Identification of T Cell-Restricted Genes, and Signatures for Different T Cell Responses, Using a Comprehensive Collection of Microarray Datasets

Tatyana Chtanova, Rebecca Newton, Sue M. Liu, Lilach Weininger, Timothy R. Young, Diego G. Silva, Francesco Bertoni, Andrea Rinaldi, Stephane Chappaz, Federica Sallusto, Michael S. Rolph and Charles R. Mackay

*J Immunol* 2005; 175:7837-7847; doi: 10.4049/jimmunol.175.12.7837

http://www.jimmunol.org/content/175/12/7837

---

### Supplementary Material

http://www.jimmunol.org/content/suppl/2005/12/06/175.12.7837.DC1

---

**Why *The JI*?**

- **Rapid Reviews! 30 days** from submission to initial decision
- **No Triage!** Every submission reviewed by practicing scientists
- **Speedy Publication!** 4 weeks from acceptance to publication

*average

---

**References**

This article cites 54 articles, 20 of which you can access for free at:

http://www.jimmunol.org/content/175/12/7837.full#ref-list-1

---

**Subscription**

Information about subscribing to *The Journal of Immunology* is online at:

http://jimmunol.org/subscription

---

**Permissions**

Submit copyright permission requests at:

http://www.aai.org/About/Publications/JI/copyright.html

---

**Email Alerts**

Receive free email-alerts when new articles cite this article. Sign up at:

http://jimmunol.org/alerts
Identification of T Cell-Restricted Genes, and Signatures for Different T Cell Responses, Using a Comprehensive Collection of Microarray Datasets

Tatyana Chtanova,*†† Rebecca Newton,* Sue M. Liu,*†† Lilach Weininger,* Timothy R. Young,* Diego G. Silva,† Francesco Bertoni,‡ Andrea Rinaldi,§ Stephane Chappaz,§ Federica Sallusto,§ Michael S. Rolph,*†† and Charles R. Mackay2*††

We used a comprehensive collection of Affymetrix microarray datasets to ascertain which genes or molecules distinguish the known major subsets of human T cells. Our strategy allowed us to identify the genes expressed in most T cell subsets: TCR αβ+, and γδ+, three effector subsets (Th1, Th2, and T follicular helper cells), T central memory, T effector memory, activated T cells, and others. Our genechip dataset also allowed for identification of genes preferentially or exclusively expressed by T cells, compared with numerous non-T cell leukocyte subsets profiled. Cross-comparisons between microarray datasets revealed important features of certain subsets. For instance, blood γδ T cells expressed no unique gene transcripts, but did differ from αβ T cells in numerous genes that were down-regulated. Hierarchical clustering of all the genes differentially expressed between T cell subsets enabled the identification of precise signatures. Moreover, the different T cell subsets could be distinguished at the level of gene expression by a smaller subset of predictor genes, most of which have not previously been associated directly with any of the individual subsets. T cell activation had the greatest influence on gene regulation, whereas central and effector memory T cells displayed surprisingly similar gene expression profiles. Knowledge of the patterns of gene expression that underlie fundamental T cell activities, such as activation, various effector functions, and immunological memory, provide the basis for a better understanding of T cells and their role in immune defense. The Journal of Immunology, 2005, 175: 7837–7847.

Specialized subsets of effector T cells participate in different types of immune responses. Th1 cells produce IFN-γ and protect against viral pathogens, whereas Th2 cells produce cytokines, such as IL-4 and IL-5, and protect against large extracellular parasites (1). The recently identified follicular B helper T cell (Tfh)3 subset provides help to B cells in B cell follicles (2, 3). All these effector T cell subsets interact with other leukocyte types, including B cells, dendritic cells, and macrophages, as part of a coordinated response to antigenic challenge. In addition, other subsets of T cells provide immune protection at a more primitive level; γδ T cells and NK T cells contribute to early cellular immune responses, particularly innate immunity.

The specialized functions of leukocyte subsets are reflected in the differing patterns of molecules and genes they express. These molecules often underlie the unique function of the subsets they mark. For instance, CD3 defines T cells because it is an essential component of the TCR complex; CD4 is associated with MHC class II recognition, mostly for Th responses, whereas CD8 is associated with MHC class I recognition and cytotoxic responses, and B cells are defined by their Ig expression or production. Immunological research during much of the 1980–1990 era was directed at identifying and characterizing leukocyte markers, first through mAb production, and then through gene sequencing. This led to the discovery of most of the CD molecules that are used today to mark leukocyte subsets and define subset function.

