

The T Cell Surface— How Well Do We Know It?

Edward J. Evans,¹ Lawrence Hene,¹ Lisa M. Sparks,¹
Tao Dong,² Christelle Retiere,² Janet A. Fennelly,¹
Raquel Manso-Sancho,¹ Jill Powell,³
Veronique M. Braud,^{2,4} Sarah L. Rowland-Jones,²
Andrew J. McMichael,² and Simon J. Davis^{1,*}

¹Nuffield Department of Clinical Medicine
The University of Oxford
John Radcliffe Hospital

²Medical Research Council Human Immunology Unit
Weatherall Institute of Molecular Medicine
The University of Oxford
John Radcliffe Hospital

Headington
Oxford OX3 9DU

³The Richard Dimbleby Department of
Cancer Research
Cancer Research UK Laboratory
Rayne Institute
St. Thomas's Hospital
Lambeth Palace Road
London SE1 7EH
United Kingdom

Summary

The overall degree of complexity of the T cell surface has been unclear, constraining our understanding of its biology. Using global gene expression analysis, we show that 111 of 374 genes encoding well-characterized leukocyte surface antigens are expressed by a resting cytotoxic T cell. Unexpectedly, of 97 stringently defined, T cell-specific transcripts with unknown functions that we identify, none encode proteins with the modular architecture characteristic of 80% of leukocyte surface antigens. Only two encode proteins with membrane topologies found exclusively in cell surface molecules. Our analysis indicates that the cell type-specific composition of the resting CD8⁺ T cell surface is now largely defined, providing an insight into the overall compositional complexity of the mammalian cell surface and a framework for formulating systematic models of T cell surface-dependent processes, such as T cell receptor triggering.

Introduction

Our understanding of cell surface biology and immunological phenomena have been driven, to a considerable extent, by the characterization of proteins expressed at leukocyte cell surfaces, which remain the best-characterized of all mammalian cells. Early work was limited to the characterization of murine surface antigens via the analysis of allotypic differences between inbred mouse strains (reviewed in Williams, 1977). Among the early

antigens identified in this way were Thy-1, which was of key importance for delineating T and B lymphocytes, and Ly-1 and Ly-2/3, which led to the discovery of distinct T cell subsets. The rate at which leukocyte surface antigens, of humans in particular, were identified was transformed in the 1970s by the application of the monoclonal antibody approach to the problem (Williams et al., 1977). This relied on the xenogeneic immunization of mice with leukocyte membrane preparations and succeeded immediately at clarifying the molecular complexity of these mixtures, allowing the identification of both minor membrane antigens and proteins expressed on small populations of leukocytes.

The development of an extremely efficient expression cloning method (Seed and Aruffo, 1987) revolutionized the identification of the genes encoding these antigens. Although the rate of cloning new genes slowed considerably in the last decade (Barclay et al., 1997), at the most recent Human Leukocyte Differentiation Antigen (HLDA) Workshop held in July 2000, the total number of clusters of differentiation (CD) antigens identified on leukocytes reached 247. Over 100 additional antigens will be considered for CD designation at the next workshop. Approximately 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 19 different protein superfamilies (Barclay et al., 1997). A further 3% of leukocyte antigens lack these domains but are characterized by the presence of either four (e.g., Fc receptor subunits) or seven (i.e., G protein-coupled receptors) transmembrane domains.

Cell surface biology increasingly seems to be characterized by the ensemble behavior of cell surface proteins, necessitating, at the very least, insights into the overall complexity of these systems. The emerging view of leukocyte interactions, for example, is that these are controlled by the cooperative movement and interactions of a relatively large set of molecules, initially at a scale below the resolution of confocal or deconvolution fluorescence microscopy (Davis and van der Merwe, 1996; van der Merwe et al., 2000), which culminates in the formation of a μm scale structure called the immunological synapse (Grakoui et al., 1999; van der Merwe et al., 2000). More generally, it is recognized that the spatial microcompartmentalization of protein kinases and phosphatases, some of which are cell surface molecules, underlies the complexity and specificity of signal transduction (Mochly-Rosen, 1995). On the other hand, the discovery of single proteins with novel activities can also have considerable impact on our understanding of these systems. An example of this is the transformation of our understanding of costimulatory signaling, which can determine the course of immune responses, by the recent discovery of the T cell surface antigen, ICOS (Hutloff et al., 1999). For these reasons, it will be of great importance to know when, for given cells and tissues, the discovery process has been completed and the development of systematic, quantitative models of cell surface function can be undertaken.

*Correspondence: sdavis@molbiol.ox.ac.uk

⁴Present address: CNRS UMR 6097, Institut de Pharmacologie Moléculaire et Cellulaire, Sophia Antipolis, 06 560 Valbonne, France.

Genome sequencing has opened up the possibility of systematic compositional analyses of complex biological systems. However, relatively few technologies fully exploit this opportunity by allowing the characterization and simultaneous, discovery-driven analysis of entire transcriptomes. Global gene expression analyses, as currently embodied in microarray technology, will clearly play a key role in the future (reviewed by Shaffer et al., 2001; Staudt, 2002; Staudt and Brown, 2000), but these and other closed architecture systems are limited to retrospective inquiries and by the extent to which transcriptome coverage has been achieved. In the light of current uncertainty regarding the actual number of genes in the human genome, and the somewhat poor consensus between *ab initio* gene prediction methods (Hogenesch et al., 2001), it may be some time before all human genes can be confidently sampled using this methodology.

Open gene expression profiling methods such as differential-display PCR and serial analysis of gene expression (SAGE), on the other hand, require no *a priori* knowledge of the genes likely to be of interest (Green et al., 2001). SAGE is based on the generation and high-throughput sequencing of short (e.g., 14 bp) tags from single, fixed positions in individual cDNA transcripts (Velculescu et al., 1995). SAGE has already been used to compare differential gene expression patterns in a variety of leukocyte populations, including CD4⁺ T cell subsets (Nagai et al., 2001; Zelenika et al., 2002), monocytes, macrophages, and dendritic cells (e.g., Hashimoto et al., 1999) and intraepithelial lymphocytes (Shires et al., 2001), and to follow gene expression changes during HIV-1 infection of CD4⁺ T cells (Ryo et al., 1999). Perhaps the most important advantage of SAGE, however, is that, because large numbers of tags can be sequenced efficiently, the method can be used to probe very deeply into cellular transcriptomes.

We have used SAGE to characterize the expression of genes encoding known cell surface molecules by a human CD8⁺ cytotoxic T cell clone and show that 111 of these genes are expressed in this cell. We also show that, although ~45% of the moderately to highly abundant, stringently defined CTL-specific transcripts have yet to be ascribed any function, a surprisingly small number of these uncharacterized transcripts encode proteins with transmembrane domains, and that none of these have the modular architecture characteristic of the great majority of leukocyte surface antigens. This indicates that most, if not all, of the cell type-specific proteins expressed on the human resting CD8⁺ T cell surface have been identified, providing an insight into its overall complexity. In the course of our analysis we also show that tissue-specific gene expression in the T cell stratifies with protein functional class, that tissue-level regulatory changes accompany the activation and/or lineage commitment of CD8⁺ and CD4⁺ T cells, and that the least well-characterized part of the T cell-specific transcriptome encodes signaling molecules.

