

Minireview

## Scratching the (T cell) surface

Joerg Ermann, Chan D Chung and C Garrison Fathman

Address: Division of Immunology and Rheumatology, Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA.

Correspondence: C Garrison Fathman. E-mail: cfathman@stanford.edu

Published: 11 December 2003

*Genome Biology* 2003, **5**:202

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2003/5/1/202>

© 2003 BioMed Central Ltd

### Abstract

Using a genome-scale approach to study transcription levels in a human CD8<sup>+</sup> T-cell clone, a recent study has suggested that the repertoire of molecules on the surface of T cells is close to being completely characterized.

Over the last few years technological advances have made it possible to study, in parallel, the expression of thousands of genes in cells, tissues or organisms. While this genome-scale approach to gene-expression analysis has been touted by some as the new 'golden era' of hypothesis-unlimited discovery-driven research [1], it is apparent that our ability to make sense of the vast accumulations of data does not always keep up with our ability to generate them. Many groups have used technologies such as microarrays or serial analysis of gene expression (SAGE) to analyze gene expression in cells of the immune system. A recent article by Evans *et al.* in *Immunity* [2] attempting to expand beyond the simple description of differential gene-expression patterns arrives at a bold conclusion. The article's title is intriguing - *The T-cell surface - how well do we know it?* - and the conclusion is that we know it quite well.

Not surprisingly, molecules on the surface of T cells are of great interest to immunologists. Much information has been generated about T-cell-specific surface molecules and their function since the first monoclonal antibodies against leukocyte surface markers were made in the late 1970s. Attempts to compare the specificities of different monoclonal antibodies and to identify their targets led to the development of the cluster of differentiation (CD) nomenclature [3]. Most of the CD antigens turned out to be proteins (a few CD antibodies detect carbohydrate modifications) and the respective genes have been cloned in both humans and mice. Over time the CD system thus transformed into a classification for leukocyte

surface molecules, rather than antibodies. Currently, there are 247 assigned CDs [4]. Almost 200 additional molecules are being considered for CD status during the current Eighth International Workshop on Human Leukocyte Differentiation Antigens that will culminate in the HLDA8 Conference in Adelaide, Australia in December 2004 [5].

The molecules on the surface of T cells belong to very diverse structural and functional classes, and include components of the immunoreceptors on T, B and NK cells, adhesion molecules, and cytokine and chemokine receptors. Some CDs have proven to be useful markers of subpopulations of cells with strikingly different functions. Most T cells express CD3 and an  $\alpha\beta$  T-cell receptor (TCR) paired with either the CD4 or the CD8 molecule as co-receptor. CD4<sup>+</sup>  $\alpha\beta$  T cells recognize antigens presented by antigen-presenting cells in the context of major histocompatibility complex (MHC) class II molecules. Their main effector function is the production of cytokines and the facilitation of immune responses. A subclass of CD4<sup>+</sup> T cells that has recently gained attention comprises regulatory/suppressor T cells, which are characterized by the constitutive expression of CD25 - the  $\alpha$ -chain of the interleukin 2 (IL-2) receptor. These cells are thought to negatively regulate immune responses and to prevent uncontrolled autoimmunity. CD8<sup>+</sup>  $\alpha\beta$  T cells recognize antigens in the context of MHC class I molecules, which are expressed on most somatic cells. Upon activation, CD8<sup>+</sup> T cells develop into cytolytic T lymphocytes (CTLs) ready to kill cells infected by intracellular pathogens, such as viruses, or eradicate tumors cells.

In their article on the T-cell surface, Evans *et al.* [2] used SAGE to study gene expression in a human CD8<sup>+</sup> CTL clone. The SAGE library that was generated (referred to as the CTL library) contained 71,174 SAGE tags representing 20,204 distinct sequences. This number was estimated to cover all of the transcripts whose expression level was at or above 0.008% of the transcriptome (that is having at least 22 copies per cell). The library included 111 genes with, or being considered for, CD status. Several pairwise comparisons with unrelated SAGE libraries were then performed.

The central analysis of the paper starts with a comparison of the CTL library with a library derived from cerebellum; 1,098 transcripts were significantly more abundant in the CTL library. Among these, about a quarter of the transcripts with known function coded for proteins involved in protein/mRNA synthesis, a result that was thought to reflect the proliferating versus non-proliferating character of the two cell/tissue types (CTL versus cerebellum). Interestingly, the set of genes that was highly differentially expressed in the CTLs was enriched for surface markers, signaling molecules and soluble mediators. In an attempt to find a core set of CTL-specific genes, additional comparisons were performed between the CTL library and SAGE libraries from ovary epithelium (as a type of proliferating cell) and a panel of tumor libraries. This resulted in a shortened list of 387 CTL-specific transcripts.