Chemokine receptors and other molecules involved in cell migration have proved useful for defining leukocyte subsets, particularly stages of T cell differentiation and function. CXCR5 marks Tfh cells and facilitates their migration to B cell follicles, where they provide help to B cells (2, 3). Th1 cells preferentially express CCR5 and CXCR3, whereas Th2 cells preferentially express CCR3, CCR4, and CRTh2 (4–6). Most human Th1 and Th2 clones maintain flexibility of cytokine gene expression and, when stimulated under opposite conditions, differentiate to Th0 capable of producing both IFN-γ and IL-4 (7). CCR7 facilitates naïve T cell homing to secondary lymphoid organs (8), but also distinguishes central memory (Tcm) from effector memory (Tem) T cells (9, 10).

More recently, gene microarrays have proven to be a powerful tool for identifying novel gene expression patterns for various subsets of leukocytes (11). We and others have successfully used microarray technology to create gene expression profiles of effector and memory T cell subsets as well as other leukocyte subsets (12–15). In this study we present a comprehensive analysis of most of

---

* Garvan Institute of Medical Research, Darlinghurst, Australia; †Australian National University Medical School, Canberra, Australia; ‡Experimental Oncology, Oncology Institute of Southern Switzerland, Bellinzona, Switzerland; §Institute for Research in Biomedicine, Bellinzona, Switzerland; and ¶Cooperative Research Center for Asthma, University of Sydney, Camperdown, Australia

Received for publication January 21, 2005. Accepted for publication September 27, 2005.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

1 This work was supported by the Clinical Research Center for Asthma, the National Health and Medical Research Council, and the Swiss National Science Foundation (Grant 3100-101962).

2 Address correspondence and reprint requests to Dr. Charles R. Mackay, Garvan Institute of Medical Research, 384 Victoria Street, Darlinghurst, New South Wales 2010, Australia. E-mail address: c.mackay@garvan.org.au

3 Abbreviations used in this paper: Tfh, follicular helper T cell; EAT-2, EWS/FLI1 activated transcript-2; PDBu, phorbol dibutyrate; SH2, Src homology 2; Tcm, T central memory; Tem, T effector memory; XLP, X-linked lymphoproliferative syndrome; IMAP, immunity-associated protein; IAN, immunity-associated nucleotide; GIMAP, GTPase, IMAP family member.
the major subsets of human T cells and identify at the level of gene expression the distinguishing features associated with T cell differentiation to effector subsets, T cell activation, and memory T cell development. In addition, we identify T cell subset-specific gene expression signatures and describe a distinctive transcriptional profile for γδ T cells. Finally, the use of a comprehensive collection of genechip datasets, representing gene expression profiles for all the major human leukocyte subsets, has allowed us to identify numerous genes that are T cell restricted and presumably of relevance for T cell-specific functions.

Materials and Methods

Generation of T cell and leukocyte subsets

The microarray experiments used in this study are summarized in Table I. Detailed descriptions of each microarray experiment are provided at (http:// linkage.garvan.unsw.edu.au/public/microarrays/). Generation of Th1 and Th2, TCM, TEM, and TFH cells has been described previously (15). Resting and activated Th0 and Th2 cells derived from peripheral blood naive T cells were generated as follows. Adult CD4+ naive T cells were isolated from PBMC using a combination of MACS and FACS sorting. Briefly, CD4+ T cells were positively selected with anti-CD4-coated MACS microbeads (Miltenyi Biotec). Cells were then stained with anti-CD45RA Abs, and naive CD45RA- T cells were sorted on a FACSVantage (BD Biosciences). The cells were cultured in RPMI 1640 medium supplemented with 1% Glutamax, 1% sodium pyruvate, 1% nonessential amino acids, 50 μg/ml streptomycin/penicillin (Invitrogen Life Technologies), 5 × 10−5 M 2-ME, and 5% human serum. Single naive T cells were distributed in 96-well plates by FACSVantage sorting and were stimulated with 105 allogeneic irradiated (40 Gy) PBMC and 1 μg/ml PHA in IL-2-containing medium in the presence of 2 ng/ml rIL-4 and 1 μg/ml neutralizing anti-IL-12 Abs (R&D Systems). Clones were expanded in IL-2-containing medium. A cytokine production assay was performed after −2 wk. After 8 wk, selected T cell clones were restimulated under the same Th2 conditions to generate Th2 clones or under opposite Th1 conditions (0.5 ng/ml rIL-12) to generate Th0 clones. Clones were expanded in IL-2-containing medium. A cytokine production assay was performed after 2 wk. RNA was extracted from resting and activated T cell clones after 4 wk using TRIzol (Invitrogen Life Technologies).