Results

We used the SAGE method to estimate the extent to which the resting cytotoxic T lymphocyte (CTL) surface

has been characterized. Our initial goal was to show that the methodology efficiently detects transcripts encoding cell surface molecules. We then established stringent criteria for identifying a set of transcripts that are highly CTL specific compared to nonimmune tissues. Our objective was then to define, using this set, the ratio of known versus previously unknown surface molecules (i.e., those encoded by uncharacterized cDNAs and EST clusters). Our analysis focused on the transcriptome of a human CTL clone in order to avoid the confounding effects of gene expression heterogeneity.

Library Construction and Properties

Clone 32, isolated from the blood of an asymptomatic, untreated HIV-infected East African woman, recognizes the HIV-1 polA peptide ETAYFILKL bound to the A6802 MHC class I molecule (Figures 1A and 1B) in a CD8-independent manner (Figure 1C). FACS analyses done at the time of library preparation, 4 weeks after the final expansion, indicate that clone 32 has the phenotype of a conventional resting CD8⁺ T cell insofar as it expresses CD2, CD3, CD5, CD8, CD11a, CD43, CD132, HLA-DR, and CD45RB (see Supplemental Figure S1 and Supplemental Data at <http://www.immunity.com/cgi/content/full/19/2/213/DC1>). As expected, the cells also express markers indicative of previous antigen exposure, i.e., CD38, CD45RO, CD69, CD70, and CD95. PCR-based HLA haplotyping confirmed that the cDNA used for library construction was not significantly contaminated with transcripts from feeder cells used in culture (Figure 2A): the clone's HLA-B42 and -B45 alleles were readily amplifiable, whereas only one of four feeder cell-specific alleles gave any detectable product (B44). Dilution analysis showed that the library is ~100,000-fold enriched for B42 (clone) versus B44 (feeder) transcripts (Figure 2A).

A total of 71,174 SAGE tags, representing 20,204 distinct sequences (excluding linker-derived tags) were identified in a library prepared from the cDNA. Assuming 300,000 transcripts/cell (Velculescu et al., 1999) and binomial sampling, all transcripts constituting 0.008% or more of the transcriptome (i.e., those expressed at 22 or more copies/cell) were sampled with 95% confidence. Log scatter plots versus cerebellum (Riggins and Strausberg, 2001; Figure 2B)- and CD56⁺CD3⁻ NK cell line (unpublished data; Figure 2C)-derived libraries confirmed sample purity: granulysin and RANTES transcripts constitute approximately 1% of the total transcripts in the clone 32 and NK libraries but are not present at all in cerebellum, whereas the expression of the NMDA glutamate receptor 2C is restricted to brain tissue (Figure 2B). Similarly, CD3 δ and CD158 (KIR family) transcripts exhibit reciprocal expression in the clone 32 and NK cell libraries (Figure 2C). The ability of SAGE to accurately mirror transcript abundance according to Northern blot analysis or real-time PCR has been confirmed by others (Shires et al., 2001; Velculescu et al., 1997).

Transcripts Encoding Cell Surface Molecules Expressed by Clone 32

SAGE tags for all 247 CD antigen clusters defined at the Seventh HLDA Workshop (www.ncbi.nlm.nih.gov/prow/guide/45277084.htm; Mason et al., 2001), except those

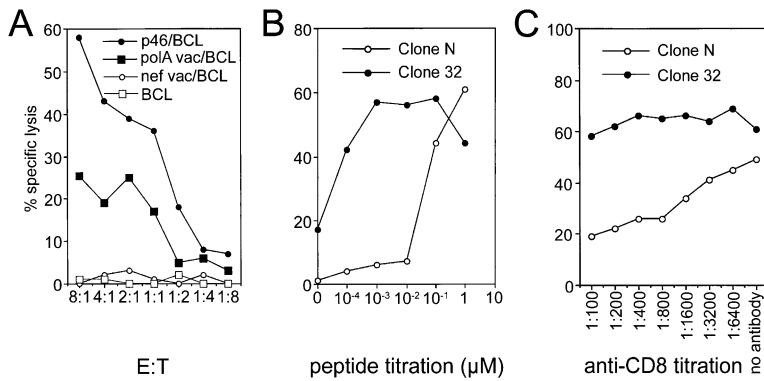


Figure 1. Characterization of CTL Clone 32 Cytolytic Activity

Standard 4 hr ^{51}Cr release cytotoxicity assays.

(A) Autologous B cell targets given no antigen (\square), pulsed with specific peptide (\bullet), or infected with a modified Vaccinia virus expressing the polA protein (\blacksquare) or an irrelevant protein (nef, \circ) were mixed with clone 32 CTL at the indicated effector:target (E:T) ratio.

(B) B cell targets were pulsed with the indicated concentration of peptide and mixed with clone 32 or clone N at an E:T ratio of 8:1. Clone 32 targets were A6802 B cells pulsed with peptide ETAYFILKL, and clone N targets were B35 B cells pulsed with the influenza peptide ASCMGLIY.

(C) B cells pulsed with $0.1 \mu\text{M}$ peptide were mixed with clone 32 or clone N as in (B), in the presence of anti-CD8 antibody.

containing carbohydrate or uncharacterized antigens, were obtained using the SAGEmap Unigene-to-Tag mapping database (www.ncbi.nlm.nih.gov/SAGE; Lash et al., 2000), with manual curation of the matches where necessary. 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 will be considered for CD status at the Eighth HLDA Workshop (www.hlda8.org/PotentialCDs.htm) were also included.

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 (see Supplemental Table S1). 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-1 α), CD43, CD45, and CD53. CD18 tags were found at high levels in the library, but one of these matched several other genes, preventing abundance determination. The absence of CD28 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. The absence of B cell (e.g., CD19, CD20, CD21, CD22) and myeloid (e.g., CD14, CD32, CD33) lineage-marker tags confirmed the purity of the starting population and the

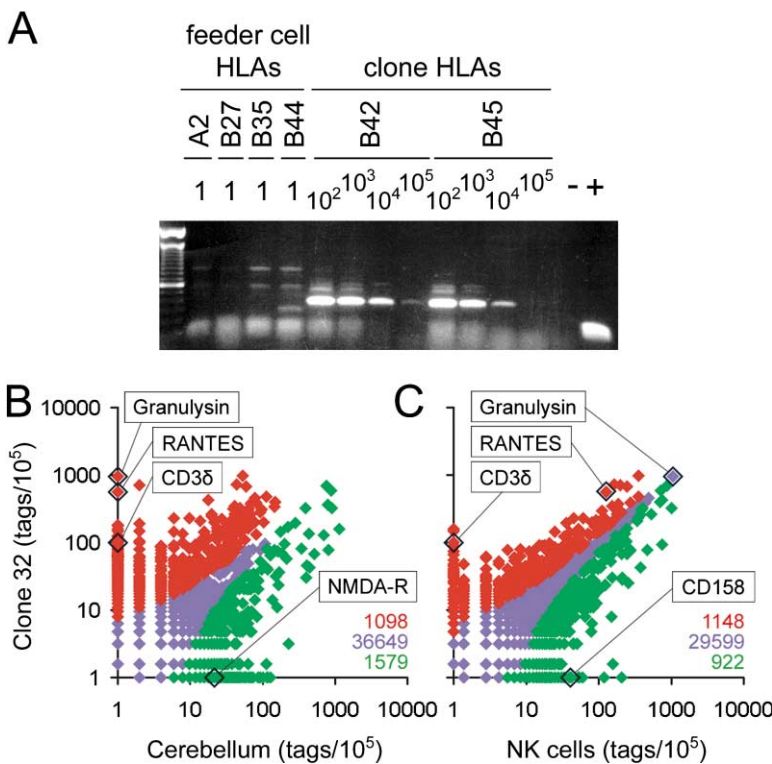


Figure 2. Assessment of the Quality of the Clone 32 SAGE Library

(A) HLA-typing PCR analysis of clone 32 cDNA used to produce the SAGE library. Clone 32 expresses HLA-B42 and -B45, whereas the irradiated mixed feeder cells expressed the other four alleles tested (A2, B27, B35, and B44). Numbers over each lane refer to the fold dilution of cDNA used in the PCR reactions. Control PCR reactions in the absence (-) and presence (+) of HLA-A2 cDNA are shown.

