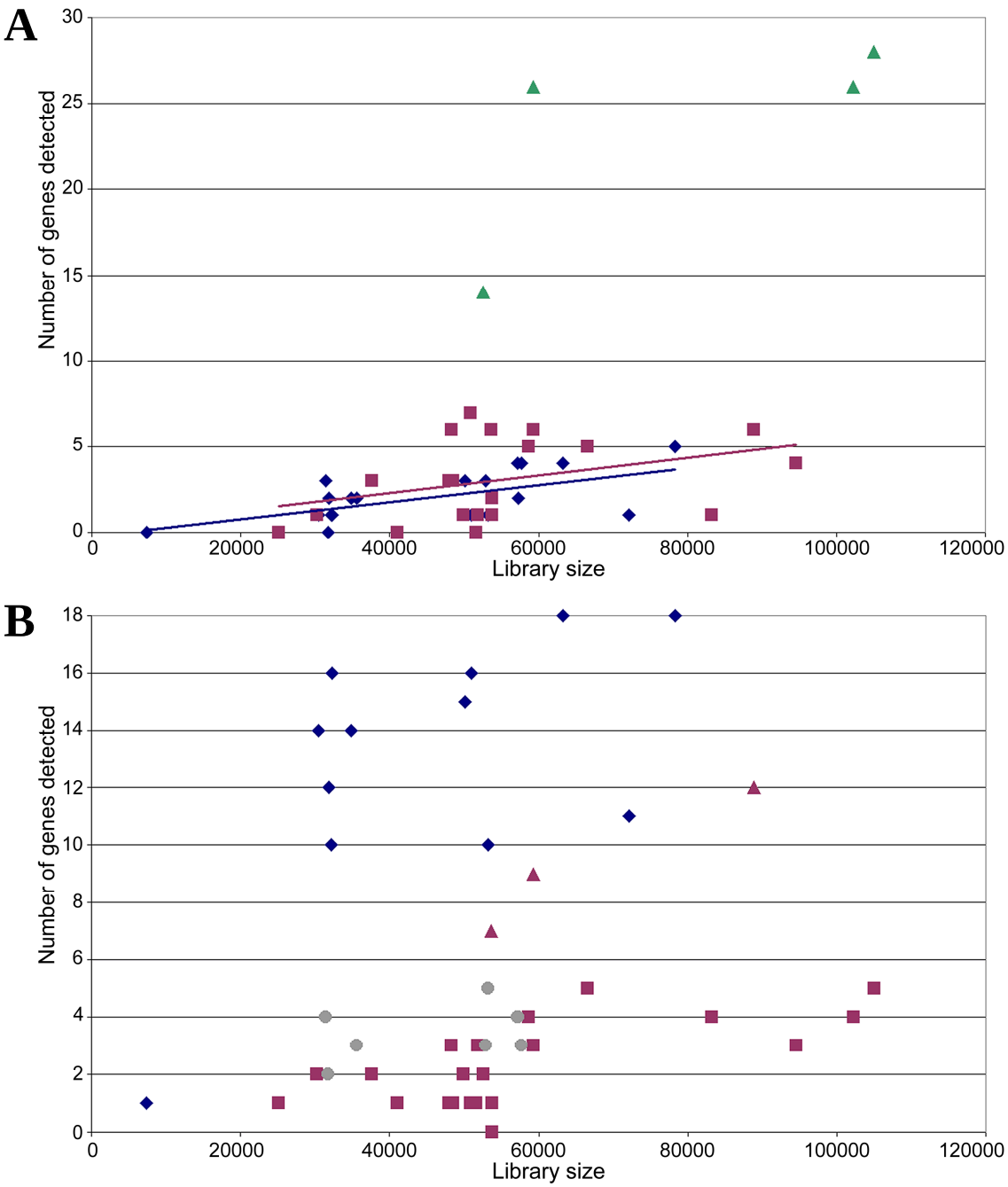
Figure 3.6 shows the effect of library size on the number of genes discovered. As expected when looking at the retinal genes (Figure 3.6 A) the immune and non-immune classes are indistinguishable. However, both classes show that with increasing library size there is a gradual increase in detection of the retinal transcripts. A different pattern emerges when examining the key T-cell genes (Figure 3.6 B). In this case it is useful to break down the immune library category into lymphoid and myeloid cells. The myeloid cells and the non-immune cells show essentially the same pattern, which is similar to that observed in the retinal genes, i.e. a gradual increase in the number of genes detected as library size increases. The lymphoid cell

libraries show a very different pattern, the number of transcripts detected increases much more rapidly with library size. As all the key markers have been detected at a library size of ~60,000, one can assume that the rapid discovery of functionally specific transcripts tails off at a larger library size.

Expanding the analysis of the cell surface proteins to include all 374 proteins in the list generated previously is revealing. As this list is not limited to genes of key importance to a T-cell, one would expect that the lymphoid and myeloid derived libraries show a similar pattern of transcript detection. This is indeed the case (Figure 3.7). It also interesting to note, that in all three categories the rate of transcript detection appears to remain constant with increasing library size. This suggests that larger libraries will be needed to profile all transcripts in a cell, rather than just the functionally specific transcripts.

Figure 3.6: Effect of library size on "tissue specific" gene detection.

The number of "tissue specific" genes detected in each library was plotted against library size to examine the effect of increasing library size on gene detection. The libraries used are described in Figure 3.5. Figure A shows the retinal genes: blue diamonds represent immune cell derived libraries, mauve squares represent non-immune derived libraries and green triangles represent retina derived libraries. Trend lines were added for both the immune and non-immune derived libraries. The gradients of the trend lines appear to be the same, suggesting there is little difference between these two classes of libraries when examining the expression of retinal specific genes. Figure B shows the key T-cell genes: blue diamonds represent lymphoid libraries, grey circles represent myeloid libraries, and mauve squares all the non-immune libraries except the three potentially contaminated libraries (lung, peritoneum mesothelium and prostate) which are represented by mauve triangles.


A
Number of genes detected
Library size
0
5
10
15
20
25
30
0
20000
40000
60000
80000
100000
120000
B
Number of genes detected
Library size
0
2
4
6
8
10
12
14
16
18
0
20000
40000
60000
80000
100000
120000

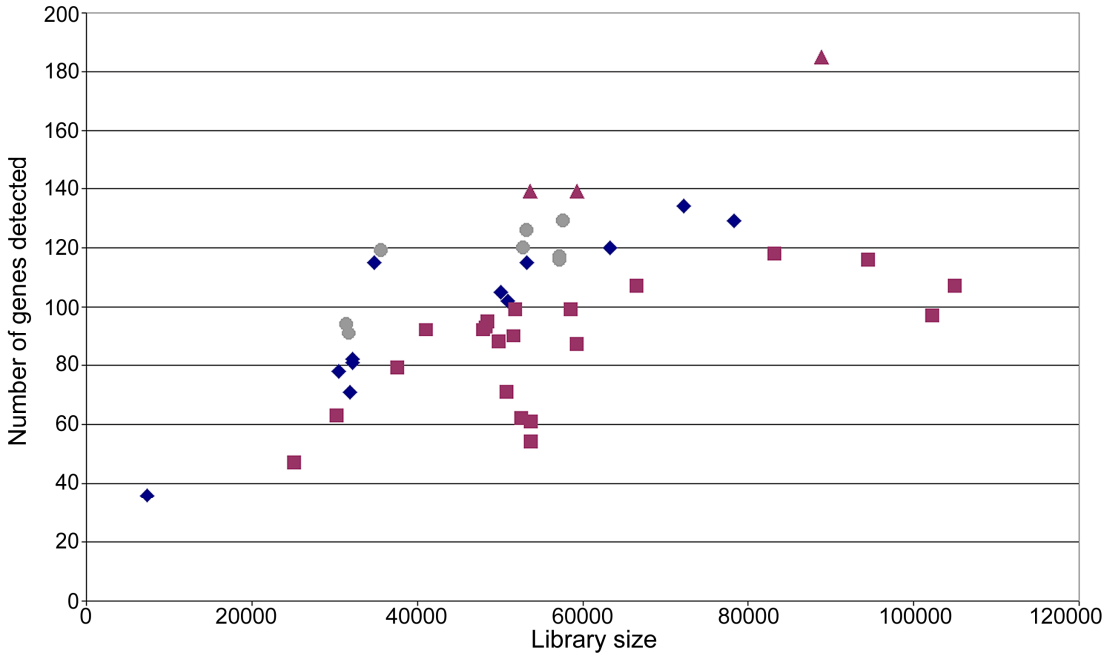


Figure 3.7: Effect of library size on the detection of 374 transcripts encoding cell surface proteins.

The expression of all 374 transcripts, which encode cell surface proteins and have manually curated SAGE tags, was examined across the panel of libraries. The numbers of transcripts detected in each library is plotted here. Blue diamonds represent lymphoid libraries, grey circles represent myeloid libraries, mauve squares represent non-immune libraries except the three potentially contaminated libraries (lung, peritoneum mesothelium and prostate) which are represented by mauve triangles.

Two key conclusions emerge from this work. The first is that functionally specific transcripts that one would expect to be tissue specific are found in a wide range of tissues (e.g. CD3 in the retina). There are three explanations for this: the genes may have an as yet unknown function in these tissues; this transcription may just be non-specific "leaky" transcription that is functionally irrelevant or the tags that correspond to these genes may also correspond to other genes that are functionally important in the tissue being studied. The second conclusion is that whilst a clonal library of 60,000 tags appears to detect all functionally specific transcripts it is clearly not large enough to detect all transcripts expressed in a library. Therefore if one wants to characterise the entire transcriptome of a cell, rather than the just the functionally specific transcripts, it will be necessary to produce much larger libraries. This is investigated further in Chapter 5.

2. Non-immune cell surface

In the previous section, 111 genes encoding cell surface proteins with a known immune function were found to be expressed by a T-cell clone. This involved using a manual strategy to assign tags to each of the 374 genes tested. To place this number in context, a partially automated strategy was used to identify transcripts corresponding to cell surface proteins, regardless of whether they had an immune function or not. Tags were then derived for these transcripts and their expression was examined in the panel of libraries used above.

The methodology used to assemble the gene list and tags is described in detail in Chapter 2. Briefly, the Genome Ontology[172] (GO) database at CGAP was used for this work, all genes corresponding to proteins classed in any of a variety of classes that potentially contained cell surface proteins were obtained. This list contained

2,818 genes. For each gene published literature was examined and those genes where there was evidence for cell surface expression were kept. Tags were then associated to genes using the automated mapping of SAGE Genie[112]. This left 1,776 genes for which the corresponding proteins had been shown to be found on the cell surface and tags could be assigned. To evaluate the quality of the tag-to-gene mapping by SAGE Genie, the mappings were compared to the manual mapping for the 374 genes corresponding to cell surface proteins used above. SAGE Genie contained mappings for 364 of the genes. Within this set, the tag identified by SAGE Genie was also found in the curated set for 309 genes. This suggests that SAGE Genie provides a satisfactory level of mapping for an automated analysis.

The number of genes expressed and the level of expression of the genes in the immune and non-immune libraries was compared. In both classes of tissue there was a wide range in the number of transcripts observed, spanning from 126 to 440 transcripts in the immune libraries and 197 to 553 transcripts in the non-immune libraries. To take into account the effect library size has on the number genes detected the data was normalised in terms of the number of unique transcripts in the library and the total library size (Figure 3.8). In terms of percentage of number of unique transcripts, the immune and non-immune libraries showed no statistical difference (2.30% vs. 2.27%, p-value 0.81). This suggests that the relative complexity of the cell surface, when compared to the overall complexity of the library, is the same in both classes. However, looking at the percentage of transcription used to express cell surface proteins, the immune libraries show a statistically significant higher level of expression compared to the non-immune libraries (8.1% vs. 5.5%, p-value $<<0.01$). This possibly reflects the importance of cell surface interactions in the function of immune cells.

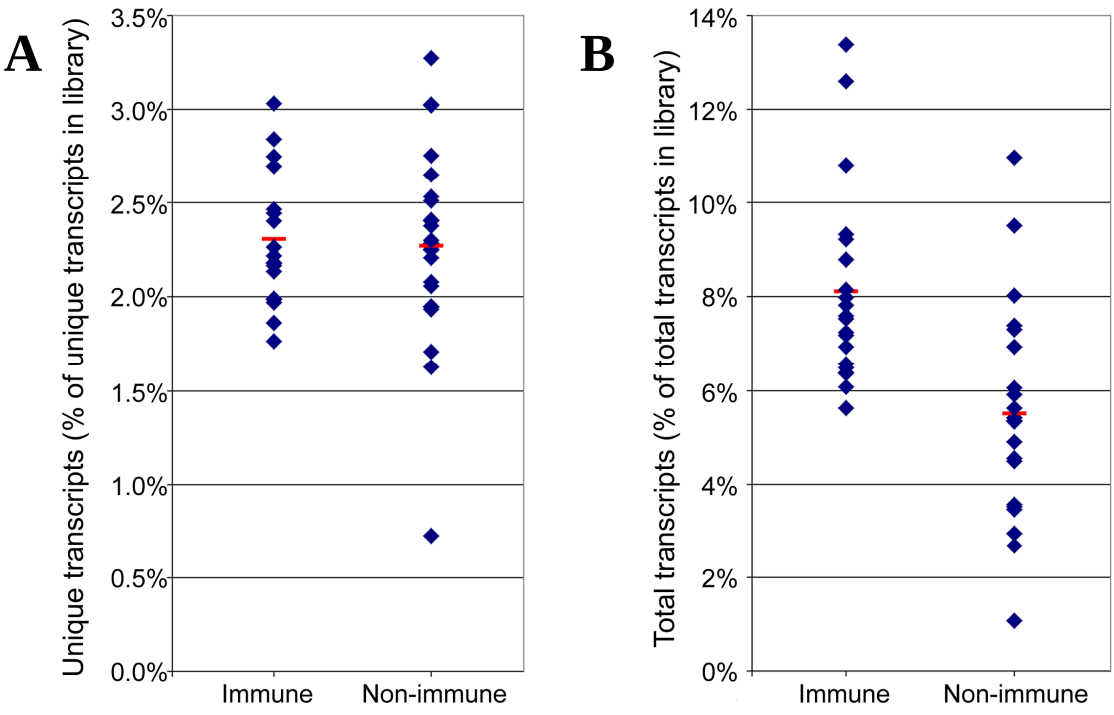


Figure 3.8: Expression of GO annotated cell surface proteins across the panel of libraries.

The expression of the transcripts encoding proteins found at the cell surface and listed in GO, for which SAGE Genie tags could be assigned, was examined across the panel of libraries. Figure A shows the percentage of unique transcripts in a library that are found in this set; this shows the relative complexity of the cell surface relative to the rest of the transcripts. Figure B shows the percentage of total transcripts in a library that are found in this set; this shows the relative amount of transcriptional effort that is being spent on this set.

Whilst the differences observed between the classes of libraries are likely to be real, the absolute numbers of cell surface transcripts identified here are likely to be underestimates for a variety reasons. Firstly, only one tag was used per gene and as shown in Chapter 6, often more than one tag is found per gene. Secondly, as described above, all the libraries are too small to reliably identify all the non-tissue specific genes. Thirdly, neither the GO database nor SAGE Genie are complete, so not all known and obviously none of the novel genes encoding cell surface proteins will have been examined by this method. For example, approximately 20% of the transcripts encoding the HLDA cell surface molecules were not found in the final GO/SAGE Genie list.

Going back to the original aim of investigating how the expression of immune cell surface proteins fitted in the context of all cell surface proteins, the clone 32 library was examined in detail and different classes of cell surface transcripts were classified. 412 transcripts were found, of which 132 had a known immune function and the remaining 280 were classed as non-immune. Again, it is perhaps more informative to compare the transcriptional effort being spent on these transcript classes. The two classes were found at similar levels (5.6% of total transcription on cell surface transcripts with a known immune function and 5.0% on those without a known immune function). The average expression level of those transcripts with an immune function was over twice that of those without a known immune function (42.3 vs. 18.0 tags per 100,000). At this level of expression there is a relatively high correlation between transcript level and protein synthesis[139]. Therefore, it is tempting to speculate that the cell surface will be populated with a larger number of cell

specific (immune) surface proteins than those cell surface proteins with a house keeping role (non-immune).

IV. Natural Killer cell transcriptome

A. Introduction

In the previous chapter SAGE was successfully used to characterise the expression of "immune specific" transcripts in a CTL clone. Surprisingly, few novel cell surface molecules were identified in the CTL library. To try to establish whether this would be true for other lymphocytes a natural killer (NK) cell library was made. NK cells were chosen as their role in the immune system is well known but the components and their individual roles are still relatively poorly understood. The activation state of NK cells is thought to be controlled by a balance of signals coming through various receptors independently, or in combination[188]. However, it is far from clear that all the key receptors have been identified. Indeed, recently a new family of receptors responsible for mediating NK cytotoxicity has been discovered (the natural cytotoxicity receptors, reviewed by Moretta *et al.*[189]). Therefore the initial aim of this work was to identify the "immune specific" cell surface components of NK cells.

As the NK cell is relatively poorly characterised it would also be useful to identify the key functional components expressed in the cell. It was shown in the previous chapter that the 3-stage filter process can be used to identify "immune specific" molecules. Therefore, the secondary aim of the work presented here is to identify key molecules expressed in an NK cell, using SAGE and the 3-stage filter process.

In the previous chapter approximately one quarter of the "immune specific" tags were classed as "no match" tags. These "no match" tags may be genuine and correspond to novel exons or transcripts, or they may be artificial and be a result of either

technical artefacts (such as sequencing errors) or problems with the tag-to-gene mapping protocol. To establish whether the "no match" tags are genuine a systematic approach was undertaken to identify the transcripts corresponding to the "no match" tags identified in the NK library.

B. Results and Discussion

As described in the materials and methods chapter, a SAGE library was made from a sample of RNA from an NK cell line. After the exclusion of linker tags, a total of 72,161 tags were sequenced, representing 18,963 unique tags. As this library is a similar size to that of the clone 32 library, one would expect this library not to have reached sampling saturation. Figure 4.1 shows that this is the case, therefore one may not detect all the lowly expressed transcripts in the NK cell with this library.

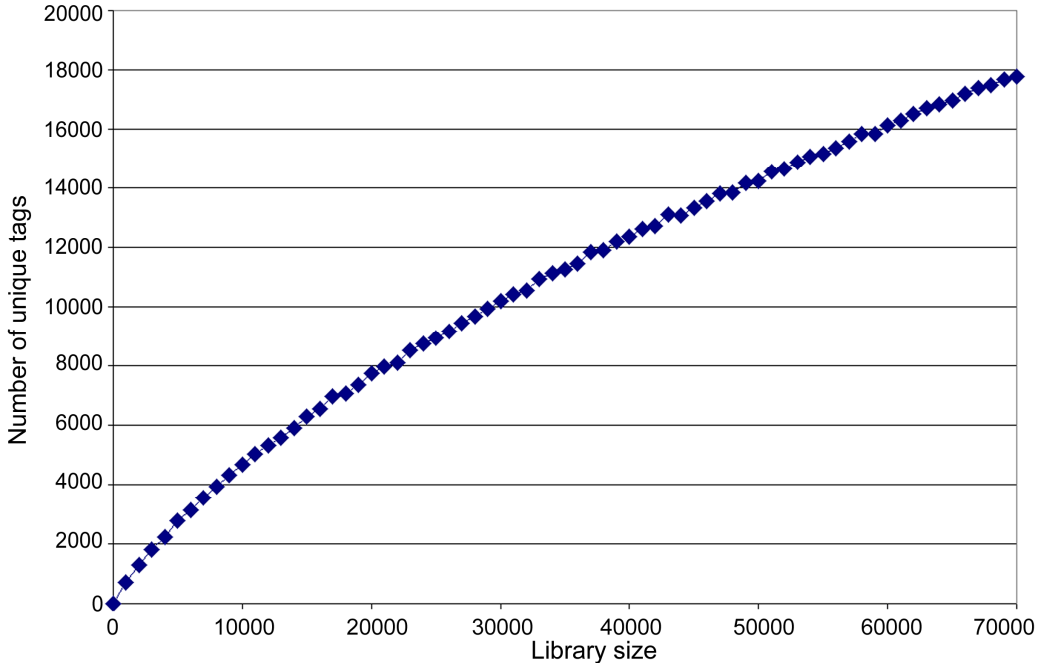


Figure 4.1: Effect of total number of tags sequenced on number of unique tags identified.

The NK library was sampled at various sizes to examine the effect of library size on the number of unique tags identified. If the library is large enough to sample all available tags, then increasing the library size will not increase the number of transcripts detected.

1. Cell surface molecules

To examine the quality of the library the FACS data was compared to the SAGE data. Examining the FACS plots shows that the starting NK population was heterogeneous (Figure 4.2). All markers detected by FACS were also found by SAGE and the two absent markers (CD3 and CD28) were absent in the SAGE library. The FACS markers were found over a wide range of expression levels, 1 to 35 tags per library. This suggested that the library is sampling deep enough to identify key cell surface markers.

To examine the cell surface in further detail and to evaluate the efficacy of the 3-stage filter process on the NK cell surface molecules, the expression of cell surface molecules highlighted as being functionally important in NK cells by Chiesa *et al.*[190] was analysed. Chiesa *et al*. listed 29 cell surface molecules, which are believed to have either an activating or inhibitory role in NK cells. From this list twenty four were in the CD list used in the previous chapter and were analysed here (Table 4.1). The expression of three other molecules of importance in NK cells was also examined (CD56 [ref:[191]], CD94 [ref:[191]] and CD161 [ref:[192]]). Twenty of the twenty seven molecules were found in the NK library, at expression levels ranging from 1 to 171 tags in the library. There has been one other published NK cell line SAGE library[146] and in this library 17 of the transcripts were found, of which there was an overlap with 14 of the transcripts in the library presented here. Having established that the majority of cell surface molecules one would expect to find in the library are present, it was possible to evaluate the efficacy of the 3-stage filter process on NK cell surface molecules.

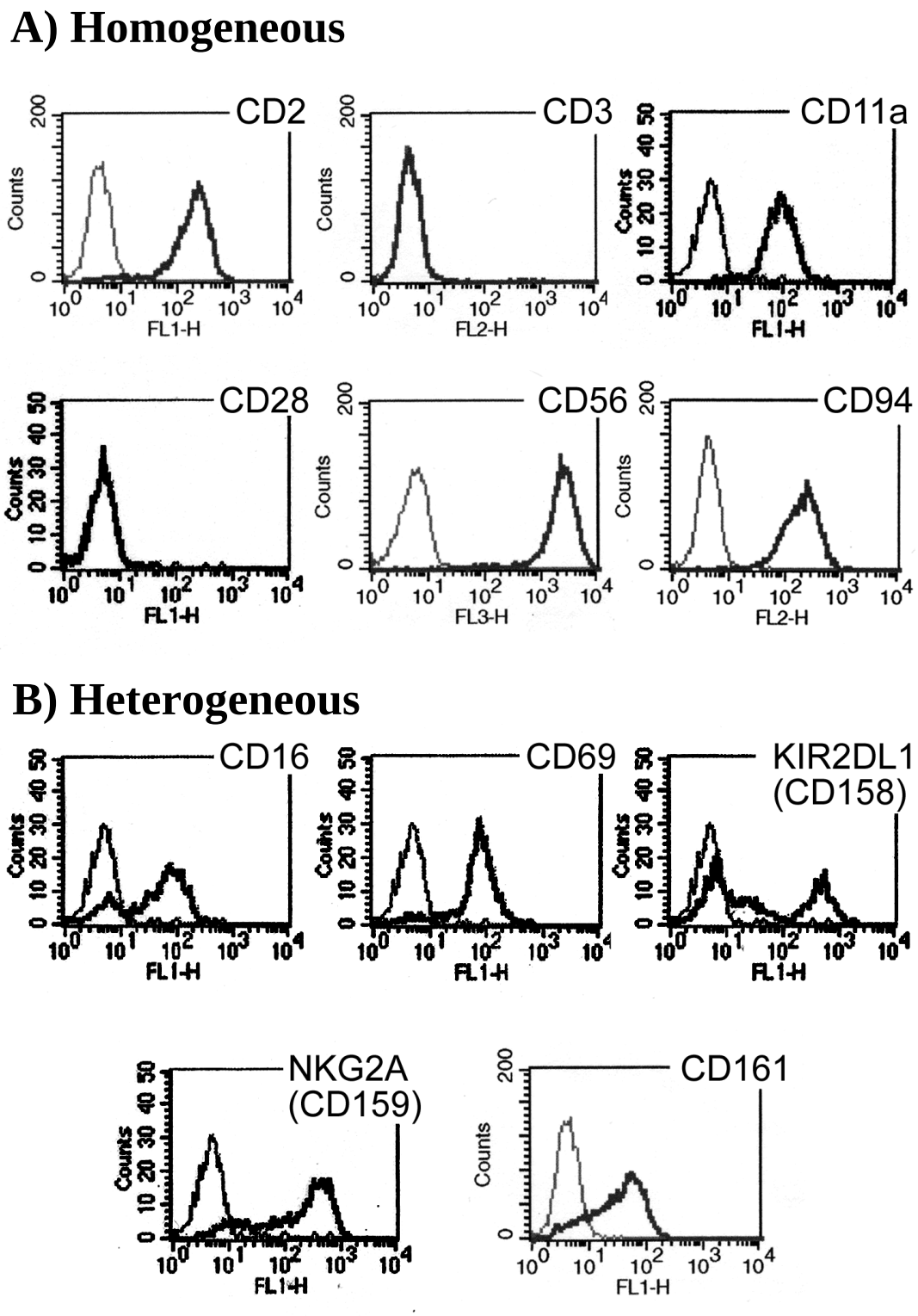


Figure 4.2: Expression of various cell surface markers on NK cells detected by FACS.

FACS staining profiles of NK cells with FITC-, PE- or APC- conjugated antibodies specific for each of the cell surface molecules are shown. The heterogeneous nature of the cell population is clearly demonstrated by the plots in B. The isotype controls are represented by the thinner lines, due to overlap they cannot be seen in the negative FACS plots (i.e., CD3 and CD28). Data from C. Retiere.

Table 4.1: Expression of NK receptors.

The expression of NK cell surface receptors (based on a list by Chiesa *et al.*[190]) was examined in the library presented here and the previously published *ex-vivo* NK library[146]. The expression level in our library was compared to that in cerebellum (marked as cereb.), ovary epithelium (marked as ovary epith.) and a panel of cancers; those transcripts that were specific according to all three criteria were deemed to be "immune specific". P-values were calculated using the Audic-Claverie test[174].

‡ This refers to the number of cancer libraries in which the transcripts were found at over one third the level found in the NK library (maximum 12).
† CD158 refers to the KIR family. Full length sequences for many members of the family were not available, therefore KIR2DL1 (which was also analysed by FACS) was chosen as a representative sequence.

Name	NK	ex vivo NK	Cereb.	Ovary epith.	P-value NK vs. cerebellum	P-value NK vs. ovary epithelium	Cancers[‡]	"Immune specific"
CD2	48.5	23	0	4.2	0.00%	0.00%	0	yes
CD11a	16.6	11.5	0	0	0.10%	0.10%	0	yes
CD11b	2.8	14.4	2	0	62.90%	36.10%	12	no
CD11c	2.8	2.9	0	0	20.20%	21.70%	12	no
CD16a	0	5.7	0	0	-	-	-	no
CD16b	12.5	54.6	0	0	0.50%	0.60%	2	no
CD56	12.5	8.6	17.7	6.2	22.00%	17.00%	2	no
CD59	4.2	0	2	25	45.30%	0.00%	11	no
CD94	1.4	0	0	0	34.40%	36.10%	12	no
CD96	23.6	2.9	0	0	0.00%	0.00%	0	yes
CD158[†]	27.7	14.4	0	0	0.00%	0.00%	1	yes
CD159 (NKG2A / NKG2B)	9.7	0	0	2.1	2.40%	10.70%	0	no
CD160	0	5.7	0	0	-	-	-	no
CD161	13.9	20.1	0	0	0.30%	0.40%	0	yes
CD226	0	0	0	0	-	-	-	no
CD244	8.3	17.2	9.8	10.4	27.60%	24.50%	6	no
CS1; CRACC; 19A24	123.3	163.7	5.9	31.2	0.00%	0.00%	2	no
LAIR-1	1.4	5.7	0	0	34.40%	36.10%	12	no
NKG2C	0	0	0	0	-	-	-	no
NKG2E / NKG2H	1.4	0	0	0	34.40%	36.10%	12	no
NKp30	43	46	0	0	0.00%	0.00%	0	yes
NKp44 (Ly95)	1.4	0	0	0	34.40%	36.10%	12	no
NKp46 (Ly94)	0	0	0	0	-	-	-	no
NKp80	1.4	2.9	2	0	37.10%	36.10%	12	no
NTBA	2.8	0	0	0	34.40%	36.10%	12	no
Siglec-7	0	2.9	0	0	-	-	-	no
Siglec-9	0	0	0	0	-	-	-	no

2. "Immune specific" molecules

In the previous chapter, the 3-stage filter process was demonstrated to enrich a subset of tags, for tags corresponding to proteins with a known immune function. In the $CD8^+$ T-cell clone, all key T-cell markers were found to be "immune specific". Of the twenty seven key NK molecules defined above, twenty were found in the NK library and only six were found to be "immune specific". To understand how representative the "immune specific" tags are, it is important to understand why fourteen molecules are filtered out. Many of the NK receptors are known to be expressed in other cell types[188] and if the receptors were found to be non-specific due to broad expression, it would mean that the 3-stage filter process was unsuitable for identifying NK receptors. Of the fourteen that were not "immune specific", nine were found at too low a level in the NK library (less than six copies per library) to be significant regardless of the expression level in the other libraries, four were found at too high a level in the cerebellum or ovary epithelium (including NCAM which one would expect to be highly expressed in cerebellum) and one was found at too high a level in the cancer libraries. Overall, it appears that the biggest weakness in detecting the NK "immune specific" cell surface molecules is their low expression level in the library. This can be partially explained by the non-clonal nature of the library, which will reduce the effective expression level of some molecules. This was demonstrated by the analysis of the non-clonal *ex vivo* $CD8^+$ T-cells[146] in the previous chapter, where the 3-stage filter process did not identify tags corresponding to all the key T-cell surface molecules. It is also tempting to speculate that this low expression level may be particular to the key immune cell surface components of NK cells. Compared to T-cells, NK cells express a large number of cell surface molecules which trigger or inhibit activation[188, 190]. Hence their activation is not

dependent on a single main receptor (i.e., like the TCR in T-cells) but on a balance of signals from a wide range of molecules. Therefore, in NK cells the breadth of molecules, with an immune function, found on the surface may be more important than having a higher expression level for a smaller number of molecules.

The fact that the majority of cell surface molecules that were detected but were not "immune specific" was due to the expression level in the NK library rather than a broad expression in the other tissues has important implications for what can be studied with the "immune specific" set. It suggests that the logic of the 3-stage filter still stands and that the "immune specific" set can be examined for important molecules in NK function. However, one should not expect to be able to characterise all the key components, as in this case the "immune specific" set will contain only those molecules that are expressed at least moderately in all the cells in the heterogeneous population or highly in one sub-population. With this is mind, the NK SAGE library was passed through the 3-stage filter process. Of the initial 18,963 tags approximately 9% were specific at the 95% confidence level against cerebellum or ovary epithelium (Figure 4.3). Over half of these tags were specific against both tissues. This suggests that transcripts corresponding to tags selected by the filter are likely to constitute a differentially expressed subset. 256 tags survived the filter process and were classed as "immune specific". As in the previous chapter, the proteins corresponding to these transcripts were identified and classified (see Figure 4.4 A for summary and Appendix D for details). However, due to the incompleteness of the "immune specific" set an attempt was not made to estimate the extent of characterisation of the NK cell surface.

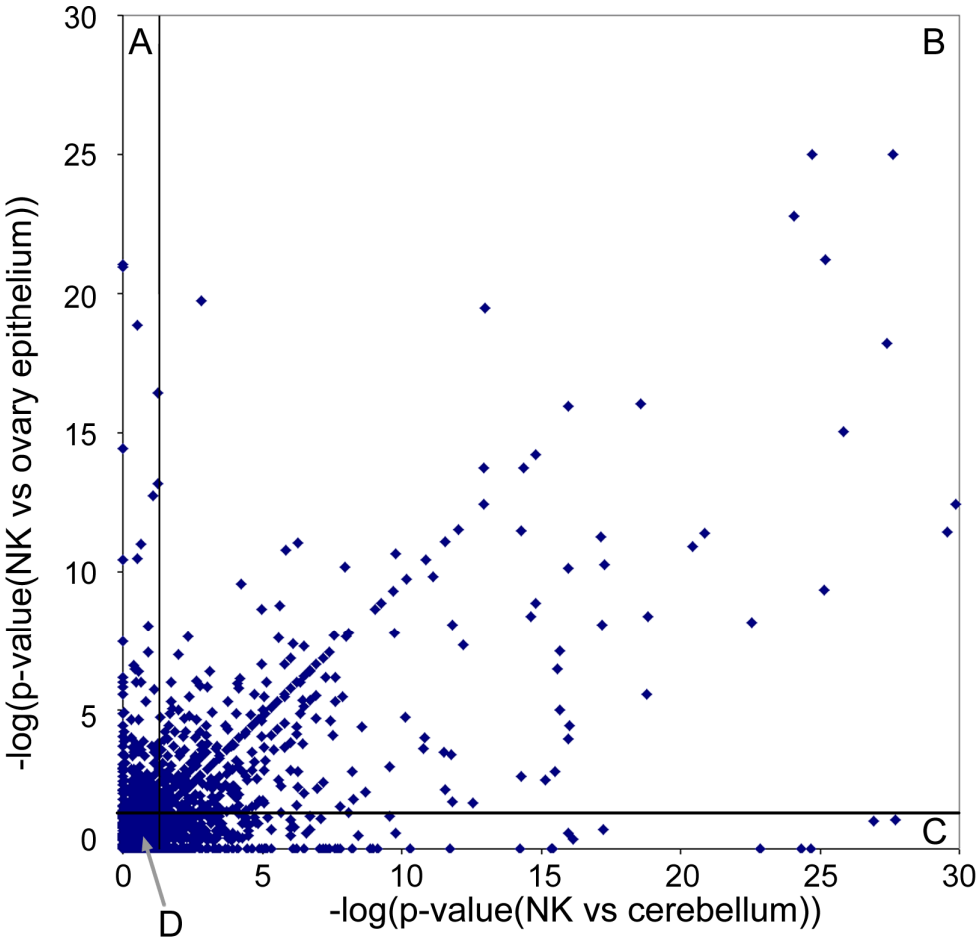


Figure 4.3: Distribution of NK-specific tags.

As part of the 3-stage filter process the NK library was compared to libraries from cerebellum and ovary epithelium. P-values of 5% were selected and the above figure shows the distribution of the upregulated tags. The lines marked in black represent the negative log of a 5% p-value. Quadrant A contains the tags specific against ovary epithelium only (2.11% of unique tags in the NK library). Quadrant B contains the tags that are specific against both cerebellum and ovary epithelium (4.58%). Quadrant C contains the tags that are specific against cerebellum only (2.53%). Quadrant D contains the large majority of tags which are not specific against either cerebellum or ovary epithelium (90.77%). Twenty six outlying tags are not shown.

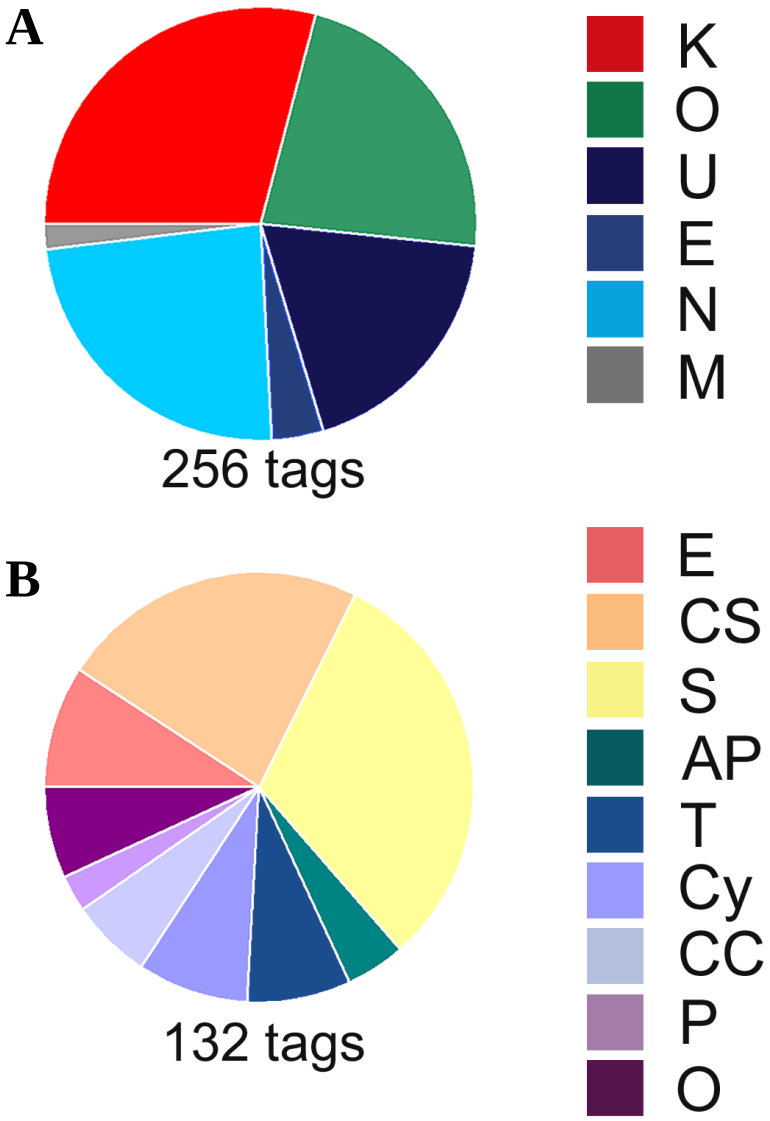


Figure 4.4: Function and extent of characterisation of NK-specific transcripts.

SAGE tags were assigned to transcripts classed as NK-specific by the 3-stage filter process. These transcripts were then categorised according to their level of characterisation (A) and broad function of their protein products (B) using LocusLink, Unigene and RefSeq database entries. Characterisation classes (A) are: K, Known immune function; O, Other known function; U, Uncharacterised cDNA; E, EST cluster; N, No true Unigene match and M, Multiple true matches. Functional classes (B) are: E, soluble Effector molecules; CS, Cell Surface molecules; S, Signalling molecules; AP, Antigen Presentation; T, Transcriptional regulation; Cy, Cytoskeleton related; CC, Cell Cycle or viability related; P, Protein synthesis related; and O, Other.

a) Tags matching Unigene clusters

Of the 256 “immune specific” tags, 132 correspond to single transcripts with a known function (Figure 4.4 A). Interestingly, when examining the function of the known proteins, a similar pattern emerges to that observed in the CTL library, with molecules encoding signalling and cell surface proteins dominating the “immune specific” tags (Figure 4.4 B). The overlap between the “immune specific” classes tags in the CTL library was examined to check that this similarity was not due to the molecules being the same. 118 of the 256 tags were also “immune specific” in the CTL library. These tags were roughly evenly distributed over all the categories, except for the “no matches” where there was only a 21% overlap and the genes encoding proteins with known immune function where there was a 68% overlap. This suggests, somewhat unsurprisingly, that whilst similar proportions of the different categories are up-regulated by a lymphocyte, many of the individual molecules change depending on the lymphocyte being examined.

Of the 58 NK-specific transcripts matching Unigene clusters but lacking any prescribed function, putative complete open reading frames were available for 48 of these transcripts (Appendix E.1). Using bioinformatics tools it was possible to assign possible functions to 31 of the transcripts. Ten of these were predicted to have one or more transmembrane domains, only two of which also contain a signal peptide. Of the other transcripts containing open reading frames, eight apparently encode signalling or adaptor proteins, one encodes a DNA binding domain, two others appear to encode cell cycling proteins and six encode domains linked to rearrangements of the actin cytoskeleton. Finally, one transcript encodes a domain associated with

mRNA processing and protein synthesis while three have domains associated with other housekeeping functions.

The remaining ten tags linked to transcripts with unknown function matched Unigene clusters containing only EST sequences. Possible full length sequences were generated and open reading frames identified for these clusters (Appendix E.2). Six of the putative encoded proteins have recognizable domains, i.e. a probable G-protein coupled receptor, two G-protein-related signalling proteins, a protein kinase, a serine protease and a probable Zinc finger-containing transcription factor.

b) No match tags

In the above analysis, sixty one tags were found not to match any Unigene clusters. These "no match" tags account for approximately one quarter of all the "immune specific" tags and a similar proportion was found in the CTL library. Some of these "no match" tags may correspond to novel transcripts or exons, whereas others will be falsely identified as such due to the tag-to-gene mapping protocol used or technical artefacts. To check whether these no match tags account for genuine novel transcripts various techniques were used. Both bench-based and *in silico* techniques were used to identify the transcripts corresponding to these tags (see Table 4.2 for results of the analyses).

Table 4.2: "No match analysis".

This table contains the results of the five different analyses (two bench-based and three *in silico*) carried out on the "no match" tags.

Tag no	Tag	GLGI product?	3' RACE carried out	3' RACE product?	Summary of GLGI/RACE success	Internal match details	Sequence error	Antisense in RefSeq
1	AAACGCCACTA	Yes			Homo sapiens, granzyme B (polymorphism of tag found in NM_004131, tag found in full length mRNA BC030195.1)			
2	AACACCCTTTC	No	Yes	Yes				Yes (NM_003479)
3	AACGCGAGCAG	Yes						
4	AAGACACCAAG	No	Yes	Yes		internal tag, Homo sapiens T cell activation, increased late expression (TACTILE), mRNA (Hs.142023)		
5	AATACGAGCAG	Yes						
6	AATATGAGCAG	Yes				internal tag, Homo sapiens mRNA for KIAA1586 protein, partial cds. (Hs.180663)		
7	AATGTGAGCAG	Yes						
8	AATTTGAGCAG	Yes				internal tag, Homo sapiens UDP-glucose pyro-phosphorylase 2 (UGP2), mRNA (Hs.77837)		
9	ACTAAGAGCCT	Yes						
10	ACTGGTGGTCA	Yes						Yes (NM_005718)
11	AGCAGCAGAAA	No	Yes	Yes		internal tag, Homo sapiens ankyrin repeat and SOCS box-containing 1 (ASB1), mRNA (Hs.153489)		Yes (NM_016114)
12	AGCTACAGGTG	Yes						
13	AGGCCGTCCCC	No	Yes	Yes		internal tag, Homo sapiens heterogeneous nuclear ribo-nucleoprotein A/B (HNRPAB), transcript variant 1, mRNA (Hs.81361)		
14	AGTGGCAAGGG	Yes						
15	ATCGCGGAGG	Yes			Novel exon 3' of penultimate exon in natural killer cell transcript 4 (NK4, Hs.943)			
16	ATGGGTGGGTG	No	Yes	Yes	Probable repeat			Yes (NM_002841)

Tag no	Tag	GLGI product?	3' RACE carried out	3' RACE product?	Summary of GLGI/RACE success	Internal match details	Sequence error	Antisense in RefSeq
17	ATTTGAGCAGA	Yes				internally primed site, DHX40: DEAH (Asp-Glu-Ala-His) box polypeptide 40 (Hs.29403)		
18	ATTTTGAGCAG	Yes				internal tag, Homo sapiens cDNA FLJ40860 fis, clone TRACH2018034(Hs. 140605)		
19	CAAATCCAAA	No	No					Yes (NM_003618)
20	CAAGGGCTTAG	Yes						
21	CACCAGCTGGA	Yes						
22	CAGATCCAAA	Yes						
23	CAGCGCTGCCG	Yes						
24	CAGTTGGTACT	Yes						
25	CATTTGAGCAG	Yes						
26	CCTTCGAGCAG	Yes						
27	CGGCCCAGGAT	Yes				internal tag, wq58d02.x1 NCI_CGAP_GC6 Homo sapiens cDNA clone IMAGE:2475459 3' similar to contains MER30.b2 MER30 repetitive element; mRNA sequence (Hs.270765)		
28	CTAGCAGAAAC	No	Yes	No				
29	CTCCTGGGCAA	Yes				internal tag, Homo sapiens granulysin (GNLY), transcript variant 519, mRNA (Hs.105806)		
30	CTGATCTCCAA	Yes			Novel 3' exon of CD43 (Hs.461934)			
31	CTGCAGTTATA	Yes			Matches internal restriction site of Homo sapiens IQ motif containing GTPase activating protein 2 (NM_006633.1)	internally primed site, Homo sapiens IQ motif containing GTPase activating protein 2 (IQGAP2), mRNA (Hs.78993)		
32	CTGGGGTGAGC	Yes						
33	CTGTGCGCCCT	Yes				internal tag, Homo sapiens serine protease inhibitor, Kunitz type 1 (SPINT1), mRNA (Hs.233950)		
34	GAAACCGGGCT	Yes				internal tag, Homo sapiens ATPase, H+ transporting, lysosomal 42kDa, V1 subunit C isoform 2 (ATP6V1C2), mRNA (Hs.372429)		

Tag no	Tag	GLGI product?	3' RACE carried out	3' RACE product?	Summary of GLGI/RACE success	Internal match details	Sequence error	Antisense in RefSeq
35	GAGCGGCTACC	No	Yes	Yes		internal tag, Homo sapiens ubiquitin specific protease 21 (USP21), transcript variant 1, mRNA (Hs.8015)		
36	GAGGCCTGGCC	Yes				internal tag, Homo sapiens cDNA FLJ12702 fis, clone NT2RP1000767 (Hs.58582)		
37	GATTGAGCATA	No	Yes	Yes				Yes (NM_002444)
38	GCAAACGACAG	Yes			Contig Match (AC007245, chromosome 7)			Yes (NM_152670)
39	GCAAGTGGGAA	Yes					Lymphotoxin beta (TNF superfamily, member 3)	
40	GCCCGAGGAAG	No	No				Ribosomal protein S12	
41	GCTGAGTGCAG	No	Yes	Yes	EVER2 - novel exon possibly internal of 3' end of shorter transcript variant (Hs.15284)	internally primed site, Homo sapiens A kinase (PRKA) anchor protein 13 (AKAP13), transcript variant 2, mRNA (Hs.301946)		
42	GCTGCCAGGC	Yes					Daughter of GLGI-43	Yes (NM_006990)
43	GCTGCCCAGGC	No	Yes	No		internal tag, Homo sapiens methylthioadenosine phosphorylase (MTAP), mRNA (Hs.152817)		
44	GCTGGGCGCGG	No	Yes	Yes	ALU Y repeat	internal tag, Homo sapiens cDNA FLJ20083 fis, clone COL03440. (Hs.306378)		
45	GGGAGGTATCA	Yes					Granulysin (also possible daughter of basic helix-loop-helix domain containing, class B, 2)	
46	GGGCCAAAGTC	Yes				internally primed site, Homo sapiens FYN binding protein (FYB), mRNA (Hs.58435)		
47	GGGCCAGGTGT	Yes				internal tag, Homo sapiens hypothetical protein FLJ14641, mRNA (Hs.245326)		
48	GGTCCAGAACT	Yes						
49	GTGCCACTGC	Yes						
50	GTGGGCCGGCT	Yes						
51	GTTCGTGCAGA	Yes					Ribosomal protein L35a	

Tag no	Tag	GLGI product?	3' RACE carried out	3' RACE product?	Summary of GLGI/RACE success	Internal match details	Sequence error	Antisense in RefSeq
52	TAAGTTACCAG	Yes			Homo sapiens, platelet activating receptor homolog, clone MGC:46113 (NM_013308)			
53	TAGACCTGACA	Yes						
54	TAGGCACCTGT	Yes						
55	TAGGGTCTTGA	Yes						Yes (NM_006144)
56	TAGGTTGCTA	No	Yes	Yes			Tumor protein, translationally controlled 1	
57	TCCCTATAGC	No	Yes	Yes				
58	TCCTATAAGC	No	Yes	Yes				
59	TCCTATTAGCC	Yes						
60	TGATCAAGGTC	Yes						
61	TGCCTCCTTTG	Yes				internal tag, Homo sapiens, clone IMAGE:3912859. (Hs.326416)		

(i) Bench-based techniques

Initially, laboratory based techniques were used, as one can have high confidence in transcripts amplified from NK mRNA. These laboratory experiments were carried out in collaboration with Janet Fennelly, a research assistant in the Davis Group. Ideally, one would use an amplification based approach to extend these SAGE tags along their corresponding transcripts. At the time of this research, three techniques had been published that attempt to extend SAGE tags (GLGI[164], RAST-PCR[193] and a method devised by Polyak *et al.*[194]). GLGI appeared to be theoretically superior as it was the only technique that cleaved the cDNA with *Nla*III before amplification with tag specific primers, mirroring the initial steps of SAGE library production[94]. This reduces the complexity of the cDNA library and should reduce the number of false positives produced. For this reason GLGI was chosen as the technique to extend the SAGE tags along their transcripts.

Primers were designed, as per the GLGI protocol, for the 61 tags and also for 4 positive controls. The 4 controls picked were tags for: granulysin; ATP synthase, H+ transporting, mitochondrial F0 complex (subunit: ATP5G3); lymphocyte specific protein tyrosine kinase and mitochondrial ribosomal protein S30. These tags were chosen as they covered a wide range of expression levels in the SAGE library (~3 tags to >1000 tags per 100,000). All four controls were identified using GLGI. For the 61 “no match” tags, 45 gave PCR products. These PCR products were then mapped to the genome using BLAST. Only six of the PCR products appeared to be genuine, in that the match to the genome was identical for the *entire* tag sequence and highly similar for rest of the amplified product. Of these six, three matched to previously characterised exons in known genes. It is revealing to examine the reason why these

tags were not mapped to their corresponding transcripts in the initial tag-to-gene mapping. GLGI-1 matched Granzyme B, however it matched a polymorphism at the tag site, which was only observed in one full length sequence and, therefore, using our criteria was not assigned. GLGI-52 also matched a full length sequence (Homo sapiens, platelet activating receptor homolog, clone MGC:46113) but again there were not enough sequences in the database to confirm the tag. Our criteria required a minimum of two cDNAs or 3′ ESTs, with at least one containing a polyA tail or site, to contain the tag for the tag match to be accepted. The fact that two sequences that were genuine failed our criteria suggests that perhaps our criteria were too strict. However, if poor quality and un-oriented sequences were not excluded a large proportion of SAGE tags would be falsely deemed to match multiple transcripts. Therefore, a small number of false negatives seems to be a reasonable compromise. The final tag (GLGI-31) that matched a known transcript (Homo sapiens IQ motif containing GTPase activating protein 2) was found to correspond to an internal restriction site. The other three tags were found to match novel genes or exons. GLGI-15 matched a novel 3′ end of penultimate exon in the 3′ UTR of natural killer cell transcript 4 (NK4). GLGI-30 matched downstream of CD43. A re-examination of the Unigene cluster found 3′ ESTs that contained the tag and are from an alternatively spliced 3′ UTR. GLGI-38 tag matches a region on chromosome 7, where there are ESTs mapped upstream of the tag, so this may be the 3′ end of a novel gene. Alternatively, it is also possible that this is genomic contamination.

Considering the high success rates presented in papers[195, 196] by the group that developed GLGI, the results above are disappointing. There are various possible reasons for this. Firstly, the GLGI library was made separately from the SAGE

library. This means that the random incomplete cleavage events, where *Nla*III does not cleave the transcripts to the 3′ most site, will be different in the SAGE and GLGI libraries preventing elongation of certain tags. Secondly, the mRNA used in GLGI, whilst being from the same NK cell line, was not from the same sample as the SAGE library. Therefore some transcripts present in the SAGE library may not have been present in the mRNA used for GLGI. Finally, some of the tags may not be genuine but in fact caused by technical artefacts, such as sequencing errors. The effect of sequencing errors and incomplete cleavage is investigated below.

For fourteen of the tags, which GLGI produced no PCR products, a 3′ RACE protocol[165] was used to try to identify the transcripts to which the tags correspond. Each tag was matched to the genome and primers were designed for all the exact matches. There were, on average, approximately nine hits to the genome for each tag. The length of primers was set at 21 base pairs to evaluate if this technique would be suitable for using directly with LongSAGE tags (which are 21 base pairs). RACE was carried out for all exact matches to the genome. RACE products were obtained for 11 of the tags. Unfortunately the majority of the products matched the genome, but the match was not complementary over the entirety of the tag, suggesting non-specific amplification. It was possible to assign transcripts to three of the tags. Two of the tags appeared to be transcribed from repeat regions, whilst the third, GLGI-41, appears to be in a novel 3′ exon of EVER 2. The success rate of this modified RACE protocol was higher than for GLGI but still not satisfactory. One reason for the low detection rate may be that the primers were designed on a 2003 build of the genome. If these experiments were repeated with the now “finished” genome sequence[81], one would expect the detection rate to be higher. Due to the fact

that many of the RACE products were incorrect I would not recommend using this approach for LongSAGE tag-to-gene mapping and an alternative approach is suggested in the following chapter.

Overall, laboratory techniques identified transcripts corresponding to ~15% of the "no match" tags. One possible way to improve the rate of transcripts identification would be to combine the GLGI and RACE protocol to perform a "nested GLGI". This would involve two rounds of amplification. For the first amplification, one would use the GLGI primers and for the second one would the used the RACE primers that were designed based on matches the genome. However, the low success rate of both laboratory techniques (compared to the 100% success of the GLGI controls) suggests that some of these tags may correspond to tags that are not derived from the 3′ most restriction site of the transcript.

(ii) In silico techniques

In silico approaches were used to examine three aspects of SAGE library production that may produce tags that do not correspond to tags derived from the 3′ most restriction site of the transcript. There are two well known problems that produce SAGE tags that do not map to 3′ end of known transcripts. The first is the production of "internal tags". These can be caused by two factors, either incomplete cleavage of the cDNA during library production leading to production of tags that are 5′ of the 3′-most tag or internal priming by a stretch of adenosines during cDNA synthesis. The second is that sequencing errors will produce tags with incorrect sequences. Finally, tags were examined to check if they corresponded to tags from antisense transcripts. These have recently been observed in high levels in SAGE libraries[154, 197].

The SAGE Genie database[112] was used to examine "internal tags". To build SAGE Genie, tags were extracted from a variety of databases, including full length mRNAs from RefSeq and MGC, Unigene consensus sequences and ESTs. The various data sources were then ranked according to reliability (this was calculated by comparing how many of the theoretical tags were observed in a panel of tags from 171 libraries). After this ranking process it was possible to assign a "best gene" to each tag. The "best gene" matches are classified, so it is possible to examine if the tag is an internal tag.

All tags in the panel of 44 libraries, described in Chapter 3, were matched to their "best gene". The percentage of unique tags categorised as "internal tags" is shown for each library in Figure 4.5. On average ~23% of the tags are classed as internal. This will be an overestimate of the true level of internal tags, as some of these tags are likely to belong to genes that were not characterised in this build of SAGE Genie. Also, restriction sites that are "internal" in current sequences may become the 3′-most site due to detection of alternatively spliced or polyadenylated transcripts. However, this does suggest that a significant proportion of tags in a library may not be derived from the 3′ most *Nla*III site in the transcript.

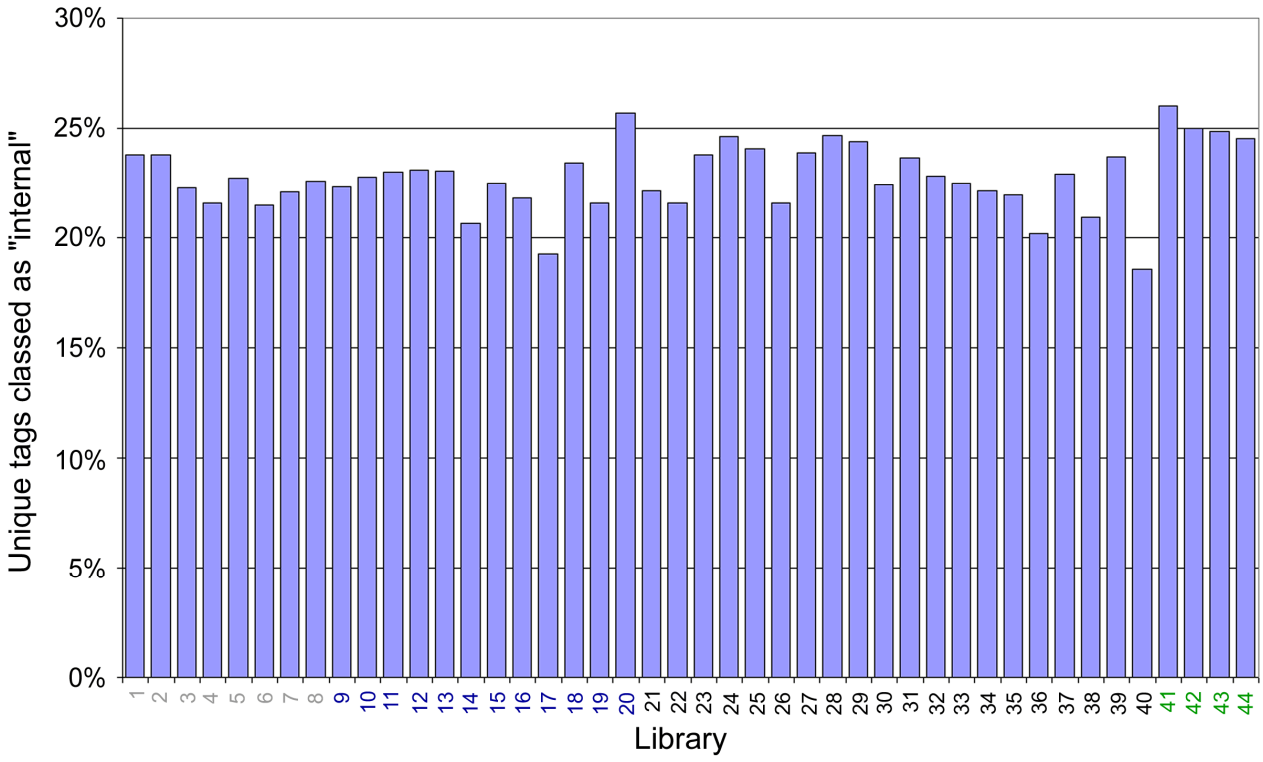


Figure 4.5: Percentage of "internal tags" across the panel of SAGE libraries.

Using the "best gene" annotation in SAGE Genie, internal tags (i.e., tags not derived from the 3′ end most *Nla*III site) were identified in each SAGE library in the panel of 44 libraries. The percentage of unique tags classed as internal was calculated for each library.

The libraries are grouped as in the previous chapter. Grey numbered libraries are myeloid derived libraries, numbers 1-8 represent: Granulocyte, Monocyte, Macrophage-Colony Stimulating Factor Macrophage, Granulocyte Macrophage-Colony Stimulating Factor Macrophage, Immature Dendritic Cells, Mature Dendritic Cells, Langerhans-like Cells and Lipopolysaccharide Monocytes. Blue numbered libraries are lymphoid derived libraries, numbers 9-20 represent: Naive T-cells, Resting Th1, Resting Th2, Activated Th1, Activated Th2, Japanese NK, Japanese CD8 T-cell, B-cell, Activated B-cell, Clone 32, NK, αCD8 activated Clone 32. Black numbered libraries are non-immune cell derived libraries, excluding the retinal derived libraries, libraries 21-40 represent: Cerebellum, Cerebellum Duke, Colon Nc1, Heart, Kidney, Liver, Lung, Peritoneum mesothelium, Prostate, Ovary epithelium, Colon Nc2, Astrocytes, White matter, Vascular endothelium, Breast epithelium, Cervix, Muscle old, Muscle young, Stomach epithelium, Thalamus. Green numbered libraries are retinal derived libraries, libraries 41-40 represent: Retina epithelium, Retina macula, Retina peripheral and Retina peripheral GSM572.

Thirty three percent of the "no match" tags (20 tags) are classed as "internal tags". This figure is higher than that for the entire library (23%), suggesting that the "no match" tags may be biased towards "internal tags". To put these numbers in context the 61 "no match" tags were each randomised ten times and classed by SAGE Genie. Eighteen percent of these random tags were classed as "internal tags". It is not surprising that the "no match" tags may be biased towards the internal tags, as the original tag-to-gene mapping was based on NCBI's SAGEmapper which does not extract internal tags. Of these 20 tags, transcripts corresponding to three had been successfully characterised using RACE or GLGI. GLGI-31 was found to correspond to Homo sapiens IQ motif containing GTPase activating protein 2 by GLGI and SAGE Genie found it to be the internal tag of the same transcript. However, for GLGI-41 RACE found the tag to be from EVER2 whereas SAGE Genie found it to be an internal tag from Homo sapiens A kinase anchor protein 13 (AKAP13). Primers were designed for the internal tag of the transcript and used in the RACE analysis, and this transcript was not identified suggesting that this is a false positive in SAGE Genie. This highlights the point above, that SAGE Genie will be overestimating the extent of internal tags, due to the absence of tags from the 3′ *Nla*III site of uncharacterised transcripts.

There have been a variety of approaches used to identify and/or remove sequencing errors from SAGE libraries (these techniques range from removing all the singletons[187] to a modelled analysis of the "nearest neighbours"[198-200]). Sequencing errors have two main effects on SAGE libraries. Firstly, the apparent complexity of the library is artificially increased, as the number of unique tag sequences is increased

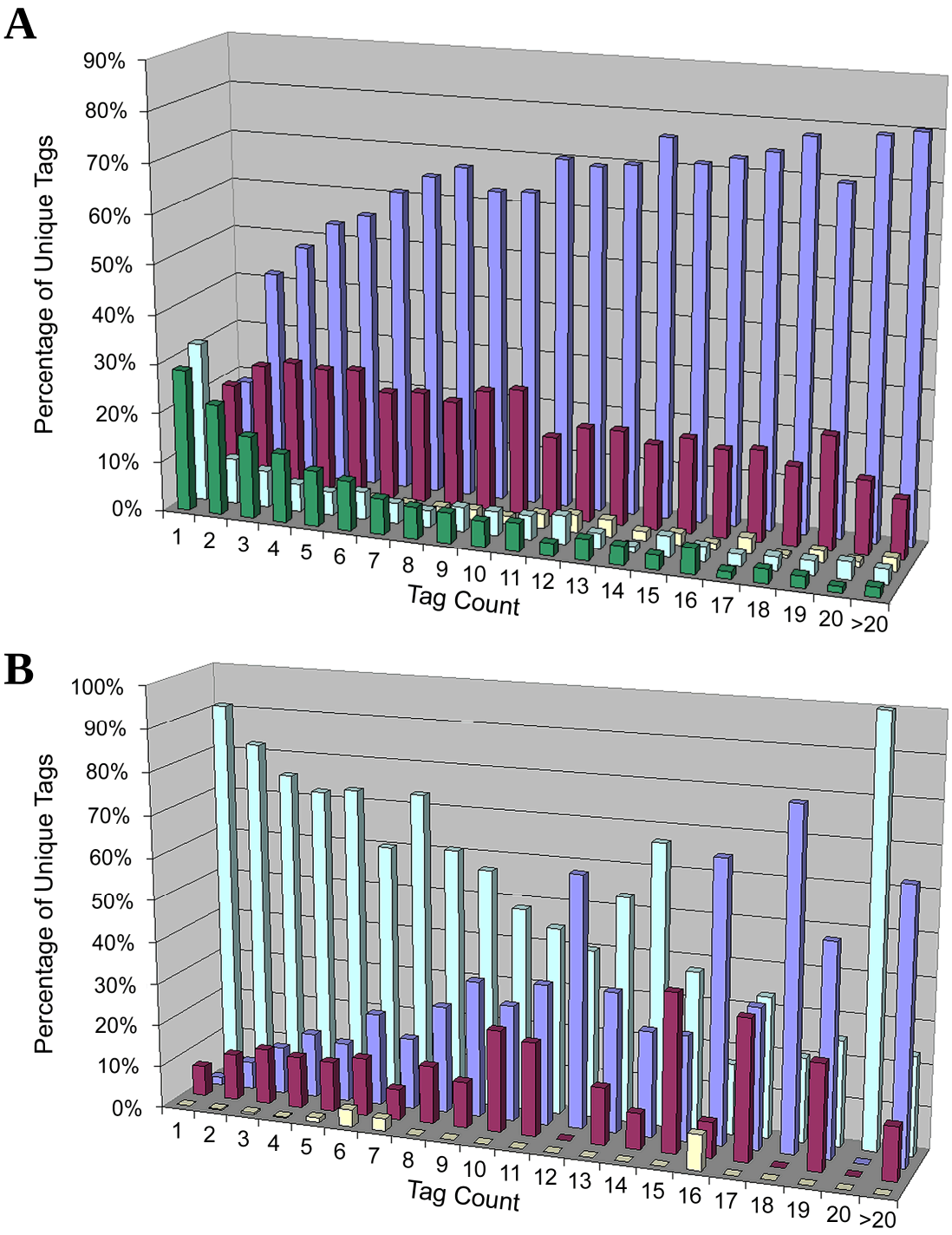
by these false tags. Secondly, the counts of the genuine tags ("parents") are reduced and the counts of artificial tags ("daughters") are increased.

One error removal technique that might seem appropriate is to calculate all possible daughters for each tag and remove those daughters from the library. This will of course be an overestimate, as tags that are similar due to the fact are they are from related genes or just by chance will erroneously be removed. To examine the extent of this effect, daughters (that were possibly due to a single base error) were identified in the largest publicly available SAGE library (SHES2, a LongSAGE library of greater than 400,000 tags derived from human embryonic stem cells, available at www.transcriptomes.org). All tags were classified by whether they matched the genome, RefSeq, genome *and* RefSeq, or nothing (Figure 4.6). One can assume that the vast majority of tags that do not match the genome or RefSeq are sequencing errors. There are two clear trends. Looking at Figure 4.6 A it is clear that as the tag count increases the percentage of daughters decreases. From Figure 4.6 B it is clear that as tag count increases the percentage of daughters that are likely to be sequencing errors decreases. This suggests that in the "no match" tags, which are expressed at least 6 copies per library, there are not likely to be a large number of sequencing errors. This pattern shows that the simplest published approach used to remove sequencing errors, which is to remove all singletons, is poor as while it will have removed many errors it will also have removed a similar number of genuine tags.

Figure 4.6: Distribution of "daughters" in the SHES2 LongSAGE library.

All daughters (i.e., potential sequencing errors) of the LongSAGE tags in the SHES2 library were identified, where possible tags were mapped to the genome or RefSeq. Figure A shows all tags in the library, with the non-daughters being grouped by class (Genome, RefSeq, Genome *and* Refseq, or no match). Figure B shows daughters grouped by class.

	Tag match class
	Genome and RefSeq
	RefSeq only
	Genome only
	No match
	Daughters (Figure A only)


A
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Percentage of Unique Tags
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 >20
Tag Count
B
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Percentage of Unique Tags
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 >20
Tag Count

Of course if the error rate in the SHES2 library is largely different to that in the NK library then the trends may not be applicable. By comparing their tags to the yeast genome, Velculescu *et al.*[201] estimated a maximum sequencing error rate of 0.7 % per base. This equates to a 5% error rate in a ten base pair tag library. To examine the error rate in the NK library presented here, tags were compared and those that were different due to a possible error at one base were grouped. Looking at tag base positions 7 to 9, this gave a per base error rate of 6.6%. This is much higher than that observed by Velculsecu *et al*. and is likely to be due to the fact that many of these putative errors were not in fact single base errors but actually tags corresponding to different transcripts. To confirm this, a LongSAGE library produced in our laboratory was also examined. (LongSAGE tags are seven bases longer than SAGE tags, and so tags corresponding to different transcripts are less likely to differ by *only* one base.) This gave an error rate for bases 7-9 of 0.9%, which is only slightly higher than that observed by Velculesu *et al*. If the tags in this LongSAGE library were artificially reduced to 10 bp, then a much larger error rate was observed (6.1%). This suggests that the high error rate observed in the NK library is due to genuine similar tags rather than errors. As both the NK library and the LongSAGE library were sequenced using the same protocol and equipment, it is fair to assume the error rate of the NK library was the same observed in the LongSAGE library (0.9% per base, which equates to a 9% error rate per ten base tag). Using the same method the error rate in the SHES2 library was found to be 1.0%. The similarity between the SHES2 error rate and the rate observed in our library suggests that the trends observed in the SHES2 are applicable to the NK library. The errors rates calculated here are higher than that observed by Velculescu *et al.*, this is likely to be because the method used here has a relatively high false positive rate (as shown by Figure 4.6 B).

If sequencing errors occur completely randomly then this 5-9% reduction in tag count is unlikely to alter the success of identifying genuinely differentially expressed genes. Equally as there are over 100 possible errors per tag (39 possible insertions, 39 possible deletions and 30 substitutions) it is unlikely that the error tags would appear to be differentially expressed. However, sequencing errors are not produced in a completely random manner[202]. Therefore some tags are more likely to produce errors than other tags and some daughter tags are more likely to be produced than other daughters. This means that it is possible that some of the "no match" tags are daughters of tags from highly expressed transcripts.

The ideal approach to dealing with sequencing errors would correctly identify every error, remove it from the library and add the count to its parent (SAGEscreen, a publicly available program attempts to do this[198]). However, this is not straightforward, with many daughters often found to have many potential parents. Also removal of false positive error tags will remove genuine tags from the analysis. As one of the benefits as using an open technology, such as SAGE, is the ability to detect novel transcripts, anything that can remove genuine tags should be used with caution. Therefore the approach taken here was to identify all daughters of a tag. If any tags in the "no match" list were found to be daughters, then their expression level was compared to that of the parent tag and the tags were categorised depending on the ratio of daughter to parent. Six tags were found to be daughters with the ratio between parent and daughter being greater than ten to one. Interestingly, all six of the potential daughters were tags that the previous analyses had failed to identify.

One final screen was used to examine the "no match" tags. Neither NCBI's SAGEmapper nor SAGE Genie account for the existence of tags derived from antisense transcripts. In Chapter Six it is shown that approximately 10% of total tags in a LongSAGE library are derived from one class of antisense transcripts (*cis*-NATs). To examine *cis*-NAT antisense expression in this SAGE library, all forward and reverse SAGE tags were extracted from every *Nla*III site in RefSeq. Those tags that were only found in the antisense direction were classed as antisense tags. This screen is far from ideal, as the lack of certain genes in RefSeq will produce false positives. Ideally, one would filter out tags that match the genome multiple times, however, the large majority of SAGE tags match the genome multiple times (see Figure 5.1). Nine tags in the "no match" set were potentially found to be antisense tags. Of these nine, five had not been annotated by any of the previous methods.

