Automated discovery of functional generality of human gene expression programs

Georg K. Gerber[1,2], Robin D. Dowell[1], Tommi S. Jaakkola[1], David K. Gifford[1,3]

[1] MIT Department of Computer Science and Electrical Engineering, 32 Vassar Street, Cambridge, MA 02139.
[2] Harvard-MIT Division of Health Sciences and Technology, 45 Carleton Street Room E25-519, Cambridge, MA 02139.
[3] Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge MA 02142.

Correspondence should be addressed to David Gifford (gifford@mit.edu), 32 Vassar Street G542, Cambridge, MA 02139. Telephone: (617) 253-6039.

Abbreviations used in the text: Expression program (EP), Gene Ontology (GO), or Kyoto Encyclopedia of Genes and Genomes (KEGG), Markov Chain Monte Carlo (MCMC), non-negative matrix factorization (NMF), Peripheral blood mononuclear cell (PBMC), and Singular Value Decomposition (SVD).

Abstract

An important research problem in computational biology is the identification of *expression programs*, sets of co-expressed genes orchestrating normal or pathological processes, and the characterization of the functional breadth of these programs. The use of human expression data compendia for discovery of such programs presents several challenges, including: 1) cellular inhomogeneity within samples, 2) genetic and environmental variation across samples, 3) uncertainty in the numbers of programs and sample populations, and 4) temporal behavior. We developed GeneProgram, a new unsupervised Bayesian computational framework based on Hierarchical Dirichlet Processes that addresses each of the above challenges. GeneProgram uses expression data to simultaneously organize tissues into groups and genes into overlapping programs with consistent temporal behavior, to produce maps of expression programs, which are sorted by generality scores that exploit the automatically learned groupings. Using synthetic and real gene expression data, we showed that GeneProgram outperformed several popular expression analysis methods. We applied GeneProgram to a compendium of sixty-two short time-series gene expression data sets exploring the responses of human cells to infectious agents and immune-modulating molecules. GeneProgram produced a map of 104 expression programs, a substantial number of which were significantly enriched for genes involved in key signaling pathways and/or bound by NF-κB transcription factors in genome-wide experiments. Further, GeneProgram discovered expression programs that appear to implicate surprising signaling pathways or receptor types in the response to infection, including Wnt signaling and neurotransmitter receptors. We believe the discovered map will be useful for guiding future biological experiments; genes from programs with low generality scores might serve as new drug targets that exhibit minimal "cross-talk," and genes from high generality programs may maintain common physiological responses that go awry in

disease states. Further, our method is multipurpose, and can be applied readily to novel compendia of biological data.

Synopsis

In recent years, DNA microarrays have been used to produce large compendia of human gene expression data, which are promising resources for discovery of *expression programs*, sets of co-expressed genes orchestrating important physiological or pathological processes. However, these compendia present particular challenges, including cellular inhomogeneity within samples, genetic and environmental variation across samples, uncertainty in the numbers of programs and sample populations, and temporal behavior. To address these challenges, we developed GeneProgram, a state-of-the-art statistical framework that automatically generates interpretable maps of expression programs from expression data. GeneProgram accomplishes this by simultaneously organizing tissues into groups and genes into overlapping programs with consistent temporal behavior, and sorting programs by a generality score. Such maps may be valuable for guiding future biological experiments; genes from programs with low generality scores might serve as new drug targets that exhibit minimal "cross-talk," and genes from high generality programs may maintain common physiological responses that go awry in disease states. Using synthetic and real data, we showed that GeneProgram outperformed several popular expression analysis methods. Further, on a compendium of time-series gene expression data measuring the responses of human cells to infectious agents, GeneProgram discovered programs that implicate surprising signaling pathways and receptor types.

Introduction

The great complexity of the human body, both in normal physiology and in pathological states, arises from the coordinated expression of genes. A fundamental challenge in computational biology is the identification of sets of co-activated genes in a given biological context and the characterization of the functional breadth of such sets. Understanding of the functional generality of gene sets has both practical and theoretical utility. Sets of genes that are very specific to a particular cell type or pathological condition may be useful as diagnostic markers or drug targets. In contrast, sets of genes that are active across diverse cell types or pathological states can give us insight into unexpected functional similarities and involvement of core common pathways.

In this study, we use a large compendium of short time-series gene expression data sets measuring the responses of human cells to infectious agents or immune-modulating molecules, to discover a set of biologically interpretable expression programs and to characterize quantitatively the specificity of each program. Such large genome-wide human expression data compendia present several new challenges that do not necessarily arise when analyzing data from simpler organisms. First, tissue samples may represent collections of diverse cell-types mixed together in different proportions. Even if a sample

consists of a relatively homogenous cell population, the cells can still behave asynchronously. Second, each tissue sample is often from a different individual, so that the compendium represents a patchwork of samples from different genetic and environmental backgrounds. Third, the number of expression programs and distinct cell populations present in a compendium is effectively unknown *a priori*. Fourth, a compendium may contain experiments measuring temporal responses over different durations or using varied sampling rates.

We present a novel methodology, GeneProgram, designed for analyzing large compendia of human expression data, which simultaneously compresses sets of genes into expression programs and sets of tissues into groups. Specific features of our algorithm address each of the above issues relating to analysis of compendia of human gene expression data. First, our method handles tissue inhomogeneity by allocating the total mRNA recovered from each tissue to different gene expression programs, which may be shared across tissues. The number of expression programs used by a tissue therefore relates to its functional homogeneity. We address the second issue, of tissue samples coming from different individuals, by explicitly modeling each tissue as a sample from a population of related tissues. That is, related tissues are assumed to use similar expression programs and to similar extents, but the precise number of genes and the identity of genes used from each program may vary in each sample. Additionally, populations of related tissues are discovered automatically, and provide a natural means for characterizing the generality of expression programs. Third, uncertainty in the numbers of tissue groups and expression programs is handled by using a non-parametric Bayesian technique, Dirichlet Processes, which provides prior distributions over numbers of sets. Fourth, GeneProgram explicitly models patterns of temporal expression change using the novel concept of program *usage modifiers*, variables that alter the manner in which each tissue uses an expression program. That is, both the genes used by a tissue from a program, and the manner in which they are expressed (e.g., early induction versus late repression) are chosen probabilistically and influenced by the behavior of similar tissues. Further, usage is consistent across a program for a particular tissue, which facilitates biological interpretation.

To understand the novel contributions of the GeneProgram algorithm, it is useful to view our framework in the context of a lineage of unsupervised learning algorithms that have previously been applied to gene expression data. These algorithms are diverse, and can be classified according to various features, such as whether they use matrix factorization methods [1], heuristic scoring functions [2], generative probabilistic models [3], statistical tests [4,5], or some combinations of these methods [6,7]. The simplest methods, such as K-means clustering, assume that all genes in a cluster are co-expressed across all tissues, and that there is no overlap among clusters. Next in this lineage are biclustering algorithms [2,5,8-10], which assume that all genes in a bicluster are co-expressed across a subset rather than across all tissues. In many such algorithms, genes can naturally belong to multiple biclusters.