(B and C) The abundance of each SAGE tag, per 100,000 tags, in the CTL library is plotted on a logarithmic scale against that in libraries from normal cerebellum (B) or CD56 $^+$ CD3 $^-$ NK cells separated from PBL using magnetic beads (C). Absent tags are assigned a value of 1 tag per 100,000. Red and green symbols represent tags that are significantly more abundant in clone 32 (p value ≤ 0.05 by AC test) and the second library, respectively. Labeled tags are from transcripts whose expression is expected to be restricted.

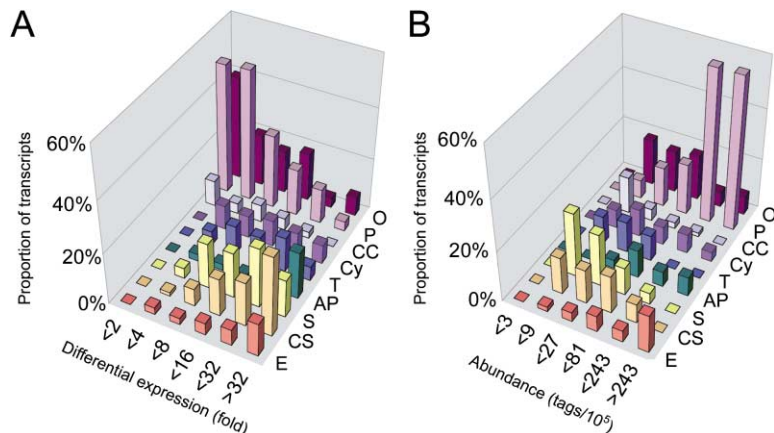


Figure 3. Proportion of CTL-Specific Transcripts in Each Functional Class at Each Expression Level

SAGE tags significantly more abundant (p value ≤ 0.05 by the AC test) in the CTL library compared to normal cerebellum were categorized by functional class and with regard to (A) fold difference in expression in clone 32 compared to cerebellum and (B) overall tag abundance (per 100,000 tags) in the clone 32 library. Functional classes are: E, soluble effector molecules; CS, cell surface molecules; S, signaling molecules; AP, antigen presentation; T, transcriptional regulation; Cy, cytoskeleton related; CC, cell cycle or viability related; P, protein synthesis related; and O, other.

absence of feeder cell-derived transcripts. The cerebellum was characterized by large numbers of transcripts encoding CD56 (NCAM), CD90 (Thy-1), CD230 (prion protein), and CD231 (TALLA-1). The reciprocal expression of transcripts encoding CD3 components and most NK cell inhibitory receptors is the most striking difference between clone 32 and the NK cell line (unpublished data).

Identification of CTL-Specific Transcripts

A series of pairwise comparisons with unrelated SAGE libraries was used to identify CTL-specific transcripts, using Audic-Claverie (AC; Audic and Claverie, 1997) or χ^2 tests of statistical significance (see Experimental Procedures). The robustness of our method was verified by examining the properties of the known genes present at each stage of the selection procedure.

An initial comparison with cerebellum indicated that 1098 transcripts are significantly more abundant ($p \leq 0.05$) in clone 32 than in this tissue (Figure 2B). These transcripts were categorized according to the function of the encoded protein, where this is known, using RefSeq (Pruitt and Maglott, 2001), Unigene (Schuler et al., 1996), and LocusLink (Pruitt and Maglott, 2001) annotations, or the literature (Supplemental Table S2). An unexpectedly simple pattern of stratified differential gene expression emerged when we examined the relationship between protein functional class and differential expression level for the set of known transcripts within this initial pool (Figure 3A). Highly differentially expressed (≥ 8 -fold) clone 32 transcripts with known functions are dominated by those encoding cell surface receptors, signaling proteins, and soluble effector molecules, whereas transcripts that are less differentially expressed (≤ 4 -fold) are mainly restricted to housekeeping functions, particularly protein synthesis. In contrast, there is no apparent relationship between functional class and absolute levels of transcript abundance (Figure 3B). The relatively high degree of differential expression of transcripts encoding cell surface molecules is significant as it implies that these transcripts will generally be identifiable at the depth of our sampling of the clone 32 transcriptome.

The inclusion of a large number of transcripts involved in protein/mRNA synthesis and processing (24% of the

genes with known function; Figure 4A, left) implied that this initial pairwise comparison reflected, to a significant degree, the nonproliferative properties of cerebellum, rather than cell type-specific differences. To avoid this complication, we subtracted transcripts that were not significantly more abundant in clone 32 than in a second library prepared from a proliferating tissue, ovary epithelium (Riggins and Strausberg, 2001). This reduced the set of transcripts to 758 in total and the fraction involved in protein synthesis from 24% to 20% (Figure 4A, center). A final filter involved comparisons with a panel of 12 different tumor-derived SAGE libraries (Riggins and Strausberg, 2001). We 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. This final step 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 4A, right).

Overall, the three-step selection process only eliminated 59 transcripts with known immune function from the initial list. Most of these are expected to be expressed in nonimmune tissues, e.g., those encoding MHC molecules and proteasome subunits (data not shown). Thirty-four of the transcripts encoding known cell surface molecules survived the selection procedure, including 17 of the 18 principal T cell markers detectable in the library (i.e., TCR α , $\beta 1$ and $\beta 2$, CD3 γ , δ and ϵ , CD2, CD5, CD6, CD7, CD8 α , $\beta 1$ and $\beta 2$, CD11a [LFA-1 α], CD45, CD53 and CD247 [ζ]). The only exception was CD43 for which the tag count fell just below the cutoff for significance when compared to cerebellum (see Supplemental Table S1 at <http://www.immunity.com/cgi/content/full/19/2/213/DC1>).

