Supplementary Note

Construction of Blood Transcription Modules

Introduction

Blood transcriptomes have always been a focus of systems biology and personalized medicine (Golub et al., 1999; Auffray et al., 2009). As such data emerge in human immunology (Chaussabel et al., 2010; Pulendran et al., 2010), their analysis starts to reveal several limitations of canonical pathways: a) pathways are not specific to the context of human blood tissue; b) lower sensitivity due to that not all genes in a pathway are regulated at transcriptional level or synchronized; c) biased toward oncology, limited coverage of immunology; d) immune response in healthy people differs from pathological observations; e) intercellular communication often plays a key role in immunological activities.

There is thus a strong motivation to learn immune response from existing data, and build new tools for analyzing blood transcriptomes. Chaussabel et al (2008) employed a K-means based clustering approach to find coexpressed gene modules, based on a collection of in-house data cross multiple diseases at Baylor Institute for Immunology Research. These modules form the basis of statistical testing between patient groups, and the significant modules project the biological meaning of the data. This approach has been applied to various datasets including systemic lupus erythematosus and turberculosis (Chaussabel et al., 2008; Berry et al., 2010; Quartier et al., 2011; Tattermusch et al., 2012).

Regulation of gene expression is condition dependent. A gene module is expected not to be present under all conditions, but to be under a subset of the conditions. This is the conceptual basis of a class of biclustering methods, in which Chaussabel's method was a special instance. Many biclustering methods have been applied to gene expression data (Cheng and Church, 2000; Ben-Dor et al., 2002; Tanay et al., 2004; Ihmels et al., 2004; Prelic et al., 2006). Most of these biclustering methods work on the variations in a 2-dimensional matrix (gene by samples), thus the quality of this input matrix is critical to the resulted modules. Tanay et al (2002, 2004) converted the expression data into a network structure, and looked for dense subnetworks as modules. The advantage of the network approach was not apparent in benchmark studies (Prelic et al., 2006), but it is far more scalable because it is impractical to standardize heterogeneous studies in the same input expression matrix.

Segal et al (2003, 2005a) took a probabilistic graph approach in finding transcriptional modules. A set of candidate regulators were predefined, and the likelihood of gene regulation is assessed by posterior bayesian method. It should be noted that this was not the module approach in their work on “cancer module map” (Segal et al., 2004, 2005b). The latter was a clustering approach using priorly curated gene sets. In fact, most above studies on biclustering and transcriptional modules were using the yeast model organism. Multicellular organism data, especially human clinical data, pose a different complexity to computational inference.

In the context of human blood transcriptomes, multiple cell types are involved. A concerted immune response commands active intercellular communications and a kinetic that lasts from hours to days and weeks. Therefore, our blood transcription modules are not limited to a set of direct transcriptional targets, but coexpressed genes as a result of a particular biological activity (Figure B1). The method has to be scalable to fully leverage the large amount of public data. We thus formulate the problem as reconstructing a high-quality gene network and finding topological modules from the network.

Overview of approach

The construction of blood transcription modules is illustrated in Figure 3 in the main text.

1. First, we reconstruct a high-quality coexpression network from a large compendium of human blood transcriptomes. The integration across studies is as easy as combining network edges from each individual study.
2. From the master network, we generate 77 subnetwork that are specific to biological contexts (using gene ontology, cell type specific gene expression, interactome and bibliome).
3. Next, modules are extracted from the coexpression networks as genes with significantly dense connections. This is intuitive because more connections in a coexpression network means a higher chance of coexpression. Data from pathway databases and transcription factor targets are integrated in this step.
4. Post-processing removes redundant modules, filters for larger modules of denser connections and adjusts for gene over-representation. The final modules are evaluated and annotated.

Compendium of human blood transcriptomes

We retrieved from NCBI Gene Expression Omnibus (GEO) all human blood transcriptomes. The search was constructed as:

(((human[Organism]) AND "rna"[Sample Type]) AND blood[MeSH Terms]) AND "Homo sapiens"[porgn:__txid9606]

This yielded 1282 datasets from GEO (November 14, 2011). They were filtered for major microarray platfroms by Affymetrix and Illumina, and only datasets of ten or more samples were retained. Some samples were present in multiple datasets (due to multiple publications, reanalysis, etc.), and datasets containing significantly redundant samples were removed. All our vaccine datasets, plus a few public test datasets, were set aside, not used in module construction. This resulted in 540 data series and 32766 samples. A few very large cohorts were split into multiple data series in GEO. Each of these data series was still very large (e.g. > 500 samples) and treated as an independent study.