For cytokine production assays, T cells were stimulated with 2 × 10−7 M PMA and 1 μg/ml ionomycin (Sigma-Aldrich) for 4 h. Brefeldin A (10 μg/ml) was added for the last 2 h. Fluorochrome-labeled anti-IFN-γ and anti-IL-4 Abs (BD Pharmingen) were used after fixation and permeabilization, performed using the Cytofix/Cytoperm kit (BD Pharmingen). T cells (10 6/condition) were left untreated or were activated for 4 h with 1 μg/ml anti-CD3 Ab (clone TR66) and 50 ng/ml phorbol dibutyrate (PdBu; Sigma-Aldrich).

Preparation of cRNA and genechip hybridizations

Total RNA was isolated from cells using the RNeasy Total RNA Isolation Kit (Qiagen) or TRIzol according to the manufacturer’s instructions. cRNA was prepared as previously described (16). Briefly, cDNA was specifically transcribed from 500 ng of mRNA using a poly-T nucleotide primer containing a T7 RNA polymerase promoter (Geneworks). Biotinylated, antisense target cRNA was subsequently synthesized in vitro transcription using the BioArray High Yield RNA Transcript Labeling kit (Enzo Diagnostics). Fifteen micrograms of biotin-labeled target cRNA was then fragmented and used to prepare a hybridization mixture, which included probe array controls and blocking agents. Hybridization to U133A and B Affymetrix arrays was conducted for 16 h at 45°C and 60 rpm. After hybridization, washing and staining of the hybridized probe array were performed by an automated fluids station, according to the manufacturer’s protocols. The stained probe array was scanned using the Agilent GeneArray Laser Scanner, and the resultant image was captured as a data image file.

Data analysis

Absolute levels of expression of genes were determined and scaled to 150 using algorithms in MicroArray Analysis suite 5.0 (MAS5) software (Affymetrix). The signal value represents the level of expression of a transcript. The signal log ratio is the change in expression level of a transcript expressed as the log2 ratio (a signal log2 ratio of 1 is equal to a fold change of 2). Annotations were extracted using the NetAffx analysis tool (www.affymetrix.com).