Characterization of CTL-Specific Transcripts

At each stage of the selection process, the CTL-specific transcripts were categorized as having (1) a known immune function, (2) some other known function, (3) known sequence but no characterized 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. For each dataset, 42%–45% of CTL-specific transcripts have yet

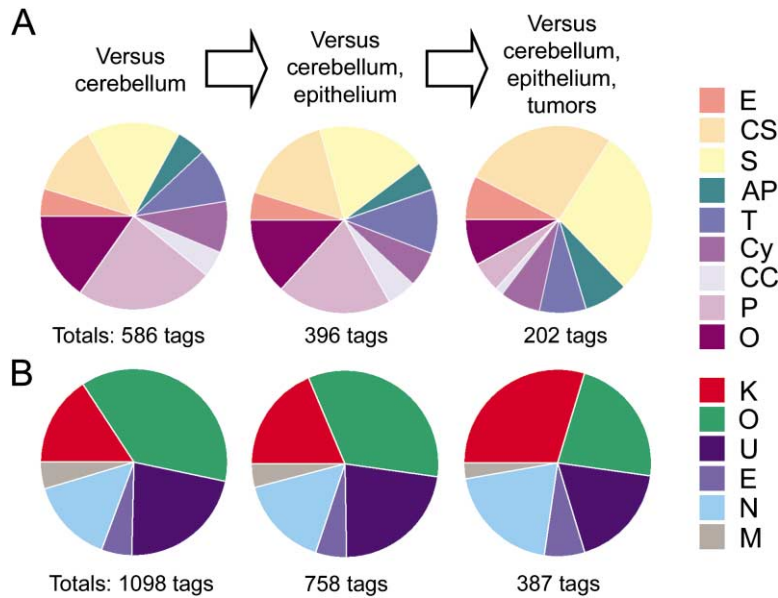


Figure 4. Function and Extent of Characterization of CTL-Specific Transcripts

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 (center), 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 12 tumor libraries (right). SAGE tags were assigned to transcripts, and these transcripts were then categorized according to the broad function of their protein products (A) and level of characterization (B) using LocusLink, Unigene, and RefSeq database entries. Functional classes (A) are as in Figure 3. Characterization classes (B) are: K, known immune function; O, other known function; U, uncharacterized cDNA; E, EST cluster; N, no true Unigene match; M, multiple true matches.

to be assigned any clear function (Figure 4B). The relative invariance of this fraction suggests that, regardless of how tissue specificity is defined, assuming that all the nonmatching tags correspond to bona fide transcripts, only ~55%–60% of the differentially regulated, moderately to highly abundant transcripts in the resting T cell transcriptome have been characterized. At the highest level of stringency, more than half the known, CTL-specific transcripts encode cell surface molecules (27%) or signaling/adaptor proteins (29%, Figure 4A, 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, 11 (5%) encode proteins involved in protein/mRNA synthesis and 19 (9%) have other miscellaneous functions.

Standard bioinformatics tools were used to screen the set of 97 stringently defined CTL-specific transcripts matching Unigene clusters but lacking any prescribed function for transcripts encoding transmembrane regions or known modular domains. Putative complete open reading frames were available for 70 of these transcripts (Supplemental Table S3). Fourteen of these have one or more transmembrane domains, only two of which also contain a signal peptide (Table 1). The putative proteins encoded by these transcripts give weak matches to Claudin, GRAM, DNAJ, and GTPase-like domains. One previously uncharacterized transcript encodes a protein that has the seven transmembrane pass topology characteristic of G protein-coupled receptors (Table 1) and is therefore likely to be a cell surface molecule. The presence of a Claudin domain in another molecule is suggestive of expression at the cell surface, as these domains are seen in, e.g., tight junctions. However, the match is weak (E value 0.02) and only extends over 75% of the Claudin profile and therefore may not be genuine. Of the other transcripts containing open reading frames, 11 apparently encode signaling or adaptor proteins (Table 1), three encode DNA binding do-

main, 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 (Supplemental Table S3).

The remaining 27 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 (Supplemental Table S4). Only seven of the encoded proteins have recognizable domains, i.e., a probable G protein-coupled receptor, two G protein-related signaling 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 are currently being investigated. Thus, in contrast to transcripts encoding the principal T cell markers, 17 out of 18 of which survive the selection procedure, none of the 97 stringently defined CTL-specific but uncharacterized transcripts encode proteins with the modular extracellular domains characteristic of leukocyte surface antigens.

Comparisons with Other Leukocyte Libraries

The similarity of the clone 32 and NK cell libraries (Figure 2C and Supplemental Table S1) prompted comparisons with other, publicly available, leukocyte-derived libraries prepared from ex vivo-activated Th1- and Th2-polarized CD4⁺ cells (Nagai et al., 2001), anti-CD8 bead-purified CD8⁺ cells, and negatively selected NK cells (Obata-Onai et al., 2002). Pearson correlation coefficients derived from comparisons of these libraries, which give a global view of similarities between such datasets, are shown in Figure 5A. Our clone 32 and NK libraries yielded a correlation coefficient (0.783) similar to that for comparisons of ex vivo-purified CD8⁺- and NK cell-derived

Table 1. Uncharacterized CTL-Specific Proteins Containing Putative Transmembrane Helices or Signaling Domains

Tag Sequence	Tag Count (per 100,000)	Unigene Cluster	Accession No.	Domains
Transcripts Encoding Transmembrane Helices				
ACCATTGGAT	103	146360	NM_003641	2 TMH
CATTTACTCT	32	17109	NM_004867	1 TMH
GACTTGGCCT	30	16291	AK057590	2 TMH
ATAAACAGAT	28	334825	AK027658	coiled-coil, 1 TMH
CTCCTCCAAG	27	15284	BC028076	3 TMH
AGCAAGAAAC	19	107393	NM_019895	3 or 4 TMH/Claudin ^a
TGTTGACTCT ^b	19	8882	CAP3 contig 1	Signal peptide, 7 TMH/GPCR-like receptor
TACGAGGCCG	17	16165	NM_007267	10 TMH
TGGGGCCGCA	17	288455	AK026923	7 TMH
GATGAAAAGG	16	343473/172847	NM_005528	DNAJ, 1 TMH
AGGCCACTGG	11	323634	NM_024070	Signal peptide, 1 TMH
GACAGATGGA	11	83575	BC014077	GRAM, 1 TMH
CTTTTCCA	10	259737	NM_020179	1 TMH
GGGCGCCTGG	8	159955	NM_130759	MMR GTPase ^a , 1 TMH
TTCTCAAGAA	8	37189	NM_007069	NLP/P60 ^a , 1 TMH
Transcripts Encoding Signaling Domains				
CACCCAATGG	65	110121	NM_012455	Sec7-like GEF, PH
GAACCGTCCT ^b	47	123164	AW512177	TBC
TAAGGACGAG	33	238707	NM_024901	DENN (AEX-3)
TGCAAGAGAG	30	238954	AL832852	RhoGAP
GGAGCTTGAG	27	288316	NM_022107	Proline-rich, GoLoco
GGTAGAATA	27	61469	NM_018990	SH3, SAM
AGGCTCCGTG	27	196914	D86976	RhoGAP, CDC15, PKC C1
GGCGGGGCCA	17	54985	AB002301/BC003646	Protein kinase (PK), PK extension, PDZ domain
TGCCAATTA ^b	16	165337	AI990569	GTPase (Rab-, Rac-, Ras-, Ran-, or Arf-like)
GTGTTAAATC	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

Transcripts that have been sequenced but whose functions have not been characterized, which 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 Experimental Procedures). Those predicted to have one or more putative transmembrane helices (TMH) by TMHMM2 (run via SMART) or that contain putative signaling domains are listed. Domains found in each protein are also listed in order from the amino to the carboxy terminus.

^aIndicates matches to incomplete domains that might therefore be false hits.

