

Gene expression microarrays: glimpses of the immunological genome

Gordon Hyatt, Rachel Melamed, Richard Park, Reuben Seguritan, Catherine Laplace, Laurent Poirot, Silvia Zucchelli, Reinhard Obst, Michael Matos, Emily Venanzi, Ananda Goldrath, Linh Nguyen, John Luckey, Tetsuya Yamagata, Ann Herman, Jonathan Jacobs, Diane Mathis & Christophe Benoist

Successful microarray experimentation can generate enormous amounts of data, potentially very rich but also very unwieldy. Bold outlooks and new methods for data analysis and presentation should yield additional insight into the complexities of the immune system.

Microarray analyses of gene expression, because of the unprecedented breadth of the data they yield, hold unique promise for elucidation of the functional organization of a collection of cell types, such as those that compose the immune system. Although 'blind' to some aspects of cellular regulation,

such as translational control or intracellular compartmentalization, the evaluation of steady-state amounts of coding mRNA provides a direct representation of transcriptional and post-transcriptional regulation^{1,2}. At present, microarrays have the capacity to represent all transcripts and splice variants, demonstrating the genome's global activity. Thus, what constitutes the 'immunological genome' can be defined — the inventory of genes expressed in different immune system cells and the ways in which those transcripts are connected in regulatory networks and vary during differentiation and immune responses. To a large extent, immunologists have only begun to scratch the surface of what may be gleaned from microarray technology, in particular from meta-analyses that encompass the immune system as a whole. This is true for conventional protein-encoding mRNA, and even more so for microRNA molecules, whose expression patterns and functional consequences are only beginning to be explored³. Here we will review the applications of microarray analyses of gene expression as tools for exploring the structure and function of the immune system, with an emphasis on meta-analyses. We will not discuss applications for leukemia and lymphoma exploration and classification, which represents a field of its own⁴. We will emphasize general conclusions about the immunological genome and 'flag' new paths that merit further exploration. To demonstrate the principles discussed here, we introduce the ImmGen website, a new interactive tool

that displays various aspects of gene expression in the immune system.

'A versus B'

In the field of immunology, as elsewhere, microarrays were initially applied to address focused and well defined issues to delineate the set of transcripts that distinguish 'condition A' from 'condition B', in which transcripts distinguish two closely related cell types, developmental stages^{5–7} or functional states (such as memory, anergy or regulatory)^{6,8–10}. Microarrays have also been used to compare transcriptional programs elicited by distinct stimuli, such as Toll-like receptor ligands or cytokines^{11–13}, or to explore the regulatory 'imprint' of a particular transcription factor^{14,15}. In case-control formats, the perturbations associated with specific immunological diseases have been explored by expression profiling^{12,16–18}, with attempts to decipher key cellular or molecular pathways from disease-specific 'signatures'.

Most of those studies succeeded in demonstrating some of the molecular framework underlying the phenomena studied. There have also been 'flops'; in several cases (including our lab!), the data did not yield the anticipated insight and resulted in little more than the dreaded 'gene list' from which no knowledge immediately emerged. Such a result may arise from a true absence of informative variation, from faulty experimental design or from timid data analysis. Too often, this analysis is limited to a ranked gene list based on wholly arbitrary cutoffs, from which one or a few 'attractive' genes at the 'tail ends' of the distribution are

Gordon Hyatt, Rachel Melamed, Richard Park, Reuben Seguritan, Catherine Laplace, Laurent Poirot, Silvia Zucchelli, Reinhard Obst, Michael Matos, Emily Venanzi, Ananda Goldrath, Linh Nguyen, John Luckey, Tetsuya Yamagata, Ann Herman, Jonathan Jacobs, Diane Mathis and Christophe Benoist are with the Section on Immunology and Immunogenetics, Joslin Diabetes Center, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02215, USA. Laurent Poirot and Ann Herman are now with the Genomics Institute of the Novartis Research Foundation, San Diego, California 92121, USA. Silvia Zucchelli is now with the SISSA-LNM, Area Science Park, Basovizza, 34012 Trieste, Italy. Michael Matos is now with Wolfboro Pediatrics, Wolfboro, New Hampshire 03894, USA. Ananda Goldrath is now in the Division of Biology, University of San Diego, La Jolla, California 92093, USA. Linh Nguyen is now with The Campbell Family Institute for Breast Cancer Research, Ontario Cancer Institute and Princess Margaret Hospital, Toronto, Canada. Tetsuya Yamagata is now in the Department of Hematology, Dokkyo University School of Medicine, Tochigi-Ken, Japan.
 e-mail: cbdm@joslin.harvard.edu

culled for additional study. Confining microarray data to the generation of hypotheses has become the norm but can generate oversimplified interpretations and leaves behind much complex but valuable information about the system being investigated. In some contexts, gene expression profiles should be considered valid endpoints in themselves; for example, when the nature of the transcriptional signature demonstrates a facet of biological reality or provides a clear answer to an experimental question.

Expression compendia: the landscape

Microarrays have the ability to define the true range of genes active in cells of the immune system, an essential 'ground map' of the genes that must be considered in understanding immune function. That definition is not entirely trivial, however. The total number of genes in mammalian genomes is still in question, although the number of remaining surprises may be limited¹⁹. Moreover, it is difficult to appreciate the exact contribution of signals from cross-hybridizing transcripts in a profile (the 'dirty secret' of the microarray world). Finally, the range of defined populations and subpopulations in the immune system quickly enters into the hundreds.

With those caveats, a few studies have attempted to address the range of gene expression in a substantial spectrum of immune cells^{20–22}. The Novartis SymAtlas project analyzed over 60 organs in both humans and mice, some of which corresponded to lymphoid tissue or sorted cell populations²³. The Genentech IRIS project²⁴ and Joslin ImmGen project (<http://www.immgen.org>) grouped data from more than 20 purified immune cell populations from humans and mice, respectively. Overall, the picture that has emerged from those compendia is that much of the genome is active in one or many types of immune system cells. Even with conservative thresholds, 67% of genes were assigned scores of having substantial expression in at least one immune cell by both the ImmGen and SymAtlas data groups. Those fractions may actually be an underestimate, as microarrays are fairly insensitive in detecting very low expression values. Thus, it seems that the immune system makes use of a large fraction of the genome's potential. The corollary of this breadth is that relatively few genes are truly specific to the immune system. Only 121 of 16,969 unique genes in the SymAtlas data group were expressed exclusively in immune cells. Notably, those fractions are very similar to those obtained for organs of the nervous system (73.4% of genes expressed; only 136 exclusively neural). Transcripts whose immune specificity was only quantitative were