GeneProgram is based on two newer unsupervised learning frameworks, the *topic model* [11,12] and the Hierarchical Dirichlet Process mixture model [13]. The topic model

formalism allows GeneProgram to further relax the assumptions of typical biclustering methods, through a probabilistic model in which each gene in an expression program has a (potentially) different chance of being co-expressed in a subset of tissues. The hierarchical structure of our model, which encodes the assumption that groups of tissues are more likely to use similar sets of expression programs in similar proportions, also provides advantages. Hierarchical models tend to be more robust to noise, because statistical strength is "borrowed" from items in the same group for estimating the parameters of clusters. Additionally, hierarchical models can often be interpreted more easily – in the context of the present application, the inferred expression programs will tend to be used by biologically coherent sets of tissues or experimental conditions. Finally, through the Dirichlet Process mixture model formalism, GeneProgram automatically infers the numbers of gene expression programs and tissue groups. Because this approach is fully Bayesian, the numbers of mixture components can be effectively integrated over during inference, and the complexity of the model is automatically penalized. This is in contrast to previous methods that either require the user to specify the number of clusters directly or produce as many clusters as are deemed significant with respect to a heuristic or statistical score without providing a global complexity penalty. We note that Medvedovic *et al.* have also applied Dirichlet Process mixture models to gene expression analysis, but not in the context of topic models, Hierarchical Dirichlet Processes, or human data [14].

A variety of algorithms have been developed to analyze time-series expression data [15], but to our knowledge, none have been specifically designed for analysis of large compendia of such data. Analysis methods for combining time-series of different durations or that use different sampling rates have focused on long time-series over a few experimental conditions [16,17], rather than short series over many conditions as we do. Jenner and Young performed a meta-analysis of a superset of the infection time-course experiments we analyze in this paper using hierarchical clustering [18]. However, their analysis was not automated or statistically principled, relying on extensive prior biological knowledge, and using visual assessment of clusters to manually assign genes to pathways of interest. Further, as described in the Results and Discussion section, our analysis implicated several surprising signaling pathways and receptor types in the response to infection that previous analyses of these data sets have not.

The remainder of this paper is organized as follows. We first present an intuitive overview of the GeneProgram algorithm and probability model. We then use synthetic data to examine the kinds of structures that GeneProgram and several other classes of algorithms can recover from noisy data. Next, we benchmark GeneProgram on two large compendia of mammalian data [19,20] by comparing our algorithm's ability to recover biologically relevant gene sets to those of two popular biclustering methods. We then apply GeneProgram to a compendium of sixty-two short time-series gene expression data sets measuring the responses of human cells to infectious agents or immune-modulating molecules [21-26], and produce a map of expression programs organized by functional generality scores. We evaluate the biological relevance of the discovered expression programs using biological process categories [27] and pathway [28] databases, as well as genome-wide data profiling binding of NF-κB family transcription factors [29]. Finally,

we provide examples of discovered expression programs involved in key pathways related to the response to infection, discuss the significance of our results, and comment on possible future research directions.

Results and Discussion

The GeneProgram algorithm and probability model

Algorithm overview. The GeneProgram algorithm consists of data pre-processing, model inference, and distribution summary steps as depicted in Figure 1. Data pre-processing steps make data from multiple tissues comparable and discretize continuous values in preparation for input to the model. The model inference step seeks to discover underlying expression programs and tissue groups in the data probabilistically. To accomplish this, we use Markov Chain Monte Carlo (MCMC) sampling [30], an approximate inference method, to estimate the model posterior probability distribution. Each posterior sample describes a configuration of expression programs and tissue groups for the entire data set; more probable configurations tend to occur in more samples. The final step of the algorithm is model summarization, which produces consensus descriptions of expression programs and tissue groups from the posterior samples. See the Methods section for details.

Probability model overview. We can understand the GeneProgram probability model intuitively as a series of "recipes" for constructing the gene expression of tissues. Figure 2 presents a cartoon of this process, in which we imagine that we are generating the expression data for the digestive tract of a person. The digestive tract is composed of a variety of cell types, with cells of a given type living in different microenvironments, and thus expressing somewhat different sets of genes. We can envision each cell in an organ choosing to express a subset of genes from relevant expression programs; some programs will be shared among many cell types and others will be more specific. As we move along the digestive tract, the cell types present will change and different expression programs will become active. However, based on the similar physiological functions of the tissues of the digestive tract, we expect more extensive sharing of expression programs than we would between dissimilar organs such as the brain and kidneys. As can be seen in Figure 2, the final steps of our imaginary data generation experiment involve organ dissection, homogenization, cell lysis and nucleic acid extraction, to yield the total mRNA expressed in the tissue, which is then measured on a DNA microarray.

The conceptual experiment described above for "constructing" collections of mRNA molecules from tissues is analogous to the *topic model*, a probabilistic method developed for information retrieval applications [12,31] and also applied to other domains, such as computer vision [32] and haploinsufficiency profiling [11]. In topic models for information retrieval applications, documents are represented as unordered collections of words, and documents are decomposed into sets of related words called topics that may be shared across documents. In hierarchical versions of such models, documents are further organized into categories and topics are preferentially shared within the same category. In the GeneProgram model, a unit of mRNA detectable on a microarray is

analogous to an individual word in the topic model. Related tissue populations (tissue groups) are analogous to document categories, tissues are analogous to documents, and topics are analogous to expression programs.

GeneProgram extends the hierarchical topic model to capture general patterns of expression changes, such as induction/repression or temporal dynamics, through the novel concept of program *usage modifiers*, which are variables that alter the manner in which each tissue uses an expression program. For instance, for time-series data, usage modifiers would take on values of particular temporal patterns, such as early, middle or late induction or repression of expression. A tissue then probabilistically chooses a set of genes from a program, and also a setting for its usage modifier (e.g., the particular temporal pattern with which to express genes from the program). Note that usage is consistent across a program for a particular tissue, which facilitates biological interpretation. Further, the choice of how to use an expression program is influenced by the group a tissue belongs to.

GeneProgram handles uncertainty in the numbers of expression programs and tissue groups by using Dirichlet Processes, a non-parametric Bayesian statistical method that provides a prior distribution over the numbers of programs and tissues groups while penalizing model complexity. More specifically, GeneProgram is based on Hierarchical Dirichlet Process mixture models [13], which allow data items to be assigned to groups. Items in the same group preferentially share mixture components, although mixture components may be used across the entire model. We note that in the original Hierarchical Dirichlet Processes formulation [13], items were required to be manually assigned to groups. The GeneProgram model extends this work, automatically determining the number of groups and tissue memberships in the groups.

The GeneProgram probability model consists of a three-level hierarchy of Dirichlet Processes, as depicted in Figure 3A. Tissue samples are at the lowest level in the hierarchy. Each tissue sample is characterized by a mixture (weighted combination) of expression programs and a set of usage modifiers that are used to describe the observed gene expression patterns in the tissue sample. An expression program represents a set of genes that are co-expressed to varying extents and used consistently by each tissue, as depicted in Figure 3B. Tissue samples differ in terms of which expression programs they employ, how the programs are weighted, and how the programs are used. The middle level of the hierarchy consists of tissue groups, in which each group represents tissue samples that are similar in their use of expression programs (in terms of both weightings and usage modifiers). The highest and root level in the hierarchy describes a base level mixture of expression programs that is not tissue sample or group specific.

Each node in our hierarchical model maintains a mixture of gene expression programs, and the mixtures at the level below are constructed on the basis of those above. Thus, a tissue sample is decomposed into a collection of gene expression programs, which are potentially shared across the entire model, but are more likely to be shared by related tissues (those in the same tissue group). Further, the usage of each program may differ between tissues, but is more likely to be the same in related tissues. Because our model

uses Dirichlet Processes, the numbers of both expression programs and tissue groups are not fixed and may vary with each sample from the model posterior distribution. See the Supplemental Methods section for complete details.