^bTag matched ESTs only.

libraries (0.776) and for libraries previously generated from the same tissue, cerebellum, in separate laboratories (0.781). The clone 32 and ex vivo-purified CD8⁺

cell comparison yielded a similar high correlation coefficient (0.736), which was somewhat larger than that obtained for the comparison of our NK cell line with the

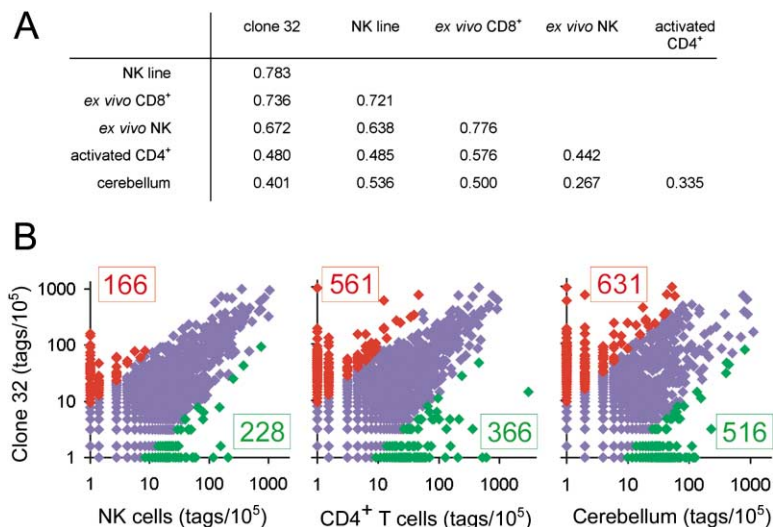


Figure 5. Comparisons of Leukocyte SAGE Libraries

(A) Pearson correlation coefficients for pairwise comparisons of the clone 32 and NK cell libraries with libraries generated from ex vivo-purified CD8⁺, NK, and CD4⁺ cells.

(B) The abundance of each SAGE tag, per 100,000 tags, in the clone 32 library is plotted on a logarithmic scale against those in libraries from CD56⁺ CD3⁻ NK cells (left), combined activated Th1 and Th2 CD4⁺ T cells (center), or normal cerebellum (right). In each case, the numbers of tags ≥ 8 -fold more abundant in clone 32 (red) or the other library (green) are shown.

ex vivo-purified NK cells (0.638). Overall, this analysis confirms the impression obtained by visual inspection of the scatter plots (Figure 2C), i.e., that NK- and CD8⁺ T cell-derived libraries are very similar.

In marked contrast, comparisons of the clone 32 or ex vivo-purified CD8⁺ cell libraries with the CD4⁺ cell library yielded much lower correlation coefficients (0.480–0.576), which overlap those obtained for comparisons with the cerebellum library (0.401–0.500). These differences are borne out in log scatter plots (Figure 5B). Using an 8-fold elevated expression threshold, which seems to represent a cutoff identifying the majority of phenotypically important molecules (Figure 3A), 561 transcripts are preferentially expressed in clone 32 versus the activated CD4⁺ T cell (Figure 5B, center). A comparable number of transcripts (631) are preferentially expressed in clone 32 versus cerebellum (Figure 5B, right), in contrast to 166 transcripts for the NK cell-clone 32 comparison (Figure 5B, left). This implies that tissue-level reprogramming may accompany the lineage commitment or activation, or both, of CD8⁺ and CD4⁺ T cells.

Discussion

We have characterized the expression of a set of 374 genes encoding leukocyte cell surface molecules in a human cytotoxic T cell clone and show that it expresses 111 of these genes. According to our analysis, summarized in Figure 6, the T cell surface appears to be 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. It is likely that the set of polypeptides shown in Figure 6 represents an upper limit to the complexity of the T cell-specific composition of the CTL surface given that 28 of the transcripts are expressed at levels of ~ 3 copies/100,000 or less. We are presently unable to place these molecules within the context of broadly expressed or housekeeping cell surface molecules as our approach was directed at identifying known and unknown immune specific genes. This is nevertheless of considerable interest and is being investigated.

Our most important observation is that of 97 stringently defined, T cell-specific transcripts with no known function, none encode proteins with the modular architecture characteristic of 80% of leukocyte surface antigens. We did find two transcripts encoding unambiguous surface receptors, each with the architecture of G protein-coupled receptors (i.e., seven transmembrane domains). It might be a coincidence that the only new proteins we have found each have this architecture, or perhaps genes encoding these types of receptors are more difficult to identify via conventional cloning methods (e.g., due to more limited immunogenicity). There is no reason to suspect that the 77 additional tags that currently do not give matches to any database corre-

spond to transcripts encoding a higher proportion of cell surface molecules than tags matching uncharacterized sequences, as neither EST nor cDNA sequencing ought to be biased against such transcripts. Overall, therefore, our results strongly suggest that the immune-specific protein composition of the human resting CD8⁺ T cell surface is largely defined.

An obvious and important caveat concerns the extent to which the entire pool of transcripts encoding cell surface molecules was sampled in our 71,174 tag SAGE library. To our knowledge, this is the first analysis of the expression of a very large set of genes encoding cell surface molecules in, effectively, a single cell. We did not anticipate that all of the transcripts encoding each of the principal T cell markers would be detected, or that all but one, CD43, would survive the three-stage selection procedure. The detection of these transcripts by our method is unlikely to be due to uniformly high expression of these particular genes, however, as in general they are moderately expressed and their expression varies by more than two orders of magnitude. Moreover, the abundance of the known transcripts is likely to be representative for cell surface molecules in general since (1) most genes encoding these molecules were identified on the basis of protein abundance or antigenicity rather than high mRNA expression levels, and (2) there is essentially no correlation between protein and transcript levels for medium- to low-abundance transcripts such as those considered here (i.e., the correlation coefficient is ~ 0.4 ; Gygi et al., 1999). In addition, our analysis suggests that, as a group, genes encoding cell surface molecules tend to be highly differentially expressed (Figure 3A). Generally, this level of differential expression is detectable with a high degree of statistical certainty at the depth of our sampling. Therefore, while we cannot formally rule out 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, with the possible exception of CD43, are expressed in this way. 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.

An unrelated issue concerns the extent to which our results, obtained with clonal T cells, are generalizable. We have no evidence that clone 32 is not a typical CTL. It has potent cytotoxic activity and expresses all of the genes encoding the effector molecules known to be present in resting CTL, including the chemokines RANTES, Mip1 α and Mip1 β , granulysin, perforin, and the granzymes A, B, H, K, and M, in addition to all of the expected surface markers. Forty-nine of the fifty most abundant tags in the library generated from ex vivo-purified CD8⁺ T cells (Obata-Onai et al., 2002) are also 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 an alternative allele with a single base substitution. In addition, global gene expression comparisons (Figure 5A) indicate that the similarity of the clone 32 and ex vivo-purified CD8⁺