### Table I. Microarray experiments used in this study

<table>
<thead>
<tr>
<th>Leukocyte Subset</th>
<th>Cell Type</th>
<th>Source</th>
<th>U133A+B Arrays</th>
</tr>
</thead>
<tbody>
<tr>
<td>T cells</td>
<td>Naive cord blood</td>
<td>CD4+CD45RA- T cells isolated from cord blood</td>
<td>1</td>
</tr>
<tr>
<td>T&lt;sub&gt;CM&lt;/sub&gt;</td>
<td></td>
<td>CCR7+CD4+CD45RO+ T cells isolated from PBMC</td>
<td>2</td>
</tr>
<tr>
<td>T&lt;sub&gt;EM&lt;/sub&gt;</td>
<td></td>
<td>CCR7+CD4+CD45RO+ T cells isolated from PBMC</td>
<td>2</td>
</tr>
<tr>
<td>T&lt;sub&gt;TH&lt;/sub&gt;</td>
<td></td>
<td>CD57+CXCR5+CD4+ T cells isolated from tonsil</td>
<td>2</td>
</tr>
<tr>
<td>cbTh1</td>
<td></td>
<td>Cord blood CD4+ T cells polarized under Th1 conditions</td>
<td>2</td>
</tr>
<tr>
<td>cbTh2</td>
<td></td>
<td>Cord blood CD4+ T cells polarized under Th2 conditions</td>
<td>2</td>
</tr>
<tr>
<td>pbTh0TNA</td>
<td></td>
<td>Th2 clones polarized under Th1 conditions</td>
<td>2</td>
</tr>
<tr>
<td>pbTh2TNA</td>
<td></td>
<td>T cell clones polarized under Th2 conditions</td>
<td>3</td>
</tr>
<tr>
<td>pbTh0ACT</td>
<td></td>
<td>Th2 clones polarized under Th1 conditions, activated</td>
<td>3</td>
</tr>
<tr>
<td>pbTh2ACT</td>
<td></td>
<td>T cell clones polarized under Th2 conditions, activated</td>
<td>3</td>
</tr>
<tr>
<td>γδ T cells</td>
<td></td>
<td>γδ T cells isolated from PBMC by FACS</td>
<td>2</td>
</tr>
<tr>
<td>CD8&lt;sup&gt;+&lt;/sup&gt; T&lt;sub&gt;EM&lt;/sub&gt;</td>
<td></td>
<td>CD8&lt;sup&gt;+&lt;/sup&gt;CCR7+CD45RO+ T cells isolated from PBMC</td>
<td>2</td>
</tr>
<tr>
<td>CD8&lt;sup&gt;+&lt;/sup&gt; T&lt;sub&gt;CM&lt;/sub&gt;</td>
<td></td>
<td>CD8&lt;sup&gt;+&lt;/sup&gt;CCR7+CD45RO+ T cells isolated from PBMC</td>
<td>1</td>
</tr>
<tr>
<td>NK</td>
<td>CD16+CD56&lt;sup&gt;+&lt;/sup&gt;</td>
<td>CD16&lt;sup&gt;+&lt;/sup&gt;CD56&lt;sup&gt;+&lt;/sup&gt; cells isolated from PBMC by FACS</td>
<td>2</td>
</tr>
<tr>
<td>B cells</td>
<td>Naive</td>
<td>Isolated from spleen by FACS</td>
<td>2</td>
</tr>
<tr>
<td>IgG IgA IgE memory</td>
<td></td>
<td>Isolated from spleen by FACS</td>
<td>2</td>
</tr>
<tr>
<td>IgM memory</td>
<td></td>
<td>Isolated from spleen by FACS</td>
<td>2</td>
</tr>
<tr>
<td>Plasma</td>
<td></td>
<td>Isolated from spleen by FACS</td>
<td>2</td>
</tr>
<tr>
<td>CD19</td>
<td>CD19&lt;sup&gt;+&lt;/sup&gt; B cells-isolated from PBMC by FACS</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Dendritic cells</td>
<td>Immature</td>
<td>Monocyte derived DC</td>
<td>2</td>
</tr>
<tr>
<td>6 h LPS stimulated</td>
<td>Immature DC activated in the presence of 100 ng/ml LPS for 6 h</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>48 h LPS stimulated</td>
<td>Immature DCs activated in the presence of 100 ng/ml LPS for 48 h</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Macrophages</td>
<td>Resting</td>
<td>Monocyte derived</td>
<td>2</td>
</tr>
<tr>
<td>4 h LPS treated</td>
<td>Macrophages activated for 4 h in the presence of 100 ng/ml LPS</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Eosinophils</td>
<td>Resting</td>
<td>Eosinophils derived from peripheral blood by MACS</td>
<td>2</td>
</tr>
<tr>
<td>Activated</td>
<td>Eosinophils activated for 2 h with PMA</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Neutrophils</td>
<td>Resting</td>
<td>Neutrophils isolated from peripheral blood by MACS</td>
<td>2</td>
</tr>
<tr>
<td>Activated</td>
<td>Neutrophils activated for 1 hr with LPS</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Mast cells</td>
<td>Resting</td>
<td>Cord blood derived mast cells</td>
<td>2</td>
</tr>
<tr>
<td>Activated</td>
<td>Mast cells activated through IgE receptor</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Basophils</td>
<td>Resting</td>
<td>Basophils isolated from PBMC by FACS</td>
<td>1</td>
</tr>
</tbody>
</table>
that showed a change of 2-fold or greater were considered differentially expressed. Software was used to map genechip analysis results where fluorescence intensity detection was arranged in a hierarchy and were displayed in a phylogenetic tree of clusters of genes in a hierarchically ordered relationship. Branch lengths represent the degree of similarity between sets. Gene expression profiles that were similar across the experimental samples were clustered together. Principal component analysis using GeneSpring software was performed using all genes that were expressed in at least two samples, to identify the components responsible for the greatest variability.

Data were imported as a Microsoft Excel file into Spotfire for graphical representation of gene expression patterns in TCM and TEM cells. Spotfire software was used to map genechip analysis results where fluorescence intensity detected on the genechips is represented by a color scale. Genes that showed a change of 2-fold or greater were considered differentially expressed.

**Real-time PCR validation of gene expression**

Total RNA was isolated from purified cells as described. RT was performed on 200 ng of total RNA per reaction using the Reverse-IT RTase Blend Kit (ABgene) according to the manufacturer’s instructions. cDNA synthesis, semiquantitative real-time RT-PCR was performed with FastStart DNA Master SYBR Green I Reagent by LightCycler (Roche), as previously described (17). Primers were designed using Primer3 software (18). mRNA expression was normalized to GAPDH, and the fold change was calculated relative to Th1 gene expression.