Overall, of the 61 "no match" tags it was possible to characterise 37. Nine were characterised by laboratory methods and for the remaining 28, possible transcripts were found by *in silico* methods. However, this still leaves ~40% of the tags that are uncharacterised. These must either be due to factors not considered here (e.g. antisense transcripts that are not *cis*-NATs), or factors that it was not possible to exhaustively analyse here (e.g., internal tags matching novel transcripts). The original aim of this work was to establish if the "no match" tags were genuine and it appears that some of them are, but that some are due to technical artefacts of SAGE or limitations of the tag-to-gene mapping protocol. It is clear that the ideal tag-to-gene mapping protocol, as well as using the 3′ most tag, needs to include internal tags, antisense tags and tags downstream of defined transcripts. The tag-to-gene mapping protocol developed in the next chapter includes all of these features.

V. CD4+ T-cell transcriptome

A. Introduction

In this chapter I present an initial characterisation of the entire transcriptome of an activated CD4+ T-cell clone. In the previous chapters SAGE has been used to examine the expression level of genes encoding cell surface proteins in a variety of tissues. However, three factors make SAGE unsuitable for characterising an entire transcriptome. These factors are: the uniqueness of tag to transcript mapping, the depth of sampling and the breadth of sampling.

1. Unique transcript identification

One of the advantages of tag based expression technologies over EST or cDNA sequencing is that they can rapidly and relatively inexpensively sample a large number of transcripts. However, reducing the tag length, whilst decreasing the cost per transcript sampled, also reduces the ability to identify the transcripts that the tags represent. For any tag based expression technology, each tag will ideally map uniquely to both the genome and the transcriptome. Theoretically the 14 bp SAGE tags should map uniquely to the transcriptome[94]. However, because the transcript sequences are non-random, 14 bp tags are too short to distinguish between similar sequences. Even more significantly, as demonstrated in the previous chapter mapping 14 bp tags to the genome generally generates multiple hits. This makes identifying novel genes difficult. However, there are two commercially available tag technologies which generate longer tags. One is LongSAGE[95], which is a modification of the standard SAGE protocol using *Mme*I rather than *Bsm*FI for the second cleavage step. This generates tags that are 21 bp. The second technology is

MPSS[93]. This generates 20 bp tags, which start with GATC rather than CATG, as *Dpn*II is used instead of *Nla*III in the first cleavage step. Figure 5.1 demonstrates the effect of increasing the tag length on transcript mapping by analysing the actual frequencies of tags in the entire human genome. In both technologies the vast majority of tags are unique in the genome and transcriptome (>94%). Therefore both technologies largely solve the problem of unique transcript identification for characterising the entire transcriptome.

2. Depth of sampling

It has been estimated that there are ~400,000 different transcripts in a mammalian cell[203]. Therefore to gain complete transcriptome coverage by any tag based expression technology it will be necessary to sample several times this number of tags. Due to the proprietary sequencing technique used in MPSS it is feasible to sequence over a million tags per sample, at a cost of about $30,000. To sequence a similar number of LongSAGE tags would cost over 10 times more. Therefore, for most labs MPSS is the only tag based technology that provides the depth of sampling required to cover the whole transcriptome.

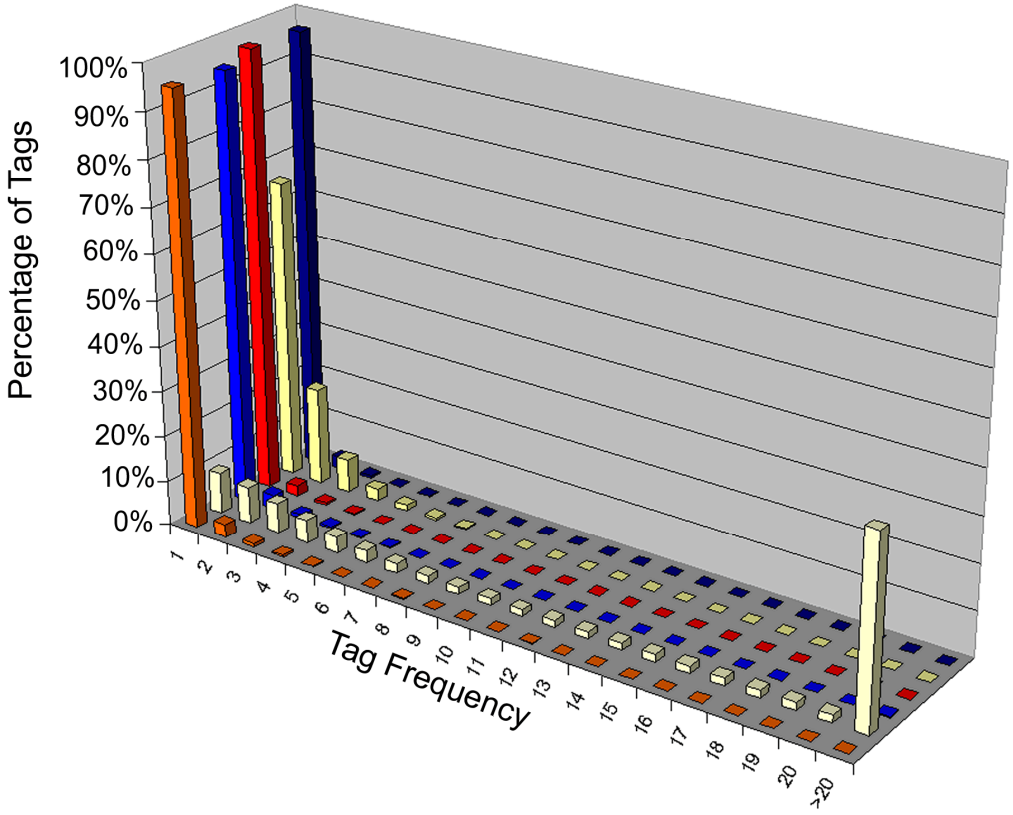


Figure 5.1: Effect of tag length on frequency of matches to the genome and transcriptome.

All tags were extracted from the Ensembl genome and transcriptome. The frequency of each tag in the genome or transcriptome was calculated and tags were grouped by frequency. The different tag lengths chosen correspond to the different techniques used in this thesis: 14 bases, SAGE; 21 bases, LongSAGE; and 20 bases, MPSS.

	Tags extracted from	Restriction site	Tag length (bases)
	Genome	CATG	21
	Genome	CATG	14
	Genome	GATC	20
	Transcriptome	CATG	21
	Transcriptome	CATG	14
	Transcriptome	GATC	20

3. Breadth of sampling

For a restriction site based technique to detect a given transcript, the starting transcript must contain the recognition site. Using *Nla*III or *Dpn*II, which have a four bp recognition site, one would expect to find the recognition site once every 256 (4^4) base pairs. However, some sequences will not contain the recognition site. As every restriction site based technology suffers from the same weakness and both LongSAGE and MPSS use 4bp recognition sites they should be both similarly affected. In the RefSeq database[167], the number of full length cDNAs that do not contain the LongSAGE recognition site is less than 1% (132/22183). While this number is small, it should not be overlooked when trying to catalog the entire transcriptome.

Neither MPSS nor LongSAGE can sample all transcripts, but they do provide an improved tag-to-gene mapping when compared to SAGE. It seemed that the ideal strategy would be to use both techniques, as this would provide both the depth and breadth of coverage required. Only 36 of the 22,183 sequences in RefSeq did not contain one of the two recognition sites.

In addition, one of the major benefits of using an open expression technology is the potential to examine the gene expression of previously uncharacterised genes. However, identification of the novel genes is not straightforward, as was demonstrated by the limited success of the GLGI in the previous chapter. Whilst GLGI has been adapted for MPSS[204] we felt a combination of LongSAGE and MPSS might be a more powerful approach. Regions of the genome that are not known to

code for any genes, but which contain tags found by *both* techniques would be likely to contain novel genes.

B. Results and Discussion

As described in the materials and methods chapter a LongSAGE library and an MPSS library were made from a single sample of RNA from a CD4+ T-cell clone. 385,475 LongSAGE tags were sequenced. 4,271,516 MPSS tags were sequenced in total, but unless stated otherwise, only 1,744,173 tags were available for the analysis presented below.

The first simple test to apply to both libraries is to check that tags corresponding to genes encoding proteins detected by FACS analysis are found. FACS analysis by Julian Sutton showed that these cells were positive for CD4, CD28, CD45 and CD69, and negative for CD27 and CD62L (ref:[162]). The LongSAGE data perfectly matched the FACS results and the MPSS data agreed with FACS for all but CD69, which was absent in the MPSS library. For CD69, the observed LongSAGE tag was in the 3′ UTR but upstream of the only *Dpn*II site in the full length transcript. Between the *Nla*III site and the *Dpn*II site there is a potential polyA signal, and therefore the absence of MPSS tags for CD69 may be due to alternative polyadenylation producing a shorter alternative transcript that lacks a *Dpn*II site.

1. Technique comparison

The ideal way to compare the libraries would be to compare the expression level of all the different transcripts present in the cell. Due to ambiguities in tag-to-gene mapping and certain transcripts lacking at least one of the required recognition sites it is not possible to compare all transcripts. However, a set of transcripts was extracted

from Ensembl[205] which contained *only Nla*III *and Dpn*II derived tags that matched to at most one position in the genome and one position in the Ensembl transcriptome. This set contained ~6000 transcripts and shall be referred to as the UTBS dataset (Unique Transcripts Both Sites, see Chapter 2 for further details). The Spearman correlation coefficient for the comparison of the expression of transcripts from the UTBS dataset in the two libraries was ~0.64. To put this into context, various other pair-wise comparisons were made, where there was a difference in at least one of the following properties in the libraries compared: (a) tissue/cell line, (b) activation state or (c) technique used (Table 5.1). A clear and expected pattern emerges, in that those comparisons with only one difference have a higher correlation than those with multiple differences. However, a larger correlation is observed when comparing two LongSAGE libraries that differ in either activation state or tissue than when comparing the two libraries presented here. As both these techniques have been shown to give highly reproducible results[107, 206, 207] it suggests that the differences observed between the two libraries must be due to some form of systematic bias in at least one of the techniques. To examine the source of this bias, the libraries were compared at two levels, the depth of sampling and the breadth of sampling.

2. Depth of sampling

To examine whether one has sampled enough tags, one can look at the rate of addition of novel tags to the library. Once all available tags have been sequenced, the rate of addition of novel tags should be zero. One would expect a curve of total tags sequenced versus unique tags to become asymptotic if all transcripts have been sampled. As expected, Figure 5.2 shows this to be the case with MPSS but not with LongSAGE.

Library 1	Library 2	Tissue/Cell line	Technique	Activation state	Correlation coefficient
CD4+ T-cell, activated, LongSAGE	CD4+ T-cell, activated, MPSS	Same	Different	Same	0.637431
CD4+ T-cell, activated, LongSAGE	CD4+ Resting, LongSAGE	Same	Same	Different	0.828269
CD8+ T-cell, α-CD3 activated, LongSAGE	CD4+ T-cell, activated, LongSAGE	Different	Same	Same	0.768597
CD8+ T-cell, α-CD3 activated, LongSAGE	CD4+ T-cell, activated, MPSS	Different	Different	Same	0.548839
CD4+ T-cell, activated, MPSS	BT-20 Cell line, MPSS	Different	Same	Different	0.449896
BT-20 Cell line, MPSS	CD4+ T-cell, activated, LongSAGE	Different	Different	Different	0.463064
CD4+ T-cell, activated, MPSS	CD4+ Resting, LongSAGE	Same	Different	Different	0.588749

Table 5.1: Comparisons of the effects of different factors on the similarity of libraries produced.

Using the UTBS dataset, it is possible to compare different techniques. Sense tag counts for all transcripts were compared and correlation coefficients produced. The correlation coefficients are Spearman coefficients on Log transformed data. The CD8+ T-cell, α-CD3 activated, LongSAGE library was produced by S.H.I. Abidi. The BT-20 MPSS library is made from an estrogen receptor-negative breast cancer cell line and is courtesy of Lynx Therapeutics[208].

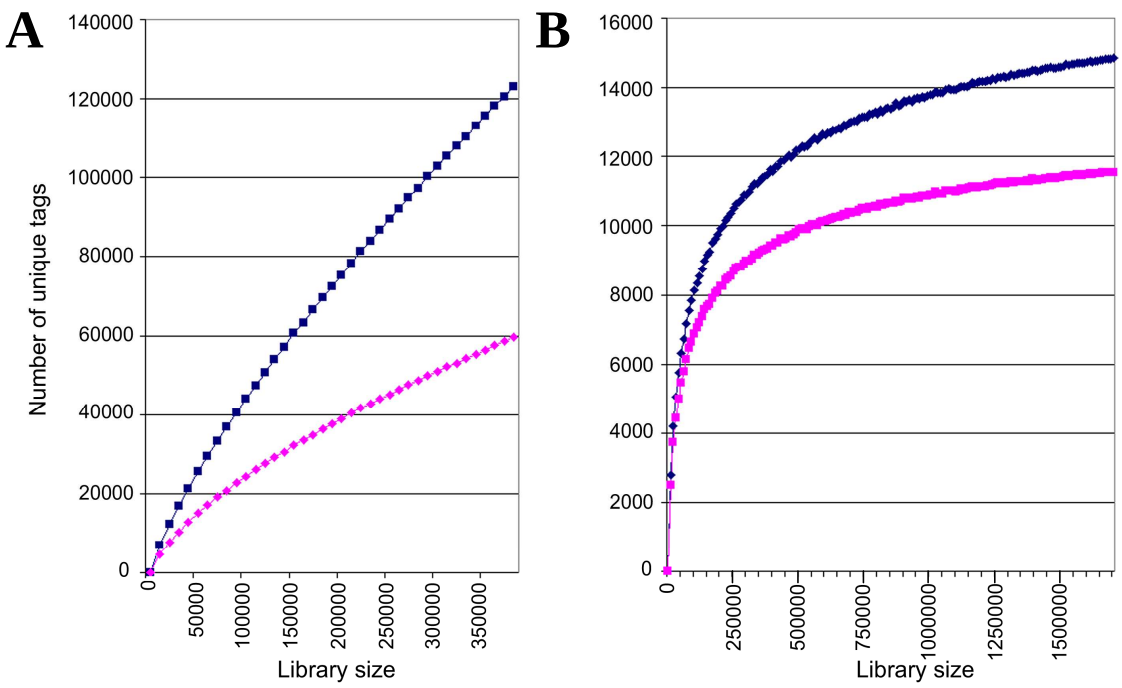


Figure 5.2: Effect of total number of tags sequenced on number of unique tags identified.

The LongSAGE and MPSS CD4+ T-cell libraries were sampled at various sizes to examine the effect of library size on the number of unique tags identified. If the library is large enough to sample all available tags, then increasing the library size will not increase the number of transcripts detected. Graph A represents the LongSAGE library, Graph B represents the MPSS library. Blue symbols represent all tags in the library. Pink symbols represent those tags that match either the genome or the transcriptome.

However, the rate of novel tag addition may be artificially increased in the LongSAGE library due to sequencing errors. As was demonstrated in the previous chapter, one of the major problems when trying to interpret SAGE data is the accumulation of sequencing errors. MPSS has been reported to have much lower error rates than LongSAGE, in one case the error rate was estimated to be ~3.2% [209]. The false tags generated by sequencing errors will increase the number of unique tags detected and prevent the curve flattening out. A simple filter was used to remove tags caused by sequencing errors; only tags that matched either the genome or the known transcriptome were kept. Some genuine tags, generated by polymorphisms, splice variants or alternative polyadenylation, may not match either of these datasets and so will be removed. However, this will not have a large effect when considering the properties of the whole transcriptome[210]. The percentage of total tags removed remains almost constant for both the MPSS and LongSAGE libraries at the different sampling points (6.1% [σ 0.05%] and 26.4% [σ 0.12%] respectively). As can be seen in Figure 5.2 once the error-derived tags are removed, the rate of novel tag addition slows dramatically in the LongSAGE library, although as expected the graph is still not asymptotic.

A different way to examine sampling depth is to look at the frequency of raw tag counts for those tags that match the genome. Assuming that the majority of transcripts are expressed at a low level, if there has been over-sampling one would expect the frequency of tags seen only once to be less than the frequency of tags seen twice. Conversely, if there has been under-sampling then one would expect the frequency of singletons to be higher than the frequency of the tags seen twice. The magnitude of the difference between these two frequencies can provide an indication

of how much further sampling is required. Figure 5.3 shows this data for the two libraries. This gives a similar result to that of the previous analysis. In the MPSS library the frequency of singletons and doubletons are similar, 6.77 % of all tags and 5.81% respectively. This contrasts with the LongSAGE data, where the frequency of the singletons is nearly an order of magnitude greater than that of the doubletons (66.1% and 11.6% respectively).

A third way to examine the depth of sampling is to look at a subset of tags that correspond to known genes. The counts for the tags that matched the transcripts in the UTBS dataset were analysed and used to examine the effect of increasing library size on the sampling of known genes, the results of which are shown in Figure 5.4. Again it is clear that the MPSS library has sampled almost all transcripts, while the rate of transcript discovery in LongSAGE is decreasing, this suggests that this library has enough tags to sample most known transcripts.

Looking at the tags that match the genome and at those in the UTBS dataset provides different answers to how many tags need to be sequenced to sample the entire transcriptome. All methods suggest that MPSS has sampled deep enough to cover the entire transcriptome. Mapping tags to the genome suggests a LongSAGE library needs to be much larger than 380,000 tags to sample all transcripts, whereas mapping tags to the UTBS dataset suggests that 380,000 tags is almost enough to sample all known transcripts. This difference is not surprising. Known genes are likely to be expressed at a higher level than novel transcripts, aiding their initial identification[82, 211]. This means that fewer tags will need to be sampled to identify all the previously known transcripts expressed in a cell.

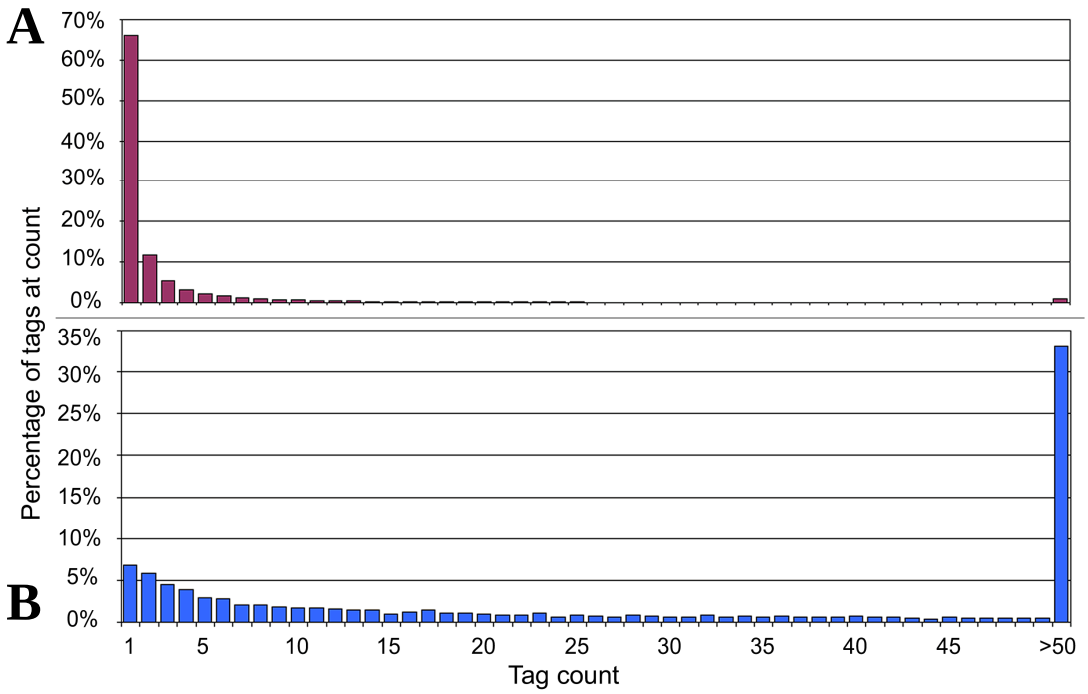


Figure 5.3: Histogram of tag counts observed in the LongSAGE and MPSS libraries.

All tags in a given library (LongSAGE or MPSS) that matched the genome were selected. The tags were then grouped together by their observed count. Graph A represents the LongSAGE library, Graph B represents the MPSS library. Assuming that the majority of transcripts are expressed at a low level, then if a library has over sampled the underlying transcript population, the percentage of tags observed with a count of 1 will be less than that of those with a count of 2. It is clear from the graphs that the MPSS library appears to be almost over-sampling the transcript population, whereas the LongSAGE library is not.

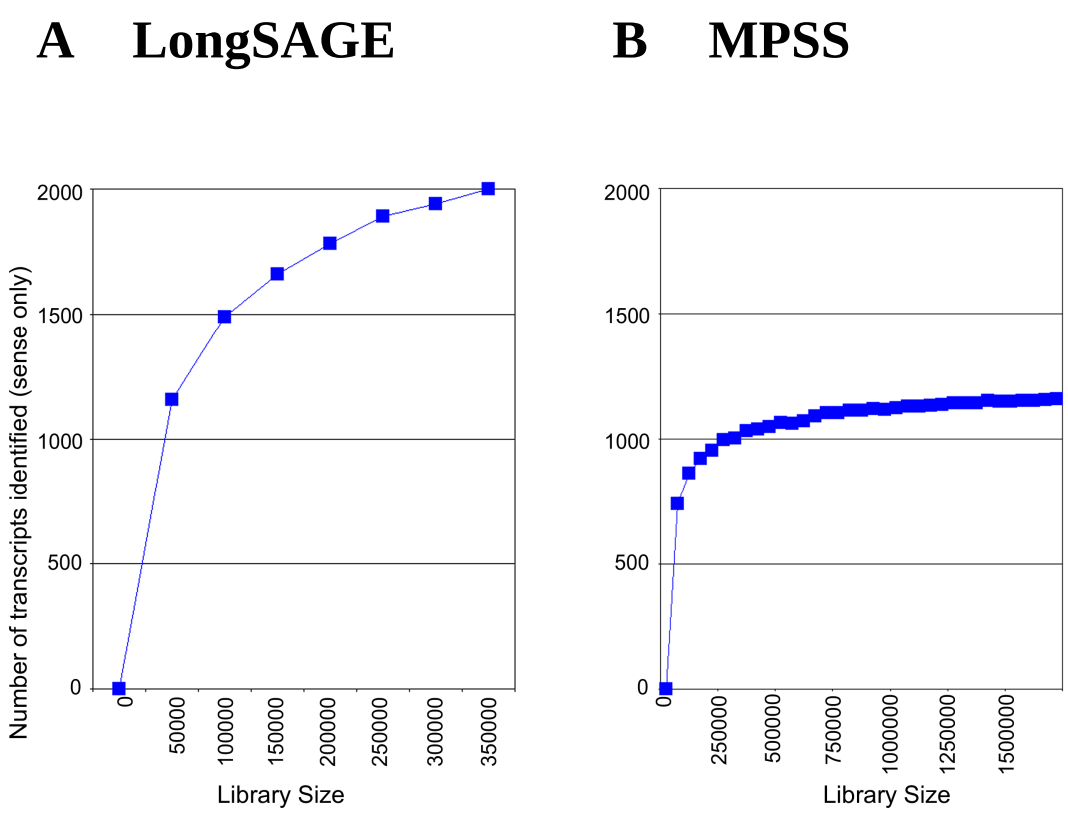


Figure 5.4: Number of transcripts in the UTBS dataset identified by LongSAGE and MPSS.

The UTBS dataset consists of transcripts for which all extracted tags are unique in both the transcriptome and the genome. The libraries were sampled at various sizes and the numbers of transcripts from the UTBS dataset that were found were calculated. Graph A represents the LongSAGE library, Graph B represents the MPSS library.

3. Breadth

The depth of sampling is only one aspect of a sequencing technology to consider. Closed expression technologies, such as microarrays, can also show deep sampling but are limited by the breadth of sampling, i.e. what is on the array. It is important to check that these open expression technologies are truly open.

Whilst it appears that the MPSS library is large enough to sample all transcripts and that the LongSAGE library needs further sequencing there is a large difference in the number of unique tags in both libraries. At the same sampling depth (380,000 tags) there are over ten times the number of unique tags in the LongSAGE library than in the MPSS library (123,202 vs. 11,478). As has been noted previously, the sequencing error rate is thought to be significantly lower in MPSS than in LongSAGE. To reduce the impact of sequencing errors one can look at the unique tags that are found in either the genome or the transcriptome. In this reduced error data one finds that LongSAGE still identifies many more unique transcripts than MPSS (59,586 vs. 11,544). This large difference in the number of unique tags suggests that either the LongSAGE library contains many spurious tags or the MPSS library is missing many genuine tags.

It is not possible to directly convert the number of unique tags found in a library to the number of genes being profiled, because there are multiple true tags per gene (due to polymorphisms, alternative polyadenylation, antisense expression and incomplete cleavage by the restriction enzymes) and, in some cases, multiple genes per tag. However, one can examine the number of transcriptional loci identified. As this was not an attempt to find specific genes or give an exact gene number, but rather to

compare the two techniques, a very simple set of criteria were used to define a transcriptional locus. Briefly, all the tags that matched the genome once were sorted by chromosome position. A tag was considered to be in a new transcriptional locus if it was greater than X bases away from the last tag or Y bases away from the first tag of the last locus. Table 5.2 shows that whatever criteria were used the LongSAGE library identified a greater number of loci than the MPSS (2.6-3.3 fold more loci in LongSAGE).

Looking at the UTBS dataset should provide a way to determine if the difference is due to spurious LongSAGE tags or missing tags in the MPSS library. This shows a similar trend to the analysis of all unique tag counts, with the LongSAGE library identifying almost twice as many expressed transcripts as the MPSS library. This can be used to provide an estimate of the number of genes being sampled in each library. 16% of the total tags in the MPSS library represented 1189 known transcripts and 11% of the LongSAGE tags represented 2082 transcripts. This gives estimates of ~7,000 transcripts being sampled in the complete MPSS library and ~19,000 in the entire LongSAGE library. These numbers are likely to be underestimates because, as stated previously, the known transcripts are likely to be expressed at a higher expression level than the novel transcripts. Interestingly, these numbers are very similar to those found when estimating the transcriptional loci, using a maximum distance between tags of 20,000 bp (Table 5.2).

Max. loci length (bp)	Max. distance between tags (bp)	Number of loci		Corrected number of loci		Ratio of loci found by LongSAGE : MPSS
		Long SAGE	MPSS	Long SAGE	MPSS	
50000	5000	22831	6997	28671	8757	3.27:1
50000	10000	18830	6385	23647	7991	2.96:1
50000	15000	16514	5954	20738	7452	2.78:1
50000	20000	15106	5654	18970	7076	2.68:1
100000	5000	22831	6997	28671	8757	3.27:1
100000	10000	18818	6385	23632	7991	2.96:1
100000	15000	16384	5951	20575	7448	2.76:1
100000	20000	14730	5640	18498	7059	2.62:1
150000	5000	22831	6997	28671	8757	3.27:1
150000	10000	18818	6385	23632	7991	2.96:1
150000	15000	16378	5951	20567	7448	2.76:1
150000	20000	14708	5640	18470	7059	2.62:1
200000	5000	22831	6997	28671	8757	3.27:1
200000	10000	18818	6385	23632	7991	2.96:1
200000	15000	16378	5951	20567	7448	2.76:1
200000	20000	14705	5640	18466	7059	2.62:1

Table 5.2: Number of transcriptional loci identified.

Using the location of tags that mapped uniquely to the genome it is possible to estimate the number of transcriptional loci. Two factors were altered to establish if they had any effect on the ratio of loci found in LongSAGE vs. MPSS. The factors were: maximum loci length (variable Y in the text) and the maximum distance between any two consecutive tags in the loci (variable X in the text). The "corrected" figures take into account that in each library only a proportion of tags match the genome uniquely. The correction factors were 1.26 for LongSAGE and 1.25 for MPSS.

To put these gene numbers in context, the total number of genes in the human genome is still under debate, but the current consensus places it under 30,000 genes[212] (possibly below 25,000, ref:[81]). Work trying to estimate the number of transcript species expressed in one cell has come up with widely varying numbers. Studies on mouse brain have estimated that there are between ~10,000 (ref:[203]) and ~100,000 (ref:[213]) transcript species per cell. Gene expression has been shown to be a stochastic process[214] and so one might expect that, if one can sample a pure population of cells deep enough, every gene could be detected.

It seems clear that whichever way you look at the data, MPSS is underestimating the true complexity of the transcriptome. To examine the differences in complexity discovered by LongSAGE and MPSS it is revealing to examine the expression level of the different classes of transcripts in the UTBS dataset. The average expression level for *all* transcripts found in the sense direction by LongSAGE was 49.3 tags per million (tpm) (2032 transcripts in total); for transcripts found by *both* techniques, it was 77.1 tpm (1082 transcripts) and for those by LongSAGE *only*, it was 17.7 tpm (950 transcripts). This suggests that MPSS is not able to detect the lower expressed transcripts. This is contrary to what one would expect and may be due to systematic bias in MPSS sequencing, as described below, or may be due to factors affecting other facets of library production.

The MPSS library analysed here is actually one part of a three part library, hereafter referred to as the "entire library". Whilst writing this chapter the final batches of data were delivered, so it was possible to compare the different libraries. Only one cDNA library was made from the mRNA, but this was then used to produce three

separate “bead libraries”. As it appears that the sampling depth of the “bead library” analysed above is enough to sample all available transcripts, one would expect that when comparing it to the other “bead libraries” they would be essentially the same. In fact, although they appear similar when looking at the total number of unique tags, with ~14,500 unique tags being found in each library; the majority of unique tags are only found in one of the three libraries (see Figure 5.5). Those tags only found in one library are expressed at a much lower level than those found in all three libraries (averages of 3.1 tpm vs. 191.4 tpm). This suggests that random sampling during the MPSS production process has a large effect on the resulting “bead library”; it appears to limit the complexity of an individual library and this leads to large differences in which of the tags corresponding to lowly expressed transcripts will appear in the library.

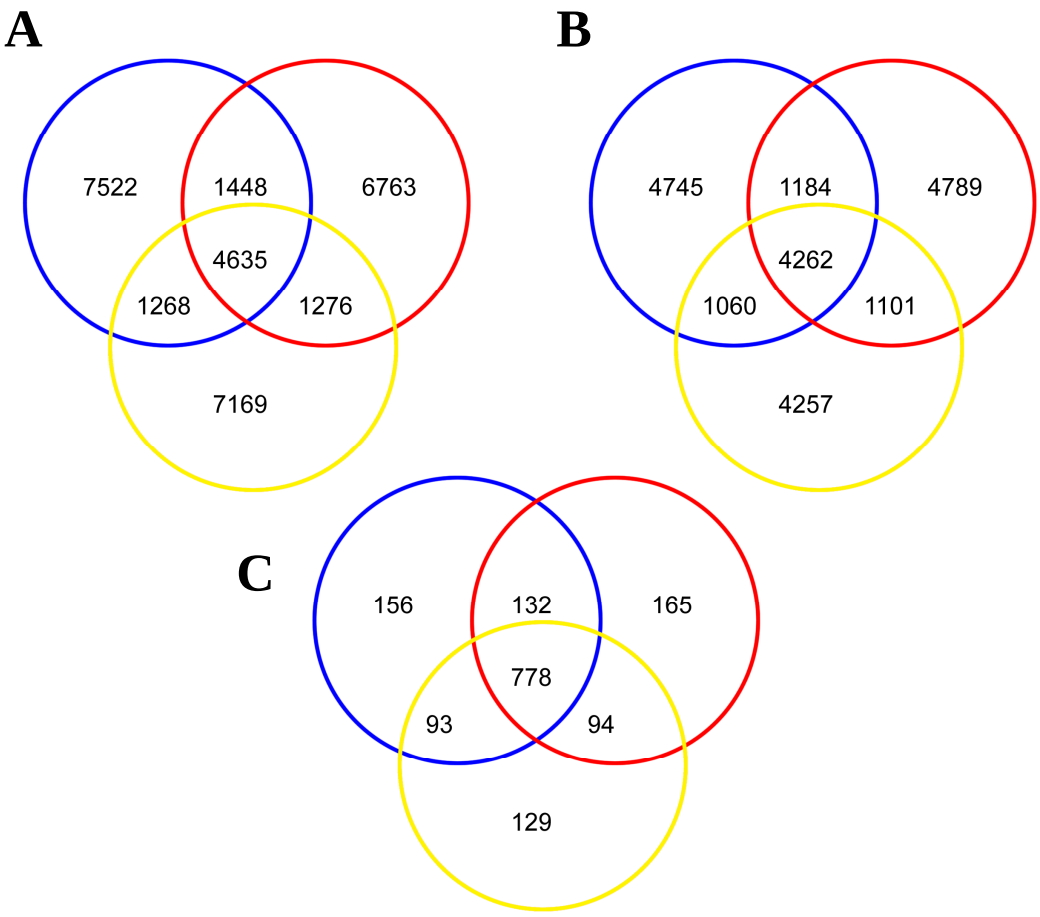


Figure 5.5: Venn diagram comparing the number of unique tags in individual MPSS "bead libraries".

The unique tags in the three individual "bead libraries" making up the "entire library" were compared. Panel A represents all unique tags in the individual libraries. Panel B represents only those unique tags that match the genome; this was to reduce the influence of sequencing errors. The pattern is the same in both comparisons, with the majority of distinct sequences being found in only one library. Panel C represents known transcripts in the UTBS dataset found expressed in the sense direction. Here the pattern is less marked, but only half the transcripts were observed in all three libraries (778/1558). The pattern is likely to be less marked because known transcripts (i.e. those in the UTBS) are likely to be expressed at a higher level than novel transcripts, and therefore have a higher chance of being sampled. The library represented by the blue circle was used in all the other analyses presented in this chapter.

There are four steps in the MPSS protocol where random sampling could have a large and limiting effect. The first is the sampling of the mRNA to make the cDNA library (this also occurs in LongSAGE library production). The second is the sampling of the cDNA to produce the tags. At this stage 1.28 million sequences are sampled from the cDNA library, converted to tags and then transformed into bacteria. The next sampling step occurs when ligating the tags to the beads. The final sampling step is the sampling of beads with tags attached when loading a flow cell for sequencing. Inefficiencies at any of the stages will have a greater effect on the detection of tags corresponding to lowly expressed transcripts than those corresponding to the higher expressed transcripts, i.e. for the lowly expressed transcripts the effect will be that they are not observed, whereas for the higher expressed transcripts the effect will be to reduce their count. The overall effect this will have on a library is to reduce its complexity and increase the percentage of tags corresponding to the more highly expressed transcripts.

This leads to the question, how many individual "bead libraries" are required for an "entire library" to have the breadth of coverage necessary to catalogue an entire transcriptome? Approximately 1,500 transcripts are identified in the UTBS dataset when using the "entire library". This is still much less than those identified by LongSAGE, therefore one can say more than three libraries are required. To answer this question it would be useful to examine other published libraries. Unfortunately, there is a paucity of published and freely available MPSS libraries made from human cells. In fact, for the six papers published using human MPSS data, no groups have currently released their data in full[107, 208, 215-218]. Therefore, at present it is only

possible to say that more than three "bead libraries" are required to characterise an entire transcriptome using MPSS.

It should also be noted that the LongSAGE library analysed here also consists of two libraries. Both libraries were made from the same source of mRNA but using different SAGE reactions. These libraries were sequenced to different depths, with the majority (>89%) of tags coming from one library. When looking at transcripts detected, 32 transcripts were found only in the smaller library. This equates to approximately 3.2% of all transcripts found in the sense direction in the smaller library. This suggests that sampling may also have a small impact on LongSAGE libraries. However, one should also note that novel LongSAGE tags were still observed as the library was sequenced deeper, therefore one might expect that these 32 transcripts would be observed in the larger library if it was sequenced deeper.

It has been demonstrated above that if enough libraries are produced then the complexity of an MPSS library may be increased. However, there are large differences in the sequencing technologies used in LongSAGE and MPSS and this also appears to impact on library complexity. The sequencing technology in MPSS has steps involving 4 bases and it has been previously noted that MPSS has problems with palindromic 4 base words (e.g. TTAA)[209, 219]. To reduce the effect of this bias, MPSS sequencing is carried out in two staggered phases, which should increase the number of tags that do not contain 4 base palindromes in at least one of the phases. However, Meyers *et al.*[209] estimated that despite the two phase sequencing approximately 8% of sequences would still be affected by bias.

MPSS data is provided in two forms, tags of 20 bp and 17 bp. Upon request a dataset containing a 14 bp tag extraction was also provided. These different tag extractions can be used to examine the effect of the sequencing technology on library complexity. A comparison of the alternate tag extractions suggests that sequencing length has an effect on the complexity of the transcriptome, the longer the tag sequence, the fewer the number of unique transcripts that are sequenced (Table 5.3). The 20 bp library was ~25% less complex than the 14 bp library. This is contrary to what one would expect; a 14 bp library generated *from* the 20 bp data is ~16% *less* complex than the 20 base library.

Tag length used in extraction (bp)	Length of tags analysed (bp)	Number of unique tags
20	20	14855
20	17	13523
20	14	12473
17	17	17593
17	14	14800
14	14	19594

Table 5.3: Effect of tag length on MPSS library complexity.

MPSS tags can be extracted from the same initial dataset to produce tags of different lengths; in this case 14, 17 and 20 bp tags were extracted. After the extractions tag lengths can be computationally shortened to see if there is a difference in complexity between the different tag extractions. Decreasing the tag length was, unexpectedly, found to increase the complexity of the library. For example, an extraction of 20 base tags produced 14,855 unique tags, if the tag length was computationally shortened to 17 bases then a library of 13,523 unique tags was observed, however, if the tags were initially extracted at 17 bases then a library of 17,593 unique tags was produced. All extractions were from libraries sampled at 1.7 million tags.

In conventional dideoxy sequencing[220], when there is an error this leads to a false sequence being produced. This will lead to an artificial increase in LongSAGE library complexity. As explained previously, these erroneous tags can either be removed by using the genome sequence as a filter or be identified and converted into their parent tags using a BLAST[168] or nearest neighbour approach[200]. In MPSS sequencing it appears that there are two classes of error. Firstly, there are errors like conventional sequencing where there is a false base-call and secondly errors where it is not possible to sequence the tag at all. The first class of errors can be treated in the same way as LongSAGE errors. For the second class of errors, if they were being produced in an entirely random fashion then by sequencing deeper eventually all sequences could be sequenced. However, as explained above, some sequences appear to be lost in a non-random manner due to the palindrome effect and perhaps other as yet unidentified sources of bias; this may affect tag identification in two ways. Firstly, there will be some tags that it is not possible to sequence at all - no matter what the expression levels of these tags are, they will not appear in the library. Secondly there may be some tags with a reduced chance of appearing in the library, in which case the tags corresponding to highly expressed sequences will appear but at a reduced level, whereas tags from transcripts expressed at a lower level will not appear in the library at all. Sequencing deeper will increase the chance of finding these tags. The predicted extent of the palindrome effect is not enough to explain the large decrease in library complexity observed when going from 14 bp tags to 20 bp tags. This suggests that either the palindrome effect is bigger than previously estimated or that there is another currently unidentified systematic bias in MPSS sequencing.

One potential cause of bias in LongSAGE library creation may be the formation of ditags. *Mme*I cleaves with a 2 base overhang and tags with the overhang are ligated together to create ditags. One might expect that due to stronger base pairing, and therefore increased efficiency of ditag formation, the last two bases of tags would be enriched for G and C. However, this does not appear to be the case. The only clear trend is the preference for AA and its complementary partner TT, with these being the first and second most common 3′ ends for tags in three quarters of the LongSAGE libraries tested (Table 5.4). There are two causes for this deviation from the expected. The first is that whilst *Mme*I cleaves with a 2 base overhang, the cleavage length is not exact; so it often produces tags that are longer than 21 bases. This means that examining the last two bases of a tag is not always the same as examining the bases in the overhang. This explains why AA and TT are not observed at exactly the same frequency. Secondly, the tag site is towards the 3′ end of the sequence, so due to the vicinity of the polyA tail one might expect tags ending in a string of As to be enriched. This is demonstrated by the fact that the tags in two conventional SAGE libraries (where ditags are produced by blunt end ligation) were also enriched with tags ending in 3′ AA but not those ending TT. This suggests that LongSAGE libraries will have an over-representation of tags ending with TT.

3′ End	LongSAGE libraries												SAGE libraries	
	1	2	3	4	5	6	7	8	9	10	CD4 Activated	CD4 Resting	Clone 32	NK
AA	1	1	2	7	1	5	3	2	2	1	1	1	2	6
AC	11	12	13	11	13	11	12	11	12	12	15	15	15	14
AG	8	4	11	10	10	10	10	10	10	10	7	9	6	8
AT	12	15	12	15	12	15	15	14	15	15	13	14	10	13
CA	15	10	14	12	11	13	11	12	14	11	11	13	1	2
CC	10	5	8	8	8	8	6	9	8	8	14	12	3	1
CG	16	16	16	16	16	16	16	16	16	16	16	16	16	16
CT	3	3	3	3	3	7	4	3	7	3	4	3	4	3
GA	6	7	4	6	6	4	7	5	5	6	3	8	9	9
GC	9	14	10	9	9	9	8	13	9	9	12	5	8	5
GG	7	6	7	2	4	2	1	7	4	4	8	4	7	7
GT	13	13	15	13	15	14	13	15	13	14	9	6	12	10
TA	14	11	9	14	14	12	14	8	11	13	6	7	14	15
TC	5	9	5	1	5	3	5	4	3	5	5	11	13	11
TG	4	8	6	5	7	6	9	6	6	7	10	10	5	4
TT	2	2	1	4	2	1	2	1	1	2	2	2	11	12

Table 5.4: Rankings of 3′ end of tags found in SAGE and LongSAGE libraries.

The last two bases of every tag in a library were examined and frequencies of each of the sixteen 2 bp combinations in each were produced. The frequencies in each library were then ranked, with 1 being the most common and 16 being the least common. These rankings corresponded to a large variation in the frequency of the different 2 bp combinations. For example, in the $CD4^+$ activated library, AA was observed at the end of 12.7% of all tags whereas CG was found at end of 1.4% of all tags.

Libraries 1-10 are libraries of >200,000 tags found in NCBI GEO. For libraries 1- 10 their respective accession numbers are: GSM31945, GSM31935, GSM41365, GSM41360, GSM41358, GSM41363, GSM41364, GSM41361, GSM41362, GSM41359.

4. Most abundant tags

Despite the major bias affecting the MPSS library, it may be informative to compare the more highly expressed genes. Many comparisons of different techniques have found that whilst the overall correlation between the two techniques may be low, looking at the higher expressed genes produced a high correlation (e.g. SAGE vs. Microarray[102], MPSS vs. microarray[101]).

The 50 most abundant tags account for ~41% of the MPSS library and ~13% of the LongSAGE library. These figures appear very different but there are two factors which artificially reduce the LongSAGE percentage. The first is that the higher error rate in LongSAGE sequencing reduces the observed count of the most common tags. Secondly, the standard tag extraction process in SAGE removes duplicate ditags to avoid PCR bias[94]. In a large library, such as the one presented here, this will reduce the observed amount of the most common tags[207]. The effect of both these factors will be to artificially reduce the absolute count of the most common tags, but this should not have an impact on whether a tag is detected or not. As described above, there are factors limiting the complexity of the MPSS library and this will artificially increase the MPSS percentage.

To compare the tags found by the different technologies, a tag-to-gene algorithm was developed to automatically match tags to their "best" gene. Briefly, the tags were initially matched to the genome and, depending on the number of hits to the genome, they were also mapped to the Ensembl Transcriptome and the Ensembl ESTGene dataset (for further details see Chapter 2). The criteria used for categorising a tag are

similar to those already published[112, 209] except that the criteria used here use the genome, rather than the transcriptome, as their framework.

In the UTBS dataset the LongSAGE library identified approximately twice as many expressed transcripts as did the MPSS library. If the chance of finding a transcript by both techniques was independent of expression level, then one would expect that approximately half of the transcripts corresponding to the top 50 LongSAGE tags would not be found in the MPSS library. For the top 50 LongSAGE tags it was possible to assign genes to 44 of the tags. The remaining six were either not found in the genome or were found greater than twenty times. (Two of the tags that were not found in the genome appear to be daughters of more abundant tags and so may be produced by single base polymorphisms or sequencing errors.) Of these 44 tags, the transcripts corresponding to 31 were among the top 200 tags in the MPSS library and all but two were found in the MPSS library. Interestingly, one of the genes that corresponded to one of these two LongSAGE tags that matched genes not found in the MPSS library did not contain a *Dpn*II recognition site. Overall, this confirms the expectation that the libraries are more similar for highly expressed genes, i.e. the technique biases having a greater effect on the detection of genes expressed at lower levels.

5. Identification of potential novel transcriptional loci

The deep sampling of both LongSAGE and MPSS means that tags can be matched to many regions where transcripts have not been previously identified or predicted by Ensembl (>10,000 loci). However, many of these loci will not be true regions of transcription, as tags may match more than one region. It is likely that loci identified by both techniques will represent genuine regions of transcription. For an initial

study, a strict criteria was developed to identify regions where transcription was detected by both techniques. Tags were required to match the genome only once, at a position where no known or predicted genes were annotated within 5000 bases upstream in either the sense or antisense direction. This left ~2500 unique LongSAGE tags and ~200 MPSS tags. These two tag lists were combined. If a tag was within either 5000 bases or five recognition site sequences (for the two restriction enzymes) of and on the same strand as a tag from the other technique, both tags were kept. This left 85 tag pairs. An initial examination of the some of the tag pairs showed that although they were in regions where no genes had been identified, there were many EST matches nearby. Therefore an additional filter was applied: if the tags were found in the Ensembl ESTGene dataset or were less than 5000 bases downstream of an ESTGene then the tags were excluded, this left 80 pairs of tags. The pairs of tags are listed in Appendix F.1.

Interestingly, the average tag count for the MPSS tags that match the genome once and have some form of gene annotation is 62.6 but for those that match the genome once and are in this novel gene list, it is 17.2. This difference is statistically significant (p value $<<0.001$) and confirms the theory presented earlier that novel genes are expressed at a lower level than those already discovered. It is also interesting to note that over half these tag pairs (43) are found in regions of the genome that are masked in Ensembl.

This technique is potentially a powerful tool for gene identification. The regions selected have been shown to be both free of known and predicted genes and also poorly identified by EST sampling. Identifying the transcripts corresponding to these

tags should be relatively simple. Using both tags as primers, standard PCR or nested 5′ RACE[165] should reveal the transcripts these tags correspond to.

6. Future Directions

The original aim of this chapter was to completely characterise the transcriptome of a $CD4^+$ T-cell clone, using the sequencing depth of MPSS to ensure there was over-sampling of the transcriptome. However, the work presented here suggests that while MPSS can be used to sample deeper than LongSAGE, systematic biases in the sequencing technology limit the breadth of sampling. This lack of breadth in MPSS negates the benefit provided by its greater depth. It appears that LongSAGE is the better open gene expression technology to use when trying to examine the whole transcriptome. To fully characterise the $CD4^+$ T-cell transcriptome, it will be necessary to sample deeper than the library presented here does. To this end further clones are being sequenced with the aim of increasing the size of the library to 500,000 LongSAGE tags and also to produce a library of 500,000 tags from the same clone but in a resting state. In addition, it should be noted that the depth of a LongSAGE library is reduced by sequencing errors and polymorphisms which lower the number of tags which exactly match the genome. A process allowing for single base mismatches or other errors should be developed to attempt to match these tags to the genome extractions presented here, thereby increasing the depth of the LongSAGE library.

VI. Antisense transcription

A. Introduction

In the previous chapter it was shown that at a library depth of approximately 400,000 LongSAGE tags it is possible to detect the expression of the majority of known transcripts. In Chapter 4 it was shown that tags possibly corresponding to antisense transcripts can be found at high levels in medium sized SAGE libraries (70,000 tags). Therefore, using the large LongSAGE libraries and a systematic approach, one should be able to characterise the antisense transcription occurring in a cell. As described in Chapter 1 there are many classes of antisense transcripts. This chapter will focus on *cis*-NATs (natural antisense transcripts); these are transcripts which are transcribed from the same loci as their corresponding sense transcripts but on the opposite strand. *Cis*-NATs have been detected for many individual genes (e.g. haemoglobin b [ref:[221]], MAGE D2 [ref:[222]] and many others as listed by Shendure and Church[223]). Global expression of *cis*-NATs has been examined using a whole range of different global gene expression technologies on a variety of species (for examples see references: LongSAGE[156], microarrays[123], MPSS[155] and SAGE[197]).

All of the previously published studies examining widespread antisense expression have initially extracted sequences from the databases and found putative sense-antisense pairs. The global gene expression technology has then been used to examine whether these pairs exist. In this chapter I aim to take a different approach from previously published work. It is possible to predict the antisense LongSAGE tags for any given transcript. To do this one just has to reverse transcribe 17 bases upstream of a recognition site in the sense strand. One can then examine the

expression level of these tags and determine whether a transcript is being expressed. There are two key advantages of this approach over the previous studies. The first is that there is no need for antisense transcription to have been previously observed for a particular sense transcript. Therefore, one is less limited by the lack of depth of sampling of EST and cDNA sequencing, as any sequence is only required to have been previously observed in one direction rather than two. Secondly, only two assumptions are made about the transcript structure; it must overlap the sense transcript and contain a polyA tail (necessary for a transcript to be detected by LongSAGE). In the other published studies assumptions about splicing were often made. As relatively little is known about the structure of *cis*-NATs, it is beneficial to make as few assumptions about transcript structure as possible. The main disadvantage of this technique is that it does not take into account genome information and therefore will not be analysing regions where the sense and antisense transcript differ due to different splicing.

Using a set of ~9,000 transcripts I aim to examine three properties of antisense transcription. Firstly, the breadth of antisense transcription will be examined, i.e. how many antisense transcripts are observed across a range of samples. Secondly, the level of antisense transcription will be examined and compared to the level of sense transcription. Finally, the position of antisense transcripts will be characterised, i.e. where does antisense transcription occur relative to the sense transcript? The sensitivity of this method will be inspected by comparing it with previously published global studies that examined antisense transcription in human cells[122, 123, 223]. Potentially novel regions of transcription will be identified; regions

will be selected that have been found to be expressed in both the sense and antisense direction.

B. Results and Discussion

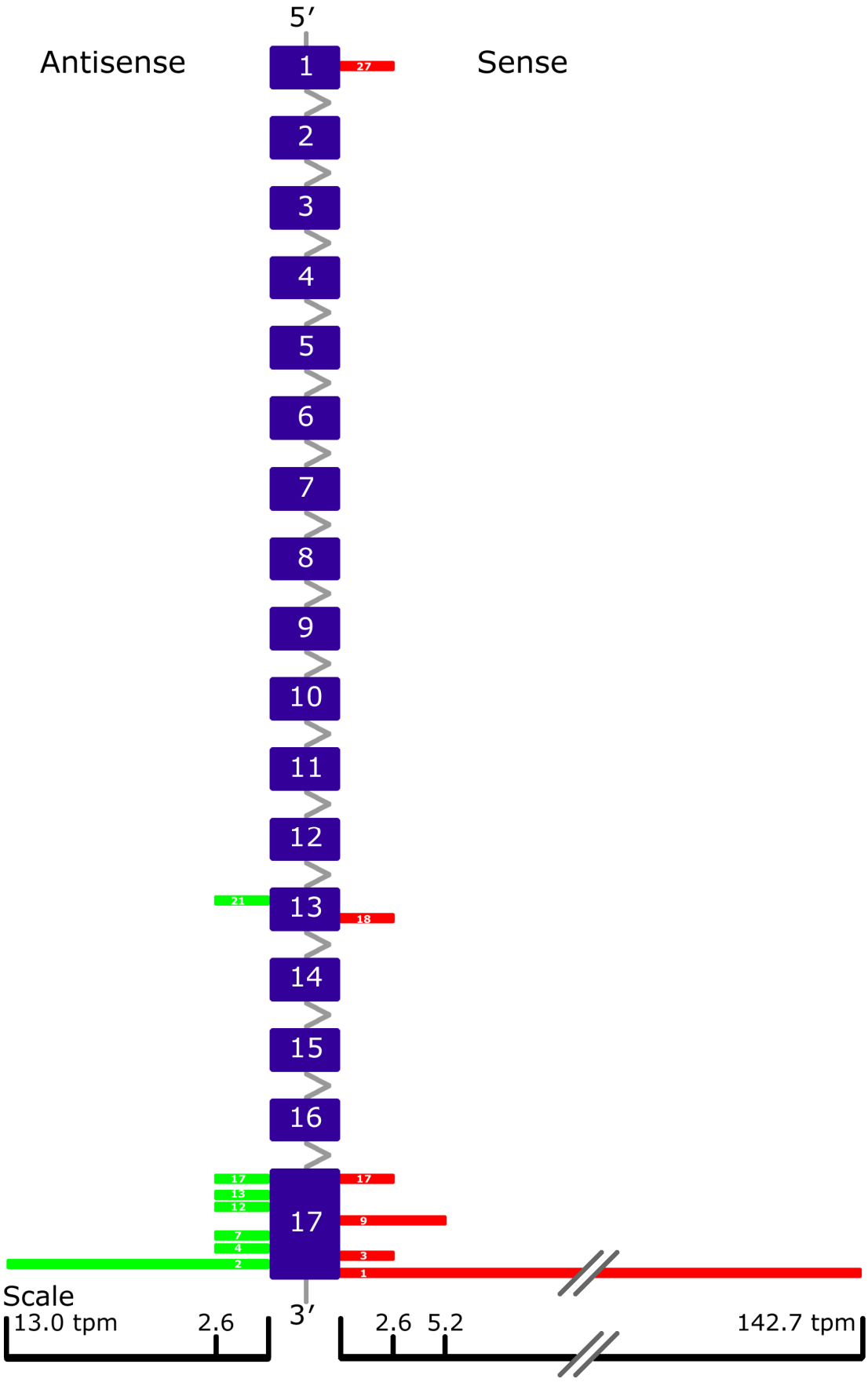
The analysis presented here was carried out on the UTLST dataset (Unique Transcripts LongSAGE Tags). The creation of this dataset is described in detail in Chapter 2. Briefly, all LongSAGE tags were extracted from Ensembl. Those transcripts for which all tags, in both the sense and antisense direction, were unique in the transcriptome and found at most once in the genome were kept. This set consisted of 8954 transcripts. Figure 6.1 shows a pictorial representation of the data collected and analysed for a representative transcript from the UTLST dataset.

1. Characteristics of *cis*-NAT expression

The counts of these transcripts were examined in twelve different libraries; ten previously described LongSAGE libraries obtained from GEO and two $CD4^+$ T-cell clone libraries (resting and activated) produced in our laboratory. In absolute numbers of transcripts detected, antisense transcripts were observed for between one third and half the number of sense transcripts detected (Figure 6.2). Looking at the overlap, that is for transcripts where both sense and antisense transcripts are expressed, ~90% of all antisense transcripts were also expressed in the sense direction in the same tissue/cell line. This is similar to recently published data where LongSAGE was used to identify antisense transcription in mouse embryonic tail; there it was found that for 78% of all antisense transcripts a sense transcript was found[156].

Figure 6.1: Pictorial representation of the sense vs. antisense tags observed for the Lyk gene (a representative transcript from the UTLST dataset).

The Lyk gene (Ensembl transcript ID: ENST00000231189) consists of 17 exons, which contain 27 *Nla*III sites. For all the *Nla*III sites, both the forward and reverse tags are only found once in the genome and transcriptome. The expression of tags in the $CD4^+$ activated library is shown, with tags corresponding to sense transcripts shown as red bars and those corresponding to *cis*-NATs shown as green bars. The numbers on the bars represent the restriction site number, starting at 1 for the 3′ most site. The bar lengths are proportional to the expression level observed for all tags (with the majority of bars representing 2.6 tpm), except for the sense tag at position 1. Exons and introns are not drawn to scale.


5′
Antisense
Sense
1
27
2
3
4
5
6
7
8
9
10
11
12
21
13
18
14
15
16
17
17
13
12
17
9
7
4
3
2
1
3′
Scale
13.0 tpm
2.6
2.6
5.2
142.7 tpm

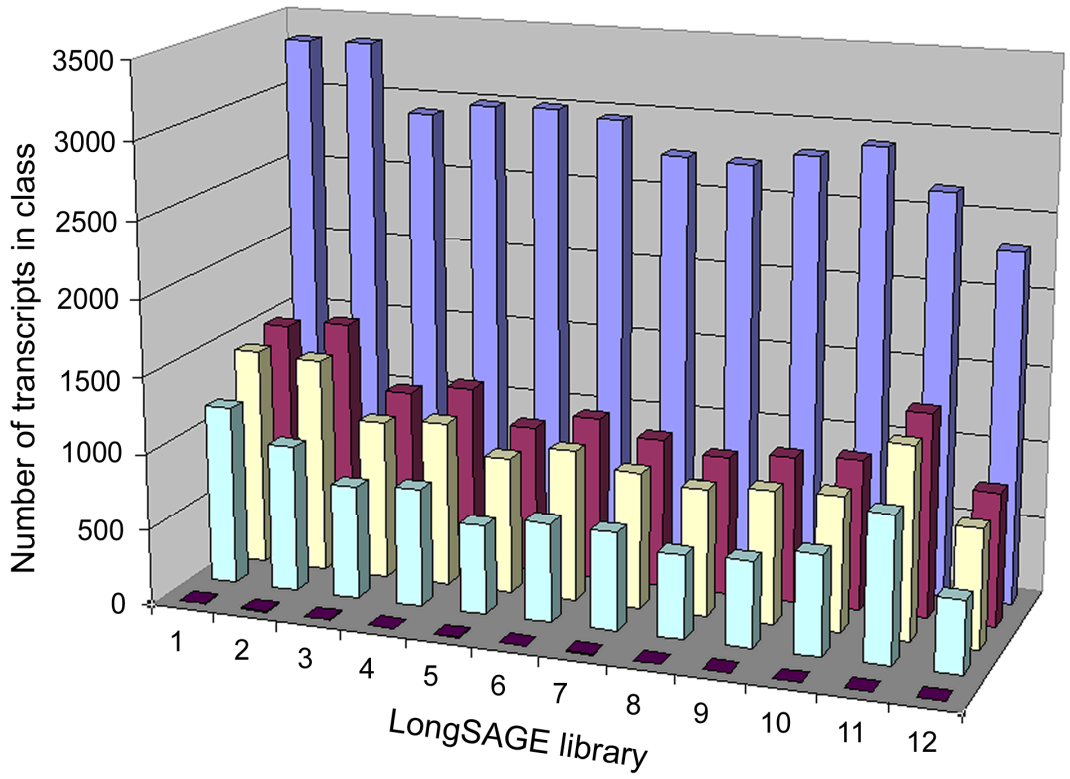


Figure 6.2: Breakdown by class of transcripts observed in the panel of LongSAGE libraries.

The level of expression of transcripts in the UTLST dataset was examined across a panel of 12 LongSAGE libraries. Transcripts were classed depending on which direction the transcripts were found to be expressed (sense, antisense or both) and by the ratio of expression level between the sense and antisense transcripts.

	Transcript class
	Transcripts found in sense direction
	Transcripts found in antisense direction
	Transcripts found in both sense and antisense direction
	Transcripts found three fold over-expressed in sense vs. antisense
	Transcripts found three fold over-expressed in antisense vs. sense

Libraries 1-10 are libraries of >200,000 tags found in NCBI GEO. For libraries 1- 10 their respective accession numbers are: GSM31945, GSM31935, GSM41365, GSM41360, GSM41358, GSM41363, GSM41364, GSM41361, GSM41362, GSM41359. Libraries 11 and 12 are libraries produced in our lab, 11 is the activated $CD4^+$ T-cell clone library and 12 is the resting $CD4^+$ T-cell clone library.

Looking at the ratios of sense to antisense transcript expression levels, where both are detected, for the vast majority of transcripts the sense transcripts are expressed at a higher level. Across the panel of libraries between 65% and 82% of transcript pairs the sense transcripts were expressed at a level over three-fold that of the antisense transcripts, whereas in all libraries, in less than 1% of the pairs were antisense transcripts expressed at three-fold the sense transcript (in absolute numbers this ranges from 2 to 8 transcript pairs per library). The transcript pairs where the antisense transcripts are found at a much higher level may be due to genuine antisense transcription or may be due to transcription of sense transcripts currently unidentified in Ensembl. The mean ratio of sense to antisense transcript expression levels in each library, where both directions were observed, ranged from 7.0 to 10.7. In a given library *cis*-NAT expression represents ~9.1% of transcription identified by LongSAGE (across the panel of libraries this figure ranged from 8.1% to 11.0%).

It can also be informative to look at the absolute expression levels for different transcript classes. All the data below has been normalised to tags per million (tpm), to enable comparisons between libraries. The average antisense transcript was expressed at 10.9 tpm (the different libraries ranged from 8.0 to 14.0 tpm) and the average sense transcript was 42.0 tpm (range 36.2 to 49.0 tpm). Similar figures were observed by Chen *et al.* in their recent paper examining intron length in antisense transcripts (unfortunately exact figures were not given; the data was taken from a graph in the supplementary information)[224]. Looking at subclasses of transcripts also reveals interesting differences. Those loci for which only sense transcripts were found were observed at significantly lower counts than those loci at which both sense and antisense transcripts were observed (means: 21.4 tpm vs. 79.4 tpm). It has

already been established that the ratio of sense to antisense transcripts is between 7 and 10 to 1. Therefore, if an antisense transcript was expressed for those sense transcripts where no antisense was detected, one would expect it to be present at a low level - approximately 3 tpm (calculated by taking the mean of sense transcripts, 21.4 tpm, and dividing this by the ratio of sense to antisense transcripts, between 7.0 and 10.7 to 1). In most of the libraries used here, 3 tpm equates to less than one tag in the library. This suggests that genuine antisense expression may not be being detected due to the libraries not being large enough. This is investigated further in the sensitivity section below.

2. Positional effects

One concern when using any technique is that the phenomena observed may be due to technical artefacts. In LongSAGE, after the first cleavage stage the free 5′ cDNA is washed away. It is possible, but unlikely, that this is not all washed away. If this is the case, then linkers will be added to both ends of the 5′ cDNA and spurious tags that correspond to sense and antisense tags will be produced in equal quantities. Comparing the counts for the theoretical forward and reverse tags at each position for a transcript should address this issue[197]. If there is a linear relationship between the forward and reverse counts then this suggests that the antisense tags are artefactual. From Figure 6.3 it is clear that there is no linear relationship between the forward and reverse tags suggesting that the observed antisense tags are genuine.

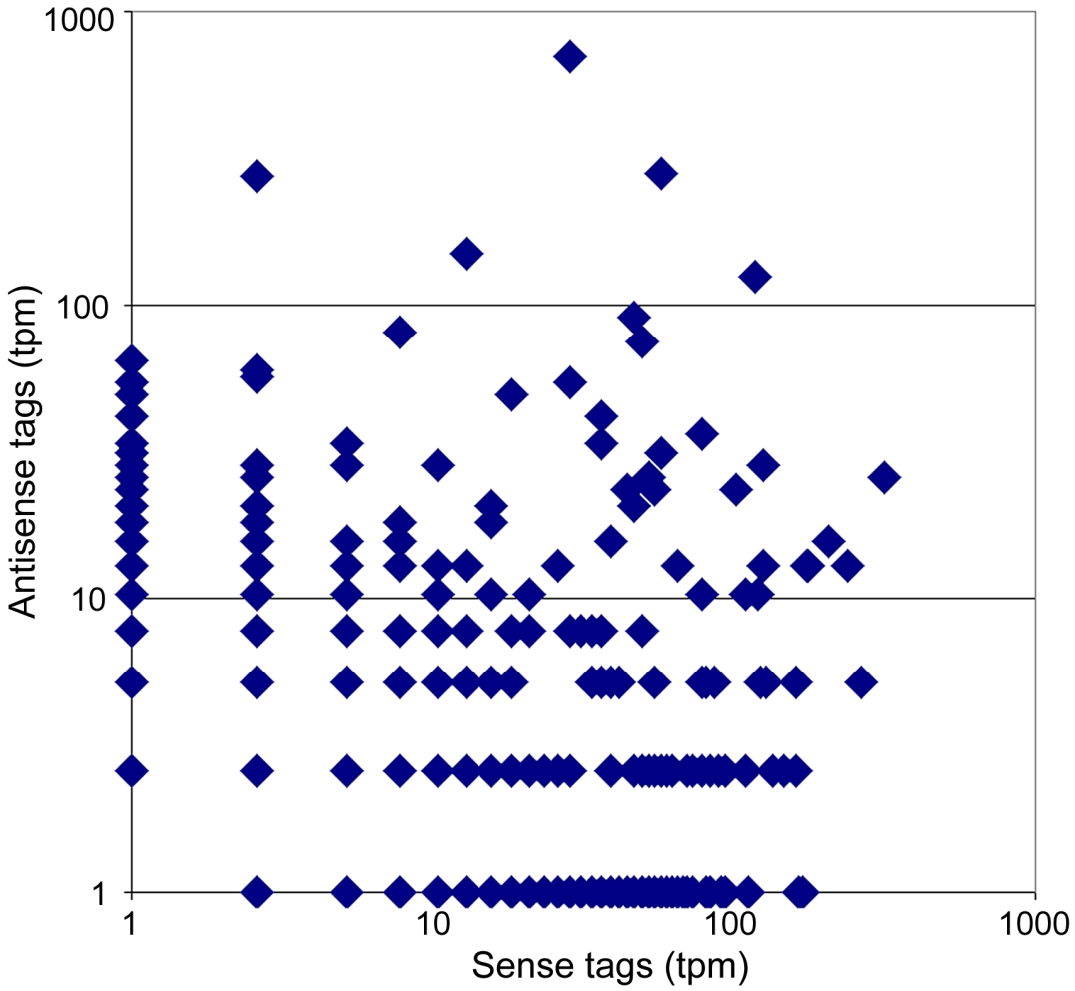


Figure 6.3: Poor correlation between sense and antisense tags from given restriction sites.

The counts of the forward and reverse tags for each position in a transcript were calculated for the activated $CD4^+$ T-cell clone library. The sense and antisense counts are plotted on a logarithmic scale, with zero values being plotted as 1 tag per million. Counts for tag position 1 (the 3′ most restriction site) were excluded to remove bias from transcripts with only 1 restriction site.

To examine whether antisense transcription is a result of random transcription, the level of antisense transcription was compared to that of both sense transcription and transcription in regions where there are no known genes. If the level of antisense transcription was found to be similar to that of transcription outside of known genes, then one could conclude that antisense transcription is the result of random transcription. Briefly, a sliding window approach was used to examine the level of transcription in the positive strand of chromosome 22 by matching tags from the activated $CD4^+$ T-cell library (further details are in Chapter 2). Using the Ensembl genes, tags were classed as matching regions of no known genes, genes in the sense direction, or genes in the antisense direction. Approximately 120,000 windows were analyzed and the average expression level of tags corresponding to sense transcripts was ~6.5 times that of putative antisense tags, which were found at ~3 times that of tags in regions of no known genes. The value for regions with no known genes will be artificially increased because these regions will contain true genes that have not yet been discovered. This suggests that antisense transcription happens at higher level than random transcription from the genome, suggesting that there may be a functional role for antisense transcription. However, the antisense transcription level may just be higher than the basal level of transcription because there is a greater chance of it occurring in regions that are already transcriptionally active (i.e. the expressed genes).

Very little is known about the structure of the "standard" *cis*-NAT. Yelin *et al.* found that the *cis*-NATs primarily overlapped with the sense transcripts at the 5′ and 3′ UTRs. Using LongSAGE it is possible to examine the approximate 3′ end of *cis*-NATs. For this analysis a subset of transcripts from the UTLST tags were used; only those transcripts containing a polyA signal were used, as one can have increased

confidence that these transcripts are full length. The definition for a polyA signal was based on the criteria of Caron *et al*.[153]. For each transcript, all of its tags in one direction were extracted and their counts in a given library totalled. The tag counts were then normalised across the transcript so that each transcript, regardless of its expression level, had the same effect on the positional data. Table 6.1 shows the results of this analysis.

There are two clear trends. The first is that for the sense transcripts the 3′ most LongSAGE tag is the most common tag, being found ten times as often as any other position, regardless of the number of restriction sites in the transcript. This is not the case with the antisense transcripts. With these transcripts the 3′ most (in the sense direction) tag is the most common, but for all but the shortest sequences (those transcripts with two or less restriction sites) it is observed less than half the time. This suggests that whilst there is a preference for *cis*-NATs that terminate at the 3′ end of the sense strand, transcripts are also found that terminate at various positions along the sense strand. It is impossible to establish by LongSAGE whether these transcripts all initiate from the same point.

	Number of transcripts expressed	Average number of tags found per transcript	Number of possible tags	Proportion of tags observed at each site (percentage)									
				Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	Site 7	Site 8	Site 9	Site >9
Sense	68	1.00	1	100	-	-	-	-	-	-	-	-	-
	83	1.36	2	91.9	8.1	-	-	-	-	-	-	-	-
	142	1.42	3	87.9	7.7	4.4	-	-	-	-	-	-	-
	158	1.46	4	88.1	6.1	2.8	3.1	-	-	-	-	-	-
	144	1.67	5	86.3	5.9	3.4	2.3	2.1	-	-	-	-	-
	152	1.71	6	82.6	6.1	3.1	4.1	1.9	2.1	-	-	-	-
	136	1.92	7	80.8	3.9	4.0	4.2	2.5	2.4	2.2	-	-	-
	124	1.72	8	84.3	4.9	3.0	1.8	2.4	1.0	1.2	1.3	-	-
	137	1.86	9	74.7	3.8	6.0	3.0	2.9	4.7	2.9	1.2	0.8	-
	973	2.32	>9	69.6	4.7	3.6	2.8	2.7	3.1	2.2	1.7	1.3	8.4
Antisense	36	1.00	1	100	-	-	-	-	-	-	-	-	-
	43	1.21	2	52.9	47.1	-	-	-	-	-	-	-	-
	71	1.45	3	44.0	31.5	24.5	-	-	-	-	-	-	-
	80	1.50	4	43.6	23.4	17.7	15.3	-	-	-	-	-	-
	82	1.57	5	33.7	15.2	15.4	18.8	16.9	-	-	-	-	-
	65	1.65	6	33.6	18.6	13.9	11.8	7.1	15.0	-	-	-	-
	75	1.61	7	34.1	14.2	18.4	7.7	9.2	7.0	9.3	-	-	-
	62	1.60	8	30.7	23.0	7.7	8.4	6.3	6.9	4.8	12.2	-	-
	60	1.53	9	26.5	9.9	10.8	6.6	15.5	4.3	9.8	7.5	9.2	-
	561	1.83	>9	27.2	17.8	10.0	9.1	5.6	3.4	3.6	2.1	4.2	16.8

Table 6.1: Distribution of tags observed for transcripts with different numbers of restriction sites.

Transcripts in the UTLTS dataset that had a polyA site were grouped by number of *Nla*III recognition sites. The tags found in the activated $CD4^+$ T-cell clone library were then analysed for their position in the transcripts.