GeneProgram accurately recovered coherent gene sets in noisy synthetic data that other algorithms could not

We used a simple synthetic data example to explore the kinds of structures GeneProgram and several other well known unsupervised learning algorithms could recover from noisy data. Our objective with these experiments was to use simulated data to illustrate the capabilities of the algorithms; whether or not particular structures are present in real data can only be answered empirically, and is addressed in the next subsection. In creating synthetic data, we sought to simulate important features of real microarray data profiling human tissues. Thus, we assumed noisy data in which there were several distinct populations of related tissues using different sets of co-expressed genes. In particular, the simulated data contained four sets of co-expressed genes used by 40 tissues divided equally among four tissue populations (see Figure 4A). Each gene set contained 40 genes with varying mRNA levels; gene sets three and four overlapped in 10 genes. See the Methods section for details. We note that this scheme for simulating data does not simply recapitulate the assumptions present in the GeneProgram model (e.g., it does not assume discrete and independent "units" of expression signal and it introduces microarray-like noise).

Hierarchical clustering is one of the most frequently used methods for clustering microarray gene expression data. Figure 4B shows the results of hierarchical clustering, using Pearson correlation as a similarity metric and the average linkage method [33], applied to sorting both rows (genes) and columns (tissues) of the synthetic data. As can be seen, hierarchical clustering did reasonably well at sorting tissues and genes independently, although it did not separate gene sets three and four correctly. But, this method's failure to consider genes and simultaneously is known to break up coherent "overhanging" blocks of genes, making interpretation difficult [2,5,8]. This issue was demonstrated in this example by gene sets one and two that "overhang," thus causing gene set two to be broken up horizontally (blue rows in Figures 4A and 4B).

The inability of hierarchical clustering to handle "overhanging" block structure in data was one of the motivations for the development of biclustering algorithms that take genes and tissues into account simultaneously [2,5,9]. To investigate the behavior of biclustering algorithms, we used Samba, an algorithm that has been shown previously to outperform other biclustering methods [5,34]. Samba produced 23 biclusters from the synthetic data (not shown). This method tended to find small subsets of genes co-expressed in some tissues, but did not recover the four gene sets as coherent biclusters. Presumably, this is because Samba does not attempt to incorporate more global constraints on biclusters.

Singular Value Decomposition (SVD) is a matrix factorization method that can be used to approximate a matrix using a smaller number of factors or components. In the context

of gene expression data, the method has previously been used to decompose data into "eigengenes" and "eigenexperiments," linear combinations of genes and experiments respectively [1,35]. However, it is generally recognized that SVD often produces components that are difficult to interpret [8,36-38]. As can be seen in Figure 4C, components produced by SVD [33] do not clearly correspond to the distinct gene sets or tissue populations in the synthetic data. For instance, the first component is to some extent a composition of gene sets one and three, and subsequent components then subtract off different combinations of the gene sets.

The development of non-negative matrix factorization (NMF) methods was in part driven by the aforementioned problems with SVD. NMF algorithms decompose a matrix into the product of non-negative matrices [38]. These algorithms generally produce more interpretable factors than does SVD, and have been successfully applied to various problems including gene expression analysis [8,36,37]. Figure 4D shows the application of an NMF-based algorithm to the synthetic data. We used a publicly available implementation [36], which searches for an optimal number of factors using a cophenetic clustering coefficient metric, and in this case found three factors to be optimal. As can be seen in the figure, NMF did fairly well at recovering genes sets one and two, although there was some overlap between the sets. However, gene sets three and four were indistinguishable.

Figure 4E demonstrates the application of a simplified version of GeneProgram in which tissue groups were not modeled (all tissues were attached to the root of the hierarchy). As can be seen, this version of the algorithm accurately recovered gene sets one and two. However, as with NMF, gene sets three and four completely overlapped.

Figure 4F shows the application of GeneProgram with full automatic learning of tissue groups enabled. As can be seen, the algorithm accurately recovered all four gene sets. By leveraging hierarchical structure in the data, the algorithm had additional information (the pattern of expression program use by related tissues), which presumably allowed it to correctly recover all the synthetic gene sets – something the other methods were not capable of.

GeneProgram outperformed biclustering algorithms in the discovery of biologically relevant gene sets in large compendia of mammalian gene expression data

Our objective was to apply GeneProgram to large compendia of mammalian gene expression data to compare our method's performance against that of other algorithms. In this regard, we used the Novartis Gene Atlas v2 [20], consisting of genome-wide Affymetrix expression measurements for 79 human and 61 mouse tissues, and a second data set collected by Shyamsundar *et al.*, consisting of cDNA expression measurements for 115 human tissues [19]. We compared GeneProgram's performance to two biclustering algorithms, Samba [5,34] and a non-negative matrix factorization (NMF) implementation [36]. We chose these two algorithms for comparison because they are popular in the gene expression analysis community, they have previously outperformed

other biclustering algorithms, and available implementations are capable of handling large data sets.

Because expression programs characterize both genes and tissues, we used both Gene Ontology (GO) categories [27] and manually derived, broadly physiologically based tissue categories to assess the algorithms' performance. However, GO categories and the manually derived tissue categories represent only limited biological knowledge. So, we were also interested in assessing the consistency of gene sets discovered by each algorithm across the two data sets. Because the two data sets used different microarray platforms and sources for tissues, similarities in discovered gene sets between data sets were likely to be biologically relevant. To perform this analysis for each algorithm, we used the gene sets discovered from one data set to compute the significance of the overlap with sets produced using the second data set. We then inverted the analysis and averaged the results to produce correspondence plots, which have previously been used for sensitive, graphical comparisons of biclustering algorithm performance [5]. See the Methods section for details.

GeneProgram clearly outperformed the other algorithms in both the tissue and gene dimensions on both data sets individually (Table 1), and also in terms of gene set consistency between the data sets (Figure 5). These results suggest several performance trends related to features of the different algorithms. As noted in the section on synthetic data experiments, Samba was successful at finding relatively small sets of genes that are co-expressed in subsets of tissues, but had difficulty uncovering larger structures in data. Presumably, our algorithm's clear dominance of both Samba and the NMF method was partly attributable to GeneProgram's hierarchical model. Both of the other algorithms lack such a model, so the assignment of tissues to biclusters was not guided by global relationships among tissues.

We note also that the algorithms differed substantially in runtimes: Samba was fastest (approximately 3 hours), GeneProgram the next fastest (approximately 3 days), and NMF the slowest (approximately 6 days), with all software running on a 3.2 GHz Intel Xenon CPU. Although these runtime differences may be attributable in part to implementation details, it is worth noting that GeneProgram, a fully Bayesian model using MCMC sampling for inference, ran faster than the NMF algorithm, which uses a more "traditional" objective maximization algorithm and searches for the appropriate number of biclusters.

GeneProgram discovered 5 tissue groups and 104 expression programs in human host-cell infection time-series data

In the previous two subsections, we used GeneProgram in a manner similar to traditional biclustering algorithms. In this subsection, we take advantage of GeneProgram's ability to find coherent gene sets that may be used with different (but consistent) temporal dynamics by each tissue sample – a capability not shared with traditional biclustering algorithms.

We applied GeneProgram to a compendium of sixty-two short time-series gene expression data sets exploring the responses of human cells to various infectious agents or immune-modulating molecules (see Tables S1 and S2). Behavior for each gene over each time-series experiment was mapped to one of six simple temporal patterns. A limited set of possible temporal patterns was intentionally chosen for two reasons. First, in the original studies, a primary feature of interest for all the experiments analyzed was the time of earliest induction or repression of each gene [21,23-25]. Thus, a small set of relevant temporal patterns aids in the biological interpretability of our results. Second, the time-series data sets analyzed had different durations, sampling rates and numbers of samples. By considering only simple temporal patterns that extract features present in all time-series, we could produce meaningful results spanning all the data sets. The six temporal patterns characterize the phase (early, middle, or late) of first induction or repression for each gene in each time-series experiment. See Figure S3 for a summary of the temporal patterns used, and Figure S4 for example genes with expression profiles corresponding to the patterns.