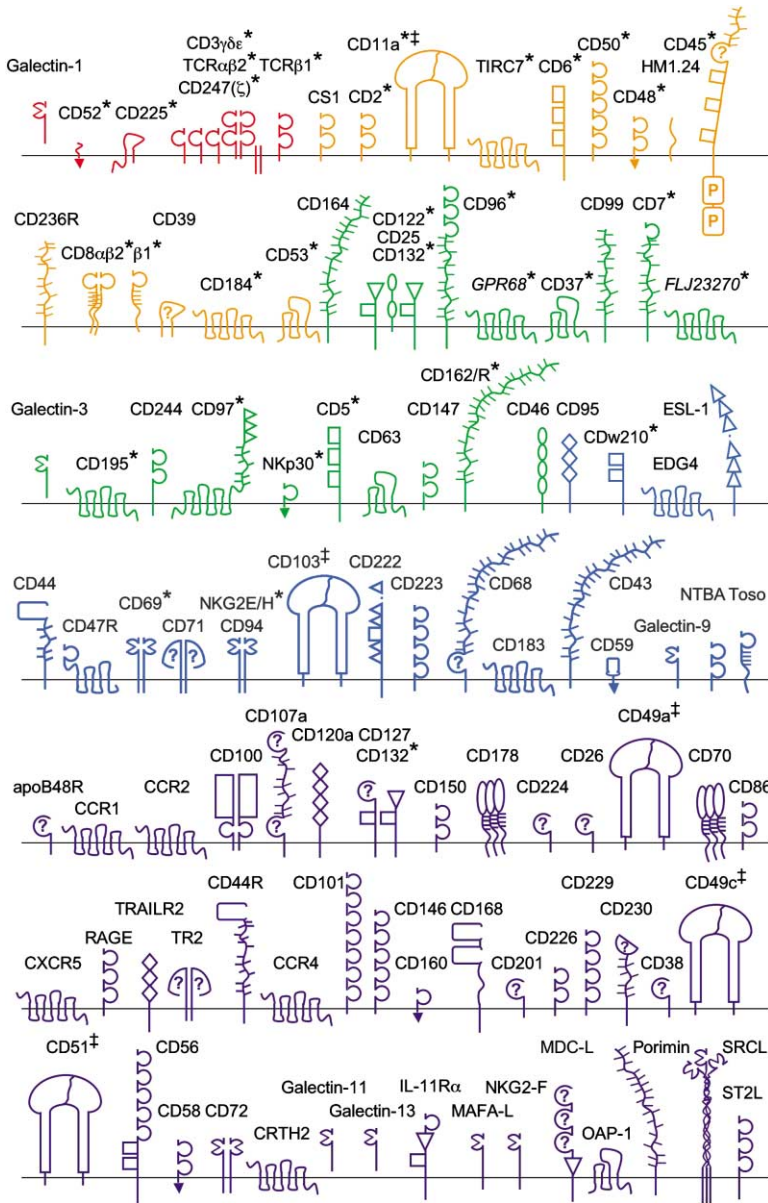


Figure 6. Cell Surface Molecules Encoded by Transcripts Identified in the Clone 32 SAGE Library

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 seven 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. (1997). Unconventional domains or those for which there are no structures are labeled “?”. The molecules are colored 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 $\beta 7$. Tags derived from CD18 and CD61 were found at high levels in the library but matched several other genes, preventing abundance determination. Integrin $\beta 7$ 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.

cell libraries is comparable to that of two CGAP cerebellum libraries prepared in separate laboratories (Riggins and Strausberg, 2001) and greater than that of the two NK cell libraries (Figure 5A). We conclude from this that clone 32 is highly representative of ex vivo CD8⁺ T cells. An important technical point is that three-stage filtering of the ex vivo-purified CD8⁺ T cell library (Obata-Onai et al., 2002) results in the loss of, among other genes, those encoding CD2, CD3 ϵ , CD3 γ , CD8 $\beta 1$, CD8 $\beta 2$, and CD45 from the list of known T cell markers (data not shown). This is only partly explained by genetic differences (i.e., polymorphisms) in the Japanese and African populations and more likely results from heterogeneity in the cell sample used to generate the library (unpublished data). This implies that the effectiveness of our analysis depended on the use of a pure, homogeneous population of cells.

We estimate that, overall, $\sim 45\%$ of moderately to

highly abundant CTL-specific transcripts have unassigned functions, suggesting that relatively large numbers of proteins have yet to be incorporated into models of T cell biology. Sequence data is available for 97 uncharacterized transcripts (Supplemental Tables S3 and S4). Since the three-stage selection process used to identify CTL-specific transcripts significantly enriched for those encoding known cell surface molecules and signaling proteins (Figure 4A), the paucity of transcripts encoding new cell surface molecules and the presence of new signaling transcripts among the 97 uncharacterized transcripts implies that the least well-understood T cell-specific processes involve intracellular signaling molecules. We expect that the 13 new T cell-specific transcripts encoding well-characterized signaling and adaptor domains (including G protein-related, protein kinase, SH2, SH3, and PH domains; Table 1) identified here will have important functions in T cells.

We also find that the differential expression of known tissue-specific genes is stratified according to the broad functions of the proteins that they encode. Thus, genes encoding cell surface antigens and signaling molecules are more highly differentially expressed (≥ 8 -fold) in clone 32 as a group when compared to nonimmune tissues than genes linked to housekeeping functions, e.g., protein synthesis (4- to 8-fold), but less differentially expressed than secreted effector molecules (≥ 32 -fold). Overall, the set of ≥ 8 -fold differentially expressed genes is highly enriched for transcripts encoding proteins responsible for the phenotypic properties of the cell. This degree of stratification, given that it occurs independently of overall transcript abundance, suggests that these sets of genes may be coordinately regulated. Others have speculated that global mechanisms for transcriptional regulation, perhaps involving changes in the transcriptional apparatus or p53, may coordinate gene expression during lymphocyte activation (Teague et al., 1999). It may be relevant to such a process that immune-related genes, particularly those encoding cell surface molecules (e.g., the CD2 subset or killer inhibitory receptors) or components of the major histocompatibility complex, are often highly clustered (Trowsdale, 2001), and that a striking correlation exists between genomic location, gene density, and transcriptional level (Caron et al., 2001). The same higher-order structural features of chromosomes that enhance transcription from gene-dense regions of the genome are likely to promote the coordinated expression of functionally related genes clustered in these regions. In contrast to the tissue-level comparison, the stratification of gene expression in clone 32 versus the NK cell line is far less pronounced (unpublished data), supporting a hierarchical model of cell-specific gene regulation in lymphocytes.

Microarray experiments have been invaluable for identifying sets of coordinately regulated genes, or "expression signatures," defining lineages, differentiation stages, and signaling pathways in lymphocytes. The impression obtained from such studies thus far, however, is that these pathways involve relatively few specific genes. For example, the T cell signature has been defined as consisting of components of the T cell receptor, the signaling proteins LAT, TRIM, and SAP, and the surface markers CD5 and CD2 (Shaffer et al., 2001). Against this background, the observation that the expression of almost 600 genes is increased 8-fold or more in clone 32 versus activated Th1- and Th2-polarized CD4⁺ cells (Nagai et al., 2001) was unexpected. Placing this into context, a similar number of transcripts (~ 650) are ≥ 8 -fold more abundant in clone 32 than in cerebellum. We have yet to establish whether the tissue-level reprogramming apparent in the CD8⁺ and CD4⁺ cell populations is the result of lineage commitment, or activation, or both. Although the robustness of quantitative microarray data is controversial (e.g., Bustin and Dorudi, 2002; Ishii et al., 2000; Kothapalli et al., 2002), it has given an indication of the numbers of genes differentially expressed on T cell activation. The CD4⁺ T cell SAGE library (Nagai et al., 2001) was produced from cells 12–14 days after in vitro stimulation, and microarray experiments by Teague et al. (1999) indicate that after 48 hr of activation in vivo, the pattern of gene expression in T cells returns to one resembling the resting profile, with

only 38 genes (of a total of ~ 2000) remaining ≥ 8 -fold differentially expressed. At face value, this suggests that a significant fraction of the changes we observe may be the result of lineage differences. What is certain is that our analysis does not support the view that T cell subsets share generally homogeneous gene expression profiles with relatively minor differences developing upon activation.