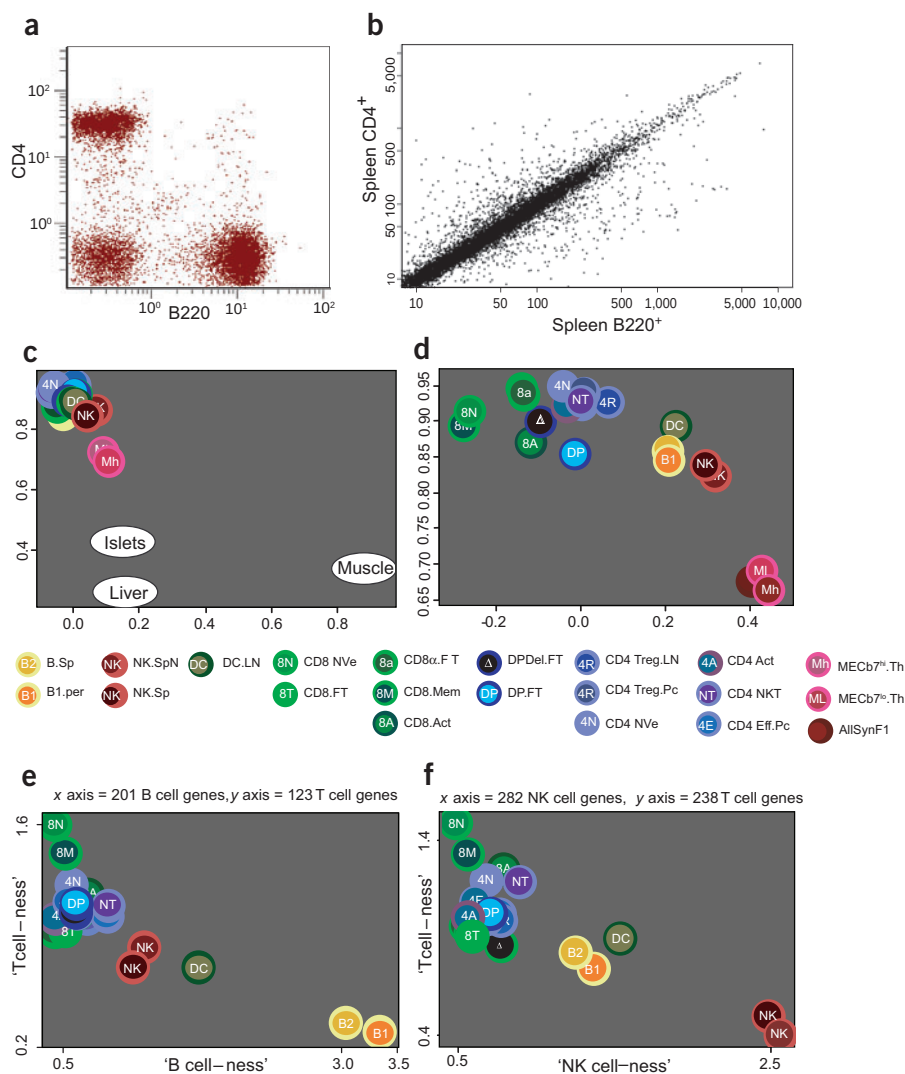


Figure 1 Differentiating cells by microarray analysis. (a,b) B lymphocytes and T lymphocytes, distinguished by flow cytometry (a) or gene expression profile (b). (c,d) Relative positioning of immune cells based on principal component analysis of their global expression profiles; the x and y coordinates of a given population on the plot derive from its coefficients for the first two principal components. In c, the ImmGen data group of immune populations (key) is complemented with data sets from muscle, liver or pancreatic islets (omitted in d for better visualization of the relationships between lymphoid populations). (e,f) Reference population plots. The genes that distinguish two reference populations are used to calculate 'likeness' indices. In e, 'B cell genes' were selected as those overexpressed in B cells (averaging profiles from splenic B cells and peritoneal B1B cells) relative to T cells (averaging profiles from CD4 and CD8 T cells). Expression values of 'B cell genes' and 'T cell genes' in experimental populations were normalized to the expression values of those genes in the reference populations and were averaged to yield the x and y coordinates of each experimental population. Populations abbreviated as in c. Details, **Supplementary Note** online. B.Sp, B lymphocytes, all, spleen; B1.per, B lymphocytes, peritoneum; NK.SpN, natural killer cells, spleen, nonobese diabetic; NK.Sp, natural killer cells, spleen; DC.LN, dendritic cells, lymph node; CD8 Nve, CD8⁺ naive cells, lymph node, OT-1-transgenic; CD8.FT, CD8 $\alpha\beta$ single-positive cells, fetal thymic organ culture; CD8 α .FT, CD8 α single-positive cells, fetal thymic organ culture; CD8.Mem, CD8⁺ memory cells, lymph node, OT-1-transgenic; CD8.Act, CD8⁺ activated cells, lymph node, OT-1-transgenic; DPDeL.FT, CD4⁺CD8⁺ double-positive cells, preapoptotic, fetal thymic organ culture, BDC2.5 TCR-transgenic; DP.FT, CD4⁺CD8⁺ double-positive cells, fetal thymic organ culture; CD4 Treg.LN, CD4⁺CD25⁺ T_{reg} cells, mesenteric lymph node, BDC2.5 TCR-transgenic, nonobese diabetic; CD4 Treg.Pc, CD4⁺CD25⁺ T_{reg} cells, pancreatic infiltrate, BDC2.5 TCR-transgenic, nonobese diabetic; CD4 Nve, CD4⁺ naive cells, lymph node; CD4 Act, CD4⁺ cells, partially activated, peripheral lymph node, BDC2.5 TCR-transgenic; CD4 NKT, CD4⁺ NKT cells, spleen; CD4 Eff.Pc, CD4⁺ cells, pancreatic infiltrate, BDC2.5 TCR-transgenic, nonobese diabetic; MEc7^{hi}.Th, medullary epithelial cells, B7^{hi}, thymus; MEc7^{lo}.Th, medullary epithelial cells, B7^{lo}, thymus; AII.Syn.FI, arthritic synovial fluid.

Thus, the microarray compendia portray expression profiles that are very broad, with little absolute specificity but much quantitative tuning. Such a view contrasts with the disproportionate importance ascribed to cell type-specific genes in mental representations of immune cells, which stem from flow cytometry images obtained with chosen cell type-specific markers. The distinction between B lymphocytes and T lymphocytes seems very different with profiling by flow cytometry versus microarray (**Fig. 1a,b**), but the latter is probably a better reflection of reality. Few of the CD markers are truly cell type specific, and it can be very difficult to identify markers unique to a given cellular phenotype (such as with the long quest for a marker truly specific for CD4⁺CD25⁺ regulatory T (T_{reg}) cells before the identification of Foxp3). This 'wide breadth and little specificity' distribution of transcripts is also compatible with classic RNA-DNA reassociation and subtractive hybridization experiments. It has been esti-

Distinguishing immune cell types

If truly cell type-specific transcripts constitute a relatively minor part of a cell's overall transcriptional program, then the definition of a differentiated cell's identity must also take into account the quantitative variations that affect shared transcripts. Indeed, analyses have shown that essentially all expressed genes have subtle variations between CD4⁺ and CD8⁺ T cells, reflecting regulatory fine-tuning ignored before because it was within the range of 'experimental noise'. Therefore, it is important to analyze distinctions between immune cell types with a genome-wide perspective, rather than focusing on a few differentially expressed transcripts. Mathematical dimensionality-reduction tools can be applied to such analyses and can generate plots in which each cell type is positioned to best reflect its integrated genome-wide profile relative to those of all other cells (**Fig. 1c,d**). When applied to the ImmGen data group and data sets from unrelated organs, the results of all hematopoietic cells clustered tightly together, away from those of muscle, liver and pancreatic islets (**Fig. 1c**; no such patterns were obtained with control randomized data groups). The results from thymic medullary epithelial cells, which might be expected to be more like other epithelia, mapped closer to

In the hematopoietic system, the same algorithm generates groupings that fit well with what might be expected. All T cell types map to the same general area of the plot and all B cell subsets are also tightly grouped. Although immature DP thymocytes are distinct, there is little distance on the plot between subsets of mature CD4⁺ T cells (which does not negate the distinct differences between CD4⁺ T cell subsets demonstrated in more focused analyses of T_{reg} cells or natural killer T (NKT) cells). That disposition is consistent with the idea that NKT cells are a close variant of conventional CD4⁺ T lymphocytes²⁷, rather than an intermediate between T cells and natural killer cells, as their acronym would suggest. Overall, dendritic cells, natural killer cells and B cells are grouped in a distinct area on the plot (**Fig 1d**). That layout is somewhat unexpected, as it might be expected that results from B lymphocytes and T lymphocytes would cluster, or that natural killer cells would be most closely related to T cells, with which they share common precursors in the thymus, cytolytic activity and the expression of several activation markers²⁸. In reference population analysis plots, in which each population is positioned in the two-dimensional