**Results**

Comprehensive leukocyte subset profiling allows the identity of T cell-restricted genes

T cells differ from other leukocyte subsets in their phenotypic and functional properties, which is reflected by different gene expression patterns. We have produced a large dataset of genechip expression profiles for all the major leukocyte subsets, including various subsets of B cells, plasma cells, NK cells, eosinophils, neutrophils, basophils, dendritic cells, macrophages, mast cells, and all major effector and memory T cell subsets (Table I).

The first question we asked was whether this comprehensive dataset of genechips for various leukocyte types might allow us to identify T cell-specific transcripts. We applied one-way ANOVA to identify genes that most reliably discriminate between T cells and other leukocytes (Fig. 1A). We also used the publicly available SymAtlas database (http://symatlas.gnf.org/SymAtlas/) (19) to assess the expression of gene transcripts in many nonlymphoid tissues. We identified a large number of genes that were preferentially expressed by T cells compared with non-T cell leukocytes (>1200 probe sets on the Affymetrix U133 A+B arrays; p < 0.0001). The top ~100 genes most intimately associated with T cells are listed (Fig. 1A), and ~400 additional genes selectively expressed by T cells (p < 0.00001) are provided as supplementary information (Table I). As expected, some T cell-restricted genes were expressed outside of the hemopoietic lineage, whereas others were highly specific to T cells (Fig. 1A). Interestingly, we found that many genes were shared between T and NK cells, indicating that the origin and functional properties of these two leukocyte types are closely related. Importantly, many of the genes identified through this approach (Fig. 1A) were well-characterized, T cell-specific genes, such as CD3 chains, CD28, ICOS, CD40L, TCR, Lck, Zap70, and GATA3, which confirmed the validity of our approach. This gave greater credence and reliability to the numerous additional genes that have not yet been necessarily associated with T cell-selective roles. We selected several genes and confirmed their T cell-specific expression using real-time PCR (Fig. 1B). These genes included IL-32 (previously known as NK-4), a recently identified inflammatory cytokine (20); BCL11b, a transcriptional repressor that controls T cell development (21, 22); EOMES, a paralog of T-bet important during embryogenesis (23), but also able to regulate CD8$^+$ effector function (24); and UBASH3A (also known as suppressor of TCR signaling-2), which is a regulator of signaling pathway involved in T cell activation. We also confirmed T cell-specific expression of two GTPase members of the immunity-associated protein (IMAP) (or immunity-associated nucleotide (IAN) family GTPase, IMAP family member (GIMAP5) (hIAN5 and IAN4L) and GIMAP7 (hIAN7). Interestingly, we also noted preferential expression of GIMAP6 (hIAN2) by T cells (supplementary Table I). This family is relatively poorly characterized to date; however, members of this family seem to be highly expressed in the spleen and lymph nodes, and there is increasing evidence of their importance in T cell development and maturation (25–27). Fig. 1A shows that in addition to genes specifically expressed by T cells, there were also many genes that were specifically absent from T cells compared with other leukocyte subsets. Not surprisingly, these included several genes associated with Ag processing and presentation, endocytosis, complement receptors, and numerous other genes. It is beyond the scope of this study to discuss all the novel genes identified as T cell specific; however, the important point of this work is that bioinformatic strategies and large gene expression datasets provide a novel way of identifying T cell-biased molecules.

One of the processes unique to T cells is TCR signaling. We found >150 TCR signaling or costimulation related genes that were significantly preferentially expressed in T cells (supplementary Table II). Selected genes involved in the development of the immunological synapse and downstream signaling in T cells are shown in Fig. 1C. The majority of genes involved in T cell signaling cascades were restricted to T cells, although some were expressed by other leukocytes. Interestingly, many of the T cell signaling molecules were expressed at a higher level in resting, rather than activated, T cells, which was noted in previous studies (28, 29) and could be linked to down-modulation of the TCR signaling apparatus after stimulation (30, 31).