The second trend observed is that with an increasing number of restriction sites available there is an increase in the number of different tags observed per transcript. There are two possible causes for this effect. The biological explanation is that as the transcript length increases (as length should be proportional to the number of restrictions sites) there is an increase in the observed alternative splicing. This seems reasonable, as increasing transcript length increases the number of exons, and hence the possibility for alternative splicing. It is also possible that with an increasing number of restriction sites there is an increasing chance of incomplete cleavage, therefore artificially increasing the number of different tags observed.

3. Novel transcriptional loci

In the previous chapter it was demonstrated that by using two different techniques it was possible to identify novel transcriptional loci. The fact that transcription at a given loci was independently identified by two techniques suggested that it was genuine transcription rather than due to technical artefact or tags that were wrongly mapped to the genome. By a similar logic if transcription is found on both the sense and antisense strand at the same location then it is likely to be genuine transcription.

A similar set of criteria was used to that in the previous chapter. Briefly, all tags from a library were mapped to the genome. Those tags that mapped only once to the genome and were in regions where genes had not been previously identified or predicted were kept. If the forward and reverse tag for the same restriction site were in the library then these tags were kept and the region was marked as a potential novel region of transcription.

Using just the tags from the activated $CD4^+$ T-cell clone, 40 regions were identified (see Appendix F.2). Interestingly there was a small overlap with the regions identified in the previous chapter; 8 regions were identified by both methods. This is likely to be due to the limited breadth of sampling of the MPSS library, as highlighted in the previous chapter. Expanding the analysis to include all the tags from all the libraries in the panel increased the number of transcriptional loci identified to 386. Again there was limited overlap with the set identified in the previous chapter, with 18 regions being common to both.

4. Sensitivity

Two approaches were taken to examine the sensitivity of the method used above. The first was to examine how the rate of *cis*-NAT detection changed as an individual LongSAGE library increased in size. It was shown in the previous chapter, using the UTBS dataset, that a library of ~380,000 tags was not large enough to sample all transcripts. Therefore the expectation was that in a single library not all *cis*-NATs will be sampled. The second approach was to compare a panel of LongSAGE libraries with the results of other global searches for antisense transcripts.

Figure 6.4 shows the number of sense and antisense transcripts detected at various library sizes. The rate of transcript discovery appears to have slowed in the sense direction, whereas this is not the case in the antisense direction. Figure 6.5 shows the frequency of the observed counts for the transcripts in the sense and antisense direction. As explained in the previous chapter, if the frequencies for counts 1 and 2 are the same then one would expect that one has sampled all the transcripts. For the sense transcripts the frequencies are similar, whereas for the antisense transcripts, the frequency at count 1 is approximately three times that of count 2.

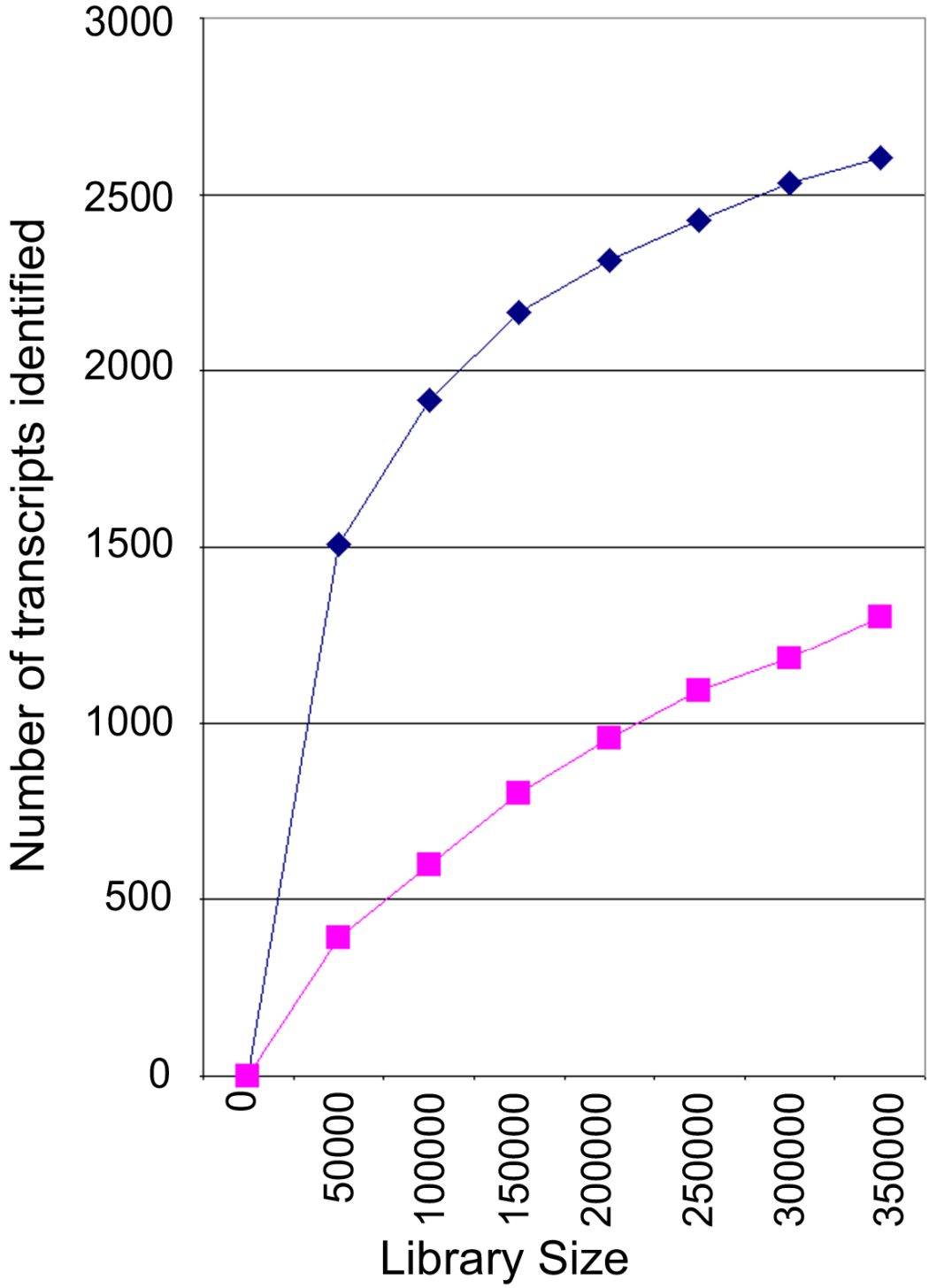


Figure 6.4: Effect of library size on number of transcripts detected.

The UTLST dataset consists of transcripts where all extracted tags are unique in both the transcriptome and found at less than two copies in the genome. The activated $CD4^+$ T-cell clone library was sampled at various sizes and the numbers of transcripts identified were calculated. The blue line represents sense transcripts and pink line represents antisense transcripts.

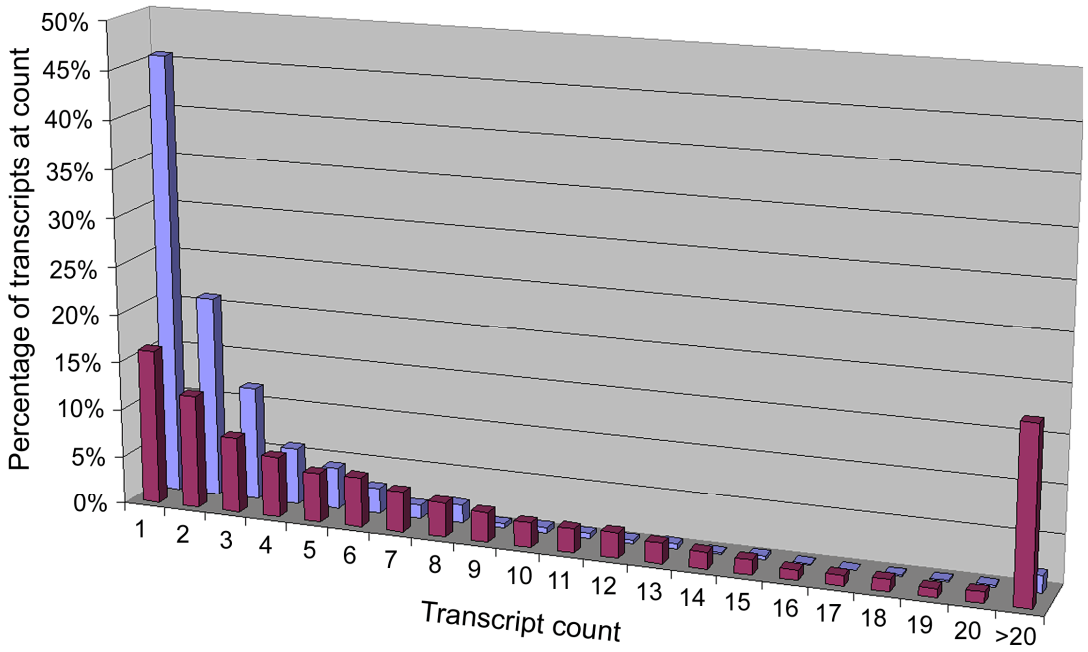


Figure 6.5: Transcript count vs. percentage of transcript at count.

Counts for transcripts were grouped together for the UTLST dataset in the activated $CD4^+$ T-cell library. The frequency of the transcripts counts was then calculated. The blue bars represent the antisense transcripts and the red bars represent the sense transcripts.

Both analyses suggest the majority of sense transcripts have been detected but that this is not the case for the *cis*-NATs. This is not surprising, given that above it was shown that the antisense transcripts are expressed at a much lower level than the sense transcripts. It should be noted that these analyses were carried out on a library of ~380,000 tags prepared from a clonal cell population. The other libraries in the panel ranged in size from 181,483 tags (the resting $CD4^+$ T-cell clone library) to 400,918 tags, with the majority (9/12) of libraries being less 300,000 tags. The fact that most of the libraries in the panel are smaller than the one analysed here and not clonal, suggests that they will also not have sampled all sense or antisense transcripts expressed in the cell.

To analyse the sensitivity of the panel of libraries, the tag counts in all the libraries were combined and then compared to the transcripts identified in three published global analyses that searched for human antisense expression. Shendure and Church[223] assembled EST clusters and identified 144 potential sense-antisense pairs. Using RT-PCR they found 33 out of 39 candidates. Chen *et al.*[122] created clusters of genes by grouping together transcripts of known orientation. They found over 20% of these clusters contained transcripts in the antisense orientation. Out of a test set of 25 transcripts, 24 were found by RT-PCR. Yelin *et al.*[123] also assembled clusters, using the genome as a framework. They identified 2,667 sense-antisense pairs. Using microarrays antisense expression was then tested in a range of tissues, for approximately 10% of their pairs. They estimate that there are "greater than 1600 sense-antisense transcriptional units in the human genome".

Table 6.2 summarises the results of the confirmatory experiments carried out in the original papers and compares the results with what was found in our dataset. It is interesting to note that in all three comparisons our rate of detection for an antisense transcript is ~70%. Two of the papers used RT-PCR to check their predictions, they had a detection rate of >80%, whereas the other paper used microarrays and had a detection rate of ~40%. It appears that the sensitivity of our method lies between the two methods. This is not surprising. Yelin *et al.* examined the expression in 19 different libraries and here only 12 were studied, so one might expect that the microarrays would be more sensitive. However, it has previously been estimated that a microarray experiment is equivalent to a SAGE library of 120,000 tags[225]. All the libraries used here consisted of at least 180,000 tags and so even using a lower number of libraries one would expect a higher rate of detection by our method. Another factor hindering both microarray and LongSAGE confirmation of sequences is the requirement for the mRNA to be polyadenylated; one recent report has suggested that the majority of antisense sequences are polyA negative[127].

Yelin *et al.* predicted that greater than 8% of all genes will show antisense expression, whilst Chen *et al.* predicted that there will be antisense transcription for over 20% of all genes. Here we find that if data from all libraries is pooled, antisense expression is found for over 60% of all genes. These numbers obviously vary widely and it is informative to look at the different criteria used.

	Pairs identified	Original papers			LongSAGE panel		
		Transcripts tested	Sense detected	Antisense detected	Transcripts tested	Sense detected	Antisense detected
Chen[122]	2736	25	23	24	334	289	232
Shendure[223]	144	39	32	33	13	11	9
Yelin[123]	264	264	-	112	38	34	26

Table 6.2: Comparison of LongSAGE antisense detection sensitivity with that of three other published works.

Publicly available transcript pair information was obtained for the three papers listed. In the original papers, sense/antisense transcript pairs were identified computationally. The expression of a subset of these pairs was then tested by bench based methods. The results of these experiments are shown in the "Original papers" columns.

The mRNA pairs identified in the original papers were mapped to Ensembl transcripts, and the pairs were kept if the transcripts could be uniquely mapped to Ensembl transcripts. The transcript pairs that contained transcripts that were found in the UTLST dataset were kept. The expression of these transcripts was then tested across the panel of twelve LongSAGE libraries. A transcript was deemed to be expressed if a corresponding tag was found in any one of the twelve LongSAGE libraries. The results of this analysis are shown in the "LongSAGE panel" columns.

As stated in the introduction, the method proposed here requires minimal previously observed information. The other papers require that antisense transcription has been seen previously by either cDNA or EST sampling. It has been shown previously that SAGE is more sensitive than ESTs for detecting low abundance transcripts[226]. Therefore it is not surprising that a purely LongSAGE based approach detects more antisense transcripts than an approach that relies on EST sampling.

Finally, it seems likely that the antisense transcription observed here is actually a lower limit on antisense transcription because of four factors which will reduce the antisense transcription observed by this method. Firstly, the requirement for a *cis*-NAT to have a polyA tail for LongSAGE to detect it may mean that a proportion of *cis*-NATs can never be identified by LongSAGE. Secondly, the libraries are not sampling deep enough to catalogue the entire transcriptome, so larger LongSAGE libraries should reveal more antisense transcripts. Thirdly, due to the stringent criteria used in transcript selection, the method employed here will exclude well characterised transcript pairs that are known to overlap. Finally, the exact structure of the antisense transcripts is unknown and many may have 3′ ends that differ greatly from the sense sequences; this means their LongSAGE tags will not be profiled by our method. This suggests that *cis*-NATs are probably produced at a low level for all transcripts.

It is hard to predict whether it is only the *cis*-NATs that are expressed at a high level that are functionally important or whether the low level transcripts also have a role. It is tempting to view the antisense system in a similar context to metabolic substrate cycling, where enzymes act at the same time on both directions of a reaction in what

at first glance appears to be an inefficient set-up (e.g. phosphofructokinase and fructose-1,6-bisphosphatase cycle fructose 6-phosphate and fructose-1,6-bisphosphate[227]). However, this enables the cell to achieve a rapid change in flux through the system when necessary. In a transcription context, this would mean that *cis*-NAT production reduces the effective sense transcript level by binding to the sense transcript. Therefore inhibiting *cis*-NAT production could lead to a rapid increase in the number of sense transcripts. Of course *cis*-NATs may have another as yet unidentified function or may merely be a result of random transcription and be functionally irrelevant in the majority of cases.

VII. General Discussion

In this thesis, tag-based gene expression technologies have been used to characterise lymphocyte transcriptomes. The results presented in the previous chapters have implications for both the immunological and gene expression fields.

The initial aim of this thesis was to characterise the T-cell surface. A tag-based gene expression profiling method (SAGE) was used, as it was the only available technology that could meet this aim. Over 100 transcripts encoding cell surface proteins with a known immunological function were found to be expressed by a $CD8^+$ T-cell clone. Surprisingly, only two previously uncharacterised T-cell specific cell surface molecules were identified (Chapter 3). This suggests that the key immunological components of the resting T-cell surface have largely been characterised and that future effort should be focussed on trying to understand how these molecules function, rather than on molecular discovery. This is reinforced by the fact that in the two and a half years since this work was completed, no new cell surface molecules have been identified on a resting T-cell.

Beyond the initial aim, examining the expression of the genes encoding cell surface molecules with known immunological functions revealed two key points. Firstly, there appears to be a correlation between functional importance and the level of gene expression. When one examines the expression levels for transcripts encoding cell surface proteins found on the resting T-cell, one finds that those proteins with critical roles in T-cell activation, such as the TCR, CD8 and CD45, are all encoded by the 20% most highly expressed transcripts for cell surface molecules. One might think that this is because these molecules have been discovered due to their high

expression. However, the initial identification of cell surface molecules generally has more to do with their antigenicity than their level of expression. This suggests that whilst there is a relatively poor correlation between protein and transcript abundance, the expression level of a transcript may be a good indicator of its functional importance in the cell. The second key point is of technical interest. Unsurprisingly, it was shown that by increasing the number of transcripts sampled (library size) the number of transcripts encoding cell surface proteins detected also increased. However, this trend appeared to continue beyond libraries of ~100,000 tags, which is the upper limit of publicly available SAGE libraries. This suggests that if one wants to characterise an entire transcriptome, rather than just the cell surface molecules, it will be necessary to use libraries of a much larger size.

Technological advances have since made it possible to produce much larger libraries. Improvements in dideoxy sequencing reduced the cost of sequencing reactions, meaning that it was feasible to sequence hundreds of thousands of tags for a LongSAGE library. The development of another, proprietary technology meant that millions of tags could be sequenced to produce MPSS libraries. Libraries of this size provide the first opportunity to attempt to profile an entire transcriptome. However, before one attempts to use these libraries to characterise entire transcriptomes it is necessary to gauge how well these libraries reflect the underlying transcriptome. To this end, an MPSS library and a LongSAGE library were made from the same batch of RNA from a $CD4^+$ T-cell clone (Chapter 5) and the two libraries compared. As far as I am aware, this is the first direct comparison of these two tag-based profiling methods. The results of this comparison were somewhat surprising.

Despite being over four times larger than the LongSAGE library, the MPSS library detected fewer transcripts than LongSAGE. The exact difference in the number of transcripts detected depended on the method of calculation, but approximately twice as many transcripts were observed in the LongSAGE library as in the MPSS library. A comparison of multiple MPSS libraries produced from the same RNA suggests that random sampling issues severely limit the efficacy of this technique and that MPSS is not appropriate for characterising entire transcriptomes. On the other hand, LongSAGE detected the expression of a large number of transcriptional loci (~20,000) and, if sequencing errors are filtered out, appears to be well suited to characterising entire transcriptomes. Sampling of the UTLST dataset, i.e., the set of annotated transcripts with unique, unambiguous SAGE tags, suggests that a SAGE library of ~400,000 tags is approaching the size necessary to sample all the transcripts of a given cell. Two caveats need to be kept in mind, however. Firstly, any set of known genes is likely to be biased to some degree toward those that are highly expressed. Secondly, the LongSAGE library presented in this thesis was still detecting novel tags for as long as the library was increasing in size. This suggests that it may be necessary to produce even larger libraries to characterise the entire transcriptome of any given cell. It would be of great help in deciding how deep it is necessary to go in such analyses to know what constitutes a functionally significant level of transcription.

As stated above, it appears that the more highly expressed transcripts encode for proteins of key functional importance. Many of the lower expressed transcripts may be due to random transcription, as it is known that transcription is a stochastic process[214]. It is perhaps surprising that this random transcription may be sufficiently

high to be widely detectable in current experiments. This raises the question of how does one determine whether the expression of a gene is of functional significance or merely the result of random processes. This can be addressed at either the protein or transcriptome level. At the transcriptome level, one approach used in this thesis to identify key genes was to examine patterns of differential expression. It was shown that using differential expression as a filter one could identify the expression of key components of the T-cell surface (Chapter 3). However, differential expression requires the availability of appropriate libraries to compare against and, of course, cannot be used to identify transcripts that are of key importance but encode for ubiquitous "housekeeping" proteins. Recently, large-scale RNAi experiments have been used to examine the importance in endocytosis of genes encoding kinase proteins[228]. Combining this type of experiment with prior knowledge of transcript expression levels may enable one to identify an approximate threshold for functionally important transcription. Further studies are also required into the control of transcription. Whilst it is a stochastic process, there are controlling factors, such as promoter regions, and perhaps once these are understood further it will be possible to predict the level of a particular transcript in a given cell.

Ideally, to understand the expression level of functionally important transcripts, one would compare the proteome to the transcriptome. However, as described in the Introduction, the key problem with proteomics is that large, quantitative experiments are not feasible with current technologies. Along with the fact that large-scale gene expression studies have only been feasible during the past decade, this means that the correlation between gene expression and protein level is poorly understood, with current data giving a very mixed picture. To fully understand whether the expression

of a transcript in a given cell has any functional importance, a technological shift in proteomics, similar in scale to that of the change from Northern blots to SAGE/microarray experiments in transcriptomics, is required to catalogue the translation of the cellular transcripts.

If it is possible to establish a threshold for functional importance, then the level of this threshold will be of key importance for deciding upon future gene expression experiments. If a high level of transcription is necessary for transcripts to be functionally important, then one would be able to use microarrays (which are estimated to profile transcription to a depth of a SAGE library of ~120,000 tags[225]). Conversely, if transcripts observed at a very low level are found to be important then it may be necessary to produce libraries larger than those presented here, and microarrays would have very limited utility.

Examination of NK specific tags indicated that antisense transcripts were potentially identifiable by SAGE (Chapter 4). Having produced a deep sampling library it was then possible to examine the extent of this antisense transcription (Chapter 6). Using a panel of LongSAGE libraries, it was shown that *cis*-NAT transcription was much higher than had previously been detected. This was to be expected and was only possible due the deep sampling of the large LongSAGE libraries, as previous work had shown that the deeper the study, the greater the amount of antisense transcription that was discovered. Across a panel of 12 tissues, *cis*-NATs were detected for almost 70% of sense transcripts, which leads me to speculate that if a larger range of libraries were available, or if the libraries were sampled more deeply, then one may be able to

detect *cis*-NATs for all transcripts. It was also observed that total *cis*-NAT transcription was found at a relatively constant level across different tissues.

This high number of *cis*-NATs leads one to question the functional importance, if any, of this type of antisense transcription. It is unclear from their abundance and distribution whether the *cis*-NATs are of key importance or are merely an unavoidable by-product of transcription. It seems plausible that *cis*-NATs are a by-product of transcription when one considers that DNA transcription is thought to occur in static factories[229, 230]. Once the RNA polymerase has disassociated from the DNA, then there will be competition for its re-binding. It may bind with high affinity to a novel region of DNA that has associated transcription factors bound, but it is also conceivable that it will re-bind with lower affinity to the local antisense strand of the unwound DNA it has just disassociated from and produce a *cis*-NAT. This could be tested by placing a promoter at the 3′ end of a gene and determining whether its activity is enhanced over that at some other location.

To establish whether *cis*-NATs have a cellular function will require further studies, possibly combining both transcriptomic and proteomic data. It may be that the ratio between sense and antisense transcripts plays a role in determining translation efficiency and protein level. It would be interesting to know how many of these *cis*-NATs are found as double stranded transcripts in the nucleus. Also, whilst LongSAGE has identified the regions of antisense transcription, it would be informative to know the structure of a typical *cis*-NAT, i.e. its average length and whether or not it is spliced.

The number of transcripts detected by LongSAGE in a single T-cell clone was unexpectedly large. Around 20,000 different loci were detected; a larger LongSAGE library is likely to have identified additional active loci. When examining the UTBS dataset, just over a third of the transcripts were found in the library, suggesting that a third of the genome might be transcribed in each cell. The latest build of the human genome at Ensembl (31.35d) has 24,194 annotated genes, suggesting, on the other hand, that the great majority of genes are expressed in a given cell. These two contrasting figures can be reconciled to some degree when one considers that approximately one quarter of the ~20,000 transcriptional units identified by SAGE were outside the genes annotated in Ensembl. This reduces the fraction of Ensembl annotated genes that are found to be expressed to just over a half. Overall, therefore, one can estimate that between one third and slightly more than half of the genes in the genome are transcribed in a given cell. It has previously been shown that transcription occurs in "ridges", i.e., regions of the genome appear to be actively transcribed whilst others appear to be switched off[153]. Combining this with the large number of distinct transcripts observed by LongSAGE, it is tempting to suggest that when a region of genome is open for transcription, all the genes in that region will be transcribed, though many will only be transcribed at levels that are functionally unimportant.

A final point to consider is the extent to which the entire human transcriptome is known. Using three different approaches, attempts were made to identify novel transcriptional units; somewhat surprisingly, all three identified very few. Firstly, in the NK library, 256 tags were identified as "immune specific" and only one of these potentially corresponded to a novel gene. Secondly, combining the LongSAGE and

MPSS results identified 80 potentially novel regions of transcription; this compares to the total of ~3,600 annotated genes identified by SAGE and MPSS tag pairs. The tag pairs mapping to novel regions of transcription are found at a lower average expression level than those that match to known genes. Thirdly, combining LongSAGE sense and antisense tags identified 40 potential novel regions of transcription in the T-cell; this compares to approximately 2,200 pairs of forward and reverse LongSAGE tags matching single annotated sites in the genome. However, due to the criteria used, one would not expect the three methods to detect all the currently unknown transcripts expressed in a cell. When one examines *only* the LongSAGE tags, one finds that ~6,000 map to genomic loci where there are no known genes. Again these novel regions are found at a lower average expression level than those loci that match known genes, but at equivalent levels to those that match Ensembl Genscan predictions. Of course, many of these transcripts may be due to either random transcription or technical artefacts. Overall, this suggests that the highly expressed transcripts are well characterised, but it is necessary to go deeper than the levels probed by techniques historically used if one wants to identify all transcripts encoded by the human genome.

The “completion” of the human genome sequence constituted a major milestone in biology. The very large numbers of genomic sites active in individual cells, and the apparent randomness of much of this activity, suggests that a full understanding of the transcriptome will take longer and be more problematic. Separating the "signal" from the "noise" in the transcriptome will depend heavily on a deeper understanding of what constitutes meaningful levels of expression for distinct classes of genes.

VIII. Appendices

Appendix A: Expression of cell surface molecules by CTL Clone 32.

SAGE tags were derived for (1) each of the CD (cluster of differentiation) antigens defined at the 7th HLDA workshop, (2) genes thought likely to be given CD designations at the 8th HLDA workshop, and (3) T cell receptor subunits. Carbohydrate antigens, uncharacterised antigens and molecules linked to tags that match multiple transcripts were not included. The total abundances of all tags derived from each transcript in the clone 32, normal cerebellum and NK cell line are shown alongside the ratio of their total abundance in clone 32 compared to that in normal cerebellum, assigning absent transcripts a total abundance of 1 tag per 100,000. The p values, calculated using the AC test, are given for the differential expression of each transcript in clone 32 compared to normal cerebellum and ovary epithelium. "Number of cancers" indicates the number of cancer derived SAGE libraries (out of 12 tested) in which the total tag abundance for the given transcript is at least one-third of that in clone 32. Values where no calculations were possible due to the absence of transcript detection in the clone 32 library are marked "-".

CD molecule	Alternate Names	LocusLink ID	Unigene Cluster	Representative Accession Number	Clone 32 tag count per 100,000	Cerebellum tag count per 100,000	NK tag count per 100,000	Ratio CTL vs. Cereb.	p-value		Number of Cancers
									Cerebellum	Ovary epithelium	
CD1a	R4; HTA1	909	1309	M28825	0.0	0.0	0.0	1.0	-	-	-
CD1b	R1	910	1310	NM_001764	0.0	0.0	0.0	1.0	-	-	-
CD1c	M241; R7	911	1311	NM_001765	0.0	0.0	0.0	1.0	-	-	-
CD1d	R3	912	1799	NM_001766	0.0	0.0	0.0	1.0	-	-	-
CD1e	R2	913	249217	NM_030893	0.0	0.0	0.0	1.0	-	-	-
CD2	CD2R; E-rosette receptor; T11; LFA-2	914	89476	NM_001767	61.6	0.0	48.5	61.6	0.0%	0.0%	0
CD3γ	CD3g	915	2259	NM_000073	11.1	0.0	0.0	11.1	0.9%	1.1%	0
CD3δ	CD3d	916	95327	NM_000732	101.2	0.0	0.0	101.2	0.0%	0.0%	0
CD3ε	CD3e	917	3003	NM_000733	36.4	0.0	0.0	36.4	0.0%	0.0%	0
CD4	L3T4; W3/25	920	17483	NM_000616	0.0	0.0	0.0	1.0	-	-	-
CD5	Leu-1; Ly-1; T1; Tp67	921	58685	NM_014207	12.6	0.0	0.0	12.6	0.5%	0.6%	0
CD6	T12	923	81226	NM_006725	47.4	0.0	0.0	47.4	0.0%	0.0%	0
CD7	gp40	924	36972	NM_006137	17.4	0.0	9.7	17.4	0.1%	0.1%	0
CD8α	Leu2; Lyt2; T cell co-receptor; T8	925	85258	NM_001768	30.0	2.0	2.8	15.3	0.0%	0.0%	0
CD8β	Leu2; CD8; Lyt3 - beta 1	926	2299	NM_004931	17.4	0.0	0.0	17.4	0.1%	0.1%	0
	Leu2; CD8; Lyt3 - beta 2			X13445	30.0	0.0	0.0	30.0	0.0%	0.0%	0
CD9	DRAP-27; MRP-1; p24	928	1244	NM_001769	0.0	3.9	0.0	0.3	-	-	-
CD10	EC 3.4.24.11; neprilysin; CALLA; enkephalinase; gp100; NEP	4311	1298	NM_007287	0.0	0.0	0.0	1.0	-	-	-
CD11a	AlphaL integrin chain; LFA-1alpha	3683	174103	NM_002209	53.7	0.0	16.6	53.7	0.0%	0.0%	0
CD11b	AlphaM integrin chain; AlphaM-beta2; C3biR; CR3; Mac-1; Mo1	3684	172631	NM_000632	0.0	2.0	2.8	0.5	-	-	-
CD11c	AlphaX integrin chain; Axb2; CR4; leukocyte surface antigen p150,95	3687	51077	NM_000887	0.0	0.0	2.8	1.0	-	-	-

CD molecule	Alternate Names	LocusLink ID	Unigene Cluster	Representative Accession Number	Clone 32 tag count per 100,000	Cerebellum tag count per 100,000	NK tag count per 100,000	Ratio CTL vs. Cereb.	p-value		Number of Cancers
									Cerebellum	Ovary epithelium	
CD13	APN; EC 3.4.11.2; gp150	290	1239	NM_001150	0.0	0.0	0.0	1.0	-	-	-
CD14	LPS-R	929	75627	NM_000591	0.0	0.0	0.0	1.0	-	-	-
CD16a	FCRIIIA	2214	334687	NM_000569	0.0	0.0	0.0	1.0	-	-	-
CD16b	FCRIIIB	2215	176663	NM_000570	0.0	0.0	12.5	1.0	-	-	-
CD19	B4	930	96023	NM_001770	0.0	0.0	0.0	1.0	-	-	-
CD20	B1; Bp35	931	89751	NM_021950	0.0	0.0	0.0	1.0	-	-	-
CD21	C3d receptor; CR2; EBV-R	1380	73792	NM_001877	0.0	0.0	0.0	1.0	-	-	-
CD22	BL-CAM; Lyb8	933	171763	NM_001771	0.0	0.0	0.0	1.0	-	-	-
CD23	B6; BLAST-2; FceRII; Leu-20; Low affinity IgE receptor	2208	1416	NM_002002	0.0	0.0	0.0	1.0	-	-	-
CD25	IL-2R alpha chain; IL-2R; Tac antigen	3559	1724	NM_000417	4.7	0.0	1.4	4.7	9.5%	10.5%	0
CD26	EC 3.4.14.5; ADA-binding protein; DPP IV ectoenzyme	1803	44926	NM_001935	3.2	0.0	6.9	3.2	17.1%	-	5
CD27	S152; T14	939	180841	NM_001242	0.0	0.0	0.0	1.0	-	-	-
CD28	T44; Tp44	940	1987	NM_006139	0.0	0.0	0.0	1.0	-	-	-
CD29	Platelet GPIIa; VLA-beta chain; beta-1 integrin chain	3688	352536	U28252	0.0	0.0	0.0	1.0	-	-	-
CD30	Ber-H2 antigen; Ki-1 antigen	943	1314	NM_001243	0.0	0.0	0.0	1.0	-	-	-
CD31	GPiia'; endocam; PECAM-1	5175	78146	NM_000442	0.0	0.0	8.3	1.0	-	-	-
CD32	FCR II; Fc gamma RII	2212	78864	NM_021642	0.0	0.0	0.0	1.0	-	-	-
CD33	gp67; p67	945	83731	BC005861	0.0	0.0	0.0	1.0	-	-	-
CD34	gp105-120	947	85289	NM_001773	0.0	0.0	0.0	1.0	-	-	-
CD36	GPIIIb; GPIV; OKM5-antigen; PASIV	948	75613	NM_000072	0.0	0.0	0.0	1.0	-	-	-
CD37	gp52-40	951	153053	NM_001774	19.0	0.0	11.1	19.0	0.0%	0.1%	0
CD38	T10; cyclic ADP-ribose hydrolase	952	66052	NM_001775	1.6	0.0	1.4	1.6	30.8%	32.3%	0
CD39		953	205353	NM_001776	30.0	5.9	11.1	5.1	0.1%	0.0%	3
CD40	Bp50	958	25648	NM_001250	0.0	0.0	0.0	1.0	-	-	-

CD molecule	Alternate Names	LocusLink ID	Unigene Cluster	Representative Accession Number	Clone 32 tag count per 100,000	Cerebellum tag count per 100,000	NK tag count per 100,000	Ratio CTL vs. Cereb.	p-value		Number of Cancers
									Cerebellum	Ovary epithelium	
CD41	GPIIb; alpha IIb integrin chain	3674	785	NM_000419	0.0	0.0	4.2	1.0	-	-	-
CD42a	GPIX	2815	1144	NM_000174	0.0	0.0	0.0	1.0	-	-	-
CD42b	GPIbalpha; Glycocalicin	2811	1472	NM_000173	0.0	0.0	0.0	1.0	-	-	-
CD42c	GPIb-beta	2812	283743	NM_000407	0.0	13.8	0.0	0.1	-	-	-
CD43	gpL115; leukocyte sialoglycoprotein; leukosialin; sialophorin	6693	80738	NM_003123	11.1	0.0	26.3	11.1	0.9%	1.1%	0
CD44	ECMR III; H-CAM; HUTCH-1; Hermes; Lu, In-related; Pgp-1; gp85	960	169610	NM_000610	7.9	2.0	1.4	4.0	2.9%	-	8
CD44R	CD44v; CD44v9				1.6	0.0	0.0	1.6	30.8%	32.3%	0
CD45	B220; CD45R; CD45RA; CD45RB; CD45RC; CD45RO; EC 3.1.3.4; LCA; T200; Ly5	5788	170121	NM_002838	34.8	0.0	13.9	34.8	0.0%	0.0%	0
CD46	MCP	4179	83532	NM_002389	11.1	0.0	0.0	11.1	0.9%	48.1%	8
CD47R	Rh-associated protein; gp42; IAP; neurophilin; OA3; MEM-133; formerly CDw149	961	82685	NM_001777	7.9	0.0	2.8	7.9	2.9%	-	3
CD48	BCM1; Blast-1; Hu Lym3; OX-45	962	901	NM_001778	36.4	0.0	12.5	36.4	0.0%	0.0%	0
CD49a	Alpha-1 integrin chain; VLA-1 alpha chain	3672	116774	X68742	3.2	0.0	0.0	3.2	17.1%	18.4%	5
CD49b	Alpha-2 integrin chain; GPIa; VLA-2 alpha chain	3673	271986	NM_002203	0.0	0.0	0.0	1.0	-	-	-
CD49c	Alpha-3 integrin chain; VLA-3 alpha chain	3675	265829	NM_002204 NM_005501	1.6	3.9	1.4	0.4	-	-	0

CD molecule	Alternate Names	LocusLink ID	Unigene Cluster	Representative Accession Number	Clone 32 tag count per 100,000	Cerebellum tag count per 100,000	NK tag count per 100,000	Ratio CTL vs. Cereb.	p-value		Number of Cancers
									Cerebellum	Ovary epithelium	
CD49d	Alpha-4 integrin chain; VLA-4 alpha chain	3676	40034	NM_000885	0.0	0.0	0.0	1.0	-	-	-
CD49f	Alpha-6 integrin chain; Platelet gpI; VLA-6 alpha chain	3655	227730	NM_000210	0.0	0.0	0.0	1.0	-	-	-
CD50	ICAM-3	3385	99995	NM_002162	41.1	2.0	18.0	20.9	0.0%	0.0%	0
CD51	VNR-alpha chain; alpha V integrin chain; vitronectin receptor	3685	295726	NM_002210	1.6	13.8	1.4	0.1	-	-	0
CD52	CAMPATH-1	1043	276770	NM_001803	123.3	0.0	95.6	123.3	0.0%	0.0%	0
CD53		963	82212	NM_000560	28.4	0.0	15.2	28.4	0.0%	0.0%	0
CD54	ICAM-1	3383	168383	NM_000201	0.0	2.0	4.2	0.5	-	-	-
CD55	DAF	1604	1369	NM_000574	0.0	2.0	1.4	0.5	-	-	-
CD56	Leu-19; NKH1; NCAM	4684	167988	NM_000615	1.6	17.7	12.5	0.1	-	-	0
CD58	LFA-3	965	75626	NM_001779	1.6	0.0	2.8	1.6	30.8%	32.3%	0
CD59	1F-5Ag; H19; HRF20; MACIF; MIRL; P-18; Protectin	966	278573	NM_000611	4.7	2.0	4.2	2.4	9.5%	-	11
CD62E	E-selectin; ELAM-1; LECAM-2	6401	89546	NM_000450	0.0	0.0	0.0	1.0	-	-	-
CD62L	L-selectin; LAM-1; LECAM-1; Leu-8; MEL-14; TQ-1	6402	82848	NM_000655	0.0	0.0	0.0	1.0	-	-	-
CD62P	P-selectin; GMP-140; PADGEM	6403	73800	NM_003005	0.0	0.0	0.0	1.0	-	-	-
CD63	LIMP; MLA1; PTLGP40; gp55; granulophysin; LAMP-3; ME491; NGA	967	76294	NM_001780	12.6	5.9	15.2	2.1	14.2%	8.2%	9
CD64	FC gammaRI; FCR I	2209	77424	NM_000566	0.0	2.0	0.0	0.5	-	-	-
CD66b	CD67; CGM6; NCA-95	1088	41	NM_001816	0.0	0.0	0.0	1.0	-	-	-
CD66c	NCA; NCA-50/90	4680	73848	NM_002483	0.0	0.0	0.0	1.0	-	-	-

CD molecule	Alternate Names	LocusLink ID	Unigene Cluster	Representative Accession Number	Clone 32 tag count per 100,000	Cerebellum tag count per 100,000	NK tag count per 100,000	Ratio CTL vs. Cereb.	p-value		Number of Cancers
									Cerebellum	Ovary epithelium	
CD66d	CGM1	1084	11	NM_001815	0.0	0.0	0.0	1.0	-	-	-
CD66e	CEA	1048	220529	NM_004363	0.0	0.0	0.0	1.0	-	-	-
CD66f	Pregnancy specific b1 glycoprotein; SP-1; PSG	5669	336423	NM_006905	1.6	0.0	5.5	1.6	30.8%	32.3%	0
CD68	gp110; macrosialin	968	246381	NM_001251	6.3	3.9	0.0	1.6	16.9%	35.4%	8
CD69	AIM; EA 1; MLR3; gp34/28; VEA	969	82401	NM_001781	7.9	0.0	1.4	7.9	2.9%	3.4%	0
CD70	CD27-ligand; Ki-24 antigen	970	99899	NM_001252	3.2	0.0	0.0	3.2	17.1%	18.4%	2
CD71	T9; transferrin receptor	7037	77356	NM_003234	7.9	2.0	5.5	4.0	2.9%	-	9
CD72	Ly-19.2; Ly-32.2; Lyb-2	971	116481	NM_001782	1.6	0.0	0.0	1.6	30.8%	32.3%	0
CD73	Ecto-5'-nucleotidase	4907	153952	NM_002526	0.0	0.0	0.0	1.0	-	-	-
CD74	Class II-specific chaperone; Ii; Invariant chain	972	84298	NM_004355	189.7	2.0	155.2	96.4	0.0%	0.0%	4
CD79a	Ig alpha; MB1	973	79630	NM_001783 NM_021601	0.0	0.0	0.0	1.0	-	-	-
CD79b	B29; Ig beta	974	89575	NM_000626 NM_021602	0.0	0.0	5.5	1.0	-	-	-
CD80	B7; BB1	941	838	NM_005191	0.0	0.0	0.0	1.0	-	-	-
CD83	HB15	9308	79197	NM_004233	0.0	0.0	0.0	1.0	-	-	-
CD85 (ILT/LIR family)	LILRA1	11024	166156	NM_006863	0.0	0.0	0.0	1.0	-	-	-
	LILRA2	11027	94498	NM_006866	0.0	0.0	0.0	1.0	-	-	-
	LILRA3	11026	113277	NM_006865	0.0	0.0	0.0	1.0	-	-	-
	LILRB2	10288	22405	NM_005874	0.0	0.0	0.0	1.0	-	-	-
	LILRB3	11025	105928	NM_006864	0.0	0.0	0.0	1.0	-	-	-
	LILRB4	11006	67846	NM_006847	0.0	0.0	0.0	1.0	-	-	-
	LILRB5	10990	77062	NM_006840	0.0	2.0	0.0	0.5	-	-	-
	ILT7	23547	48647	NM_012276	0.0	0.0	0.0	1.0	-	-	-
	ILT8	79168	386230	NM_024318	0.0	0.0	0.0	1.0	-	-	-
	ILT9	79167	-	AF072102	0.0	0.0	0.0	1.0	-	-	-
	ILT10	79166	202680	NM_024317	0.0	0.0	0.0	1.0	-	-	-
	ILT11	58534	284190	AF324830	0.0	0.0	0.0	1.0	-	-	-
CD86	B7-2;B70	942	27954	NM_006889	3.2	0.0	0.0	3.2	17.1%	18.4%	2

CD molecule	Alternate Names	LocusLink ID	Unigene Cluster	Representative Accession Number	Clone 32 tag count per 100,000	Cerebellum tag count per 100,000	NK tag count per 100,000	Ratio CTL vs. Cereb.	p-value		Number of Cancers
									Cerebellum	Ovary epithelium	
CD87	Fcalpha-R; IgA Fc receptor;IgA receptor	5329	179657	NM_002659	0.0	3.9	0.0	0.3	-	-	-
CD88	C5aR	728	2161	NM_001736	0.0	0.0	0.0	1.0	-	-	-
CD90	Thy-1	7070	125359	NM_006288	0.0	15.7	0.0	0.1	-	-	-
CD91	ALPHA2M-R; LRP	4035	89137	NM_002332	0.0	3.9	0.0	0.3	-	-	-
CD92	CTL1; formerly CDw92	23446	179902	NM_080546 NM_022109	0.0	0.0	0.0	1.0	-	-	-
CD94	Kp43; killer cell lectin-like receptor subfamily D, member 1	3824	41682	NM_002262 NM_007334	6.3	0.0	1.4	6.3	5.2%	5.9%	0
CD95	APO-1; Fas; TNFRSF6; APT1	355	82359	NM_000043	9.5	0.0	1.4	9.5	1.6%	1.9%	2
CD96	TACTILE	10225	142023	NM_005816	20.5	0.0	23.6	20.5	0.0%	0.0%	0
CD97		976	3107	NM_078481 NM_001784	14.2	0.0	15.2	14.2	0.3%	1.9%	0
CD99	CD99R; E2; MIC2 gene product	4267	177543	NM_002414	19.0	2.0	13.9	9.6	0.0%	-	8
CD100	SEMA4D	10507	79089	NM_006378	3.2	0.0	6.9	3.2	17.1%	18.4%	3
CD101	IGSF2; P126; V7	9398	74117	NM_004258	1.6	0.0	1.4	1.6	30.8%	32.3%	0
CD102	ICAM-2	3384	347326	NM_000873	0.0	0.0	2.8	1.0	-	-	-
CD103	ITGAE; HML-1; integrin alphaE chain	3682	851	NM_002208	6.3	3.9	16.6	1.6	16.9%	35.4%	9
CD104	beta 4 integrin chain; TSP-1180; beta 4	3691	85266	NM_000213	0.0	0.0	0.0	1.0	-	-	-
CD105	endoglin	2022	76753	NM_000118	0.0	5.9	0.0	0.2	-	-	-
CD106	INCAM-110; VCAM-1	7412	109225	NM_001078 NM_080682	0.0	0.0	0.0	1.0	-	-	-
CD107a	LAMP-1	3916	150101	NM_005561	3.2	3.9	4.2	0.8	-	-	12
CD108	SEMA7A; JMH human blood group antigen; formerly CDw108	8482	24640	NM_003612	0.0	0.0	0.0	1.0	-	-	-
CD109	8A3; E123; 7D1	135228	-	NM_133493	0.0	0.0	0.0	1.0	-	-	-
CD110	MPL; TPO-R; C-MPL	4352	84171	NM_005373	0.0	0.0	0.0	1.0	-	-	-

CD molecule	Alternate Names	LocusLink ID	Unigene Cluster	Representative Accession Number	Clone 32 tag count per 100,000	Cerebellum tag count per 100,000	NK tag count per 100,000	Ratio CTL vs. Cereb.	p-value		Number of Cancers
									Cerebellum	Ovary epithelium	
CD111	PVRL1; PRR1; HevC; nectin-1; HIgR	5818	334846	NM_002855	0.0	0.0	0.0	1.0	-	-	-
CD112	HVEB; PRR2; PVRL2; nectin 2	5819	183986	NM_002856	0.0	0.0	0.0	1.0	-	-	-
CD114	CSF3R; HG-CSFR; G-CSFR	1441	2175	NM_000760	0.0	0.0	0.0	1.0	-	-	-
CD115	c-fms; CSF-1R; M-CSFR	1436	174142	NM_005211	0.0	0.0	0.0	1.0	-	-	-
CD117	c-KIT; SCFR	3815	81665	NM_000222	0.0	9.8	0.0	0.1	-	-	-
CDw119	IFNgR; IFNgRa	3459	180866	NM_000416	0.0	2.0	0.0	0.5	-	-	-
CD120a	TNFRI; p55	7132	159	NM_001065	3.2	2.0	2.8	1.6	17.1%	42.2%	8
CD121a	IL-1R; type 1 IL-1R	3554	82112	NM_000877	0.0	0.0	0.0	1.0	-	-	-
CDw121b	IL-1R, type 2	7850	25333	NM_004633	0.0	0.0	0.0	1.0	-	-	-
CD122	IL-2Rbeta	3560	75596	NM_000878	20.5	0.0	76.2	20.5	0.0%	0.0%	0
CD123	IL-3Ralpha	3563	172689	NM_002183	0.0	0.0	0.0	1.0	-	-	-
CD124	IL-4R	3566	75545	NM_000418	0.0	0.0	1.4	1.0	-	-	-
CDw125	IL-5Ralpha	3568	68876	NM_000564	0.0	0.0	0.0	1.0	-	-	-
CD126	IL-6R	3570	193400	NM_000565	0.0	2.0	0.0	0.5	-	-	-
CD127	IL-7R; IL-7R alpha; p90 Il7 R	3575	237868	NM_002185	3.2	0.0	2.8	3.2	17.1%	18.4%	6
CDw128a	CXCR1; IL-8RA	3577	194778	NM_000634	0.0	0.0	0.0	1.0	-	-	-
CDw128b	CXCR2; IL-8RB	3579	846	NM_001557	0.0	0.0	0.0	1.0	-	-	-
CD130	gp130	3572	82065	NM_002184	0.0	2.0	0.0	0.5	-	-	-
CD131	common beta subunit	1439	285401	NM_000395	0.0	0.0	5.5	1.0	-	-	-
CD132	IL2RG; common cytokine receptor gamma chain; common gamma chain	3561	84	NM_000206	15.8	0.0	2.8	15.8	0.2%	0.2%	1
CD133	PROML1; AC133; hematopoietic stem cell antigen; prominin-like 1	8842	112360	NM_006017	0.0	0.0	0.0	1.0	-	-	-
CD134	OX40	7293	129780	NM_003327	0.0	0.0	4.2	1.0	-	-	-
CD135	flt3; Flk-2; STK-1	2322	385	NM_004119	0.0	2.0	0.0	0.5	-	-	-
CDw136	msp receptor; ron; p158-ron	4486	2942	NM_002447	0.0	0.0	0.0	1.0	-	-	-

CD molecule	Alternate Names	LocusLink ID	Unigene Cluster	Representative Accession Number	Clone 32 tag count per 100,000	Cerebellum tag count per 100,000	NK tag count per 100,000	Ratio CTL vs. Cereb.	p-value		Number of Cancers
									Cerebellum	Ovary epithelium	
CD138	heparan sulfate proteoglycan; syndecan-1	6382	82109	NM_002997	0.0	2.0	0.0	0.5	-	-	-
CD140a	PDGF-R; PDGFRa	5156	74615	NM_006206	0.0	0.0	0.0	1.0	-	-	-
CD140b	PDGFRb	5159	76144	NM_002609	0.0	0.0	0.0	1.0	-	-	-
CD141	fetomodulin; TM	7056	2030	NM_000361	0.0	0.0	0.0	1.0	-	-	-
CD142	F3; coagulation Factor III; thromboplastin; TF	2152	62192	NM_001993	0.0	2.0	0.0	0.5	-	-	-
CD143	EC 3.4.15.1; ACE; kininase II; peptidyl dipeptidase A	1636	298469	NM_000789	0.0	0.0	0.0	1.0	-	-	-
CD146	MCAM; A32; MUC18; Mel-CAM; S-endo	4162	211579	NM_006500	1.6	5.9	1.4	0.3	-	32.3%	0
CD147	5A11; Basigin; CE9; HT7; M6; Neurothelin; OX-47; EMMPRIN; gp42	682	74631	NM_001728	11.1	19.7	27.7	0.6	-	-	12
CD148	HPTP-eta; DEP-1; p260	5795	172991	NM_002843	0.0	0.0	0.0	1.0	-	-	-
CD150	SLAM; IPO-3; fomerly CDw150	6504	32970	NM_003037	3.2	0.0	1.4	3.2	17.1%	18.4%	0
CD151	PETA-3; SFA-1	977	75564	NM_004357	0.0	0.0	22.2	1.0	-	-	-
CD152	CTLA-4	1493	247824	NM_005214	0.0	0.0	0.0	1.0	-	-	-
CD153	CD30L	944	1313	NM_001244	0.0	0.0	0.0	1.0	-	-	-
CD154	CD40L; T-BAM; TRAP; gp39	959	652	NM_000074	0.0	0.0	0.0	1.0	-	-	-
CD155	PVR	5817	171844	NM_006505	0.0	0.0	1.4	1.0	-	-	-
CD156a	ADAM8; MS2 human; fomerly CD156	101	86947	NM_001109	0.0	0.0	6.9	1.0	-	-	-
CD157	BP-3/IF-7; BST-1; Mo5	683	169998	NM_004334	0.0	0.0	0.0	1.0	-	-	-
CD158	KIR-023GB	51344	274484	NM_015868	0.0	0.0	41.6	1.0	-	-	-
	KIR2DL1 (NKAT 1)	3802	278453	NM_014218	0.0	0.0	0.0	1.0	-	-	-
	KIR2DL2 (NKAT 6)	3803	278454	NM_014219	0.0	0.0	0.0	1.0	-	-	-
	KIR2DL3 (NKAT 2)	3804	274540	NM_014511	0.0	0.0	43.0	1.0	-	-	-
	KIR2DL4	3805	166085	NM_002255	0.0	0.0	16.6	1.0	-	-	-

CD molecule	Alternate Names	LocusLink ID	Unigene Cluster	Representative Accession Number	Clone 32 tag count per 100,000	Cerebellum tag count per 100,000	NK tag count per 100,000	Ratio CTL vs. Cereb.	p-value		Number of Cancers
									Cerebellum	Ovary epithelium	
	KIR2DL5	57292	283815	NM_020535	0.0	0.0	0.0	1.0	-	-	-
	KIR2DS1 (NKAT10 or Eb6-ActI)	3806	278455	NM_014512	0.0	0.0	0.0	1.0	-	-	-
	KIR2DS2 (NKAT 5)	3807	74134	NM_012312	0.0	0.0	41.6	1.0	-	-	-
	KIR2DS3 (NKAT7)	3808	258572	NM_012313	0.0	0.0	0.0	1.0	-	-	-
	KIR2DS4 (NKAT 8)	3809	258612	NM_012314	0.0	0.0	0.0	1.0	-	-	-
	KIR2DS5 (NKAT 9)	3810	278456	NM_014513	0.0	0.0	0.0	1.0	-	-	-
	KIR3DL1 (NKAT 3)	3811	274601	NM_013289	0.0	0.0	41.6	1.0	-	-	-
	KIR3DL2 (NKAT 4)	3812	56328	NM_006737	0.0	0.0	41.6	1.0	-	-	-
	KIR3DS1	3813	278457	NM_014514	0.0	0.0	0.0	1.0	-	-	-
CD159	NKG2A / NKG2B	3821	74082	NM_002259 NM_007328	0.0	0.0	9.7	1.0	-	-	-
	NKG2C	3822	117605	NM_002260	0.0	0.0	0.0	1.0	-	-	-
	NKG2E / NKG2H	3823	258850	NM_002261	7.9	0.0	1.4	7.9	2.9%	3.4%	0
	NKG2F	8302	26851	NM_013431	1.6	0.0	0.0	1.6	30.8%	32.3%	0
CD160	BY55 antigen; NK1; NK28	11126	81743	NM_007053	1.6	0.0	0.0	1.6	30.8%	32.3%	0
CD161	KLRB1; NKR-P1A; killer cell lectin-like receptor subfamily B, member 1	3820	169824	NM_002258	0.0	0.0	13.9	1.0	-	-	-
CD162/R	PSGL-1 with/without post-translational modification PEN5	6404	79283	NM_003006	11.1	0.0	2.8	11.1	0.9%	1.1%	0
CD163	GHI/61; M130; RM3/1	9332	74046	NM_004244	0.0	0.0	0.0	1.0	-	-	-
CD164	MUC-24; MGC-24v	8763	43910	NM_006016	25.3	0.0	0.0	25.3	0.0%	-	7
CD166	BEN; DM-GRASP; KG-CAM; Neurolin; SC-1; ALCAM	214	10247	NM_001627	0.0	5.9	1.4	0.2	-	-	-
CD167a	trkE; trk6; cak; eddr1; DDR1; MCK10; RTK6; NTRK4	780	75562	NM_001954 NM_013993 NM_013994	0.0	0.0	0.0	1.0	-	-	-
CD168	HMMR; IHABP; RHAMM	3161	72550	NM_012484 NM_012485	1.6	2.0	8.3	0.8	-	32.3%	0
CD169	sialoadhesin; siglec-1	6614	31869	NM_023068	0.0	0.0	0.0	1.0	-	-	-

CD molecule	Alternate Names	LocusLink ID	Unigene Cluster	Representative Accession Number	Clone 32 tag count per 100,000	Cerebellum tag count per 100,000	NK tag count per 100,000	Ratio CTL vs. Cereb.	p-value		Number of Cancers
									Cerebellum	Ovary epithelium	
CD170	Siglec-5	8778	117005	NM_003830	0.0	0.0	0.0	1.0	-	-	-
CD171	L1; L1CAM; N-CAM L1	3897	1757	NM_000425 NM_024003	0.0	9.8	1.4	0.1	-	-	-
CD172a	SIRP alpha	8194	352136	NM_004648	0.0	0.0	0.0	1.0	-	-	-
CD178	fas-L; TNFSF6; APT1LG1; CD95-L	356	2007	NM_000639	3.2	2.0	2.8	1.6	17.1%	18.4%	2
CD179a	VpreB; VPREB1; IGVPB	7441	247979	NM_007128	0.0	0.0	0.0	1.0	-	-	-
CD179b	IGLL1; lambda5; immunoglobulin omega polypeptide; IGVPB; 14.1 chain	3543	348935	NM_020070	0.0	0.0	0.0	1.0	-	-	-
CD180	LY64; RP105	4064	87205	NM_005582	0.0	0.0	0.0	1.0	-	-	-
CD183	CXCR3; GPR9; CKR-L2; IP10-R; Mig-R	2833	198252	NM_001504	4.7	0.0	5.5	4.7	9.5%	10.5%	0
CD184	CXCR4; fusin; LESTR; NPY3R; HM89; FB22	7852	89414	NM_003467	28.4	0.0	5.5	28.4	0.0%	0.9%	1
CD195	CCR5	1234	54443	NM_000579	14.2	0.0	4.2	14.2	0.3%	0.4%	0
CDw197	CCR7	1236	1652	NM_001838	0.0	0.0	0.0	1.0	-	-	-
CD200	OX2	4345	79015	NM_005944	0.0	0.0	0.0	1.0	-	-	-
CD201	EPC R	10544	82353	NM_006404	1.6	0.0	0.0	1.6	30.8%	32.3%	0
CD202b	tie2; tek	7010	89640	NM_000459	0.0	0.0	0.0	1.0	-	-	-
CD203c	NPP3; PDNP3; PD-Ibeta; B10; gp130RB13-6; ENPP3; bovine intestinal phosphodiesterase	5169	264750	NM_005021	0.0	0.0	0.0	1.0	-	-	-
CD204	macrophage scavenger R	4481	49	NM_002445	0.0	0.0	0.0	1.0	-	-	-
CD205	DEC205	4065	153563	NM_002349	0.0	0.0	0.0	1.0	-	-	-
CD206	MRC1; MMR	4360	75182	NM_002438	0.0	0.0	0.0	1.0	-	-	-
CD207	Langerin	50489	199731	NM_015717	0.0	0.0	0.0	1.0	-	-	-
CD208	DC-LAMP	27074	10887	NM_014398	0.0	0.0	0.0	1.0	-	-	-
CD209	DC-SIGN	30385	278694	NM_021155	0.0	0.0	0.0	1.0	-	-	-
CDw210	IL-10 R alpha	3587	327	NM_001558	9.5	0.0	0.0	9.5	1.6%	1.9%	0
CD212	IL-12 R beta 1	3594	121544	NM_005535	0.0	0.0	0.0	1.0	-	-	-

CD molecule	Alternate Names	LocusLink ID	Unigene Cluster	Representative Accession Number	Clone 32 tag count per 100,000	Cerebellum tag count per 100,000	NK tag count per 100,000	Ratio CTL vs. Cereb.	p-value		Number of Cancers
									Cerebellum	Ovary epithelium	
CD213a1	IL-13 R alpha 1	3597	285115	NM_001560	0.0	0.0	0.0	1.0	-	-	-
CD213a2	IL-13 R alpha 2	3598	25952	NM_000640	0.0	0.0	0.0	1.0	-	-	-
CD220	Insulin R	3643	89695	NM_000208	0.0	0.0	0.0	1.0	-	-	-
CD221	IGF1 R	3480	239176	NM_000875	0.0	0.0	0.0	1.0	-	-	-
CD222	Mannose-6-phosphate/IGF2 R	3482	76473	NM_000876	6.3	0.0	0.0	6.3	5.2%	5.9%	5
CD223	LAG-3	3902	74011	NM_002286	6.3	0.0	0.0	6.3	5.2%	5.9%	0
CD224	GGT; EC2.3.2.2	2678	284380	NM_005265 NM_013421 NM_013430	3.2	0.0	2.8	3.2	17.1%	18.4%	4
CD225	Leu13	8519	146360	NM_003641	102.7	0.0	41.6	102.7	0.0%	0.0%	0
CD226	DNAM-1; PTA1; TLiSA1	10666	57699	NM_006566	1.6	0.0	0.0	1.6	30.8%	32.3%	0
CD227	MUC1; episialin; PUM; PEM; EMA; DF3 antigen; H23 antigen	4582	89603	NM_002456	0.0	0.0	4.2	1.0	-	-	-
CD228	melanotransferrin	4241	271966	NM_005929 NM_033316	0.0	0.0	0.0	1.0	-	-	-
CD229	Ly9	4063	153042	AF244129	1.6	0.0	0.0	1.6	30.8%	32.3%	0
CD230	Prion protein	5621	74621	NM_000311	1.6	47.2	1.4	0.0	-	-	0
CD231	TM4SF2; A15; TALLA-1; MXS1; CCG-B7; TALLA	7102	82749	NM_004615	0.0	49.2	0.0	0.0	-	-	-
CD232	VESP R	10154	286229	NM_005761	0.0	0.0	0.0	1.0	-	-	-
CD233	band 3; erythrocyte membrane protein band 3;AE1; SLC4A1; Diego blood group; EPB3	6521	185923	NM_000342	0.0	0.0	0.0	1.0	-	-	-
CD234	Fy-glycoprotein; Duffy antigen	2532	183	NM_002036	0.0	3.9	0.0	0.3	-	-	-
CD235a	Glycophorin A	2993	108694	NM_002099	0.0	0.0	0.0	1.0	-	-	-
CD235b	Glycophorin B	2994	343871	NM_002100	0.0	0.0	0.0	1.0	-	-	-
CD236R	Glycophorin C	2995	81994	NM_002101 NM_016815	34.8	0.0	29.1	34.8	0.0%	0.0%	2
CD238	Kell	3792	157	NM_000420	0.0	0.0	1.4	1.0	-	-	-
CD239	B-CAM	4059	155048	NM_005581	0.0	0.0	0.0	1.0	-	-	-
CD240CE	Rh30CE	6006	278994	NM_020485	0.0	0.0	0.0	1.0	-	-	-

CD molecule	Alternate Names	LocusLink ID	Unigene Cluster	Representative Accession Number	Clone 32 tag count per 100,000	Cerebellum tag count per 100,000	NK tag count per 100,000	Ratio CTL vs. Cereb.	p-value		Number of Cancers
									Cerebellum	Ovary epithelium	
CD240D	Rh30D	6007	283822	NM_016124 NM_016225	0.0	0.0	0.0	1.0	-	-	-
CD241	RhAg	6005	169536	NM_000324	0.0	2.0	0.0	0.5	-	-	-
CD242	ICAM-4	3386	108207	NM_001544 NM_022377	0.0	0.0	0.0	1.0	-	-	-
CD243	MDR-1	5243	21330	NM_000927	0.0	0.0	0.0	1.0	-	-	-
CD244	2B4; NAIL; p38	51744	157872	NM_016382	14.2	9.8	8.3	1.4	17.6%	31.0%	2
CD246	Anaplastic lymphoma kinase	238	278572	NM_004304	0.0	0.0	0.0	1.0	-	-	-
CD247	Zeta chain	919	97087	NM_000734	12.6	0.0	19.4	12.6	0.5%	0.6%	0
Antigens likely to be given a CD designation at the 8th HLDA workshop (from www.hlda8.org/PotentialCDs)											
	A33	10223	143131	NM_005814	0.0	0.0	0.0	1.0	-	-	-
	Apolipoprotein B48 receptor (apoB48R)	55911	200333	NM_018690	3.2	0.0	19.4	3.2	17.1%	18.4%	0
	APRIL (TNFSF13)	8741	54673	NM_003808	0.0	0.0	0.0	1.0	-	-	-
	B cell maturation factor (BCMA)	608	2556	NM_001192	0.0	2.0	0.0	0.5	-	-	-
	B7-H2; LICOS; ICOS-L; TU-D; B7RP-1	23308	14155	AF289028	0.0	0.0	0.0	1.0	-	-	-
	BAFF-R (BR-3)	115650	344088	NM_052945	0.0	0.0	0.0	1.0	-	-	-
	BDCA-2	170482	351812	NM_130441	0.0	0.0	0.0	1.0	-	-	-
	BENE	7851	185055	NM_005434	0.0	0.0	0.0	1.0	-	-	-
	BLAME; BCM-like	56833	20450	NM_020125 NM_014036	0.0	0.0	0.0	1.0	-	-	-
	BlyS	10673	270737	NM_006573	0.0	0.0	1.4	1.0	-	-	-
	CAR	1525	79187	NM_001338	0.0	0.0	0.0	1.0	-	-	-
	CCR1	1230	301921	NM_001295	3.2	0.0	9.7	3.2	17.1%	18.4%	0
	CCR2	1231	395	NM_000647	3.2	0.0	0.0	3.2	17.1%	18.4%	1
	CCR3	1232	158324	NM_001837	0.0	0.0	0.0	1.0	-	-	-
	CCR4	1233	184926	NM_005508	1.6	0.0	0.0	1.6	30.8%	32.3%	0
	CCR7	1236	1652	NM_001838	0.0	0.0	0.0	1.0	-	-	-
	CCR8	1237	113222	NM_005201	0.0	0.0	0.0	1.0	-	-	-
	CCR9	10803	225946	NM_031200	0.0	0.0	0.0	1.0	-	-	-
		51338	325960	NM_024021	0.0	0.0	0.0	1.0	-	-	-
	CLEC-1	51267	29549	NM_016511	0.0	0.0	0.0	1.0	-	-	-
	CMRF-35A	10871	2605	NM_006678	0.0	0.0	2.8	1.0	-	-	-
	CMRF35H	11314	9688	AF020314	0.0	0.0	24.9	1.0	-	-	-

CD molecule	Alternate Names	LocusLink ID	Unigene Cluster	Representative Accession Number	Clone 32 tag count per 100,000	Cerebellum tag count per 100,000	NK tag count per 100,000	Ratio CTL vs. Cereb.	p-value		Number of Cancers
									Cerebellum	Ovary epithelium	
	Connexin 43	2697	74471	NM_000165	0.0	13.8	0.0	0.1	-	-	-
	CRTH2	11251	4253	NM_004778	1.6	0.0	5.5	1.6	30.8%	-	0
	CS1; CRACC; 19A24	57823	132906	NM_021181	68.0	0.0	38.8	68.0	0.0%	0.4%	4
	CTH	23584	122377	NM_014312	0.0	0.0	0.0	1.0	-	-	-
	CX3CR1	1524	78913	NM_001337	0.0	0.0	1.4	1.0	-	-	-
	CXCR5	643	113916	NM_001716	3.2	3.9	0.0	0.8	-	-	11
	CXCR6	10663	34526	NM_006564	0.0	0.0	2.8	1.0	-	-	-
	DCIR	50856	115515	NM_016184	0.0	0.0	0.0	1.0	-	-	-
	DC-SIGNR (L-SIGN)	10332	23759	NM_014257	0.0	0.0	0.0	1.0	-	-	-
	DC-STAMP	81501	211458	NM_030788	0.0	0.0	0.0	1.0	-	-	-
	Dectin-1	64581	161786	NM_022570	0.0	0.0	0.0	1.0	-	-	-
	E-Cadherin	999	356360 (NB also contains a ribosonal protein with no similarity to E-Cadherin)	NM_004360	0.0	0.0	0.0	1.0	-	-	-
	EMR2	30817	137354	NM_013447	0.0	0.0	0.0	1.0	-	-	-
	EMR3	84658	326777	NM_032571	0.0	0.0	0.0	1.0	-	-	-
	Ep-CAM	4072	692	NM_002354	0.0	0.0	0.0	1.0	-	-	-
	ESL-1	2734	78979	NM_012201	9.5	7.9	11.1	1.2	27.6%	7.7%	10
	F4/80 (EMR1)	2015	2375	NM_001974	0.0	0.0	0.0	1.0	-	-	-
	FDF03	29992	122591	NM_013439	0.0	0.0	0.0	1.0	-	-	-
	Flt-1 (VEGFR-1)	2321	381093	NM_002019	0.0	0.0	0.0	1.0	-	-	-
	Galectin-1	3956	227751	NM_002305	145.4	5.9	167.7	24.6	0.0%	-	6
	Galectin-2	3957	113987	NM_006498	0.0	0.0	0.0	1.0	-	-	-
	Galectin-3	3958	621	NM_002306	15.8	3.9	0.0	4.0	0.9%	34.6%	9
	Galectin-4	3960	5302	NM_006149	0.0	0.0	0.0	1.0	-	-	-
	Galectin-7	3963	99923	NM_002307	0.0	0.0	0.0	1.0	-	-	-
	Galectin-8	3964	4082	NM_006499	0.0	0.0	1.4	1.0	-	-	-
	Galectin-9	3965	81337	NM_009587	4.7	0.0	2.8	4.7	9.5%	10.5%	1
	Galectin-10	1178	889	NM_001828	0.0	0.0	0.0	1.0	-	-	-

CD molecule	Alternate Names	LocusLink ID	Unigene Cluster	Representative Accession Number	Clone 32 tag count per 100,000	Cerebellum tag count per 100,000	NK tag count per 100,000	Ratio CTL vs. Cereb.	p-value		Number of Cancers
									Cerebellum	Ovary epithelium	
	Galectin-11 (Antigen identified in sheep. Closest human gene is analysed.)	56891	24236	NM_020129	1.6	0.0	0.0	1.6	30.8%	32.3%	0
	Galectin-12	85329	284183	NM_033101	0.0	0.0	0.0	1.0	-	-	0
	Galectin-13	29124	23671	NM_013268	1.6	0.0	0.0	1.6	30.8%	32.3%	0
	gp200-MR6	4065	153563	NM_002349	0.0	0.0	0.0	1.0	-	-	-
	GPRv53	59340	287388	NM_021624	0.0	2.0	0.0	0.5	-	-	-
	hBRAG	51363	6079	NM_014863	0.0	0.0	0.0	1.0	-	-	-
	HM1.24	684	118110	NM_004335	36.4	2.0	26.3	18.5	0.0%	0.0%	3
	HTm4	932	99960	NM_006138	0.0	0.0	0.0	1.0	-	-	-
	ICOS	29851	56247	NM_012092	0.0	0.0	0.0	1.0	-	-	-
	Il11R alpha (E27)	3590	64310	NM_004512	1.6	0.0	0.0	1.6	30.8%	32.3%	0
	IRTA1	83417	120260	NM_031282	0.0	0.0	0.0	1.0	-	-	-
	IRTA2	83416	191958	NM_031281	0.0	0.0	0.0	1.0	-	-	-
	JAM-2	58494	54650	NM_021219	0.0	0.0	0.0	1.0	-	-	-
	LAIR-1	3903	115808	NM_002287	0.0	0.0	1.4	1.0	-	-	-
	LAR	5792	75216	NM_002840	0.0	2.0	1.4	0.5	-	-	-
	Layilin	143903	133015	BC025407	0.0	0.0	0.0	1.0	-	-	-
	EDG2, endothelial differentiation, lysophosphatidic acid G-protein-coupled receptor, 2	1902	75794	NM_001401	0.0	2.0	0.0	0.5	-	-	-
	EDG4 (endothelial differentiation, lysophosphatidic acid G-protein-coupled receptor, 4)	9170	122575	NM_004720	9.5	0.0	2.8	9.5	1.6%	1.9%	2
	EDG7 (endothelial differentiation, lysophosphatidic acid G-protein-coupled receptor, 7)	23566	258583	NM_012152	0.0	0.0	0.0	1.0	-	-	-
	LYVE-1	10894	17917	NM_006691	0.0	0.0	0.0	1.0	-	-	-
	M160	192150	49636	AF264014	0.0	2.0	0.0	0.5	-	-	-
	MAdCAM	8174	102598	NM_130760	0.0	0.0	0.0	1.0	-	-	-

CD molecule	Alternate Names	LocusLink ID	Unigene Cluster	Representative Accession Number	Clone 32 tag count per 100,000	Cerebellum tag count per 100,000	NK tag count per 100,000	Ratio CTL vs. Cereb.	p-value		Number of Cancers
									Cerebellum	Ovary epithelium	
	MAFA-L	10219	87224	Cluster appears to contain at least three different genes. Tag given is supported by refseq but may not be the only tag. NM_005810	1.6	0.0	6.9	1.6	30.8%	-	0
	MARCO (Macrophage receptor with collagenous structure)	8685	67726	NM_006770	0.0	0.0	0.0	1.0	-	-	-
	MD-2	23643	69328	NM_015364	0.0	0.0	0.0	1.0	-	-	-
	MDC-L (ADAM-23)	10863	174030	NM_014265	1.6	2.0	1.4	0.8	-	32.3%	0
	MDL-1	23601	126355	NM_013252	0.0	2.0	0.0	0.5	-	-	-
	N-Cadherin (cadherin 2, type 1, N-cadherin (neuronal))	1000	161	NM_001792	0.0	0.0	0.0	1.0	-	-	-
	NKp30 (Ly117)	259197	88411	NM_147130	14.2	0.0	43.0	14.2	0.3%	0.4%	0
	NKp44 (Ly95)	9436	194721	NM_004828	0.0	0.0	1.4	1.0	-	-	-
	NKp46 (Ly94)	9437	97084	NM_004829	0.0	0.0	0.0	1.0	-	-	-
	NKp80; killer cell lectin-like receptor F1	51348	183125	NM_016523	0.0	2.0	1.4	0.5	-	-	-
	NTBA; Ly108; KALI	114836	293286	NM_052931	4.7	0.0	2.8	4.7	9.5%	10.5%	1
	OAP-1/Tspan-3	10099	100090	NM_005724	1.6	23.6	2.8	0.1	-	-	0
	PC-1(CD203a - reserved)	5167	11951	NM_006208	0.0	0.0	0.0	1.0	-	-	-
	PD-1	5133	158297	NM_005018	0.0	0.0	0.0	1.0	-	-	-
	PD-1L (B7-H1)	29126	97269	NM_014143	0.0	0.0	0.0	1.0	-	-	-
	Phosphatidylserine receptor	23210	Hs.72660	NM_015167	0.0	0.0	0.0	1.0	-	-	-
	Porimin	114908	393431	NM_052932	1.6	0.0	2.8	1.6	30.8%	32.3%	0
	PZR	9019	287832	NM_003953	0.0	0.0	0.0	1.0	-	-	-
	RAGE	177	184	NM_001136	3.2	0.0	0.0	3.2	17.1%	18.4%	2

CD molecule	Alternate Names	LocusLink ID	Unigene Cluster	Representative Accession Number	Clone 32 tag count per 100,000	Cerebellum tag count per 100,000	NK tag count per 100,000	Ratio CTL vs. Cereb.	p-value		Number of Cancers
									Cerebellum	Ovary epithelium	
	RANK (tumor necrosis factor receptor superfamily, member 11a, activator of NFKB)	8792	114676	NM_003839	0.0	0.0	0.0	1.0	-	-	-
	Siglec-10	89790	335337	NM_033130	0.0	0.0	1.4	1.0	-	-	-
	Siglec-4a (MAG)	4099	1780	NM_002361	0.0	2.0	0.0	0.5	-	-	-
	Siglec-6 (OB-BP1)	946	117992	NM_001245	0.0	2.0	0.0	0.5	-	-	-
	Siglec-7	27036	274470	NM_014385	0.0	0.0	0.0	1.0	-	-	-
	Siglec-9	27180	245828	NM_014441	0.0	0.0	0.0	1.0	-	-	-
	SIRPbeta-1	10326	194784	NM_006065	0.0	0.0	0.0	1.0	-	-	-
	SIRPbeta-2	55423	50716	NM_018556	0.0	0.0	0.0	1.0	-	-	-
	Sphingosine 1-P (S1P) receptor EDG1	1901	154210	NM_001400	0.0	0.0	0.0	1.0	-	-	-
	Sphingosine 1-P (S1P) receptor EDG3	1903	392259	NM_005226	0.0	0.0	0.0	1.0	-	-	-
	Sphingosine 1-P (S1P) receptor EDG5	9294	202672	NM_004230	0.0	0.0	0.0	1.0	-	-	-
	Sphingosine 1-P (S1P) receptor EDG8	53637	302161	NM_030760	0.0	0.0	0.0	1.0	-	-	-
	SRCL (collectin sub-family member 12)	81035	29423	NM_030781	1.6	0.0	0.0	1.6	30.8%	-	0
	SR-PSOX	58191	82407	NM_022059	0.0	0.0	0.0	1.0	-	-	-
	ST2L	9173	66	NM_003856	1.6	9.8	0.0	0.2	-	-	0
	TACI (tumor necrosis factor receptor superfamily, member 13B)	23495	158341	NM_012452	0.0	0.0	0.0	1.0	-	-	-
	TIRC7	10312	46465	NM_006019	53.7	0.0	22.2	53.7	0.0%	0.0%	0
	TLR-2	7097	63668	NM_003264	0.0	0.0	0.0	1.0	-	-	-
	TLR-4	7099	159239	NM_003266	0.0	0.0	0.0	1.0	-	-	-
	TLR-5	7100	114408	NM_003268	0.0	0.0	1.4	1.0	-	-	-

CD molecule	Alternate Names	LocusLink ID	Unigene Cluster	Representative Accession Number	Clone 32 tag count per 100,000	Cerebellum tag count per 100,000	NK tag count per 100,000	Ratio CTL vs. Cereb.	p-value		Number of Cancers
									Cerebellum	Ovary epithelium	
	TLR-6	10333	227105	NM_006068	0.0	0.0	0.0	1.0	-	-	-
	TLR-9	54106	87968	NM_017442	0.0	0.0	0.0	1.0	-	-	-
	Toso	9214	58831	NM_005449	4.7	0.0	1.4	4.7	9.5%	10.5%	1
	TRAIL R1 (tumor necrosis factor receptor superfamily, member 10a)	8797	Placed in wrong cluster	NM_003844	0.0	0.0	0.0	1.0	-	-	-
	TRAIL R2 (tumor necrosis factor receptor superfamily, member 10b)	8795	51233	NM_003842	3.2	0.0	0.0	3.2	17.1%	-	12
	TRAIL R3 (Tumor necrosis factor receptor superfamily, member 10c, decoy without an intracellular domain)	8794	119684	NM_003841	0.0	0.0	0.0	1.0	-	-	-
	TRAIL R4 (tumor necrosis factor receptor superfamily, member 10d, decoy with truncated death domain)	8793	129844	NM_003840	0.0	0.0	0.0	1.0	-	-	-
	TRAMP (DR3,LARD, tumor necrosis factor receptor superfamily, member 12 (translocating chain-association membrane protein))	8718	180338	NM_003790	0.0	19.7	6.9	0.1	-	-	-

CD molecule	Alternate Names	LocusLink ID	Unigene Cluster	Representative Accession Number	Clone 32 tag count per 100,000	Cerebellum tag count per 100,000	NK tag count per 100,000	Ratio CTL vs. Cereb.	p-value		Number of Cancers
									Cerebellum	Ovary epithelium	
	TRANCE (RANKL, Tumor necrosis factor (ligand) superfamily, member 11)	8600	115770	NM_003701	0.0	0.0	1.4	1.0	-	-	-
	Transferrin Receptor 2	7036	63758	NM_003227	3.2	0.0	0.0	3.2	17.1%	18.4%	4
	TREM-1	54210	283022	NM_018643	0.0	0.0	1.4	1.0	-	-	-
	TREM-2 (Triggering receptors expressed on myeloid cells)	54209	44234	NM_018965	0.0	0.0	0.0	1.0	-	-	-
	VAP-1 (Vascular adhesion protein)	8639	198241	NM_003734	0.0	0.0	0.0	1.0	-	-	-
	Yac-1 (ADP-ribosyltransferase 1 (ART1))	417	73139	NM_004314	0.0	0.0	0.0	1.0	-	-	-
T-cell receptor chains											
	T-cell receptor alpha locus	6955	74647	M12423	91.7	0.0	18.0	91.7	0.0%	0.0%	0
	T-cell receptor beta locus (Cb1)	6957	303157	K02885	74.3	0.0	12.5	74.3	0.0%	0.0%	0
	T-cell receptor beta locus (Cb2)			X01411	79.0	0.0	8.3	79.0	0.0%	0.0%	0
	T-cell receptor gamma locus	6965	112259	M27334 Y00790	0.0	2.0	5.5	0.5	-	-	-
	T-cell receptor delta locus	6964	2014	M21624 X73617	0.0	0.0	40.2	1.0	-	-	-

Appendix B: SAGE tags significantly more abundant in clone 32 than in cerebellum.