Figures S5-S7 (programs 1-75) and Figure 6 (programs 76-104) provide graphical summaries, and Table S8 provides complete details, for the tissue groups and expression programs discovered by GeneProgram from the infection data. The tissue groups essentially corresponded to the different host-cell types used for the infection experiments, although there was some intermingling of the dendritic and peripheral blood mononuclear cell experiments and those using other host-cell types. Expression programs were used by 2-27 experiments (median = 7) and contained 101-410 genes (median = 201). All six temporal patterns were used, although late induction and late repression were used least frequently, likely in part because the corresponding time intervals were the most sparsely sampled in the time-series analyzed. In many cases, patterns for a single program were uniformly inductive or repressive, although programs were sometimes used with differently phased patterns by different experiments. However, in other interesting cases, usage patterns were not uniformly inductive or repressive. Specific examples of expression programs with different temporal pattern usage are discussed below.

Discovered expression programs overlapped extensively with key human signaling pathways and biological processes

To evaluate the biological relevance of expression programs, we used two external sources of information about gene function: GO biological process categories [27] and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways [28]. Computation of enrichment scores, significance testing and correction for multiple hypotheses were done as described in the Methods Section.

A large number of expression programs were significantly enriched for GO categories (50%) or KEGG pathways (59%). As expected, many significant GO categories and KEGG pathways were specifically involved with response to infection. Interestingly, a substantial number of significant categories or pathways corresponded to signaling

cascades. Further, there were also a number of significantly enriched biological processes or pathways not directly labeled as being infection-related, but that are involved with alterations in cellular physiology consistent with infection. Finally, there were some unexpected significant categories or pathways. See Tables S9 and S10 for details. Specific examples of expression programs enriched for genes involved in key pathways and biological processes are discussed below.

A surprising number of expression programs contained significant numbers of genes bound by NF-κB transcription factor family members

We further evaluated the biological relevance of discovered expression programs using genome-wide data profiling static binding of NF-κB family transcription factors in untreated or lipopolysaccharide (LPS) stimulated human cell-culture derived macrophages [29]. Although the genome-wide binding data is static, we still expected programs discovered by GeneProgram from dynamic expression data to be enriched for genes bound by NF-κB family members, because these transcription factors are key controllers of mammalian immune and inflammatory responses [39]

Fifteen of the expression programs discovered by GeneProgram were significantly enriched for sets of genes bound by at least one of the NF-κB family members. This overlap with the static binding data is quite large, considering that the most genes bound in any one experiment (p53 in LPS stimulated cells) was 193 out of 9492 genes on the microarray [29]. Overall, the fifteen significantly enriched programs tended to be used by a diverse array of host-cell types exposed to a range of pathogens and their components. This suggests that these expression programs may represent common processes that are strongly induced or repressed during infection via NF-κB family member regulatory activity. Examples of such programs are discussed below.

The generality score naturally categorized programs into a spectrum of responses to infection

We developed a score for assessing the functional generality of expression programs, and demonstrated its utility for automatically characterizing the spectrum of discovered programs – from gene sets involved in response to a specific pathogen in one host-cell type, to those mediating common inflammatory pathways. The *generality score* is the entropy of the normalized distribution of usage of an expression program by all tissues in each tissue group. Because the distribution employed for calculating the score is normalized, tissue groups that only use an expression program a relatively small amount will have little effect on the score. Thus, a high generality score indicates that an expression program is used fairly evenly across many tissue groups; a low score indicates the program is used by tissues from a small number of groups. We note that the generality score requires a global organization of tissues into groups, rather than just the local associations of subsets of tissues with individual gene sets provided by biclustering algorithms. Because there is uncertainty in the number of and membership in tissue groups, GeneProgram's Dirichlet Process-based model provides a natural framework for

computing the generality score. See the Supplemental Methods section for complete details.

Programs with low generality scores were used by experiments spanning a limited number of host-cell and infection types

Experiments involving exposure of gastric epithelial cells to *H. pylori* used several low generality programs. This is biologically plausible, because gastric epithelial cells are not involved in principle immune system functions, unlike all other host cells profiled in the data set [22]. As an example of a program used exclusively by *H. pylori* infected epithelial cells, EP 22 (generality = 0.011, 5 experiments) was enriched for genes involved in regulation of the actin cytoskeleton (KEGG: hsa04810), and was used with a middle induction modifier by all associated experiments. The induction of this pathway is consistent with extensive host-cell shape changes known to occur in *H. pylori* infection; delayed induction likely reflects the time necessary for bacterial attachment and secretion of proteins that induce host-cell cytoskeletal rearrangements [22].

Peripheral blood mononuclear cell (PBMC) and whole blood experiments also used several low generality programs. This is biologically reasonable, because PBMCs and whole blood represent mixtures of innate and adaptive immunity mediating cell types, and thus contain cell types not profiled in the other experiments analyzed. Further, the diversity of cell types in PBMC and whole blood cultures allows for critical interactions that are necessary to trigger certain cellular responses [21]. As an example, EP 33 (generality = 0.027, 9 experiments) was used by experiments involving PBMCs and whole blood exposed to the Gram-negative bacteria *N. meningitides* or *B. pertussis*, and was enriched for genes with anti-apoptotic function (GO: 0006916). We hypothesize that this program, which was generally induced in the middle of the time-courses, may involve stabilization of an anti-apoptotic state necessary for maturation and differentiation of peripheral immune cells following infection.

Programs with intermediate generality scores were used by experiments involving host cells exposed to a wider variety of agents

Several intermediate generality programs appeared to represent coordinated down-regulation of proteolytic and antigen presentation pathways. For example, EP 68 (generality = 0.227, 14 experiments) was used by experiments involving exposure of primary macrophages to a variety of interleukins/interferons, pathogens or their components. This program was enriched for genes involved in proteasome function (KEGG: hsa03050), and was generally repressed early in the time-series. As another example, EP 77 (generality = 0.298, 6 experiments) was used by several experiments involving PBMCs or cell-culture derived macrophages exposed to different bacteria or immune modulating chemicals. This program was enriched for genes involved in both MHC I and MHC II antigen processing and presentation pathways (KEGG: hsa04612), and was used with a middle repression modifier by all associated experiments. Downregulation of proteasome and antigen presentation pathways subsequent to

infection may reflect commitment of phagocytic cells to presentation of antigens from a pathogen that has just been encountered [21].

Several expression programs revealed differences in temporal phasing of the response of host cells exposed to different classes of pathogenic organisms. For example, EP 88 (generality = 0.386, 13 experiments) was enriched for genes involved in ribosomal structure or function (KEGG: hsa03010). Induction of ribosomal genes may be a prelude to production of critical signaling and defensive proteins, such as chemokines and cytokines. EP 88 was induced early in macrophages or dendritic cells exposed to several varieties of Gram-negative bacteria, but induced in the middle of the time-series in host cells exposed to Gram-positive bacteria. As another example, EP 91 (generality = 0.402, 13 experiments), was enriched for genes involved in mRNA splicing (GO: 0000398) and oxidative phosphorylation (KEGG: hsa00190). This program was induced early in time-courses in dendritic cells exposed to bacteria or viruses, but not induced until the middle phase in experiments involving dendritic cells exposed to live fungi or fungal cell components. Interestingly, the program was repressed early in macrophages exposed to various interferons/interleukins. We hypothesize that the phasing differences observed in induction of EPs 88 and 91 may be due to the lesser ability of Gram-positive or fungal organisms to induce critical signaling pathways in innate immune cells. We also note that both programs were significantly enriched for genes bound by NF-κB family members. Interestingly, a number of the NF-κB targets in EP 88 were ribosomal genes, suggesting a direct role for this transcription factor in ribosome activity induction.