Notably, at all stages of comparison 42-45% of the transcripts lacked an assigned function. Of the known genes in the final list of 387 specific transcripts in the CTL library, 27% were cell-surface molecules, including TCR components, CD2, CD5, and CD8. Evans *et al.* [2] then asked how many of the unknown CTL-specific transcripts encoded surface molecules. Sequences representing UniGene clusters [6] were analyzed for signatures of surface molecules by domain analysis, looking for transmembrane regions or other domains characteristic of leukocyte surface molecules, and by BLAST searches for related genes with known function. Surprisingly, only 2 of the 97 (2%) UniGene clusters analyzed showed some potential for encoding novel surface molecules. The authors therefore concluded that “the cell-type-specific composition of the resting CD8<sup>+</sup> T-cell surface is now largely defined.”

How complete and cell-type-specific is this list of 387 genes? The CTL SAGE library of 20,204 transcripts represents the transcriptome of the CTL clone. A detailed discussion of the limitations of the SAGE method is beyond the scope of this article; suffice it to say that it is possible that functionally important genes may be missing from the library due to low mRNA expression levels, chance or because they lack the target sequence for the tagging enzyme used in the SAGE protocol [7]. In order to get to the final set of 378 CTL-specific genes Evans *et al.* [2] eliminated all transcripts that were present at comparable levels in unrelated libraries.

This powerful approach seems to have validated itself by the fact that TCR components and other principal T-cell markers were present in the shortlist of 387 genes. The method used is not unbiased, however. The choice of libraries (in this case cerebellum, ovary epithelium, and a panel of tumor cell lines that were not specified further), data quality, and the algorithms used for comparison can be expected to have an enormous impact on the results obtained. It is easy to see how functionally important genes might get lost because they are expressed at sufficiently high levels in one of the cell or tissue types used for comparison. A list of cell-type-specific genes derived by this method of successive *in silico* subtraction defines a cell-type-specific gene-expression pattern against the transcriptional background of the cell or tissue types used for comparison. It is not a list of all genes relevant for cell-type-specific function.

The major finding of the Evans *et al.* study [2] is that among the 387 CTL-specific transcripts 27% of the known genes encoded cell-surface molecules, whereas only 2% of the unknown genes showed some potential in that regard. The implication is that the catalog of CTL surface molecules is close to being complete. While it is not unreasonable to assume that the concerted efforts over the last two decades to characterize surface molecules on leukocytes have led to a situation where most CTL-specific surface molecules are known [8], some questions remain. Is this finding unique for CTLs (or for leukocytes in general)? What would be the result of a similar analysis in, for instance, ovary epithelium? Were there unknown cell-surface molecules in the CTL SAGE library? If so, at what point of the stepwise subtraction process did these transcripts get eliminated? It has been noted that leukocytes share many surface molecules with neuronal cells and epithelial cells [8], the very cells used for subtraction by Evans *et al.* [2]. An alternative experimental approach to analyzing the incidence of unknown cell-surface molecules might be to generate SAGE libraries from microsomal and free-ribosomal mRNA pools generated through equilibrium density centrifugation. This approach has been demonstrated to discriminate secretory and cell-surface molecules from nonsecretory proteins quite efficiently [9,10]. One would expect a significantly lower percentage of unknown transcripts in the secretory/surface molecule fraction.

Only the CTL SAGE library was actually generated by Evans *et al.* [2]. The other libraries used for comparison were derived from publicly available databases. Open access to primary gene-expression data is essential, not only for enabling researchers to reproduce published analyses, but also to allow for novel experimental approaches that incorporate relevant data generated by others. Important information can be gained by comparing genome-wide expression data across large numbers of samples. In a recent, extreme example, 3,283 DNA microarrays from human, *Drosophila*, *Caenorhabditis elegans* and yeast were

used to define evolutionarily conserved genetic modules of co-expressed genes [11]. SAGE data have been publicly available on SAGEMap [12,13] for a number of years. Microarray data are far more complex, but a standard for the annotation of microarray data (Minimal Information About a Microarray Experiment; MIAME) [14] and a platform-independent data exchange format (Microarray Gene Expression Markup Language; MAGE-ML) [15] have been developed. Furthermore, public repositories for microarray data such as ArrayExpress [16,17] and GeoBus [18,19] are now available.

SAGE has the advantage over current microarray technology of measuring absolute transcript abundance. Nevertheless, there are some limitations as to what can be said about the T-cell surface by studying mRNA levels. First, for a number of surface molecules, such as CD45, a variety of functionally important splice variants have been described [20] that cannot be distinguished by the 3' SAGE tag. Second, mRNA levels correlate poorly with protein abundance [21]. Third, posttranslational protein modifications can be functionally relevant; for example, glycosylation of CD8 has been demonstrated to affect thymocyte selection by influencing activation thresholds [22]. Fourth, T-cell activation involves re-localization of surface molecules leading to the formation of the immunological synapse, a supramolecular cluster at the contact zone between antigen-presenting cell and the T cell [23]. These early events precede changes in gene expression.