Structural and functional data, where this is known, for proteins corresponding to all 1098 SAGE tags that are significantly more abundant in clone 32 than cerebellum (p value = 0.05 by the AC test) are listed. The list is ordered by decreasing tag count in clone 32. The p values associated with the differential expression of the tag in clone 32 compared to normal cerebellum and ovary epithelium, calculated using the AC test, are also shown. "Number of cancers" indicates the number of cancer-derived SAGE libraries, out of 12 tested, in which the tag abundance is at least one-third of that in clone 32. SAGE tags were assigned to transcripts, and these transcripts were assessed using their LocusLink, Unigene, and RefSeq database entries and the literature. In the Knowledge column, transcripts are categorised according to their current level of characterisation, and well-characterised transcripts are divided according to whether they have been assigned any immune function in the literature. In the Function column, transcripts are categorised according to their broad molecular function.

Characterisation classes are: K, known immune function; O, other known function; E, tag matches EST sequences only; U, uncharacterised (but sequenced) transcript; M, tag matches multiple transcripts; N, no match to Unigene. Functional classes are: E, soluble effector molecules; CS, cell surface molecules; S, signalling molecules; AP, antigen presentation; T, transcriptional regulation; Cy, cytoskeleton related; CC, cell cycle or viability related; P, protein synthesis related (including mRNA processing); O, other (including housekeeping transcripts such as metabolic enzymes).

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
CACAAACGGTA	978.3	18.4	0.0%	0.0%	3	ribosomal protein S27 (metallopanstimulin 1)	195453	O	P
AGGAGGTATCA	949.9	949.9	0.0%	0.0%	0	granulysin	105806	K	E
ATGAAACCCCA	725.5	16.8	0.0%	0.0%	0	small inducible cytokine A5 (RANTES) / chromosome 1 open reading frame 29	241392 / 75470	xs	xs
CTGACCTGTGT	703.3	357.4	0.0%	0.0%	0	major histocompatibility complex, class I, B	77961	K	AP
CTGGGTTAATA	676.5	11.5	0.0%	65.0%	4	ribosomal protein S19	298262	O	P
AAAAATCGGCT	565.8	565.8	0.0%	0.0%	0	small inducible cytokine A5 (RANTES)	241392	K	E
TTGGTCCTCTG	478.9	4.3	0.0%	0.0%	10	ribosomal protein L41	324406	O	P
CGCCGCCGGCT	458.4	5.8	0.0%	0.0%	8	ribosomal protein L35	182825	O	P
GTTGTGGTTAA	425.2	24.0	0.0%	0.0%	6	beta-2-microglobulin	75415	K	AP
TGTGTTGAGAG	401.5	2.7	0.0%	0.0%	12			N	N
AGGGCTTCCAA	388.8	5.5	0.0%	0.0%	9	ribosomal protein L10	29797	O	P
CCCGTCCGGAA	385.6	4.8	0.0%	0.0%	10	ribosomal protein L13	180842	O	P
GAGGGAGTTTC	374.6	5.4	0.0%	15.3%	7	ribosomal protein L27a	76064	O	P
AAGGAGATGGG	344.6	8.3	0.0%	0.0%	7	ribosomal protein L31	184014	O	P
TACCATCAATA	341.4	2.3	0.0%	0.0%	8	glyceraldehyde-3-phosphate dehydrogenase	169476	O	O
GCCGAGGAAGG	322.4	6.6	0.0%	6.8%	9	ribosomal protein S12	339696	O	P
CGCTGGTTCCA	316.1	8.0	0.0%	0.0%	7	ribosomal protein L11	179943	O	P
TAGGTTGTCTA	312.9	4.8	0.0%	0.0%	7	tumor protein, translationally-controlled 1	279860	K	E
AGGCTACGGAA	311.4	3.2	0.0%	4.9%	12	ribosomal protein L13a	119122	O	P
GCAGCCATCCG	298.7	6.9	0.0%	0.0%	9	ribosomal protein L28	4437	O	P
CCCTGGGTTCT	294.0	4.7	0.0%	0.0%	6	ferritin, light polypeptide	111334	O	O
GCATAATAGGT	279.8	3.0	0.0%	0.0%	9	ribosomal protein L21	350077	O	P
TGGTGTTGAGG	273.4	6.0	0.0%	0.0%	9	ribosomal protein S18	275865	O	P
TAAGGAGCTGA	263.9	13.4	0.0%	0.0%	7	ribosomal protein S26	299465	O	P
TTGGTGAAGGA	260.8	6.3	0.0%	0.0%	8	thymosin, beta 4, X chromosome / Homo sapiens cDNA FLJ31414 fis, clone NT2NE2000260, weakly similar to THYMOSIN BETA-4	75968 / 356629	K	E
AGCTCTCCCTG	259.2	4.2	0.0%	0.1%	10	ribosomal protein L17	82202	O	P
AAGGTGGAGGA	259.2	3.9	0.0%	0.0%	8	ribosomal protein L18a	163593	O	P
GCTTTATTTGT	254.5	5.2	0.0%	0.0%	8	actin, beta	288061	O	Cy
ATGGCTGGTAT	243.4	12.4	0.0%	0.0%	11	ribosomal protein S2	182426	O	P
ATAATTCTTTG	235.5	5.2	0.0%	0.0%	8	ribosomal protein S29	539	O	P
TGCAGCACGAG	230.8	39.1	0.0%	0.0%	1	major histocompatibility complex, class I, F	110309	K	AP
GGGCTGGGGTC	230.8	3.9	0.0%	0.0%	12			N	N
GGGGAAATCGC	227.6	23.1	0.0%	70.1%	6	thymosin, beta 10	76293	O	Cy
CAATAAATGTT	227.6	5.8	0.0%	0.0%	7	ribosomal protein L37	337445	O	P

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
ATTCTCCAGTA	226.0	4.8	0.0%	0.1%	9	ribosomal protein L23	234518	O	P
CCCATCCGAAA	224.4	8.1	0.0%	4.8%	5	ribosomal protein L26	91379	O	P
AGGAAAGCTGC	211.8	5.1	0.0%	0.1%	9	ribosomal protein L36	343443	O	P
TTACCATATCA	210.2	15.3	0.0%	0.0%	6	ribosomal protein L39	300141	O	P
TGCACGTTTTC	210.2	3.1	0.0%	0.0%	12	ribosomal protein L32	169793	O	P
GGATTTGGCCT	208.6	3.3	0.0%	0.0%	11	ribosomal protein, large P2	351937	O	P
CTGTTGGTGAT	203.9	4.0	0.0%	0.0%	9	ribosomal protein S23	3463	O	P
AAGACAGTGGC	200.7	2.0	0.0%	8.4%	12	ribosomal protein L37a	296290	O	P
GAACACATCCA	199.1	9.2	0.0%	0.0%	8	ribosomal protein L19	252723	O	P
TCAGATCTTTG	199.1	3.4	0.0%	0.0%	10	ribosomal protein S4, X-linked	108124	O	P
GTGAAGGCAGT	199.1	2.5	0.0%	0.8%	8	ribosomal protein S3A	77039	O	P
CCGTCCAAGGG	192.8	5.4	0.0%	0.3%	9	ribosomal protein S16	80617	O	P
GTTCACATTAG	189.7	96.4	0.0%	0.0%	4	CD74 antigen (invariant polypeptide of major histocompatibility complex, class II antigen-associated)	84298	K	AP
GCTGCCCAGGC	183.3	183.3	0.0%	0.0%	0			N	N
GGCAAGCCCCA	181.8	3.4	0.0%	0.0%	10	ribosomal protein L10a	334895	O	P
GCGGTGTACAC	178.6	178.6	0.0%	0.0%	0	natural killer cell group 7 sequence	10306	K	CS
AACGCGGCCAA	173.9	5.5	0.0%	0.0%	10	macrophage migration inhibitory factor (glycosylation-inhibiting factor)	73798	K	E
TTCAATAAAAA	173.9	4.4	0.0%	0.0%	8	ribosomal protein, large, P1 / Homo sapiens cDNA: FLJ21550 fis, clone COL06258	177592 / 141269	xs	xs
TCACCCACACC	167.5	7.1	0.0%	0.0%	11			N	N
GTTCGTGCCAA	167.5	4.3	0.0%	0.0%	6	Ribosomal protein L35a	287361	O	P
ATCAAGGGTGT	162.8	3.1	0.0%	0.0%	6	ribosomal protein L9	157850	O	P
TGTGCTAAATG	162.8	2.6	0.0%	0.0%	8	ribosomal protein L34	250895	O	P
CGCCGGAACAC	161.2	3.0	0.0%	4.0%	10	ribosomal protein L4	286	O	P
GTCATAGCTGT	159.6	159.6	0.0%	0.0%	0			N	N
ACTCCAAAAAA	159.6	3.5	0.0%	15.8%	7	ribosomal protein S15	133230	O	P
GTTCCCTGGCC	156.5	3.5	0.0%	0.0%	6	Finkel-Biskis-Reilly murine sarcoma virus (FBR-MuSV) ubiquitously expressed (fox derived); ribosomal protein S30	177415	O	P
AGAACCTTCCA	148.6	75.5	0.0%	5.7%	0	eukaryotic translation elongation factor 1 alpha 1, (match ESTs from this cluster)	181165	E	Est
CTGGCGCGAGG	147.0	147.0	0.0%	0.0%	0	Rho GDP dissociation inhibitor (GDI) beta	83656	O	S
GTGCACTGAGC	147.0	6.2	0.0%	0.0%	7	major histocompatibility complex, class I, C / major histocompatibility complex, class I, A	277477 / 181244	K	AP
ACATCATCGAT	147.0	4.7	0.0%	0.1%	11	ribosomal protein L12	182979	O	P
GCCCCCAATAA	145.4	24.6	0.0%	0.0%	6	lectin, galactoside-binding, soluble, 1 (galectin 1)	227751	O	E

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
GGAGTGGACAT	145.4	9.2	0.0%	0.5%	7	ribosomal protein L18	343354	O	P
AGAAAAAAAA	145.4	4.6	0.0%	0.1%	1	major histocompatibility complex, class I, A / enolase 1, (alpha) / pumilio (Drosophila) homolog 1 / KIAA1588 protein / spindle pole body protein / hypothetical protein dJ1181N3.1 / hypothetical protein P1 p373c6 / NB not checked them all	181244 / 254105 / 153834 / 18587 / 9884 / 11114 / 44720	xs	xs
GCTTTTAAGGA	142.2	9.0	0.0%	55.8%	5	ribosomal protein S20	8102	O	P
CCAGAACAGAC	137.5	4.4	0.0%	0.0%	12	ribosomal protein L30	334807	O	P
GAAAAATGGTT	137.5	2.8	0.0%	0.0%	12	laminin receptor 1 (67kD, ribosomal protein SA)	181357	O	CS
GCGACCGTCAC	135.9	3.6	0.0%	1.2%	9	aldolase A, fructose-bisphosphate	273415	O	O
TTATGGGATCT	134.3	17.1	0.0%	0.0%	9	guanine nucleotide binding protein (G protein), beta polypeptide 2-like 1	5662	K	S
GGACCACTGAA	132.8	1.6	0.4%	5.9%	12	ribosomal protein L3	119598	O	P
CCTCGGAAAAT	128.0	3.6	0.0%	0.0%	7	ribosomal protein L38	2017	O	P
TGGAACCCTGA	126.4	126.4	0.0%	0.0%	0	ESTs, Highly similar to hypothetical protein FLJ13725; KIAA1930 protein [Homo sapiens] [H.sapiens]	355944	E	Est
GCCTGTATGAG	126.4	3.8	0.0%	5.5%	10	ribosomal protein S24	180450	O	P
CAGCTCACTGA	124.9	5.3	0.0%	0.0%	7	ribosomal protein L14	738	O	P
ACTTTTTCAAA	124.9	4.2	0.0%	0.0%	10	ESTs / ESTs / ESTs / ESTs, Weakly similar to 810024C cytochrome oxidase I [H.sapiens] / serine/arginine repetitive matrix 2 / KIAA0377 gene product	136824 / 135424 / 133430 / 252338 / 197114 / 156814	xs	xs
TCCTCTTTCCA	123.3	123.3	0.0%	0.0%	0	natural killer cell transcript 4	943	H	U
GGGGCAACAG	123.3	123.3	0.0%	0.0%	0	CDW52 antigen (CAMPATH-1 antigen)	276770	K	CS
AATAGGTCCAA	123.3	3.3	0.0%	8.9%	11	ribosomal protein S25	113029	O	P
TGTACCCCGCT	121.7	121.7	0.0%	0.0%	0	protein tyrosine phosphatase, receptor type, C-associated protein	155975	K	S
GTGTTAACCAG	121.7	5.6	0.0%	0.0%	10	ribosomal protein L15	74267	O	P
GCCGTGTCCGC	115.4	3.3	0.0%	49.1%	11	ribosomal protein S6	350166	O	P
CCCCAGCCAGT	113.8	6.4	0.0%	0.7%	10	ribosomal protein S3	252259	O	P
TCACAAGCAAA	110.6	1.9	0.1%	0.0%	10	nascent-polypeptide-associated complex alpha polypeptide	32916	O	P
CTCAACATCTC	109.1	4.6	0.0%	0.2%	10	ribosomal protein, large, P0	194676	O	P
TGGCTCCTCCC	105.9	105.9	0.0%	0.0%	0	lymphocyte cytosolic protein 1 (L-plastin)	76506	K	Cy
ACTGTAAAAAA	105.9	53.8	0.0%	0.0%	0	novel C3HC4 type Zinc finger (ring finger) / granzyme A (granzyme 1, cytotoxic T-lymphocyte-associated serine esterase 3)	318584 / 90708	xs	xs

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
GATGCTGCCAA	105.9	7.7	0.0%	0.0%	8	ribosomal protein L22	326249	O	P
ACCATTGGATT	102.7	102.7	0.0%	0.0%	0	interferon induced transmembrane protein 1 (9-27)	146360	H	U
AGACTGGAAGG	101.2	101.2	0.0%	0.0%	0	CD3D antigen, delta polypeptide (TiT3 complex)	95327	K	CS
TCCAAATCGAT	101.2	51.4	0.0%	0.2%	6	vimentin	297753	O	Cy
GTACTGTGGCT	101.2	51.4	0.0%	0.1%	2	chloride intracellular channel 1	74276	O	O
TCAGACGCAGC	101.2	12.8	0.0%	0.0%	8	prothymosin, alpha (gene sequence 28)	250655	O	CC
GACGACACGAG	99.6	3.0	0.0%	0.2%	10	ribosomal protein S28	153177	O	P
AACTAAAAAAA	98.0	2.3	0.0%	17.0%	7	ribosomal protein S27a / glutamyl-prolyl-tRNA synthetase	3297 / 55921	O	P
AAACGCTACTA	93.3	93.3	0.0%	0.0%	0	granzyme B (granzyme 2, cytotoxic T-lymphocyte-associated serine esterase 1)	1051	K	E
GGCTGGGGGCC	93.3	7.9	0.0%	0.0%	10	profilin 1	75721	O	Cy
ACGCTGCGGCT	91.7	91.7	0.0%	0.0%	0	T cell receptor alpha locus	74647	K	CS
GACAAAAAAAA	91.7	15.5	0.0%	14.6%	3	ribosomal protein S15a / likely ortholog of mouse and rat twist related bHLH protein Dermo 1	343665 / 32366	xs	xs
GGCCTTTTTTT	91.7	5.8	0.0%	0.0%	0	H1 histone family, member X	109804	O	O
CTCTTCACCTG	90.1	90.1	0.0%	0.0%	0	ribosomal protein L13a	119122	O	P
GCACCAAAGCC	90.1	90.1	0.0%	0.0%	0	small inducible cytokine A3 (homologous to mouse Mip-1a) / hypothetical protein FLJ12154	73817 / 6839	xs	xs
GGGCATCTCTT	90.1	22.9	0.0%	0.0%	4	major histocompatibility complex, class II, DR alpha	76807	K	AP
AACTAACAAAA	90.1	6.5	0.0%	0.1%	6	ribosomal protein S27a	3297	O	P
CCAGTGGCCCG	88.5	6.4	0.0%	8.1%	8	ribosomal protein S9	180920	O	P
GTGCTGAATGG	88.5	2.8	0.0%	0.0%	11	Myosin, light polypeptide 6, alkali, smooth muscle and non-muscle	77385	O	Cy
ACCGCCGTGGT	86.9	86.9	0.0%	0.0%	3	cytochrome b-245, alpha polypeptide	68877	K	O
GGGCAGGCGTG	86.9	22.1	0.0%	0.0%	2	immediate early protein	737	H	U
GTGTTGCACAA	86.9	1.9	0.4%	14.7%	11	ribosomal protein S13	165590	O	P
GCCCAGCTGGA	85.3	21.7	0.0%	0.0%	1	eukaryotic translation elongation factor 1 delta (guanine nucleotide exchange protein)	223241	O	P
AAGGACCTTTT	83.8	14.2	0.0%	0.0%	2	SH3 domain binding glutamic acid-rich protein like 3	109051	H	U
CTTGTAATCC	82.2	3.2	0.0%	0.0%	6	nucleolar RNA associated protein / ESTs, Weakly similar to ALU2_HUMAN ALU SUBFAMILY SB SEQUENCE CONTAMINATION WARNING ENTRY [H.sapiens] / 351454 / NB matches many other genes, need to sort out if important.	183253 / 351454	xs	xs
GCAGTTCTGAC	80.6	80.6	0.0%	0.0%	0	major histocompatibility complex, class II, DR beta 5 / major histocompatibility complex, class II, DR beta 1	308026	K	AP
GCACAAGAAGA	80.6	8.2	0.0%	0.0%	7	growth arrest-specific 5, matches ESTs in cluster	289721	E	Est
TTCTTGTGGCG	80.6	6.8	0.0%	0.0%	9			N	N
AAGAAGATAGA	80.6	2.4	0.1%	0.0%	10	ribosomal protein L23a	350046	O	P

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
TGAAGGAGCCG	80.6	2.0	0.4%	0.0%	8	ATP synthase, H+ transporting, mitochondrial F0 complex, subunit c (subunit 9), isoform 2	89399	O	O
AATACTTCTC	79.0	79.0	0.0%	0.0%	0	T cell receptor beta locus	303157	K	CS
CCAGGCAGGGG	79.0	40.2	0.0%	0.0%	0	DNAX-activation protein 10	117339	K	CS
GGCAAGAAGAA	79.0	4.0	0.0%	0.0%	11	ribosomal protein L27	111611	O	P
AATCCTGTGGA	79.0	2.0	0.3%	0.0%	12	ribosomal protein L8	178551	O	P
ACCCTTTAACA	77.4	19.7	0.0%	0.0%	1	transketolase (Wernicke-Korsakoff syndrome) NB poor cluster, actually matches different gene in this cluster: major histocompatibility complex, class I, E	89643	K	AP
CTAGCCTCACG	77.4	3.9	0.0%	26.2%	11	actin, gamma 1	14376	O	Cy
TTGGCTTTTCT	75.9	6.4	0.0%	0.0%	2	hypothetical protein DJ328E19.C1.1	218329	H	U
CCTTCGAGATC	75.9	2.8	0.0%	0.0%	10	ribosomal protein S5	76194	O	P
CCTAAGTGACT	74.3	74.3	0.0%	0.0%	0	T cell receptor beta locus	303157	K	CS
TCAAGCCATCA	74.3	74.3	0.0%	0.0%	1			N	N
TACAAGAGGAA	74.3	4.2	0.0%	0.0%	10	ribosomal protein L6	349961	O	P
CAAAAAAAAA	74.3	2.0	0.5%	42.5%	5	Too many matches to list	50842 / 11342 / 271623 / 132071 / 2551	xs	xs
TCCGGCCGCGA	72.7	2.3	0.1%	0.0%	3	hypothetical protein	171774	H	U
GCAGAGAAAAA	71.1	71.1	0.0%	0.0%	0	coronin, actin-binding protein, 1A	109606	O	Cy
ATTGTTTATGG	71.1	2.3	0.2%	0.0%	8	high-mobility group (nonhistone chromosomal) protein 17	181163	O	O
TGGCCCCACCC	71.1	2.1	0.3%	26.7%	9	pyruvate kinase, muscle	198281	O	O
CTGCTATACGA	71.1	1.7	1.9%	4.6%	10	ribosomal protein L5	180946	O	P
CAGCGCTGCCG	69.5	69.5	0.0%	0.0%	0			N	N
GGCCCTAGGCA	69.5	17.7	0.0%	0.0%	2	zinc finger protein 36, C3H type-like 2	78909	O	T
TGTACCTGTAA	68.0	2.3	0.2%	28.2%	11	tubulin, alpha, ubiquitous	334842	O	Cy
GCAAAAAAAAA	68.0	1.9	0.9%	0.0%	8	Human DNA sequence from clone RP3-527B10 on chromosome 6q25.1-25.3 Contains a pseudogene similar to HMG (high mobility group) protein, STSs and GSSs / phosphodiesterase 6D, cGMP-specific, rod, delta / thymosin, beta 10 / suppressor of potassium transport defect 3 / KIAA0284 protein / similar to HYPOTHETICAL 34.0 KDA PROTEIN ZK795.3 IN CHROMOSOME IV / RNA binding motif protein, X chromosome / hypothetical protein FLJ21324 / AE-binding protein 1 / NB not checked.	287762 / 48291 / 76293 / 21263 / 182536 / 91579 / 146381 / 4746 / 118397	xs	xs
GTGCCCGTGCC	66.4	66.4	0.0%	0.0%	1	triosephosphate isomerase 1	83848	O	O
CCAAACGTGTA	66.4	2.4	0.1%	0.0%	12	H3 histone, family 3A	181307	O	O
CACCCAATGGG	64.8	64.8	0.0%	0.0%	0	SEC7 homolog	110121	H	U

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
GCGGTTGTGGC	64.8	64.8	0.0%	0.0%	0	Lysosomal-associated multispanning membrane protein-5	79356	H	U
CGAGCCTGTTA	64.8	64.8	0.0%	0.0%	0	zeta-chain (TCR) associated protein kinase (70 kD)	234569	K	S
ACGATGGCCGA	64.8	64.8	0.0%	0.0%	0	glia maturation factor, gamma	5210	O	S
TTTGGGCCTA	64.8	64.8	0.0%	0.0%	3	cysteine-rich protein 1 (intestinal)	17409	O	T
CCTGTCATCCC	64.8	3.0	0.0%	0.0%	2	RNB6 / hypothetical protein FLJ23322	241471 / 285932	H	U
GGGAAGGGGAA	63.2	63.2	0.0%	0.0%	0	19A24 protein	132906	K	CS
TGAAAACTAC	63.2	32.1	0.0%	0.0%	0	major histocompatibility complex, class II, DP alpha 1	914	K	AP
TTGTAATCGTG	63.2	2.9	0.0%	0.3%	10	ESTs, Highly similar to 2102231B Orn decarboxylase antizyme [Homo sapiens] [H.sapiens]	125078	E	Est
TTACCTCCTTC	63.2	1.8	2.0%	18.8%	10			N	N
TGTTTTCATAA	61.6	61.6	0.0%	0.0%	0	small inducible cytokine A4 (homologous to mouse Mip-1b)(nb ESTs that have this Tag approx 90% similar to refseq sequence)	75703	K	E
GGCTATGCCAA	61.6	31.3	0.0%	0.0%	2	cisplatin resistance-associated overexpressed protein	3688	H	U
CTGGGCGTGTC	61.6	15.7	0.0%	0.0%	6	ESTs, Moderately similar to I60307 beta-galactosidase, alpha peptide	355926	E	Est
GGGCTAGTGGG	61.6	15.7	0.0%	0.0%	0	RAS guanyl releasing protein 1 (calcium and DAG-regulated)	182591	O	S
AGACTAACCTT	60.1	60.1	0.0%	0.0%	0	similar to granzyme B, in fact Granzyme H	348264	K	E
CGCCGACGATG	60.1	15.3	0.0%	0.0%	5	interferon, alpha-inducible protein (clone IFI-6-16)	265827	H	U
AGACAAGCTGG	60.1	15.3	0.0%	0.0%	1	splicing factor, arginine/serine-rich 5	166975	O	P
ATGGTGGGGGA	58.5	29.7	0.0%	0.0%	0	zinc finger protein homologous to Zfp-36 in mouse	343586	K	P
CAGTCTCTCAA	58.5	7.4	0.0%	0.0%	7	ribosomal protein S10	76230	O	P
CACCACGGTGT	58.5	3.0	0.1%	0.0%	1	RNB6	241471	H	U
TGTAGATGCGA	56.9	56.9	0.0%	0.0%	0	CD2 antigen (p50), sheep red blood cell receptor	89476	K	CS
CAGATCTTTGT	56.9	14.5	0.0%	3.7%	9	ubiquitin A-52 residue ribosomal protein fusion product 1	5308	O	P
AGGTCCTAGCC	56.9	9.6	0.0%	0.3%	9	glutathione S-transferase pi	226795	O	O
GTTAACGTCCC	56.9	4.8	0.0%	0.0%	8	ribosomal protein L44	178391	O	P
TAGTTGAAGTC	56.9	4.1	0.0%	0.0%	9	ubiquinol-cytochrome c reductase binding protein	131255	O	O
GTAGCACCTAA	55.3	55.3	0.0%	0.0%	0			N	N
ACTTACCTGCT	55.3	1.7	4.4%	15.0%	9	cytochrome c oxidase subunit VIb	174031	O	O
GTGATGCGCAT	53.7	53.7	0.0%	0.0%	0	T-cell, immune regulator 1	46465	K	CS
TATTTATCCAA	53.7	53.7	0.0%	0.0%	0	integrin, alpha L (antigen CD11A (p180), lymphocyte function-associated antigen 1; alpha polypeptide)	174103	K	CS
TTCCCTTCTTC	53.7	53.7	0.0%	0.0%	2	major histocompatibility complex, class II, DP beta 1	814	K	AP
CCCTTAGCTTT	53.7	27.3	0.0%	4.9%	5	myosin regulatory light chain / myosin, light polypeptide, regulatory, non-sarcomeric (20kD)	180224 / 233936	O	Cy

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
AGAACCTTAAA	53.7	13.7	0.0%	1.0%	0	major histocompatibility complex, class I, A	181244	K	AP
CAGGAGTTCAA	53.7	9.1	0.0%	5.8%	7	actin related protein 2/3 complex, subunit 2 (34 kD)	83583	O	Cy
GTCTGGGGCTT	53.7	9.1	0.0%	0.0%	7	thiopurine S-methyltransferase	75725	O	O
TGAAATAAAAC	53.7	5.5	0.0%	2.2%	10	nucleophosmin (nucleolar phosphoprotein B23, numatrin)	9614	O	O
GACCCTGCCCT	53.7	5.5	0.0%	0.0%	6	FK506-binding protein 8 (38kD)	173464	H	U
ATGAAACCCTG	53.7	2.0	1.6%	12.5%	6	SOCS box containing WD protein SWiP 1	187991	H	U
GCCATAAAATG	52.2	26.5	0.0%	12.9%	2	proteoglycan 1, secretory granule	1908	K	E
ACCATTCTGCT	52.2	13.3	0.0%	0.0%	3	interferon induced transmembrane protein 2 (1-8D)	174195	H	U
AACAGAAGCAA	52.2	5.3	0.0%	0.0%	8	Homo sapiens cDNA FLJ32554 fis, clone SPLEN1000106	372680	H	U
ATCCGGCGCCA	52.2	3.8	0.0%	0.0%	10	transcription elongation factor B (SIII), polypeptide 2 (18kD, elongin B)	172772	O	P
CACTGTGACCT	52.2	2.9	0.1%	0.0%	1	hypothetical protein MGC10500	271599	H	U
TCACAGCTGTG	52.2	1.8	3.3%	0.0%	3	B-cell translocation gene 1, anti-proliferative	77054	O	CC
ACCTCCACACG	50.6	50.6	0.0%	0.0%	0	centaurin, beta 1	108947	O	S
CTGTTGGCATT	50.6	50.6	0.0%	0.0%	5	ribosomal protein L21	184108	O	P
TCTGAAGTCAA	50.6	6.4	0.0%	0.0%	1	inhibitor of DNA binding 2, dominant negative helix-loop-helix protein	180919	O	T
CTCCTCACCTG	50.6	3.2	0.1%	0.1%	11	ribosomal protein L13a / BCL2-antagonist/killer 1	119122 / 93213	xs	xs
TGGACCCCCC	50.6	2.9	0.2%	0.0%	0	Hypothetical protein MGC:5244	374608	H	U
GTGACCTCCTT	50.6	2.0	1.8%	4.0%	11	cytochrome c oxidase subunit VIII	374980	O	O
CCCCAGTTGCT	50.6	1.8	2.8%	0.0%	12	calpain, small subunit 1	74451	O	O
TCTTGTGCATA	49.0	8.3	0.0%	0.7%	6	lactate dehydrogenase A	2795	O	O
CCCCCTCGTGC	49.0	5.0	0.0%	0.0%	0	adrenergic, beta, receptor kinase 1	83636	O	S
GAGATCCGCAA	47.4	47.4	0.0%	6.8%	5	proteasome (prosome, macropain) activator subunit 1 (PA28 alpha)	75348	K	AP
GGCTCAGACCA	47.4	47.4	0.0%	0.0%	0	CD6 antigen	81226	K	CS
GTAGCGCCTCC	47.4	47.4	0.0%	0.0%	0	cystatin F (leukocystatin)	143212	K	AP
GAACCGTCCTG	47.4	24.1	0.0%	0.0%	0	ESTs, Weakly similar to T51376 plant adhesion molecule 1 (PAM1) - Arabidopsis thaliana [A.thaliana]	123164	E	Est
TAGAAAGGCAG	47.4	24.1	0.0%	0.0%	1	zinc finger protein 36, C3H type-like 2	78909	O	T
TGGCCTGCCCA	47.4	6.0	0.0%	0.0%	1	MLL septin-like fusion	181002	H	U
TGTAGGTCATT	47.4	6.0	0.0%	0.0%	0	G protein-coupled receptor 87 / ADP-ribosylation factor-like 7	58561 / 111554	xs	xs

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
TGAAAAAAAA	47.4	4.0	0.0%	22.4%	5	mago-nashi (Drosophila) homolog, proliferation-associated / ankyrin repeat and SOCS box-containing 2 / Homo sapiens uncharacterised gastric protein ZA43P mRNA, partial cds / Homo sapiens, Similar to hypothetical protein FLJ20783, clone MGC:1005 IMAGE:3139876, mRNA, complete cds / interleukin enhancer binding factor 3, 90kD / Homo sapiens unknown mRNA, alternatively spliced / Bardet-Biedl syndrome 2 / enhancer of polycomb 1 / cation-chloride cotransporter-interacting protein / spastic paraplegia 7, paraplegin (pure and complicated autosomal recessive) / hypothetical protein FLJ21963 (NB many many matches, not checked them all)	57904 / 182416 / 203594 / 307033 / 256583 / 117226 / 332633 / 129998 / 119178 / 296847 / 13222	xs	xs
CAGCAGAAGCA	47.4	3.4	0.1%	46.2%	11	small EDRK-rich factor 2 / pinin, desmosome associated protein	323806 / 44499	O	P
TACATAATTAC	47.4	2.4	0.6%	0.0%	4	Matches ESTs in multiple endocrine neoplasia I	240443	E	Est
CAGGAACGGGG	47.4	2.2	1.1%	0.0%	4	mitogen-activated protein kinase kinase 2	72241	O	S
GGGGGACGGCT	47.4	2.0	1.9%	0.0%	4	hypothetical protein LOC58481 / CXYorf1 pseudoautosomal region gene / Homo sapiens mRNA for FLJ00075 protein, partial cds / Homo sapiens, Similar to RIKEN cDNA 1110049F14 gene, clone IMAGE:4156973, mRNA, partial cds	356298 / 349333 / 367663 / 356295	H	U
CAAGAGATGCT	45.8	45.8	0.0%	0.0%	0	Septin 1	99741	O	Cy
TCTGCTAAAGA	45.8	45.8	0.0%	0.0%	6	high-mobility group (nonhistone chromosomal) protein 1	337757	O	O
TGGCAACCTTT	45.8	23.3	0.0%	0.8%	2	interleukin enhancer binding factor 2, 45kD / glutathione S-transferase subunit 13 homolog	75117 / 279952	xs	xs
TAGAAAAATAA	45.8	3.9	0.0%	0.4%	1	glucose phosphate isomerase	279789	O	O
TCCTATAAGC	44.3	44.3	0.0%	0.0%	1			N	N
CCTCCAGCAGC	44.3	44.3	0.0%	0.0%	1	retinoic acid receptor responder (tazarotene induced) 3	17466	O	T
CAGATTAGTTA	44.3	22.5	0.0%	0.0%	0	DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide 17 (72kD)	343411	O	P
TGTTAATGTTA	44.3	11.2	0.0%	0.0%	1	G protein-coupled receptor kinase 7 / hypothetical protein FLJ22833 / Homo sapiens mRNA; cDNA DKFZp434N0211 (from clone DKFZp434N0211)	372455 / 118183 / 261828	xs	xs
GACTCTGGTGC	44.3	7.5	0.0%	0.0%	7	ribosomal protein S15a	343665	O	P
GTGCTGGACCT	42.7	42.7	0.0%	39.3%	7	proteasome (prosome, macropain) activator subunit 2 (PA28 beta)	179774	K	AP
ACTTTAGCCTC	42.7	42.7	0.0%	0.0%	0	mitogen-activated protein kinase kinase kinase kinase 1	86575	O	S
CACCCCTGGGG	42.7	42.7	0.0%	0.0%	0	acyloxyacyl hydrolase (neutrophil)	82542	O	E
TACATAGCTTG	41.1	41.1	0.0%	0.0%	0	perforin 1 (pore forming protein)	2200	K	E
CTCCTGGGCAA	41.1	41.1	0.0%	0.0%	0			N	N
GAGGCCTGGCC	41.1	41.1	0.0%	0.0%	0			N	N

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
CGACGAGGAGG	41.1	41.1	0.0%	0.5%	5	epithelial membrane protein 3	9999	O	CS
TTAGGGAGGAG	41.1	20.9	0.0%	0.0%	0	intercellular adhesion molecule 3	99995	K	CS
TTTGTTAAAA	41.1	10.4	0.0%	0.0%	4	HIF-1 responsive RTP801	111244	H	U
CTGAAGCCAAA	41.1	10.4	0.0%	0.0%	0	interleukin 16 (lymphocyte chemoattractant factor)	82127	K	E
TCTGTACACCT	41.1	10.4	0.0%	1.9%	3	ribosomal protein S11	182740	O	P
CAATAAACTG	41.1	2.6	0.7%	4.4%	11	putative translation initiation factor	150580	O	P
TCCGCGAGAAG	41.1	2.1	2.2%	11.4%	9	ESTs in zinc finger protein homologous to Zfp-36 in mouse	343586	E	Est
TCCGCAAGGTG	39.5	39.5	0.0%	0.0%	0	Homo sapiens mRNA; cDNA DKFZp667F1219 (from clone DKFZp667F1219)	37617	H	U
CCTGCAACCT	39.5	39.5	0.0%	0.0%	0	integrin, beta 7	1741	K	CS
GGGCCAAAGTC	39.5	39.5	0.0%	0.0%	0			N	N
TTTGCCCTAGT	39.5	20.1	0.0%	0.0%	0	eomesodermin (Xenopus laevis) homolog	301704	O	T
GCTTTTTAGAA	39.5	5.0	0.0%	2.4%	9	high-mobility group (nonhistone chromosomal) protein 14	251064	O	O
GAGTGGGGGCT	39.5	5.0	0.0%	0.0%	6	dipeptidylpeptidase 7	14089	K	S
TGTGATCAGAC	39.5	4.0	0.1%	0.0%	9	ATP synthase, H+ transporting, mitochondrial F0 complex, subunit g	107476	O	O
GAAACAAGATG	39.5	4.0	0.1%	0.0%	10	phosphoglycerate kinase 1	78771	O	O
ATCCCTCAGTG	39.5	2.9	0.5%	0.0%	9	activating transcription factor 4 (tax-responsive enhancer element B67)	181243	O	T
CGTGTTAATGG	39.5	2.5	0.9%	0.0%	9	zinc finger protein 9 (a cellular retroviral nucleic acid binding protein)	2110	O	T
AAGGTCGAGCT	39.5	2.5	0.9%	0.1%	11	ribosomal protein L24	184582	O	P
ACTTCATAGCA	37.9	37.9	0.0%	0.0%	0	Homo sapiens, similar to hypothetical protein FLJ11110, clone MGC:27027 IMAGE:4837773, mRNA, complete cds	124675	H	U
CTGAGACGAAG	37.9	37.9	0.0%	0.0%	3	basic transcription factor 3	101025	O	T
TCTGCAAAGGA	37.9	19.3	0.0%	0.0%	4	high-mobility group (nonhistone chromosomal) protein 2	80684	O	O
GCCGCCCTGCA	37.9	9.6	0.0%	0.0%	5	acyl-Coenzyme A dehydrogenase, very long chain	82208	O	O
TCTGTTTATCA	37.9	2.4	1.3%	0.0%	3	signal recognition particle 14kD (homologous Alu RNA-binding protein)	180394	O	O
GGCTGGTCTGG	36.4	36.4	0.0%	15.4%	4	Homo sapiens, clone MGC:4677 IMAGE:3532809, mRNA, complete cds	337986	H	U
TAAGTTGTCCC	36.4	36.4	0.0%	0.0%	0	CD3E antigen, epsilon polypeptide (TiT3 complex)	3003	K	CS
TGTATGGCTGG	36.4	36.4	0.0%	0.0%	0	calpain 3, (p94)	40300	O	S
TGCTGCCTGTT	36.4	18.5	0.0%	0.0%	3	HCGIV-6 protein / bone marrow stromal cell antigen 2	145477 / 118110	xs	xs
CTCATCAGCTT	36.4	6.2	0.0%	26.9%	4	adenylyl cyclase-associated protein	104125	O	S
ACCAAGGAGG	36.4	6.2	0.0%	0.0%	7	SRp25 nuclear protein / polymerase (RNA) II (DNA directed) polypeptide E (25kD	103561 / 24301	xs	xs

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
GCCTGGGACTC	36.4	3.1	0.5%	0.1%	7	hypothetical protein MGC2803	239894	H	U
GTGCGCTAGGG	36.4	2.3	1.8%	40.7%	9	Homo sapiens, clone IMAGE:5417844, mRNA, partial cds	9408	H	U
AAGGAAGATGG	34.8	34.8	0.0%	0.0%	0	proteasome (prosome, macropain) subunit, beta type, 8 (large multifunctional protease 7)	180062	K	AP
CTTTTTTCCCA	34.8	34.8	0.0%	0.0%	0	CD48 antigen (B-cell membrane protein)	901	K	CS
CCGCCGAAGTT	34.8	34.8	0.0%	0.0%	6			N	N
TGCCTTAATGC	34.8	34.8	0.0%	0.0%	0	Ras association (RalGDS/AF-6) domain family 1	26931	O	S
GGGCCCCCTGG	34.8	34.8	0.0%	0.0%	2	glycophorin C (Gerbich blood group)	81994	O	CS
AGCCCTCCCTG	34.8	5.9	0.0%	2.0%	7	RNA-binding protein (autoantigenic)	74111	H	U
CACTTCAAGGG	34.8	4.4	0.1%	46.1%	6	lymphocyte antigen 6 complex, locus E	77667	K	CS
TTCATTATAAT	34.8	3.5	0.3%	0.0%	9	prothymosin, alpha (gene sequence 28)	250655	O	CC
GTATTCCCCTT	34.8	2.2	2.5%	0.0%	9	poly(A)-binding protein, nuclear 1	117176	O	P
TAAGGACGAGA	33.2	33.2	0.0%	0.0%	0	hypothetical protein FLJ22457	238707	H	U
GGATATGTGGT	33.2	33.2	0.0%	0.0%	3	early growth response 1 / ESTs	326035 / 48285	xs	xs
CCCCGCCAAGT	33.2	8.4	0.0%	36.7%	5	calponin 2	169718	O	Cy
TTTCAGATTGG	33.2	8.4	0.0%	0.0%	3	activated RNA polymerase II transcription cofactor 4	356473	O	T
GCAAAACCTCA	33.2	4.2	0.2%	0.0%	4	ESTs / Homo sapiens mRNA for FLJ00121 protein	293218 / 283365	xs	xs
GTTCTGGTTTA	33.2	3.4	0.4%	6.3%	7	ATPase inhibitor precursor	241336	O	O
GGCCAGCCCTT	33.2	3.4	0.4%	36.7%	8	hypothetical protein MGC15429 / phosphofructokinase, liver	79 / 155455	xs	xs
TTAAAAAAAA	33.2	2.8	1.0%	2.8%	3	DNA fragmentation factor, 45 kD, alpha polypeptide / matrin 3 / DNA segment, numerous copies, expressed probes (GS1 gene) / chromosome 5 open reading frame 3 / Homo sapiens mRNA; cDNA DKFZp434D1835 (from clone DKFZp434D1835) / activating transcription factor 7 / yeast Sec31p homolog / hypothetical protein PRO2521 / candidate tumor suppressor p33 ING1 homolog / small nuclear RNA activating complex, polypeptide 5, 19kD/ NB too many to check, have not checked them all	105658 / 78825 / 78991 / 166551 / 278393 / 55888 / 70266 / 19054 / 108183 / 30174	xs	xs
AGAACCTTCAA	31.6	31.6	0.0%	35.6%	0	major histocompatibility complex, class I, A (Matches ESTs which match refseq 95%)	181244	K	AP
ATCCTGAGTTA	31.6	31.6	0.0%	0.0%	1	major histocompatibility complex, class II, DQ beta 1	73931	K	AP
TCCTATTAGC	31.6	31.6	0.0%	0.0%	1			N	N
CATTTACTCTA	31.6	16.1	0.0%	0.0%	0	integral membrane protein 2A	17109	H	U
CATCCAAAACA	31.6	16.1	0.0%	0.0%	4			N	N
TAATGTAAAGG	31.6	16.1	0.0%	0.0%	0	chromosome condensation 1-like	27007	O	S
ACCCTTCCCTC	31.6	16.1	0.0%	0.0%	6	signal sequence receptor, beta (translocon-associated protein beta)	74564	O	P

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
AATAAAGGCTA	31.6	8.0	0.0%	0.0%	7	ras homolog gene family, member C (Refseq sequence changed since release of SAGEmap, therefore cluster should be different)	179735	H	U
TGCTAAAAAAA	31.6	8.0	0.0%	27.3%	4	myosin, heavy polypeptide 9, non-muscle	146550	O	Cy
TTGCTGGAGAA	31.6	8.0	0.0%	0.0%	4	hypothetical protein FLJ21120 / serine/arginine repetitive matrix 2	133546 / 197114	xs	xs
CTGAGACAAAG	31.6	5.4	0.1%	10.1%	11	basic transcription factor 3	101025	O	T
ATGACAGATGG	31.6	5.4	0.1%	0.0%	0	lung cancer-associated Y protein	13775	H	U
CAGCTCATCTA	31.6	5.4	0.1%	0.0%	6	jumping translocation breakpoint	6396	H	U
GCTGTTGCGCG	31.6	3.2	0.7%	4.7%	9			N	N
TAACTGGAGGA	31.6	3.2	0.7%	0.0%	4	leukocyte receptor cluster (LRC) member 8 / KIAA1932 protein	348571 / 306121	xs	xs
CTCATAGCAGT	31.6	2.7	1.4%	0.0%	10	tumor protein, translationally-controlled 1	279860	K	E
TAATTTTGGA	30.0	30.0	0.0%	0.0%	3	Homo sapiens, clone MGC:16362 IMAGE:3927795, mRNA, complete cds	292457	H	U
AGCAGATCAGG	30.0	30.0	0.0%	0.0%	7	S100 calcium-binding protein A10 (annexin II ligand, calpactin I, light polypeptide (p11))	119301	O	Cy
TGCAAGAGAGG	30.0	30.0	0.0%	0.0%	0	Homo sapiens mRNA; cDNA DKFZp667K0625 (from clone DKFZp667K0625)	238954	H	U
GACTTGGCCT	30.0	30.0	0.0%	0.0%	0	Homo sapiens cDNA FLJ33028 fis, clone THYMU2000140	16291	H	U
CTGACAGTGAA	30.0	30.0	0.0%	0.0%	0	major histocompatibility complex, class II, DM alpha	77522	K	AP
GATAACACATT	30.0	30.0	0.0%	0.0%	0	small inducible cytokine A4 (homologous to mouse Mip-1b)	75703	K	E
GATCCTGAAAA	30.0	30.0	0.0%	0.0%	0	CD8 antigen, beta polypeptide 1 (p37)	2299	K	CS
CCTGGGGTAAG	30.0	30.0	0.0%	0.0%	1	major histocompatibility complex, class II, DQ alpha 1	198253	K	AP
CTGGGCCTGAA	30.0	30.0	0.0%	0.1%	1	LPS-induced TNF-alpha factor	76507	K	T
TGTGATCACAA	30.0	30.0	0.0%	0.0%	1	proteasome (prosome, macropain) subunit, beta type, 10	9661	K	AP
CTCAGGAAGCT	30.0	30.0	0.0%	0.0%	0	signal-induced proliferation-associated gene 1	7019	O	S
GCGACGAGGCG	30.0	15.3	0.0%	0.4%	12			N	N
TAAAATATGTC	30.0	15.3	0.0%	0.0%	0	CD8 antigen, alpha polypeptide (p32)	85258	K	CS
GTTGCAGATAA	30.0	15.3	0.0%	0.6%	3	O-linked N-acetylglucosamine (GlcNAc) transferase (UDP-N-acetylglucosamine:polypeptide-N-acetylglucosaminyl transferase)	100293	O	P
CAGCGCTGCAT	30.0	7.6	0.0%	44.3%	7	CDC37 (cell division cycle 37, S. cerevisiae, homolog)	160958	O	CC
TAGATAATGGC	30.0	7.6	0.0%	1.4%	7	FK506-binding protein 1A (12kD)	179661	K	S
TAGCTGCTGGT	30.0	7.6	0.0%	0.0%	4	splicing factor, arginine/serine-rich 11	11482	O	P
CGGATAAGGCC	30.0	5.1	0.1%	0.2%	1	nuclear prelamin A recognition factor	256526	H	U

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
ACAAAAAAAA	30.0	3.8	0.4%	0.0%	0	potassium channel, subfamily K, member 10 (TREK 2) / ESTs, Weakly similar to hypothetical protein FLJ20378 [Homo sapiens] [H.sapiens] / plus many ohers	288618 / 188128	xs	xs
GGGCCCAGGAG	28.4	28.4	0.0%	10.9%	5	hypothetical protein FLJ12150	118983	H	U
TAAACTGTTTC	28.4	28.4	0.0%	6.5%	5	ribosomal protein S14	244621	O	P
CGCCCCGGCGG	28.4	28.4	0.0%	0.0%	0	ESTs	196244	E	Est
TCTCTCAAAGT	28.4	28.4	0.0%	0.0%	0	CD53 antigen	82212	K	CS
TTTTAAGCTG	28.4	28.4	0.0%	0.0%	0	Arachidonate 5-lipoxygenase-activating protein	100194	K	O
TTAAACTTAAA	28.4	28.4	0.0%	0.9%	1	chemokine (C-X-C motif), receptor 4 (fusin)	89414	K	CS
GCTGAGTGCAG	28.4	28.4	0.0%	0.0%	0			N	N
CCAGTAATCCC	28.4	14.5	0.0%	0.0%	2	hypothetical protein MGC4368 / Homo sapiens cDNA FLJ13295 fis, clone OVARC1001240	9732 / 237078	H	U
CTAAACTTTT	28.4	14.5	0.0%	0.0%	0	inhibitor of DNA binding 2, dominant negative helix-loop-helix protein	180919	O	T
ATAAACAGATG	28.4	7.2	0.1%	0.1%	0	Homo sapiens cDNA FLJ14752 fis, clone NT2RP3003071 / SR rich protein	334825 / 18368	H	U
TCTGCAATGAA	28.4	3.6	0.6%	6.8%	8	hypothetical protein YR-29	8170	H	U
GCATTTAAATA	28.4	3.6	0.6%	0.0%	8	eukaryotic translation elongation factor 1 beta 2	275959	O	P
ACCGCCTGTGG	28.4	2.9	1.4%	6.8%	6	Chromosome 20 open reading frame 149	79625	H	U
GTGACAGAAGA	28.4	2.4	2.9%	3.4%	11	eukaryotic translation initiation factor 4A, isoform 1	129673	O	P
GGGAAGCAGAT	26.9	26.9	0.0%	0.0%	6			N	N
CGTTTTCTGAT	26.9	26.9	0.0%	9.1%	3	protein tyrosine phosphatase type IVA, member 2	82911	O	S
CCCAGCCTATT	26.9	26.9	0.0%	0.0%	2	ESTs, Moderately similar to hypothetical protein FLJ20489 [Homo sapiens] [H.sapiens]	173725	E	Est
CTCCTCCAAGT	26.9	26.9	0.0%	0.0%	0	Homo sapiens, clone MGC:40121 IMAGE:5216355, mRNA, complete cds	15284	H	U
GGAGCTTGAGG	26.9	26.9	0.0%	0.0%	0	chromosome 6 open reading frame 9	288316	H	U
GGTAGAACTAA	26.9	26.9	0.0%	0.0%	0	chromosome X open reading frame 9	61469	H	U
ATGGAAGTCTG	26.9	26.9	0.0%	0.0%	0	inositol polyphosphate-5-phosphatase, 145kD	155939	K	S
GTCACTCGATG	26.9	26.9	0.0%	0.0%	0	protein tyrosine phosphatase, receptor type, C	170121	K	CS
GTGCTAATATT	26.9	26.9	0.0%	0.0%	0	lectin-like NK cell receptor	136748	K	CS
TATGGCTGGTG	26.9	26.9	0.0%	0.0%	0	src family associated phosphoprotein 1	19126	K	S
TTGCGTGTGTC	26.9	26.9	0.0%	0.0%	1	dual specificity phosphatase 2	1183	K	S
ATATTTGGGAA	26.9	26.9	0.0%	0.0%	0			N	N
AAGGTCTCCGC	26.9	26.9	0.0%	0.0%	0			N	N
GCTGCTGCGCG	26.9	26.9	0.0%	0.0%	1			N	N
TGGCCTCCCCG	26.9	26.9	0.0%	0.0%	5	Rho GDP dissociation inhibitor (GDI) alpha	159161	O	S

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
AGAAAGATGTC	26.9	13.7	0.0%	0.0%	10	annexin A1	78225	K	S
CTCAGACAGTG	26.9	13.7	0.0%	1.1%	8	ribosomal protein S27-like	108957	O	P
AGGCTCCGTGG	26.9	13.7	0.0%	0.0%	0	minor histocompatibility antigen HA-1	196914	H	U
GATTTTCTGGG	26.9	13.7	0.0%	0.0%	0	pleckstrin homology, Sec7 and coiled/coil domains 4	7189	O	S
TGCTTGTCAAG	26.9	13.7	0.0%	0.0%	0	butyrophilin, subfamily 3, member A1	284283	O	CS
GTGCCATATTT	26.9	13.7	0.0%	1.3%	6	isocitrate dehydrogenase 2 (NADP+), mitochondrial	5337	O	O
CAGATGCAAAA	26.9	6.8	0.1%	0.0%	2	TLH29 protein precursor	94695	H	U
GGCCGCGTTCG	26.9	6.8	0.1%	4.8%	11	ribosomal protein S17	5174	O	P
GCGATTCCGGA	26.9	6.8	0.1%	1.3%	5	nuclear LIM interactor-interacting factor	283724	H	U
GCAGGTGGTTT	26.9	6.8	0.1%	0.5%	5	RNA-binding region (RNP1, RRM) containing 2	145696	O	P
CTGAGTCTCCC	26.9	4.6	0.4%	2.8%	6	guanine nucleotide binding protein (G protein), alpha inhibiting activity polypeptide 2	77269	O	S
AAGATCCCCGC	26.9	3.4	0.9%	27.6%	10	divalent cation tolerant protein CUTA	107187	H	U
GCGAAACTCC	26.9	2.7	2.1%	0.0%	4	ction unknown protein 1 (NB matches 2 cDNA but not the same as refseq)	13809	H	U
TTTGCGGTCCG	26.9	2.7	2.1%	0.0%	4	TSC-22-like	102447	O	T
TCTCCAAGGAG	25.3	25.3	0.0%	0.0%	0	ESTs, Weakly similar to PSF_HUMAN PTB-ASSOCIATED SPLICING FACTOR [H.sapiens]	189284	E	Est
GGACCAGCGCC	25.3	25.3	0.0%	0.0%	0	hypothetical protein FLJ21438	136979	H	U
GGAGCTGTCTG	25.3	25.3	0.0%	0.0%	0	differentially expressed in FDCP (mouse homolog) 6	15476	H	U
CTGGCCATCGA	25.3	25.3	0.0%	0.0%	2	Homo sapiens mRNA for FLJ00043 protein, partial cds	356684	H	U
AGACAGTCCCA	25.3	25.3	0.0%	0.0%	0	linker for activation of T cells	83496	K	S
GCTAGGAAACT	25.3	25.3	0.0%	1.9%	1			N	N
TCCCTATAGC	25.3	25.3	0.0%	0.0%	1			N	N
TCCCGTAATCG	25.3	25.3	0.0%	0.2%	3			N	N
GAGCCCCGAAT	25.3	25.3	0.0%	0.0%	0	parvin, gamma	120243	O	Cy
GCCAAGCCTGA	25.3	12.8	0.0%	7.2%	3	annexin A6	118796	O	S
CACTTTTGGGC	25.3	12.8	0.0%	33.1%	4	LIM and SH3 protein 1	334851	O	Cy
AACTGCTTCAA	25.3	12.8	0.0%	33.1%	6	actin related protein 2/3 complex, subunit 1A (41 kD)	11538	O	Cy
TTCTCTACAAG	25.3	12.8	0.0%	0.0%	2	likely ortholog of rat CDK5 activator-binding protein C53	20157	H	U
CCTCCCTGATG	25.3	12.8	0.0%	0.1%	2	dynamin 2	167013	O	Cy
AGAATAAAGCT	25.3	6.4	0.2%	0.0%	0	hypothetical protein DKFZp434G0920	98564	H	U
AGCCGGGATGG	25.3	6.4	0.2%	0.8%	1	proteasome (prosome, macropain) subunit, beta type, 9 (large multifunctional protease 2)	9280	K	AP
GGTGATGAGGA	25.3	4.3	0.6%	0.1%	3	putative breast adenocarcinoma marker (32kD)	12107	H	U
ACGTGGTGATG	25.3	4.3	0.6%	0.0%	5	HSPC023 protein	279945	H	U
AGGGCTGCCAT	25.3	4.3	0.6%	0.0%	1	adaptor-related protein complex 1, gamma 2 subunit	343244	O	Cy

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
CTAACCAGACA	25.3	3.2	1.4%	11.9%	6	capping protein (actin filament) muscle Z-line, beta	333417	O	Cy
GCTCTCTATGC	25.3	3.2	1.4%	7.2%	9	signal sequence receptor, delta (translocon-associated protein delta)	102135	O	P
TCAGGCATTTT	25.3	3.2	1.4%	0.1%	8	RAB1B, member RAS oncogene family	300816	O	S
GACTCACTTTT	25.3	2.6	3.0%	2.4%	12	peptidylprolyl isomerase B (cyclophilin B)	699	K	E
ACTGTTTCCCA	23.7	23.7	0.0%	0.0%	0	SHP2 interacting transmembrane adaptor	88012	K	S
ATATGTCAGGG	23.7	23.7	0.0%	0.0%	0	integrin, alpha L (antigen CD11A (p180), lymphocyte function-associated antigen 1; alpha polypeptide)	174103	K	CS
AGGCCAGGCCG	23.7	23.7	0.0%	0.1%	1	AT-hook transcription factor AKNA	159578	K	T
AGTGCCGTGTG	23.7	23.7	0.0%	0.0%	2	myxovirus (influenza) resistance 1, homolog of murine (interferon-inducible protein p78)	76391	K	O
GTGTGCCTCCA	23.7	23.7	0.0%	0.1%	2	interferon regulatory factor 3	75254	K	T
CGCACCATTGC	23.7	12.0	0.1%	9.7%	5	GCN5 (general control of amino-acid synthesis, yeast, homolog)-like 1	94672	O	T
CAGGAGGAGTT	23.7	6.0	0.3%	5.5%	10	glucose regulated protein, 58kD	289101	K	AP
AAGCTGCTGGA	23.7	6.0	0.3%	0.0%	7	HCNP protein; XPA-binding protein 2	9822	O	O
ATCGGGCCCGG	23.7	4.0	0.9%	30.5%	8	SCAN domain-containing 1	274411	O	T
ACTTGTTCGCT	23.7	4.0	0.9%	0.4%	4	Homo sapiens cDNA: FLJ22050 fis, clone HEP09454	173705	H	U
AAAGTGAAGA	23.7	3.0	2.1%	15.4%	9	ESTs / Homo sapiens mRNA; cDNA DKFZp762B195 (from clone DKFZp762B195) / FLJ23277 protein	374646 / 356766 / 334477	xs	xs
AACTCTTGAAG	23.7	3.0	2.1%	0.4%	8	eukaryotic translation initiation factor 3, subunit 3 (gamma, 40kD)	58189	O	P
GCGCTGGAGTG	23.7	2.4	4.3%	35.3%	10	hypothetical protein MGC3133	110695	H	U
GCCTTGATCTC	22.1	22.1	0.0%	7.6%	2	protein kinase D2	91146	O	S
GCATAATAAGG	22.1	22.1	0.0%	0.2%	0	ESTs	11594	E	Est
TGCCCCTGAAC	22.1	22.1	0.0%	0.0%	0	ESTs	154993	E	Est
AGTGATGTAAA	22.1	22.1	0.0%	0.0%	0	Homo sapiens clone CDABP0095 mRNA sequence	46919	H	U
CTGGTTTTATT	22.1	22.1	0.0%	0.0%	0	granzyme K (serine protease, granzyme 3; tryptase II)	3066	K	E
GGCAAGTGCAA	22.1	22.1	0.0%	0.0%	0	interferon regulatory factor 7	166120	K	T
TTAAATCCCAT	22.1	22.1	0.0%	0.0%	0	protein tyrosine phosphatase, receptor type, C	170121	K	CS
AAACGCCCAAT	22.1	22.1	0.0%	0.0%	7	interleukin enhancer binding factor 3, 90kD	256583	K	T
TCCTTTAAGCC	22.1	22.1	0.0%	0.0%	2			N	N
TTCCCTCGTGA	22.1	22.1	0.0%	0.6%	3	aspartyl-tRNA synthetase	80758	O	P
TGTCCTGGTTC	22.1	22.1	0.0%	4.0%	7	cyclin-dependent kinase inhibitor 1A (p21, Cip1)	179665	O	CC

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
TTTTAAAAAAA	22.1	22.1	0.0%	4.0%	0	arachidonate 5-lipoxygenase-activating protein / hypothetical protein FLJ13782 / ESTs / EST / EST	100194 / 257924 / 202287 / 335650 / 335747	xs	xs
TTGGCAGCCCA	22.1	11.2	0.1%	0.4%	7			N	N
TCCCTTAAGC	22.1	11.2	0.1%	0.0%	4			N	N
CTAATGCAAAA	22.1	5.6	0.5%	0.0%	3	PNAS-123	40092	H	U
AATACCTCGTG	22.1	5.6	0.5%	0.2%	6	Scotin	24220	H	U
GACTCTGTGTT	22.1	5.6	0.5%	0.0%	1	Rho GTPase activating protein 4	3109	O	S
GCCCGCCTTGT	22.1	3.7	1.3%	1.5%	10	polymerase (RNA) II (DNA directed) polypeptide J (13.3kD)	80475	O	P
TGTCGCTGGGG	22.1	3.7	1.3%	4.0%	10	Hypothetical protein MGC2198	227152	H	U
GTCTGCGTGCC	22.1	2.8	3.1%	12.9%	10	proteasome (prosome, macropain) subunit, alpha type, 1	82159	K	AP
GCCAGACCCCT	22.1	2.8	3.1%	0.0%	6	KIAA0515 protein	108945	H	U
CCGATCACCGG	20.5	20.5	0.0%	47.2%	8	eukaryotic translation initiation factor 2, subunit 2 (beta, 38kD)	12163	O	P
TGGAAGCATCT	20.5	20.5	0.0%	0.0%	0	HSPC022 protein	367740	H	U
GACGACTGACC	20.5	20.5	0.0%	0.3%	2	interferon, gamma-inducible protein 16	155530	H	U
CAGGATGCTTG	20.5	20.5	0.0%	0.0%	0	lymphocyte-specific protein 1	56729	K	Cy
CCCCCTCCTGC	20.5	20.5	0.0%	0.0%	0	interleukin 2 receptor, beta (CD122)	75596	K	CS
CCTCAGCCCTG	20.5	20.5	0.0%	0.0%	0	protein tyrosine phosphatase, non-receptor type 6	63489	K	S
AAGAAGCTGA	20.5	20.5	0.0%	0.0%	0			N	N
TCCCTACATCG	20.5	20.5	0.0%	2.6%	3			N	N
GCACAGAGCA	20.5	20.5	0.0%	0.0%	0	basic leucine zipper transcription factor, ATF-like	41691	O	T
TTTCTGTCTGG	20.5	20.5	0.0%	2.6%	0	phosphoinositide-3-kinase, catalytic, delta polypeptide	162808	O	S
ATCCTTTTAT	20.5	20.5	0.0%	0.0%	1	golgin-67	182982	O	P
GGCGTGAACCT	20.5	20.5	0.0%	1.0%	7	proliferating cell nuclear antigen	78996	O	CC
GACTCTGAAAA	20.5	10.4	0.2%	9.5%	4	ribosomal protein S15a	343665	O	P
TGGCTGGGAAA	20.5	10.4	0.2%	24.8%	5	vesicle-associated membrane protein 8 (endobrevin)	172684	O	Cy
TCTCAATTCT	20.5	10.4	0.2%	43.3%	11	Sec23 (S. cerevisiae) homolog B / cell division cycle 42 (GTP-binding protein, 25kD)	173497 / 146409	xs	xs
GGGGTCCTTCA	20.5	10.4	0.2%	0.3%	2	KIAA0082 protein	154045	H	U
CTGGGTGCCCC	20.5	10.4	0.2%	0.0%	3	zinc finger protein 335	165983	H	U
TGGAGAAGAG	20.5	10.4	0.2%	0.0%	5	thioredoxin interacting protein	179526	H	U
TCCCATAAGC	20.5	10.4	0.2%	0.0%	3			N	N
TTTGTTTTTGA	20.5	10.4	0.2%	0.3%	5	SLC2A4 regulator	170088	O	T
GCGAAAACCC	20.5	5.2	0.7%	0.0%	3	ESTs, Weakly similar to hypothetical protein FLJ20958 [Homo sapiens] [H.sapiens]	104720	E	Est