The considerable similarity between the clone 32 and NK cell libraries was also unexpected. Regarding surface markers, the major difference between the NK and clone 32 cells consists of the reciprocal expression of most inhibitory NK receptors on the one hand and the T cell receptor complex on the other (unpublished data). It is currently uncertain whether NK cells only differentiate from bone marrow or whether the common T/NK progenitors described in fetal thymus (in humans and mice) are also an important source of NK cells (Ikawa et al., 1999; Sanchez et al., 1994). Our initial comparisons support the view that the ontologies of CD8⁺ and NK cells may be closely linked.

The present analysis was undertaken in anticipation of the "collision" between model-driven immunology and the discovery-driven field of genomics (Staudt and Brown, 2000). Rather than the incremental addition of new genes and proteins to existing models of biological function, genome sequencing has greatly accelerated the identification of all the molecular components of complex systems, such as T cells. The task of integrating these elements into systematic, quantitative models nevertheless remains a formidable goal. A worthwhile, short-term objective is to identify all the components of important systems and to reconsider the validity of existing functional models within this larger context. We deliberately chose a CD8-independent, cytotoxic T cell clone because it represents the simplest possible T cell receptor-triggering system. Our analysis strongly suggests that all the key cell surface molecules constituting the basic triggering apparatus of this cell have been identified. Existing models of this process are therefore unlikely to be wrong because key elements are missing.

Experimental Procedures

CTL Clone and SAGE Library Generation

Clone 32 was generated from an early HIV-1-specific CTL line grown from the peripheral blood lymphocytes (PBL) of an asymptomatic, untreated HIV-1-infected individual as described (Rowland-Jones et al., 1995). Cytolytic activity was assessed using standard 4 hr ⁵¹Cr release assays. The SAGE library was produced from 5.7 μ g mRNA using the original protocol (Velculescu et al., 1995) with the modifications of Powell (1998). A second SAGE library was produced from CD3⁻ CD56⁺ NK cells purified from PBL using magnetic beads (unpublished data).

Analysis of SAGE Sequences

Tag sequences were extracted from raw sequence data using the SAGE Program Group software. SAGE libraries from normal cerebellum, ovary epithelium, and 12 tumors were produced by CGAP (Riggins and Strausberg, 2001), obtained via SAGEmap (Lash et al., 2000). Libraries from activated CD4⁺ T cells (Nagai et al., 2001), anti-CD8 bead-purified, ex vivo CD8⁺, and negatively selected ex vivo NK cells (Obata-Onai et al., 2002) were obtained from the University of Tokyo (www.prevent.m.u-tokyo.ac.jp/sage.html). SAGE libraries were compared pairwise, and ratios of tag expression between tissues were calculated using a value of 1 per 100,000 for

tags absent in either of the libraries. The statistical significance of the observed differences in expression was calculated using the AC statistic, designed for comparing digital gene expression profiles (Audic and Claverie, 1997), which can, in principle, be used for analyzing differences at all gene expression levels. However, software limitations meant that AC tests could not be used for tags occurring more than 140 times in total. For these tags, the $\chi^2 2 \times 2$ test was used. Pearson correlation coefficients for comparisons of normalized tag abundances between libraries were calculated using SPSS 11.0 (SPSS Inc., Chicago, IL).

Identification of Transcripts Matching SAGE Tags

Tag files were linked to the SAGEmap Tag to UniGene Mapper (Lash et al., 2000). Apparent Unigene matches were rejected if these were based on (1) a single EST, (2) only 5' ESTs, (3) ESTs which had been incorrectly clustered, or (4) ESTs that contain a single base error within the apparent tag sequence, compared to the majority of sequences in the cluster. Matches were also checked to ensure that they had the correct 11th base, where it could be confidently assigned. Transcripts matching each tag were categorized according to their level of characterization and the broad molecular function of their products, based on Unigene (Schuler et al., 1996), Locuslink and Refseq (Pruitt and Maglott, 2001) entries, and information in publications linked to these databases.

Bioinformatic Analysis of Uncharacterized Transcripts

Unigene EST clusters were assembled using CAP3 (Huang and Madan, 1999; run at bio.ifom-firc.it/ASSEMBLY/assemble.html), and the resulting contigs were used in BLASTn and BLASTx searches (Altschul et al., 1997) to identify full-length sequences and possible homologs. For both these sequences and uncharacterized cDNA sequences, the longest forward-reading open reading frame was identified using the NCBI ORF Finder tool (www.ncbi.nlm.nih.gov/gorf), where protein sequence was not already given. Protein sequences were analyzed using BLASTp (Altschul et al., 1997), InterPro (Apweiler et al., 2000), the NCBI conserved domain database (Marchler-Bauer et al., 2002), and SMART (Schultz et al., 2000). Signal peptides and transmembrane domains were predicted via the SMART interface.

Acknowledgments

The authors thank John F. Elliott for helpful early discussions. This work was funded by the Wellcome Trust, the Royal Society, and the United Kingdom Medical Research Council.

Received: January 22, 2003

Revised: April 16, 2003

Accepted: May 21, 2003

Published: August 19, 2003

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.

Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D., et al. (2000). InterPro—an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics* **16**, 1145–1150.

Audic, S., and Claverie, J.M. (1997). The significance of digital gene expression profiles. *Genome Res.* **7**, 986–995.

Barclay, A.N., Brown, M., Law, S.K.A., McKnight, A.J., Tomlinson, M.G., and van der Merwe, P.A. (1997). *The Leucocyte Antigen Factsbook*, Second Edition, Volume One (London: Academic Press).

Bustin, S.A., and Dorudi, S. (2002). The value of microarray techniques for quantitative gene profiling in molecular diagnostics. *Trends Mol. Med.* **8**, 269–272.

Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M.C., van Asperen, R., Boon, K., Voute, P.A.,

et al. (2001). The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* **291**, 1289–1292.

Davis, S.J., and van der Merwe, P.A. (1996). The structure and ligand interactions of CD2: implications for T-cell function. *Immunol. Today* **17**, 177–187.

Grakoui, A., Bromley, S.K., Sumen, C., Davis, M.M., Shaw, A.S., Allen, P.M., and Dustin, M.L. (1999). The immunological synapse: a molecular machine controlling T cell activation. *Science* **285**, 221–227.

Green, C.D., Simons, J.F., Taillon, B.E., and Lewin, D.A. (2001). Open systems: panoramic views of gene expression. *J. Immunol. Methods* **250**, 67–79.

Gygi, S.P., Rochon, Y., Franza, B.R., and Aebersold, R. (1999). Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* **19**, 1720–1730.

Hashimoto, S., Suzuki, T., Dong, H.Y., Nagai, S., Yamazaki, N., and Matsushima, K. (1999). Serial analysis of gene expression in human monocyte-derived dendritic cells. *Blood* **94**, 845–852.

Hogenesch, J.B., Ching, K.A., Batalov, S., Su, A.I., Walker, J.R., Zhou, Y., Kay, S.A., Schultz, P.G., and Cooke, M.P. (2001). A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell* **106**, 413–415.

Huang, X., and Madan, A. (1999). CAP3: a DNA sequence assembly program. *Genome Res.* **9**, 868–877.

Hutloff, A., Dittrich, A.M., Beier, K.C., Eljaschewitsch, B., Kraft, R., Anagnostopoulos, I., and Kroczeck, R.A. (1999). ICOS is an inducible T-cell co-stimulator structurally and functionally related to CD28. *Nature* **397**, 263–266.

