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 explored[3]. 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 own[4]. 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.

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 Wolfeboro Pediatrics, Wolfeboro, 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

'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

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 limited[19]. 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 populations[23]. The Genentech IRIS project[24] 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

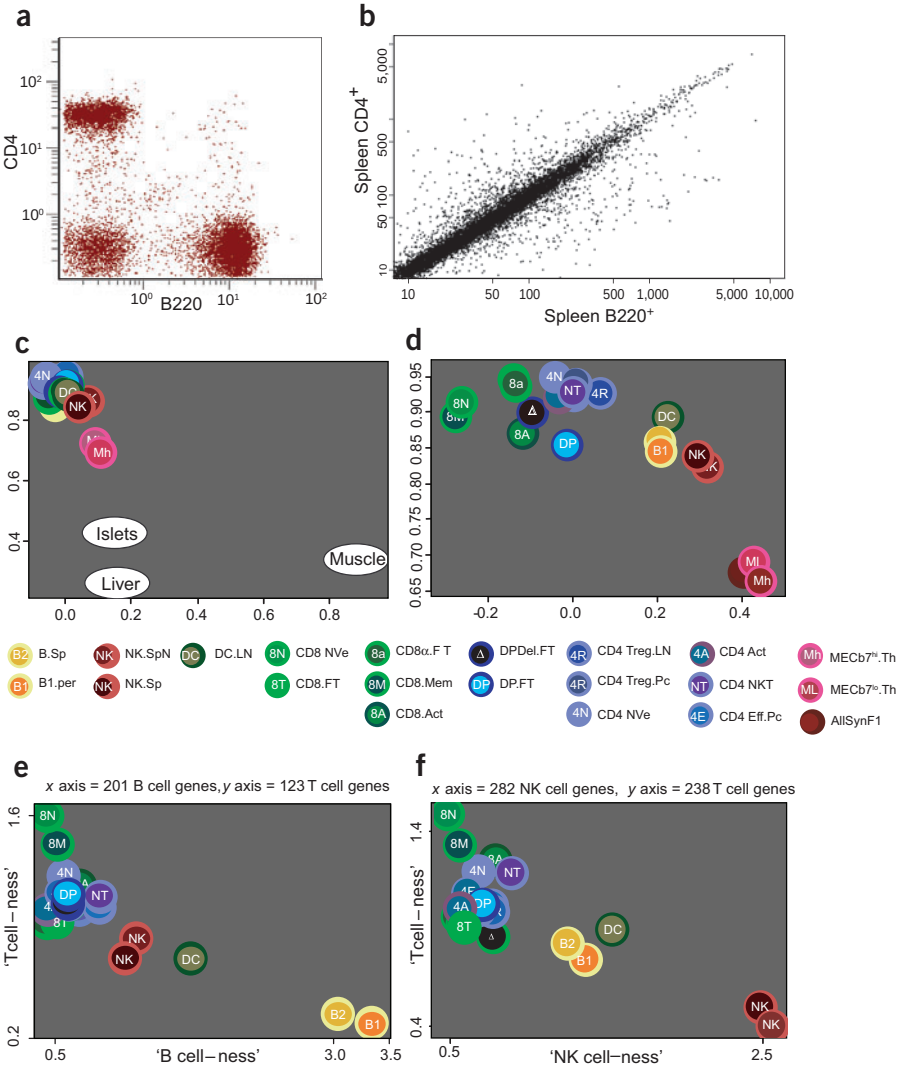

Figure 1 Differentiating cells by microarray analysis. (**a**,**b**) B lymphocytes and T lymphocytes, distinguished by flow cytometry (**a**) or gene expression profile (**b**). (**c**,**d**) Relative positioning of immune cells based on principal component analysis of their global expression profiles; 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 abbereviated 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, CD8αβ single-positive cells, fetal thymic organ culture; CD8α.FT, CD8αα 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; $MECb7^{hi}$.Th, medullary epithelial cells, $B7^{hi}$, thymus; $MECb7^{lo}$.Th, medullary epithelial cells, $B7^{lo}$, thymus; AllSynFl, arthritic synovial fluid.

somewhat more common. The IRIS study found 1,688 of 19,054 (8.8%) human genes to be 'preferentially' expressed in immune cells. A rarity of specific transcripts also applies in distinguishing between types of immune cells. Of the 67% of genes expressed in immune cells, most were detected in essentially all or shared between several immune cell types (36% and 21%, respectively). Only a minority (9.5%) was expressed exclusively in a single immune cell type.

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 estimated that 10,000-15,000 individual genes are expressed in a given cell, but that only 2% of transcripts differ qualitatively between B cells and T cells[25,26].

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 hematopeoietic 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 those of the cluster of cells of hematopoietic origin, suggesting that functional interactions with lymphocytes require the sharing of a particular range of transcripts and proteins (such as adhesion molecules) or that a cell's location in an organ may modify the transcriptional patterns dictated by its developmental origin.