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
AGACAGTCATT	20.5	3.5	2.0%	0.0%	3			N	N
AGCACATTTGA	20.5	2.6	4.5%	13.1%	8	coactosin-like protein	289092	O	Cy
GAGCGGGATCA	20.5	2.6	4.5%	2.6%	9	splicing factor, arginine/serine-rich 1 (splicing factor 2, alternate splicing factor) / Splicing factor, arginine/serine-rich, 46kD	73737 / 155160	O	P
CCTTTGAACAG	19.0	19.0	0.0%	18.3%	7	Homo sapiens cDNA: FLJ23602 fis, clone LNG15735	181634	H	U
TCCCGACATC	19.0	19.0	0.0%	21.8%	4			N	N
AAGGAATCGGG	19.0	19.0	0.0%	40.7%	9			N	N
CTCCAATAAAA	19.0	19.0	0.0%	9.6%	4	HLA B associated transcript 1 (NB poorly assembled cluster actually matches gene talin 1) / ESTs	55296 / 176860	xs	xs
TCAGTTGCCTC	19.0	19.0	0.0%	0.1%	0	ESTs, Weakly similar to activation-induced cytidine deaminase; activation induced cytidine deaminase [Homo sapiens] [H.sapiens]	236533	E	Est
AGCAAGAAACT	19.0	19.0	0.0%	0.1%	0	chromosome 3 open reading frame 4	107393	H	U
ATTATCCAGCG	19.0	19.0	0.0%	0.1%	2	RNA binding motif protein 3	301404	H	U
ACGCTCTCGAT	19.0	19.0	0.0%	0.1%	0	CD37 antigen	153053	K	CS
ATAAAGAGGTT	19.0	19.0	0.0%	0.1%	0	proline-serine-threonine phosphatase interacting protein 1	129758	K	S
GAGGCTCGGCT	19.0	19.0	0.0%	0.1%	0	protein tyrosine phosphatase, non-receptor type 7	35	K	S
AGAGGGAGTGA	19.0	19.0	0.0%	0.1%	1	C-type (calcium dependent, carbohydrate-recognition domain) lectin, superfamily member 2 (activation-induced)	85201	K	CS
GGCCTGCAGGA	19.0	19.0	0.0%	0.1%	1	apoptosis-associated speck-like protein containing a CARD	71869	K	S
CGAAGGCTGTA	19.0	19.0	0.0%	0.4%	2	nuclear factor, interleukin 3 regulated	79334	K	T
CTGGGGTGAGC	19.0	19.0	0.0%	0.1%	0			N	N
TCTGTGTTGAA	19.0	19.0	0.0%	0.1%	0			N	N
CCTCGGAGATC	19.0	19.0	0.0%	0.1%	4	splicing factor, arginine/serine-rich 7 (35kD)	184167	O	P
GCGGGGTGGAG	19.0	19.0	0.0%	1.5%	6	zinc finger protein 36, C3H type-like 1	85155	O	T
AGTATCTGGGA	19.0	19.0	0.0%	0.1%	7	actin related protein 2/3 complex, subunit 1A (41 kD)	11538	O	Cy
GGATGTGAAAG	19.0	9.6	0.3%	9.6%	8	antigen identified by monoclonal antibodies 12E7, F21 and O13	177543	K	CS
AGGAAAAGATG	19.0	9.6	0.3%	14.0%	2	polymerase (DNA-directed), delta 4	82520	O	CC
ACTGGTAAAAA	19.0	9.6	0.3%	0.4%	6	ATP synthase, H+ transporting, mitochondrial F0 complex, subunit f, isoform 2	155751	O	O
TTCTATTTCAG	19.0	9.6	0.3%	50.6%	7	moesin	170328	O	Cy
CCAAAAAAAAA	19.0	9.6	0.3%	0.2%	8	ninjurin 1 / interferon-induced protein 35 / adrenergic, beta-2-, receptor, surface / nucleoporin 50kD / ovarian carcinoma immunoreactive antigen	11342 / 50842 / 2551 / 271623 / 132071	xs	xs
TGTTGACTCTG	19.0	9.6	0.3%	0.1%	0	ESTs	8882	E	Est
GTGCCACCAGT	19.0	9.6	0.3%	0.1%	1	KIAA1554 protein	17767	H	U

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
GTCTGCCTGGC	19.0	9.6	0.3%	0.1%	1	granzyme M (lymphocyte met-ase 1)	268531	K	E
GCCGCCTGCCT	19.0	9.6	0.3%	0.4%	3	IMP (inosine monophosphate) dehydrogenase 1	850	O	O
CCACGAGTAAA	19.0	9.6	0.3%	0.1%	4	SMC4 (structural maintenance of chromosomes 4, yeast)-like 1	50758	O	CC
AAAGCAAACCA	19.0	4.8	1.2%	8.0%	7	ESTs in E2F transcription factor 5, p130-binding	2331	E	Est
GGGCGAGAACA	19.0	4.8	1.2%	21.8%	2	Optineurin	278898	H	U
GTTTTTCATT	19.0	4.8	1.2%	30.9%	8	cyclin-E binding protein 1	26663	H	U
GAGTAGAGAAA	19.0	4.8	1.2%	14.0%	9			N	N
TGCAGCGCCTG	19.0	4.8	1.2%	0.1%	2	uridine phosphorylase	77573	O	O
AATATTGAGA	19.0	4.8	1.2%	0.4%	7	eukaryotic translation initiation factor 3, subunit 6 (48kD)	106673	O	T
GTAAAACCCTG	19.0	4.8	1.2%	0.4%	5	Homo sapiens cDNA FLJ30541 fis, clone BRAWH2001355 / Peroxisomal trans 2-enoyl CoA reductase; putative short chain alcohol dehydrogenase	351853 / 281680	xs	xs
AAGAAGCAGGG	19.0	3.2	3.1%	21.8%	5	chromosome 1 open reading frame 8	11441	H	U
GATGAGTCTCG	19.0	3.2	3.1%	40.1%	12	proteasome (prosome, macropain) subunit, alpha type, 7	233952	K	AP
CCACTCCTCAA	19.0	3.2	3.1%	0.0%	10	defender against cell death 1	82890	O	P
CCCGACGTGCC	19.0	3.2	3.1%	50.6%	11	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 3 (9kD, B9)	198269	O	O
GAAATTTAAAG	19.0	3.2	3.1%	31.6%	12	high-mobility group (nonhistone chromosomal) protein 1	337757	O	O
TTGAAGTGGTT	19.0	3.2	3.1%	0.1%	2	hypothetical protein FLJ10154	179972	H	U
CTCCTTAAGAG	19.0	3.2	3.1%	0.1%	3			N	N
AAATAAAAAGT	19.0	3.2	3.1%	0.1%	0	myosin IXB	159629	O	Cy
GGCGCCAAAAA	19.0	3.2	3.1%	0.1%	4	kinesin-like 4	119324	O	CC
GCTAGTGATGT	17.4	17.4	0.1%	37.8%	4	chromosome 15 open reading frame 15	284162	H	U
TGAAGCAGTAA	17.4	17.4	0.1%	27.7%	6	programmed cell death 4 (neoplastic transformation inhibitor)	326248	H	U
TACAGCACAAA	17.4	17.4	0.1%	0.1%	0	ESTs	260395	E	Est
TCAGTGACCAG	17.4	17.4	0.1%	0.1%	0	hypothetical protein MGC5363	1880	H	U
GGCGGGGCCAG	17.4	17.4	0.1%	0.1%	1	KIAA0303 protein	54985	H	U
GGTAGCCCACG	17.4	17.4	0.1%	2.4%	1	trinucleotide repeat containing 5	56828	H	U
TACGAGGCCGG	17.4	17.4	0.1%	0.1%	1	expressed in activated T/LAK lymphocytes	16165	H	U
TGGGGCCGCAG	17.4	17.4	0.1%	0.1%	1	Homo sapiens cDNA: FLJ23270 fis, clone COL10309, highly similar to HSU33271 Human normal keratinocyte mRNA	288455	H	U
ACAAAGCCCCA	17.4	17.4	0.1%	0.7%	2	similar to APOBEC1	8583	H	U
ATGGCCTCCTC	17.4	17.4	0.1%	0.1%	3	syntaxin 4A (placental)	83734	H	U
TGATCTGCCTG	17.4	17.4	0.1%	0.1%	3	hypothetical protein MGC5178	326067	H	U
AACATTTCTCT	17.4	17.4	0.1%	0.1%	0	hematopoietic cell-specific Lyn substrate 1	14601	K	S
CCTCTCCAACA	17.4	17.4	0.1%	0.1%	0	major histocompatibility complex, class II, DM beta	1162	K	AP

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
CTTCTTTCATT	17.4	17.4	0.1%	0.1%	0	CD8 antigen, beta polypeptide 1 (p37)	2299	K	CS
GCCAAGGAGG	17.4	17.4	0.1%	0.1%	0	CD7 antigen (p41)	36972	K	CS
TTGAGATAACT	17.4	17.4	0.1%	0.1%	0	lymphocyte cytosolic protein 2 (SH2 domain-containing leukocyte protein of 76kD)	2488	K	S
AAGGTGGCCAT	17.4	17.4	0.1%	0.7%	2	Homo sapiens, transmembrane activator and CAML interactor, clone MGC:39952 IMAGE:5213128, mRNA, complete cds protein sequence same as: Homo sapiens tumor necrosis factor receptor superfamily, member 13B)	374625	K	CS
GTGCAGGCTCC	17.4	17.4	0.1%	0.7%	2	transporter 1, ATP-binding cassette, sub-family B (MDR/TAP)	158164	K	AP
CACAAACGGG	17.4	17.4	0.1%	0.1%	0			N	N
TTGGCTGGTGT	17.4	17.4	0.1%	0.1%	0			N	N
TTTTCTCTGCA	17.4	17.4	0.1%	0.1%	1			N	N
TCCCGTCATC	17.4	17.4	0.1%	2.4%	3			N	N
GCAGACATTGA	17.4	17.4	0.1%	2.4%	6			N	N
GCACCCAACAC	17.4	17.4	0.1%	0.1%	0	lectin, galactoside-binding, soluble, 8 (galectin 8)	4082	O	E
AGGAGGGATAA	17.4	17.4	0.1%	0.1%	1	RAB24 Homo sapiens, clone MGC:29471 IMAGE:4329216, mRNA, complete cds	16258	O	Cy
ATGTGTAACG	17.4	17.4	0.1%	0.1%	7	S100 calcium-binding protein A4 (calcium protein, calvasculin, metastasin, murine placental homolog)	81256	O	S
CTTAAGGATT	17.4	17.4	0.1%	0.1%	8	PAI-1 mRNA-binding protein	165998	O	P
ATGGCAACAGA	17.4	17.4	0.1%	0.1%	5	aquaporin 1 (channel-forming integral protein, 28kD) / integrin, alpha 5 (fibronectin receptor, alpha polypeptide)	74602 / 149609	xs	xs
GCTCCAGCCAT	17.4	8.8	0.5%	18.6%	3	interferon-stimulated transcription factor 3, gamma (48kD), Refseq changed since assembly of SAGEmap, therefore now maps a cDNA which is not similar to the cluster	1706	H	U
CTGACCCCCTT	17.4	8.8	0.5%	18.8%	5	beta-1,3-glucuronyltransferase 3 (glucuronosyltransferase I)	26492	O	P
TTGATGCCCGA	17.4	8.8	0.5%	0.7%	2	serine (or cysteine) proteinase inhibitor, clade B (ovalbumin), member 1	183583	K	E
CCCCTCTGAG	17.4	8.8	0.5%	0.7%	8	adenosine deaminase, RNA-specific	7957	K	O
ACCCACGTCAG	17.4	8.8	0.5%	0.1%	6	jun B proto-oncogene	198951	O	T
GTGCTGCGTGA	17.4	4.4	1.8%	11.1%	7	mitochondrial ribosomal protein L37	4209	O	P
CTTTCAGATGT	17.4	4.4	1.8%	6.8%	10	phosphofructokinase, platelet	99910	O	O
AAAAGAAACT	17.4	4.4	1.8%	37.8%	12	poly(A)-binding protein, cytoplasmic 1	172182	O	P
GGGCTGCTTTT	17.4	4.4	1.8%	0.7%	3	Homo sapiens cDNA FLJ14845 fis, clone PLACE1000308 / Homo sapiens cDNA FLJ14752 fis, clone NT2RP3003071	334711 / 334825	H	U
TCAACTTCTGG	17.4	4.4	1.8%	0.1%	0	solute carrier family 2 (facilitated glucose transporter), member 3	7594	O	CS
GCGACCAACAT	17.4	4.4	1.8%	0.1%	4	SON DNA binding protein	92909	O	T

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
AGGGAAAGAGG	17.4	4.4	1.8%	0.7%	5	maternal G10 transcript	348416	O	T
TGTTTGTGTGT	17.4	2.9	4.6%	5.7%	3	Homo sapiens, clone MGC:19762 IMAGE:3636045, mRNA, complete cds / sema domain, immunoglobulin domain (Ig), transmembrane domain TM) and short cytoplasmic domain, (semaphorin) 4C	343214 / 7188	H	U
CAGCTCCGCTT	17.4	2.9	4.6%	42.0%	6	dUTP pyrophosphatase	82113	O	CC
AGGCTGGATGC	17.4	2.9	4.6%	2.4%	4	KIAA0668 protein	5898	H	U
GGAAGGGAGGC	17.4	2.9	4.6%	0.7%	8	hypothetical protein FLJ20568	279581	H	U
CCCTGAATCCC	17.4	2.9	4.6%	0.1%	4	protein kinase, lysine deficient 1	184592	O	S
GTGGACCCTGT	17.4	2.9	4.6%	0.1%	4	eNOS interacting protein	7236	O	O
CCCACTTGTAA	15.8	15.8	0.2%	1.3%	3	ESTs, Highly similar to CH60_HUMAN 60 KDA HEAT SHOCK PROTEIN, MITOCHONDRIAL PRECURSOR [H.sapiens] / ESTs	331803 / 356388	E	Est
GTGCTGGTCAG	15.8	15.8	0.2%	26.1%	2			N	N
TTAATGCGTCT	15.8	15.8	0.2%	0.2%	0	inhibitor of DNA binding 2, dominant negative helix-loop-helix protein (Matches ESTs in this cluster but not the contig with the Refseq)	180919	E	Est
TGCCAATTAAG	15.8	15.8	0.2%	0.2%	1	ESTs	165337	E	Est
GGTTCAAGGCC	15.8	15.8	0.2%	0.2%	2	hypothetical protein FLJ22635	353181	H	U
TGTGGGAACCA	15.8	15.8	0.2%	0.2%	4	hypothetical protein AL133206	7750	H	U
AATCTGCGCCT	15.8	15.8	0.2%	0.2%	8	interferon-stimulated protein, 15 kDa	833	H	U
GCGTCCTGCCC	15.8	15.8	0.2%	0.2%	0	linker for activation of T cells	83496	K	S
TAAGGGAGCCA	15.8	15.8	0.2%	0.2%	0	T cell activation, increased late expression	142023	K	CS
TTCACTTCCAA	15.8	15.8	0.2%	0.2%	0	SH2 domain protein 1A, Duncan's disease (lymphoproliferative syndrome)	151544	K	S
AAGCACCTTGA	15.8	15.8	0.2%	1.2%	1	TNFRSF1A-associated via death domain	89862	K	S
TTTTTCTTCTC	15.8	15.8	0.2%	0.2%	1	interleukin 2 receptor, gamma (severe combined immunodeficiency)	84	K	CS
GGTCAAAGGAA	15.8	15.8	0.2%	1.2%	3	natural killer-tumor recognition sequence	241493	K	CS
TGGCAGAACTG	15.8	15.8	0.2%	0.2%	0			N	N
ATCTTGGCCCA	15.8	15.8	0.2%	0.2%	0	XIAP associated factor-1	139262	O	CC
CAGGACAGGGT	15.8	15.8	0.2%	0.2%	0	NESH protein	130719	O	S
AGTTGGACGGA	15.8	15.8	0.2%	0.2%	2	ligase I, DNA, ATP-dependent	1770	O	O
AGACTTGGCAT	15.8	15.8	0.2%	0.2%	3	ARP3 (actin-related protein 3, yeast) homolog	5321	O	Cy
CCTTTTTGTCC	15.8	8.0	0.9%	15.2%	2	hypothetical protein MGC4090 / Homo sapiens mRNA; cDNA DKFZp586K1318 (from clone DKFZp586K1318)	58389 / 62601	H	U
CAGCCCAACCG	15.8	8.0	0.9%	34.4%	11	eukaryotic translation initiation factor 3, subunit 4 (delta, 44kD)	28081	O	P
GGAAGTTTCGA	15.8	8.0	0.9%	34.6%	11	mitochondrial ribosomal protein 64	55847	O	P

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
GTTCTCCCACT	15.8	8.0	0.9%	15.2%	11	protein transport protein SEC61 alpha subunit isoform 1	306079	O	P
GAGAATCAGAG	15.8	8.0	0.9%	0.2%	0	ESTs, Weakly similar to I55214 salivary proline-rich glycoprotein precursor - rat [R.norvegicus] / tropomodulin 3 (ubiquitous), matches ESTs not Refseq	355961 / 22826	E	Est
GCGGCTGACAG	15.8	8.0	0.9%	0.2%	2	KIAA1966 protein, matches ESTs from contig without mRNA	158184	E	Est
GATGAAAAGGA	15.8	8.0	0.9%	0.2%	1	Homo sapiens, clone MGC:19482 IMAGE:4309314, mRNA, complete cds / DnaJ (Hsp40) homolog, subfamily C, member 4, two clusters, treat as DnaJ homolog	343473 / 172847	H	U
AAGACCGAGGG	15.8	8.0	0.9%	0.2%	2	Homo sapiens, clone IMAGE:3611719, mRNA, partial cds	244482	H	U
GGGGCTTCCAG	15.8	8.0	0.9%	3.6%	2	KIAA0239 protein	9729	H	U
TCGGGTGTGGG	15.8	8.0	0.9%	0.2%	2	hypothetical protein MGC15906	104938	H	U
TCTAACTCACC	15.8	8.0	0.9%	0.2%	2	Homo sapiens mRNA; cDNA DKFZp586J101 (from clone DKFZp586J101)	322645	H	U
CAGGGAGCGCC	15.8	8.0	0.9%	0.2%	4	PC2 (positive cofactor 2, multiprotein complex) glutamine/Q-rich-associated protein	8657	O	T
CTGGCGAGCGC	15.8	8.0	0.9%	0.2%	10	ubiquitin carrier protein	174070	O	O
GAACCCTGGGA	15.8	8.0	0.9%	3.6%	10	mitochondrial ribosomal protein S34	157160	O	P
ACTGGCTGCTG	15.8	4.0	2.9%	15.2%	9	ESTs, Weakly similar to A42442 integrin beta-1 chain, splice form beta-1-S [H.sapiens]	349092	E	Est
AAGGAACTTGG	15.8	4.0	2.9%	15.2%	6	Hypothetical protein FLJ12443	179882	H	U
GCTGGGCGGCT	15.8	4.0	2.9%	1.2%	3	hypothetical protein FLJ10140	250671	H	U
TGTGTGGGGCC	15.8	4.0	2.9%	0.2%	4	Homo sapiens, clone IMAGE:3629896, mRNA, partial cds	21497	H	U
CTGCCTTCTTG	15.8	4.0	2.9%	0.2%	7	protein phosphatase 1, catalytic subunit, gamma isoform	79081	O	S
AGGGTGAAACT	15.8	4.0	2.9%	1.2%	8	splicing factor, arginine/serine-rich 9	77608	O	P
AAAGGGGGCAG	15.8	4.0	2.9%	1.2%	9	heterogeneous nuclear ribonucleoprotein A3	249247	O	P
AAGGTAGCAGA	15.8	4.0	2.9%	3.6%	9	adenylyl cyclase-associated protein	104125	O	S
GTGTCTCATC	15.8	4.0	2.9%	3.6%	6	nuclear receptor co-repressor 1 / Homo sapiens, clone IMAGE:3633225, mRNA / ESTs	144904 / 356377 / 131563	xs	xs
TTTCAATAGA	14.2	14.2	0.3%	5.5%	1	Homo sapiens cDNA: FLJ21920 fis, clone HEP04049 / hypothetical protein MGC10986	288025 / 50601	H	U
TAATTTTGAAA	14.2	14.2	0.3%	0.0%	5	Homo sapiens, clone MGC:16362 IMAGE:3927795, mRNA, complete cds	292457	H	U
TTGCTAGAGGG	14.2	14.2	0.3%	19.8%	10	ubiquitously-expressed transcript	172791	H	U
AGAACCAAAAA	14.2	14.2	0.3%	31.0%	2	major histocompatibility complex, class I, A	181244	K	AP
CATCTTCACCA	14.2	14.2	0.3%	19.8%	12			N	N
GGGAGGTAGCA	14.2	14.2	0.3%	11.7%	2	basic helix-loop-helix domain containing, class B, 2	171825	O	CC

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
AGGATGACCCC	14.2	14.2	0.3%	31.0%	5	FXYD domain-containing ion transport regulator 5	333418	O	O
CCGTGGTCACC	14.2	14.2	0.3%	31.0%	10	adaptor-related protein complex 2, sigma 1 subunit	119591	O	Cy
GGGGGGGGTC	14.2	14.2	0.3%	5.5%	1	protein phosphatase 2, regulatory subunit B (B56), gamma isoform / Homo sapiens cDNA FLJ31993 fis, clone NT2RP7009168	171734 / 182648	xs	xs
CTTCCAGCTA	14.2	14.2	0.3%	0.0%	12	annexin A2 / Homo sapiens mRNA; cDNA DKFZp434C107 (from clone DKFZp434C107)	217493 / 101651	xs	xs
GGTCTTCTAGT	14.2	14.2	0.3%	1.9%	1	hypothetical protein FLJ10330	342307	H	U
TACCGCCCGTA	14.2	14.2	0.3%	0.4%	4	hypothetical protein MGC5442	238513	H	U
AAGAGACCCTG	14.2	14.2	0.3%	0.4%	0	ectonucleoside triphosphate diphosphohydrolase 1 (matches ESTs in diff contig to Refseq)	205353	K	CS
ATGGAGCGCAC	14.2	14.2	0.3%	0.4%	0	tumor necrosis factor receptor superfamily, member 1B	256278	K	CS
CCTGGTGCTTC	14.2	14.2	0.3%	0.4%	0	interferon, gamma	856	K	E
TATGAGGAGG	14.2	14.2	0.3%	0.4%	0	matrix metalloproteinase 25 (leukolysin)	198265	K	E
TATGTCTTGGA	14.2	14.2	0.3%	0.4%	0	lymphocyte-specific protein tyrosine kinase	1765	K	S
AGCAGCTGCTG	14.2	14.2	0.3%	0.4%	1	megakaryocyte-associated tyrosine kinase	274	K	S
ACCGGCCTGTG	14.2	14.2	0.3%	0.4%	0			N	N
CACCAAGGATC	14.2	14.2	0.3%	0.4%	0			N	N
CTTAATCTTGT	14.2	14.2	0.3%	1.9%	2			N	N
TCCCTTTAGCC	14.2	14.2	0.3%	0.4%	2			N	N
TCCTATTAAC	14.2	14.2	0.3%	0.4%	2			N	N
ACCCCCAAGCA	14.2	14.2	0.3%	0.4%	0	pleckstrin	77436	O	S
AGATTTGTGGC	14.2	14.2	0.3%	0.4%	0	FGD1 family, member 3	5013	O	S
CCTGTAGCCTT	14.2	14.2	0.3%	0.4%	0	Gardner-Rasheed feline sarcoma viral (v-fgr) oncogene homolog	1422	O	S
CGCGTCAGAGC	14.2	14.2	0.3%	0.4%	0	golgin-67	182982	O	P
AGCCTCGGCCA	14.2	14.2	0.3%	1.9%	1	Rho guanine nucleotide exchange factor (GEF) 1	252280	O	S
CCGCTTACTCT	14.2	14.2	0.3%	1.9%	1	v-ets erythroblastosis virus E26 oncogene homolog 1 (avian)	18063	O	T
GTCCCTCTCAA	14.2	14.2	0.3%	0.4%	4	calcium-regulated heat-stable protein (24kD)	92198	O	S
CTTTCTTCCCT	14.2	14.2	0.3%	1.9%	8	zinc finger protein 36, C3H type-like 1	85155	O	T
TTTGAGAATA	14.2	14.2	0.3%	0.4%	1	ESTs, Highly similar to A47328 natural killer cell tumor-recognition protein [H.sapiens] / ecotropic viral integration site 2B	310646 / 5509	xs	xs
AGCCTGCCTGA	14.2	7.2	1.5%	46.3%	4	heat shock transcription factor 1	1499	O	T
ATGCAGCCATA	14.2	7.2	1.5%	5.5%	9	ornithine decarboxylase 1	75212	O	O
ATTTGTCCCAG	14.2	7.2	1.5%	2.9%	9	high-mobility group (nonhistone chromosomal) protein isoforms I and Y	139800	O	T
CAAGGGCTTGC	14.2	7.2	1.5%	14.1%	9	RAP1B, member of RAS oncogene family	156764	O	S

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
CCGTGGTCGTG	14.2	7.2	1.5%	11.7%	9	fibrillarin	99853	O	P
CGCGGCAGCTC	14.2	7.2	1.5%	0.4%	0	ESTs, Weakly similar to MASL1 [H.sapiens]	132216	E	Est
TTGATGCCCTA	14.2	7.2	1.5%	0.4%	1	hypothetical protein FLJ10081	7871	H	U
AGCTCTATGAT	14.2	7.2	1.5%	0.4%	3	COBW-like protein	7535	H	U
TGCTGCTGCTT	14.2	7.2	1.5%	0.4%	5	hypothetical protein FLJ14166 / hypothetical protein FLJ20396	14070 / 283685	H	U
ATCCTCCCTAT	14.2	7.2	1.5%	0.4%	1	RAP1A, member of RAS oncogene family	865	K	S
TGGCCCTCCAG	14.2	7.2	1.5%	1.9%	5	signal transducer and activator of transcription 6, interleukin-4 induced	181015	K	T
TTGTCACAAAT	14.2	7.2	1.5%	0.4%	0			N	N
ACAAATTAACA	14.2	7.2	1.5%	1.9%	1	splicing factor, arginine/serine-rich 11	11482	O	P
GCTGGCTGTTT	14.2	7.2	1.5%	1.9%	1	SKIP for skeletal muscle and kidney enriched inositol phosphatase	178347	O	S
TGTGCACCCC	14.2	7.2	1.5%	1.9%	3	Mouse Mammary Turmor Virus Receptor homolog	18686	O	CS
CTGGCCCGGAG	14.2	7.2	1.5%	0.4%	6	vasodilator-stimulated phosphoprotein	93183	O	Cy
AACGCTGCCTG	14.2	7.2	1.5%	0.4%	7	adenine phosphoribosyltransferase	28914	O	O
TCTGCTAAAA	14.2	7.2	1.5%	1.9%	3	ubiquitin-conjugating enzyme E2G 2 (homologous to yeast UBC7) / Hypothetical protein FLJ32332	192853 / 373560	xs	xs
GTGAGTGTGTC	14.2	3.6	4.5%	5.5%	4	hypothetical protein FLJ12953 similar to Mus musculus D3Mm3e	323537	H	U
GGCCAAAGAGG	14.2	3.6	4.5%	5.5%	4			N	N
GGCGTCCTGGC	14.2	3.6	4.5%	31.0%	8			N	N
CTGCTGTGATA	14.2	3.6	4.5%	5.5%	7	small nuclear ribonucleoprotein polypeptide C	1063	O	P
TCCAAGGAAGG	14.2	3.6	4.5%	42.4%	7	peroxisomal D3,D2-enoyl-CoA isomerase	15250	O	O
CAACTTAGTTT	14.2	3.6	4.5%	0.0%	10	myosin regulatory light chain	180224	O	Cy
GGGTTTTTATT	14.2	3.6	4.5%	0.0%	12	nuclease sensitive element binding protein 1	74497	O	T
AGCCTGGAGAG	14.2	3.6	4.5%	0.4%	4	ESTs	33184	E	Est
GTGACATCTCC	14.2	3.6	4.5%	0.4%	1			N	N
GCGACGGCCGT	14.2	3.6	4.5%	0.4%	5			N	N
TCCCGTACAC	14.2	3.6	4.5%	0.4%	6			N	N
TACAATAATTT	14.2	3.6	4.5%	0.4%	7			N	N
TCTCAAAAAA	14.2	3.6	4.5%	0.4%	1	apolipoprotein F	2388	O	O
TGGGCAGCTGG	14.2	3.6	4.5%	1.9%	2	ribosomal protein S9 / ESTs	180920	O	P
TTGTTGGATAT	14.2	3.6	4.5%	0.4%	2	nardilysin (N-arginine dibasic convertase)	4099	O	O
CCCCCAATGCT	14.2	3.6	4.5%	0.4%	4	splicing factor 3a, subunit 2, 66kD	115232	O	P
ATGCGAAAGG	14.2	3.6	4.5%	0.4%	7	dodecenoyl-Coenzyme A delta isomerase (3,2 trans-enoyl-Coenzyme A isomerase)	89466	O	O

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
CAAACTCAAAA	12.6	12.6	0.5%	8.2%	2	WW domain-containing adapter with a coiled-coil region (NB matches ESTs in contig but not refseq sequences)	70333	E	Est
GAAGTCATTTT	12.6	12.6	0.5%	27.1%	2	Matches ESTs in clustee: hypothetical protein FLJ21939 similar to 5-azacytidine induced gene 2	164478	E	Est
CACACCCATTA	12.6	12.6	0.5%	8.2%	1	hypothetical protein MGC15875	315054	H	U
ATTTTTTAACA	12.6	12.6	0.5%	51.1%	3	HSPCO34 protein	46967	H	U
GAAGTGGAAGC	12.6	12.6	0.5%	27.1%	8	hypothetical protein MGC2477	9061	H	U
GGGGCTGTATT	12.6	12.6	0.5%	27.1%	8	transforming growth factor, beta 1	1103	K	E
TCCGTACATCG	12.6	12.6	0.5%	8.2%	2			N	N
TGTGTCAAAGT	12.6	12.6	0.5%	8.2%	8			N	N
TTTTTGATAAA	12.6	12.6	0.5%	8.2%	8			N	N
ATACTTTAATC	12.6	12.6	0.5%	27.1%	11			N	N
CAGGCCCCACC	12.6	12.6	0.5%	0.0%	9	S100 calcium-binding protein A11 (calgizzarin)	256290	O	S
CCCCCACCTAA	12.6	12.6	0.5%	8.2%	9	proteolipid protein 2 (colonic epithelium-enriched)	77422	O	CS
AGTTAAAACCA	12.6	12.6	0.5%	0.6%	0	ESTs, Weakly similar to I57588 HSrel-1 [H.sapiens]	283364	E	Est
CCCCAAAGCTG	12.6	12.6	0.5%	0.6%	0	ESTs	192855	E	Est
TGCCAGGTGCA	12.6	12.6	0.5%	0.6%	1	ESTs (possibly fits in albumin, cluster too large to assemble)	356560	E	Est
TCCCTAAAGC	12.6	12.6	0.5%	0.6%	2	ESTs	196128	E	Est
TTGCAATGCA	12.6	12.6	0.5%	3.0%	6	ESTs	355952	E	Est
AGTCTTTAGTT	12.6	12.6	0.5%	0.6%	0	Homo sapiens BIC noncoding mRNA, complete sequence	89104	H	U
GTACCAGAAAA	12.6	12.6	0.5%	0.6%	0	bridging integrator 2	14770	H	U
GTGTTAAATCG	12.6	12.6	0.5%	0.6%	0	KIAA1373 protein	16229	H	U
CGGCACCTTAA	12.6	12.6	0.5%	0.6%	1	DKFZP434C171 protein	209100	H	U
ACTGGGTGCAG	12.6	12.6	0.5%	0.6%	2	small optic lobes (Drosophila) homolog	55836	H	U
GTTTCAAACGA	12.6	12.6	0.5%	0.6%	2	hypothetical protein MGC10966	180535	H	U
GGGAAGATGAA	12.6	12.6	0.5%	3.0%	3	hypothetical protein MGC2668	56851	H	U
GCTGGAGCGCC	12.6	12.6	0.5%	3.0%	4	Homo sapiens, clone IMAGE:2989556, mRNA, partial cds	12284	H	U
GCTTGCTGGCC	12.6	12.6	0.5%	3.0%	4	HpaII tiny fragments locus 9C	63609	H	U
CCGGCGCGTG	12.6	12.6	0.5%	3.0%	7	CGI-20 protein	107387	H	U
AAACAGTGTAT	12.6	12.6	0.5%	0.6%	0	tumor necrosis factor receptor superfamily, member 14 (herpesvirus entry mediator)	279899	K	CS
ACTGTTTCTCT	12.6	12.6	0.5%	0.6%	0	DNA segment on chromosome 12 (unique) 2489 expressed sequence	74085	K	CS
CTAAAGCCTTC	12.6	12.6	0.5%	0.6%	0	CD3Z antigen, zeta polypeptide (TiT3 complex)	97087	K	CS
GCCTACTTAAA	12.6	12.6	0.5%	0.6%	0	CD5 antigen (p56-62)	58685	K	CS
GCGCTGGTACG	12.6	12.6	0.5%	0.6%	0	2'-5'-oligoadenylate synthetase 3 (100 kD)	56009	K	O
GGTTGTTGGGA	12.6	12.6	0.5%	0.6%	0	chemokine (C-C motif) receptor 5	54443	K	CS

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
TGGGAGCTCAG	12.6	12.6	0.5%	0.6%	0	lymphocyte antigen 117	88411	K	CS
CATACAGAAAA	12.6	12.6	0.5%	3.0%	1	CD97 antigen	3107	K	CS
GCTTCCCCTTG	12.6	12.6	0.5%	0.6%	1	SH2 domain protein 2A	103527	K	S
GTGTGTCTGAA	12.6	12.6	0.5%	0.6%	1	major histocompatibility complex, class II, DR beta 5 / major histocompatibility complex, class II, DR beta 3	308026 / 279930	K	AP
TGATTTTTCTG	12.6	12.6	0.5%	3.0%	2	tripartite motif-containing 22	318501	K	T
CAGTGACTAGA	12.6	12.6	0.5%	0.6%	0			N	N
TCTAGCACAGT	12.6	12.6	0.5%	0.6%	0			N	N
TGGGCTGCTTT	12.6	12.6	0.5%	0.6%	0			N	N
TCCATTAAGCC	12.6	12.6	0.5%	3.0%	1			N	N
ACACAGCAAGA	12.6	12.6	0.5%	0.6%	5			N	N
GATGGGCTGA	12.6	12.6	0.5%	0.6%	5			N	N
CGGAGACCCTA	12.6	12.6	0.5%	0.6%	7			N	N
AAGGAAAGGCC	12.6	12.6	0.5%	0.6%	1	manic fringe (Drosophila) homolog	31939	O	S
GACAAGCCCAG	12.6	12.6	0.5%	0.6%	1	RAB5 interacting protein 3	180040	O	Cy
GAGACCTTTGT	12.6	12.6	0.5%	0.6%	1	DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide 11 (S.cerevisiae CHL1-like helicase)	27424	O	S
GGATCAAGTCC	12.6	12.6	0.5%	3.0%	1	damage-specific DNA binding protein 2 (48kD)	77602	O	O
TATTTTATTTG	12.6	12.6	0.5%	0.6%	1	purinergic receptor (family A group 5)	18999	O	CS
TTTCACAACAG	12.6	12.6	0.5%	0.6%	1	arginase, type II	172851	O	O
CCGCTATCGAA	12.6	12.6	0.5%	0.6%	3	nuclear receptor subfamily 3, group C, member 1	75772	O	T
GGGCTCACCTG	12.6	12.6	0.5%	0.6%	3	DNA replication factor	122908	O	CC
TGTGTGAGCTC	12.6	12.6	0.5%	0.6%	3	FH1/FH2 domain-containing protein	95231	O	T
CGTCCCGGAGC	12.6	12.6	0.5%	0.6%	4	MAD1 (mitotic arrest deficient, yeast, homolog)-like 1	7345	O	CC
AGTGGAGGGAA	12.6	12.6	0.5%	0.6%	5	ataxin 2 related protein	43509	O	Cy
CCTAGGACCTG	12.6	12.6	0.5%	3.0%	7	actin related protein 2/3 complex, subunit 4 (20 kD)	323342	O	Cy
ATTGACCGCTG	12.6	12.6	0.5%	0.6%	9	ADP-ribosyltransferase (NAD+; poly (ADP-ribose) polymerase)	177766	O	T
AAGGCCCAGGA	12.6	12.6	0.5%	0.6%	0	death-associated protein kinase 2 / hypothetical protein FLJ14665	129208 / 334517	xs	xs
AGGACAGAAGG	12.6	12.6	0.5%	3.0%	5	matrix metalloproteinase 25 / ribosomal protein L29	198265 / 183698	xs	xs
GGAGACAGAGT	12.6	6.4	2.5%	16.5%	4	ESTs, Weakly similar to neuronal thread protein [Homo sapiens] [H.sapiens]	355724	E	Est
GACTTCACTTT	12.6	6.4	2.5%	20.4%	11	ESTs	356605	E	Est
TTTCTGTTAAA	12.6	6.4	2.5%	39.1%	3	hypothetical protein	12101	H	U
GTGCTGGTCCC	12.6	6.4	2.5%	8.2%	6	hypothetical protein BC003515	284207	H	U

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
TACATTCTGTG	12.6	6.4	2.5%	16.5%	8	myeloid cell leukemia sequence 1 (BCL2-related)	86386	H	U
TGAAACTCATC	12.6	6.4	2.5%	8.2%	9	Matches cDNA BC000771 (Homo sapiens, Similar to tropomyosin 4, clone MGC:3261 IMAGE:3506357, mRNA, complete cds.)	85844	H	U
TTCCGGTTCCA	12.6	6.4	2.5%	4.3%	12	nucleobindin 1	172609	K	AP
GAGCAGCTGGA	12.6	6.4	2.5%	16.5%	5	copine I	166887	O	Cy
TCTTTACTTGA	12.6	6.4	2.5%	27.1%	7	actin related protein 2/3 complex, subunit 3 (21 kD)	6895	O	Cy
GCCTCTGTCTC	12.6	6.4	2.5%	16.5%	8	ribosomal protein, large, P1	177592	O	P
CGAGGGGCCAG	12.6	6.4	2.5%	16.5%	9	actinin, alpha 4	182485	O	Cy
CTAAAAGGAGA	12.6	6.4	2.5%	8.2%	9	small nuclear ribonucleoprotein polypeptide E	334612	O	P
TACCTGTCTGT	12.6	6.4	2.5%	0.6%	0	ESTs	356517	E	Est
ACTTGCGAATA	12.6	6.4	2.5%	0.6%	6	hypothetical protein FLJ11618	77735	H	U
TGAGTCTGGCT	12.6	6.4	2.5%	3.0%	7	Homo sapiens mRNA; cDNA DKFZp564C2063 (from clone DKFZp564C2063)	4055	H	U
CTGGGAGAGGC	12.6	6.4	2.5%	3.0%	8	arginyl aminopeptidase (aminopeptidase B)-like 1	5345	H	U
GCAGACACTTA	12.6	6.4	2.5%	0.6%	1			N	N
ACCACCTACTT	12.6	6.4	2.5%	0.6%	1	septin 6	90998	O	Cy
AGAAGGCTGCT	12.6	6.4	2.5%	3.0%	6	protein kinase C-like 1	2499	O	S
GATGGGGACA	12.6	6.4	2.5%	3.0%	8	ESTs / DR1-associated protein 1 (negative cofactor 2 alpha) / ESTs	92195 / 295362 / 372541	xs	xs
GTGGTATGGCT	11.1	11.1	0.9%	12.2%	3	hypothetical protein	25199	H	U
CCTGGATAAAT	11.1	11.1	0.9%	14.5%	5	Homo sapiens cDNA: FLJ23602 fis, clone LNG15735	181634	H	U
GTGCCAGCCCT	11.1	11.1	0.9%	12.2%	5	FK506 binding protein precursor	24048	H	U
GATTTTCTACT	11.1	11.1	0.9%	12.2%	6	KIAA0663 gene product	17969	H	U
GCCCCGAGCCC	11.1	11.1	0.9%	29.4%	6	DNA segment, single copy probe LNS-CAI/LNS-CAII (deleted in polyposis	178112	H	U
TATCACTCTGT	11.1	11.1	0.9%	35.2%	6	male-enhanced antigen	278362	H	U
GATTTTGTAGC	11.1	11.1	0.9%	12.2%	8	acidic (leucine-rich) nuclear phosphoprotein 32 family, member B	84264	H	U
GGAAGGGGAGG	11.1	11.1	0.9%	2.6%	5	nuclear factor of kappa light polypeptide gene enhancer in B-cells 2 (p49/p100)	73090	K	T
TTTATTGAATT	11.1	11.1	0.9%	9.7%	5	CD164 antigen, sialomucin	43910	K	CS
CTTTTCAAGAA	11.1	11.1	0.9%	48.1%	7	membrane cofactor protein (CD46, trophoblast-lymphocyte cross-reactive antigen)	83532	K	CS
TTTATTGAAAA	11.1	11.1	0.9%	14.5%	7	CD164 antigen, sialomucin	43910	K	CS
AGCCTGCAGAA	11.1	11.1	0.9%	6.4%	9	IL25 Likely ortholog of mouse interleukin 25	10927	K	E
TTCACAAAGGA	11.1	11.1	0.9%	22.7%	11	proteasome (prosome, macropain) subunit, alpha type, 5	76913	K	AP
CAAGTTCTTTC	11.1	11.1	0.9%	29.4%	9			N	N

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
TCTGGTTTGTC	11.1	11.1	0.9%	1.6%	10			N	N
CCACTCTGGCT	11.1	11.1	0.9%	48.1%	4	glucosidase I	83919	O	O
CTGTATTTGAA	11.1	11.1	0.9%	12.2%	5	transformer-2 alpha (htra-2 alpha)	24937	O	P
GGCGGCTGCAG	11.1	11.1	0.9%	22.7%	5	excision repair cross-complementing rodent repair deficiency, complementation group 1 (includes overlapping antisense sequence)	59544	O	O
TGGAAATAAAA	11.1	11.1	0.9%	12.2%	5	mitochondrial ribosomal protein S6	6945	O	P
GAAAATTAACC	11.1	11.1	0.9%	35.2%	7	aspartyl-tRNA synthetase	80758	O	P
GTGTGTAAAAA	11.1	11.1	0.9%	4.1%	7	accessory proteins BAP31/BAP29	291904	O	S
GACCAGGCCCT	11.1	11.1	0.9%	0.0%	8	tropomyosin 2 (beta)	300772	O	Cy
CGTGAAAGATA	11.1	11.1	0.9%	1.1%	0	ESTs	60416	E	Est
TCACGTGGGCA	11.1	11.1	0.9%	1.1%	0	ESTs	271985	E	Est
AACATTTGTGT	11.1	11.1	0.9%	1.1%	1	ESTs	279619	E	Est
GTGAGAACCC	11.1	11.1	0.9%	1.1%	1	ESTs	170498	E	Est
TTCCTTTTACT	11.1	11.1	0.9%	1.1%	1	Matches ESTs from RAD50 (S. cerevisiae) homolog	41587	E	Est
AGCGGCTACAC	11.1	11.1	0.9%	1.1%	0	interferon stimulated gene (20kD)	183487	H	U
AGGCCACTGGG	11.1	11.1	0.9%	1.1%	0	hypothetical protein MGC2463	323634	H	U
AAGAGGCTGA	11.1	11.1	0.9%	1.1%	1	DKFZP434B103 protein	289010	H	U
ACCTGCAGGCA	11.1	11.1	0.9%	1.1%	1	Homo sapiens, Similar to RAB37, member of RAS oncogene family, clone MGC:21391 IMAGE:4520191, mRNA, complete cds	147066	H	U
CTCAGTGAACT	11.1	11.1	0.9%	1.1%	1	hypothetical protein FLJ10803 (tag found in ESTs which are found repetitive region of one cDNA, BC001743)	8173	H	U
GACAGATGGAC	11.1	11.1	0.9%	4.9%	1	KIAA1533 protein	83575	H	U
GGATGCGCAGG	11.1	11.1	0.9%	4.9%	1	Homo sapiens mRNA full length insert cDNA clone EUROIMAGE 50374	302741	H	U
TTACCCAGTGT	11.1	11.1	0.9%	1.1%	1	hypothetical protein FLJ21709	10888	H	U
TTTGGGACCCT	11.1	11.1	0.9%	1.1%	1	pleckstrin homology, Sec7 and coiled/coil domains, binding protein	270	H	U
ACGGGAACCTA	11.1	11.1	0.9%	1.1%	2	hypothetical protein MGC10744	25092	H	U
TTTGGAGCATT	11.1	11.1	0.9%	1.1%	3	Homo sapiens cDNA FLJ31762 fis, clone NT2RI2007754, weakly similar to INTESTINAL MEMBRANE A4 PROTEIN	7773	H	U
ACCTGCCGACA	11.1	11.1	0.9%	1.1%	5	Homo sapiens cDNA FLJ14866 fis, clone PLACE1002066 / tumor suppressor deleted in oral cancer-related 1	288941 / 25664	H	U
GATGACGACTC	11.1	11.1	0.9%	1.1%	5	symplekin; Huntingtin interacting protein I	107019	H	U
TCAGAGATGAG	11.1	11.1	0.9%	1.1%	5	syntaxin binding protein 2	90535	H	U
TGGCCATCTGC	11.1	11.1	0.9%	4.9%	7	PP1201 protein	184052	H	U
AAGGCCGAGTA	11.1	11.1	0.9%	4.9%	9	DKFZP564J0123 protein	31387	H	U

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
AATTTCCAGTG	11.1	11.1	0.9%	1.1%	0	Ksp37 protein	98785	K	S
ATACCACTTTT	11.1	11.1	0.9%	1.1%	0	leupaxin	49587	K	S
CAGATGCAGTC	11.1	11.1	0.9%	1.1%	0	natural killer cell receptor 2B4	157872	K	CS
CCTCTCTCCTT	11.1	11.1	0.9%	1.1%	0	tripartite motif-containing 22	318501	K	T
GGCCAGTGAGG	11.1	11.1	0.9%	4.9%	0	tumor necrosis factor receptor superfamily, member 14 (herpesvirus entry mediator)	279899	K	CS
GTCACCAAACA	11.1	11.1	0.9%	1.1%	0	selectin P ligand	79283	K	CS
GTTTAAAGATG	11.1	11.1	0.9%	1.1%	0	CD3G antigen, gamma polypeptide (TiT3 complex)	2259	K	CS
TACATTTCTGA	11.1	11.1	0.9%	1.1%	0	protein tyrosine phosphatase, receptor type, C	170121	K	CS
TCGTCAAAGCT	11.1	11.1	0.9%	1.1%	0	SH2 domain protein 1A, Duncan's disease (lymphoproliferative syndrome)	151544	K	S
AGCCTGTGCTT	11.1	11.1	0.9%	1.1%	1	leukotriene b4 receptor (chemokine receptor-like 1)	28408	K	CS
TGAATACTACT	11.1	11.1	0.9%	1.1%	3	LPS-induced TNF-alpha factor	76507	K	T
GGGCCCTGGCC	11.1	11.1	0.9%	1.1%	4	NADPH oxidase-related, C2 domain-containing protein	25895	K	S
ATTCCTGAGCG	11.1	11.1	0.9%	1.1%	0			N	N
CATTTTCTACT	11.1	11.1	0.9%	1.1%	0			N	N
CTTGACCTGTG	11.1	11.1	0.9%	1.1%	0			N	N
GAGAAGCACAG	11.1	11.1	0.9%	1.1%	0			N	N
GAGGCCTGGGT	11.1	11.1	0.9%	1.1%	0			N	N
GCCCGAGGAAG	11.1	11.1	0.9%	1.1%	0			N	N
TGGGTAACAGT	11.1	11.1	0.9%	1.1%	0			N	N
AAAAGACAAAT	11.1	11.1	0.9%	4.9%	1			N	N
TGAACCCGGGA	11.1	11.1	0.9%	1.1%	1			N	N
AAGAAAAAGCC	11.1	11.1	0.9%	1.1%	6			N	N
ATGGAAAGGAA	11.1	11.1	0.9%	4.9%	8			N	N
GTGAGCCCAT	11.1	11.1	0.9%	1.1%	8			N	N
AGGAGATGGAG	11.1	11.1	0.9%	4.9%	0	CDC-like kinase 3 / Homo sapiens, clone MGC:16360 IMAGE:3927645, mRNA, complete cds	73987 / 339840	O	P
CAACCCACGCT	11.1	11.1	0.9%	1.1%	0	histone deacetylase 10	26593	O	T
GAGGCCGCCTG	11.1	11.1	0.9%	1.1%	0	ATP-binding cassette, sub-family A (ABC1), member 7	134514	O	O
GGTTGATTCTG	11.1	11.1	0.9%	1.1%	0	phosphodiesterase 4D, cAMP-specific (dunce (Drosophila)-homolog phosphodiesterase E3)	172081	O	S
TGTATGGCTAA	11.1	11.1	0.9%	1.1%	0	calpain 3, (p94)	40300	O	O
ACAAGATATTT	11.1	11.1	0.9%	4.9%	1	caspase 4, apoptosis-related cysteine protease	74122	O	CC
CGCTCTCTTTC	11.1	11.1	0.9%	1.1%	1	lymphoid blast crisis oncogene	301946	O	S

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
CGGTGGATTT	11.1	11.1	0.9%	1.1%	1	excision repair cross-complementing rodent repair deficiency, complementation group 5 (xeroderma pigmentosum, complementation group G (Cockayne syndrome))	48576	O	O
CTGCGGCTGTA	11.1	11.1	0.9%	1.1%	1	chondroitin 4-O-sulfotransferase 2	25204	O	O
CTGTACATACT	11.1	11.1	0.9%	1.1%	1	NS1-binding protein	197298	O	P
CTTTCCTTTTC	11.1	11.1	0.9%	1.1%	1	uncoupling protein 2 (mitochondrial, proton carrier)	80658	O	O
GACAATGTATG	11.1	11.1	0.9%	1.1%	1	guanine nucleotide binding protein (G protein), gamma 2	289026	O	S
TGTTTCACACC	11.1	11.1	0.9%	1.1%	1	synaptotagmin-like 2	92254	O	Cy
TTCTGTAGCCC	11.1	11.1	0.9%	1.1%	1	ATPase, Ca++ transporting, ubiquitous	5541	O	O
GGCAGGCTGTG	11.1	11.1	0.9%	4.9%	2	peptidylprolyl isomerase E (cyclophilin E)	33251	O	O
AATGGGGGTTA	11.1	11.1	0.9%	1.1%	3	RAB35, member RAS oncogene family	94308	O	Cy
AATTGTCCGTA	11.1	11.1	0.9%	1.1%	3	ribonucleotide reductase M2 polypeptide	75319	O	O
GGACCTGCGCC	11.1	11.1	0.9%	4.9%	4	ribonuclease 6 precursor	8297	O	O
GCCCTTCCCC	11.1	11.1	0.9%	1.1%	5	transcriptional adaptor 3 (ADA3, yeast homolog)-like (PCAF histone acetylase complex)	158196	O	T
TGCAGATATTC	11.1	11.1	0.9%	1.1%	7	cyclin-dependent kinase inhibitor 3 (CDK2-associated dual specificity phosphatase)	84113	O	CC
GGGTCTGCTG	11.1	11.1	0.9%	1.1%	1	Homo sapiens, clone MGC:14381 IMAGE:4299817, mRNA, complete cds / ESTs	105280 / 260844	xs	xs
AGTGTGGAATA	11.1	5.6	4.1%	12.2%	3	hypothetical protein FLJ10330	342307	H	U
CAAGGGCCAAG	11.1	5.6	4.1%	12.2%	5	RAB2, member RAS oncogene family-like	170160	K	S
GCTCCGAGCGT	11.1	5.6	4.1%	14.5%	11			N	N
GGGGGGGTCTC	11.1	5.6	4.1%	35.2%	5	protein phosphatase 2, regulatory subunit B (B56), gamma isoform	171734	O	S
CCTAAGGCTAA	11.1	5.6	4.1%	35.2%	6	E2F transcription factor 4, p107/p130-binding	108371	O	T
GGGCCTGTGCC	11.1	5.6	4.1%	0.0%	9	solute carrier family 16 (monocarboxylic acid transporters), member 3	85838	O	O
AAGGTAATGCT	11.1	5.6	4.1%	12.2%	11	nucleolar protein family A, member 2 (H/ACA small nucleolar RNPs)	23990	O	P
ACCTCAGGAAA	11.1	5.6	4.1%	0.0%	11	high density lipoprotein binding protein (vigilin)	177516	O	O
GGCCCTGAGCG	11.1	5.6	4.1%	0.0%	12	polymerase (RNA) II (DNA directed) polypeptide L (7.6kD)	71618	O	T
TCAAAAAAAGA	11.1	5.6	4.1%	22.7%	4	splicing factor 30, survival of motor neuron-related / capping protein (actin filament) muscle Z-line, alpha 1	79968 / 184270	xs	xs
TATGGTACCAA	11.1	5.6	4.1%	1.1%	2	ESTs	332869	E	Est
ACCAGCCAAAG	11.1	5.6	4.1%	1.1%	0	Homo sapiens cDNA FLJ31087 fis, clone IMR321000074	377879	H	U
CTTCAAGGCCG	11.1	5.6	4.1%	1.1%	1	hypothetical protein FLJ22127	59457	H	U
TTTGATTTTAG	11.1	5.6	4.1%	4.9%	1	Homo sapiens cDNA FLJ10590 fis, clone NT2RP2004392, weakly similar to MNN4 PROTEIN	183779	H	U

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
GTGTGTCTCGA	11.1	5.6	4.1%	1.1%	2	hypothetical protein from BCRA2 region	23518	H	U
GTTTCAGTTAC	11.1	5.6	4.1%	1.1%	2	hypothetical protein FLJ20950	285673	H	U
ACGTGAGTGCT	11.1	5.6	4.1%	4.9%	4	CGI-105 protein	279932	H	U
AAGCCCCTTCC	11.1	5.6	4.1%	1.1%	5	Hypothetical protein BC017335	292570	H	U
GCCCAGCCCTG	11.1	5.6	4.1%	1.1%	5	hypothetical protein 384D8_6	180903	H	U
GGGCAGAATTG	11.1	5.6	4.1%	1.1%	6	KIAA0370 protein	70500	H	U
GCTCAGGATGA	11.1	5.6	4.1%	1.1%	0			N	N
TGATACTTTTG	11.1	5.6	4.1%	1.1%	4			N	N
AAGGATGTAGA	11.1	5.6	4.1%	1.1%	0	Golgi-associated, gamma-adaptin ear containing, ARF-binding protein 2	155546	O	Cy
GGGAATGATGA	11.1	5.6	4.1%	1.1%	0	myeloid/lymphoid or mixed-lineage leukemia (trithorax (Drosophila) homolog)	199160	O	T
GTATTTTTAAA	11.1	5.6	4.1%	1.1%	0	Rho GTPase activating protein 9	19807	O	S
TACCAAGAAAA	11.1	5.6	4.1%	4.9%	2	tubulin-specific chaperone d	12570	O	Cy
TGTAAAGATTT	11.1	5.6	4.1%	1.1%	4	cyclin L ania-6a	4859	O	CC
AAAGCAGTTT	11.1	5.6	4.1%	4.9%	5	apolipoprotein L, 1 / apolipoprotein L, 2	114309 / 241412	O	O
GCCCGCAAGCT	11.1	5.6	4.1%	1.1%	6	bromodomain-containing 4	278675	O	CC
GATGGTCAGTC	11.1	5.6	4.1%	4.9%	7	component of oligomeric golgi complex 4	108779	O	P
TGTGACCTCTC	11.1	5.6	4.1%	1.1%	8	dolichyl-phosphate mannosyltransferase polypeptide 2, regulatory subunit	108973	O	P
CCTGTGACAGC	11.1	5.6	4.1%	4.9%	10	anti-oxidant protein 2 (non-selenium glutathione peroxidase, acidic calcium-independent phospholipase A2)	120	O	O
TGGTGACAGTT	11.1	5.6	4.1%	4.9%	11	histone H2A.F/Z variant	301005	O	O
TGTCAGAAAA	11.1	5.6	4.1%	1.1%	2	phosphoserine phosphatase-like / ESTs	76845 / 354609	xs	xs
CTTAAATCAG	9.5	9.5	1.6%	7.7%	0	Homo sapiens cDNA FLJ13634 fis, clone PLACE1011133, NB matches ESTs but not cDNA in cluster	279607	E	Est
TGACTTTGAAA	9.5	9.5	1.6%	17.7%	1	butyrophilin, subfamily 3, member A2, matches ESTs but not cDNA	87497	E	Est
CAAAATGCAAA	9.5	9.5	1.6%	6.1%	3	cathepsin C matches ESTs from this cluster	10029	E	Est
TTGTAAATGCG	9.5	9.5	1.6%	44.7%	9	kinesin family member 5B matches ESTs in this cluster	149436	E	Est
GATCTCCTGCG	9.5	9.5	1.6%	7.7%	1	hypothetical protein FLJ14972 / hypothetical protein FLJ20406	11900 / 149227	H	U
TTGGAGCAAAG	9.5	9.5	1.6%	7.7%	1	Homo sapiens, clone MGC:32104 IMAGE:4873910, mRNA, complete cds	147025	H	U
ACTTCACAAAG	9.5	9.5	1.6%	7.7%	2	hypothetical protein FLJ22405	27556	H	U
CCAATGGACA	9.5	9.5	1.6%	7.7%	2	hypothetical protein FLJ20559	98135	H	U
GCCTCCAGCC	9.5	9.5	1.6%	7.7%	2	hypothetical protein FLJ12592	23100	H	U

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
AGGATGGGTGC	9.5	9.5	1.6%	7.7%	3	Homo sapiens mRNA; cDNA DKFZp586N1323 (from clone DKFZp586N1323)	24064	H	U
CCTGCACACT	9.5	9.5	1.6%	7.7%	3	SPPL2b	284161	H	U
CGGTCCCATTG	9.5	9.5	1.6%	17.7%	3	hypothetical protein MGC11134	326586	H	U
CTCCGCAGCTG	9.5	9.5	1.6%	17.7%	3	Homo sapiens, clone MGC:2299 IMAGE:2967519, mRNA, complete cds	306981	H	U
GCTGCCAAAAG	9.5	9.5	1.6%	7.7%	3	Homo sapiens cDNA FLJ11840 fis, clone HEMBA1006639	251946	H	U
GCCCTGACCTT	9.5	9.5	1.6%	7.7%	5	Homo sapiens, clone IMAGE:3845253, mRNA, partial cds	97871	H	U
GCTGGCAGGCC	9.5	9.5	1.6%	30.6%	5	choline kinase-like	154886	H	U
TCACAAAAGAG	9.5	9.5	1.6%	42.1%	5	hypothetical protein FLJ22693	12646	H	U
TCCACTACCA	9.5	9.5	1.6%	21.5%	5	CGI-116 protein	18885	H	U
TGGTGAGGGGA	9.5	9.5	1.6%	17.7%	6	hypothetical protein	83530	H	U
ATCAACTGGAG	9.5	9.5	1.6%	44.7%	7	nucleobindin 2	3164	H	U
GACTCTGGGAT	9.5	9.5	1.6%	21.5%	8	cisplatin resistance related protein CRR9p	323769	H	U
GAGGGCCGGTG	9.5	9.5	1.6%	42.1%	8	hypothetical protein FLJ10903, (matches BC003602 contig cointaining Homo sapiens, Similar to H2A histone family, member O)	36727	H	U
GTGCCTAGGGA	9.5	9.5	1.6%	9.6%	9	angiotensin II, type I receptor-associated protein	12854	H	U
TGGTCTGGAGG	9.5	9.5	1.6%	7.7%	0	TGFB1-induced anti-apoptotic factor 1	75822	K	S
GACCTGGTGCC	9.5	9.5	1.6%	7.7%	3	c-src tyrosine kinase	77793	K	S
TGCGCTGGCCC	9.5	9.5	1.6%	17.7%	9	latent transforming growth factor beta binding protein 3	289019	K	E
CCTTACTTTAT	9.5	9.5	1.6%	7.7%	3			N	N
GCAGGCATCA	9.5	9.5	1.6%	7.7%	3			N	N
TCCCATTAGC	9.5	9.5	1.6%	7.7%	3			N	N
CCTTGCTTTTA	9.5	9.5	1.6%	44.7%	4			N	N
TCCCGTACTC	9.5	9.5	1.6%	7.7%	5			N	N
AGTGCACGTGC	9.5	9.5	1.6%	42.1%	6			N	N
CCTTTGGCTAG	9.5	9.5	1.6%	2.3%	8			N	N
GTCCTGGAGGT	9.5	9.5	1.6%	7.7%	9			N	N
TGAAGTGTATA	9.5	9.5	1.6%	17.7%	0	O-linked N-acetylglucosamine (GlcNAc) transferase (UDP-N-acetylglucosamine:polypeptide-N-acetylglucosaminyl transferase)	100293	O	P
AAGAATCAAAA	9.5	9.5	1.6%	21.5%	3	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 1 (7kD, MNLL)	183435	O	O
GGCAGGCACAA	9.5	9.5	1.6%	7.7%	3	emopamil-binding protein (sterol isomerase)	75105	O	O
TCACTGTGGG	9.5	9.5	1.6%	7.7%	3	spectrin, alpha, non-erythrocytic 1 (alpha-fodrin)	77196	O	Cy
GATTACCTGTG	9.5	9.5	1.6%	44.7%	4	hexosaminidase A (alpha polypeptide)	119403	O	O

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
GGGCTGAACAC	9.5	9.5	1.6%	7.7%	5	U4/U6-associated RNA splicing factor	11776	O	P
ATGTAGAGTGT	9.5	9.5	1.6%	7.7%	6	thymidylate synthetase	82962	O	O
CAGACTATGTT	9.5	9.5	1.6%	17.7%	7	ADP-ribosylation factor 6	89474	O	Cy
GGCGGCCTGG	9.5	9.5	1.6%	7.7%	7	mitochondrial ribosomal protein S11	111286	O	P
TGTTAGCCTG	9.5	9.5	1.6%	30.7%	7	vitamin A responsive; cytoskeleton related	92384	O	Cy
CCCCCTCCGGG	9.5	9.5	1.6%	21.5%	8	small nuclear ribonucleoprotein polypeptides B and B1	83753	O	P
CTCAGCCTGA	9.5	9.5	1.6%	30.6%	9	U7 snRNP-specific Sm-like protein LSM10	3496	O	P
GTGTCCCTGTT	9.5	9.5	1.6%	7.7%	10	ubiquitin-conjugating enzyme E2G 2 (homologous to yeast UBC7)	192853	O	O
CTCAACAGCAA	9.5	9.5	1.6%	14.5%	12	eukaryotic translation initiation factor 3, subunit 5 (epsilon, 47kD)	7811	O	P
GGGCTGGGCCC	9.5	9.5	1.6%	1.4%	12	6-phosphogluconolactonase	100071	O	O
ATGTTTGCCC	9.5	9.5	1.6%	44.7%	5	nuclear transcription factor Y, beta / RAD54 (S.cerevisiae)-like	84928 / 66718	xs	xs
AATATTTTCTC	9.5	9.5	1.6%	1.9%	0	ESTs in Homo sapiens cDNA FLJ35429 fis, clone SMINT2002126	377850	E	Est
ATGCTAAGCTA	9.5	9.5	1.6%	1.9%	0	ESTs	233071	E	Est
GCGAACTCCGT	9.5	9.5	1.6%	1.9%	0	ESTs, Weakly similar to ALU1_HUMAN ALU SUBFAMILY J SEQUENCE CONTAMINATION WARNING ENTRY [H.sapiens]	248844	E	Est
TGTTCTCCCCC	9.5	9.5	1.6%	1.9%	0	ESTs	126630	E	Est
AAACTGTCAG	9.5	9.5	1.6%	1.9%	1	ESTs, Weakly similar to JC1405 6-pyruvoyltetrahydropterin synthase [H.sapiens]	14204	E	Est
GCAGTCCCAGG	9.5	9.5	1.6%	1.9%	4	ESTs, Weakly similar to T28919 hypothetical protein C13F10.7	356259	E	Est
TGGCAAGATGA	9.5	9.5	1.6%	1.9%	5	polymyositis/scleroderma autoantigen 1 (75kD), matches ESTs that do not match main contig	91728	E	Est
GCCTCTTCCCC	9.5	9.5	1.6%	1.9%	7	ESTs, Weakly similar to CYL1_HUMAN CYLICIN I [H.sapiens]	294142	E	Est
AGCAGGCTCTG	9.5	9.5	1.6%	1.9%	0	Homo sapiens cDNA FLJ25887 fis, clone CBR02996	355961	H	U
AGCTTTGAAGT	9.5	9.5	1.6%	1.9%	0	Homo sapiens cDNA FLJ35637 fis, clone SPLEN2012115	111377	H	U
TATAAAATATG	9.5	9.5	1.6%	1.9%	0	Homo sapiens, LOC205472, clone MGC:40441 IMAGE:4385178, mRNA, complete cds	49614	H	U
CAGAGCTGTGC	9.5	9.5	1.6%	1.9%	0	phorbolin-like protein MDS019	250619	H	U
GGGTTTTAAA	9.5	9.5	1.6%	1.9%	0	retinal degeneration B beta	333212	H	U
GGTCTTGCAGT	9.5	9.5	1.6%	1.9%	0	Homo sapiens mRNA for FLJ00180 protein	128357	H	U
CTTCTTTCCAA	9.5	9.5	1.6%	1.9%	1	FN5 protein	259737	H	U
GTGATGCACTT	9.5	9.5	1.6%	1.9%	1	hypothetical protein FLJ10803	8173	H	U
AGAGCTAATTT	9.5	9.5	1.6%	1.9%	3	Homo sapiens mRNA; cDNA DKFZp566P1124 (from clone DKFZp566P1124)	321022	H	U

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
ATTTAGTCATA	9.5	9.5	1.6%	1.9%	3	interferon-induced protein 44	82316	H	U
TGACATCTGAT	9.5	9.5	1.6%	1.9%	3	hypothetical protein MGC5306	301732	H	U
AAGGTGGAGA	9.5	9.5	1.6%	1.9%	4	chromosome 20 open reading frame 34	306044	H	U
TGCCCTTCGG	9.5	9.5	1.6%	1.9%	4	KIAA0551 protein / Homo sapiens, Similar to zinc finger protein 296, clone MGC:13093 IMAGE:3942704, mRNA, complete cds	170204 / 192237	H	U
AAGACTGGCTT	9.5	9.5	1.6%	1.9%	6	surfeit 4	284296	H	U
GCCACGTTGTC	9.5	9.5	1.6%	1.9%	6	hypothetical protein DKFZp434K1210	32352	H	U
GGCCTCTGAGC	9.5	9.5	1.6%	1.9%	9	KIAA1096 protein	69559	H	U
AAAAATCGGG	9.5	9.5	1.6%	1.9%	0	small inducible cytokine A5 (RANTES)	241392	K	E
AGCTGACGGTG	9.5	9.5	1.6%	1.9%	0	neutrophil cytosolic factor 4 (40kD)	196352	K	S
GCCAGTGATTT	9.5	9.5	1.6%	1.9%	0	2'-5'-oligoadenylate synthetase 2 (69-71 kD)	264981	K	O
TTTTTTATTCC	9.5	9.5	1.6%	1.9%	0	interleukin 10 receptor, alpha	327	K	CS
AGACAGCAGGA	9.5	9.5	1.6%	1.9%	0			N	N
CAAGCAGAAA	9.5	9.5	1.6%	1.9%	0			N	N
CAAGGGCTTAG	9.5	9.5	1.6%	1.9%	0			N	N
CAGAACTTGA	9.5	9.5	1.6%	1.9%	0			N	N
GACCTTTTCGA	9.5	9.5	1.6%	1.9%	0			N	N
GCAAAACGCAG	9.5	9.5	1.6%	1.9%	0			N	N
GCAAGCCAGCG	9.5	9.5	1.6%	1.9%	0			N	N
GGTGTTTCTTT	9.5	9.5	1.6%	1.9%	0			N	N
TCCCCTCATC	9.5	9.5	1.6%	1.9%	0			N	N
TGATGGATCAG	9.5	9.5	1.6%	1.9%	0			N	N
TTCTCAGGTTC	9.5	9.5	1.6%	1.9%	0			N	N
GATTCAACAAA	9.5	9.5	1.6%	1.9%	1			N	N
TCTATTAAGC	9.5	9.5	1.6%	1.9%	1			N	N
ACTCCAGTCAA	9.5	9.5	1.6%	1.9%	2			N	N
GGGGCGGGGGG	9.5	9.5	1.6%	1.9%	2			N	N
GGCCCCCTAAG	9.5	9.5	1.6%	1.9%	3			N	N
GCGGCGCTGCT	9.5	9.5	1.6%	1.9%	5			N	N
ATAAATGCAGA	9.5	9.5	1.6%	1.9%	9			N	N
AAAGTTATCCA	9.5	9.5	1.6%	1.9%	0	regulator of G-protein signalling 1	75256	O	S
CCTCCGAAAAT	9.5	9.5	1.6%	1.9%	0	G protein-coupled receptor 15	159900	O	CS
GCTTTGCAGCG	9.5	9.5	1.6%	1.9%	0	Src-like-adapter	75367	O	S
TTGCAAACCTC	9.5	9.5	1.6%	1.9%	0	phospholipase C, gamma 2 (phosphatidylinositol-specific)	75648	O	S
TTGTAATTTTG	9.5	9.5	1.6%	1.9%	0	death-associated protein kinase 2	129208	O	S
GCTCAAAGATT	9.5	9.5	1.6%	1.9%	1	WD40 protein Ciao1	12109	O	T