**Gene expression signatures of T cell subsets**

After identification of all the genes that distinguish T cells from other leukocytes, we focused on the differences between the T cell subsets. Our dataset of genechip profiles contains a diverse array of T cell subsets, including resting and activated Th0 and Th2 cells; naive, central, and effector memory CD4$^+$ T cells; CD8$^+$ T cells; and αβ and γδ T cells (Table I). A one-way ANOVA with Student-Newman-Keuls post-hoc testing was used to identify the genes that distinguish particular subsets of T cells from all others. NK cells were also included in this study, because we noted previously that T cells share many genes with NK cells. Unsupervised clustering of these genes (Fig. 2A) revealed easily identifiable and distinct gene expression signatures that distinguished each subset. For instance, the T_fol specific signature consisted of well-known markers of this subset, such as CXCR5 and CXCL13, and many genes not previously associated with this subset (2, 3). Of note, gene expression signatures were not only characterized by highly preferentially expressed genes, but also by the genes whose expression was particularly low or absent in a particular subset, compared with others. This was especially true among closely related
subsets, where few genes were expressed exclusively by a single subset. For instance, we identified few genes that were specific to TEM or TCM cells, and these memory T cell populations appeared to be highly related based on the limited differences in gene expression between the two. We have listed select genes from signature clusters, including genes up- and down-regulated in TEM cells (Fig. 2A, i and v), down-regulated in γδ T cells (Fig. 2A, ii), and down-regulated in TCM (Fig. 2A, v) and TEM (Fig. 2A, vii) and others. Genes contributing to Th1 and Th2 signatures have been discussed previously (15). Importantly, gene expression profiling of the numerous T cell subsets allowed us to distinguish clear and discriminating signatures based on the presence and the absence of certain transcripts.

A signature for γδ T cells

A perplexing question in T cell biology is the relevance and functions of the γδ T cell subset (32). Certain transcriptional profiling approaches such as serial analysis of gene expression (33, 34) have identified preferentially expressed genes in the mouse; however, to date, sheep and cattle are the only species in which γδ T cells are known to express a distinctive marker, T19 (35). We sought to identify genes specific to human γδ T cells that might underlie a function unique to this enigmatic subset. Surprisingly, we found few genes expressed specifically by γδ T cells. A number of genes highly expressed by γδ T cells were also expressed by NK cells, in particular, a number of KIR genes (Fig. 2A, iv). The expression of these genes, however, was not confined to NK and γδ T cells, because some αβ effector T cells also expressed transcripts for these molecules. The gene expression signature of γδ T cells could be distinguished more easily by low or absent expression of a number of αβ T cell transcripts, rather than by the expression of any γδ-specific genes (Fig. 2A, ii). Even the γ-chain of the TCR was expressed in some αβ T cells. Among the genes down-regulated in γδ T cells were several genes involved in pre-mRNA splicing processes, including transformer 2β, suppressor of white apricot homologue 2 (splicing factors, arginine/serine rich 10 and 16, respectively), and pre-mRNA splicing factor 16 (the conserved motif Asp-Glu-Ala-Asp/H box polypeptide 38). The down-regulation of these splicing factors, compared with that in αβ T cells, may be due to γδ T cells not requiring extensive alternative splicing of the TCR. It should be noted that to date, only blood γδ T cells have been assessed, not those from epithelial or lymphoid tissues.

T cell subset predictor genes

As an extension of our signature gene analysis for the different T cell subsets, we next sought to identify small numbers of predictor genes that could reliably distinguish between all the different T cell subsets in this study. We selected one or two genes with the lowest p value from each subset that most reliably distinguished that particular T cell subset from all other T cells, creating a list of 13 genes (Fig. 2B). We found that using this small subset of genes it was possible to correctly identify the T cell type used in each microarray experiment using a k-nearest neighbors method. As few as 13 genes were sufficient to distinguish between 11 different T cell subsets and NK cells, demonstrating the remarkable power of this approach for cell type classification. Interestingly, the genes that made up the optimal predictor set were not necessarily well-known markers and included intracellular signaling molecule (such as phosphatases (DUSP4)) and obscure uncharacterized genes (such as BC006146 and A1766311, which do not show homology to any known gene families). These results demonstrate the utility of gene expression profiling for T cell type classification, similar to the way in which predictor genes have been used to distinguish cancer subtypes (36–39). An important qualification, however, is that a much larger number of replicates will be required for a more reliable classification of unknown subsets of T cells, and probably the predictor set of genes will be refined with increased sample numbers for each subset and additional subsets, such as regulatory T cells. Nevertheless, our analysis demonstrates the potential of microarray dataset analysis for identification of unexpected molecules as biomarkers for T cell subset classification.

T cell activation induces an extensive transcriptional program

T cell subsets examined in this study represent many of the major paradigms of T cell biology, for instance, TCR γδ vs αβ T cells, Th1 vs Th2, and T cell activation and memory cell development. Principal component analysis was undertaken to identify the most distinctive expression patterns in all our T cell subsets. Principal component analysis involves a mathematical procedure that transforms a number of possibly correlated variables (in this case, gene expression data points) into a smaller number of uncorrelated variables, called principal components. The first three principal components, which accounted for more than half the variability in gene expression, clearly separated T cells based on the isolation method and activation state. Thus, T cell activation and culture were the two major influences on gene expression patterns in the T cell subsets in this study (Fig. 3). This was not totally surprising, because previous studies have documented the profound changes in T cell transcription upon activation (28).