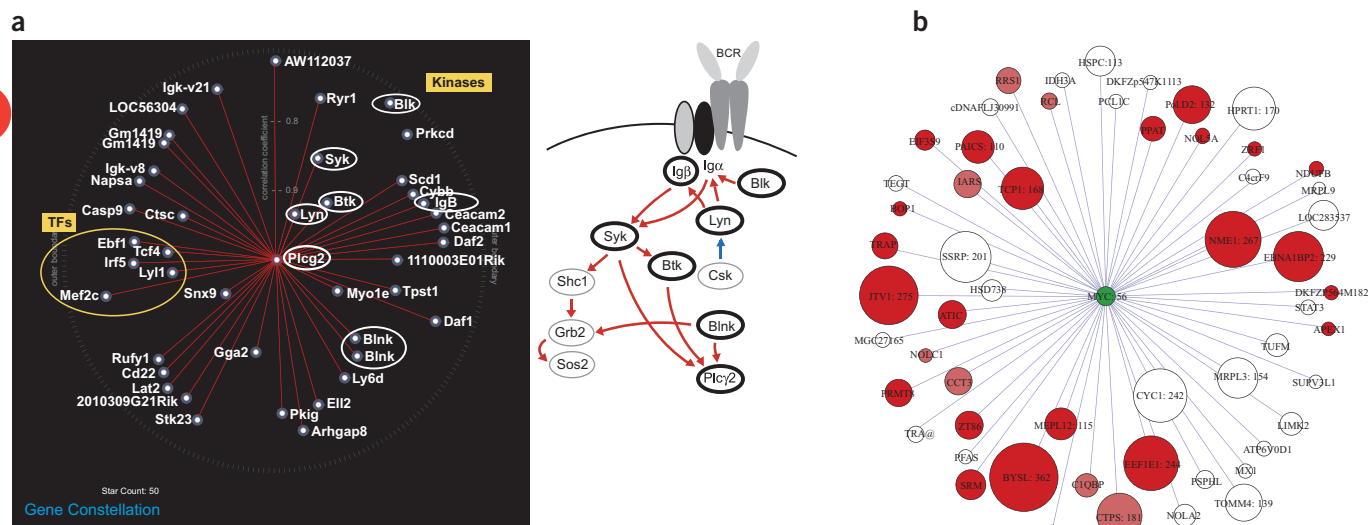


Figure 2 Gene expression networks in the immune system. **(a)** Network of correlated expression in immune cells. Data represent genes whose expression profiles are most correlated with that of *Plcg2* (center), in the 76 data sets of the ImmGen data, which includes all of the main cell types of the mouse immune system. The distance from the center is proportional to the tightness of the correlation, and the correlated genes are positioned according to shared GeneOntology identifiers, such that genes sharing identifiers are grouped in the same angular quadrants. Right, signal transduction pathways moving from the B cell receptor (BCR) toward *Plcg2* (interactive display, <http://www.immgen.org>). **(b)** The *MYC* subnetwork identified during analysis of 336 data sets from human normal or leukemic B cells⁴⁵. The ARACNe algorithm was used to identify the genes with regulatory connections to *MYC*, one of the most connected genes in the reconstructed B cell network. The size of each circle is proportional to the number of interactions in which each of the neighbor genes is itself involved.

plane relative to its expression of a set of 'reference genes' that distinguish two reference populations²⁹, B1 and B2 cells map closer to natural killer cells and dendritic cells than to the cluster of T cells (**Fig. 1e,f**) based on genes differentially expressed in B lymphocytes and T lymphocytes. The genomic similarity between B cells, natural killer cells and dendritic cells is also consistent with a SAGE (serial analysis of gene expression) tag analysis of human peripheral blood lymphocytes²¹ and with the existence of 'natural killer–dendritic' cells³⁰. Thus, relationships between cell lineages based on these global expression patterns depart to some extent from classical views in which differentiation potential, and hence gene expression capabilities, become progressively restricted at each lineage fork³¹. The picture obtained from these genome-wide analyses is more compatible with views arguing for considerable plasticity in lymphoid and myeloid cell differentiation³², with substantial overlap between the expression profiles of different lineages.

Modules, signatures and networks

The pattern of gene expression that defines a unique state of cellular differentiation and activation reflects an intricate network of regulatory interactions. Although a full grasp of this regulatory network remains remote, genome-scale expression data groups have the potential to unmask some aspects of the connectivity in the network. These meta-analyses have created new virtual 'objects' (borrowing from the semantics of computer science), such as modules, signatures or networks. These objects allow a reduction and organization of the bewildering complexity of the genome and its expression. They correspond to molecular and cellular realities but are not necessarily intuitive from immunologists' usual perspective and require particular tools for analysis, representation and publication.

Modules

The smallest of genomic's new objects, a 'module' is defined as a group of genes that form a regulatory unit, sharing regulatory controls that influence their expression^{33,34}. A module typically includes genes of different functional categories. The definition of modules is intrinsic to the expression data, as generated by unsupervised computational algorithms without *a priori* biological definition. The computational definition of such modules from gene expression compendia can also incorporate input from different data types (such as sequence motifs or protein-protein interactions). Biological relevance is then inferred from the composition of the module (insight is generated from the realization that a module 'makes sense').

Signatures

Conceptually different from modules, 'signatures' are sets of genes generated by comparison of transcriptional profiles in conditions that are defined from an *a priori* perceived biological logic (a differentiated state or a response to a drug)³⁵. The 'T_{reg} cell signature' is the set of transcripts that distinguishes Foxp3⁺ T_{reg} cells from conventional T cells, and the 'interferon signature' is the group of transcripts modified by interferon exposure^{10,12}. Signatures are fluid and complex objects. Their boundaries and composition depend on nonpermanent elements that guide their definition (knowledge in the field, investigator bias, composition of the data group and formulation of the query). The 'plasma cell signature' will vary with the type(s) of precursor B cells used as a reference. Moreover, signatures intersect and overlap considerably. For example, the T_{reg} cell signature defined by comparison of naive CD4⁺ T cells and T_{reg} cells^{10,36} contains genes that also belong to 'T cell activation', 'transforming growth factor-β' and 'plasma cell' signatures. Furthermore, signatures are quantitative, rather than simply binary. This includes the direction of the transcriptional perturbation (for example, the T_{reg} cell signature includes roughly equal numbers of genes that are under- and overexpressed relative to those of conventional CD4⁺ cells, but the former always receive much less attention) and the extent of differential expression. Indeed, akin to Sherlock Holmes' silent hound, transcripts whose abundance does not vary can be an important element of a signature (the absence of cell cycle genes is a key component of the signature of T cell anergy⁹). Finally, a signature can vary with the state or anatomical location of a population. For example, several transcripts of the T_{reg} cell signature vary among T_{reg} cell subsets or after infiltration into autoimmune lesions^{10,37,38}.

Signatures are the most intuitively useful objects for the immunologist. Yet because of the complexity and fluidity of signatures, algorithms to analyze their composition, variation and overlap are only now emerging. Similarly, standardized formats for publishing (beyond the typical 'supplementary gene list') and sharing signatures are still limited. However, some databases of expression signatures that should be of immediate use to the community are emerging³⁹ (http://www.immgen.org/index_content.html, Population Signature).

Networks

'Networks' represent the highest level of integration, in which the ultimate aim is to link all genes and transcripts in the immunological genome. Several distinct types of networks

have been defined. In coexpression networks⁴⁰, each gene is linked to the genes with which its expression is most closely correlated. In such a network, the modules described above are basically local density maxima of the network space. Sets of genes with correlated expression may denote a functional pathway or may indicate regulation by the same transcriptional or mRNA (de)stabilization elements. Gene-gene correlations that are conserved across species emphasize key functional pathways and genes^{40,41}. Defining and demonstrating coregulation in the immune system should yield functional insight. Representing a genome-wide network presents serious challenges, however, because of the scale and complexity of genome-wide networks. The interactive interface of the ImmGen site presents the genes most closely connected to a query gene. Each of these 'connected' genes can in turn be 'interrogated' dynamically. Positioning of genes in a two-dimensional plane presents additional information, including the degree of correlation, shared GeneOntology identifiers and chromosomal location. For example, correlations with the B cell signaling molecule Plcg2 emphasize protein kinases or adaptors, signaling adaptors and transcription factors, which bear considerable similarity to the diagram of signaling pathways that converge on Plcg2 after B cell receptor triggering (many genes in **Fig. 2**, right, are found in the set of genes correlated with Plcg2).