Ikawa, T., Kawamoto, H., Fujimoto, S., and Katsura, Y. (1999). Commitment of common T/natural killer (NK) progenitors to unipotent T and NK progenitors in the murine fetal thymus revealed by a single progenitor assay. *J. Exp. Med.* **190**, 1617–1626.

Ishii, M., Hashimoto, S., Tsutsumi, S., Wada, Y., Matsushima, K., Kodama, T., and Aburatani, H. (2000). Direct comparison of GeneChip and SAGE on the quantitative accuracy in transcript profiling analysis. *Genomics* **68**, 136–143.

Kothapalli, R., Yoder, S.J., Mane, S., and Loughran, T.P., Jr. (2002). Microarray results: how accurate are they? *BMC Bioinformatics* **3**, 22.

Lash, A.E., Tolstoshev, C.M., Wagner, L., Schuler, G.D., Strausberg, R.L., Riggins, G.J., and Altschul, S.F. (2000). SAGEmap: a public gene expression resource. *Genome Res.* **10**, 1051–1060.

Marchler-Bauer, A., Panchenko, A.R., Shoemaker, B.A., Thiessen, P.A., Geer, L.Y., and Bryant, S.H. (2002). CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.* **30**, 281–283.

Mason, D.Y., Andre, P., Bensussan, A., Buckley, C., Civin, C., Clark, E., de Haas, M., Goyert, S., Hadam, M., Hart, D., et al. (2001). CD antigens 2001. *Tissue Antigens* **58**, 425–430.

Mochly-Rosen, D. (1995). Localization of protein kinases by anchoring proteins: a theme in signal transduction. *Science* **268**, 247–251.

Nagai, S., Hashimoto, S., Yamashita, T., Toyoda, N., Satoh, T., Suzuki, T., and Matsushima, K. (2001). Comprehensive gene expression profile of human activated T(h)1- and T(h)2-polarized cells. *Int. Immunol.* **13**, 367–376.

Obata-Onai, A., Hashimoto, S., Onai, N., Kurachi, M., Nagai, S., Shizuno, K., Nagahata, T., and Mathushima, K. (2002). Comprehensive gene expression analysis of human NK cells and CD8(+) T lymphocytes. *Int. Immunol.* **14**, 1085–1098.

Powell, J. (1998). Enhanced concatemer cloning—a modification to the SAGE (Serial Analysis of Gene Expression) technique. *Nucleic Acids Res.* **26**, 3445–3446.

Pruitt, K.D., and Maglott, D.R. (2001). RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29**, 137–140.

Riggins, G.J., and Strausberg, R.L. (2001). Genome and genetic resources from the Cancer Genome Anatomy Project. *Hum. Mol. Genet.* **10**, 663–667.

Rowland-Jones, S., Sutton, J., Ariyoshi, K., Dong, T., Gotch, F.,

- McAdam, S., Whitby, D., Sabally, S., Gallimore, A., Corrah, T., et al. (1995). HIV-specific cytotoxic T-cells in HIV-exposed but uninfected Gambian women. *Nat. Med.* *1*, 59–64.
- Ryo, A., Suzuki, Y., Ichiyama, K., Wakatsuki, T., Kondoh, N., Hada, A., Yamamoto, M., and Yamamoto, N. (1999). Serial analysis of gene expression in HIV-1-infected T cell lines. *FEBS Lett.* *462*, 182–186.
- Sanchez, M.J., Muench, M.O., Roncarolo, M.G., Lanier, L.L., and Phillips, J.H. (1994). Identification of a common T/natural killer cell progenitor in human fetal thymus. *J. Exp. Med.* *180*, 569–576.
- Schuler, G.D., Boguski, M.S., Stewart, E.A., Stein, L.D., Gyapay, G., Rice, K., White, R.E., Rodriguez-Tome, P., Aggarwal, A., Bajorek, E., et al. (1996). A gene map of the human genome. *Science* *274*, 540–546.
- Schultz, J., Copley, R.R., Doerks, T., Ponting, C.P., and Bork, P. (2000). SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.* *28*, 231–234.
- Seed, B., and Aruffo, A. (1987). Molecular cloning of the CD2 antigen, the T-cell erythrocyte receptor, by a rapid immunoselection procedure. *Proc. Natl. Acad. Sci. USA* *84*, 3365–3369.
- Shaffer, A.L., Rosenwald, A., Hurt, E.M., Giltner, J.M., Lam, L.T., Pickeral, O.K., and Staudt, L.M. (2001). Signatures of the immune response. *Immunity* *15*, 375–385.
- Shires, J., Theodoridis, E., and Hayday, A.C. (2001). Biological insights into TCR $\gamma\delta^+$ and TCR $\alpha\beta^+$ intraepithelial lymphocytes provided by serial analysis of gene expression (SAGE). *Immunity* *15*, 419–434.
- Staudt, L.M. (2002). Gene expression profiling of lymphoid malignancies. *Annu. Rev. Med.* *53*, 303–318.
- Staudt, L.M., and Brown, P.O. (2000). Genomic views of the immune system. *Annu. Rev. Immunol.* *18*, 829–859.
- Teague, T.K., Hildeman, D., Kedl, R.M., Mitchell, T., Rees, W., Schaefer, B.C., Bender, J., Kappler, J., and Marrack, P. (1999). Activation changes the spectrum but not the diversity of genes expressed by T cells. *Proc. Natl. Acad. Sci. USA* *96*, 12691–12696.
- Trowsdale, J. (2001). Genetic and functional relationships between MHC and NK receptor genes. *Immunity* *15*, 363–374.
- van der Merwe, P.A., Davis, S.J., Shaw, A.S., and Dustin, M.L. (2000). Cytoskeletal polarization and redistribution of cell-surface molecules during T cell antigen recognition. *Semin. Immunol.* *12*, 5–21.
- Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. (1995). Serial analysis of gene expression. *Science* *270*, 484–487.
- Velculescu, V.E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M.A., Bassett, D.E., Jr., Hieter, P., Vogelstein, B., and Kinzler, K.W. (1997). Characterization of the yeast transcriptome. *Cell* *88*, 243–251.
- Velculescu, V.E., Madden, S.L., Zhang, L., Lash, A.E., Yu, J., Rago, C., Lal, A., Wang, C.J., Beaudry, G.A., Ciriello, K.M., et al. (1999). Analysis of human transcriptomes. *Nat. Genet.* *23*, 387–388.
- Williams, A.F. (1977). Differentiation antigens of the lymphocyte cell surface. *Contemp. Top. Mol. Immunol.* *6*, 83–116.
- Williams, A.F., Galfre, G., and Milstein, C. (1977). Analysis of cell surfaces by xenogeneic myeloma-hybrid antibodies: differentiation antigens of rat lymphocytes. *Cell* *12*, 663–673.
- Zelenika, D., Adams, E., Humm, S., Graca, L., Thompson, S., Cobbold, S.P., and Waldmann, H. (2002). Regulatory T cells overexpress a subset of Th2 gene transcripts. *J. Immunol.* *168*, 1069–1079.

Note Added in Proof

Further sequencing of the 3' untranslated region of the CD43 gene has identified an additional SAGE tag. Taking this tag into account, the CD43 transcript now joins the list of those selected by the three-stage filtering process. Therefore, transcripts encoding all 18 classical T cell markers were selected by the three-stage filtering process, implying that the process identifies all cell surface genes with key immune functions in the CTL clone.