Several intermediate generality programs were significantly enriched for surprising signaling pathways or cell-host receptor types. For instance, EP 50 (generality = 0.134, 13 experiments) and EP 84 (generality = 0.331, 12 experiments) were significantly enriched for genes involved in the Wnt signaling pathway (KEGG: hsa04310), including WNT7A and FZD5 in EP 50, and WNT1, WNT5A and MMP7 in EP 84 (see Figure S11). Wnt signaling pathways have been traditionally implicated in developmental processes [40], and have only recently been shown to be involved in immune system functions [41-43]. For instance, Lobov *et al*. demonstrated that macrophages can secrete WNT7B, which induces apoptosis in vascular endothelial target cells via the canonical Wnt signaling pathway [42]. Signaling via the WNT5A receptor FZD5 has been implicated in stimulation of pro-inflammatory molecules (e.g., MMP7, TNF-α, IL-12) in macrophages, possibly via both canonical and non-canonical pathways [41,43]. Consistent with these reports of Wnt activity in macrophages, EP 50 was used by macrophages exposed to a variety of bacteria and stimulatory molecules. Interestingly, the program was repressed in macrophages infected with bacteria and induced in cells treated with interleukins or interferons. This difference in program usage may reflect the ability of bacteria to downregulate pro-inflammatory Wnt pathways. In contrast, EP 84, which was mostly used by macrophages and dendritic cells exposed to bacteria or microbial components, was uniformly induced in the middle of the infection time-series. Because the two programs contain different Wnt pathway genes, they may be involved in different inflammatory functions. Further, we note that only some of the Wnt pathway associated genes in EPs 50 and 84 have previously been implicated in macrophage function, making these attractive candidates for future experimental biology work.

EP 55 (generality = 0.161, 12 experiments), which was significantly enriched for genes coding for neurotransmitter or hormonal receptors (KEGG: hsa04080), was another surprising finding. This program was induced in macrophages treated with various interleukins or interferons, and repressed in macrophages exposed to various microbial components. The program contained genes coding for a variety of receptors including those for acetylcholine (CHRM5), cannabinoids (CNR2), dopamine (DRD2), histamine (HRH1) and somatostatin (SSR4). Although such receptor types are typically found on neurons, they have also been found on macrophages and T-cells, and recent studies suggest they may have important pro- and anti-inflammatory properties [44-47]. The different use of this program by macrophages exposed to interleukin/interferon (induction) versus that of those exposed to bacterial components (repression) may reflect bias toward pro- or anti-inflammatory states mediated by different neuroactive receptor signaling pathways.

Programs with high generality scores were used by the full spectrum of experiments involving exposure of different host cell types to a wide variety of agents

Most programs with high generality scores were enriched for genes bound by NF-κB family members and involved in a range of signaling pathways. For instance, EP 99 (generality = 0.585, 27 experiments) appeared to represent a "common infection response" program, which was used by experiments from every tissue group and almost always induced early. This EP contained 203 genes, 15% of which code for cytokines or cytokine receptor, and was also significantly enriched for a number of signaling pathways. Of particular interest, it contained many genes involved in the Toll-like receptor signaling pathway (KEGG: hsa04620, see Figure S12). As expected, many of the genes in the overlap between EP 99 and this pathway code for cytokines, but various downstream signaling molecules were also in the set, including TRAF6, AP-1, NF-κB (p105) and IκBα. Further, EP 99 was significantly enriched for genes bound by NF-κB family members RELB, p65, p50, p52 or c-REL in LPS stimulation experiments. In contrast, EP 102 (generality = 0.634, 15 experiments) was also used by a wide variety of experiments, but was induced in the middle of the time-series of all associated experiments, and was not significantly enriched for binding of any NF-κB family members. This program contained a large number of genes coding for interferons or interferon induced chemokines [18]. Interestingly, many interferon sensitive genes (ISGs) are activated by transcription factors other than NF-κB, and a delay in production of ISGs has previously been noted, presumably due to the time needed to establish autocrine and paracrine signaling loops [18].

Conclusions

We presented a new computational methodology, GeneProgram, specifically designed for analyzing large compendia of human expression data. In our first two applications, we used GeneProgram in a manner similar to traditional biclustering algorithms. First, we used synthetic data experiments to show that GeneProgram was able to correctly recover

gene sets that other popular analysis methods could not. Second, we analyzed two large compendia of mammalian gene expression data from representative normal tissue samples, and demonstrated that GeneProgram outperformed other biclustering methods in the discovery of biologically relevant gene sets. In our third application, we took advantage of GeneProgram's ability to find coherent gene sets used with different temporal dynamics by each tissue sample – a capability traditional biclustering algorithms do not have. We applied GeneProgram to a compendium of sixty-two short time-series gene expression data sets exploring the responses of human cells to various infectious agents or immune-modulating molecules. Using this data set, GeneProgram discovered 5 tissue groups and 104 expression programs, a substantial number of which were significantly enriched for genes involved in key signaling pathways and/or bound by NF-κB transcription factor family members in genome-wide experiments. We introduced an expression program generality score that exploits the tissue groupings automatically learned by GeneProgram, and showed that this score characterizes the functional spectrum of discovered expression programs – from gene sets involved in response to specific pathogens in one host cell type, to those mediating common inflammatory pathways.

We provided many examples of discovered expression programs involved in key pathways related to the response to infection. Interestingly, some programs were used with different temporal patterns by certain types of experiments, such as early induction of programs enriched for genes involved in ribosomal function or energy production in host cells exposed to Gram-negative organisms versus later induction of the same programs in host cells exposed to Gram-positive organisms or fungi. Some of the discovered programs overlapped considerably with previously described gene sets derived from the same expression data sets, such as EP 99 and the "common host response" genes discussed by Jenner and Young [18]. However, previous meta-analyses of these data sets relied on extensive prior biological knowledge and manual inspection of the data [18]. In contrast, our method was automatic, discovering expression programs, associating them with consistent temporal patterns, and finding significantly overlapping biological pathways and NF-κB binding.