Coexpression networks remain limited in that they cannot show true regulatory interactions. In regulatory networks⁴², each gene is linked to the genes it controls or is controlled by. Underlying the construction of genome-wide networks are the ideas that all regulatory controls are ultimately interconnected and that a given perturbation has local and predictable effects as well as consequences that are more distant and less intuitively obvious. These genomic networks share some structural and conceptual aspects with the 'idiotypic networks' of immunology's past⁴³. Specific computational tools, derived from applied mathematics (pattern recognition, system dynamics and Bayesian networks) have been developed to 'reverse-engineer' network structure and to predict 'regulator-regulated' relationships^{44–47}. These reverse-engineering approaches have proven powerful in microbial systems but face daunting challenges when faced with the larger mammalian genomes. All share the requirement for very large data groups (hundreds or thousands of data sets) in which discrete variation is the substrate used by the algorithms to predict regulatory connections between genes. Variation can be introduced through gene knock-out, RNA interference 'knockdown', or drug

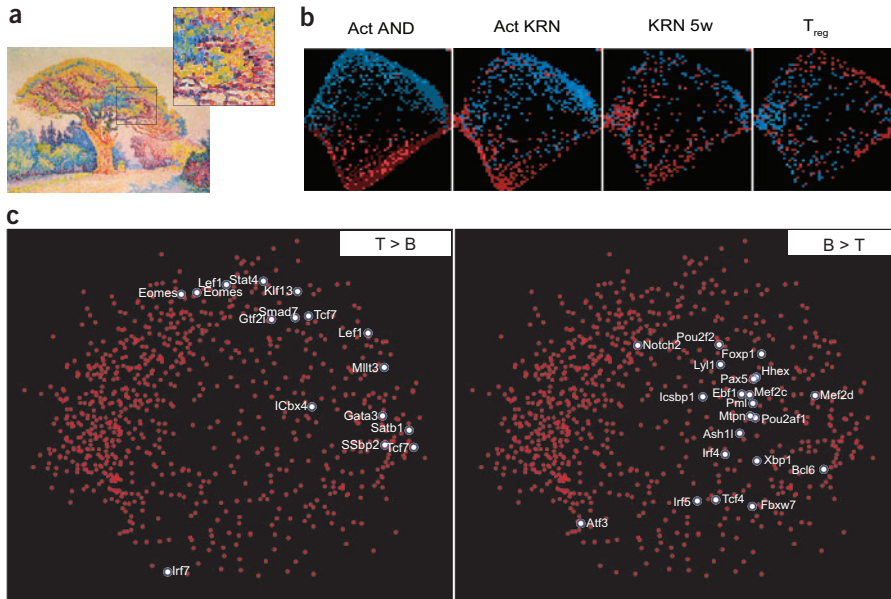


Figure 3 Complex data representations in art and immunogenomics. (a) Pine Tree by Paul Signac (1909; Pushkin Museum, Moscow). Inset, enlargement of boxed area at left. (b) Abstract representation of expression signatures in subsets of CD4⁺ T cells. Expression profiles of CD4⁺ T cell populations are compared with those of naive CD4⁺ cells, and the 'fold change' for each differentially expressed gene is presented as a dot in a classic heat map (underexpression, blue; overexpression, red). The genes are positioned similarly in each two-dimensional plot, which is optimized by multidimensional scaling such that the relative positions best reflect the correlation between the 'fold changes' for each condition¹⁰ (R.O. and L.N., unpublished data). Act, T cell receptor–transgenic T cells (AND or KRN) acutely activated by cognate antigen *in vivo*; KRN 5w, chronically activated KRN T cells. Acute activation of different T cell receptors elicits different activation responses, which are amplified in the chronically activated KRN T cells. Most genes overexpressed in T_{reg} cells (red; far right) are repressed in activated KRN and AND cells, but a portion of the activation signature is upregulated in T_{reg} cells. (c) Multidimensional scaling is used to calculate optimal positions for transcription factor genes, as a 'best projection' of the correlations of their expression across the ImmGen data group. This display names only those transcription factors that characterize a cell population (here, those overexpressed in B cells or T cells; interactive display, <http://www.immgen.org>).

useful in displaying gene expression data. For instance, Impressionist painters of the pointillist school used the juxtaposition of multiple points of paint to generate complex images. Signac's painting incorporates a many such dots to create an image that 'speaks' to the viewer on several levels (Fig. 3a). It incorporates fine details, integrates them in complex structures (such as a tree) and conveys the artist's overall impression and emotion. Similar techniques can be used to compare the signatures of different states of CD4⁺ T cells, such as transcriptional programs elicited by engagement of different T cell receptors or chronic activation, or the overlap between the T_{reg} cell and conventional T cell activation signatures (Fig. 3b). Each point 'color codes' the expression of a gene, and their arrangement aims to bring out the swath of subsignatures that distinguish or connect the different states. In the ImmGen 'GeneFamily' representation (Fig. 3c), the constellation of transcription factor genes is arrayed according to overall expression, showing members of particular subfamilies or genes that distinguish lineages.

Impressionists did not have web access. Like modern art, representations of complex signature and network structures can benefit from incorporation of the dazzling capabilities of electronic display, of interactive environments and of virtual reality. Also, the meta-analyses described here only hint at the more ambitious opportunities afforded by 'systems immunology'. Gene expression networks will be connected to other types of 'meta-data', such as representations of the proteome and its modifications, genome-wide analyses of chromatin structure and compartmentalization of transcripts and proteins in cells. In addition, genomic and proteomic data are mostly static, mainly representing steady-state conditions in a cell. Systems-level understanding will require the integration of its dynamic and quantitative aspects, taking into account the complex thresholds and feedback loops that give the immune system its amazing discriminatory power⁵¹.

Note: Supplementary information is available on the Nature Immunology website.

ACKNOWLEDGMENTS

We thank D. Laidlaw, J.J. Collins and D. Koller for discussions; E. Hyatt, Q.M. Pham, G. Losyev, R. Saccone and J. Johnson for assistance with mice, cell sorting and microarray processing; and the S+ Listserv members for advice. Supported by the National Institutes of Health (AR046580, AI51530, AI52343, DK60027, DK59658 to C.B. and D.M.) and the William T. Young Chair.

treatment, or by exploitation of natural genetic polymorphism through joint analysis of gene expression and genetic variation. Anchoring the reverse engineering of the expression network on underlying polymorphism makes for more robust analyses and demonstrates genetic association^{47–49}. For immunology, the most demonstrative study so far deciphered regulatory connectivity among normal and transformed human B cells using an algorithm based on 'mutual information' techniques⁴⁵. This showed that the regulatory network is not uniform but that a limited number of genes serve as 'hubs' of regulation, and account for most of the connectivity (a typical 'scale-free network' in the terminology of network theory, analogous to networks of social interactions or airline traffic). As shown for MYC (Fig. 2b), many of a hub's first neighbors are themselves large hubs, which suggests that a single gene can influence many cellular processes. That idea makes teleological sense, as the cell cycle triggering exerted by Myc requires profound changes in cell metabolism and biosynthesis.

Representation: impressionist microarrays

Immunologists in the 'molecular biology decades' have become accustomed to working in relative intellectual comfort, testing one (or a few) isolated variable(s) against the backdrop of a more nebulous whole system. Experiments can be 'read out' visually by comparing the intensity of a few bands on a gel or of a few cell populations on a cytometry profile. With the thousands of simultaneous variables provided by microarray and other systems-level analyses, this situation is radically changing. Because the human mind cannot grasp much more than a single page of data at one time, a new method of graphic representation is needed to present in a human-accessible form the characteristics of genes, signatures and networks as they evolve in different cell or phenotypic states. It is from such projections that true knowledge can be gained.