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 lymphocytes[27], 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 markers[28]. In reference population analysis plots, in which each population is positioned in the two-dimensional

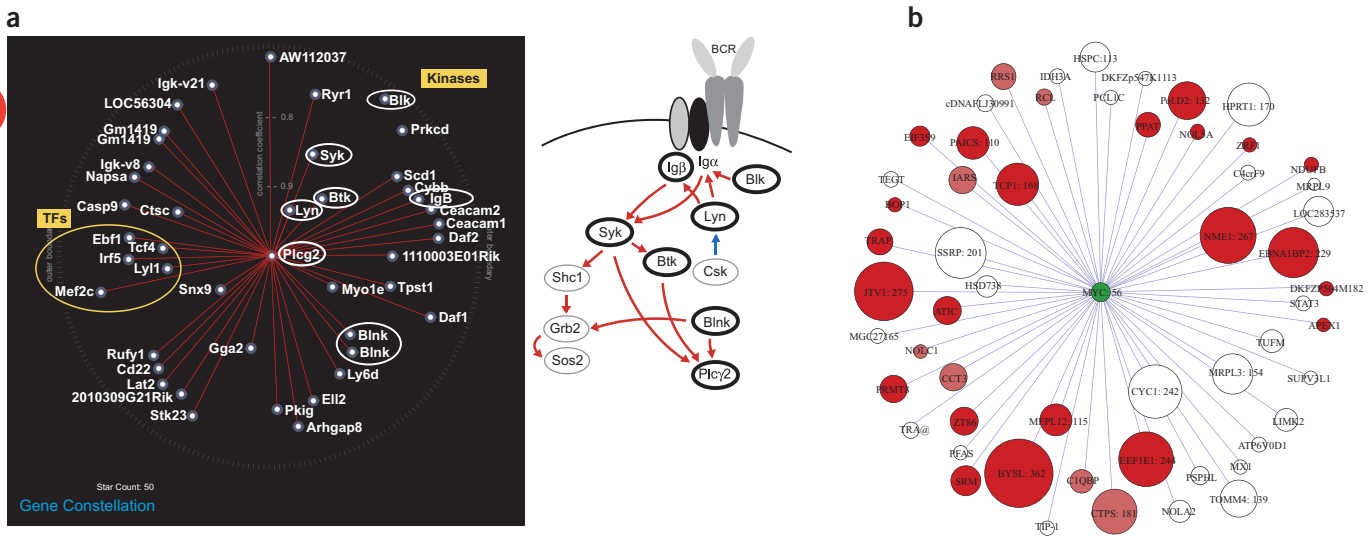

Figure 2 Gene expression networks in the immune system. (**a**) Network of correlated expression in immune cells. Data represent genes whose expression profiles are most correlated with that of *Plcg2* (center), in the 76 data sets of the ImmGen data, which includes all of the main cell types of the 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 cells[45]. 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 populations[29], 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 lymphocytes[21] and with the existence of 'natural killer–dendritic' cells[30]. 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 fork[31]. The picture obtained from these genome-wide analyses is more compatible with views arguing for considerable plasticity in lymphoid and myeloid cell differentiation[32], 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 anergy[9]). 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 emerging[39] (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 networks[40], 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 networks[42], 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 past[43]. 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 knockout, RNA interference 'knockdown', or drug

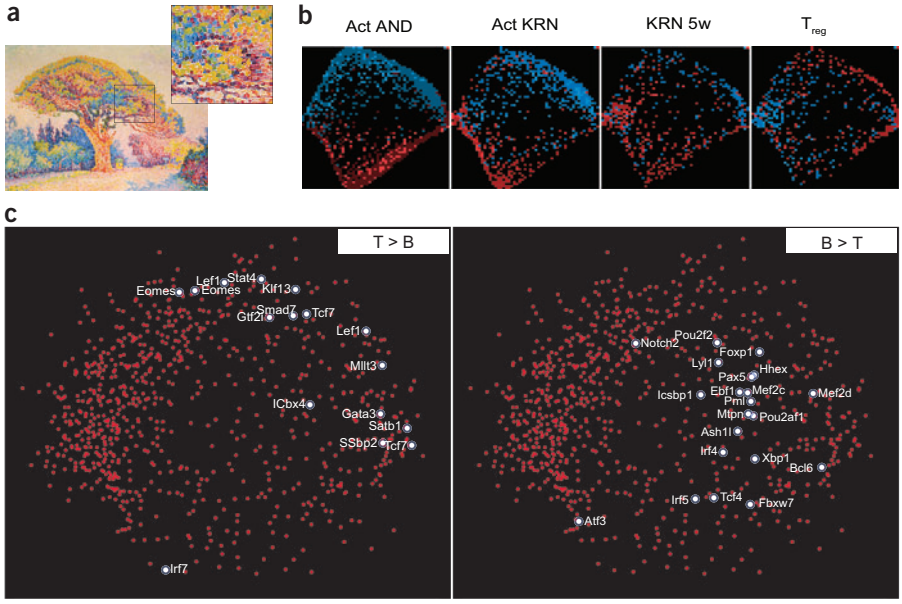

Figure 3 Complex data representations in art and immunogenomics. (**a**) Pine Tree by Paul Signac (1909; Pushkin Museum, Moscow). Inset, enlargement of boxed area at left. (**b**) Abstract representation of expression signatures in subsets of CD4$^+$ T cells. Expression profiles of CD4$^+$ T cell populations 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 condition[10] (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).

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' techniques[45]. 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 data[50], some of which should be 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 power[51].

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.

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., McCaughtry, 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. Crampton, 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).