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
CGCCTCAGAGC	9.5	9.5	1.6%	1.9%	2	golgin-67	182982	O	P
CTACCCGGTAT	9.5	9.5	1.6%	1.9%	2	endothelial differentiation, lysophosphatidic acid G-protein-coupled receptor, 4	122575	O	CS
TGGCCCGACGA	9.5	9.5	1.6%	1.9%	2	nudix (nucleoside diphosphate linked moiety X)-type motif 1	388	O	O
AAAGCTGGCTC	9.5	9.5	1.6%	1.9%	3	deoxycytidine kinase	709	O	O
AGGCCCACAAG	9.5	9.5	1.6%	1.9%	3	mannosidase, alpha, class 2C, member 1	26232	O	P
GGAGCAGACGC	9.5	9.5	1.6%	1.9%	3	Transient receptor potential cation channel, subfamily V, member 2	279746	O	CS
TGTAGTATTT	9.5	9.5	1.6%	1.9%	3	protein kinase C and casein kinase substrate in neurons 2	18842	O	S
AAGATCAAGAT	9.5	9.5	1.6%	1.9%	4	actin, gamma 1 / actin, alpha 1, skeletal muscle / actin, alpha 2, smooth muscle, aorta	14376 / 1288 / 195851	O	Cy
ACTCAAAAAA	9.5	9.5	1.6%	1.9%	4	deoxycytidine kinase	709	O	O
CGCCGCTTCTT	9.5	9.5	1.6%	1.9%	4	bromodomain adjacent to zinc finger domain, 2A	277401	O	T
TGTCCGTCAC	9.5	9.5	1.6%	1.9%	4	E74-like factor 1 (ets domain transcription factor)	154365	O	T
TGGACTTTGTG	9.5	9.5	1.6%	1.9%	5	ubiquitin-activating enzyme E1-like	16695	O	O
TTCCCCTTCCT	9.5	9.5	1.6%	1.9%	5	signal recognition particle receptor ('docking protein')	75730	O	P
GCTAAGAGGGA	9.5	9.5	1.6%	1.9%	1	hypothetical protein BC010734 / EST	48821 / 290674	xs	xs
TTGTAAAAAA	9.5	9.5	1.6%	1.9%	4	taxol resistance associated gene 3 / pinin, desmosome associated protein / ATP-binding cassette, sub-family E (OABP), member 1 / S164 protein	251377 / 44499 / 12013 / 180789	xs	xs
CTTATGTAGAT	7.9	7.9	2.9%	25.3%	2	ESTs from tumor protein p53-binding protein	179982	E	Est
TCGTTACGCAG	7.9	7.9	2.9%	25.3%	4	ESTs	356036	E	Est
ATTAAGAAAAT	7.9	7.9	2.9%	44.7%	6	AHNAK nucleoprotein (desmoyokin), matches ESTs not cDNA	301417	E	Est
AGTCAAGCCCC	7.9	7.9	2.9%	12.1%	2	four and a half LIM domains 3	57687	H	U
CTGCTAACCC	7.9	7.9	2.9%	12.1%	4	cat eye syndrome chromosome region, candidate 1	170310	H	U
ATGGGCTTGAT	7.9	7.9	2.9%	44.7%	5	CGI-44 protein; sulfide dehydrogenase like (yeast)	8185	H	U
CGTGTGCCTG	7.9	7.9	2.9%	25.3%	5	hypothetical protein MGC4342	301342	H	U
AGCTACCGGG	7.9	7.9	2.9%	25.3%	6	EGF-containing fibulin-like extracellular matrix protein 2	6059	H	U
GCCGGGCACGG	7.9	7.9	2.9%	12.1%	6	hypothetical protein FLJ20686	29032	H	U
GGAGCCAGCT	7.9	7.9	2.9%	12.1%	6	hypothetical protein DKFZp762A227	274453	H	U
TGTGTGCCAC	7.9	7.9	2.9%	12.1%	6	chromosome 11 open reading frame 13	72925	H	U
AGTTTAAGCAT	7.9	7.9	2.9%	5.8%	7	phosphatidylinositol binding clathrin assembly protein	7885	H	U
CTGAGGGCCGG	7.9	7.9	2.9%	12.1%	8	WD repeat domain 13	12142	H	U
TCCTAGCCTGT	7.9	7.9	2.9%	44.7%	9	DnaJ (Hsp40) homolog, subfamily C, member 8	74711	H	U
GCTTTACTTTG	7.9	7.9	2.9%	44.7%	10	integral membrane protein 1	287850	H	U
GCCTTGATGAT	7.9	7.9	2.9%	3.5%	3	CD47 antigen (Rh-related antigen, integrin-associated signal transducer)	82685	K	CS

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
GTATTCTCCAG	7.9	7.9	2.9%	12.1%	3	integrin, alpha 2 (CD49B, alpha 2 subunit of VLA-2 receptor)	271986	K	CS
GTTGGGTAGA	7.9	7.9	2.9%	12.1%	3	putative protein tyrosine kinase	282990	K	S
TGTGAACACA	7.9	7.9	2.9%	12.1%	4	interferon regulatory factor 1	80645	K	T
AAGATTGGGGT	7.9	7.9	2.9%	21.9%	8	CD44 antigen (homing function and Indian blood group system) / Homo sapiens CD44 isoform RC (CD44) mRNA, complete cds	169610 / 306278	K	CS
CGCCGGTTCTG	7.9	7.9	2.9%	12.1%	0			N	N
CCCTATTAGC	7.9	7.9	2.9%	12.1%	1			N	N
CGCAAACGGTA	7.9	7.9	2.9%	25.3%	1			N	N
GCACTTACAA	7.9	7.9	2.9%	12.1%	4			N	N
TCCCCTACAC	7.9	7.9	2.9%	12.1%	5			N	N
CTTCTGGGGAC	7.9	7.9	2.9%	12.1%	6			N	N
AAACTCGGGTC	7.9	7.9	2.9%	12.1%	7			N	N
AATGAGAAGGT	7.9	7.9	2.9%	44.7%	7			N	N
GAGGGGAAACG	7.9	7.9	2.9%	25.3%	7			N	N
GTTGATTTTA	7.9	7.9	2.9%	12.1%	7			N	N
GGGGGAATTTT	7.9	7.9	2.9%	12.1%	8			N	N
GCGGAGAGAGG	7.9	7.9	2.9%	14.5%	10			N	N
AGCTGTTTCTT	7.9	7.9	2.9%	12.1%	12			N	N
GCCTGCAGGGC	7.9	7.9	2.9%	12.1%	1	histone deacetylase 10	26593	O	T
TGGCTGTTAAT	7.9	7.9	2.9%	12.1%	1	septin 6	90998	O	Cy
ACTTAAGTACC	7.9	7.9	2.9%	12.1%	3	protein phosphatase 1, regulatory (inhibitor) subunit 2	267819	O	S
CACGGAGGCGG	7.9	7.9	2.9%	12.1%	4	xeroderma pigmentosum, complementation group C	320	O	O
ACAGCCAAGA	7.9	7.9	2.9%	12.1%	5	nudix (nucleoside diphosphate linked moiety X)-type motif 2	14142	O	O
GAGAGAACGGA	7.9	7.9	2.9%	12.1%	5	retinoblastoma-binding protein 2	76272	O	T
TGTTCAGAAAA	7.9	7.9	2.9%	25.3%	5	vaccinia related kinase 1	48269	O	S
AATGAGCAACT	7.9	7.9	2.9%	40.5%	6	guanylate binding protein 2, interferon-inducible	171862	O	S
CTCCCCTGCC	7.9	7.9	2.9%	0.4%	6	capping protein (actin filament), gelsolin-like	82422	O	Cy
GAAATCCGCAC	7.9	7.9	2.9%	12.1%	7	mannosidase, alpha, class 2B, member 1	279854	O	O
GGGAAGTCACC	7.9	7.9	2.9%	12.1%	7	tissue specific transplantation antigen P35B	264428	O	P
GACTCAGGGAT	7.9	7.9	2.9%	40.5%	8	GTP binding protein 2	13011	O	S
GCGTGATCCTG	7.9	7.9	2.9%	40.5%	8	aldo-keto reductase family 1, member A1 (aldehyde reductase)	89529	O	O
TTTTGCTACAG	7.9	7.9	2.9%	25.3%	8	HIV-1 Rev binding protein	171545	O	P
CCGGGCCCAGC	7.9	7.9	2.9%	44.7%	9	thyroid hormone receptor interactor 6	119498	O	S
CTCTGATGCAG	7.9	7.9	2.9%	25.3%	9	polymerase (DNA directed), gamma	80961	O	CC
GCCTGGTGAC	7.9	7.9	2.9%	32.0%	9	death-associated protein 6	336916	O	S

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
TCTTCTCCCT	7.9	7.9	2.9%	12.1%	10	hepatoma-derived growth factor (high-mobility group protein 1-like)	89525	O	CC
AGATCCTACTT	7.9	7.9	2.9%	25.3%	11	farnesyl-diphosphate farnesyltransferase 1	48876	O	O
GCCGCCATCTC	7.9	7.9	2.9%	25.3%	11	transketolase (Wernicke-Korsakoff syndrome)	89643	O	O
CAACATTCCTG	7.9	7.9	2.9%	0.0%	12	D-dopachrome tautomerase	180015	O	O
CTGTTAATAAA	7.9	7.9	2.9%	44.7%	6	nuclear localization signal deleted in velocardiofacial syndrome / ESTs, Highly similar to DDP_HUMAN X-LINKED DEAFNESS DYSTONIA PROTEIN [H.sapiens]	19500 / 355429	xs	xs
AAATAAAAGCT	7.9	7.9	2.9%	0.4%	9	villin 2 (ezrin) / thyroid stimulating hormone receptor	155191 / 123078	xs	xs
CACCCCCAGGC	7.9	7.9	2.9%	25.3%	10	G protein pathway suppressor 2 / putative mitochondrial outer membrane protein import receptor	7301 / 31334	xs	xs
GAATGCGAAGA	7.9	7.9	2.9%	3.4%	0	ESTs	92448	E	Est
GACTGTTAAAA	7.9	7.9	2.9%	3.4%	2	ESTs	208669	E	Est
GCTGGTCCCAG	7.9	7.9	2.9%	3.4%	3	ESTs, Weakly similar to T31613 hypothetical protein Y50E8A.i - Caenorhabditis elegans [C.elegans]	296234	E	Est
ATCTCCTTTTA	7.9	7.9	2.9%	3.4%	0	hypothetical protein FLJ12788	20242	H	U
GGGCGCCTGGC	7.9	7.9	2.9%	3.4%	0	immunity associated protein 1	159955	H	U
TTGTACATTTC	7.9	7.9	2.9%	3.4%	0	KIAA0449 protein	169182	H	U
CAGGTTAAGCT	7.9	7.9	2.9%	3.4%	1	Janus kinase 3 (a protein tyrosine kinase, leukocyte), tag matches mRNA which is similar to Janus kinase 3	99877	H	U
CCTTTCTTTA	7.9	7.9	2.9%	3.4%	1	Homo sapiens, Similar to hypothetical protein, MGC:7036, clone MGC:4797 IMAGE:3544761, mRNA, complete cds	50535	H	U
GGCCTCTCCGA	7.9	7.9	2.9%	3.4%	1	hematopoietic protein 1	132834	H	U
GTCTGATATCT	7.9	7.9	2.9%	3.4%	1	hypothetical protein FLJ11712	14920	H	U
TGTCAAAAGAG	7.9	7.9	2.9%	3.4%	1	KIAA0874 protein	27973	H	U
TTCTCAAGAAA	7.9	7.9	2.9%	3.4%	1	HRAS like suppressor 3	37189	H	U
AACAGACACAA	7.9	7.9	2.9%	3.4%	2	hypothetical protein FLJ10631	238944	H	U
CAGGGGTTGGG	7.9	7.9	2.9%	3.4%	2	hypothetical protein FLJ00012	21051	H	U
GATCCGCTGTC	7.9	7.9	2.9%	3.4%	2	Homo sapiens, clone IMAGE:3161564, mRNA, partial cds	7133	H	U
GCGGATACTGT	7.9	7.9	2.9%	3.4%	2	hypothetical protein MGC2650	61273	H	U
TCTCTACTCTG	7.9	7.9	2.9%	3.4%	2	cell division cycle associated 1	234545	H	U
TTGTAAGAGGG	7.9	7.9	2.9%	3.4%	2	hypothetical protein BC009231	65907	H	U
ACCACCCTGTT	7.9	7.9	2.9%	3.4%	3	KIAA0892 protein	112751	H	U
GGCAGATTGCT	7.9	7.9	2.9%	3.4%	3	chromosome 6 open reading frame 35	173259	H	U
GGCCAGACCTG	7.9	7.9	2.9%	3.4%	3	hypothetical protein FLJ20637	179669	H	U
TCTCTGGTTTC	7.9	7.9	2.9%	3.4%	3	KIAA0675 gene product	165662	H	U
ACTACCTCCCC	7.9	7.9	2.9%	3.4%	4	Mov10 (Moloney leukemia virus 10, mouse) homolog	20725	H	U

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
CAGGGCTCGCG	7.9	7.9	2.9%	3.4%	4	hypothetical protein FLJ21865	29288	H	U
TTTCAAAGATA	7.9	7.9	2.9%	3.4%	4	KIAA0073 protein	1191	H	U
CTGAGGAACAA	7.9	7.9	2.9%	3.4%	5	VRK3 for vaccinia related kinase 3	98289	H	U
CTGTCTGTGGC	7.9	7.9	2.9%	3.4%	5	hypothetical protein FLJ20748	91973	H	U
TAGCAGCTGGG	7.9	7.9	2.9%	3.4%	5	hypothetical protein MGC14439	107001	H	U
TATCTGCTGAA	7.9	7.9	2.9%	3.4%	5	hypothetical protein similar to actin related protein 2/3 complex, subunit 5	315164	H	U
TTGTACAACAG	7.9	7.9	2.9%	3.4%	5	C21orf19 like protein	20814	H	U
CCTGGCAGTTG	7.9	7.9	2.9%	3.4%	7	hypothetical protein GL009	24054	H	U
ACTAAGTGCT	7.9	7.9	2.9%	3.4%	0	HIC: I-mfa domain-containing protein, no unigene cluster but ESTs from this cluster match the HIC cDNA.	132739	K	T
AGTACCCTCAT	7.9	7.9	2.9%	3.4%	0	2',5'-oligoadenylate synthetase 1 (40-46 kD)	82396	K	O
ATGGAATGCTA	7.9	7.9	2.9%	3.4%	0	receptor-interacting serine-threonine kinase 3	268551	K	S
CATTTGCACTA	7.9	7.9	2.9%	3.4%	0	CD69 antigen (p60, early T-cell activation antigen)	82401	K	CS
TAAAACCTGCT	7.9	7.9	2.9%	3.4%	0	tumor necrosis factor receptor superfamily, member 6	82359	K	CS
TACCATTTAGC	7.9	7.9	2.9%	3.4%	0	killer cell lectin-like receptor subfamily C, member 3	258850	K	CS
AGTGGAAATT	7.9	7.9	2.9%	3.4%	2	interferon (alpha, beta and omega) receptor 2	86958	K	CS
GCAGTGGGAAA	7.9	7.9	2.9%	3.4%	3	lymphotoxin beta (TNF superfamily, member 3)	890	K	CS
GTATCCAGCTC	7.9	7.9	2.9%	3.4%	4	colony stimulating factor 1 (macrophage)	173894	K	E
GATTGGCGGCT	7.9	7.9	2.9%	3.4%	5	BAF53	274350	K	T
AAAAATCGGT	7.9	7.9	2.9%	3.4%	0			N	N
AAGAGAAACC	7.9	7.9	2.9%	3.4%	0			N	N
AGAAGGTATCA	7.9	7.9	2.9%	3.4%	0			N	N
ATCGCGGAGG	7.9	7.9	2.9%	3.4%	0			N	N
ATGAGCCTCG	7.9	7.9	2.9%	3.4%	0			N	N
CAGCGGGTAAG	7.9	7.9	2.9%	3.4%	0			N	N
CAGTGAAGCTC	7.9	7.9	2.9%	3.4%	0			N	N
CCCCTACATCG	7.9	7.9	2.9%	3.4%	0			N	N
CGCCCCCAAA	7.9	7.9	2.9%	3.4%	0			N	N
GAATTTTGTC	7.9	7.9	2.9%	3.4%	0			N	N
GCAACCCCGCC	7.9	7.9	2.9%	3.4%	0			N	N
GCCACGTATTA	7.9	7.9	2.9%	3.4%	0			N	N
GGGATAAGCTC	7.9	7.9	2.9%	3.4%	0			N	N
TACCGCTCGGC	7.9	7.9	2.9%	3.4%	0			N	N
TGGCTCAAGCC	7.9	7.9	2.9%	3.4%	0			N	N
TGTGGTGCTGC	7.9	7.9	2.9%	3.4%	0			N	N
CCCTTGCATTG	7.9	7.9	2.9%	3.4%	1			N	N

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
CCGAGGAAGG	7.9	7.9	2.9%	3.4%	1			N	N
GATTCAAAAAA	7.9	7.9	2.9%	3.4%	1			N	N
GGGAAGATTGT	7.9	7.9	2.9%	3.4%	1			N	N
TCACATCGCT	7.9	7.9	2.9%	3.4%	1			N	N
TCTTATTAAG	7.9	7.9	2.9%	3.4%	1			N	N
TTGGCCCCAAA	7.9	7.9	2.9%	3.4%	2			N	N
GGGAGCGGGTT	7.9	7.9	2.9%	3.4%	3			N	N
TCCCCTAATCG	7.9	7.9	2.9%	3.4%	3			N	N
CCAAGCGGCCT	7.9	7.9	2.9%	3.4%	4			N	N
TTTGGACAATA	7.9	7.9	2.9%	3.4%	4			N	N
AAGAGCTAATG	7.9	7.9	2.9%	3.4%	5			N	N
AGGAAAGGATG	7.9	7.9	2.9%	3.4%	5			N	N
CCTGGGCACT	7.9	7.9	2.9%	3.4%	5			N	N
AGGGTGAACGT	7.9	7.9	2.9%	3.4%	6			N	N
GAGGCGAGGCC	7.9	7.9	2.9%	3.4%	7			N	N
GCGGGTGTGG	7.9	7.9	2.9%	3.4%	9			N	N
CAATATTACAG	7.9	7.9	2.9%	3.4%	0	Homo sapiens guanylate binding protein 5 mRNA, complete cds	237809	O	S
GAACCATTTGC	7.9	7.9	2.9%	3.4%	0	death effector filament-forming Ced-4-like apoptosis protein	104305	O	CC
GAAGGGAAGAA	7.9	7.9	2.9%	3.4%	0	ras homolog gene family, member H	109918	O	S
GCCAGAGGCCC	7.9	7.9	2.9%	3.4%	0	troponin T3, skeletal, fast	73454	O	Cy
TATCTAAACTG	7.9	7.9	2.9%	3.4%	0	RAB27A, member RAS oncogene family	50477	O	S
TATGGGCAGG	7.9	7.9	2.9%	3.4%	0	Wiskott-Aldrich syndrome protein interacting protein	24143	O	Cy
CTGAGGTGTGA	7.9	7.9	2.9%	3.4%	1	runt-related transcription factor 3	170019	O	T
GAGGCCTGTGC	7.9	7.9	2.9%	3.4%	1	RAB27A, member RAS oncogene family	50477	O	S
GTTTTAATTTT	7.9	7.9	2.9%	3.4%	1	muscleblind (Drosophila)-like	28578	O	P
GCCCTCATTAA	7.9	7.9	2.9%	3.4%	2	FYN oncogene related to SRC, FGR, YES	169370	O	S
GCTTAAGAATG	7.9	7.9	2.9%	3.4%	2	CDC16 (cell division cycle 16, S. cerevisiae, homolog)	1592	O	CC
AGAGCTCACTA	7.9	7.9	2.9%	3.4%	3	solute carrier family 25 (carnitine/acylcarnitine translocase), member 20	13845	O	O
GCAGACTCACC	7.9	7.9	2.9%	3.4%	3	aminomethyltransferase (glycine cleavage system protein T)	102	O	O
AGTAGTCTGC	7.9	7.9	2.9%	3.4%	4	mitochondrial ribosomal protein 63	182695	O	P
CCAAAATTTGC	7.9	7.9	2.9%	3.4%	4	Cell division cycle 2, G1 to S and G2 to M	334562	O	CC
CTAACCAAAGA	7.9	7.9	2.9%	3.4%	4	Rho-associated, coiled-coil containing protein kinase 1	17820	O	S
AGAAGTACTG	7.9	7.9	2.9%	3.4%	6	ribonucleotide reductase M1 polypeptide	2934	O	CC
GTAGGAAAGCT	7.9	7.9	2.9%	3.4%	7	aminopeptidase puromycin sensitive	293007	O	CC
TTTATTTGGCA	7.9	7.9	2.9%	3.4%	7	lamin B receptor	152931	O	CC

Tag sequence	Clone 32 tag count per 100,000	Ratio CTL vs. Cerebellum	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
			Cerebellum	Ovary Epithelium					
AATTTACTTCC	7.9	7.9	2.9%	3.4%	4	3-phosphoinositide dependent protein kinase-1, matches ESTs in cluster but not Refseq / IQ motif containing GTPase activating protein 1 / ESTs, Weakly similar to hypothetical protein FLJ20489 [Homo sapiens] [H.sapiens]	154729 / 1742 / 356549	xs	xs
TCAAAAACTTG	7.9	7.9	2.9%	3.4%	5	leucine rich repeat containing 5 / cell division cycle 2, G1 to S and G2 to M	44672 / 334562	xs	xs
GCAGGAATTG	7.9	7.9	2.9%	3.4%	9	farnesyl diphosphate synthase (farnesyl pyrophosphate synthetase, dimethylallyltranstransferase, geranyltranstransferase) / ESTs	77393 / 354385	xs	xs

Appendix C.1: CTL-specific tags that match uncharacterised but sequenced transcripts.

CTL-specific tags were defined as those significantly more abundant (p value ≤ 0.05 by the AC test) in the clone 32 library compared to both normal cerebellum (marked cereb.) and ovary epithelium (marked as ovary epith.) libraries and which were not found at similar levels (more than one-third that in the clone 32 library) in more than 1 of 12 cancer SAGE libraries. The Unigene clusters matching these tags are listed, together with a representative GenBank accession number. The results of the bioinformatic analysis are shown, with the main points of interest in bold type. "Notes on transcript match" indicates any problems with the transcript assignment to the SAGE tag. In particular, for cases where it was not obvious which protein sequence ought to be used in the analysis the column is highlighted red; for cases where the transcript had been previously characterised in some way, it is highlighted light blue.

Tag Sequence	Clone 32 tag count per 100K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
TCCTCTTTCC	123.3	123.3	0.0%	0.0%	0	natural killer cell transcript 4	943	NM_004221	Further investigation shows that other researchers have characterised this transcript to some extent	**This transcript is selectively expressed in lymphocytes and is up regulated on T or NK cell activation. It has a signal peptide and publications suggest it contains an RGD sequence**	
ACCATTGGAT	102.7	102.7	0.0%	0.0%	0	interferon induced transmembrane protein 1 (9-27)	146360	NM_003641	Further investigation shows that other researchers have characterised this transcript to some extent	One of three members of a family of transmebrane (TM) proteins that are **upregulated by interferons**, with unknown function. This particular protein is 17kDa and has been **implicated in relaying antiproliferative and homotypic adhesion signals**	**CS**
CACCCAATGG	64.8	64.8	0.0%	0.0%	0	SEC7 homolog	110121	NM_012455		All domain detection programs identify a **Sec7-like guanine-nucleotide exchange factor (GEF) domain** and a **pleckstrin homology (PH) domain**. BLASTp shows that a large region of this protein is **similar to several GEFs of the cytohesin family** that regulate the ARF-family G-proteins	**S**

Tag Sequence	Clone 32 tag count per 100K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
GCGGTTGTGG	64.8	64.8	0.0%	0.0%	0	Lysosomal-associated multispanning membrane protein-5	79356	NM_006762	Previously identified as a novel lysosomal associated transmembrane protein that is preferentially expressed in haematopoietic cells. Functionally uncharacterised	BLASTp matches to several other lysosomal proteins. SMART detects 5 TM domains. High probability match to a **golgi 4TM spanning transporter** in InterProScan and NCBI conserved domain (CD) search	**P**

Tag Sequence	Clone 32 tag count per 100K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
TCCGCAAGGT	39.5	39.5	0.0%	0.0%	0	Homo sapiens mRNA; cDNA DKFZp667F1219 (from clone DKFZp667F1219)	37617	AL832302	No amino acid sequence identified, used ORF in frame +2, 245 - 2956	All domain prediction programs detect an N terminal MYSc Myosin domain (an ATPase molecular motor). SMART finds this domain, followed by an IQ domain (short calmodulin-binding motif containing conserved Ile and Gln residues) while InterProScan finds a central Bipartite nuclear localisation signal (although this profile has a very high false positive rate). BLASTp search finds this protein to be identical to "Similar to myosin Ic; Unconventional myosin from rat 4 for myosin I heavy chain" [Homo sapiens] (XM_166579) and "Unconventional myosin 1G valine form" [Homo sapiens] (AF380932) and other uncharacterised myosins	**Cy**
ACTTCATAGC	37.9	37.9	0.0%	0.0%	0	Homo sapiens, similar to hypothetical protein FLJ11110, clone MGC:27027 IMAGE:4837773, mRNA, complete cds	124675	BC027613		CD search detects an N terminal match to the C terminus of an MMR_HSR1 domain (GTPase of unknown function). BLASTp search matches other uncharacterised sequences (including mouse "immune associated nucleotide 4 like 1" (NM_018384) 44% identity (91/204) total length 300.) SMART detects C terminal coiled coil and an N terminal AAA (ATPase) domain	

Tag Sequence	Clone 32 tag count per 100K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
TAAGGACGAG	33.2	33.2	0.0%	0.0%	0	hypothetical protein FLJ22457	238707	NM_024901		BLASTp shows this transcript is similar to several uncharacterised sequences and unclassified fly proteins (e.g. sbf). All domain detection programs detect a DENN (AEX-3) domain. This domain has unknown function but is found in both a MAP-kinase activating death domain protein and a GDP/GTP exchange factor (GEF)	**S**
GACTTGGCCT	30.0	30.0	0.0%	0.0%	0	Homo sapiens cDNA FLJ33028 fis, clone THYMU2000140	16291	AK057590		SMART predicts **two TM domains**	**CS**
TGCAAGAGAG	30.0	30.0	0.0%	0.0%	0	Homo sapiens mRNA; cDNA DKFZp667K0625 (from clone DKFZp667K0625)	238954	AL832852	No amino acid sequence identified, used ORF in frame +1, 226-3087	All domain detection programs detect an N terminal **RhoGAP domain** (GTPase-activator protein for Rho-like GTPases). BLASTp detects C terminal of protein is **identical to "Similar to retinitis pigmentosa GTPase regulator"** [Homo sapiens] (XM_170910, an protein predicted from the genome, supported by EST alignments) and that the N terminus is similar to the N terminal region of "Cdc42 GTPase-activating protein" [Mus musculus] (NM_020260)	**S**

Tag Sequence	Clone 32 tag count per 100K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
CTCCTCCAAG	26.9	26.9	0.0%	0.0%	0	Homo sapiens, clone MGC:40121 IMAGE:5216355, mRNA, complete cds	15284	BC028076		BLASTp matches one other uncharacterised sequence. SMART detects 3 TM domains	**CS**
GGAGCTTGAG	26.9	26.9	0.0%	0.0%	0	chromosome 6 open reading frame 9	288316	NM_022107		BLASTp shows this protein is identical to G18 (human) and NG1 (mouse) - both have uncharacterised functions. The transcript contains a proline-rich domain and three GoLoco domains. GoLoco domains contain an LGN motif and are found in putative GEFs specific for Gα GTPase and regulators of G-protein signalling	**S**
GGTAGAACTA	26.9	26.9	0.0%	0.0%	0	chromosome X open reading frame 9	61469	NM_018990	SH3 and SAM domains previously identified. This transcript is encoded on chromosome X in close proximity to genes involved in various immune disorders	CD search and SMART identify an SH3 domain and all 3 detection programs find a SAM domain. BLASTp search detects matches to many other proteins (mainly uncharacterised) that contain SAM or SH3 domains.	**S**
GGACCAGCGC	25.3	25.3	0.0%	0.0%	0	hypothetical protein FLJ21438	136979	AK025091		BLASTp matches to other uncharacterised sequences. SMART predicts a coiled-coil region at the C-terminus	

Tag Sequence	Clone 32 tag count per 100K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
GGAGCTGTCT	25.3	25.3	0.0%	0.0%	0	differentially expressed in FDCP (mouse homolog) 6	15476	NM_022047		BLASTp shows this protein is **similar to SWAP70** (70% identity to N terminal region and 44% identity to C-terminal region), a protein that functions in B-cell isotype switching. All domain detection programs identify a **PH domain**, an N-terminal **EF hand** and **4 tropomyosin domains** suggesting a role in regulating the cytoskeleton. InterProScan also detects two domains that are similar to the **Tubby N-terminal domain**. The function of Tubby is unknown but its disruption can lead to obesity in mice	**Cy**
AGTGATGTAA	22.1	22.1	0.0%	0.0%	0	Homo sapiens clone CDABP0095 mRNA sequence	46919	AY007155	No amino acid sequence identified, used ORF in frame +3, 228-401	BLASTp shows the central region of this protein is identical to the C-terminal half of a protein predicted from the genome (XM_168849)	
TGGAAGCATC	20.5	20.5	0.0%	0.0%	0	HSPC022 protein	367740	NM_014029		BLASTp matches to other uncharacterised sequences. SMART predicts an N-terminal signal peptide	

Tag Sequence	Clone 32 tag count per 100K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
AGCAAGAAAC	19.0	19.0	0.0%	0.1%	0	chromosome 3 open reading frame 4	107393	NM_019895		BLASTp shows this transcript is weakly similar to mouse/human Claudin 19. CD search and SMART detect a **weak match to a PMP22/Claudin domain**. SMART also predicts an **N-terminal signal peptide and 3 TM domains** or 4 TM domains (without the signal peptide)	**CS**
TCAGTGACCA	17.4	17.4	0.1%	0.1%	0	hypothetical protein MGC5363	1880	NM_024064		BLASTp matches other uncharacterised sequences. No conserved domains detected	
GGCGGGGCCA	17.4	17.4	0.1%	0.1%	1	KIAA0303 protein	54985	AB002301 (long) AND BC003646 (short)	There are two predicted amino acid sequences for this transcript, one long and one short	(1) BLASTp shows this protein is similar to mouse "microtubule associated testis specific serine/threonine protein kinase" (56% identity) and "syntrophin associated S/T kinase" (52% identity). All domain identification programs identify a protein kinase domain at N-terminus (could be S/T or Y specific), a protein kinase extension domain and a PDZ domain (type of protein interaction domain, involved in signalling). (2) BLASTp matches only uncharacterised sequences. SMART predicts an N-terminal signal peptide	**S**

Tag Sequence	Clone 32 tag count per 100K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
GGTAGCCCAC	17.4	17.4	0.1%	2.4%	1	trinucleotide repeat containing 5	56828	BC008961		BLASTp shows this transcript is weakly similar to pryruvate phosphate dikinase from Rickettsia conorii (43% similar over a short region). SMART predicts an **N-terminal signal peptide**, and InterProScan identifies an ATP/GTP binding site motif A (**P-loop**)	
TACGAGGCCG	17.4	17.4	0.1%	0.1%	1	expressed in activated T/LAK lymphocytes	16165	NM_007267		SMART detects 7 TM domains. BLASTp search detects matches to uncharacterised sequences, including weak matches to the family (in both mice and humans): TM, cochlear expressed	**CS**
TGGGGCCGCA	17.4	17.4	0.1%	0.1%	1	Homo sapiens cDNA: FLJ23270 fis, clone COL10309, highly similar to HSU33271 Human normal keratinocyte mRNA	288455	AK026923	No amino acid sequence identified, used ORF in frame +3, 33-2381	SMART detects 7 TM domains. BLASTp search detects matches to uncharacterised sequences.	**CS**

Tag Sequence	Clone 32 tag count per 100K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
CATTTACTCT	31.6	16.1	0.0%	0.0%	0	integral membrane protein 2A	17109	NM_004867	The chromosomal location and gene structure of this transcript have previously been characterised. It was cloned as a marker for chondro-osteogenic differentiation and also from CD34+ haematopoietic stem/progenitor cells	BLASTp shows this transcript is similar to the BRI gene (also called integral membrane protein 2B) that is disrupted in familial British dementia. SMART predicts **a single TM domain**	**CS**
GGTCTTCTAG	14.2	14.2	0.3%	1.9%	1	hypothetical protein FLJ10330	342307	NM_018061		CD search detects a PRP38 family match. (Members of this family are related to the pre mRNA splicing factor PRP38 from yeast. This conserved region could be involved in RNA binding) BLASTp search matches only uncharacterised sequences. SMART detects a central coiled coil region	**P**

Tag Sequence	Clone 32 tag count per 100K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
AGGCTCCGTG	26.9	13.7	0.0%	0.0%	0	minor histocompatibility antigen HA-1	196914	D86976	Previously described as an antigen (see cluster name) but functionally uncharacterised.	This transcript contains a **RhoGAP domain**, **a Fes/CIP4/CDC15 domain** and a protein kinase C conserved region 1 (C1) domain (also called a **phorbol ester/diacylglycerol binding domain**)	**S**
AGTCTTTAGT	12.6	12.6	0.5%	0.6%	0	Homo sapiens BIC noncoding mRNA, complete sequence	89104	AF402776	This transcript is uncharacterised although its name suggests it may be a non-coding RNA. No amino acid sequence identified, no clear ORFs		
GTACCAGAAA	12.6	12.6	0.5%	0.6%	0	bridging integrator 2	14770	NM_016293	The sequence of this transcript is published and the BAR domain was identified. It has been shown to associate with BIN1 but has no known function	BLASTp shows this transcript has **strong similarity to BIN1** (69% similarity) **and amphiphysin** (70% similarity) over the ~240aa of the predicted **BAR domain**. This domain is detected by all domain detection programs and is found in amphiphysin and clathrin binding protein. Its function is unknown but it is likely to be an adaptor/bridging domain. InterProScan also identifies **several extensin-like proline rich domains** at the C-terminus of the transcript	**Cy**

Tag Sequence	Clone 32 tag count per 100K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
GTGTTAAATC	12.6	12.6	0.5%	0.6%	0	KIAA1373 protein	16229	AB037794	The correct start codon for this transcript has not been identified, it could therefore be incomplete	CD search detects overlapping **Mov 34** (found in proteasome regulatory subunits, eukaryotic initiation factor 3 subunits and regulators of transcription factors) **and JAB/MPN domains** (domain in Jun kinase activation domain binding protein and proteasomal subunits. This domain is of unknown function.) BLASTp search detects similarity to "Associated molecule with the SH3 domain of STAM" [Homo sapiens] (NM_006463, this protein interacts with the SH3 domain of STAM, plays a critical role in cytokine-mediated intracellular signal transduction and acts downstream of the Jak2/Jak3 STAM complex.) SMART detects a central **coiled coil region** followed by a JAB/MPN domain	**S T**
CGGCACCTTA	12.6	12.6	0.5%	0.6%	1	DKFZP434C171 protein	209100	AL080169		BLASTp gives matches to other uncharacterised sequences. InterProScan detects a bipartite nuclear localization signal, but this profile has a very high false positive rate	

Tag Sequence	Clone 32 tag count per 100K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
AGCGGCTACA	11.1	11.1	0.9%	1.1%	0	interferon stimulated gene (20kD)	183487	NM_002201		All domain detection programs detect an EXOIII domain (exonuclease domain in DNA-polymerase α and ε chain, ribonuclease T and other exonucleases). BLASTp search finds matches to many uncharacterised sequences	**CC**
AGGCCACTGG	11.1	11.1	0.9%	1.1%	0	hypothetical protein MGC2463	323634	NM_024070		SMART predicts an **N-terminal signal peptide** and a **single TM domain**	**CS**
AAGAGGCTGA	11.1	11.1	0.9%	1.1%	1	DKFZP434B103 protein	289010	NM_015644		BLASTp shows this transcript is **similar to Tubulin-tyrosine ligase** (TTL), which catalyses the post-translational addition of a tyrosine to the C-terminus of detyrosinated alpha tubulin (49% similarity over ~220aa). InterProScan and CD search also identify a **TTL family domain**	**Cy P**
ACCTGCAGGC	11.1	11.1	0.9%	1.1%	1	Homo sapiens, Similar to RAB37, member of RAS oncogene family, clone MGC:21391 IMAGE:4520191, mRNA, complete cds	147066	BC016615		CD search detects overlapping RAB/RAS/RHO domains, SMART detects a **RAB domain**. BLASTp search finds it is **almost identical to "RAB37**, member of RAS oncogene family; GTPase Rab37" [Mus musculus] (NM_021411)	**S**

Tag Sequence	Clone 32 tag count per 100K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
CTCAGTGAAC	11.1	11.1	0.9%	1.1%	1	hypothetical protein FLJ10803 (tag found in ESTs which are found repetitive region of one cDNA, BC001743)	8173	BC001743	This tag is produced when cDNA priming occurs at an internal polyA stretch. The true tag for the full cDNA is GTGATGCACTT	BLASTp matrches to other uncharacterised sequences. SMART predicts either a signal peptide or a TM domain at the N-terminus	
GACAGATGGA	11.1	11.1	0.9%	4.9%	1	KIAA1533 protein	83575	BC014077		BLASTp matches to other uncharacterised sequences. SMART and CD search detect an N-terminal GRAM domain (domain in glucosyltransferases, myotubularins and other putative membrane-associated proteins). SMART predicts a single TM domain near the C-terminus	**CS**
GGATGCGCAG	11.1	11.1	0.9%	4.9%	1	Homo sapiens mRNA full length insert cDNA clone EUROIMAGE 50374	302741	AL133429	No amino acid sequence identified, used ORF in frame +3, 909-1109	SMART predicts an **N-terminal signal peptide**	

Tag Sequence	Clone 32 tag count per 100K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
TTACCCAGTG	11.1	11.1	0.9%	1.1%	1	hypothetical protein FLJ21709	10888	NM_032206		BLASTp shows that this transcript is **similar to several known proteins, all with high leucine content**: (1) Delerium A, a leucine rich repeat molecule required for PKA expression in Dictyostelium (42% similarity) (2) Ran GTPase activating protein from the plant Medicago sativa (44% similarity) and (3) Pig hepatic ribonuclease inhibitor (42% similarity). InterProScan detects 1**2 leucine rich repeats (LLRs)** in this transcript and SMART detects 14, with all but two identified as the **ribonuclease inhibitor subtype**	

Tag Sequence	Clone 32 tag count per 100K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
TTTGGGACCC	11.1	11.1	0.9%	1.1%	1	pleckstrin homology, Sec7 and coiled/coil domains, binding protein	270	NM_004288	This sequence has been published. **2 leucine zippers** and a **nuclear localisation signal** have been identified. It has been **shown to interact with GEFs and expression in resting NK and T cells has been noted**	BLASTp shows this transcript is **similar to mouse GRASP** (GRP1-associated scaffold protein, 54% similarity). All domain detection programs identified a **PDZ signalling domain**	**S**
GTGCCACCAG	19.0	9.6	0.3%	0.1%	1	KIAA1554 protein	17767	AB046774	The correct start codon for this transcript has not been identified, it could therefore be incomplete. This transcript also contains several polyA sites so would be expected to produce several different SAGE tags	BLASTp matches only uncharacterised sequences. All domain detection programs find a RING finger domain (Zn fingers involved in protein-protein interactions) near the N-terminus	

Tag Sequence	Clone 32 tag count per 100K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
CAGAGCTGTG	9.5	9.5	1.6%	1.9%	0	phorbolin-like protein MDS019	250619	NM_021822		BLASTp shows this transcript is most **similar to Phorbolin-3** (68% similarity). It is also similar to Phorbolin 1 and 2. Phorbolins are upregulated in psioratic keratinocytes; they have no known function but are **thought to be involved in post-transcriptional modification of RNA**, possibly analogous to that of apobec-1 which deaminates cytosines in certain mRNAs (leaving uracil). InterProScan detects a **cytidine and deoxycytidylate deaminase zinc-binding region** (this region is found in other other-wise unrelated proteins as well as these enzymes)	**P**
GGGTTTTAAA	9.5	9.5	1.6%	1.9%	0	retinal degeneration B beta	333212	NM_012417	Previously identified as containing an amino-terminal PITP-like domain and a short carboxyl-terminal domain, but function remains unknown	CD search and SMART detect a Phosphatidylinositol transfer protein domain. BLASTp search finds matches to several PITP proteins and other uncharacterised sequences	

Tag Sequence	Clone 32 tag count per 100K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
GGTCTTGCAG	9.5	9.5	1.6%	1.9%	0	Homo sapiens mRNA for FLJ00180 protein	128357	AK074109	No amino acid sequence identified, used ORF in frame +1, 61-1500	CD search finds multiple matches to C terminal domain of Tropomodulin. (Tropomodulin is a novel tropomyosin regulatory protein that binds to the end of erythrocyte tropomyosin and blocks head-to-tail association of tropomyosin along actin filaments) and one match to Leucine rich repeat, ribonuclease inhibitor type. BLASTp search detects matches to many other protein (several uncharacterised) that contain LRRs. InterProScan detects an Aldo/keto reductase family putative active site signature. SMART detects N terminal signal peptide followed by 14 Leucine Rich Repeats	**Cy**
CTTCTTTCCA	9.5	9.5	1.6%	1.9%	1	FN5 protein	259737	NM_020179	Short sequence (59 amino acids)	BLASTp matches only uncharacterised sequences. No conserved domains detected. SMART predicts a single TM domain near the C-terminus	**CS**

Tag Sequence	Clone 32 tag count per 100K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
GTGATGCACT	9.5	9.5	1.6%	1.9%	1	hypothetical protein FLJ10803	8173	BC001743	This transcript may not be a true match for this tag. Only one cDNA contains the tag, it has the same coding sequence as other cDNAs but has a very different 3' UTR. This may represent a rare splice variant or an incorrectly spliced cDNA	SMART predicts an **N-terminal signal peptide**	
AGCAGGCTCT	9.5	9.5	1.6%	1.9%	0	Homo sapiens cDNA FLJ25887 fis, clone CBR02996	355961	AK095686	No amino acid sequence identified, used ORF in frame +2, 1196-1480		
TATAAAATAT	9.5	9.5	1.6%	1.9%	0	Homo sapiens, LOC205472, clone MGC:40441 IMAGE:4385178, mRNA, complete cds	49614	BC030506		BLASTp matches to other uncharacterised sequences	

Tag Sequence	Clone 32 tag count per 100K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
AGCTTTGAAG	9.5	9.5	1.6%	1.9%	0	Homo sapiens cDNA FLJ35637 fis, clone SPLEN2012115	111377	AK092956	No amino acid sequence identified, used ORF in frame +3, 3-389 (no start codon so may be incomplete)	InterProScan detects a weak match (only 11bp long) to an **EGF-like domain** which overlaps with another weak match to **thiolase domain**	
GAGAATCAGA	15.8	8.0	0.9%	0.2%	0	Homo sapiens cDNA FLJ25887 fis, clone CBR02996	355961	AK095686	No amino acid sequence identified, used ORF in frame +2, 1196-1480		
GATGAAAAGG	15.8	8.0	0.9%	0.2%	1	Homo sapiens, clone MGC:19482 IMAGE:4309314, mRNA, complete cds / DnaJ (Hsp40) homolog, subfamily C, member 4, two clusters, treat as DnaJ homolog….	343473 / 172847	NM_005528	This tag matches two unigene clusters. They appear to be alternative splice variants of the same gene	BLASTp shows that a section of this transcript is identical to a section of DNAJ/Hsp40 but the flanking regions are unrelated. All domain identification programs identify a **DNAJ domain** (DNAJ proteins are involved in the Hsp70 heat-shock system, associating via this domain - thus they may be chaperone proteins). SMART also predicts a **single TM domain** C-terminal of the DNAJ domain	**CS**
ATCTCCTTTT	7.9	7.9	2.9%	3.4%	0	hypothetical protein FLJ12788	20242	NM_022492		BLASTp matches to other uncharacterised sequences. InterProScan detects a bipartite nuclear localization signal, but this profile has a very high false positive rate. SMART also predicts a coiled-coil region in the sequence	

Tag Sequence	Clone 32 tag count per 100K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
GGGCGCCTGG	7.9	7.9	2.9%	3.4%	0	immunity associated protein 1 (IMAP1)	159955	NM_130759		CD search finds a weak match at N terminus to the central region of a MMR_HSR1 domain (GTPase of unknown function). BLASTp search detects matches to other IMAP proteins and many uncharacterised sequences. SMART detects a C terminal single TM domain	**CS**
TTGTACATTT	7.9	7.9	2.9%	3.4%	0	KIAA0449 protein	169182	NM_017596		BLASTp matches to other uncharacterised sequences. InterProScan and SMART detect a leucine rich repeat N-terminal domain (LLRNT). This repeat is a recently characterised structural motif used in molecular recognition processes as diverse as signal transduction, cell adhesion, cell development, DNA repair and RNA processing. All proteins containing these repeats are thought to be involved in protein-protein interactions	

Tag Sequence	Clone 32 tag count per 100K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
CAGGTTAAGC	7.9	7.9	2.9%	3.4%	1	Janus kinase 3 (a protein tyrosine kinase, leukocyte), tag matches mRNA which is similar to Janus kinase 3	99877	BC028068		SMART and CD search detect a **Band 4.1 domain** (this domain is found in a number of cytoskeletal-associated proteins that associate with various proteins at the interface between the plasma membrane and the cytoskeleton) followed by a **SH2 domain**, followed by a match to one quarter of a **Tyrosine kinase domain**. BLASTp finds the transcript to be **identical to the N termini of JAK3 and several other kinases**	**S**
CCTTTCTTTA	7.9	7.9	2.9%	3.4%	1	Homo sapiens, Similar to hypothetical protein, MGC:7036, clone MGC:4797 IMAGE:3544761, mRNA, complete cds	50535	NM_145058		SMART detects 2 coiled coil regions. BLASTp detects matches to many other uncharacterised sequences	

Tag Sequence	Clone 32 tag count per 100K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
GGCCTCTCCG	7.9	7.9	2.9%	3.4%	1	hematopoietic protein 1	132834	NM_005337	Further investigation shows that other researchers have characterised this transcript to some extent	This transcript is **annotated as a member of the HEM family of tissue-specific TM proteins** which are highly conserved from invertebrates through mammals. This gene is **only expressed in haematopoietic cells**. However, note that **no TM regions were predicted by SMART**. BLASTp shows that the transcript is **similar to NCK-associated protein 1** (HEM-2, NM_013436, 58% identical), a protein **thought to be involved in resistance to apoptosis**	**CC**

Tag Sequence	Clone 32 tag count per 100K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
GTCTGATATC	7.9	7.9	2.9%	3.4%	1	hypothetical protein FLJ11712	14920	NM_024570*	This tag is probably generated by priming of cDNA synthesis from an internal polyA stretch in this gene. One apparently complete cDNA ends at this polyA stretch but may not be a true transcript. Thus the full sequence of the other cDNAs in the cluster were used	BLASTp matches to other uncharacterised sequences. SMART predicts an N-terminal signal peptide	
TGTCAAAAGA	7.9	7.9	2.9%	3.4%	1	KIAA0874 protein	27973	NM_015208*	This tag is probably generated by priming of cDNA synthesis from an internal polyA stretch in this gene. It is a long protein - 2062aa	BLASTp shows the N-terminus of the transcript is **similar to a region of rat BRCA1-associated RING domain protein 1** (NM_022622, 44% identical). All domain detection programs identify **3 Ankyrin repeats** (these probably function in protein-protein interactions), in the region homologous to this rat protein. SMART predicts **several coiled-coil regions** throughout the remainder of the protein	

Tag Sequence	Clone 32 tag count per 100K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
TTCTCAAGAA	7.9	7.9	2.9%	3.4%	1	HRAS like suppressor 3	37189	NM_007069		BLASTp shows that this transcript is 74% identical to rat p18H-rev 107 (X76453) . CD search identifies an **NLP/P60 family domain** (function unknown) but with weak homology. SMART predicts a **single C-terminal TM domain**	**CS**
ATAAACAGAT	28.4	7.2	0.1%	0.1%	0	Homo sapiens cDNA FLJ14752 fis, clone NT2RP3003071 / SR rich protein	334825 / 18368	1) AK027658 / 2) NM_032870	Tag matches 2 unigene clusters, containing highly similar sequences	1) BLASTp matches other uncharacterised sequences. SMART predicts a short coiled-coil region and a single TM domain 2) BLASTp matches other uncharacterised sequences. SMART predicts many coiled coil regions and InterProScan predicts an N terminal Proline rich domain and C terminal contains a Bipartite nuclear localization signal (although this profile has a very high false positive rate)	**CS**
TTGATGCCCT	14.2	7.2	1.5%	0.4%	1	hypothetical protein FLJ10081	7871	AY050169	This is a large protein - 791 aa	BLASTp shows this protein is identical to several uncharacterised sequences with different names such as testis development protein PRTD (AF311326) and serum inhibited-related protein (AY050169)	
AGAATAAAGC	25.3	6.4	0.2%	0.0%	0	hypothetical protein DKFZp434G0920	98564	NM_032251		BLASTp matches to other uncharacterised sequences	

Tag Sequence	Clone 32 tag count per 100K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
TGGCCTGCCC	47.4	6.0	0.0%	0.0%	1	MLL septin-like fusion	181002	NM_006640	Publication suggest that this protein is **related to the septin family** of GTPase genes involved in cytokinesis.	Both SMART and CD search detect a C terminal **GTP_CDC domain** (Found in cell division protein. Members of this family include CDC3, CDC10, CDC11 and CDC12/Septin. Members of this family bind GTP.) InterProScan detects an ATP/GTP-binding site motif A (**P-loop**). BLASTp search shows this protein is almost **identical to other members of the septin (and septin-like) family**. C terminal similar (47% identity) to cell division control protein CDC10 - yeast (Candida albicans)	**CC**
ACCAGCCAAA	11.1	5.6	4.1%	1.1%	0	Homo sapiens cDNA FLJ31087 fis, clone IMR32100007 4	377879	AK055649	No amino acid sequence identified, used ORF in frame +3, 1215-1616	SMART detects an **N terminal signal peptide**	

Tag Sequence	Clone 32 tag count per 100K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
CTTCAAGGCC	11.1	5.6	4.1%	1.1%	1	hypothetical protein FLJ22127	59457	NM_022775		All domain detection programs identify an N terminal **WW Domain** (has 2 conserved Trp residues, also known as the WWP or rsp5 domain. Binds proline-rich polypeptides) Both InterProScan and SMART detect at least one **Double-stranded RNA binding domain**. (The DsRBD domain is found in a variety of RNA-binding proteins with different structures and exhibiting a diversity of functions. It is involved in localization of at least five different mRNAs in the early Drosophila embryo and by interferon-induced protein kinase in humans, which is part of the cellular response to dsRNA.)	
TTTGATTTTA	11.1	5.6	4.1%	4.9%	1	Homo sapiens cDNA FLJ10590 fis, clone NT2RP2004392, weakly similar to MNN4 PROTEIN	183779	AK001452		BLASTp shows that this protein is similar to 3 segments of Dictyostelium discoideum development protein DG1124 (AF111943). Disruption of this gene results in a morphological defect. All domain detection programs identify a **SAP domain** (a putative DNA-binding (bihelical) motif predicted to be involved in chromosomal organisation). SMART also predicts **2 C-terminal coiled-coil regions**	**O T**

Tag Sequence	Clone 32 tag count per 100K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
ATGACAGATG	31.6	5.4	0.1%	0.0%	0	lung cancer-associated Y protein	13775	NM_032495	This transcript encodes a short protein - only 73 aa	BLASTp shows this transcript is **similar to Goosecoid and Pax6** from several species - these are homeobox containing proteins. All domain detection programs identify a **single homeobox domain** in this transcript	**T**

Tag Sequence	Clone 32 tag count per 100K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
CGGATAAGGC	30.0	5.1	0.1%	0.2%	1	nuclear prelamin A recognition factor	256526	NM_012336	Although functionally uncharacterised, much is known about this protein. NARF binds to the prenylated c-terminal tail of prelamin A. It may therefore be a component of a prelamin A endoprotease complex. It is located in the nucleus, where it partially colocalizes with the nuclear lamina. Transcript variants encoding different isoforms and/or utilizing alternative polyadenylation sites exist	BLASTp shows that this transcript is **similar to several bacterial hydrogenases** and all domain detection programs identify domains in the transcript similar to both the **large and small subunits of iron-only hydrogenases**. These enzymes are usually involved in reduction reactions in bacterial electron transport chains	**O**

Tag Sequence	Clone 32 tag count per 100K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
CACCACGGTG	58.5	3.0	0.1%	0.0%	1	RNB6	241471	NM_016337	Previously published homology to murine EVL, a member of **Ena/VASP protein family** that is implicated to be involved in the control of cell motility through actin filament assembly by their GP5 motifs.	CD search detect an N terminal **WH1 domain**, this overlaps with a possible **Ran-binding domain** (domain of approximately 150 amino acids that stabilises the GTP-bound form of Ran (a Ras-like nuclear small GTPase)). BLASTp search detects high similarity to "Ena-vasodilator stimulated phosphoprotein" [Mus musculus] (NM_007965), this protein seems to be involved in actin rearrangements. InterProScan detects an N terminal **EVH1 (RanBP1-WASP) domain**, (many EVH1-containing proteins associate with actin-based structures and play a role in cytoskeletal organisation) followed by a **proline rich domain**. SMART detects an N terminal **WASP homology region 1**, function unknown	**Cy**
CACTGTGACC	52.2	2.9	0.1%	0.0%	1	hypothetical protein MGC10500	271599	NM_031477		BLASTp shows **similarity to fly Yippee protein** (53% similarity) - a Zn binding protein conserved in eukaryotes	

Tag Sequence	Clone 32 tag count per 100K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
TGGACCCCCC	50.6	2.9	0.2%	0.0%	0	Hypothetical protein MGC:5244	374608	NM_031213		SMART predicts an N terminal signal peptide. Both SMART and CD search detect a C terminal Phospholipase/Carboxylesterase domain (this family consists of both phospholipases and carboxylesterases with broad substrate specificity, and is structurally related to alpha/beta hydrolases abhydrolase). Whereas InterProScan detects a central esterase/lipase/thioesterase active site (this profile covers the active site serine of a wide variety of enzymes including esterases, lipases, pepitdases etc. Proteolytic enzymes that exploit serine in their catalytic activity are ubiquitous, being found in viruses, bacteria and eukaryotes. They include a wide range of peptidase activity.) BLASTp search detects matches to uncharacterised sequences only	**O**

Appendix C.2: CTL-specific tags that match Unigene clusters containing only EST Sequences.

CTL-specific tags were defined as those significantly more abundant (p value ≤ 0.05) in the clone 32 library compared to both normal cerebellum (marked as cereb.) and ovary epithelium (marked as ovary epith.) libraries and which were not found at similar levels (more than one-third that in the clone 32 library) in more than 1 of 12 cancer derived SAGE libraries. The EST clusters were assembled using the CAP3 programme. Where possible, full-length transcripts were identified by BLAST, and possible protein sequences were searched for conserved domains or homology to known proteins. The last column indicates possible functions of the transcripts, given the results of this analysis.

Tag sequence	Clone 32 tag count per 100 K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	CAP3 Contig	BLASTx	BLASTn	Further analysis of matches.	Possible function
			Cereb.	Ovary Epith.								
TGGAACCCTG	126.4	126.4	0.00%	0.00%	0	ESTs, Highly similar to hypothetical protein FLJ13725; KIAA1930 protein [Homo sapiens] [H.sapiens]	355944	1	Matches uncharacterised sequences. 5′ end of contig is 39% identical to central region of "CCAAT/enhancer binding protein (C/EBP), alpha; CAAT/enhancer-binding protein, DNA-binding protein" [Rattus norvegicus] (NM_012524). 5′ end is also 28% identical to central region of protein tyrosine phosphatase HD-PTP [Homo sapiens] (AB025194)	3' two-thirds of this contig matches chromosome 17 sequence		

Tag sequence	Clone 32 tag count per 100 K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	CAP3 Contig	BLASTx	BLASTn	Further analysis of matches.	Possible function
			Cereb.	Ovary Epith.								
CGCCCCGGCG	28.4	28.4	0.00%	0.00%	0	ESTs	196244	1		Matches 3' end of "Homo sapiens candidate tumour suppressor **HIC-1 (HIC-1) gene**, complete cds" (99% identity)	HIC-1 contains a **BTB/POZ homomeric dimerisation domain** and **5 Zinc fingers**. POZ domains from several zinc finger proteins have been shown to mediate transcriptional repression and to interact with components of histone deacetylase co-repressor complexes including N-CoR and SMART. BLASTp shows that it is **similar to several transcription factors / repressors**	**T**
TCTCCAAGGA	25.3	25.3	0.00%	0.01%	0	ESTs, Weakly similar to PSF_HUMAN PTB-ASSOCIATED SPLICING FACTOR [H.sapiens]	189284	1	Matches (37%) N terminal end of AF077549 antigen [Clonorchis sinensis]	Matches short section at 5' end of "Homo sapiens mRNA for KIAA1618 protein, partial cds" (AB046838)	Note: no PolyA signal/tail in ESTs, so this cluster may be a false match to the tags	

Tag sequence	Clone 32 tag count per 100 K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	CAP3 Contig	BLASTx	BLASTn	Further analysis of matches.	Possible function
			Cereb.	Ovary Epith.								
GAACCGTCCT	47.4	24.1	0.00%	0.00%	0	ESTs, Weakly similar to T51376 plant adhesion molecule 1 (PAM1) - Arabidopsis thaliana [A.thaliana]	123164	1		Approximately two-thirds of contig matches "**Homo sapiens, clone IMAGE:4849486, mRNA**" (BC022202)	"Homo sapiens, clone IMAGE:4849486, mRNA" has no identified amino acid sequence (used ORF in frame +2, 2 to 1342). All domain detection programs identify a central TBC domain (Probable Rab-GAP; Widespread domain present in Gyp6 and Gyp7, thereby giving rise to the notion that it performs a GTP-activator activity on Rab-like GTPases.) Similar to other uncharacterised sequences and proteins containing TBC domains	**S**

Tag sequence	Clone 32 tag count per 100 K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	CAP3 Contig	BLASTx	BLASTn	Further analysis of matches.	Possible function
			Cereb.	Ovary Epith.								
GCATAATAAG	22.1	22.1	0.01%	0.16%	0	ESTs	11594	1 and 2	1) 5′ end of contig 77% identical to C terminal half of "RIKEN cDNA 4930424G05" [Mus musculus] (NM_026251) 2) Central fifth of contig identical to C terminal half of similar to "RIKEN cDNA 4930424G05" [Homo sapiens] (XM_058682)	1) Identical to 3' three-quarters of "Homo sapiens, clone IMAGE:5227850, mRNA" (BC036924). Matches Chromosome 15 sequence. 3' of contig identical to last four-fifths of "Homo sapiens similar to RIKEN cDNA 4930424G05 (LOC123120), mRNA" (XM_058682) 2) Matches Chromosome 15 sequence. 3' of contig is identical to last quarter of "clone IMAGE:5227850, mRNA, BC036924". 3' identical to last four-fifths of "Homo sapiens similar to RIKEN cDNA 4930424G05 (LOC123120), mRNA" (XM_058682)	Tag found in two contigs in the cluster. 3' half of both contigs is the same, i.e. Unigene cluster contains different splice variants. Homo sapiens, clone IMAGE:5227850, mRNA, BC036924 has no identified amino acid sequence (used ORF in frame +2, 581 to 1645). InterProScan detects ProDom domain PD416875 (DEVELOPMENTAL P100 EGG-SPECIFIC DNA-BINDING POLYMORPHISM 4930424G05RIK). BLAST search detects matches (77%) to C terminal three-fifths of "Data source:SPTR, source key:Q91817, evidence:ISS~putative~related to EGG-SPECIFIC PROTEIN (P100)" [Mus musculus] (AK015195)	

Tag sequence	Clone 32 tag count per 100 K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	CAP3 Contig	BLASTx	BLASTn	Further analysis of matches.	Possible function
			Cereb.	Ovary Epith.								
TGCCCCTGAA	22.1	22.1	0.01%	0.02%	0	ESTs	154993	1	Small regions of contig matches various uncharacterised sequences	Matches various BACs and PACs		

Tag sequence	Clone 32 tag count per 100 K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	CAP3 Contig	BLASTx	BLASTn	Further analysis of matches.	Possible function
			Cereb.	Ovary Epith.								
TCAGTTGCCT	19.0	19.0	0.05%	0.06%	0	ESTs, Weakly similar to activation-induced cytidine deaminase; activation induced cytidine deaminase [Homo sapiens]	236533	1	5′ end of contig matches (97%) C terminal of "**dJ742C19.2 APOBEC1 (Apolipoprotein B mRNA editing protein) and Phorbolin 1 like**" [Homo sapiens] (AL031846, protein ID: CAB36553.1)	Matches Human DNA sequence from clone RP4-742C19 on chromosome 22, which contains a pseudogene similar to Cytochrome C Oxidase Polypeptide VB and (parts of) up to four novel genes, two with homology to Phorbolin genes and one a novel Chromobox protein gene	dJ742C19.2 (APOBEC1 and Phorbolin 1 LIKE) has no defined start codon, notes state "could be part of gene dJ742C19.1 and/or dJ742C19.3". CD search detects **Cytidine and deoxycytidylate deaminase zinc-binding region**. BLAST search finds matches to other homologues of APOBEC1, one of which, NP_055323, is described as a member of the cytidine deaminase gene family. Members of this family encode proteins that are structurally and **functionally related to the C to U RNA-editing cytidine deaminase APOBEC1**. It is thought that the proteins may be RNA editing enzymes and have roles in growth or cell cycle control	**P**

Tag sequence	Clone 32 tag count per 100 K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	CAP3 Contig	BLASTx	BLASTn	Further analysis of matches.	Possible function
			Cereb.	Ovary Epith.								
TACAGCACAA	17.4	17.4	0.08%	0.11%	0	ESTs	260395	1	Centre of contig matches (45%) N terminus of "uncharacterised sequence XP_098831" [Homo sapiens] (XM_098831)	Centre of contig matches (84%) centre of "Homo sapiens cDNA FLJ14121 fis, clone MAMMA1002009" (AK024183). Flanking sequences do not show similarity		
TTAATGCGTC	15.8	15.8	0.15%	0.20%	0	inhibitor of DNA binding 2, dominant negative helix-loop-helix protein (Matches ESTs in this cluster but not the contig containing the Refseq)	180919	2	5′ end of contig matches DNA-binding protein inhibitor ID-2	5' half of contig matches chromosome 2 sequence. It also matches inhibitor of DNA binding 2 as seen described on right	There is a group of ESTs for which 5' end matches the centre of the refseq cDNA but their 3' end does not match, so either the ESTs are from different genes/splice variants or they are chimeric/badly sequenced ESTs	
TGCCAATTAA	15.8	15.8	0.15%	0.20%	1	ESTs	165337	1		Large central region of contig matches 3' region of "**Homo sapiens Ras-like GTP-binding protein (RAB27A) gene**" (98% identity)	RAB27A is a **RAB like GTPase**. All domain detection programs identify a single domain most similar to a RAB like GTPase domain, but could also be a Rac/Ras/Ran/Arf like GTPase. BLASTp find matches to many GTPase proteins	**S**

Tag sequence	Clone 32 tag count per 100 K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	CAP3 Contig	BLASTx	BLASTn	Further analysis of matches.	Possible function
			Cereb.	Ovary Epith.								
AGTTAAAACC	12.6	12.6	0.50%	0.62%	0	ESTs, Weakly similar to I57588 HSrel-1 [H.sapiens]	283364	1		Matches unassigned genomic sequence	Note: no PolyA signal/tail in ESTs, so this cluster may be a false match to the tag	
CCCCAAAGCT	12.6	12.6	0.50%	0.62%	0	ESTs	192855	1		Matches chromosome 19 sequence		
TGCCAGGTGC	12.6	12.6	0.50%	0.62%	1	ESTs (possibly fits in albumin, cluster too large to assemble)	356560	1	Short tyrosine rich region, in frame -3 matches other uncharacterised sequences	Matches chromosome 5 sequence. Poly A site region matches 5' of "Homo sapiens cDNA FLJ39969 fis, clone SPLEN2027868" (AK097288, first 744 bases of 2789)		
CGTGAAAGAT	11.1	11.1	0.89%	1.09%	0	ESTs	60416	1		Matches unassigned genomic sequence		
TCACGTGGGC	11.1	11.1	0.89%	1.09%	0	ESTs	271985	1	Matches ALU CLASS C WARNING ENTRY in many places	Matches unassigned genomic sequence	Note: no PolyA signal/tail PolyA signal/tail is present in the contig, so this cluster may be a false match to the tag	

Tag sequence	Clone 32 tag count per 100 K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	CAP3 Contig	BLASTx	BLASTn	Further analysis of matches.	Possible function
			Cereb.	Ovary Epith.								
AACATTTGTG	11.1	11.1	0.89%	1.09%	1	ESTs	279619	1		Almost identical to 3' end of "**Homo sapiens LY49L pseudogene**, complete sequence"	ESTs appear to match the Ly49L pseudogene 400bp after the end of the only sequenced cDNA but may represent the true 3' end. The human Ly49L gene appears to be a **truncated form of a Ly49 lectin-like MHC-receptor** (these are the main MHC recognising receptors in rodent NK cells). It's **truncation probably renders it inactive**. It may be a mutated form of the primordial Ly49 gene or the last of a deleted family in humans. Human NK cells recognise MHC with the KIR/KAR IgSF proteins so do not need functional Ly49 genes for this purpose	**evoluti-onary interest**
GTGAGAACCC	11.1	11.1	0.89%	1.09%	1	ESTs	170498	1	3′ eighth of contig matches uncharacterised sequences	3' end of contig matches sequence from several chromosomes		

Tag sequence	Clone 32 tag count per 100 K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	CAP3 Contig	BLASTx	BLASTn	Further analysis of matches.	Possible function
			Cereb.	Ovary Epith.								
TTCCTTTTAC	11.1	11.1	0.89%	1.09%	1	Matches ESTs from RAD50 (S. cerevisiae) homolog	41587	3	Identical to C-terminus of uncharacterised protein XP_098434 [Homo sapiens] (XM_098434)	Approx three-quarters of the contig is identical to 3' end of "Homo sapiens hypothetical gene LOC153852 (LOC153852)" which encodes **predicted protein XP_098434**	Sequence XP_098434 was removed from GenBank at the submitters request and no domains were detected in this protein.	

Tag sequence	Clone 32 tag count per 100 K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	CAP3 Contig	BLASTx	BLASTn	Further analysis of matches.	Possible function
			Cereb.	Ovary Epith.								
TGTTGACTCT	19.0	9.6	0.32%	0.06%	0	ESTs	8882	1		Identical to sequence from Chromosome 14q31. 3′ third of contig matches (82%) 3′ end of the **Bos. taurus orphan G protein-coupled receptor bRGR1 mRNA** (BTU88366). Central region of contig (290-362) also matches (87%) slightly 5′ of the 3′ end of this gene. The matches are approximately 1900 bp 3′ of the protein coding region of the gene	The human homologue of Bos. taurus orphan G protein-coupled receptor bRGR1 mRNA (BTU88366) is "GP68_HUMAN **Probable G protein-coupled receptor GPR68** (Ovarian cancer G protein-coupled receptor 1) (OGR-1)" (U48405). The sequence for this gene is identical to a region 800bp upstream of the match to this contig on chromosome 14 (in the same BAC clone), suggesting that these ESTs are from the 3' UTR for this gene. The ligand for this receptor and the effects of its signalling are unknown	**CS**

Tag sequence	Clone 32 tag count per 100 K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	CAP3 Contig	BLASTx	BLASTn	Further analysis of matches.	Possible function
			Cereb.	Ovary Epith.								
AATATTTTCT	9.5	9.5	1.61%	1.92%	0	ESTs in Homo sapiens cDNA FLJ35429 fis, clone SMINT2002126	377850	1		Matches many genomic clones containing MHC molecules (mainly HLA-F) but appears to match both intronic and exonic sequences, with the SAGE tag found in intronic sequence		
ATGCTAAGCT	9.5	9.5	1.61%	1.92%	0	ESTs	233071	1		Matches unassigned genomic sequence		
GCGAACTCCG	9.5	9.5	1.61%	1.92%	0	ESTs, Weakly similar to ALU1_HUMAN ALU SUBFAMILY J SEQUENCE CONTAMINATION WARNING ENTRY [H.sapiens]	248844	1	Matches "Alu subfamily J sequence contamination warning entry" in many places	Matches unassigned genomic sequence	Note: no PolyA signal/tail PolyA signal/tail is present in the contig, so this cluster may be a false match to the tag	
TGTTCTCCCC	9.5	9.5	1.61%	1.92%	0	ESTs	126630	2		Matches unassigned genomic sequence		

Tag sequence	Clone 32 tag count per 100 K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	CAP3 Contig	BLASTx	BLASTn	Further analysis of matches.	Possible function
			Cereb.	Ovary Epith.								
AAACTGTCAG	9.5	9.5	1.61%	1.92%	1	ESTs, Weakly similar to JC1405 6-pyruvoyltetrahydropterin synthase [H.sapiens]	14204	1	Central region of contig is identical to the C-terminus of uncharacterised protein sequence XP_094531 (XM_094531). Central region of contig also matches (~40% identity) 6-pyruvoyl-tetrahydropterin synthase from various organisms	Central region of contig is identical to 3' end of "**Homo sapiens similar to 6-pyruvoyl-tetrahydropterin synthase** (LOC167514)" mRNA	The contig contains additional 3' UTR sequence after the 6-pyruvoyl-tetrahydropterin synthase match. However, the region of the contig 5′ of this match is not similar to the predicted gene, suggesting the prediction has the wrong 5' exon or that the contig contains chimeric ESTs. The predicted gene contains a domain that aligns poorly to a **6-Pyruvoyl tetrahydrobiopterin synthase domain** in an CD search and is similar to the PDB entry for rat 6-pyruvoyl tetrahydrobiopterin synthase in SMART. This gene is the second enzyme in the biosynthetic pathway of tetrahydrobiopterin. Tetrahydrobiopterin (BH4) is the redox cofactor for the aromatic amino acid hydroxylases such as phenylalanine hydroxylase	**O**

Tag sequence	Clone 32 tag count per 100 K	Ratio CTL vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	CAP3 Contig	BLASTx	BLASTn	Further analysis of matches.	Possible function
			Cereb.	Ovary Epith.								
GAATGCGAAG	7.9	7.9	2.91%	3.38%	0	ESTs	92448	1		Matches unassigned genomic sequence		
CGCGGCAGCT	14.2	7.2	1.50%	0.35%	0	ESTs, Weakly similar to MASL1 [H.sapiens]	132216	2	Central region of contig is identical to large central region of uncharacterised sequence XP_084712 [Homo sapiens] (XM_084712)	3' end of contig matches (99%) entire length of "**Homo sapiens LOC144059** (LOC144059), mRNA". 2 regions of the contig also match "Homo sapiens LOC144058 (LOC144058), mRNA"	No domains were detected in predicted LOC144058 protein and BLASTp gave no additional matches. No PolyA signal/tail is present in the contig, so this cluster may be a false match to the tags	
TACCTGTCTG	12.6	6.4	2.49%	0.62%	0	ESTs	356517	1	No significant matches	Matches genomic sequence from chromosome 7p11.2-q11.2. Central region almost identical to Macaca fascicularis brain cDNA clone:QflA-12743, full insert sequence		

Appendix D: Immune specific NK SAGE tags.