The field of scientific visualization has used art analogies to display multiple facets of complex data⁵⁰, some of which should be

1. De Risi, J.L., Iyer, V.R. & Brown, P.O. *Science* **278**, 680–686 (1997).
2. Staudt, L.M. & Brown, P.O. *Annu. Rev. Immunol.* **18**, 829–859 (2000).
3. Plasterk, R.H. *Cell* **124**, 877–881 (2006).
4. Staudt, L.M. & Dave, S. *Adv. Immunol.* **87**, 163–208 (2005).
5. Hoffmann, R., Bruno, L., Seidl, T., Rolink, A. & Melchers, F. *J. Immunol.* **170**, 1339–1353 (2003).
6. Mick, V.E., Starr, T.K., McCaughy, T.M., McNeil, L.K. & Hogquist, K.A. *J. Immunol.* **173**, 5434–5444 (2004).
7. Edwards, A.D. *et al. J. Immunol.* **171**, 47–60 (2003).
8. Kaech, S.M., Hemby, S., Kersh, E. & Ahmed, R. *Cell* **111**, 837–851 (2002).
9. Macian, F. *et al. Cell* **109**, 719–731 (2002).
10. Fontenot, J.D. *et al. Immunity* **22**, 329–341 (2005).
11. Huang, Q. *et al. Science* **294**, 870–875 (2001).
12. Bennett, L. *et al. J. Exp. Med.* **197**, 711–723 (2003).
13. Zeng, R. *et al. J. Exp. Med.* **201**, 139–148 (2005).
14. Wong, A.W. *et al. Nat. Immunol.* **4**, 891–898 (2003).
15. Anderson, M.S. *et al. Science* **298**, 1395–1401 (2002).
16. Adarichev, V.A. *et al. Arthritis Res. Ther.* **7**, R196–R207 (2005).
17. Matos, M., Park, R., Mathis, D. & Benoist, C. *Diabetes* **53**, 2310–2321 (2004).
18. Poirot, L., Benoist, C. & Mathis, D. *Proc. Natl. Acad. Sci. USA* **101**, 8102–8107 (2004).
19. Carninci, P. *et al. Science* **309**, 1559–1563 (2005).
20. Kluger, Y. *et al. Proc. Natl. Acad. Sci. USA* **101**, 6508–6513 (2004).
21. Hashimoto, S. *et al. Blood* **101**, 3509–3513 (2003).
22. Hutton, J.J. *et al. BMC Genomics* **5**, 82 (2004).
23. Su, A.I. *et al. Proc. Natl. Acad. Sci. USA* **101**, 6062–6067 (2004).
24. Abbas, A.R. *et al. Genes Immun.* **6**, 319–331 (2005).
25. Crompton, J., Humphries, S., Woods, D. & Williamson, R. *Nucleic Acids Res.* **8**, 6007–6017 (1980).
26. Davis, M.M., Cohen, D.I., DeFranco, A.L. & Paul, W.E. *UCLA Symp. Mol. Cell. Biol.* **24**, 215–220 (1982).
27. Wilson, S.B. & Byrne, M.C. *Curr. Opin. Immunol.* **13**, 555–561 (2001).
28. McMahon, C.W. & Raulet, D.H. *Curr. Opin. Immunol.* **13**, 465–470 (2001).
29. Yamagata, T., Benoist, C. & Mathis, D. *Immunol. Rev.* **210**, 52–66 (2006).
30. Shortman, K. & Villadangos, J.A. *Nat. Med.* **12**, 167–168 (2006).
31. Akashi, K. *et al. Blood* **101**, 383–389 (2003).
32. Rothenberg, E.V. & Dionne, C.J. *Immunol. Rev.* **187**, 96–115 (2002).
33. Segal, E., Friedman, N., Koller, D. & Regev, A. *Nat. Genet.* **36**, 1090–1098 (2004).
34. Brunet, J.P., Tamayo, P., Golub, T.R. & Mesirov, J.P. *Proc. Natl. Acad. Sci. USA* **101**, 4164–4169 (2004).
35. Shaffer, A.L. *et al. Immunity* **15**, 375–385 (2001).
36. McHugh, R.S. *et al. Immunity* **16**, 311–323 (2002).
37. Huehn, J. *et al. J. Exp. Med.* **199**, 303–313 (2004).
38. Chen Z., Herman A., Matos M., Mathis D. & Benoist C. *J. Exp. Med.* **202**, 1387–1397 (2005).
39. Shaffer, A.L. *et al. Immunol. Rev.* **210**, 67–85 (2006).
40. Stuart, J.M., Segal, E., Koller, D. & Kim, S.K. *Science* **302**, 249–255 (2003).
41. Bergmann, S., Ihmels, J. & Barkai, N. *PLoS Biol.* **2**, 85–93 (2004).
42. Levine, M. & Davidson, E.H. *Proc. Natl. Acad. Sci. USA* **102**, 4936–4942 (2005).
43. Jerne, N.K. *Ann. Immunol. (Paris)* **125C**, 373–389 (1974).
44. Segal, E. *et al. Nat. Genet.* **34**, 166–176 (2003).
45. Basso, K. *et al. Nat. Genet.* **37**, 382–390 (2005).
46. Gardner, T.S., di Bernardo, D., Lorenz, D. & Collins, J.J. *Science* **301**, 102–105 (2003).
47. Battle, A., Segal, E. & Koller, D. *J. Comput. Biol.* **12**, 909–927 (2005).
48. Tegner, J., Yeung, M.K., Hasty, J. & Collins, J.J. *Proc. Natl. Acad. Sci. USA* **100**, 5944–5949 (2003).
49. Bystrykh, L. *et al. Nat. Genet.* **37**, 225–232 (2005).
50. Keefe, D., Karelitz, D.V.E. & Laidlaw, D.H. *IEEE Comput. Graph. Appl.* **25**, 18–23 (2005).
51. Altan-Bonnet, G. & Germain, R.N. *PLoS Biol.* **3**, 1925–1938 (2005).

Supplemental Material to Hyatt et al, “Gene-expression microarrays: glimpses at the immunological genome”

A. Description and Use of the ImmGen Website.

The ImmGen website (<http://ImmGen.org>) can be accessed with standard web browsers on Windows or Mac systems (not tested for Linux-based browsers). It requires the current version of Macromedia Flash for its operation (if not present, the user will be prompted for a free download from the Macromedia website).

On-line help functions are available for each of the site’s pages, as well as a tutorial which guides users through the sites operation.

ImmGen is an interactive browser, which consists of four main pages at present:

- The **Gene Skyline** page displays expression values of individual genes across cell types, with a choice of datagroups
- The **Gene Constellation** page displays correlations between genes
- The **Population Signature** page shows relations between populations and the genes that characterize them
- The **Gene Family** page displays a 2D projection of members of particular gene families according to their relative expression patterns across all populations

Terminology conventions:

“Gene” is meant as one element of the microarray. A true gene in the molecular biology sense may be represented by several “genes” on the array (usually displaying very correlated expression patterns, as expected, but not always; for example the array “genes” may correspond to alternative transcripts, have different efficacies, or have different degrees of off-target crosshybridization.

The **“Gene Skyline”** page presents, as a classic bar graph, the expression profiles of a selected gene(s) in one chosen datagroup. Basic annotation information on the gene and links to external databases are also provided.

The page allows the user to search for the genes to display, based on gene names, symbols, or other common identifiers (when more than one gene is returned, by scrolling between the different genes).

User Settings:

Setting	Description	Possible Values
Datagroup	Sets the datagroup for which the expression values of the gene are displayed	ImmGen (all immunological datasets) ImmGenPlus (as ImmGen, with a few additional datasets from unrelated organs for reference) SymAtlas (from the GNF collection of normal mouse organs)
Scaling	Sets the scale type and maximum for the bar graph	Local: Linear scale, maximum set to the maximum expression value for the gene displayed Global: Linear scales, constant scale for all genes; the scale maximum is chosen to include 98% of genes. Best for a general perspective, but crushes variations in low expression range Log: Logarithmic scaling, constant for all genes
Search in field	Field to be searched	ProbeID, Gene Symbol, Gene Name, NCBI GeneID, Chromosome, Unigene, etc
Search For	String of characters to be searched for.	Any string of numbers or characters. The database engine will search for these characters within the specified field. Note: only 100 genes will be returned, and searches generating many matches will be returned more slowly.
Search Type	Stringency of the search	Contains, Exact Match, Begins With.