To identify transcriptional changes induced by T cell activation, we analyzed gene expression in Th0 and Th2 cells before and after activation in vitro. TCR engagement (anti-CD3 and PdBu treatment) was accompanied by a vast change in gene expression in Th0 and Th2 cells. The majority of genes affected by activation...
were regulated in a similar manner in both Th0 and Th2 cells (Fig. 4A). Activation induced an extensive change in gene expression, with >7000 genes significantly different between resting and activated T cells. A detailed description of many (or even some) of the individual genes is not possible in this study; however, global analyses that reflects changes in biological processed can be more

### Figure 2.

Distinct gene expression signatures characterize effector/memory T cell subsets. Gene expression profiles of different T cell subsets were analyzed using one-way ANOVA with Benjamini multiple testing correction and Student-Newman-Keuls post-hoc tests to identify genes with significant changes in expression among the subsets. A, The union of all genes that distinguish the various cell types was created by combining genes that differentiate a single subset from all others. Unsupervised clustering using Pearson similarity measure was used to arrange the genes (y-axis) and microarrays (x-axis).

**B**

![Diagram showing gene expression patterns](image-url)
Gene expression in memory T cell subsets: $T_{CM}$ vs $T_{EM}$

Gene signatures for T cell subsets are not necessarily informative for revealing all the differences between two particular subsets when a proportion of the genes is shared by numerous other T cell types. One interesting subdivision of the memory pool is the TCM when a proportion of the genes is shared by numerous other T cell types. For instance, naive T cell-specific genes (at least among the leukocytes) were also expressed in brain (approximately five genes; Fig. 1A); perhaps the genes serve the same function in both cell types, but, more likely, the T cell immune system may have simply adopted receptors and ligands from other systems. For instance, naive T cells use CXCR4 for cell migration; however, CXCR4 is a widely expressed molecule and in evolution system may have simply adopted receptors and ligands from other systems. One of the main aims of this study was to establish datasets and bioinformatic strategies for the identification of new T cell-restricted genes that probably play important functional roles in particular T cell subsets.

An in-depth examination of gene expression profiles of numerous effector and memory T cell subsets revealed specific signatures that distinguished each subset. Interestingly, some subsets were more easily distinguished by the genes that were down-regulated or absent. We also showed that T cell subsets could be easily distinguished from each other through the use of a small number of predictor genes. Previous studies have demonstrated the application of gene expression profiling to distinguish subtypes of neoplastic disease (36–39), and gene microarrays have been used for more precise diagnosis and subgrouping of cancers (43–45). Moreover, gene microarrays have proven useful for predicting disease outcomes, particularly for cancers (46–49). Similarly, leukocyte-specific signatures can be used for classification of unknown cells or cancer cell types based on their gene expression patterns.

**Discussion**

In this study we used an extensive collection of Affymetrix gene-chip datasets, representing gene expression profiles of various leukocyte and T cell subsets. This dataset collection allowed us to identify T cell-specific gene transcription based on gene expression in the T cell subsets and absence in the non-T leukocyte subsets. Importantly, this signature comprised both well-characterized markers of T cell function and differentiation as well as a number of more obscure genes previously not associated with T cell biology. Several of these more obscure genes were validated by PCR, and we (15) and others (12) generally found a reasonable concordance between Affymetrix expression data and PCR results. Thus, we believe that our approach is valid for the identification of additional factors that are intimately involved in the biology of T cells. Of the numerous T cell-expressed genes that were absent from all other leukocyte types, many, in fact, were expressed in other tissues. For instance, some T cell specific genes (at least among the leukocytes) were also expressed in brain (approximately five genes; Fig. 1A); perhaps the genes serve the same function in both cell types, but, more likely, the T cell immune system may have simply adopted receptors and ligands from other systems. One of the main aims of this study was to establish datasets and bioinformatic strategies for the identification of new T cell-restricted genes and molecules. The relevance of many of these genes for T cell responses must await studies with gene-deficient mice or other systems that suitably address molecular functions in T cells. Regardless, the datasets described in this study should focus the attention of T cell biologists to many poorly characterized genes that probably play important functional roles in particular T cell subsets.
Although leukocytes were previously characterized by their expression of cell surface markers, gene expression by microarrays provides a new basis for leukocyte subset classification. Precise signatures for many of the T cell subsets examined in this study must await further analyses with a greater numbers of genechips per subset. However, we established that the use of predictor signatures is applicable to T cell subset classification, and we expect that the predictor genes reported in this study will be refined to accurately distinguish various stages of T cell differentiation and function.