Structural and functional data, where this is known, for proteins corresponding to all 256 SAGE tags that are "immune specific" (i.e., up-reg against cerebellum, ovary epithelium and a panel of cancers libraries) are listed. The p values associated with the differential expression of the tag in the NK cell line compared to normal cerebellum (cereb.) and ovary epithelium (ovary epith.), calculated using the AC test, are also shown. "Number of cancers" indicates the number of cancer-derived SAGE libraries, out of 12 tested, in which the tag abundance is at least one-third of that in the NK cell line. SAGE tags were assigned to transcripts, and these transcripts were assessed using their LocusLink, Unigene, and RefSeq database entries and the literature. In the Knowledge column, transcripts are categorised according to their current level of characterisation, and well-characterised transcripts are divided according to whether they have been assigned any immune function in the literature. In the Function column, transcripts are categorised according to their broad molecular function.

Characterisation classes are: K, known immune function; O, other known function; E, tag matches EST sequences only; U, uncharacterised (but sequenced) transcript; M, tag matches multiple transcripts; N, no match to Unigene. Functional classes are: E, soluble effector molecules; CS, cell surface molecules; S, signalling molecules; AP, antigen presentation; T, transcriptional regulation; Cy, cytoskeleton related; CC, cell cycle or viability related; P, protein synthesis related (including mRNA processing); O, other (including housekeeping transcripts such as metabolic enzymes).

Tag sequence	NK tag count per 100,000	Ratio NK vs. Cereb.	Ratio NK vs. Ovary Epith.	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
				Cereb.	Ovary Epith.					
AGGAGGTATCA	1056.0	1056.0	1056.0	0.00%	0.00%	0	granulysin	105806	K	E
GGCTGGGGGCC	756.6	64.1	1.9	0.00%	0.00%	1	profilin 1	75721	O	Cy
TGGCTCCTCCC	239.7	239.7	57.5	0.00%	0.00%	0	lymphocyte cytosolic protein 1 (L-plastin)	76506	K	Cy
CTGGCCCGAGG	207.9	207.9	207.9	0.00%	0.00%	0	Rho GDP dissociation inhibitor (GDI) beta	83656	O	S
CTGACCTGTGT	199.6	101.4	2.9	0.00%	0.00%	1	major histocompatibility complex, class I, B	77961	K	Ap
GCAGTGGGAAA	149.7	149.7	149.7	0.00%	0.00%	0	lymphotoxin beta (TNF superfamily, member 3)	890	K	CS
CAGCGCTGCCG	142.7	142.7	142.7	0.00%	0.00%	0			N	N
GCTGCCCAGGC	140.0	140.0	67.2	0.00%	0.00%	0			N	N
AAAAATCGGCT	126.1	126.1	126.1	0.00%	0.00%	0	small inducible cytokine A5 (RANTES)	241392	K	E
AAACGCCACTA	119.2	119.2	119.2	0.00%	0.00%	0			N	N
GCGGTGTACAC	116.4	116.4	116.4	0.00%	0.00%	0	natural killer cell group 7 sequence	10306	K	CS
GCACCAAAGCC	110.9	110.9	110.9	0.00%	0.00%	0	small inducible cytokine A3 (homologous to mouse Mip-1a) / hypothetical protein FLJ12154	73817 / 6839	xs	xs
GTGCTGGACCT	106.7	106.7	2.3	0.00%	0.01%	1	proteasome (prosome, macropain) activator subunit 2 (PA28 beta)	179774	K	Ap
CCAGGCAGGGG	103.9	52.8	103.9	0.00%	0.00%	0	DNAX-activation protein 10	117339	K	CS
GAGGCTCGGCT	98.4	98.4	98.4	0.00%	0.00%	0	protein tyrosine phosphatase, non-receptor type 7	35	K	S
TACATAGCTTG	98.4	98.4	98.4	0.00%	0.00%	0	perforin 1 (pore forming protein)	2200	K	E
GGGGCAACAG	95.6	95.6	95.6	0.00%	0.00%	0	CDW52 antigen (CAMPATH-1 antigen)	276770	K	CS
TGTACCCCGCT	95.6	95.6	95.6	0.00%	0.00%	0	protein tyrosine phosphatase, receptor type, C-associated protein	155975	K	S
GCAGAGAAAAA	87.3	87.3	87.3	0.00%	0.00%	0	coronin, actin-binding protein, 1A	109606	O	Cy
AAGCCCCTTCA	84.5	43.0	84.5	0.00%	0.00%	0	cathepsin W (lymphopain)	87450	K	O
CCTGCAACCTT	84.5	84.5	84.5	0.00%	0.00%	0	integrin, beta 7	1741	K	CS
AAACGCTACTA	76.2	76.2	76.2	0.00%	0.00%	0	granzyme B (granzyme 2, cytotoxic T-lymphocyte-associated serine esterase 1)	1051	K	E
CCCCCTCCTGC	76.2	76.2	76.2	0.00%	0.00%	0	interleukin 2 receptor, beta (CD122)	75596	K	CS
CTCCTGGGCAA	70.7	70.7	70.7	0.00%	0.00%	0			N	N
GGGAGGTAGCA	69.3	69.3	11.1	0.00%	0.00%	0	basic helix-loop-helix domain containing, class B, 2	171825	O	CC
ACGATGGCCGA	67.9	67.9	67.9	0.00%	0.00%	0	glia maturation factor, gamma	5210	O	S
CACCACGGTGT	66.5	3.4	31.9	0.01%	0.00%	1	RNB6	241471	H	U
ATTTGAGCAGA	65.1	33.1	65.1	0.00%	0.00%	0			N	N
AGCAGCTGCTG	63.7	63.7	63.7	0.00%	0.00%	0	megakaryocyte-associated tyrosine kinase	274	K	S
GTAGCACCTCC	59.6	59.6	59.6	0.00%	0.00%	0	cystatin F (leukocystatin)	143212	K	Ap

Tag sequence	NK tag count per 100,000	Ratio NK vs. Cereb.	Ratio NK vs. Ovary Epith.	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
				Cereb.	Ovary Epith.					
CGGATAACCAG	59.6	7.6	4.8	0.00%	0.00%	1	proliferation-associated 2G4, 38kD / Human erbB3 binding protein EBP1 mRNA, complete cds NB both clusters cDNA are basically the same.	374491 / 343258	O	CC
TGGAACCCTGA	54.0	54.0	54.0	0.00%	0.00%	0	ESTs, Highly similar to hypothetical protein FLJ13725; KIAA1930 protein [Homo sapiens] [H.sapiens]	355944	E	Est
ACTAAGAGCCT	54.0	54.0	54.0	0.00%	0.00%	0			N	N
ACTGTAAAAAA	54.0	27.5	25.9	0.00%	0.00%	0	novel C3HC4 type Zinc finger (ring finger) / granzyme A (granzyme 1, cytotoxic T-lymphocyte-associated serine esterase 3)	318584 / 90708	xs	xs
AGTATCTGGGA	52.7	52.7	52.7	0.00%	0.00%	1	actin related protein 2/3 complex, subunit 1A (41 kD)	11538	O	Cy
GATAACACATT	47.1	47.1	47.1	0.00%	0.00%	0	small inducible cytokine A4 (homologous to mouse Mip-1b)	75703	K	E
CAAATCCAAA	47.1	3.0	3.8	0.14%	0.04%	1			N	N
CCGGACCTGTG	45.7	7.7	11.0	0.00%	0.00%	0	Homo sapiens cDNA FLJ31238 fis, clone KIDNE2004864	375208	H	U
TCCTATAAGC	45.7	45.7	11.0	0.00%	0.00%	1			N	N
AGGCCAGGCCG	44.3	44.3	21.3	0.00%	0.00%	0	AT-hook transcription factor AKNA	159578	K	T
CCCCCTCGTGC	44.3	4.5	10.6	0.03%	0.00%	0	adrenergic, beta, receptor kinase 1	83636	O	S
CTGGCCCGGAG	44.3	22.5	44.3	0.00%	0.00%	1	vasodilator-stimulated phosphoprotein	93183	O	Cy
TGGAAGCATCT	43.0	43.0	43.0	0.00%	0.00%	0	HSPC022 protein	367740	H	U
TGGCCTGCCCA	43.0	5.5	20.6	0.01%	0.00%	1	MLL septin-like fusion	181002	H	U
GAACCGTCCTG	41.6	21.1	41.6	0.00%	0.00%	0	ESTs, Weakly similar to T51376 plant adhesion molecule 1 (PAM1) - Arabidopsis thaliana [A.thaliana]	123164	E	Est
ACCATTGGATT	41.6	41.6	41.6	0.00%	0.00%	1	interferon induced transmembrane protein 1 (9-27)	146360	H	U
TCTGCCCTCAA	41.6	41.6	41.6	0.00%	0.00%	0	killer cell immunoglobulin-like receptor, three domains, long cytoplasmic tail, 1 / killer cell immunoglobulin-like receptor, three domains, long cytoplasmic tail, 2 / NK-receptor	274601 / 56328 / 274484	K	CS
TGGGAGCTCAG	41.6	41.6	41.6	0.00%	0.00%	0	lymphocyte antigen 117	88411	K	CS
TGTAGATGCGA	41.6	41.6	41.6	0.00%	0.00%	0	CD2 antigen (p50), sheep red blood cell receptor	89476	K	CS
AGGAAGCTACA	40.2	40.2	40.2	0.00%	0.00%	0	T cell receptor delta locus / T cell receptor delta diversity 3	2014 / 367739	K	CS
ACCTCCACACG	40.2	40.2	9.6	0.00%	0.00%	0	centaurin, beta 1	108947	O	S
ATGTAGAGTGT	40.2	40.2	19.3	0.00%	0.00%	1	thymidylate synthetase	82962	O	O

Tag sequence	NK tag count per 100,000	Ratio NK vs. Cereb.	Ratio NK vs. Ovary Epith.	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
				Cereb.	Ovary Epith.					
GACCTGGTGCC	38.8	38.8	18.6	0.00%	0.00%	0	c-src tyrosine kinase	77793	K	S
TAACAGCCAGG	37.4	37.4	18.0	0.00%	0.00%	1	Nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, alpha	81328	K	T
GGGAAGGGGAA	36.0	36.0	36.0	0.00%	0.00%	0	19A24 protein	132906	K	CS
TATGAGGAGG	36.0	36.0	36.0	0.00%	0.00%	0	matrix metalloproteinase 25 (leukolysin)	198265	K	E
TGTAGGTCATT	36.0	4.6	17.3	0.06%	0.00%	0	G protein-coupled receptor 87 / ADP-ribosylation factor-like 7	58561 / 111554	xs	xs
ACCTGCCGACA	34.6	34.6	34.6	0.00%	0.00%	0	1) Homo sapiens cDNA FLJ14866 fis, clone PLACE1002066 / 2) Tumor suppressor deleted in oral cancer-related 1	1) 288941 / 2) 379039	H	U
GGGCTAGTGGG	34.6	8.8	34.6	0.01%	0.00%	0	RAS guanyl releasing protein 1 (calcium and DAG-regulated)	182591	O	S
CACCCAATGGG	33.3	33.3	16.0	0.00%	0.00%	0	SEC7 homolog	110121	H	U
TCCTCTTTCCA	33.3	33.3	16.0	0.00%	0.00%	1	natural killer cell transcript 4	943	H	U
CCTCAGCCCTG	33.3	33.3	33.3	0.00%	0.00%	0	protein tyrosine phosphatase, non-receptor type 6	63489	K	S
GCCCCCTTCCT	33.3	16.9	33.3	0.00%	0.00%	0	Tumor necrosis factor receptor superfamily, member 18	212680	K	CS
ACCCCCAAGCA	33.3	33.3	33.3	0.00%	0.00%	0	pleckstrin	77436	O	S
GCCTCTGTCTC	33.3	16.9	5.3	0.00%	0.07%	1	ribosomal protein, large, P1	177592	O	P
CTGGCCATCGA	31.9	31.9	31.9	0.00%	0.00%	1	Homo sapiens mRNA for FLJ00043 protein, partial cds	356684	H	U
GGCAGGCACAA	31.9	31.9	15.3	0.00%	0.01%	0	emopamil-binding protein (sterol isomerase)	75105	O	O
CGCCCCGGCGG	30.5	30.5	30.5	0.00%	0.00%	0	ESTs	196244	E	Est
CCCCGGGCCTC	30.5	30.5	30.5	0.00%	0.00%	0	phosphodiesterase 4A, cAMP-specific (dunce (Drosophila)-homolog phosphodiesterase E2)	89901	O	S
AGGCTCCGTGG	29.1	14.8	29.1	0.01%	0.00%	0	minor histocompatibility antigen HA-1	196914	H	U
CTCCTCCAAGT	29.1	29.1	29.1	0.00%	0.00%	0	Homo sapiens, clone MGC:40121 IMAGE:5216355, mRNA, complete cds	15284	H	U
GGACCAGCGCC	29.1	29.1	29.1	0.00%	0.00%	0	hypothetical protein FLJ21438	136979	H	U
GGGGGTGGGTG	29.1	14.8	2.8	0.01%	2.17%	1	Mago-nashi homolog, proliferation-associated (Drosophila)	57904	H	U
GTGCCACCAGT	29.1	14.8	29.1	0.01%	0.00%	1	KIAA1554 protein	17767	H	U
AATTTGAGCAG	29.1	29.1	29.1	0.00%	0.00%	0			N	N
ATGACAGATGG	27.7	4.7	27.7	0.36%	0.00%	0	lung cancer-associated Y protein	13775	H	U
TCCCTATAGC	27.7	27.7	27.7	0.00%	0.00%	1			N	N
TCTGCCCCCAA	27.7	27.7	27.7	0.00%	0.00%	1	Homo sapiens extracellular signal-regulated kinase 8 mRNA, complete cds	133017	O	S
TACGAGGCCGG	26.3	26.3	26.3	0.00%	0.00%	1	expressed in activated T/LAK lymphocytes	16165	H	U

Tag sequence	NK tag count per 100,000	Ratio NK vs. Cereb.	Ratio NK vs. Ovary Epith.	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
				Cereb.	Ovary Epith.					
GCTGAGTGCAG	26.3	26.3	26.3	0.00%	0.00%	0			N	N
CCCCCAATGCT	26.3	6.7	26.3	0.11%	0.00%	1	splicing factor 3a, subunit 2, 66kD	115232	O	P
CCGCTTACTCT	26.3	26.3	12.6	0.00%	0.03%	1	v-ets erythroblastosis virus E26 oncogene homolog 1 (avian)	18063	O	T
GCCGGCCGGAC	24.9	24.9	2.4	0.00%	4.04%	0	B-cell CLL/lymphoma 7C	303197	H	U
GCTCCAGAACT	24.9	24.9	24.9	0.00%	0.01%	0	leukocyte membrane antigen	9688	K	CS
CTGATCTCCAA	24.9	24.9	24.9	0.00%	0.01%	0			N	N
GTGGGCCGGCT	24.9	6.3	12.0	0.17%	0.05%	1			N	N
CCTGTAGCCTT	24.9	24.9	24.9	0.00%	0.01%	0	Gardner-Rasheed feline sarcoma viral (v-fgr) oncogene homolog	1422	O	S
CAGAGCTGTGC	23.6	23.6	23.6	0.01%	0.01%	0	phorbolin-like protein MDS019	250619	H	U
GGAGCTTGAGG	23.6	23.6	11.3	0.01%	0.08%	0	chromosome 6 open reading frame 9	288316	H	U
GGTAGAACTAA	23.6	23.6	23.6	0.01%	0.01%	0	chromosome X open reading frame 9	61469	H	U
TGCAAGAGAGG	23.6	23.6	23.6	0.01%	0.01%	0	Homo sapiens mRNA; cDNA DKFZp667K0625 (from clone DKFZp667K0625)	238954	H	U
TGTGGGAACCA	23.6	23.6	23.6	0.01%	0.01%	1	hypothetical protein AL133206	7750	H	U
CGAGCCTGTTA	23.6	23.6	23.6	0.01%	0.01%	0	zeta-chain (TCR) associated protein kinase (70 kD)	234569	K	S
CTGAAGCCAAA	23.6	6.0	23.6	0.26%	0.01%	0	interleukin 16 (lymphocyte chemoattractant factor)	82127	K	E
AGCCGGGATGG	23.6	6.0	3.8	0.26%	1.12%	1	proteasome (prosome, macropain) subunit, beta type, 9 (large multifunctional protease 2)	9280	K	Ap
GGCCTGCAGGA	23.6	23.6	23.6	0.01%	0.01%	1	apoptosis-associated speck-like protein containing a CARD	71869	K	S
CTGTGCGCCCT	23.6	23.6	23.6	0.01%	0.01%	0			N	N
CACTTTACCAG	23.6	23.6	23.6	0.01%	0.01%	0	runt-related transcription factor 3	170019	O	T
TTCTGTAGCCC	23.6	23.6	23.6	0.01%	0.01%	0	ATPase, Ca++ transporting, ubiquitous	5541	O	H
TTTCTGTCTGG	23.6	23.6	3.8	0.01%	1.12%	0	phosphoinositide-3-kinase, catalytic, delta polypeptide	162808	O	S
TGTTGACTCTG	22.2	11.3	22.2	0.15%	0.03%	0	ESTs	8882	E	Est
GTGATGGGGGC	22.2	22.2	22.2	0.02%	0.03%	1	EST in Homo sapiens mRNA for FLJ00067 protein, partial cds	41045	E	Est
TCCGCAAGGTG	22.2	22.2	22.2	0.02%	0.03%	0	Homo sapiens mRNA; cDNA DKFZp667F1219 (from clone DKFZp667F1219)	37617	H	U
TAGCCCCCTGG	22.2	22.2	22.2	0.02%	0.03%	0	tumor necrosis factor (TNF superfamily, member 2)	241570	K	E
TGTGATCACAA	22.2	22.2	10.6	0.02%	0.21%	1	proteasome (prosome, macropain) subunit, beta type, 10	9661	K	Ap

Tag sequence	NK tag count per 100,000	Ratio NK vs. Cereb.	Ratio NK vs. Ovary Epith.	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
				Cereb.	Ovary Epith.					
AATATGAGCAG	22.2	22.2	22.2	0.02%	0.03%	0			N	N
GCTGGGTAACC	22.2	22.2	22.2	0.02%	0.03%	0	septin 6	90998	O	Cy
CCTAGGACCTG	22.2	22.2	10.6	0.02%	0.21%	1	actin related protein 2/3 complex, subunit 4 (20 kD)	323342	O	Cy
TACTTGGTCTT	20.8	20.8	10.0	0.02%	0.21%	1	ESTs in thymopoietin	11355	E	Est
GCCCAGCCCTG	20.8	10.6	20.8	0.15%	0.03%	0	hypothetical protein 384D8_6	180903	H	U
TAACCTCAGGT	20.8	20.8	20.8	0.02%	0.03%	0	KIAA0870 protein	18166	H	U
TGATCTGCCTG	20.8	20.8	20.8	0.02%	0.03%	1	hypothetical protein MGC5178	326067	H	U
TAGGCACCTGT	20.8	20.8	20.8	0.02%	0.03%	0			N	N
TCCTATTAGCC	20.8	20.8	20.8	0.02%	0.03%	1			N	N
AGGGCCGACTG	20.8	20.8	10.0	0.02%	0.21%	0	antigen identified by monoclonal antibody Ki-67	80976	O	CC
GGAGCTGTCTG	19.4	19.4	19.4	0.03%	0.05%	0	differentially expressed in FDCP (mouse homolog) 6	15476	H	U
CAGCCTTGCGG	19.4	9.9	9.3	0.24%	0.33%	1	hypothetical protein similar to actin related protein 2/3 complex, subunit 5	315164	H	U
TGGTCCACGGC	19.4	19.4	19.4	0.03%	0.05%	1	HSPC037 protein	108196	H	U
CTAAAGCCTTC	19.4	19.4	19.4	0.03%	0.05%	0	CD3Z antigen, zeta polypeptide (TiT3 complex)	97087	K	CS
GCAGAAGCACA	19.4	4.9	19.4	0.93%	0.05%	0	serine/threonine kinase 10	16134	K	S
GCTTCCCCTTG	19.4	19.4	19.4	0.03%	0.05%	0	SH2 domain protein 2A	103527	K	S
CAGGACAGGGT	19.4	19.4	19.4	0.03%	0.05%	0	NESH protein	130719	O	S
GAGCAGGCAAA	19.4	19.4	19.4	0.03%	0.05%	0	apolipoprotein B48 receptor	200333	K	CS
GGCACCTCGGG	19.4	9.9	19.4	0.24%	0.05%	0	Collagen, type IX, alpha 2	37165	O	O
GTTCGGGCCGG	19.4	19.4	4.7	0.03%	1.25%	1	dipeptidylpeptidase III	22880	O	O
TCAGTGACCAG	18.0	18.0	18.0	0.06%	0.08%	0	hypothetical protein MGC5363	1880	H	U
ACGCTGCGGCT	18.0	18.0	18.0	0.06%	0.08%	0	T cell receptor alpha locus	74647	K	CS
AGCAGCCGCTC	18.0	9.2	18.0	0.39%	0.08%	0	Kruppel-like factor 13	7104	K	T
AGGGATCACAG	18.0	18.0	18.0	0.06%	0.08%	0	lymphotoxin alpha (TNF superfamily, member 1)	36	K	E
ATGGAAGTCTG	18.0	18.0	18.0	0.06%	0.08%	1	inositol polyphosphate-5-phosphatase, 145kD	155939	K	S
GCAGTTCTGAC	18.0	18.0	18.0	0.06%	0.08%	1	major histocompatibility complex, class II, DR beta 5 / major histocompatibility complex, class II, DR beta 1	308026	K	Ap
TTAGGGAGGAG	18.0	9.2	18.0	0.39%	0.08%	1	intercellular adhesion molecule 3	99995	K	CS
GCACTACTCGA	16.6	8.4	16.6	0.62%	0.13%	0	hypothetical protein FLJ22573	352548	H	U
GCCGCCCGTCG	16.6	16.6	16.6	0.10%	0.13%	0	tumor necrosis factor (ligand) superfamily, member 12	26401	K	CS
TATTTATCCAA	16.6	16.6	16.6	0.10%	0.13%	0	integrin, alpha L (antigen CD11A (p180), lymphocyte function-associated antigen 1; alpha polypeptide)	174103	K	CS

Tag sequence	NK tag count per 100,000	Ratio NK vs. Cereb.	Ratio NK vs. Ovary Epith.	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
				Cereb.	Ovary Epith.					
TCTGCCTTCA	16.6	16.6	16.6	0.10%	0.13%	0	Killer cell immunoglobulin-like receptor, two domains, long cytoplasmic tail, 4	166085	K	CS
GAGGCCTGGCC	16.6	16.6	16.6	0.10%	0.13%	0			N	N
TGGCCCGACGA	16.6	16.6	16.6	0.10%	0.13%	0	nudix (nucleoside diphosphate linked moiety X)-type motif 1	388	O	O
GACGGCCAGAG	15.2	15.2	15.2	0.17%	0.22%	0	ESTs, Weakly similar to extensin-like protein [Arabidopsis thaliana] [A.thaliana]	294142	E	Est
AGGCCACTGGG	15.2	15.2	15.2	0.17%	0.22%	0	hypothetical protein MGC2463	323634	H	U
GGAGGGAGCTG	15.2	15.2	15.2	0.17%	0.22%	0	lens intrinsic membrane protein 2 (19kD) / Homo sapiens cDNA FLJ31501 fis, clone NT2NE2005537	162754 / 311208	H	U
TGGTTAATCTT	15.2	15.2	15.2	0.17%	0.22%	0	Homo sapiens mRNA; cDNA DKFZp667O1616 (from clone DKFZp667O1616)	365655	H	U
TGTGTGTGACA	15.2	15.2	15.2	0.17%	0.22%	1	hypothetical protein FLJ14466	55148	H	U
CATACAGAAAA	15.2	15.2	7.3	0.17%	1.27%	0	CD97 antigen	3107	K	CS
TCTCTCAAAGT	15.2	15.2	15.2	0.17%	0.22%	0	CD53 antigen	82212	K	CS
ATTTTGAGCAG	15.2	15.2	15.2	0.17%	0.22%	0			N	N
CACCCCTGGGG	15.2	15.2	15.2	0.17%	0.22%	0	acyloxyacyl hydrolase (neutrophil)	82542	O	E
GACAAGCCCAG	15.2	15.2	15.2	0.17%	0.22%	0	RAB5 interacting protein 3	180040	O	Cy
AGCCTCGGCCA	15.2	15.2	7.3	0.17%	1.27%	1	Rho guanine nucleotide exchange factor (GEF) 1	252280	O	S
CTGGTGGTGCC	15.2	15.2	7.3	0.17%	1.27%	1	Gem-interacting protein	49427	O	S
TTGCTCTGCGA	13.9	13.9	13.9	0.48%	0.61%	0	ESTs from myelin basic protein, matches one cDNA (AK074315) which does not match refseq	69547	H	U
GACCAACAAGC	13.9	13.9	13.9	0.48%	0.61%	0	killer cell lectin-like receptor subfamily B, member 1	169824	K	CS
GAGGCACTGAA	13.9	13.9	13.9	0.48%	0.61%	0	zinc finger protein, subfamily 1A, 1 (Ikaros)	54452	K	T
GCAGAAGAATG	13.9	13.9	6.7	0.48%	3.05%	0	colony stimulating factor 2 (granulocyte-macrophage)	1349	K	E
ATCGCGGAGG	13.9	13.9	13.9	0.48%	0.61%	0			N	N
ATGGGTGGGTG	13.9	13.9	13.9	0.48%	0.61%	0			N	N
CTAGCAGAAAC	13.9	13.9	13.9	0.48%	0.61%	0			N	N
GATTGAGCATA	13.9	13.9	6.7	0.48%	3.05%	0			N	N
GCCCGAGGAAG	13.9	13.9	13.9	0.48%	0.61%	0			N	N
GCTGGGCGCGG	13.9	7.0	13.9	2.48%	0.61%	0			N	N
CTGAGGTGTGA	13.9	13.9	13.9	0.48%	0.61%	0	runt-related transcription factor 3	170019	O	T
GATTTTCTGGG	13.9	7.0	13.9	2.48%	0.61%	0	pleckstrin homology, Sec7 and coiled/coil domains 4	7189	O	S
TGCCTGAGATC	13.9	7.0	13.9	2.48%	0.61%	0	antigen identified by monoclonal antibody Ki-67	80976	O	CC

Tag sequence	NK tag count per 100,000	Ratio NK vs. Cereb.	Ratio NK vs. Ovary Epith.	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
				Cereb.	Ovary Epith.					
GACAATGTATG	13.9	13.9	13.9	0.48%	0.61%	1	guanine nucleotide binding protein (G protein), gamma 2	289026	O	S
GTGCAGGTCTC	13.9	13.9	6.7	0.48%	3.05%	1	GDP dissociation inhibitor 2	56845	O	S
GGCTTGGGGG	12.5	12.5	12.5	0.48%	0.61%	1	Homo sapiens cDNA FLJ40207 fis, clone TESTI2020946 / ESTs	298296 / 364622	xs	xs
AAATCAGGAAC	12.5	12.5	12.5	0.48%	0.61%	1	MOB-LAK	180549	H	U
ACTGCTCATTG	12.5	12.5	6.0	0.48%	3.05%	1	NEDD8-conjugating enzyme	289051	K	T
CCTGGAGCGCC	12.5	6.3	12.5	2.48%	0.61%	1	flightless I (Drosophila) homolog	83849	H	U
GTTTCAAACGA	12.5	12.5	12.5	0.48%	0.61%	1	hypothetical protein MGC10966	180535	H	U
CAGGATGCTT	12.5	12.5	12.5	0.48%	0.61%	0	lymphocyte-specific protein 1	56729	K	Cy
CTTTTTTCCCA	12.5	12.5	12.5	0.48%	0.61%	0	CD48 antigen (B-cell membrane protein)	901	K	CS
TCTCCAAGTGT	12.5	12.5	12.5	0.48%	0.61%	0	signal transducer and activator of transcription 4	80642	K	T
AACATTTCTCT	12.5	12.5	12.5	0.48%	0.61%	1	hematopoietic cell-specific Lyn substrate 1	14601	K	S
CCTAAGTGACT	12.5	12.5	12.5	0.48%	0.61%	1	T cell receptor beta locus	303157	K	CS
AATACGAGCAG	12.5	12.5	12.5	0.48%	0.61%	0			N	N
AGTGGCAAGGG	12.5	12.5	12.5	0.48%	0.61%	0			N	N
CACCAGCTGGA	12.5	12.5	12.5	0.48%	0.61%	0			N	N
CTGGGGTGAGC	12.5	12.5	12.5	0.48%	0.61%	0			N	N
GGGAGGTATCA	12.5	12.5	12.5	0.48%	0.61%	0			N	N
GTGCCACTGC	12.5	6.3	6.0	2.48%	3.05%	0			N	N
TGATCAAGGTC	12.5	12.5	12.5	0.48%	0.61%	0			N	N
ACTGGTGGTCA	12.5	6.3	12.5	2.48%	0.61%	1			N	N
CAAGAGATGCT	12.5	12.5	12.5	0.48%	0.61%	0	Septin 1	99741	O	Cy
CTTGCAAACCA	12.5	12.5	12.5	0.48%	0.61%	0	baculoviral IAP repeat-containing 3	127799	O	CC
TCCAGCCAGCC	12.5	12.5	12.5	0.48%	0.61%	0	docking protein 2, 56kD	71215	O	S
AAGGAAAGGCC	12.5	12.5	12.5	0.48%	0.61%	1	manic fringe (Drosophila) homolog	31939	O	S
AGGAGGGTGGG	12.5	12.5	12.5	0.48%	0.61%	1	lamin B1	89497	O	Cy
CTGGCAGATTG	12.5	12.5	12.5	0.48%	0.61%	1	splicing factor 3a, subunit 3, 60kD	77897	O	P
CTGTACATACT	12.5	12.5	12.5	0.48%	0.61%	1	NS1-binding protein	197298	O	P
GAACATAGCCA	12.5	12.5	12.5	0.48%	0.61%	1	Rac GTPase activating protein 1	23900	O	S
TCCTACAATCT	12.5	12.5	6.0	0.48%	3.05%	1	f-box and leucine-rich repeat protein 11 / Homo sapiens mRNA; cDNA DKFZp451H072 (from clone DKFZp451H072); complete cds	219614 / 351547	xs	xs
GGGCGCCTGGC	11.1	11.1	11.1	1.41%	1.69%	0	immunity associated protein 1	159955	H	U
GGAGCACACAG	11.1	11.1	11.1	1.41%	1.69%	1	Hypothetical protein FLJ31952	193490	H	U
TGCCTATAGCC	11.1	11.1	11.1	1.41%	1.69%	1	hypothetical protein FLJ11029	274448	H	U
TGCTGCCGTGA	11.1	11.1	11.1	1.41%	1.69%	1	Hypothetical protein FLJ13491	33533	H	U

Tag sequence	NK tag count per 100,000	Ratio NK vs. Cereb.	Ratio NK vs. Ovary Epith.	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
				Cereb.	Ovary Epith.					
ACGCTCTCGAT	11.1	11.1	11.1	1.41%	1.69%	0	CD37 antigen	153053	K	CS
ACTAAGTGCTA	11.1	11.1	11.1	1.41%	1.69%	0	ESTs	132739	K	T
ACTGTTTCTCT	11.1	11.1	11.1	1.41%	1.69%	0	DNA segment on chromosome 12 (unique) 2489 expressed sequence	74085	K	CS
ATACCACTTTT	11.1	11.1	11.1	1.41%	1.69%	0	leupaxin	49587	K	S
GCTCCCCCTCC	11.1	11.1	11.1	1.41%	1.69%	0	Wiskott-Aldrich syndrome (eczema-thrombocytopenia)	2157	K	S
GTCACTCGATG	11.1	11.1	11.1	1.41%	1.69%	0	protein tyrosine phosphatase, receptor type, C	170121	K	CS
TAAGGGAGCCA	11.1	11.1	11.1	1.41%	1.69%	0	T cell activation, increased late expression	142023	K	CS
CTGGTTTTATT	11.1	11.1	11.1	1.41%	1.69%	1	granzyme K (serine protease, granzyme 3; tryptase II)	3066	K	E
GCCATTGTCCC	11.1	11.1	11.1	1.41%	1.69%	1	apolipoprotein L, 3	241535	K	O
TTGAGATAACT	11.1	11.1	11.1	1.41%	1.69%	1	lymphocyte cytosolic protein 2 (SH2 domain-containing leukocyte protein of 76kD)	2488	K	S
AACGCGAGCAG	11.1	11.1	11.1	1.41%	1.69%	0			N	N
CATTTGAGCAG	11.1	11.1	11.1	1.41%	1.69%	0			N	N
CTGCAGTTATA	11.1	11.1	11.1	1.41%	1.69%	0			N	N
CGGCCCAGGAT	11.1	11.1	11.1	1.41%	1.69%	1			N	N
CAGTGGAGGGG	11.1	11.1	11.1	1.41%	1.69%	0	CIDEB Cell death-inducing DFFA-like effector b	288835	O	CC
CGGTGGATTT	11.1	11.1	11.1	1.41%	1.69%	1	excision repair cross-complementing rodent repair deficiency, complementation group 5 (xeroderma pigmentosum, complementation group G (Cockayne syndrome))	48576	O	O
GCGGGCGCCTT	9.7	9.7	9.7	1.41%	1.69%	1	Cold inducible RNA binding protein	119475	O	CC
TGCCAATTAAG	9.7	9.7	9.7	1.41%	1.69%	1	ESTs	165337	E	Est
CAGGTTAAGCT	9.7	9.7	9.7	1.41%	1.69%	0	Janus kinase 3 (a protein tyrosine kinase, leukocyte), tag matches mRNA which is similar to Janus kinase 3	99877	H	U
ATGACTGCTGT	9.7	9.7	9.7	1.41%	1.69%	1	hypothetical protein FLJ22794	19525	H	U
AGCTGACGGTG	9.7	9.7	9.7	1.41%	1.69%	0	neutrophil cytosolic factor 4 (40kD)	196352	K	S
GCCAAGGAGGG	9.7	9.7	9.7	1.41%	1.69%	0	CD7 antigen (p41)	36972	K	CS
AAGACACCAAG	9.7	9.7	9.7	1.41%	1.69%	0			N	N
AATGTGAGCAG	9.7	9.7	9.7	1.41%	1.69%	0			N	N
CAGATCCAAA	9.7	9.7	9.7	1.41%	1.69%	0			N	N
CAGTTGGTACT	9.7	9.7	9.7	1.41%	1.69%	0			N	N
GCAAACGACAG	9.7	9.7	9.7	1.41%	1.69%	0			N	N
GCAAGTGGGAA	9.7	9.7	9.7	1.41%	1.69%	0			N	N
GGGCCAAAGTC	9.7	9.7	9.7	1.41%	1.69%	0			N	N

Tag sequence	NK tag count per 100,000	Ratio NK vs. Cereb.	Ratio NK vs. Ovary Epith.	p value		Number of Cancers	Unigene Description	Unigene Cluster	Knowledge	Function
				Cereb.	Ovary Epith.					
GTTCGTGCAGA	9.7	9.7	9.7	1.41%	1.69%	0			N	N
TAGGGTCTTGA	9.7	9.7	9.7	1.41%	1.69%	0			N	N
ATGAGCAACTT	9.7	9.7	9.7	1.41%	1.69%	0	dual specificity phosphatase 4	2359	O	S
GGAGGAGCTGT	9.7	9.7	9.7	1.41%	1.69%	1	KIAA1898 protein	22410	O	S
GAGTCCGGGAT	8.3	8.3	8.3	2.40%	2.82%	0	ESTs, Weakly similar to TRYG_HUMAN Tryptase gamma precursor (Transmembrane tryptase) [H.sapiens]	177971	E	Est
TCGGACCTGAG	8.3	8.3	8.3	2.40%	2.82%	0	ESTs	60137	E	Est
CAGGTCCCTGA	8.3	8.3	8.3	2.40%	2.82%	0	Immune associated nucleotide 4 like 1 (mouse)	26194	H	U
CCCTCTGTGCT	8.3	8.3	8.3	2.40%	2.82%	1	Homo sapiens cDNA FLJ38402 fis, clone FEBRA2008210	169061	H	U
GTCTGATATCT	8.3	8.3	8.3	2.40%	2.82%	1	hypothetical protein FLJ11712	14920	H	U
TATGTCTTGGA	8.3	8.3	8.3	2.40%	2.82%	0	lymphocyte-specific protein tyrosine kinase	1765	K	S
TCGTCAAAGCT	8.3	8.3	8.3	2.40%	2.82%	0	SH2 domain protein 1A, Duncan's disease (lymphoproliferative syndrome)	151544	K	S
AGCAGCAGAAA	8.3	8.3	8.3	2.40%	2.82%	0			N	N
CAAGGGCTTAG	8.3	8.3	8.3	2.40%	2.82%	0			N	N
CCTTCGAGCAG	8.3	8.3	8.3	2.40%	2.82%	0			N	N
GAGCGGCTACC	8.3	8.3	8.3	2.40%	2.82%	0			N	N
GCTGCCAGGC	8.3	8.3	8.3	2.40%	2.82%	0			N	N
TAAGTTACCAG	8.3	8.3	8.3	2.40%	2.82%	0			N	N
TAGACCTGACA	8.3	8.3	8.3	2.40%	2.82%	0			N	N
TAGGTTGCTA	8.3	8.3	8.3	2.40%	2.82%	0			N	N
TGCCTCCTTTG	8.3	8.3	8.3	2.40%	2.82%	0			N	N
AACACCCTTTC	8.3	8.3	8.3	2.40%	2.82%	1			N	N
AGCTACAGGTG	8.3	8.3	8.3	2.40%	2.82%	1			N	N
AGGCCGTCCCC	8.3	8.3	8.3	2.40%	2.82%	1			N	N
GAAACCGGGCT	8.3	8.3	8.3	2.40%	2.82%	1			N	N
GGGCCAGGTGT	8.3	8.3	8.3	2.40%	2.82%	1			N	N
GGTCCAGAACT	8.3	8.3	8.3	2.40%	2.82%	1			N	N
ACTTTAGCCTC	8.3	8.3	8.3	2.40%	2.82%	0	mitogen-activated protein kinase kinase kinase kinase 1	86575	O	S
GAGCAGTGCTG	8.3	8.3	8.3	2.40%	2.82%	0	Feline sarcoma oncogene	7636	O	S
GGTTGATTCTG	8.3	8.3	8.3	2.40%	2.82%	0	phosphodiesterase 4D, cAMP-specific (dunce (Drosophila)-homolog phosphodiesterase E3)	172081	O	S
TAATGGTATCT	8.3	8.3	8.3	2.40%	2.82%	1	caspase 3, apoptosis-related cysteine protease	74552	O	CC
TCTACTCAGCA	8.3	8.3	8.3	2.40%	2.82%	1	Hyaluronan-mediated motility receptor (RHAMM)	72550	O	CS

Appendix E.1: NK-specific tags that match uncharacterised but sequenced transcripts.

NK-specific tags were defined as those significantly more abundant (p value ≤ 0.05 by the AC test) in the NK library compared to both normal cerebellum (marked cereb.) and ovary epithelium (marked as ovary epith.) libraries and which were not found at similar levels (more than one-third that in the NK library) in more than 1 of 12 cancer SAGE libraries. The Unigene clusters matching these tags are listed, together with a representative GenBank accession number. The results of the bioinformatic analysis are shown, with the main points of interest in bold type. "Notes on transcript match" indicates any problems with the transcript assignment to the SAGE tag. In particular, for cases where it was not obvious which protein sequence ought to be used in the analysis the column is highlighted red; for cases where the transcript had been previously characterised in some way, it is highlighted light blue.

Tag sequence	NK tag count per 100,000	Ratio NK vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
ACCTGCCGACA	34.6	34.6	0.00%	0.00%	0	1) Homo sapiens cDNA FLJ14866 fis, clone PLACE1002066 / 2) Tumor suppressor deleted in oral cancer-related 1	1) 288941 / 2) 379039	1)AK021576 2)NM_005851	No defined CDS, no ORF found by ORF finder	BLAST search detects similarity to (57% identity) CDK2-associated protein 1 [Homo sapiens] (NM_004642) . The protein encoded by this gene is a specific **CDK2-associated protein** which has a **putative regulatory role in DNA replication** during S phase of the cell cycle	**CC**
AGGCCACTGGG	15.2	15.2	0.17%	0.22%	0	hypothetical protein MGC2463	323634	NM_024070		SMART predicts an **N-terminal signal peptide** and a **single TM domain**	**CS**
AGGCTCCGTGG	29.1	14.8	0.01%	0.00%	0	minor histocompatibility antigen HA-1	196914	D86976	Previously described as an antigen (see cluster name) but functionally uncharacterised	This transcript contains a **RhoGAP domain**, **a Fes/CIP4/CDC15 domain** and a protein kinase C conserved region 1 (C1) domain (also called a **phorbol ester/diacylglycerol binding domain**)	**S**
ATGACAGATGG	27.7	4.7	0.36%	0.00%	0	lung cancer-associated Y protein	13775	NM_032495	This transcript encodes a short protein - only 73 aa	BLASTp shows this transcript is **similar to Goosecoid and Pax6** from several species - these are homeobox containing proteins. All domain detection programs identify a **single homeobox domain** in this transcript	**T**
CACCCAATGGG	33.3	33.3	0.00%	0.00%	0	SEC7 homolog	110121	NM_012455		All domain detection programs identify a **Sec7-like guanine-nucleotide exchange factor (GEF) domain** and a **pleckstrin homology (PH) domain**. BLASTp shows that a large region of this protein is **similar to several GEFs of the cytohesin family** that regulate the ARF-family G-proteins	**S**

305

Tag sequence	NK tag count per 100,000	Ratio NK vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
CAGAGCTGTGC	23.6	23.6	0.01%	0.01%	0	phorbolin-like protein MDS019	250619	NM_021822		BLASTp shows this transcript is most **similar to Phorbolin-3** (68% similarity). It is also similar to Phorbolin 1 and 2. Phorbolins are upregulated in psioratic keratinocytes; they have no known function but are **thought to be involved in post-transcriptional modification of RNA**, possibly analogous to that of apobec-1 which deaminates cytosines in certain mRNAs (leaving uracil). InterProScan detects a **cytidine and deoxycytidylate deaminase zinc-binding region** (this region is found in other other-wise unrelated proteins as well as these enzymes)	**P**
CAGGTCCCTGA	8.3	8.3	2.40%	2.82%	0	Immune associated nucleotide 4 like 1 (mouse)	26194	NM_018384		SMART detects a N terminal P-loop containing nucleotide triphosphate hydrolases domain followed by a **C terminal TM domain.** BLASTp finds matches to many uncharacterised cDNAs including other members of the immune associated protein family	**CS**
CAGGTTAAGCT	9.7	9.7	1.41%	1.69%	0	Janus kinase 3 (a protein tyrosine kinase, leukocyte), tag matches mRNA which is similar to Janus kinase 3	99877	BC028068		SMART and CD search detect a **Band 4.1 domain** (this domain is found in a number of cytoskeletal-associated proteins that associate with various proteins at the interface between the plasma membrane and the cytoskeleton) followed by a **SH2 domain**, followed by a match to one quarter of a **Tyrosine kinase domain**. BLASTp finds the transcript to be **identical to the N termini of JAK3 and several other kinases**	**S**

306

Tag sequence	NK tag count per 100,000	Ratio NK vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
CCGGACCTGTG	45.7	7.7	0.00%	0.00%	0	Homo sapiens cDNA FLJ31238 fis, clone KIDNE2004864	375208	AK055800	No defined CDS, used ORF in frame +3 from 1437 to 1775 (112aa)	BLAST search finds C terminal region almost identical to other uncharacterised proteins	
CTCCTCCAAGT	29.1	29.1	0.00%	0.00%	0	Homo sapiens, clone MGC:40121 IMAGE:5216355, mRNA, complete cds	15284	BC028076		BLASTp matches one other uncharacterised sequence. SMART detects **3 TM domains**	**CS**
GCACTACTCGA	16.6	8.4	0.62%	0.13%	0	hypothetical protein FLJ22573	352548	NM_024660		SMART detects an N terminal signal peptide followed by a central transmembrane domain. BLASTp only detects matches to other uncharacterised cDNAs	**CS**
GCCCAGCCCTG	20.8	10.6	0.15%	0.03%	0	hypothetical protein 384D8_6	180903	NM_014551		BLASTp detects to other uncharacterised proteins	
GCCGGCCGGAC	24.9	24.9	0.00%	4.04%	0	B-cell CLL/lymphoma 7C	303197	NM_004765		InterProScan detects a central Bipartite nuclear localization signal (although this profile has a very high false positive rate) and an N terminal Proline rich region. BLASTp search finds matches to other uncharacterised proteins including other B-cell CLL/lymphoma 7 family members	
GGACCAGCGCC	29.1	29.1	0.00%	0.00%	0	hypothetical protein FLJ21438	136979	AK025091		BLASTp matches to other uncharacterised sequences. SMART predicts a **coiled-coil region** at the C-terminus	

307

Tag sequence	NK tag count per 100,000	Ratio NK vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
GGAGCTGTCTG	19.4	19.4	0.03%	0.05%	0	differentially expressed in FDCP (mouse homolog) 6	15476	NM_022047		BLASTp shows this protein is **similar to SWAP70** (70% identity to N terminal region and 44% identity to C-terminal region), a protein that functions in B-cell isotype switching. All domain detection programs identify a **PH domain**, an N-terminal **EF hand** and **4 tropomyosin domains** suggesting a role in regulating the cytoskeleton. InterProScan also detects two domains that are similar to the **Tubby N-terminal domain**. The function of Tubby is unknown but its disruption can lead to obesity in mice	**Cy**
GGAGCTTGAGG	23.6	23.6	0.01%	0.08%	0	chromosome 6 open reading frame 9	288316	NM_022107		BLASTp shows this protein is identical to G18 (human) and NG1 (mouse) - both have uncharacterised functions. The transcript contains a proline-rich domain and three GoLoco domains. GoLoco domains contain an LGN motif and are found in putative GEFs specific for Gα GTPase and regulators of G-protein signalling	**S**
GGAGGGAGCTG	15.2	15.2	0.17%	0.22%	0	lens intrinsic membrane protein 2 (19kD) / Homo sapiens cDNA FLJ31501 fis, clone NT2NE2005537	162754 / 311208	1) NM_030657 2) AK056063		Both SMART and CD search detect **PMP-22/EMP/MP20/Claudin family** match. BLAST finds it **identical or almost identical to other members of the lens intrinsic membrane protein**. 24 % identity to (M32231) calcium channel gamma-subunit precursor [Oryctolagus cuniculus]	**CS**

308

Tag sequence	NK tag count per 100,000	Ratio NK vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
GGGCGCCTGGC	11.1	11.1	1.41%	1.69%	0	immunity associated protein 1	159955	NM_130759		CD search finds a weak match at N terminus to the central region of a MMR_HSR1 domain (GTPase of unknown function). BLASTp search detects matches to other IMAP proteins and many uncharacterised sequences. SMART detects a C terminal single TM domain	**CS**
GGTAGAACTAA	23.6	23.6	0.01%	0.01%	0	chromosome X open reading frame 9	61469	NM_018990	SH3 and SAM domains previously identified. This transcript is encoded on chromosome X in close proximity to genes involved in various immune disorders	CD search and SMART identify an SH3 domain and all 3 detection programs find a SAM domain. BLASTp search detects matches to many other proteins (mainly uncharacterised) that contain SAM or SH3 domains	**S**
TAACCTCAGGT	20.8	20.8	0.02%	0.03%	0	KIAA0870 protein	18166	AB020677	No defined CDs, used ORF in frame +3, 165-3045	CD search and InterProScan detect an N terminal DENN (AEX-3) domain. SMART and CD search detect a **dDENN domain** (this region is always found associated with DENN) downstream of the DENN domain. Both SMART and InterProScan **WD40 repeats** in the C terminal third. BLAST search finds that the N terminus and C terminus match many uncharacterised proteins. N terminal weakly matches (~45% positives) central region of (NM_080341) Calmodulin-binding protein related to a Rab3 GDP/GTP exchange protein [Drosophila melanogaster] and N terminal of (U93181) nuclear dual-specificity phosphatase [Homo sapiens]	**S**

309

Tag sequence	NK tag count per 100,000	Ratio NK vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
TCAGTGACCAG	18.0	18.0	0.06%	0.08%	0	hypothetical protein MGC5363	1880	NM_024064		BLASTp matches other uncharacterised sequences. No conserved domains detected	
TCCGCAAGGTG	22.2	22.2	0.02%	0.03%	0	Homo sapiens mRNA; cDNA DKFZp667F1219 (from clone DKFZp667F1219)	37617	AL832302	No defined CDS, used ORF in frame +2, 245 - 2956	All domain prediction programs detect an N terminal MYSc Myosin domain (an ATPase molecular motor). SMART finds this domain, followed by an IQ domain (short calmodulin-binding motif containing conserved Ile and Gln residues) while InterProScan finds a central Bipartite nuclear localisation signal (although this profile has a very high false positive rate). BLASTp search finds this protein to be identical to "Similar to myosin Ic; Unconventional myosin from rat 4 for myosin I heavy chain" [Homo sapiens] (XM_166579) and "Unconventional myosin 1G valine form" [Homo sapiens] (AF380932) and other uncharacterised myosins	**Cy**
TGCAAGAGAGG	23.6	23.6	0.01%	0.01%	0	Homo sapiens mRNA; cDNA DKFZp667K0625 (from clone DKFZp667K0625)	238954	AL832852	No defined CDS, used ORF in frame +1, 226-3087	All domain detection programs detect an N terminal **RhoGAP domain** (GTPase-activator protein for Rho-like GTPases). BLASTp detects C terminal of protein is **identical to "Similar to retinitis pigmentosa GTPase regulator"** [Homo sapiens] (XM_170910, an protein predicted from the genome, supported by EST alignments) and that the N terminus is similar to the N terminal region of "Cdc42 GTPase-activating protein" [Mus musculus] (NM_020260)	**S**
TGGAAGCATCT	43.0	43.0	0.00%	0.00%	0	HSPC022 protein	367740	NM_014029		BLASTp matches to other uncharacterised sequences. SMART predicts an **N-terminal signal peptide**	

310

Tag sequence	NK tag count per 100,000	Ratio NK vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
TGGTTAATCTT	15.2	15.2	0.17%	0.22%	0	Homo sapiens mRNA; cDNA DKFZp667O1616 (from clone DKFZp667O1616)	365655	AL713722	No defined CDS, no ORF found by ORF finder		
TTGCTCTGCGA	13.9	13.9	0.48%	0.61%	0	ESTs from myelin basic protein, matches one cDNA (AK074315) which does not match refseq	69547	AK074315	No defined CDS, no meaningful CDS in ORF finder.		
AAATCAGGAAC	12.5	12.5	0.48%	0.61%	1	MOB-LAK	180549	NM_130807		CD search and SMART detect to the pfam **Mob1 family protein domain**. BLAST finds similarity to many uncharacterised proteins including cell cycle associated protein Mob1-2 (AY046920) [Trypanosoma brucei]	**CC**
ACCATTGGATT	41.6	41.6	0.00%	0.00%	1	interferon induced transmembrane protein 1 (9-27)	146360	NM_003641	Further investigation shows that other researchers have characterised this transcript to some extent	One of three members of a family of transmebrane (TM) proteins that are **upregulated by interferons**, with unknown function. This particular protein is 17kDa and has been **implicated in relaying antiproliferative and homotypic adhesion signals**	**CS**
ATGACTGCTGT	9.7	9.7	1.41%	1.69%	1	hypothetical protein FLJ22794	19525	NM_022074		No domains detected by any of the programs. BLASTp finds matches to only uncharacterised proteins	

311

Tag sequence	NK tag count per 100,000	Ratio NK vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
CACCACGGTGT	66.5	3.4	0.01%	0.00%	1	RNB6	241471	NM_016337	Previously published homology to murine EVL, a member of **Ena/VASP protein family** that is implicated to be involved in the control of cell motility through actin filament assembly by their GP5 motifs.	CD search detect an N terminal **WH1 domain**, this overlaps with a possible **Ran-binding domain** (domain of approximately 150 amino acids that stabilises the GTP-bound form of Ran (a Ras-like nuclear small GTPase)). BLASTp search detects high similarity to "Ena-vasodilator stimulated phosphoprotein" [Mus musculus] (NM_007965), this protein seems to be involved in actin rearrangements. InterProScan detects an N terminal **EVH1 (RanBP1-WASP) domain**, (many EVH1-containing proteins associate with actin-based structures and play a role in cytoskeletal organisation) followed by a **proline rich domain**. SMART detects an N terminal **WASP homology region 1**, function unknown	**Cy**
CAGCCTTGCGG	19.4	9.9	0.24%	0.33%	1	hypothetical protein similar to actin related protein 2/3 complex, subunit 5	315164	NM_030978		SMART finds a match to PDB Crystal structure of **arp2/3 complex**.	**Cy**
CCCTCTGTGCT	8.3	8.3	2.40%	2.82%	1	Homo sapiens cDNA FLJ38402 fis, clone FEBRA2008210	169061	AK095721	No defined CDS, two putative ones: 1) Frame +1, 1042 to 1524 2) Frame +3, 735 to 1127	SMART finds an N terminal signal peptide. InterProScan finds a central ATP/GTP-binding site motif A (P-loop)	

312

Tag sequence	NK tag count per 100,000	Ratio NK vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions	
			Cereb.	Ovary Epith.								
CCTGGAGCGCC	12.5	6.3	2.48%	0.61%	1	flightless I (Drosophila) homolog	83849	NM_002018	LLR and Gelsolin domains identifed in Refseq.	SMART detects 11 N terminal LLRs (Leucine-rich repeats (LRRs), all 3 programs detect at least **5 Gelsolin/severin/villin homology** domains. BLAST finds it highly similar to many other Flightless homologues and less similar to other Gelsolin containing proteins	**Cy**	
CTGGCCATCGA	31.9	31.9	0.00%	0.00%	1	Homo sapiens mRNA for FLJ00043 protein, partial cds	356684	AK024451	Partial CDS only	SMART detects a central "C-terminal cystine knot-like domain (CTCK)", this is followed by a **Calponin homology domain** which is predicted by all 3 programs. InterProSCAN predicts this is followed by a Proline-rich region. SMART predicts C terminal coiled coils.	**Cy**	**S**
GGAGCACACAG	11.1	11.1	1.41%	1.69%	1	Hypothetical protein FLJ31952	193490	NM_144682		SMART detects a central match to P-loop containing **nucleotide triphosphate hydrolases**. BLASTp detects an N terminal match C terminal of schlafen 3 (NM_011409) [Mus musculus], 28% identity	**O**	

313

Tag sequence	NK tag count per 100,000	Ratio NK vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
GGGGGTGGGTG	29.1	14.8	0.01%	2.17%	1	Mago-nashi homolog, proliferation-associated (Drosophila)	57904	BC018211	Published as homolog of drospophila gene. Drosophila that have mutations in their mago nashi (grandchildless) gene produce progeny with defects in germplasm assembly and germline development. This gene encodes the mammalian mago nashi homolog. In mammals, mRNA expression is not limited to the germ plasm, but is expressed ubiquitously in adult tissues and can be induced by serum stimulation of quiescent fibroblasts	All 3 prediction programs detect Mago nashi protein domain. Blast search detects many other Mago nashi like proteins and also the (U03559) mago nashi protein [Drosophila melanogaster]. 91% identical (131/143, total length 146aa)	

314

Tag sequence	NK tag count per 100,000	Ratio NK vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
GTCTGATATCT	8.3	8.3	2.40%	2.82%	1	hypothetical protein FLJ11712	14920	NM_024570*	This tag is probably generated by priming of cDNA synthesis from an internal polyA stretch in this gene. One apparently complete cDNA ends at this polyA stretch but may not be a true transcript. Thus the full sequence of the other cDNAs in the cluster were used	BLASTp matches to other uncharacterised sequences. SMART predicts an **N-terminal signal peptide**	
GTGCCACCAGT	29.1	14.8	0.01%	0.00%	1	KIAA1554 protein	17767	AB046774	The correct start codon for this transcript has not been identified, it could therefore be incomplete. This transcript also contains several polyA sites so would be expected to produce several different SAGE tags	BLASTp matches only uncharacterised sequences. All domain detection programs find a **RING finger domain** (Zn fingers involved in protein-protein interactions) near the N-terminus	

315

Tag sequence	NK tag count per 100,000	Ratio NK vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
GTTTCAAACGA	12.5	12.5	0.48%	0.61%	1	hypothetical protein MGC10966	180535	NM_031471		Both InterProScan and SMART detect a **Band 4.1 domain**. BLAST detects a 100 aa match with **Talin** (which has nearly 2500 aa) - a protein involved in the actin **cytoskeleton** (25% identity, 57% similarity)	**Cy**
TACGAGGCCGG	26.3	26.3	0.00%	0.00%	1	expressed in activated T/LAK lymphocytes	16165	NM_007267		SMART detects **7 TM domains**. BLASTp search detects matches to uncharacterised sequences, including weak matches to the family (in both mice and humans): TM, cochlear expressed	**CS**
TCCTCTTTCCA	33.3	33.3	0.00%	0.00%	1	natural killer cell transcript 4	943	NM_004221	Further investigation shows that other researchers have characterised this transcript to some extent	**This transcript is selectively expressed in lymphocytes and is up regulated on T or NK cell activation. It has a signal peptide and publications suggest it contains an RGD sequence**	
TGATCTGCCTG	20.8	20.8	0.02%	0.03%	1	hypothetical protein MGC5178	326067	NM_024044		Both CD search and SMART detect an **GIY-YIG type nucleases (URI domain)**. InterProScan finds a central proine rich region. BLAST finds matches to hypothetical proteins only	**O**
TGCCTATAGCC	11.1	11.1	1.41%	1.69%	1	hypothetical protein FLJ11029	274448	NM_018304		InterProScan find the N terminal contains a central Bipartite nuclear localization signal (although this profile has a very high false positive rate) and a central Proline rich region. BLASTp finds matches to uncharacterised proteins only	

316

Tag sequence	NK tag count per 100,000	Ratio NK vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
TGCTGCCGTGA	11.1	11.1	1.41%	1.69%	1	Hypothetical protein FLJ13491	33533	NM_024623		Both CD search and SMART detect a central **Prolyl 4-hydroxylase alpha subunit homologues**. (In CD search this overlaps with a 2OG-Fe(II) oxygenase superfamily domain). BLAST finds that the central region of protein is similar (30% identity) to C terminal region of procollagen-lysine, 2-oxoglutarate 5-dioxygenase (lysine hydroxylase, Ehlers-Danlos syndrome type VI) (BC016657) [Homo sapiens]	**O**
TGGCCTGCCCA	43.0	5.5	0.01%	0.00%	1	MLL septin-like fusion	181002	NM_006640	Publication suggest that this protein is **related to the septin family** of GTPase genes involved in cytokinesis	Both SMART and CD search detect a C terminal **GTP_CDC domain** (Found in cell division protein. Members of this family include CDC3, CDC10, CDC11 and CDC12/Septin. Members of this family bind GTP.) InterProScan detects an ATP/GTP-binding site motif A (**P-loop**). BLASTp search shows this protein is almost **identical to other members of the septin (and septin-like) family**. C terminal similar (47% identity) to cell division control protein CDC10 - yeast (Candida albicans)	**CC**
TGGTCCACGGC	19.4	19.4	0.03%	0.05%	1	HSPC037 protein	108196	NM_016095		BLAST finds matches to other uncharacterised proteins only	
TGTGGGAACCA	23.6	23.6	0.01%	0.01%	1	hypothetical protein AL133206	7750	AL133206		All 3 domain prediction programs an N terminal **ArfGAP domain.** BLASTp detects similarity to stromal membrane-associated proteins SMAP18 (44% identity, 55% similarity) and SMAP1A (40% and 50%)	**S**

317

Tag sequence	NK tag count per 100,000	Ratio NK vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	Representative Accession Number	Notes on transcript match	Summary of domain search and BLAST analysis	Possible functions
			Cereb.	Ovary Epith.							
TGTGTGTGACA	15.2	15.2	0.17%	0.22%	1	hypothetical protein FLJ14466	55148	NM_032790		SMART finds **3 TM domains**. BLAST finds the protein very highly similar to many hypothetical proteins. Highly similar (67%) to (NM_032831) CAP-binding protein complex interacting protein 2 [Homo sapiens], (function unknown)	**CS**

318

Appendix E.2: NK-specific tags that match Unigene clusters containing only EST sequences.

NK-specific tags were defined as those significantly more abundant (p value ≤ 0.05) in the NK library compared to both normal cerebellum (marked as cereb.) and ovary epithelium (marked as ovary epith.) libraries and which were not found at similar levels (more than one-third that in the NK library) in more than 1 of 12 cancer derived SAGE libraries. The EST clusters were assembled using the CAP3 programme. Where possible, full-length transcripts were identified by BLAST, and possible protein sequences were searched for conserved domains or homology to known proteins. The last column indicates possible functions of the transcripts, given the results of this analysis.

Tag sequence	NK tag count per 100 K	Ratio NK vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	CAP3 Contig	BLASTx	BLASTn	Further analysis of matches.	Possible function
			Cereb.	Ovary Epith.								
CGCCCCGGCGG	30.5	30.5	0.00%	0.00%	0	ESTs	196244	1		Matches 3' end of "Homo sapiens candidate tumour suppressor **HIC-1 (HIC-1) gene**, complete cds" (99% identity)	HIC-1 contains a **BTB/POZ homomeric dimerisation domain** and **5 Zinc fingers**. POZ domains from several zinc finger proteins have been shown to mediate transcriptional repression and to interact with components of histone deacetylase co-repressor complexes including N-CoR and SMART. BLASTp shows that it is **similar to several transcription factors / repressors**	**T**

Tag sequence	NK tag count per 100 K	Ratio NK vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	CAP3 Contig	BLASTx	BLASTn	Further analysis of matches.	Possible function
			Cereb.	Ovary Epith.								
GAACCGTCCTG	41.6	21.1	0.00%	0.00%	0	ESTs, Weakly similar to T51376 plant adhesion molecule 1 (PAM1) - Arabidopsis thaliana [A.thaliana]	123164	1		Approximately two-thirds of contig matches "**Homo sapiens, clone IMAGE:4849486, mRNA**" (BC022202)	"Homo sapiens, clone IMAGE:4849486, mRNA" has no identified amino acid sequence (used ORF in frame +2, 2 to 1342). All domain detection programs identify a central TBC domain (Probable Rab-GAP; Widespread domain present in Gyp6 and Gyp7, thereby giving rise to the notion that it performs a GTP-activator activity on Rab-like GTPases.) Similar to other uncharacterised sequences and proteins containing TBC domains	**S**

Tag sequence	NK tag count per 100 K	Ratio NK vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	CAP3 Contig	BLASTx	BLASTn	Further analysis of matches.	Possible function
			Cereb.	Ovary Epith.								
GACGGCCAGAG	15.2	15.2	0.17%	0.22%	0	ESTs, Weakly similar to extensin-like protein [Arabidopsis thaliana] [A.thaliana]	294142	1	Near 5' end of contig is identical to (BC035511) Similar to LOC124402 [Homo sapiens] (164 aa). 5' is proline rich and also matches (30%) proline rich regions of some glycoproteins.	Identical to Homo sapiens, Similar to LOC124402, clone MGC:29814 IMAGE:5091979, mRNA, complete cds. Matches chromsome 17 clone. Identical (with probable exons in our EST contig) Homo sapiens LOC124402 (LOC254110), mRNA	The CDS of Homo sapiens LOC124402 (LOC254110), mRNA is short (94aa) and contains no domains predicted by CD search, SMART or InterProScan	
GAGTCCGGGAT	8.3	8.3	2.40%	2.82%	0	ESTs, Weakly similar to TRYG_HUMAN Tryptase gamma precursor (Transmembrane tryptase) [H.sapiens]	177971	1	Central region (~3/4) of contig is almost identical to entire length of (XM_058785) similar to distal intestinal serine protease [Homo sapiens]. 67% identity to C terminal half of (NM_013921) distal intestinal serine protease [Mus musculus], similar to C terminal half of many other proteases.	Almost identical to the predicted gene, Homo sapiens similar to distal intestinal serine protease (LOC124221), mRNA. Matches clone of chromosome 16	CDS of (XM_058785) similar to distal intestinal serine protease [Homo sapiens] is predicted to contain a **Trypsin-like serine protease domain** by SMART, CD search and InterProScan. CD search suggests the CDS matches only C terminal (~1/3) of the domain, suggesting the gene may be incomplete	**O**

Tag sequence	NK tag count per 100 K	Ratio NK vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	CAP3 Contig	BLASTx	BLASTn	Further analysis of matches.	Possible function
			Cereb.	Ovary Epith.								
TCGGACCTGAG	8.3	8.3	2.40%	2.82%	0	ESTs	60137	1	No meaningful matches in the correct orientation. In the wrong orientation there is a proline rich central region	3' 90% is almost identical to chromosome 17 Pac		
TGGAACCCTGA	54.0	54.0	0.00%	0.00%	0	ESTs, Highly similar to hypothetical protein FLJ13725; KIAA1930 protein [Homo sapiens] [H.sapiens]	355944	1	Matches uncharacterised sequences. 5′ end of contig is 39% identical to central region of "CCAAT/enhancer binding protein (C/EBP), alpha; CAAT/enhancer-binding protein, DNA-binding protein" [Rattus norvegicus] (NM_012524). 5′ end is also 28% identical to central region of protein tyrosine phosphatase HD-PTP [Homo sapiens] (AB025194)	3' two-thirds of this contig matches chromosome 17 sequence		

Tag sequence	NK tag count per 100 K	Ratio NK vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	CAP3 Contig	BLASTx	BLASTn	Further analysis of matches.	Possible function
			Cereb.	Ovary Epith.								
TGTTGACTCTG	22.2	11.3	0.15%	0.03%	0	ESTs	8882	1		Identical to sequence from Chromosome 14q31. 3′ third of contig matches (82%) 3′ end of the **Bos. taurus orphan G protein-coupled receptor bRGR1 mRNA** (BTU88366). Central region of contig (290-362) also matches (87%) slightly 5′ of the 3′ end of this gene. The matches are approximately 1900 bp 3′ of the protein coding region of the gene	The human homologue of Bos. taurus orphan G protein-coupled receptor bRGR1 mRNA (BTU88366) is "GP68_HUMAN **Probable G protein-coupled receptor GPR68** (Ovarian cancer G protein-coupled receptor 1) (OGR-1)" (U48405). The sequence for this gene is identical to a region 800bp upstream of the match to this contig on chromosome 14 (in the same BAC clone), suggesting that these ESTs are from the 3' UTR for this gene. The ligand for this receptor and the effects of its signalling are unknown	**CS**

Tag sequence	NK tag count per 100 K	Ratio NK vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	CAP3 Contig	BLASTx	BLASTn	Further analysis of matches.	Possible function
			Cereb.	Ovary Epith.								
GTGATGGGGGC	22.2	22.2	0.02%	0.03%	1	EST in Homo sapiens mRNA for FLJ00067 protein, partial cds	41045	6	5' of contig matches (77%) to short (54aa) region of C terminal of (NM_138844) Munc13-4 protein [Rattus norvegicus] (function unknown). 3' region of contig matches (28% identity) central 3/4 of (AF350280) major ampullate spidroin 2 [Nephila senegalensis] (spider silk protein)	Contig almost identical to BC020376 Homo sapiens, clone IMAGE:3458340, mRNA except for what appears to be exonic region present in clone. NB match is plus/plus therefore mRNA probably wrong way round, look for CDS in negative frames. No defined CDS in clone but using ORF finder find ORF in frame -1 (790-1568). All three domain search programs find a central Protein kinase C conserved region 2 domain	No defined CDS in clone but using ORF finder finds ORF in frame -1 (790-1568). All three domain search programs find a central **Protein kinase C** conserved region 2 domain	**S**
TACTTGGTCTT	20.8	20.8	0.02%	0.21%	1	ESTs in thymopoietin	11355	2	5' third of sequence matches (45% identity) Alu subfamily SQ sequence contamination warning entry. Rest of contig has no matches	5' third matches various chromosomes		

Tag sequence	NK tag count per 100 K	Ratio NK vs. Cereb.	p-value		Number of cancers	Unigene Description	Unigene Cluster	CAP3 Contig	BLASTx	BLASTn	Further analysis of matches.	Possible function
			Cereb.	Ovary Epith.								
TGCCAATTAAG	9.7	9.7	1.41%	1.69%	1	ESTs	165337	1		Large central region of contig matches 3' region of "**Homo sapiens Ras-like GTP-binding protein (RAB27A) gene**" (98% identity)	RAB27A is a **RAB like GTPase**. All domain detection programs identify a single domain most similar to a RAB like GTPase domain, but could also be a Rac/Ras/Ran/Arf like GTPase. BLASTp find matches to many GTPase proteins	**S**

Appendix F.1: Novel transcriptional loci identified by LongSAGE and MPSS.

Both MPSS and LongSAGE libraries were made for an activated $CD4^+$ T-cell clone. To identify potentially novel region of transcription, tags from either library that matched uniquely to the genome and were not near (>5000 bases) any gene annotation were retained. If a LongSAGE tag was within 5000 bases (or five restriction sites) from a MPSS tag then both tags were kept and the region marked as a potentially novel transcriptional loci. 85 pairs were retained, of which 5 were found to be in regions predicted to be genes based on the Ensembl ESTGenes dataset.