Gene Skyline Actions

Choose datagroup (Immgen, SymAtlas)

Gene search settings

Starts the search

Execute

View Constellation

View SymAtlas

Switch to the Constellation view for this gene

Database Links

Scroll up and down between matching genes

Switch between ImmGen pages pages

The screenshot shows the Gene Skyline web application interface. The top section contains two main panels: 'OPTIONS' and 'SEARCH'. The 'OPTIONS' panel has a dropdown for 'data group' set to 'ImmGen' and a 'scaling' dropdown set to 'Local'. The 'SEARCH' panel has a 'search in field' dropdown set to 'Gene Symbol', a 'search for' input field containing 'cd2', a 'search type' dropdown set to 'Contains', and an 'Execute' button. Below the search bar, a horizontal bar chart displays gene expression data across various cell types. The y-axis lists cell types such as B.Sp, B1.per, NK.SpN, NK.Sp, DC.LN, CD8Nve.LN, CD8Act.LN, CD8Mem.LN, CD8.FT, CD8aa.FT, CD4.Sp, CD4Act.LN, CD4NKT.Sp, CD4Treg.LN, CD4Eff.PcN, CD4Treg.PcN, DP.FT, DPDeI.FT, MECb7lo.Th, MEC.Th, and AllSynFi. The x-axis represents expression levels from 0 to 1603. To the right of the chart is a detailed view for the selected gene, CD24a. This view includes fields for 'Probe Set' (100600_at), 'Gene Symbol' (Cd24a), 'Title' (CD24a antigen), 'Aliases' (HSA // Ly-52), 'Chromosome' (10), 'Location' (43,489,862), 'NCBI' (12484), 'Unigene' (Mm.29742), 'Ensembl' (ENSMUSG000000047139), 'GO' (membrane (16020), external side of plasma membrane (9897)), and 'InterPro'. Below the detailed view is a vertical scrollbar. At the bottom of the interface is a navigation bar with four tabs: 'Gene Skyline', 'Gene Constellation', 'Population Signatures', and 'Gene Families'. The 'Gene Skyline' tab is currently selected.

Cell Type	Expression Level (approx.)
B.Sp	1000
B1.per	1100
NK.SpN	300
NK.Sp	200
DC.LN	250
CD8Nve.LN	50
CD8Act.LN	100
CD8Mem.LN	20
CD8.FT	10
CD8aa.FT	20
CD4.Sp	50
CD4Act.LN	50
CD4NKT.Sp	50
CD4Treg.LN	20
CD4Eff.PcN	150
CD4Treg.PcN	50
DP.FT	600
DPDeI.FT	550
MECb7lo.Th	1300
MEC.Th	500
AllSynFi	300

Field	Value
Probe Set	100600_at
Gene Symbol	Cd24a
Title	CD24a antigen
Aliases	HSA // Ly-52
Chromosome	10
Location	43,489,862
NCBI	12484
Unigene	Mm.29742
Ensembl	ENSMUSG000000047139
GO	membrane (16020) external side of plasma membrane (9897)
InterPro	

Gene.Skyline
Joslin Diabetes Center
rev 12

Gene Skyline | Gene Constellation | Population Signatures | Gene Families

The **“Gene Constellation”** page presents the genes which are most correlated to a central input gene. Spatial coordinates are used to depict the tightness of this correlation, as well as secondary attributes of these correlated genes: the position of these correlated genes can be chosen to represent genomic position, functional relationships, or secondary correlations.

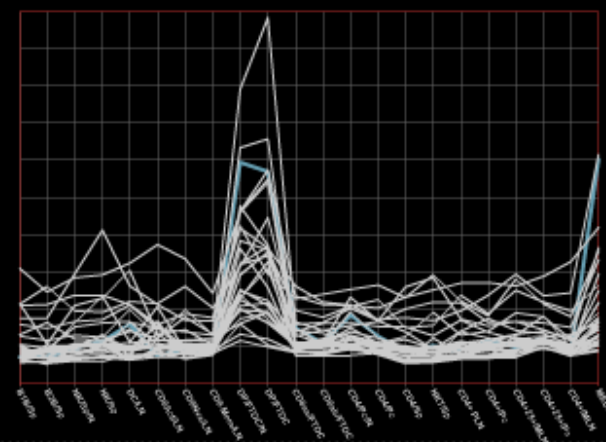
User Settings:

Setting	Description	Values
Datagroup	Sets the datagroup in which the correlations between genes are calculated	ImmGen (all immunological datasets) ImmGenShort (only one representative of the major cell types: B, T4, T8, NK, DC) ImmGenT (T lymphocytes only)
Distance metric	The measure of “similarity” between genes used to evaluate their distance	Standard Correlation: Similarities calculated as correlation coefficients across the datagroup (default) Two-tiered/T : Primary correlation coefficients are first calculated across the whole ImmGen dataset, and secondarily validated by correlation across a smaller subset (T lymphocytes only; avoids artificially high correlations that only correspond to skewed cell distribution). Only valid for ImmGen as a datagroup. Two-tiered/SymAtlas : As above, except that primary correlations in ImmGen are secondarily filtered by correlation in SymAtlas. Only valid for ImmGen as a datagroup
Arrange by	What parameter governs the angular position on the circle of correlated genes	GO cluster: The position of all the correlated genes reflects the function of the gene, as deduced by their Gene Ontology attributions. Genes with shared functions and cellular localizations tend to be grouped together (but still quite imperfect, as GO annotation is far from complete). Genes of unknown function are mapped to the “empty quadrant” (>270 degrees, top left) Gene Location: The position reflects the chromosomal position of the genes, clockwise. Chr1 has the smallest angle, then Chr2, etc. Genes of unknown location are mapped to the “empty quadrant” (>270 degrees) Second. Correl.: Secondary correlations between the secondary genes (by PCA or hierarchical clustering)
Back	Navigation button	Returns to the previous search page (useful when performing iterative browsing)
Search	Set the ProbeID to use as the central gene	An Affymetrix probe name (“at number”). It is easier, in practice, to call the correlation of one gene from the Skyline page with the “Show Constellation for this gene” action
Show	The number of correlated genes shown	Max 80 genes (the display gets busy beyond 50 genes showing)
Zoom	Sets the correlation coefficient that corresponds to the full radius of the correlation disc	Between 0.5 and 1. Small values display a broad range of correlations, but tend to bunch highly correlated genes at the center; higher values provide better resolution for the best correlated genes, but lesser fits project outside the screen.

The screenshot displays the Gene Constellation tool interface. At the top, a navigation bar includes a 'data group' dropdown set to 'All immune, with su', a 'distance metric' dropdown set to 'Standard Correlation', and an 'angle' dropdown set to 'GO Cluster'. Below this is a search bar containing '101963_at', a 'show' button set to '30', a 'zoom' slider set to '0.7', and an 'Execute' button. A yellow callout points to the 'Back' button, stating 'Return to the previous search configurations'. Another yellow callout points to the 'Execute' button, stating 'Starts the search'. The main area features a circular gene constellation with 'AA536749' at the center. Various genes are connected to the center by lines, including 'Plxnd1', '2510015F01Rik', 'D13Wsu64e', 'Ptov1', 'Ptpfr', 'H1f0', 'Ncam1', 'Rorc', 'Zipro1', 'Rag2', 'Plk4', 'Ccr9', 'Cldn4', 'Npc1', 'Ltb4', 'Aldh3a1', 'Gih2', 'Ssx2ip', 'Ssxzip', 'Cdca7', 'Morf4l2', 'Tubb2', 'Fbln2', and 'A730011O11Rik'. A yellow callout points to 'Npc1', stating 'Clicking on any of the related genes will start a search for all genes related to this one'. A red callout points to the 'Npc1' node, stating '“mouse over” on any gene will bring up (red) its expression profile on the display at right, and a pop-up with the basic annotation for this gene'. On the right, a 'View Skyline' button is circled in yellow. Below it, a table displays gene annotations for 'Ctsl' (cathepsin L), including its title, aliases, chromosome, location, locus link, unigene, and Ensembl ID. The GO terms listed are 'cysteine-type endopeptidase activity (4197)', 'cathepsin L activity (4217)', and 'extracellular space (5615)'. The InterPro entry is 'Peptidase, eukaryotic cysteine peptidase active site (IPR000169)' and 'Peptidase C1A, papain (IPR000668)'. At the bottom right, a gene skyline plot shows expression profiles for various genes across different conditions. The bottom of the interface includes a 'Star Count: 31' and a 'Gene Constellation' tab selected among 'Gene Skyline' and 'Population Signatures'.