Some of the gene sets discovered by GeneProgram implicate surprising signaling pathways or host-cell receptor types in the response to infection. In particular, EPs 50 and 82 were significantly enriched for genes involved in the Wnt signaling pathway, and EP 55 was significantly enriched for genes coding for neurotransmitter or hormonal receptors. Wnt signaling pathways [41-43] and neurotransmitter receptors [44-47] have only recently been implicated in the response to infection. To our knowledge, our work is the first to uncover the activity of these pathways in the data sets analyzed, and to characterize the different temporal behaviors of these pathways in response to a variety of infectious agents and immunomodulatory molecules. We believe that the genes in the above-mentioned expression programs constitute particularly attractive candidates for further biological characterization

GeneProgram encodes certain assumptions that differ from some previous methods for analyzing expression data and so merit further discussion. First, we model expression

data in a semi-quantitative fashion, assuming that discrete levels of mRNA correspond to biologically interpretable expression differences. We believe this is appropriate because popular array technologies can only reliably measure semi-quantitative, relative changes in expression; many relevant consequences of gene expression are threshold phenomena [48-50]; and it is difficult to assign a clear biological interpretation to a full spectrum of continuous expression levels. Second, GeneProgram assumes that discrete "units" of mRNA are independently allocated to expression programs, which captures the phenomena that mRNA transcribed from the same gene may participate in different biological processes throughout a cell or tissue. Independence of mRNA units is an unrealistic assumption, but this approximation, which is important for efficient inference, has worked well in practice for many other applications of topic models [11,12,31]. Finally, GeneProgram handles temporal data by collapsing time-series into pre-defined, discrete patterns. Overall, we believe that this is a very useful approach for finding interpretable gene expression programs, particularly when analyzing short time-series experiments, in which there are a limited number of clearly meaningful temporal patterns. Further, this method allows us to extract features present in a compendium of time-series, even when the series have different durations, sampling rates, and numbers of samples. In the case of the infection time-series data analyzed in this work, we defined relevant temporal patterns manually, based on prior biological knowledge [21,23-25]. However, temporal patterns could be derived in more automated ways, such as through pre-processing steps that apply time-series clustering algorithms [15,51] to individual series in the data compendium.

Our method produced a large map of human infection response expression programs with several distinguishing features. First, by simultaneously using information across a variety of host cell types exposed to a wide range of pathogens and immunomodulatory molecules, GeneProgram was able to finely dissect the data, automatically splitting genes differentially expressed in response to infection among both general and specific programs. Second, because our model explicitly operates on probabilistically ranked gene sets throughout the entire inference process, rather than finding individual differentially expressed genes, our results are more robust to noise. Third, the fact that expression programs provide probabilistically ranked sets of genes also provides a logical means for prioritizing directed biological experiments. Fourth, because our model is fully Bayesian, providing a global penalty for model complexity including for the number of tissue groups and expression programs, the generated map represents a mathematically principled compression of gene expression information throughout the experiments. Finally, although such a large, comprehensive map is inherently complicated, we believe that GeneProgram's automatic grouping of tissues and the associated expression program generality score aid greatly in its interpretation.

GeneProgam is a general method suitable for analyzing many large expression data compendium, and we believe that the maps of expression programs discovered by our algorithm will be particularly useful for guiding future biological experiments. Highly specific expression programs can provide candidate genes for diagnostic markers or drug targets that exhibit minimal unintended "cross-talk." General expression programs may be useful for identifying genes important in regulating and maintaining general biological

responses, which may go awry in disease states such as inflammation or malignancy. Additionally, our framework is flexible, and could accommodate other genome-wide sources of biological data in future work, such as DNA-protein binding or DNA sequence motif information. GeneProgram's ability to discover tissue groups and coherent expression programs *de novo* using a principled probabilistic method, as well as its use of data in a semi-quantitative manner, makes it especially valuable for novel "meta-analysis" applications involving large data sets of unknown complexity in which direct fully quantitative comparisons are difficult.

Methods

Data sets and pre-processing

For our benchmark data sets, we used two data sources. The first source was data from Shyamsundar *et al*. [19], consisting of 115 human tissue samples obtained from surgeries or autopsies, with expression measured on custom cDNA microarrays. The reference channel on the microarrays consisted of mRNA pooled from 11 established human cell lines. We considered a gene expressed if its ratio was greater than 2.0. Our second source of benchmark data was the Novartis Gene Atlas v2, consisting of 79 human and 61 mouse tissue samples, with expression measured on Affymetrix microarrays [20]. We retained pairs of related genes from mouse and human samples using Homologene (build 47) mappings [52]. Arrays were pre-processed using the GC content-adjusted robust multi-array algorithm (GC-RMA) [53]. To correct for probe specific intensity differences, the intensity of each probe was normalized by dividing by its geometric mean in the 31 matched tissues. For genes represented by more than one probe, we used the maximum of the normalized intensities. A gene was considered expressed if its normalized level was greater than 2.0. For both the Novartis Gene Atlas v2 and Shyamsundar *et al.* data, we mapped genes to common identifiers using the IDConverter software [54], and retained only the 7,404 genes present in both data sets.

For the infectious disease time-series analysis, we used expression data from six microarray studies [21-26] that had previously been combined and standardized by Jenner and Young [18]. The data used consisted of 347 microarray experiments measuring expression of 5042 genes in 62 short time-series (see Tables S1 and S2 for summaries of experiments). Behavior for each gene over each time-series experiment was mapped to one of six simple temporal patterns (see Figure S3). Time-points for all experiments were divided into three general phases: early (less than two hours), middle (greater than two hours and not more than twelve hours) and late (greater than twelve hours). For each gene in each time-series experiment, the gene was assigned to the pattern corresponding to the earliest phase in which the gene's expression value exceeded a two-fold increase (decrease) in at least one experiment in the respective time interval. Further, to be assigned to a pattern, a gene's expression across the time-series was required to be consistent in direction (either a two-fold increase or decrease in expression but not both). If a gene's expression profile did not meet all these criteria, it was not assigned to any pattern. The absolute expression value for a gene's earliest induction

(repression) in each time-series was then used to represent the magnitude of differential expression.

Expression data was discretized into three levels using a mutual information-based greedy agglomerative merging algorithm, as described in Hartemink *et al.* [55]. In brief, continuous expression levels were first uniformly discretized into a large number of levels. The algorithm then repeatedly found the best two adjacent levels to merge by minimizing the reduction in the pair-wise mutual information between all expression vectors. The appropriate number of levels to stop at was determined by choosing the inflection point on the curve obtained by plotting the score against the number of levels. We obtained three levels for all the data sets. To analyze the sensitivity of our results to the number of discretization levels used, we created correspondence plots for the largest data set (the Novartis Gene Atlas v2) using two, three, and four discretization levels (see Figure S13). As can be seen, at all the discretization levels examined, GeneProgram outperformed the other biclustering methods, and fluctuations across results at different numbers of discretization levels were considerationably smaller than differences between GeneProgram and the other algorithms.

To validate expression programs discovered in the infectious disease data, we used static genome-wide data profiling binding of five transcription factors in the NF-κB family in untreated or lipopolysaccharide (LPS) stimulated cell-culture derived human macrophages [29]. The experimenters obtained binding data after one hour of LPS stimulation, and 9492 genes were arrayed in the ChIP-chip analysis. We used the same criteria for determining binding events as in the original study (a *p*-value threshold of 0.002).

Probability model

The GeneProgram model is an extension of the Hierarchical Dirichlet Process mixture model as described in Teh *et al* [13]. In Hierarchical Dirichlet Processes, dependencies are specified among a set of Dirichlet Processes by arranging them in a tree structure. At each level in the tree, the base distribution for the Dirichlet Process is a sample from the parent Dirichlet Process above. In GeneProgram, each expression program corresponds to a mixture component in the Hierarchical Dirichlet Process model. Because the model is hierarchical, expression programs are shared by all Dirichlet Processes in the model. An expression program specifies a multinomial distribution over genes, and a set of usage modifier variables (one for each tissue). Discrete expression levels are treated analogously to word occurrences in document-topic models. Thus, a tissue's vector of gene expression levels is converted into a collection of expression events, in which the number of events for a given gene equals the discrete expression level of that gene in the tissue. A pattern type (e.g., induction or repression) is associated with each expression event. The model assumes that each gene expression event in a tissue is independently generated by an expression program. One usage modifier variable is associated with each tissue for each program, and thus specifies the possible pattern type for genes from the program used by the tissue. Usage variables are sampled hierarchically, and thus influenced by other tissues. In the original Hierarchical Dirichlet Process formulation

[13], the entire tree structure was assumed to be pre-specified. We extend this work, by allowing the model to learn the number of groups and the memberships of tissues in these groups. Thus, groups themselves are generated by a Dirichlet Process, which uses samples from the root of the Hierarchical Dirichlet Process as the base distribution. See the Supplemental Methods for complete details.