Finally, it seems important to note that the T-cell surface is an abstraction. T cells comprise quite different subsets of cells at variable activation states. As pointed out by Evans *et al.* [2], the finding that most of the molecules on the T-cell surface appear to be known applies strictly only to a resting CD8<sup>+</sup> T-cell clone *in vitro*. 'The T-cell surface - how well do we know it?' is an important question on our way into the post-genomic era of immunology. But even with complete lists of the genes expressed in certain T-cell subpopulations, much more needs to be learned about the regulation and complex interactions of the proteins they encode. We are just scratching the (T cell) surface.

## References

1. Staudt LM, Brown PO: **Genomic views of the immune system.** *Annu Rev Immunol* 2000, **18**:829-859.
2. Evans EJ, Hene L, Sparks LM, Dong T, Retiere C, Fennelly JA, Manso-Sancho R, Powell J, Braud VM, Rowland-Jones SL, *et al.*: **The T cell surface - how well do we know it?** *Immunity* 2003, **19**:213-223.
3. Bernard A, Boumsell L: **The clusters of differentiation (CD) defined by the first international workshop on human leukocyte differentiation antigens.** *Hum Immunol* 1984, **11**:1-10.
4. **Protein reviews on the web** [<http://www.ncbi.nlm.nih.gov/prow/>]
5. **8<sup>th</sup> International workshop on human leukocyte differentiation antigens 2000-2004** [<http://www.hlda8.org/PotentialCDs.htm>]
6. **UniGene** [<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene>]
7. Yamamoto M, Wakatsuki T, Hada A, Ryo A: **Use of serial analysis of gene expression (SAGE) technology.** *J Immunol Methods* 2001, **250**:45-66.
8. Mason D, Andre P, Bensussan A, Buckley C, Civin C, Clark E, de Haas M, Goyert S, Hadam M, Hart D, *et al.*: **Reference: CD anti-gens 2002.** *J Immunol* 2002, **168**:2083-2086.
9. Diehn M, Eisen MB, Botstein D, Brown PO: **Large-scale identification of secreted and membrane-associated gene products using DNA microarrays.** *Nat Genet* 2000, **25**:58-62.
10. Toyoda N, Nagai S, Terashima Y, Motomura K, Haino M, Hashimoto S, Takizawa H, Matsushima K: **Analysis of mRNA with microsomal fractionation using a SAGE-based DNA microarray system facilitates identification of the genes encoding secretory proteins.** *Genome Res* 2003, **13**:1728-1736.
11. Stuart JM, Segal E, Koller D, Kim SK: **A gene-coexpression network for global discovery of conserved genetic modules.** *Science* 2003, **302**:249-255.
12. Lash AE, Tolstoshev CM, Wagner L, Schuler GD, Strausberg RL, Riggins GJ, Altschul SF: **SAGEmap: a public gene expression resource.** *Genome Res* 2000, **10**:1051-1060.
13. **SAGEMap** [<http://www.ncbi.nlm.nih.gov/SAGE>]
14. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, *et al.*: **Minimum information about a microarray experiment (MIAME) - toward standards for microarray data.** *Nat Genet* 2001, **29**:365-371.
15. Spellman PT, Miller M, Stewart J, Troup C, Sarkans U, Chervitz S, Bernhart D, Sherlock G, Ball C, Lepage M, *et al.*: **Design and implementation of microarray gene expression markup language (MAGE-ML).** *Genome Biol* 2002, **3**:research0046.1-0046.9.
16. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, *et al.*: **ArrayExpress - a public repository for microarray gene expression data at the EBI.** *Nucleic Acids Res* 2003, **31**:68-71.
17. **EBI databases - ArrayExpress** [<http://www.ebi.ac.uk/arrayexpress/>]
18. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.** *Nucleic Acids Res* 2002, **30**:207-210.
19. **Gene expression omnibus** [<http://www.ncbi.nlm.nih.gov/geo/>]
20. Penninger JM, Irie-Sasaki J, Sasaki T, Oliveira-dos-Santos AJ: **CD45: new jobs for an old acquaintance.** *Nat Immunol* 2001, **2**:389-396.
21. Greenbaum D, Colangelo C, Williams K, Gerstein M: **Comparing protein abundance and mRNA expression levels on a genomic scale.** *Genome Biol* 2003, **4**:117.
22. Daniels MA, Hogquist KA, Jameson SC: **Sweet 'n' sour: the impact of differential glycosylation on T cell responses.** *Nat Immunol* 2002, **3**:903-910.
23. Krummel MF, Davis MM: **Dynamics of the immunological synapse: finding, establishing and solidifying a connection.** *Curr Opin Immunol* 2002, **14**:66-74.