[View Skyline](#)

Gene Symbol:	Ctsl
Title:	cathepsin L
Aliases:	
Chromosome:	13
Location:	63,268,268
LocusLink:	13039
Unigene:	Mm.930
Ensembl:	ENSMUSG000000021477
GO:	cysteine-type endopeptidase activity (4197) cathepsin L activity (4217) extracellular space (5615)
InterPro:	Peptidase, eukaryotic cysteine peptidase active site (IPR000169) Peptidase C1A, papain (IPR000668)



Star Count: 31

Gene Constellation
Joslin Diabetes Center

Gene Skyline

Gene Constellation

Population Signatures

The **“Population Signatures”** compares individual populations in the ImmGen datagroup for their overall relatedness, and can be queried to display the genes that most distinguish two different populations.

User Settings:

Setting	Description	Values
Datagroup	Sets the datagroup in which the correlations between genes are calculated	ImmGen (all immunological datasets, only one active at present)

Population Signature Actions



The **“Gene Families”** page displays all members of a gene family (e.g. Transcription Factors) in a 2-dimensional space, in a manner that best represents their relative distances (projection by multidimensional scaling). All genes are shown as a dot, and can be highlighted to display a unique gene, or several members of a subfamily, or several genes whose expression distinguishes a chosen cell population. This page can take 10-15 seconds for Flash to display, particularly when >30 genes are called and when working on a sluggish computer.

User Settings:

Setting	Description	Possible Values
Datagroup	Sets the datagroup for which the expression values of the gene are displayed	ImmGen (all immunological datasets)
Gene Family	Sets the gene family to display (e.g. transcription factor)	Transcription factor (only setting at present)
Genes where	Field to be searched	ProbeID, Gene Symbol, Gene Name, NCBI GeneID, Chromosome, Unigene, etc
Search Type	Stringency of the search	Contains, Exact Match, Begins With.
String	String of characters to be searched for.	Any string of numbers or characters. The database engine will search for these characters within the specified field.
Genes distinguishing	Sets the cell population(s) for which the genes have distinguishing expression	All, B lymphocytes, NK cells, DCs, DP over mature T, mature T over DPs, T over B.

Zoom display: As the display can be quite dense in certain areas, a zoom allows expansion of selected regions of the display. A depressed mouse (left on PC), followed by a movement of the mouse, generates a marquee selection; upon releasing the mouse, the graph is rescaled such that the screen encompasses the whole selected region. To return to full screen view, click on the “negative magnifying lens”, top left.

Mouse-over on a gene brings its annotation in the fields at right, and its expression profile at bottom right. Full click on a gene fixes its annotation (for navigation to additional information on this gene, using the “View Skyline” and “View Constellation” buttons, top right).

Gene Families Actions

Choose the Gene Family to display

Search settings for the genes to highlight

Zoom out

Clicking on a gene fixes its annotation (top right) and expression profile (bottom right)

"mouse over" any gene will bring up (red) its annotation (top right) and expression profile (bottom right)

View Skyline
View Constellation

Switch to the Constellation view for this gene

data group: immGen

gene family: Transcription Factors

genes where Gene Symbol Contains the string irf

OR genes distinguishing -- select list --

Execute

Probe Set:	
Gene Symbol:	
Title:	
Aliases:	
Chromosome:	
Location:	
NCBI:	
Unigene:	
Ensembl:	
GO:	
InterPro:	

View Skyline
View Constellation

Switch to the Constellation view for this gene

Gene.Families
Joslin Diabetes Center
rev 4

Gene Skyline Gene Constellation Population Signatures Gene Families

B. Computations for the ImmGen website

Essentially all computations were performed in S-Plus 6.1, applying native S+ functions and the Insightful ArrayAnalyzer package, on Dell Precision or Falcon Northwest workstations under Windows 2000. All scripts are available upon request, and computational details are outlined in their documentation. The general steps are described below.

Data preprocessing. Individual .cel files were uploaded using ArrayAnalyzer's Import Affymetrix Data dialog. Expression values were background-corrected using the MAS algorithm, where only perfect match probes were taken into consideration given the error that can be introduced by signals from "mismatched" probes (Irizarry et al, *Biostatistics*, (2003) 4:249-264). Each population was represented by several independent replicates, and the expression value for each gene was calculated by simple averaging.

All annotation values downloaded from the Affymetrix Net-Affx web site (most computations used the December 2004 release for Mu74aV2). Genomic locations of the chip's elements and GeneOntology (GO) identifiers were parsed into matrices of numeric values for S+ computations. When multiple assignments were given for a probe, we took the one with the highest identity value (thus implicitly accepting that some of the expression values may be confounded by dual reactivity, and that some of the genomic assignments may be an oversimplification).

Estimates of the numbers of expressed genes. A probabilistic approach was taken to determine whether a given gene is expressed in a given dataset. The matrix of expression values was log transformed for better normality, and the distribution of negatives was estimated from the control features on the Mu74Av2 and GNF1 chips (ignoring the control yeast genes which gave clearly positive expression signals, most likely from cross-hybridization). Mean and standard deviation were calculated from this set, with which we calculated the probability that the expression value for each gene in each population belonged to the distribution of negative controls. A gene were considered to be expressed if its probability of non-expression was less than 0.02 in 1/20th or more of the populations, or if its negative probability was less than 0.001 in a single population. The "breadth" of expression was then calculated for genes passing this filter as the proportion of cell populations in which the gene was expressed at a negative probability of <0.1.

Population analysis – Principal component Analysis (Fig1) Principal Component Analysis was performed with a custom S+ script ("PCAPopulationAnalysis"). Data matrices were prepared by population- and gene-wise means-centering, and the S+ "princomp"

function was used to extract the Principal Components. A population's correlation for a pair of components were then used as its coordinates in 2-dimensional representations. Although some distinct patterns were observed with components of lower order (particularly when non-immunological populations were included), the pairing of components 1 and 2 was used most commonly (Fig. 2), as these are the components with the most informativity. In most instances, the algorithm was run on gene sets selected to be the most informative, i.e. after filtering on expression value and variability ($CV > 0.3$). In other cases, the analysis was performed after selecting gene subsets 1) on the basis of their molecular function (i.e. by sharing a common GeneOntology (GO) identifier) or 2) as those best differentiating two input cell populations or two groups of populations (i.e. the genes showing the most extreme FoldChange between the reference populations). To generate randomized control plots, the expression values for each gene in the input data matrix were randomly reassorted row-wise, and the same script was applied.

Population analysis – ReferencePopulation Analysis (Fig1) Custom S+ scripts (“ReferencePopulationsPopulationAnalysis”) were used to generate plots where individual populations are positionned according to their expression of defined sets of genes. One reference population (or group of populations) is chosen by the user as the X-population, another as the Y-population. The X-genes are then selected as those that are overexpressed in the X-population relative to the Y-population (thresholding on fold-change), and vice-versa for the Y-genes. For each population in the datagroup, its expression value for X-genes is then scaled, from 0 (value in the Y-population) to 1 (value in the X-population), and the converse is calculated for Y-genes. These scaled X-gene for this population are then averaged, as are the scaled Y-gene values, generating the x and y coordinates of the population.

Gene correlations (Fig2, ImmGen “Gene Constellation” page) A table showing the best correlations for each gene was prepared with the “ImmGenGeneCorrelTopN” script. Expression values were sequentially normalized by population-wise and gene-wise means, and the genes filtered with thresholds on expression values (to eliminate genes whose expression values are too close to background and hence subject to experimental noise) and on variation across populations (to eliminate “housekeeping” genes which yield erroneously high correlation coefficients). The Pearson correlation coefficient of each gene against all others was then calculated using the vectorized “cor” function of S+, and the best values (typically the top 100) were stored.