Tag Pair ID	Tag sequence	Tag start	Tag end	Tag No	Distance between tags (bp)	No for site	Direction	Chromosome	Masked region	Tag count	Tag pair found in EST genes
1_A	CATGTCAGTGGTGGTGTTTCT	20256389	20256369	152003	3	97877	negative	1	no	5.1884	no
1_B	GATCATGTCAGTGGTGGTGT	20256392	20256373	152004	3	54127	negative	1	no	39.5603	no
2_A	CATGTTATTTGCTCGATTGTA	21247139	21247119	159630	225	102649	negative	1	no	2.5942	no
2_B	GATCTGAGGAGGAAATGTGA	21247364	21247345	159631	225	56982	negative	1	no	17.7735	no
3_A	GATCTGAGGAGGAAATGTGA	21247364	21247345	159631	381	56982	negative	1	no	17.7735	no
3_B	CATGCTTCTGCAGCCCCTTCC	21247745	21247725	159632	381	102650	negative	1	no	5.1884	no
4_A	CATGTAATAAAAGACTCCTGA	33061097	33061077	250921	2731	159050	negative	1	yes	2.5942	no
4_B	GATCCTCCCCCACAGCCCGC	33063828	33063809	250943	2731	91879	negative	1	yes	2.86669	no
5_A	CATGAAGGTTGGGTAACTCCA	67470422	67470442	509249	3318	326303	positive	1	yes	2.5942	no
5_B	GATCAGTCTTGGACCTGTGC	67473740	67473759	509267	3318	182954	positive	1	no	1.72001	no
6_A	GATCAGTCTTGGACCTGTGC	67473740	67473759	509267	4188	182954	positive	1	no	1.72001	no
6_B	CATGCAACATTCACAATCTCA	67477928	67477948	509288	4188	326325	positive	1	yes	2.5942	no
7_A	GATCACGAGAAAATTTGTTC	85169110	85169091	631199	1589	223079	negative	1	no	2.86669	no
7_B	CATGCAACAGAGAAATATTTT	85170699	85170679	631211	1589	408129	negative	1	yes	2.5942	no
8_A	CATGTTCTGGGTTATGTGCAA	110913790	110913770	810675	345	526373	negative	1	no	7.78261	no
8_B	GATCCATCTGCCAGCACATA	110914135	110914116	810677	345	284303	negative	1	no	12.0401	no
9_A	CATGTGTCTTGTTTGGTTGCT	117049201	117049181	855204	1611	555208	negative	1	yes	2.5942	no
9_B	GATCCTGCCCTGTAAGGAAG	117050812	117050793	855217	1611	300001	negative	1	no	9.74674	no
10_A	CATGTCAGTTTTGTGAAATAG	118118022	118118002	862992	386	560269	negative	1	no	2.5942	no
10_B	GATCTGACTCCAAAGCGCAT	118118408	118118389	862997	386	302725	negative	1	yes	1.14668	no
11_A	GATCTTTTGCCGCTTCTGTA	164953518	164953537	1050858	136	370496	positive	1	no	2.29335	no
11_B	CATGAATCCATTTTTGTAATT	164953654	164953674	1050860	136	680364	positive	1	no	10.3768	no
12_A	CATGTAACCAAAGTAAAAAGG	112224267	112224247	804265	643	529204	negative	10	no	2.5942	no
12_B	GATCAGGACGGAATTCCAGA	112224910	112224891	804267	643	275062	negative	10	yes	0.573338	no
13_A	GATCAGGACGGAATTCCAGA	112224910	112224891	804267	13	275062	negative	10	yes	0.573338	no
13_B	CATGAGAAAAGCAGATCAGGA	112224923	112224903	804268	13	529206	negative	10	yes	5.1884	no
14_A	GATCAGTGTGGTTTGGTGGT	112272330	112272349	804590	541	275192	positive	10	yes	1.14668	no
14_B	CATGATACGTATATTGTAATA	112272871	112272891	804593	541	529401	positive	10	yes	7.78261	no
15_A	GATCTGCAATGTCTAAGTTT	8647052	8647033	61742	3862	20942	negative	11	yes	55.6138	no
15_B	CATGTGGTGGCATACACCTGT	8650914	8650894	61770	3862	40821	negative	11	yes	2.5942	no
16_A	CATGATTTTAGTCATTTTAGT	58015767	58015747	391599	2361	259076	negative	11	yes	2.5942	no
16_B	GATCTATAGCAGACTCTTCC	58018128	58018109	391616	2361	132532	negative	11	no	10.8934	no
17_A	GATCTATAGCAGACTCTTCC	58018128	58018109	391616	28	132532	negative	11	no	10.8934	no
17_B	CATGGAAGAGGTAACTGAGAA	58018156	58018136	391617	28	259085	negative	11	yes	5.1884	no
18_A	GATCGCTTGTGCCTAGGCAT	60552487	60552468	410287	343	139001	negative	11	yes	14.3334	no
18_B	CATGGATTACAGTTTTGACAA	60552830	60552810	410290	343	271289	negative	11	no	2.5942	no
19_A	CATGGATCTCAGCCTGCAGTG	85675844	85675864	598188	4	392051	positive	11	yes	2.5942	no
19_B	GATCTCAGCCTGCAGTGTGA	85675848	85675867	598189	4	206138	positive	11	yes	45.2937	no

328

Tag Pair ID	Tag sequence	Tag start	Tag end	Tag No	Distance between tags (bp)	No for site	Direction	Chromosome	Masked region	Tag count	Tag pair found in EST genes
20_A	CATGGCAGCCGACCTCGGTTT	95551095	95551075	668561	4087	438874	negative	11	no	2.5942	no
20_B	GATCTGCCTGATACACAGTA	95555182	95555163	668585	4087	229699	negative	11	yes	22.3602	no
21_A	CATGGAAATAAAAATTAATCA	45812379	45812359	314105	33	204785	negative	12	no	2.5942	yes
21_B	GATCAAAGCTCCAGATTGCC	45812412	45812393	314106	33	109321	negative	12	no	0.573338	yes
22_A	CATGTGTTCGTGGTGCAAAAA	45815387	45815367	314130	170	204804	negative	12	yes	31.1304	yes
22_B	GATCTTCCAAGTCTCTCGAT	45815557	45815538	314132	170	109327	negative	12	no	5.16004	yes
23_A	GATCTTCCAAGTCTCTCGAT	45815557	45815538	314132	617	109327	negative	12	no	5.16004	yes
23_B	CATGATATATCTATCAAATTG	45816174	45816154	314139	617	204811	negative	12	yes	5.1884	yes
24_A	CATGTCCACGGAGCGTTTCTG	45937520	45937540	315065	20	205376	positive	12	no	2.5942	no
24_B	GATCCAGCTCCCCGTTTTTA	45937540	45937559	315066	20	109690	positive	12	no	18.9201	no
25_A	GATCCCAGGAAGTTTGGCTG	45941607	45941626	315096	267	109702	positive	12	yes	14.9068	no
25_B	CATGACCATCAGTGGATGCCA	45941874	45941894	315097	267	205395	positive	12	yes	2.5942	no
26_A	CATGACGAAAGAGCGAAACTC	48291881	48291901	332772	4858	216435	positive	12	yes	2.5942	no
26_B	GATCAGTAGATGCCAGAAGC	48296739	48296758	332826	4858	116361	positive	12	yes	2.29335	no
27_A	GATCGTGTTTACAGAGGACT	72180666	72180647	381283	966	128080	negative	13	no	40.707	yes
27_B	CATGCAAAATTCCTATTTACT	72181632	72181612	381285	966	253204	negative	13	no	25.942	yes
28_A	GATCTCGCCATTGTGGTCCA	73097774	73097793	392643	3982	135547	positive	14	yes	0.573338	no
28_B	CATGTGAGAGACAGTTAAAGC	73101756	73101776	392672	3982	257113	positive	14	yes	2.5942	no
29_A	CATGACAAGGATAATGCAGGA	99569061	99569041	589470	2763	387058	negative	14	no	7.78261	no
29_B	GATCCTAGCAGACAAATAGG	99571824	99571805	589489	2763	202419	negative	14	no	0.573338	no
30_A	GATCTCTTTGTGCCAGGCCA	56940781	56940800	279983	453	97465	positive	15	no	7.45339	no
30_B	CATGTTTGTTATTGTACCACA	56941234	56941254	279984	453	182519	positive	15	no	2.5942	no
31_A	GATCCTCATAGGCACTTCTA	62223048	62223067	318797	242	111077	positive	15	no	8.60006	no
31_B	CATGATATTCACATTTATATA	62223290	62223310	318798	242	207721	positive	15	no	2.5942	no
32_A	CATGTGATTCTATTAGGCAAT	56200724	56200704	361900	245	233169	negative	16	no	2.5942	no
32_B	GATCTCAGCAATTTGGGGCT	56200969	56200950	361902	245	128732	negative	16	no	16.6268	no
33_A	CATGTGTCCAAGAATAATCCA	24901085	24901105	188196	237	117324	positive	17	no	2.5942	no
33_B	GATCCTTCTCTGCTAGAACA	24901322	24901341	188197	237	70873	positive	17	no	10.8934	no
34_A	GATCGACACATAGTGGGCGC	34112615	34112634	257134	618	96240	positive	17	no	11.4668	no
34_B	CATGATTTAGGACTGACATTT	34113233	34113253	257135	618	160895	positive	17	no	7.78261	no
35_A	CATGTCATAATATACCTGCCT	40650153	40650133	307756	502	191323	negative	17	yes	25.942	no
35_B	GATCTCTTGCCAATCAGCCT	40650655	40650636	307760	502	116436	negative	17	yes	5.16004	no
36_A	GATCGCTAAAACCCAGGAGT	40651354	40651335	307768	1052	116440	negative	17	yes	34.9736	no
36_B	CATGCTGATCAGAACTGTGAG	40652406	40652386	307778	1052	191331	negative	17	no	7.78261	no
37_A	GATCACCCCAGGGAATGTTG	53529953	53529972	402927	256	150550	positive	17	yes	17.2001	no
37_B	CATGTTGTTTTTCTGTGTATC	53530209	53530229	402930	256	252380	positive	17	yes	2.5942	no
38_A	GATCCCAGGCCACGAACTGT	40513210	40513229	281203	424	94790	positive	18	no	7.45339	no
38_B	CATGGAGCAGCAGCGGCGGCG	40513634	40513654	281206	424	186414	positive	18	yes	2.5942	no

329

Tag Pair ID	Tag sequence	Tag start	Tag end	Tag No	Distance between tags (bp)	No for site	Direction	Chromosome	Masked region	Tag count	Tag pair found in EST genes
39_A	CATGCAAATTTCCAGCAAACC	6530825	6530805	49532	651	28560	negative	19	no	5.1884	no
39_B	GATCACCCGCCTCGGCTTGG	6531476	6531457	49539	651	20974	negative	19	yes	7.45339	no
40_A	CATGTGTTCTTGCACTGATGT	46734857	46734837	305633	601	187954	negative	19	yes	12.971	no
40_B	GATCCCAGACGAGCCACCAG	46735458	46735439	305637	601	117680	negative	19	no	1.72001	no
41_A	GATCATCCTGGCCTGCTAAG	46738213	46738232	305655	136	117686	positive	19	yes	81.4139	no
41_B	CATGACAACAGGCACCAGCAC	46738349	46738369	305657	136	187971	positive	19	yes	5.1884	no
42_A	CATGGCTGTTCCTGGAGTTCT	46738915	46738935	305659	57	187973	positive	19	yes	103.768	no
42_B	GATCACTGTATGTTATGGAT	46738972	46738991	305660	57	117687	positive	19	yes	1.14668	no
43_A	GATCACTGTATGTTATGGAT	46738972	46738991	305660	656	117687	positive	19	yes	1.14668	no
43_B	CATGTCTTTTTATAACTTTTT	46739628	46739648	305663	656	187976	positive	19	yes	2.5942	no
44_A	GATCTCAGCAATGCCCATGG	46749220	46749239	305725	15	117708	positive	19	no	112.948	yes
44_B	CATGGGGCAATTTACATCGGG	46749235	46749255	305726	15	188018	positive	19	no	179	yes
45_A	GATCAGTAACCCTGTACAGG	46751915	46751934	305744	716	117713	positive	19	yes	8.02673	yes
45_B	CATGACAATTATAATTATGAC	46752631	46752651	305749	716	188034	positive	19	yes	2.5942	yes
46_A	GATCCCTAGTCCACAGAAAT	7851080	7851099	57996	470	17594	positive	2	yes	1.14668	no
46_B	CATGCATAAATAAACCTTTAT	7851550	7851570	57997	470	40403	positive	2	no	5.1884	no
47_A	GATCAAGGCAAATTGTGGGA	8374562	8374543	61919	3543	18796	negative	2	no	1.72001	no
47_B	CATGGAGGGTTATCCAGCACC	8378105	8378085	61957	3543	43152	negative	2	no	2.5942	no
48_A	GATCATCTGCTGTCTGCCAG	148616444	148616463	1043417	182	350577	positive	2	no	22.9335	no
48_B	CATGTTTAAGAAAGATATTGC	148616626	148616646	1043418	182	692841	positive	2	no	2.5942	no
49_A	CATGGCTGCTTAATATTCCAC	175365265	175365245	1229999	1386	815886	negative	2	yes	2.5942	no
49_B	GATCACTCAAGCCTAGGGAA	175366651	175366632	1230014	1386	414120	negative	2	yes	14.3334	no
50_A	GATCACTCAAGCCTAGGGAA	175366651	175366632	1230014	863	414120	negative	2	yes	14.3334	no
50_B	CATGAAAGCCAAACAGATCTT	175367514	175367494	1230021	863	815897	negative	2	no	2.5942	no
51_A	GATCACCCTTAGTCATCACT	179205000	179205019	1257010	1214	423340	positive	2	no	0.573338	no
51_B	CATGAATTTTTTTGGCCTTTT	179206214	179206234	1257014	1214	833674	positive	2	no	2.5942	no
52_A	GATCCATGGTGTTAGAAGCC	3802882	3802901	29124	4	10511	positive	20	yes	1.14668	no
52_B	CATGGTGTTAGAAGCCAGGGG	3802886	3802906	29125	4	18614	positive	20	yes	2.5942	no
53_A	GATCCCATTAGACCAGGGGC	30402541	30402560	207260	4596	69379	positive	20	no	3.44003	no
53_B	CATGGTTTGACCCAGCTGAAG	30407137	30407157	207300	4596	137911	positive	20	no	2.5942	no
54_A	GATCCCAAGGGAGGCAGTGC	47644602	47644621	343106	1690	118283	positive	20	no	10.3201	no
54_B	CATGTGGGAACTGCCTGTGTG	47646292	47646312	343116	1690	224831	positive	20	no	2.5942	no
55_A	CATGCTGGATACCTGAGGACT	58332937	58332957	424996	1598	278747	positive	20	no	2.5942	no
55_B	GATCTTGTGGACATCTCGCT	58334535	58334554	425010	1598	146254	positive	20	no	2.86669	no
56_A	CATGGATTTCTATTTGTTTTT	79470596	79470576	547352	347	365718	negative	5	no	38.913	yes
56_B	GATCCTGGGAAGTTCTCTCA	79470943	79470924	547353	347	181635	negative	5	no	55.6138	yes
57_A	CATGTAGTCTTTGTGGGCTCA	118720751	118720771	815595	2188	544996	positive	5	no	2.5942	no
57_B	GATCCCCTGTCGCCCTCTCT	118722939	118722958	815603	2188	270603	positive	5	no	4.01336	no

330

Tag Pair ID	Tag sequence	Tag start	Tag end	Tag No	Distance between tags (bp)	No for site	Direction	Chromosome	Masked region	Tag count	Tag pair found in EST genes
58_A	GATCCCCTGTCGCCCTCTCT	118722939	118722958	815603	3544	270603	positive	5	no	4.01336	no
58_B	CATGAGTTGGTCTGGAGGGCA	118726483	118726503	815626	3544	545014	positive	5	no	2.5942	no
59_A	GATCCAGGCCTGATGAGAGC	118729657	118729676	815644	2812	270615	positive	5	no	2.86669	no
59_B	CATGATGGGAGCAGGCCTGAG	118732469	118732489	815660	2812	545045	positive	5	no	2.5942	no
60_A	GATCAGGATTCCGCAAGCCT	141902249	141902268	983578	1412	328435	positive	5	no	5.16004	no
60_B	CATGGTTCTAAGTAAATGATT	141903661	141903681	983585	1412	655150	positive	5	yes	2.5942	no
61_A	CATGTATGTGTTCTCTCTCCC	151170993	151171013	1051798	50	700176	positive	5	yes	20.7536	no
61_B	GATCTGTGTCTGAGTCATCT	151171043	151171062	1051799	50	351623	positive	5	yes	60.7738	no
62_A	GATCAGGACACGAGGAAGAG	30367483	30367464	222500	19	77445	negative	6	no	47.587	yes
62_B	CATGTGGACAAGTCAGCAGGA	30367502	30367482	222501	19	145056	negative	6	no	15.5652	yes
63_A	GATCTCTCTACCCAAGTTGT	32977232	32977251	242154	265	84960	positive	6	yes	1.72001	no
63_B	CATGTCATACCTTTGTCAAAA	32977497	32977517	242157	265	157197	positive	6	yes	2.5942	no
64_A	CATGAAACTGCAAGTGGTTCT	115429456	115429476	799242	187	526023	positive	6	yes	7.78261	no
64_B	GATCATTGCCCAACTTCTCA	115429643	115429662	799244	187	273221	positive	6	yes	108.934	no
65_A	GATCCCTGCTGCTTGTTGTT	136002609	136002590	941941	318	320936	negative	6	yes	3.44003	no
65_B	CATGTGAAATACTGATTGCAG	136002927	136002907	941942	318	621006	negative	6	no	2.5942	no
66_A	GATCAAAAACGACGGCTGGG	149680553	149680534	1040330	1764	354218	negative	6	no	9.1734	no
66_B	CATGTATCTAACAATTCTTCC	149682317	149682297	1040340	1764	686117	negative	6	no	2.5942	no
67_A	CATGGTTTTGCACGTTTCGGT	44690271	44690251	319791	644	210194	negative	7	no	5.1884	no
67_B	GATCAGGCTGGTTGGAAGAA	44690915	44690896	319799	644	109600	negative	7	no	3.44003	no
68_A	GATCAGGCTGGTTGGAAGAA	44690915	44690896	319799	134	109600	negative	7	no	3.44003	no
68_B	CATGAGTTTGTCCTTTTTTCC	44691049	44691029	319800	134	210200	negative	7	no	2.5942	no
69_A	CATGAAATCGAGCATTTTTTT	130088642	130088622	914870	29	598928	negative	7	yes	2.5942	no
69_B	GATCGGTGAAATGAATTTCG	130088671	130088652	914871	29	315943	negative	7	no	36.1203	no
70_A	GATCGGTGAAATGAATTTCG	130088671	130088652	914871	1350	315943	negative	7	no	36.1203	no
70_B	CATGGACTCTTGCCTTACTCT	130090021	130090001	914882	1350	598936	negative	7	no	2.5942	no
71_A	CATGTTGTGCTAATTCCCTTT	148272436	148272456	1049601	3055	686659	positive	7	yes	2.5942	no
71_B	GATCCCTGGCTGTGCGTGGT	148275491	148275510	1049624	3055	362956	positive	7	yes	1.14668	no
72_A	CATGGATCAGTACTAAAAGGC	79894272	79894292	558472	2582	369530	positive	8	yes	2.5942	no
72_B	GATCTGTTTCTGGGCCATTT	79896854	79896873	558488	2582	188950	positive	8	no	0.573338	no
73_A	CATGAATTTTTAATAAAACGT	101777196	101777176	712148	84	471364	negative	8	no	2.5942	no
73_B	GATCATTTTGCTGAGCTTGT	101777280	101777261	712149	84	240785	negative	8	no	1.14668	no
74_A	GATCAGCTGCCAGAAATTGC	129002524	129002543	906490	4106	305571	positive	8	no	3.44003	no
74_B	CATGTCATTTCTCCTTAGGGC	129006630	129006650	906526	4106	600939	positive	8	yes	2.5942	no
75_A	CATGTAGCATTTTGTTTTGCT	141600341	141600321	1002565	170	666352	negative	8	yes	23.3478	no
75_B	GATCAGCATTCTTGCACTTT	141600511	141600492	1002566	170	336214	negative	8	no	13.1868	no
76_A	GATCAGCATTCTTGCACTTT	141600511	141600492	1002566	379	336214	negative	8	no	13.1868	no
76_B	CATGCTTTTATATTTGGCAGT	141600890	141600870	1002571	379	666353	negative	8	no	5.1884	no

331

Tag Pair ID	Tag sequence	Tag start	Tag end	Tag No	Distance between tags (bp)	No for site	Direction	Chromosome	Masked region	Tag count	Tag pair found in EST genes
77_A	CATGGAATTTGATAGACTTTT	141601368	141601348	1002574	237	666356	negative	8	no	28.5362	no
77_B	GATCTACCTCAGTTAAACAG	141601605	141601586	1002576	237	336219	negative	8	no	3.44003	no
78_A	GATCTACCTCAGTTAAACAG	141601605	141601586	1002576	133	336219	negative	8	no	3.44003	no
78_B	CATGTTCTAAGCTTTCTTCAG	141601738	141601718	1002577	133	666358	negative	8	no	2.5942	no
79_A	CATGTACCAAGCCAGCTATAA	69554107	69554087	351588	353	231507	negative	9	no	23.3478	no
79_B	GATCAGCCTGTTCCAATTTG	69554460	69554441	351593	353	120082	negative	9	no	29.2402	no
80_A	CATGACTTCGATGCTGGCTGC	69555739	69555719	351598	662	231515	negative	9	no	5.1884	no
80_B	GATCTAACATCAAACTAACA	69556401	69556382	351601	662	120084	negative	9	no	0.573338	no
81_A	CATGACAACTGATCCAATTAC	90890659	90890679	507924	10	334942	positive	9	yes	2.5942	no
81_B	GATCCAATTACCCAGCCAGC	90890669	90890688	507925	10	172983	positive	9	yes	4.5867	no
82_A	GATCATTGTTGAGATGAGGA	72824414	72824395	509597	18	173857	negative	X	no	112.948	no
82_B	CATGCAGTTCCAATAGCTGAT	72824432	72824412	509598	18	335741	negative	X	no	15.5652	no
83_A	CATGCGTTGTAGAGTGGGAAT	72825682	72825662	509610	12	335752	negative	X	no	2.5942	yes
83_B	GATCACCTTTGACATGCGTT	72825694	72825675	509611	12	173859	negative	X	no	30.9602	yes
84_A	GATCACTGACAGCTGAAGAT	72826323	72826304	509612	245	173860	negative	X	no	0.573338	yes
84_B	CATGCTTCCTGGCCTGATTGA	72826568	72826548	509613	245	335753	negative	X	no	28.5362	yes
85_A	CATGAGGGACTCATATTCCTA	72997659	72997679	510942	3272	336589	positive	X	no	2.5942	no
85_B	GATCAGTATACCCTCTAGCT	73000931	73000950	510965	3272	174361	positive	X	yes	1.14668	no

332

Appendix F.2: Novel transcriptional loci identified by LongSAGE.

A LongSAGE library was made for an activated $CD4^+$ T-cell clone. LongSAGE can be used to profile sense and antisense transcription. To identify potentially novel regions of transcription, sites where tags that matched uniquely to the genome were found in both the positive and negative direction and were not near (>5000 bases) any gene annotation were retained. 40 pairs were retained, of which 8 were identified using a combination of LongSAGE and MPSS data (see Appendix F.1).

Tag pair ID	Tag sequence	Tag start	Tag end	Dist between tags (bp)	Tag No	Direction	Chromosome	Masked regiono	Tag Count	Tag pair found in EST genes	Tag pair also in MPSS/SAGE novel loci
1_A	CATGATGAAATGTTATTCTCA	21246976	21246996	3	159628	positive	1	no	2.5942	no	no
1_B	CATGCATCACCTCCTCCCTAC	21246979	21246959	3	159628	negative	1	no	2.5942	no	no
2_A	CATGACCTGACGTTTTCTTTT	60553067	60553087	3	410291	positive	11	yes	5.1884	no	no
2_B	CATGCCTTGCCTTCACTCGGG	60553070	60553050	3	410291	negative	11	yes	116.739	no	no
3_A	CATGTGTCACTAAGTGAAGTT	113631716	113631736	3	796413	positive	11	no	10.3768	no	no
3_B	CATGGAGAGAGAAAAAAACCT	113631719	113631699	3	796413	negative	11	no	5.1884	no	no
4_A	CATGGACGGAATAAAGGATTT	113632669	113632689	3	796418	positive	11	no	7.78261	no	no
4_B	CATGACAATGCAGAGCTTGCA	113632672	113632652	3	796418	negative	11	no	2.5942	no	no
5_A	CATGCACTTTATTGAATGACA	45815384	45815404	3	314130	positive	12	no	5.1884	yes	yes
5_B	CATGTGTTCGTGGTGCAAAAA	45815387	45815367	3	314130	negative	12	yes	31.1304	yes	yes
6_A	CATGCTGAGGCTCTTTTTAAG	68922450	68922470	3	484408	positive	12	no	2.5942	no	no
6_B	CATGTGTATTTTGGAATTGTG	68922453	68922433	3	484408	negative	12	no	10.3768	no	no
7_A	CATGGGCTGGTCTTGACGAGA	49428856	49428876	3	218930	positive	14	no	5.1884	no	no
7_B	CATGTTTCTTAAAAACTGCAG	49428859	49428839	3	218930	negative	14	no	10.3768	no	no
8_A	CATGCTCCGTGACATCCTCGG	78197891	78197911	3	528821	positive	16	no	18.1594	no	no
8_B	CATGTCTGGTGTTTTCAGTGC	78197894	78197874	3	528821	negative	16	no	2.5942	no	no
9_A	CATGATTTAGGACTGACATTT	34113233	34113253	3	257135	positive	17	no	7.78261	no	yes
9_B	CATGCACGAGTTTCAGAATAT	34113236	34113216	3	257135	negative	17	no	2.5942	no	yes
10_A	CATGCCAGTTAGAGACTTTTT	34663633	34663653	3	261466	positive	17	no	2.5942	no	no
10_B	CATGATTTGATTACTTTTTTC	34663636	34663616	3	261466	negative	17	no	2.5942	no	no
11_A	CATGAACAGCTGGTCCCATAG	40652661	40652681	3	307779	positive	17	no	5.1884	no	no
11_B	CATGCTTCCTCTTGCCCAGAA	40652664	40652644	3	307779	negative	17	no	10.3768	no	no
12_A	CATGAATCTTTGAAAACAAAA	62497085	62497105	3	471527	positive	17	no	2.5942	no	no
12_B	CATGGTAACAAAATGTATTTT	62497088	62497068	3	471527	negative	17	no	25.942	no	no
13_A	CATGGAGTCTTCTAATTCTCA	65671710	65671730	3	462788	positive	18	no	2.5942	no	no
13_B	CATGGGTAAGAGGAGCCACAG	65671713	65671693	3	462788	negative	18	no	5.1884	no	no
14_A	CATGCGTGAGATAAGAAATTT	46734261	46734281	3	305627	positive	19	no	5.1884	no	no
14_B	CATGACAAAATCAGCCGCAGT	46734264	46734244	3	305627	negative	19	no	7.78261	no	no
15_A	CATGGGTGGCTGCCACTTCTC	46734854	46734874	3	305633	positive	19	yes	181.594	no	yes
15_B	CATGTGTTCTTGCACTGATGT	46734857	46734837	3	305633	negative	19	yes	12.971	no	yes
16_A	CATGACTCTTGTTTTCATCTG	46738817	46738837	3	305658	positive	19	no	20.7536	no	no
16_B	CATGATGAAAATTACCCAGCA	46738820	46738800	3	305658	negative	19	no	5.1884	no	no
17_A	CATGCTGCAGTGATGATCTTG	46749198	46749218	3	305723	positive	19	no	2.5942	yes	no
17_B	CATGCAGCACAATAGTAAATT	46749201	46749181	3	305723	negative	19	no	2.5942	yes	no
18_A	CATGGGGCAATTTACATCGGG	46749235	46749255	3	305726	positive	19	no	179	yes	yes
18_B	CATGGGCATTGCTGAGATCAC	46749238	46749218	3	305726	negative	19	no	5.1884	yes	yes
19_A	CATGTGTTTCCTTTACCTATA	46764494	46764514	3	305833	positive	19	no	2.5942	no	no
19_B	CATGCATCCTCACAGAGCTAC	46764497	46764477	3	305833	negative	19	no	2.5942	no	no

Tag pair ID	Tag sequence	Tag start	Tag end	Dist between tags (bp)	Tag No	Direction	Chromosome	Masked regiono	Tag Count	Tag pair found in EST genes	Tag pair also in MPSS/SAGE novel loci
20_A	CATGTTGCTTCTTCAACAAAT	61069701	61069721	3	442918	positive	2	no	31.1304	no	no
20_B	CATGCTGTACTAATTACTGTG	61069704	61069684	3	442918	negative	2	no	5.1884	no	no
21_A	CATGTGAAATGTAAGAGATGT	127774352	127774372	3	896030	positive	2	no	2.5942	no	no
21_B	CATGGTTCTTTTTATTCATTG	127774355	127774335	3	896030	negative	2	no	7.78261	no	no
22_A	CATGTTTAAGAAAGATATTGC	148616626	148616646	3	1043418	positive	2	no	2.5942	no	yes
22_B	CATGCATTTTTCCACCTAACA	148616629	148616609	3	1043418	negative	2	no	5.1884	no	yes
23_A	CATGTTGGGGCGGTTTTTGGG	157024188	157024208	3	1101567	positive	2	no	2.5942	no	no
23_B	CATGTCCAGAGAGGGGCATCT	157024191	157024171	3	1101567	negative	2	no	2.5942	no	no
24_A	CATGAATACATTTACATTTAA	40483478	40483498	3	292452	positive	3	no	7.78261	yes	no
24_B	CATGATTTAGCCTAGATTAAA	40483481	40483461	3	292452	negative	3	yes	2.5942	yes	no
25_A	CATGCTAAAGGGAGTCATTCA	104378681	104378701	3	718897	positive	3	yes	2.5942	no	no
25_B	CATGTCTTGTAGGATAGGCCT	104378684	104378664	3	718897	negative	3	yes	2.5942	no	no
26_A	CATGTTTTAAATAAAGTTTTT	112402952	112402972	3	773736	positive	3	no	23.3478	no	no
26_B	CATGTAAAGAATCATATAGTT	112402955	112402935	3	773736	negative	3	no	7.78261	no	no
27_A	CATGGTACCTTCACGGTGTTT	161695534	161695554	3	1126798	positive	3	no	2.5942	no	no
27_B	CATGAGTTCTATGGGTACTCA	161695537	161695517	3	1126798	negative	3	no	5.1884	no	no
28_A	CATGTTTTTGTTTTGTTTCTT	192598937	192598957	3	1342467	positive	3	no	2.5942	no	no
28_B	CATGAGACAGAGACCACTGAC	192598940	192598920	3	1342467	negative	3	no	5.1884	no	no
29_A	CATGTATTTATAAAAAAATGC	197562898	197562918	3	1376835	positive	3	no	2.5942	no	no
29_B	CATGTAAATGTCCTGTTGTAA	197562901	197562881	3	1376835	negative	3	no	23.3478	no	no
30_A	CATGCAGAGAATATATATTGT	72245855	72245875	3	494949	positive	5	no	49.2898	no	no
30_B	CATGATCCAAAATATTTGGTG	72245858	72245838	3	494949	negative	5	no	2.5942	no	no
31_A	CATGACTAAAGGGGGAAAATA	79470593	79470613	3	547352	positive	5	no	5.1884	yes	yes
31_B	CATGGATTTCTATTTGTTTTT	79470596	79470576	3	547352	negative	5	no	38.913	yes	yes
32_A	CATGTGTGTGCGTGCTGGGGA	90638786	90638806	3	623862	positive	5	yes	2.5942	no	no
32_B	CATGCCAGAGTCCAACACCAA	90638789	90638769	3	623862	negative	5	yes	2.5942	no	no
33_A	CATGGCAGGTGAAAAGATAAT	133964789	133964809	3	924465	positive	5	no	2.5942	no	no
33_B	CATGAATTTGAGTGTTAGGAA	133964792	133964772	3	924465	negative	5	no	12.971	no	no
34_A	CATGGCGACAGCTCCACAGAA	43032442	43032462	3	318107	positive	6	no	2.5942	no	no
34_B	CATGAAGGTCGAGCTGTGCAG	43032445	43032425	3	318107	negative	6	no	223.101	no	no
35_A	CATGTGACAATACGACGGTGG	135543396	135543416	3	938778	positive	6	no	2.5942	no	no
35_B	CATGTCCTTCACTCTCACCCC	135543399	135543379	3	938778	negative	6	yes	2.5942	no	no
36_A	CATGGCTGCAACTGTTTTTTT	135998483	135998503	3	941920	positive	6	yes	5.1884	no	no
36_B	CATGTTGTTGACATTTTGAAT	135998486	135998466	3	941920	negative	6	no	33.7246	no	no
37_A	CATGCCAATTTGCAACCCCCA	135998651	135998671	3	941921	positive	6	no	7.78261	no	no
37_B	CATGGAAATTTAAACAGGTTC	135998654	135998634	3	941921	negative	6	no	10.3768	no	no
38_A	CATGTAATTTCAAACTTGGTT	136000292	136000312	3	941929	positive	6	yes	2.5942	no	no
38_B	CATGTGTAAATTTCATTCTTT	136000295	136000275	3	941929	negative	6	yes	2.5942	no	no

Tag pair ID	Tag sequence	Tag start	Tag end	Dist between tags (bp)	Tag No	Direction	Chromosome	Masked regiono	Tag Count	Tag pair found in EST genes	Tag pair also in MPSS/SAGE novel loci
39_A	CATGTAAAAGCTAAAGAAAGA	141600338	141600358	3	1002565	positive	8	yes	5.1884	no	yes
39_B	CATGTAGCATTTTGTTTTGCT	141600341	141600321	3	1002565	negative	8	yes	23.3478	no	yes
40_A	CATGTCAAAGGTGATCTGTTT	72825679	72825699	3	509610	positive	X	no	5.1884	yes	yes
40_B	CATGCGTTGTAGAGTGGGAAT	72825682	72825662	3	509610	negative	X	no	2.5942	yes	yes

IX. Bibliography

1. Rothenberg, E. V., Telfer, J. C. & Anderson, M. K. Transcriptional regulation of lymphocyte lineage commitment. Bioessays 21, 726-42 (1999).
2. Barclay, A. N. et al. The Leucocyte Antigen Factsbook, Second Edition (Academic Press, London, 1997).
3. Davis, M. M. et al. Dynamics of cell surface molecules during T cell recognition. Annu Rev Biochem 72, 717-42 (2003).
4. Zola, H. & Swart, B. The human leucocyte differentiation antigens (HLDA) workshops: the evolving role of antibodies in research, diagnosis and therapy. Cell Res 15, 691-4 (2005).
5. Zola, H. et al. CD molecules 2005: human cell differentiation molecules. Blood 106, 3123-6 (2005).
6. Garboczi, D. N. et al. Structure of the complex between human T-cell receptor, viral peptide and HLA-A2. Nature 384, 134-41 (1996).
7. Garcia, K. C. et al. An alphabeta T cell receptor structure at 2.5 A and its orientation in the TCR-MHC complex. Science 274, 209-19 (1996).
8. Chan, A. C., Desai, D. M. & Weiss, A. The role of protein tyrosine kinases and protein tyrosine phosphatases in T cell antigen receptor signal transduction. Annu Rev Immunol 12, 555-92 (1994).
9. Davis, S. J. & van der Merwe, P. A. The structure and ligand interactions of CD2: implications for T-cell function. Immunol Today 17, 177-87 (1996).
10. Parnes, J. R. Molecular biology and function of CD4 and CD8. Adv Immunol 44, 265-311 (1989).
11. Collins, T. L. et al. p56lck association with CD4 is required for the interaction between CD4 and the TCR/CD3 complex and for optimal antigen stimulation. J Immunol 148, 2159-62 (1992).
12. Brossard, C., Semichon, M., Trautmann, A. & Bismuth, G. CD5 inhibits signaling at the immunological synapse without impairing its formation. J Immunol 170, 4623-9 (2003).
13. Tarakhovsky, A. et al. A role for CD5 in TCR-mediated signal transduction and thymocyte selection. Science 269, 535-7 (1995).
14. Robinson, W. H., Neuman de Vegvar, H. E., Prohaska, S. S., Rhee, J. W. & Parnes, J. R. Human CD6 possesses a large, alternatively spliced cytoplasmic domain. Eur J Immunol 25, 2765-9 (1995).
15. Gimferrer, I. et al. Relevance of CD6-mediated interactions in T cell activation and proliferation. J Immunol 173, 2262-70 (2004).
16. Hassan, N. J., Barclay, A. N. & Brown, M. H. Frontline: Optimal T cell activation requires the engagement of CD6 and CD166. Eur J Immunol 34, 930-40 (2004).
17. Stillwell, R. & Bierer, B. E. T cell signal transduction and the role of CD7 in costimulation. Immunol Res 24, 31-52 (2001).
18. Subrahmanyam, G., Rudd, C. E. & Schneider, H. Association of T cell antigen CD7 with type II phosphatidylinositol-4 kinase, a key component in pathways of inositol phosphate turnover. Eur J Immunol 33, 46-52 (2003).
19. Aandahl, E. M. et al. CD7 is a differentiation marker that identifies multiple CD8 T cell effector subsets. J Immunol 170, 2349-55 (2003).
20. Zamoyska, R. The CD8 coreceptor revisited: one chain good, two chains better. Immunity 1, 243-6 (1994).

21. Binnerts, M. E., van Kooyk, Y., Simmons, D. L. & Figdor, C. G. Distinct binding of T lymphocytes to ICAM-1, -2 or -3 upon activation of LFA-1. Eur J Immunol 24, 2155-60 (1994).
22. Lub, M., van Kooyk, Y. & Figdor, C. G. Ins and outs of LFA-1. Immunol Today 16, 479-83 (1995).
23. Wacholtz, M. C., Patel, S. S. & Lipsky, P. E. Leukocyte function-associated antigen 1 is an activation molecule for human T cells. J Exp Med 170, 431-48 (1989).
24. Manjunath, N., Correa, M., Ardman, M. & Ardman, B. Negative regulation of T-cell adhesion and activation by CD43. Nature 377, 535-8 (1995).
25. Cruz-Munoz, M. E. et al. The CD43 coreceptor molecule recruits the zeta-chain as part of its signaling pathway. J Immunol 171, 1901-8 (2003).
26. Sperling, A. I. et al. CD43 is a murine T cell costimulatory receptor that functions independently of CD28. J Exp Med 182, 139-46 (1995).
27. Hall, L. R., Streuli, M., Schlossman, S. F. & Saito, H. Complete exon-intron organization of the human leukocyte common antigen (CD45) gene. J Immunol 141, 2781-7 (1988).
28. Pingel, J. T. & Thomas, M. L. Evidence that the leukocyte-common antigen is required for antigen-induced T lymphocyte proliferation. Cell 58, 1055-65 (1989).
29. Ostergaard, H. L. et al. Expression of CD45 alters phosphorylation of the lck-encoded tyrosine protein kinase in murine lymphoma T-cell lines. Proc Natl Acad Sci U S A 86, 8959-63 (1989).
30. Turka, L. A., Kanner, S. B., Schieven, G. L., Thompson, C. B. & Ledbetter, J. A. CD45 modulates T cell receptor/CD3-induced activation of human thymocytes via regulation of tyrosine phosphorylation. Eur J Immunol 22, 551-7 (1992).
31. Johnson, K. G., Bromley, S. K., Dustin, M. L. & Thomas, M. L. A supramolecular basis for CD45 tyrosine phosphatase regulation in sustained T cell activation. Proc Natl Acad Sci U S A 97, 10138-43 (2000).
32. Leupin, O., Zaru, R., Laroche, T., Muller, S. & Valitutti, S. Exclusion of CD45 from the T-cell receptor signaling area in antigen-stimulated T lymphocytes. Curr Biol 10, 277-80 (2000).
33. Wright, M. D. & Tomlinson, M. G. The ins and outs of the transmembrane 4 superfamily. Immunol Today 15, 588-94 (1994).
34. Lagaudriere-Gesbert, C. et al. Functional analysis of four tetraspans, CD9, CD53, CD81, and CD82, suggests a common role in costimulation, cell adhesion, and migration: only CD9 upregulates HB-EGF activity. Cell Immunol 182, 105-12 (1997).
35. Chen, C. P., Kernytsky, A. & Rost, B. Transmembrane helix predictions revisited. Protein Sci 11, 2774-91 (2002).
36. Moller, S., Croning, M. D. & Apweiler, R. Evaluation of methods for the prediction of membrane spanning regions. Bioinformatics 17, 646-53 (2001).
37. Labeit, S. & Kolmerer, B. Titins: giant proteins in charge of muscle ultrastructure and elasticity. Science 270, 293-6 (1995).
38. Schultz, J., Copley, R. R., Doerks, T., Ponting, C. P. & Bork, P. SMART: a web-based tool for the study of genetically mobile domains. Nucleic Acids Res 28, 231-4 (2000).

39. Sonnhammer, E. L., Eddy, S. R. & Durbin, R. Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins 28, 405-20 (1997).
40. Eddy, S. R. Hidden Markov models. Curr Opin Struct Biol 6, 361-5 (1996).
41. Eddy, S. R. Profile hidden Markov models. Bioinformatics 14, 755-63 (1998).
42. Eddy, S. R. What is a hidden Markov model? Nat Biotechnol 22, 1315-6 (2004).
43. Apweiler, R. et al. InterPro--an integrated documentation resource for protein families, domains and functional sites. Bioinformatics 16, 1145-50 (2000).
44. Barclay, A. N. Membrane proteins with immunoglobulin-like domains--a master superfamily of interaction molecules. Semin Immunol 15, 215-23 (2003).
45. Harpaz, Y. & Chothia, C. Many of the immunoglobulin superfamily domains in cell adhesion molecules and surface receptors belong to a new structural set which is close to that containing variable domains. J Mol Biol 238, 528-39 (1994).
46. Smith, D. K. & Xue, H. Sequence profiles of immunoglobulin and immunoglobulin-like domains. J Mol Biol 274, 530-45 (1997).
47. Evans, E. J. D.Phil Thesis: Protein structures and interactions at the leukocyte cell surface (University of Oxford, Oxford, 2002).
48. Fennelly, J. A., Tiwari, B., Davis, S. J. & Evans, E. J. CD2F-10: a new member of the CD2 subset of the immunoglobulin superfamily. Immunogenetics 53, 599-602 (2001).
49. Choudhuri, A., Kearney, A., Bakker, T. R. & van der Merwe, P. A. Immunology: how do T cells recognize antigen? Curr Biol 15, R382-5 (2005).
50. Meuer, S. C. et al. Antigen-like effects of monoclonal antibodies directed at receptors on human T cell clones. J Exp Med 158, 988-993 (1983).
51. Krogsgaard, M. & Davis, M. M. How T cells 'see' antigen. Nat Immunol 6, 239-45 (2005).
52. Unwin, N., Miyazawa, A., Li, J. & Fujiyoshi, Y. Activation of the nicotinic acetylcholine receptor involves a switch in conformation of the alpha subunits. J Mol Biol 319, 1165-76 (2002).
53. Gil, D., Schamel, W. W., Montoya, M., Sanchez-Madrid, F. & Alarcon, B. Recruitment of Nck by CD3 epsilon reveals a ligand-induced conformational change essential for T cell receptor signaling and synapse formation. Cell 109, 901-12 (2002).
54. Irles, C. et al. CD45 ectodomain controls interaction with GEMs and Lck activity for optimal TCR signaling. Nat Immunol 4, 189-97 (2003).
55. Davis, S. J. & van der Merwe, P. A. TCR triggering: co-receptor-dependent or -independent? Trends Immunol 24, 624-6; author reply 626-7 (2003).
56. Chan, A. C., Irving, B. A., Fraser, J. D. & Weiss, A. The zeta chain is associated with a tyrosine kinase and upon T-cell antigen receptor stimulation associates with ZAP-70, a 70-kDa tyrosine phosphoprotein. Proc Natl Acad Sci U S A 88, 9166-70 (1991).
57. Rotnes, J. S. & Bogen, B. Ca2+ mobilization in physiologically stimulated single T cells gradually increases with peptide concentration (analog signaling). Eur J Immunol 24, 851-8 (1994).
58. Lenschow, D. J., Walunas, T. L. & Bluestone, J. A. CD28/B7 system of T cell costimulation. Annu Rev Immunol 14, 233-58 (1996).

59. Davis, S. J. & van der Merwe, P. A. The immunological synapse: required for T cell receptor signalling or directing T cell effector function? Curr Biol 11, R289-91 (2001).
60. Delon, J. & Germain, R. N. Information transfer at the immunological synapse. Curr Biol 10, R923-33 (2000).
61. Lee, K. H. et al. T cell receptor signaling precedes immunological synapse formation. Science 295, 1539-42 (2002).
62. Monks, C. R., Freiberg, B. A., Kupfer, H., Sciaky, N. & Kupfer, A. Three-dimensional segregation of supramolecular activation clusters in T cells. Nature 395, 82-6 (1998).
63. Grakoui, A. et al. The immunological synapse: a molecular machine controlling T cell activation. Science 285, 221-7 (1999).
64. Bromley, S. K. et al. The immunological synapse and CD28-CD80 interactions. Nat Immunol 2, 1159-66 (2001).
65. Stinchcombe, J. C., Bossi, G., Booth, S. & Griffiths, G. M. The immunological synapse of CTL contains a secretory domain and membrane bridges. Immunity 15, 751-61 (2001).
66. van Der Merwe, P. A. & Davis, S. J. Immunology. The immunological synapse--a multitasking system. Science 295, 1479-80 (2002).
67. Orr, H. T., Lancet, D., Robb, R. J., Lopez de Castro, J. A. & Strominger, J. L. The heavy chain of human histocompatibility antigen HLA-B7 contains an immunoglobulin-like region. Nature 282, 266-70 (1979).
68. Kratzin, H. et al. [Primary structure of class II human histocompatibility antigens. 1st communication. Amino acid sequence of the N-terminal 198 residues of the beta chain of a HLA-Dw2,2;DR2,2-alloantigen (author's transl)]. Hoppe Seylers Z Physiol Chem 362, 1665-9 (1981).
69. Kohler, G. & Milstein, C. Continuous cultures of fused cells secreting antibody of predefined specificity. Nature 256, 495-7 (1975).
70. Committee on Methods of Producing Monoclonal Antibodies. Monoclonal Antibody Production (National Academy Press, Washington, D.C., 1999).
71. Seed, B. & Aruffo, A. Molecular cloning of the CD2 antigen, the T-cell erythrocyte receptor, by a rapid immunoselection procedure. Proc Natl Acad Sci U S A 84, 3365-9 (1987).
72. Yamasaki, K. et al. Cloning and expression of the human interleukin-6 (BSF-2/IFN beta 2) receptor. Science 241, 825-8 (1988).
73. Aruffo, A. & Seed, B. Molecular cloning of two CD7 (T-cell leukemia antigen) cDNAs by a COS cell expression system. Embo J 6, 3313-6 (1987).
74. Knappik, A. et al. Fully synthetic human combinatorial antibody libraries (HuCAL) based on modular consensus frameworks and CDRs randomized with trinucleotides. J Mol Biol 296, 57-86 (2000).
75. McCafferty, J., Griffiths, A. D., Winter, G. & Chiswell, D. J. Phage antibodies: filamentous phage displaying antibody variable domains. Nature 348, 552-4 (1990).
76. Holt, L. J., Enever, C., de Wildt, R. M. & Tomlinson, I. M. The use of recombinant antibodies in proteomics. Curr Opin Biotechnol 11, 445-9 (2000).
77. Kretzschmar, T. & von Ruden, T. Antibody discovery: phage display. Curr Opin Biotechnol 13, 598-602 (2002).
78. Lefkovits, I., Kettman, J. R. & Frey, J. R. Global analysis of gene expression in cells of the immune system I. Analytical limitations in obtaining sequence

information on polypeptides in two-dimensional gel spots. Electrophoresis 21, 2688-93 (2000).
79. Strachan, T. & Read, A. P. Human Molecular Genetics 2 (BIOS Scientific Publishers, Ltd, Oxford, 1999).
80. Pennisi, E. Human genome. A low number wins the GeneSweep Pool. Science 300, 1484 (2003).
81. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature 431, 931-45 (2004).
82. Schadt, E. E. et al. A comprehensive transcript index of the human genome generated using microarrays and computational approaches. Genome Biol 5, R73 (2004).
83. Cheng, J. et al. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science 308, 1149-54 (2005).
84. Bishop, J. O., Morton, J. G., Rosbash, M. & Richardson, M. Three abundance classes in HeLa cell messenger RNA. Nature 250, 199-204 (1974).
85. Resing, K. A. et al. Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. Anal Chem 76, 3556-68 (2004).
86. Resing, K. A. & Ahn, N. G. Proteomics strategies for protein identification. FEBS Lett 579, 885-9 (2005).
87. Wang, J. et al. Opinion: Vertebrate gene predictions and the problem of large genes. Nat Rev Genet 4, 741-749 (2003).
88. Schutte, B. C. et al. Discovery of five conserved beta -defensin gene clusters using a computational search strategy. Proc Natl Acad Sci U S A 99, 2129-33 (2002).
89. Blake, W. J., Kærn, M., Cantor, C. R. & Collins, J. J. Noise in eukaryotic gene expression. Nature 422, 633-7 (2003).
90. Raghava, G. P. & Han, J. H. Correlation and prediction of gene expression level from amino acid and dipeptide composition of its protein. BMC Bioinformatics 6, 59 (2005).
91. Sun, G. et al. The zinc finger protein cKrox directs CD4 lineage differentiation during intrathymic T cell positive selection. Nat Immunol 6, 373-81 (2005).
92. Eikhom, T. S., Abraham, K. A. & Dowben, R. M. Ribosomal RNA metabolism in synchronized plasmacytoma cells. Exp Cell Res 91, 301-9 (1975).
93. Brenner, S. et al. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol 18, 630-4 (2000).
94. Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. Serial analysis of gene expression. Science 270, 484-7 (1995).
95. Saha, S. et al. Using the transcriptome to annotate the genome. Nat Biotechnol 20, 508-12 (2002).
96. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270, 467-70 (1995).
97. Lockhart, D. J. et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 14, 1675-80 (1996).
98. Wang, H., He, X., Band, M., Wilson, C. & Liu, L. A study of inter-lab and inter-platform agreement of DNA microarray data. BMC Genomics 6, 71 (2005).

99. Gnatenko, D. V. et al. Transcript profiling of human platelets using microarray and serial analysis of gene expression. Blood 101, 2285-93 (2003).
100. Kuo, W. P., Jenssen, T. K., Butte, A. J., Ohno-Machado, L. & Kohane, I. S. Analysis of matched mRNA measurements from two different microarray technologies. Bioinformatics 18, 405-12 (2002).
101. Coughlan, S., Agrawal, V. & Meyers, B. A comparison of global gene expression measurement technologies in Arabidopsis thaliana. Comparative and Functional Genomics 5, 245-52 (2005).
102. Ibrahim, A. F. et al. A comparative analysis of transcript abundance using SAGE and Affymetrix arrays. Funct Integr Genomics (2005).
103. Kothapalli, R., Yoder, S. J., Mane, S. & Loughran, T. P., Jr. Microarray results: how accurate are they? BMC Bioinformatics 3, 22 (2002).
104. Zhang, J., Finney, R. P., Clifford, R. J., Derr, L. K. & Buetow, K. H. Detecting false expression signals in high-density oligonucleotide arrays by an in silico approach. Genomics 85, 297-308 (2005).
105. van Ruissen, F. et al. Evaluation of the similarity of gene expression data estimated with SAGE and Affymetrix GeneChips. BMC Genomics 6, 91 (2005).
106. Harbers, M. & Carninci, P. Tag-based approaches for transcriptome research and genome annotation. Nat Methods 2, 495-502 (2005).
107. Reinartz, J. et al. Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms. Brief Funct Genomic Proteomic 1, 95-104 (2002).
108. Unneberg, P., Wennborg, A. & Larsson, M. Transcript identification by analysis of short sequence tags--influence of tag length, restriction site and transcript database. Nucleic Acids Res 31, 2217-26 (2003).
109. Pleasance, E. D., Marra, M. A. & Jones, S. J. Assessment of SAGE in Transcript Identification. Genome Res (2003).
110. Schuler, G. D. Pieces of the puzzle: expressed sequence tags and the catalog of human genes. J Mol Med 75, 694-8 (1997).
111. Pontius, J. U., Wagner, L. & Schuler, G. D. UniGene: a unified view of the transcriptome. In: The NCBI Handbook. (2003).
112. Boon, K. et al. An anatomy of normal and malignant gene expression. Proc Natl Acad Sci U S A 99, 11287-92 (2002).
113. Versteeg, R. et al. The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. Genome Res 13, 1998-2004 (2003).
114. Sonenshein, G. E., Geoghegan, T. E. & Brawerman, G. A major species of mammalian messenger RNA lacking a polyadenylate segment. Proc Natl Acad Sci U S A 73, 3088-92 (1976).
115. Vanhee-Brossollet, C. & Vaquero, C. Do natural antisense transcripts make sense in eukaryotes? Gene 211, 1-9 (1998).
116. Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116, 281-97 (2004).
117. Wightman, B., Burglin, T. R., Gatto, J., Arasu, P. & Ruvkun, G. Negative regulatory sequences in the lin-14 3'-untranslated region are necessary to generate a temporal switch during Caenorhabditis elegans development. Genes Dev 5, 1813-24 (1991).

118. Cai, X., Hagedorn, C. H. & Cullen, B. R. Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. Rna 10, 1957-66 (2004).
119. Bracht, J., Hunter, S., Eachus, R., Weeks, P. & Pasquinelli, A. E. Trans-splicing and polyadenylation of let-7 microRNA primary transcripts. Rna 10, 1586-94 (2004).
120. Ambros, V. et al. A uniform system for microRNA annotation. Rna 9, 277-9 (2003).
121. Sijen, T. & Plasterk, R. H. Transposon silencing in the Caenorhabditis elegans germ line by natural RNAi. Nature 426, 310-4 (2003).
122. Chen, J. et al. Over 20% of human transcripts might form sense-antisense pairs. Nucleic Acids Res 32, 4812-20 (2004).
123. Yelin, R. et al. Widespread occurrence of antisense transcription in the human genome. Nat Biotechnol 21, 379-86 (2003).
124. Kiyosawa, H., Yamanaka, I., Osato, N., Kondo, S. & Hayashizaki, Y. Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. Genome Res 13, 1324-34 (2003).
125. Dahary, D., Elroy-Stein, O. & Sorek, R. Naturally occurring antisense: transcriptional leakage or real overlap? Genome Res 15, 364-8 (2005).
126. Fahey, M. E., Moore, T. F. & Higgins, D. G. Overlapping antisense transcription in the human genome. Comp Funct Genom 3, 244-253 (2002).
127. Kiyosawa, H., Mise, N., Iwase, S., Hayashizaki, Y. & Abe, K. Disclosing hidden transcripts: mouse natural sense-antisense transcripts tend to be poly(A) negative and nuclear localized. Genome Res 15, 463-74 (2005).
128. Bass, B. L. & Weintraub, H. An unwinding activity that covalently modifies its double-stranded RNA substrate. Cell 55, 1089-98 (1988).
129. Zhang, Z. & Carmichael, G. G. The fate of dsRNA in the nucleus: a p54(nrb)-containing complex mediates the nuclear retention of promiscuously A-to-I edited RNAs. Cell 106, 465-75 (2001).
130. Jen, C. H., Michalopoulos, I., Westhead, D. R. & Meyer, P. Natural antisense transcripts with coding capacity in Arabidopsis may have a regulatory role that is not linked to double-stranded RNA degradation. Genome Biol 6, R51 (2005).
131. Lehner, B., Williams, G., Campbell, R. D. & Sanderson, C. M. Antisense transcripts in the human genome. Trends Genet 18, 63-5 (2002).
132. Mestas, J., Crampton, S. P., Hori, T. & Hughes, C. C. Endothelial cell co-stimulation through OX40 augments and prolongs T cell cytokine synthesis by stabilization of cytokine mRNA. Int Immunol (2005).
133. Albrecht, G., Mosch, H. U., Hoffmann, B., Reusser, U. & Braus, G. H. Monitoring the Gcn4 protein-mediated response in the yeast Saccharomyces cerevisiae. J Biol Chem 273, 12696-702 (1998).
134. Varshavsky, A. The N-end rule: functions, mysteries, uses. Proc Natl Acad Sci U S A 93, 12142-9 (1996).
135. Beyer, A., Hollunder, J., Nasheuer, H. P. & Wilhelm, T. Post-transcriptional expression regulation in the yeast Saccharomyces cerevisiae on a genomic scale. Mol Cell Proteomics 3, 1083-92 (2004).
136. Tourriere, H., Chebli, K. & Tazi, J. mRNA degradation machines in eukaryotic cells. Biochimie 84, 821-37 (2002).
137. Guhaniyogi, J. & Brewer, G. Regulation of mRNA stability in mammalian cells. Gene 265, 11-23 (2001).

138. Kern, W. et al. Correlation of protein expression and gene expression in acute leukemia. Cytometry B Clin Cytom 55, 29-36 (2003).
139. Gygi, S. P., Rochon, Y., Franza, B. R. & Aebersold, R. Correlation between protein and mRNA abundance in yeast. Mol Cell Biol 19, 1720-30 (1999).
140. Greenbaum, D., Colangelo, C., Williams, K. & Gerstein, M. Comparing protein abundance and mRNA expression levels on a genomic scale. Genome Biol 4, 117 (2003).
141. Orntoft, T. F., Thykjaer, T., Waldman, F. M., Wolf, H. & Celis, J. E. Genome-wide study of gene copy numbers, transcripts, and protein levels in pairs of non-invasive and invasive human transitional cell carcinomas. Mol Cell Proteomics 1, 37-45 (2002).
142. Griffin, T. J. et al. Complementary profiling of gene expression at the transcriptome and proteome levels in Saccharomyces cerevisiae. Mol Cell Proteomics 1, 323-33 (2002).
143. Hashimoto, S. et al. Gene expression profile in human leukocytes. Blood 101, 3509-13 (2003).
144. Hashimoto, S. et al. Serial analysis of gene expression in human monocyte-derived dendritic cells. Blood 94, 845-52 (1999).
145. Nagai, S. et al. Comprehensive gene expression profile of human activated T(h)1- and T(h)2-polarized cells. Int Immunol 13, 367-76 (2001).
146. Obata-Onai, A. et al. Comprehensive gene expression analysis of human NK cells and CD8(+) T lymphocytes. Int Immunol 14, 1085-98 (2002).
147. Ryo, A. et al. Serial analysis of gene expression in HIV-1-infected T cell lines. FEBS Lett 462, 182-6 (1999).
148. Cobbold, S. P. et al. Regulatory T cells and dendritic cells in transplantation tolerance: molecular markers and mechanisms. Immunol Rev 196, 109-24 (2003).
149. Graca, L. et al. Both CD4(+)CD25(+) and CD4(+)CD25(-) regulatory cells mediate dominant transplantation tolerance. J Immunol 168, 5558-65 (2002).
150. Cobbold, S. P., Adams, E., Graca, L. & Waldmann, H. Serial analysis of gene expression provides new insights into regulatory T cells. Semin Immunol 15, 209-14 (2003).
151. Zelenika, D. et al. The role of CD4+ T-cell subsets in determining transplantation rejection or tolerance. Immunol Rev 182, 164-79 (2001).
152. Shires, J., Theodoridis, E. & Hayday, A. C. Biological insights into TCRgammadelta+ and TCRalphabeta+ intraepithelial lymphocytes provided by serial analysis of gene expression (SAGE). Immunity 15, 419-34 (2001).
153. Caron, H. et al. The human transcriptome map: clustering of highly expressed genes in chromosomal domains. Science 291, 1289-92 (2001).
154. Patankar, S., Munasinghe, A., Shoaibi, A., Cummings, L. M. & Wirth, D. F. Serial analysis of gene expression in Plasmodium falciparum reveals the global expression profile of erythrocytic stages and the presence of anti-sense transcripts in the malarial parasite. Mol Biol Cell 12, 3114-25 (2001).
155. Wang, X. J., Gaasterland, T. & Chua, N. H. Genome-wide prediction and identification of cis-natural antisense transcripts in Arabidopsis thaliana. Genome Biol 6, R30 (2005).
156. Wahl, M. B., Heinzmann, U. & Imai, K. LongSAGE analysis revealed the presence of a large number of novel antisense genes in the mouse genome. Bioinformatics 21, 1389-92 (2005).

157. Shiraki, T. et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci U S A 100, 15776-81 (2003).
158. Hashimoto, S. et al. 5'-end SAGE for the analysis of transcriptional start sites. Nat Biotechnol 22, 1146-9 (2004).
159. Wei, C. L. et al. 5' Long serial analysis of gene expression (LongSAGE) and 3' LongSAGE for transcriptome characterization and genome annotation. Proc Natl Acad Sci U S A 101, 11701-6 (2004).
160. Curwen, V. et al. The Ensembl automatic gene annotation system. Genome Res 14, 942-50 (2004).
161. Moss, P. A. et al. Persistent high frequency of human immunodeficiency virus-specific cytotoxic T cells in peripheral blood of infected donors. Proc Natl Acad Sci U S A 92, 5773-7 (1995).
162. Sutton, J. K. D.Phil. Thesis: CD4+ T cell responses to HIV-1 (University of Oxford, Oxford, 2004).
163. Wilson, C. C. et al. Identification and antigenicity of broadly cross-reactive and conserved human immunodeficiency virus type 1-derived helper T-lymphocyte epitopes. J Virol 75, 4195-207 (2001).
164. Chen, J. J., Rowley, J. D. & Wang, S. M. Generation of longer cDNA fragments from serial analysis of gene expression tags for gene identification. Proc Natl Acad Sci U S A 97, 349-53 (2000).
165. Frohman, M. A., Dush, M. K. & Martin, G. R. Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. Proc Natl Acad Sci U S A 85, 8998-9002 (1988).
166. Lash, A. E. et al. SAGEmap: a public gene expression resource. Genome Res 10, 1051-60 (2000).
167. Pruitt, K. D. & Maglott, D. R. RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res 29, 137-40 (2001).
168. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389-402 (1997).
169. Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng 10, 1-6 (1997).
170. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305, 567-80 (2001).
171. Huang, X. & Madan, A. CAP3: A DNA sequence assembly program. Genome Res 9, 868-77 (1999).
172. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25, 25-9 (2000).
173. Romualdi, C., Bortoluzzi, S. & Danieli, G. A. Detecting differentially expressed genes in multiple tag sampling experiments: comparative evaluation of statistical tests.PG - 2133-41. Hum Mol Genet 10 (2001).
174. Audic, S. & Claverie, J. M. The significance of digital gene expression profiles. Genome Res 7, 986-95 (1997).
175. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8, 175-85 (1998).

176. Prosdocimi, F., Peixoto, F. C. & Ortega, J. M. Evaluation of window cohabitation of DNA sequencing errors and lowest PHRED quality values. Genet Mol Res 3, 483-92 (2004).
177. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30, 207-10 (2002).
178. Birney, E. et al. An overview of Ensembl. Genome Res 14, 925-8 (2004).
179. Eyras, E., Caccamo, M., Curwen, V. & Clamp, M. ESTGenes: alternative splicing from ESTs in Ensembl. Genome Res 14, 976-87 (2004).
180. R Development Core Team. R: A Language and Environment for Statistical Computing (Vienna, 2005).
181. Stekel, D. Microarray Bioinformatics (2003).
182. Fernandez-Miguel, G. et al. Multivalent structure of an alphabetaT cell receptor. Proc Natl Acad Sci U S A 96, 1547-52 (1999).
183. Harder, T. Raft membrane domains and immunoreceptor functions. Adv Immunol 77, 45-92 (2001).
184. Mason, D. Y. et al. CD antigens 2001. Tissue Antigens 58, 425-30 (2001).
185. Riggins, G. J. & Strausberg, R. L. Genome and genetic resources from the Cancer Genome Anatomy Project. Hum Mol Genet 10, 663-7 (2001).
186. Schulz, H. L., Goetz, T., Kaschkoetoe, J. & Weber, B. H. The Retinome - Defining a reference transcriptome of the adult mammalian retina/retinal pigment epithelium. BMC Genomics 5, 50 (2004).
187. Sharon, D., Blackshaw, S., Cepko, C. L. & Dryja, T. P. Profile of the genes expressed in the human peripheral retina, macula, and retinal pigment epithelium determined through serial analysis of gene expression (SAGE). Proc Natl Acad Sci U S A 99, 315-20 (2002).
188. Raulet, D. H. Interplay of natural killer cells and their receptors with the adaptive immune response. Nat Immunol 5, 996-1002 (2004).
189. Moretta, A., Biassoni, R., Bottino, C., Mingari, M. C. & Moretta, L. Natural cytotoxicity receptors that trigger human NK-cell-mediated cytolysis. Immunol Today 21, 228-34 (2000).
190. Chiesa, S., Tomasello, E., Vivier, E. & Vely, F. Coordination of activating and inhibitory signals in natural killer cells. Mol Immunol 42, 477-84 (2005).
191. Moretta, L., Biassoni, R., Bottino, C., Mingari, M. C. & Moretta, A. Natural killer cells: a mystery no more. Scand J Immunol 55, 229-32 (2002).
192. Lanier, L. L., Chang, C. & Phillips, J. H. Human NKR-P1A. A disulfide-linked homodimer of the C-type lectin superfamily expressed by a subset of NK and T lymphocytes. J Immunol 153, 2417-28 (1994).
193. van den Berg, A., van der Leij, J. & Poppema, S. Serial analysis of gene expression: rapid RT-PCR analysis of unknown SAGE tags. Nucleic Acids Res 27, e17 (1999).
194. Polyak, K., Xia, Y., Zweier, J. L., Kinzler, K. W. & Vogelstein, B. A model for p53-induced apoptosis. Nature 389, 300-5 (1997).
195. Chen, J., Lee, S., Zhou, G. & Wang, S. M. High-throughput GLGI procedure for converting a large number of serial analysis of gene expression tag sequences into 3' complementary DNAs. Genes Chromosomes Cancer 33, 252-61 (2002).
196. Zhou, G. et al. The pattern of gene expression in human CD34(+) stem/progenitor cells. Proc Natl Acad Sci U S A 98, 13966-71 (2001).

197. Quere, R. et al. Mining SAGE data allows large-scale, sensitive screening of antisense transcript expression. Nucleic Acids Res 32, e163 (2004).
198. Akmaev, V. R. & Wang, C. J. Correction of sequence-based artifacts in serial analysis of gene expression. Bioinformatics 20, 1254-63 (2004).
199. Colinge, J. & Feger, G. Detecting the impact of sequencing errors on SAGE data. Bioinformatics 17, 840-2 (2001).
200. Beissbarth, T. et al. Statistical modeling of sequencing errors in SAGE libraries. Bioinformatics 20 Suppl 1, I31-I39 (2004).
201. Velculescu, V. E. et al. Characterization of the yeast transcriptome. Cell 88, 243-51 (1997).
202. Medigue, C., Rose, M., Viari, A. & Danchin, A. Detecting and analyzing DNA sequencing errors: toward a higher quality of the Bacillus subtilis genome sequence. Genome Res 9, 1116-27 (1999).
203. Hastie, N. D. & Bishop, J. O. The expression of three abundance classes of messenger RNA in mouse tissues. Cell 9, 761-74 (1976).
204. Silva, A. P., Chen, J., Carraro, D. M., Wang, S. M. & Camargo, A. A. Generation of longer 3' cDNA fragments from massively parallel signature sequencing tags. Nucleic Acids Res 32, e94 (2004).
205. Hubbard, T. et al. Ensembl 2005. Nucleic Acids Res 33, D447-53 (2005).
206. Blackshaw, S. et al. MicroSAGE is highly representative and reproducible but reveals major differences in gene expression among samples obtained from similar tissues. Genome Biol 4, R17 (2003).
207. Dinel, S. et al. Reproducibility, bioinformatic analysis and power of the SAGE method to evaluate changes in transcriptome. Nucleic Acids Res 33, e26 (2005).
208. Stolovitzky, G. A. et al. Statistical analysis of MPSS measurements: application to the study of LPS-activated macrophage gene expression. Proc Natl Acad Sci U S A 102, 1402-7 (2005).
209. Meyers, B. C. et al. The use of MPSS for whole-genome transcriptional analysis in Arabidopsis. Genome Res 14, 1641-53 (2004).
210. Silva, A. P. et al. The impact of SNPs on the interpretation of SAGE and MPSS experimental data. Nucleic Acids Res 32, 6104-6110 (2004).
211. Lee, S. et al. Detecting novel low-abundant transcripts in Drosophila. Rna 11, 939-46 (2005).
212. Southan, C. Has the yo-yo stopped? An assessment of human protein-coding gene number. Proteomics 4, 1712-26 (2004).
213. Bantle, J. A. & Hahn, W. E. Complexity and characterization of polyadenylated RNA in the mouse brain. Cell 8, 139-50 (1976).
214. Ross, I. L., Browne, C. M. & Hume, D. A. Transcription of individual genes in eukaryotic cells occurs randomly and infrequently. Immunol Cell Biol 72, 177-85 (1994).
215. Jongeneel, C. V. et al. Comprehensive sampling of gene expression in human cell lines with massively parallel signature sequencing. Proc Natl Acad Sci U S A 100, 4702-5 (2003).
216. Brandenberger, R. et al. MPSS profiling of human embryonic stem cells. BMC Dev Biol 4, 10 (2004).
217. Tanino, M. et al. The Human Anatomic Gene Expression Library (H-ANGEL), the H-Inv integrative display of human gene expression across disparate technologies and platforms. Nucleic Acids Res 33, D567-72 (2005).

218. Miura, T. et al. Monitoring early differentiation events in human embryonic stem cells by massively parallel signature sequencing and expressed sequence tag scan. Stem Cells Dev 13, 694-715 (2004).
219. Meyers, B. C. et al. Analysis of the transcriptional complexity of Arabidopsis thaliana by massively parallel signature sequencing. Nat Biotechnol 22, 1006-11 (2004).
220. Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol 94, 441-8 (1975).
221. Bonafoux, B. et al. Analysis of remnant reticulocyte mRNA reveals new genes and antisense transcripts expressed in the human erythroid lineage. Haematologica 89, 1434-8 (2004).
222. Harper, R. et al. Identification of a novel MAGE D2 antisense RNA transcript in human tissues. Biochem Biophys Res Commun 324, 199-204 (2004).
223. Shendure, J. & Church, G. M. Computational discovery of sense-antisense transcription in the human and mouse genomes. Genome Biol 3, RESEARCH0044 (2002).
224. Chen, J., Sun, M., Hurst, L. D., Carmichael, G. G. & Rowley, J. D. Human antisense genes have unusually short introns: evidence for selection for rapid transcription. Trends Genet 21, 203-7 (2005).
225. Lu, J., Lal, A., Merriman, B., Nelson, S. & Riggins, G. A comparison of gene expression profiles produced by SAGE, long SAGE, and oligonucleotide chips. Genomics 84, 631-6 (2004).
226. Sun, M. et al. SAGE is far more sensitive than EST for detecting low-abundance transcripts. BMC Genomics 5, 1 (2004).
227. Frayn, K. N. Metabolic Regulation: A Human Perspective, 2nd edition (Blackwell Science, Oxford, 2003).
228. Pelkmans, L. et al. Genome-wide analysis of human kinases in clathrin- and caveolae/raft-mediated endocytosis. Nature 436, 78-86 (2005).
229. Iborra, F. J., Pombo, A., Jackson, D. A. & Cook, P. R. Active RNA polymerases are localized within discrete transcription "factories' in human nuclei. J Cell Sci 109 (Pt 6), 1427-36 (1996).
230. Osborne, C. S. et al. Active genes dynamically colocalize to shared sites of ongoing transcription. Nat Genet 36, 1065-71 (2004).