Model inference and summarization

The posterior distribution for the model is approximated via Markov Chain Monte Carlo (MCMC) sampling [30]. Our inference method is based on the technique described for efficient MCMC sampling in Hierarchical Dirichlet Processes [13], with additional steps added for sampling the tissue group assignments and usage variables. See the Supplemental Methods section for complete details.

The final step of the GeneProgram algorithm summarizes the approximated model posterior probability distribution with *consensus tissue groups* and *recurrent expression programs*. The posterior distributions of Dirichlet Process mixture models are particularly challenging to summarize because the number of mixture components may differ for each sample. Previous approaches for summarizing Dirichlet Process mixture model components have used pair-wise co-clustering probabilities as a similarity measure for input into an agglomerative clustering algorithm [14]. This method is feasible if there are a relatively small number of items to be clustered, and we employ it for producing consensus tissue groups. Consensus tissue groups are constructed by first computing the empirical probability that a pair of tissues will be assigned to the same tissue group. The empirical co-grouping probabilities are then used as pair-wise similarity measures in a standard bottom-up agglomerative hierarchical clustering algorithm using complete linkage [56]. See the Supplemental Methods section for complete details. However, this method is not feasible for summarizing expression programs in large data sets because of the number of pair-wise probabilities that would need to be calculated for each sample.

We developed a novel method for summarization of the model posterior distribution, which discovers recurrent expression programs by combining information from similar expression programs that reoccur across multiple MCMC model posterior samples. Each expression program in each sample is compared against recurrent programs derived from previous samples. If the Hellinger distance between a program's distribution of gene occurrences and that of a recurrent program is sufficiently small, the programs are merged; otherwise a new recurrent program is instantiated. Statistics are tracked for each recurrent expression program, including the number of posterior samples it occurs in, its average usage by tissues, and average expression levels of genes in the program. After all recurrent expression programs have been collected, infrequently occurring programs are filtered out. See the Supplemental Methods section for complete details.

Synthetic data

We generated four synthetic gene sets used by 40 tissue samples divided equally among four tissue populations. Each gene set contained 40 genes with varying mRNA levels;

gene sets three and four overlapped in 10 genes. Each tissues population specifies the mean number of genes from each gene set used by tissue samples in the population. For each tissue sample, the number of genes used from each gene set was sampled from the population mean, and then genes were picked from each gene set with a probability weighted by the gene's underlying mRNA level. Finally, the mRNA level for each gene was perturbed by additive and multiplicative noise to simulate microarray noise. See the Supplemental Methods section for complete details.

Tissue and gene dimension category enrichment analysis and correspondence plots

We manually classified tissues from the Novartis Gene Atlas v2 and Shyamsundar *et al*. data sets into ten and eight high-level, physiologically based categories respectively (see Tables S14 and S15).

We mapped genes to Gene Ontology (GO) biological process categories [27] and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways [28] using the IDConverter software [54]. For calculating enrichment scores, we considered human GO categories or KEGG pathways containing between 5 and 200 genes.

For determining enrichment significance of GO categories, KEGG pathways, or NF-κB family transcription factor binding, we used the hypergeometric distribution to compute *p*-values. We corrected for multiple hypothesis tests using the procedure of Benjamini and Hochberg [57]. We used a false-discovery rate cut-off of 0.05.

Correspondence plots depict log *p*-values on the horizontal axis and the fraction of biclusters with *p*-values below a given value on the vertical axis. Depicted *p*-values are from the most abundant class for each bicluster (i.e., that with the largest number of genes or tissue in the overlap) and calculated using the hypergeometric distribution. Note that biclusters with large *p*-values are not significantly enriched for any class, and may represent noise.

Software availability

Complete Java source code is available upon request.

Acknowledgements

We thank Timothy Danford and Kenzie MacIssac for help with earlier versions of this work. GKG and RDD were supported by NIH Grant 2R01 HG002668-04A1.

Supplemental Materials

Table S1: Summary of data sources used for the infection time-series compendium.

Table S2: Summary of individual experiments in the infection time-series data compendium.

Figure S3: Schematic of six defined temporal patterns used in analyzing the infection time-series data.

Figure S4: Examples of genes from the infection time-series data that match the six defined temporal patterns.

Figure S5: Expression programs 1-25 discovered in the infection time-series data compendium (see Figure 6 for figure legends).

Figure S6: Expression programs 26-50 discovered in the infection time-series data compendium (see Figure 6 for figure legends).

Figure S7: Expression programs 51-75 discovered in the infection time-series data compendium (see Figure 6 for figure legends).

Table S8: Details for all 104 expression programs discovered in the infection time-series data.

Table S9: Summary of selected GO biological process categories and KEGG pathways for which expression programs discovered in the infection time-series data were significantly enriched.

Table S10: Complete list of GO biological process categories and KEGG pathways for which expression programs discovered in the infection time-series data were significantly enriched.

Figure S11: Overview of the Wnt signaling pathways and relevant genes in EP#50 (pink shading) and EP#82 (green shading); adapted from the KEGG pathway graphic.

Figure S12: Overview of the Toll-like receptor signaling pathways and relevant genes in EP#99 (pink shading); adapted from the KEGG pathway graphic.

Figure S13: Correspondence plots for gene and tissue dimensions from analysis of the Novartis Gene Atlas v2 using Samba, non-negative matrix factorization (NMF), and two, three, and four discretization levels in GeneProgram.

Table S14: Manual classification of Novartis Gene Atlas v2 tissue samples into 10 high-level, physiologically based categories.

Table S15: Manual classification of tissue samples from the Shyamsundar *et al*. data set into 8 high-level, physiologically based categories.

1. Alter O, Brown PO, Botstein D (2000) Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci U S A 97: 10101-10106.
2. Cheng Y, Church GM. Biclustering of Expression Data; 2000. AAAI Press. pp. 93-103.
3. Sheng Q, Moreau Y, De Moor B (2003) Biclustering microarray data by Gibbs sampling. Bioinformatics 19 Suppl 2: II196-II205.
4. Segal E, Friedman N, Koller D, Regev A (2004) A module map showing conditional activity of expression modules in cancer. Nat Genet 36: 1090-1098.
5. Tanay A, Sharan R, Shamir R (2002) Discovering statistically significant biclusters in gene expression data. Bioinformatics 18 Suppl 1: S136-144.
6. Battle A, Segal E, Koller D (2005) Probabilistic discovery of overlapping cellular processes and their regulation. J Comput Biol 12: 909-927.
7. Dueck D, Morris QD, Frey BJ (2005) Multi-way clustering of microarray data using probabilistic sparse matrix factorization. Bioinformatics 21 Suppl 1: i144-151.
8. Carmona-Saez P, Pascual-Marqui RD, Tirado F, Carazo JM, Pascual-Montano A (2006) Biclustering of gene expression data by Non-smooth Non-negative Matrix Factorization. BMC Bioinformatics 7: 78.
9. Madeira SC, Oliveira AL (2004) Biclustering Algorithms for Biological Data Analysis: A Survey. IEEE Transactions on Computational Biology and Bioinformatics 1: 24-45.
10. Tanay A, Sharan R, Shamir R (2005) Biclustering Algorithms: A Survey. In: Aluru S, editor. Handbook of Computational Molecular Biology: Chapman & Hall/CRC.
11. Flaherty P, Giaever G, Kumm J, Jordan MI, Arkin AP (2005) A latent variable model for chemogenomic profiling. Bioinformatics 21: 3286-3293.
12. Griffiths TL, Steyvers M (2004) Finding scientific topics. Proc Natl Acad Sci U S A 101 Suppl 1: 5228-5235.
13. Teh YW, Jordan MI, Beal MJ, Blei DM (2006) Hierarchical Dirichlet Processes. Journal of the American Statistical Association.
14. Medvedovic M, Yeung KY, Bumgarner RE (2004) Bayesian mixture model based clustering of replicated microarray data. Bioinformatics 20: 1222-1232.
15. Bar-Joseph Z (2004) Analyzing time series gene expression data. Bioinformatics 20: 2493-2503.
16. Aach J, Church GM (2001) Aligning gene expression time series with time warping algorithms. Bioinformatics 17: 495-508.
17. Bar-Joseph Z, Gerber GK, Gifford DK, Jaakkola TS, Simon I (2003) Continuous representations of time-series gene expression data. J Comput Biol 10: 341-356.
18. Jenner RG, Young RA (2005) Insights into host responses against pathogens from transcriptional profiling. Nat Rev Microbiol 3: 281-294.
19. Shyamsundar R, Kim YH, Higgins JP, Montgomery K, Jorden M, et al. (2005) A DNA microarray survey of gene expression in normal human tissues. Genome Biol 6: R22.
20. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A 101: 6062-6067.