Two values are used to position each correlated gene within the 2-dimensional circular representation of the GeneConstellation page: d , the distance from the input gene represented at the center, and θ , the angular position in the circle around the input gene (clockwise, with 0 being the position immediately above the input gene). The distance from the central input gene is a linear transformation from the correlation coefficient. The angular momentes at which each gene is shown vary with the display mode, and is calculated as follows:

- To represent the secondary correlations within the set of genes whose expression is correlated to that of an input gene. The normalized expression values of the selected genes across all populations was used as input for hierarchical clustering (“hclust”

function, using the overall Euclidian distance between normalized expression values). The order vector resulting from the clustering was used to position each of the genes, equally spaced between 20 and 340 degrees.

- To represent common functionalities within the set of genes related to a primary gene, based on GO attributes of the genes. The pool of all GO identifiers tied to the genes in the correlated set was retrieved from the annotation tables, and used to generate a logical (gene x GO) matrix of annotations (set to T if the annotation for gene i included the GO identifier j, and to F if not). This matrix then formed the input for hierarchical clustering (hclust(dist)). The order vector returned by the clustering was then used to position the genes, as above, from 10 to 270 degrees. Since the order created by hierarchical clustering includes “jumps”, the actual spacing between neighboring genes was refined as a function of the number of GO identifiers shared. Genes devoid of GO identifiers (un-annotated genes or ESTs) were positioned at random between 300 and 360 degrees.
- To represent genomic locations, the angle for each gene was calculated as a linear function of its chromosomal position, such that genes encoded on Chr1-19 and X/Y were evenly spaced between 20 and 270 degrees. Genes without a known location (ESTs or genes with ambiguous locations) were positioned at random in the 300-360 quadrant.

Spatial representation of gene families (Fig3, ImmGen “Gene Family” page). The “GeneFamilyMDS1” script was used to calculate spatial coordinates for each transcription factor, using the isoMDS function in the MASS library of S+. Probes on the chip corresponding to transcription factors were identified in the annotation tables (04/12 release of the Mu74Av2 chip) by matching with the GO identifiers 3700, 5667 and 45449. Expression values were normalized, and genes filtered on expression value and variability ($cv > 0.2$ over the entire datagroup) to avoid noise from low or invariant expression values. A square matrix of pairwise correlation coefficients between each of the 702 genes was calculated, and this matrix was used as input for multi-dimensional scaling with the isoMDS function (50 iterations); performing the scaling on the correlation matrix was found preferable to scaling from the Euclidian distance between genes, which leads to an alternative representation where expression values dominate one of the axes. This first computation generated spatial coordinates for all transcription factors, which are shown as dots of the two-dimensional space. Particular gene lists can be highlighted by the user, whose identities were pre-computed (“GeneFamilyMDS2” script); i) selecting those genes that distinguished groups of cell populations (filtering on fold-change between selected hi and low population, 2.0 in most cases) or ii) distinguish single populations compared to all others (in that case, selecting those TF genes with a foldchange > 1.4 in the designated population(s)). Aside from these “population specific” TFs, the Gene Family page of the ImmGen site can also display selectively those TFs that match a search criterion (by gene name, by chromosomal location, by identifier number).

C. Architecture of the ImmGen website

The ImmGen website (<http://ImmGen.org>) does not, as of this publication, perform computations in response to user queries but displays pre-computed information on gene expression or correlations. The primary data matrices and those generated by the S-Plus computations described above were converted into SQL tables, and incorporated into a standard three tier architecture consisting of a Presentation Layer, a Business Logic Layer, and a Data Management Layer, each is responsible for one specific task, simplifying the development and maintenance of the project.

The Data Management Layer, responsible for storing and retrieving the requested data, is a version 8.0.2 PostgreSQL Relational Database Management Server (RDBMS), (open source database available at www.postgresql.org), running on a Windows 2000 server. The PostgreSQL database was chosen to host the data for this project for cost and because it is supported on a wide variety of platforms (testing indicated a performance equal or superior to a parallel database running on MS SQL Server)..

The Presentation Layer, responsible for the user interface displaying the data was created using Macromedia Flash, version 7. Flash was chosen for two reasons: 1) it is normally already installed on most computers in the target audience, and 2) it provides a better basic graphical functionality on which to build an interactive application than other tools. Other technologies were considered for this interface; however, most require the building of the basic graphical functionality that Flash already provides.

The Business Logic Layer, responsible for translating a user's interactions at the Presentation Layer into data requests from the Data Management Layer and transforming this data prior to responding to the original request. Jakarta Tomcat version 5.0.28, a Java Servlet Container, (available from <http://jakarta.apache.org/tomcat>), (choice based on open-source and support on a wide variety of platforms). Tomcat also provides simple methods for transforming XML (eXtensible Markup Language)-formatted requests from the Flash interface into actionable commands to the database, and then transforming the received data back into XML to return to the user interface.

Although Tomcat includes a fully functional web server, we chose to serve the static content, including the Flash .SWF files that compose the user interface from Microsoft's Internet Information Server (IIS). This choice divides the task of serving the content between products that complement each other: IIS is highly optimized to quickly and efficiently deliver static files to the clients' web browsers; whereas, Tomcat is optimized to support Java Servlets. The Java servlet was written using version 3.6 of the NetBeans open-source Integrated Development Environment (IDE), (available at www.netbeans.org).

When the Java Servlet starts, some (constant) data are cached locally to reduce the number of database lookups when servicing a user request. These data include the complete description of the datagroups, distance metrics, cell populations, and the constant constellation lookup information.

The datagroup is the key table for all user queries – it determines which set of tables to use in the queries. Each datagroup identifies which chip (as of publication, data from either the GNF1 or MU74Av2 chips are displayed) was used to obtain the data, and which annotation table to use for the retrieved set of probes. Also, each datagroup identifies what tables contain the expression values for the

Skyline page, Population Signatures (color codes and gene signatures), and the Gene Families (gene family and gene family data). The Constellation lookup information includes what table contains the data for each datagroup/distance metric pair (top correlation). The response to all user queries contains the list of cell populations that define a datagroup. Each cell population includes the symbol, name, description, authorship (individual, lab, and institution), and a reference URL.

All the data, including the output from the various S-Plus scripts and the annotations, were transformed into CSV (Comma Separated Values) formatted files, then imported into the corresponding database tables, most with little or no further processing (in some cases, an additional identifier column was used to create the proper relationships between the data). The annotation, datagroups, cell populations, and authorship data were all normalized during import into the database.

All queries to the database are dynamically built – no stored procedures are used (simplifying, theoretically at least, conversion to a different database engine). The Java database driver and the PostgreSQL database engine co-operate to optimize queries to give the best performance, including switching to server-side cursors when running the same query repeatedly. (A configuration parameter in the driver controls when to attempt to switch to server-side cursors based on the number of times a query is repeated.) This change in cursor type has the effect of increasing performance similar to, but without the transactional overhead of, stored procedures.

The Skyline data are retrieved from one of two expression values tables and one of two annotation tables depending on which datagroup the user selected in the Flash interface. The columns used to retrieve the expression value data are based on the cell populations associated with the chosen datagroup. The retrieved data is then transformed into XML and sent back to the Flash interface, where the individual Skyline bar graphs are dynamically created.

The Constellation data depend not only on the chosen datagroup, but also on the selected distance metric: these two items determine which table from which to retrieve the data. Included with the Constellation data is the expression values data for the returned set of probes, for display in a line graph. The user may limit the number of probes to be returned (to a maximum of 100) in the Constellation data.

The color code data for the Population Signatures, like the Skyline data, is keyed off of the chosen datagroup, returned in ascending cell population display order and displayed in a dynamically generated colors table. The gene signature data likewise uses the chosen datagroup along with the selected cell from the color table to retrieve the associated expression values and annotation data; the expression values are displayed in a line graph similar to that for the Constellation data.

The Gene Families data, like the others, are based on the selected datagroup. The list of gene families is specific to a particular datagroup; the list of predefined highlight subsets is, in turn, keyed off of the selected gene family. The search results highlight only those probes that are in the intersection of the selected datagroup and the selected gene family. The expression values of any visible highlighted probes are displayed in a line graph similar to that shown for the Constellation data.