The γδ T cell subset remains a largely enigmatic population. Recent attempts to identify a specific transcriptional profile for γδ T cells (34, 50–52) revealed many interesting features of γδ T cells, such as their activated, yet resting, phenotype (34); however,
a unique γδ molecular profile remains to be identified. Our study, which is one of the first to examine by microarray gene expression the profiles of circulating human γδ T cells, showed that γδ T cells shared many features with αβ T cells and NK cells, yet expressed few, if any, molecules that were γδ specific. This was surprising, because a γδ-specific marker, termed T19 (aka WC1), has been identified in sheep and cattle (35). This result was, however, in accordance with the results of a recent study that showed that γδ T cells in mice share many features with an unconventional subset of αβ T cells (33). Although we were unable to identify any γδ-specific markers, their gene expression signature did provide some interesting insights into the γδ T cell biology. Although γδ T cells have the potential for extensive receptor diversity, it appears that this potential is not fully realized (53). We found that, compared with conventional αβ T cells, γδ T cells showed reduced expression of several proteins involved in pre-mRNA splicing, which could explain the lack of TCR diversity in these cells. The failure to identify any γδ-specific transcripts confounds attempts to resolve the functional relevance of this subset. In accordance with mouse studies, our human expression data support the view that αβ and γδ subsets show overlapping molecular and functional profiles, and the reason why different vertebrates show varying levels of γδ T cells in blood or at epithelial surfaces, especially around birth, is still unanswered.

Memory T cells have been divided into effector and central subsets based on their expression of homing molecules as well as other attributes. TEM cells express tissue-homing receptors and provide immediate effector function after stimulation, whereas...
T_{CM} cells express lymphoid-homing receptors and provide long-term systemic protection (9). We found surprisingly few differences between T_{EM} and T_{CM} cells isolated from human peripheral blood, at least at the level of gene expression. T_{EM} did express higher levels of several molecules associated with T cell effector function, whereas T_{CM} cells expressed, predictably, higher levels of genes encoding CCR7 and CD62L (L-selectin). The similar gene expression profiles of T_{EM} and T_{CM} suggest that these two memory subsets may serve largely similar functions, albeit in different locations. The functional distinction between T_{EM} and T_{CM} subsets is probably less clearly defined than previously thought, because T_{CM} cells in mice are capable of providing immediate protection (54). In our study we examined gene expression profiles of circulating memory T cells; however, it is possible that T_{EM} cells in tissues display a more distinctive phenotype, closer to that of an effector cell. Others have shown that T_{EM} and T_{CM} cells are heterogeneous (10) and include some Th1- and Th2-like cells as well as possibly some NK T and γδ T cells and skin- and gut-homing T cells. With recent advances in cRNA production techniques and multicolor cell sorting, isolation of even smaller subsets of T_{EM} and T_{CM} cells should be feasible, and this should provide additional insight into the nature of immunological memory in the T cell system.

One reservation regarding our approach is that not every conceivable leukocyte subset, at all stages of differentiation and stimulation, was included in our datasets. For instance, CD5 came up with ours. The datasets used in this study are freely available for download from (http://linkage.garvan.unsw.edu.au/public/microarrays/) and will also be deposited in public databases. Because of the large numbers of genes identified in the various analyses described in this study, it would be impossible to validate all of these, and we urge investigators with an interest in a particular gene to validate, by PCR methods, genes of interest.

In conclusion, we have used an extensive dataset of leukocyte gene expression profiles to identify genes expressed selectively by T cells and to gain an understanding of the molecular changes that accompany T cell activation and differentiation. The identity of numerous novel genes selectively expressed by T cells will provide important insight into T cell function in general as well as the specialized roles of subsets of T cells. In addition, strategies used in this study for T cell and leukocyte microarray dataset analysis clearly identify interesting subset-restricted genes, which should provide important new insight into T cell functions.

Acknowledgments
We thank Stuart Tangye, Kim Good, Mary Sisavan, Sabine Zimmer, Melinda Frost, and Trina So for the use of microarray data, and Mark Hughes for help with data analysis. We also thank Ian MacKay for helpful suggestions.

Disclosures
The authors have no financial conflict of interest.

References