21. Boldrick JC, Alizadeh AA, Diehn M, Dudoit S, Liu CL, et al. (2002) Stereotyped and specific gene expression programs in human innate immune responses to bacteria. Proc Natl Acad Sci U S A 99: 972-977.
22. Guillemin K, Salama NR, Tompkins LS, Falkow S (2002) Cag pathogenicity island-specific responses of gastric epithelial cells to Helicobacter pylori infection. Proc Natl Acad Sci U S A 99: 15136-15141.
23. Huang Q, Liu D, Majewski P, Schulte LC, Korn JM, et al. (2001) The plasticity of dendritic cell responses to pathogens and their components. Science 294: 870-875.
24. Nau GJ, Richmond JF, Schlesinger A, Jennings EG, Lander ES, et al. (2002) Human macrophage activation programs induced by bacterial pathogens. Proc Natl Acad Sci U S A 99: 1503-1508.
25. Nau GJ, Schlesinger A, Richmond JF, Young RA (2003) Cumulative Toll-like receptor activation in human macrophages treated with whole bacteria. J Immunol 170: 5203-5209.
26. Pathan N, Hemingway CA, Alizadeh AA, Stephens AC, Boldrick JC, et al. (2004) Role of interleukin 6 in myocardial dysfunction of meningococcal septic shock. Lancet 363: 203-209.
27. (2006) The Gene Ontology (GO) project in 2006. Nucleic Acids Res 34: D322-326.
28. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, et al. (2006) From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 34: D354-357.
29. Schreiber J, Jenner RG, Murray HL, Gerber GK, Gifford DK, et al. (2006) Coordinated binding of NF-kappaB family members in the response of human cells to lipopolysaccharide. Proc Natl Acad Sci U S A 103: 5899-5904.
30. Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian Data Analysis: Chapman & Hall.
31. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet Allocation. Journal of Machine Learning Research 3: 993-1022.
32. Sudderth EB, Torralba A, Freeman WT, Wilsky AS. Learning Hierarchical Models of Scences, Objects, and Parts; 2005. pp. 1331-1338.
33. Matlab. R14 ed. Natick: The MathWorks.
34. Shamir R, Maron-Katz A, Tanay A, Linhart C, Steinfeld I, et al. (2005) EXPANDER--an integrative program suite for microarray data analysis. BMC Bioinformatics 6: 232.
35. Alter O, Brown PO, Botstein D (2003) Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. Proc Natl Acad Sci U S A 100: 3351-3356.
36. Brunet JP, Tamayo P, Golub TR, Mesirov JP (2004) Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci U S A 101: 4164-4169.
37. Kim PM, Tidor B (2003) Subsystem identification through dimensionality reduction of large-scale gene expression data. Genome Res 13: 1706-1718.
38. Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401: 788-791.

39. Chen LF, Greene WC (2004) Shaping the nuclear action of NF-kappaB. Nat Rev Mol Cell Biol 5: 392-401.
40. Huelsken J, Birchmeier W (2001) New aspects of Wnt signaling pathways in higher vertebrates. Curr Opin Genet Dev 11: 547-553.
41. Blumenthal A, Ehlers S, Lauber J, Buer J, Lange C, et al. (2006) The Wingless homolog WNT5A and its receptor Frizzled-5 regulate inflammatory responses of human mononuclear cells induced by microbial stimulation. Blood 108: 965-973.
42. Lobov IB, Rao S, Carroll TJ, Vallance JE, Ito M, et al. (2005) WNT7b mediates macrophage-induced programmed cell death in patterning of the vasculature. Nature 437: 417-421.
43. Pukrop T, Klemm F, Hagemann T, Gradl D, Schulz M, et al. (2006) Wnt 5a signaling is critical for macrophage-induced invasion of breast cancer cell lines. Proc Natl Acad Sci U S A 103: 5454-5459.
44. Galvis G, Lips KS, Kummer W (2006) Expression of nicotinic acetylcholine receptors on murine alveolar macrophages. J Mol Neurosci 30: 107-108.
45. Ofek O, Karsak M, Leclerc N, Fogel M, Frenkel B, et al. (2006) Peripheral cannabinoid receptor, CB2, regulates bone mass. Proc Natl Acad Sci U S A 103: 696-701.
46. Tan KS, Nackley AG, Satterfield K, Maixner W, Diatchenko L, et al. (2007) Beta2 adrenergic receptor activation stimulates pro-inflammatory cytokine production in macrophages via PKA- and NF-kappaB-independent mechanisms. Cell Signal 19: 251-260.
47. Triggiani M, Petraroli A, Loffredo S, Frattini A, Granata F, et al. (2007) Differentiation of monocytes into macrophages induces the upregulation of histamine H1 receptor. J Allergy Clin Immunol 119: 472-481.
48. Hume DA (2000) Probability in transcriptional regulation and its implications for leukocyte differentiation and inducible gene expression. Blood 96: 2323-2328.
49. Little JW (2005) Threshold effects in gene regulation: when some is not enough. Proc Natl Acad Sci U S A 102: 5310-5311.
50. Zhang Q, Andersen ME, Conolly RB (2006) Binary gene induction and protein expression in individual cells. Theor Biol Med Model 3: 18.
51. Ernst J, Bar-Joseph Z (2006) STEM: a tool for the analysis of short time series gene expression data. BMC Bioinformatics 7: 191.
52. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, et al. (2006) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 34: D173-180.
53. Wu Z, Irizarry RA (2005) Stochastic models inspired by hybridization theory for short oligonucleotide arrays. J Comput Biol 12: 882-893.
54. Alibes A, Yankilevich P, Canada A, Diaz-Uriarte R (2007) IDconverter and IDClight: conversion and annotation of gene and protein IDs. BMC Bioinformatics 8: 9.
55. Hartemink AJ, Gifford DK, Jaakkola TS, Young RA (2001) Using Graphical Models and Genomic Expression Data to Statistically Validate Models of Genetic Regulatory Networks. Pacific Symposium on Biocomputing Hawaii.
56. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95: 14863-14868.

57. Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society, Series B (Methodological) 57: 289-300.