We used "Series Matrix Files" from GEO as our input data, as our approach does not attempt to normalize data cross studies and the normalization by the original authors should suffice. Any further irregularity in the data will be ruled out in a later step, where 30,000 permutations were performed on each dataset. In a large collection of public data, it

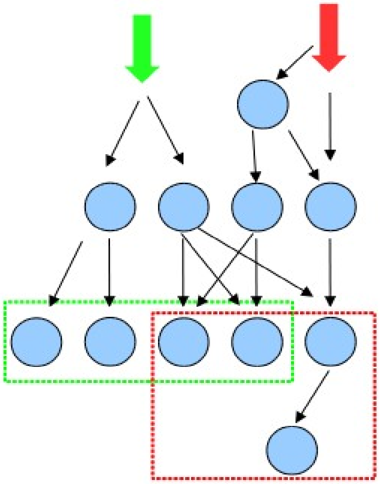


Figure B1: Concept of blood transcription modules. Each circle represents a gene. The genes in the green box respond to the green stimulus, while the genes in red box respond to the red stimulus. The module construction is to recover the green and red boxes. Because an immune response usually involves multiple cell types and lasts many cell cycles, a module does not have to be the targets of the same transcription factor, nor do they have to be a direct response.

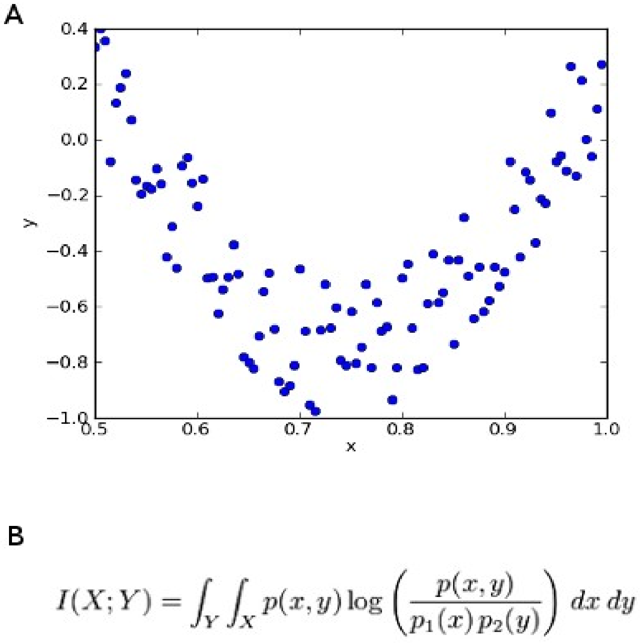
is inevitable to find misannotations and varying data quality. Therefore, on top of significance based on permutations in a single study, an edge in our reconstructed gene network required significance in at least three studies.

Reverse engineering of high quality gene network

Multiple probesets were often present for the same gene, the data were thus collapsed into gene level by using the probeset of the highest expression values in each study. Within a study, coefficient of variation for each gene was calculated and only genes with variation above the mean value were kept for the next step, in order to reduce the computational burden. We took a well established methodology of mutual information in reverse engineering a high quality gene network from this blood transcriptome compendium.

Mutual information is derived from Claude Shannon's Information theory. The statistical dependency between two variables can be identified from their probabilistic distributions, even when linear correlations fail to capture that (Figure B2). Mutual information has been long applied to the reconstruction of gene network (Basso et al., 2005; Margolin et al., 2006; Lefebvre et al., 2010, 2012). In this study, we used an algorithm by Kraskov et al (2004), implemented by Sales and Romualdi (2011) in a parallel computing package of R.

Within each study in the blood transcriptome compendium, mutual information (MI) of each gene pair was computed, and the significance of MI value is estimated over 30,000 permutations. If a gene pair yields p-value < 0.001, an edge between these two genes is reported. After significant edges were collected for all studies, the edges that appeared in three or more studies are selected to build the master reference network. This master network consists of 17,397 genes and 604,363 edges.



$$I(X;Y) = \int_Y \int_X p(x,y) \log\left(\frac{p(x,y)}{p_1(x)\,p_2(y)}\right) dx\, dy$$

Figure B2. Mutual information.
A) Relationship between X and Y is clear and can be captured by mutual information. But Pearson correlation between X and Y is zero. B) Mathematical definition of mutual information when X and Y are continuous variables, computed on their joint and separate distributions.

The quality of this master network can be assessed by comparison to known protein complexes, which tend to have concerted expression of their subunit genes. From the 920 protein complexes reported in the CORUM database (Ruepp et al., 2010), 282 are found in our master network with at least one edge. This number is very encouraging, considering that our master network is specific to human blood tissue while the CORUM protein complexes are for all mammalian tissues. Many protein complexes are found in the master network with dense connections, e.g. 2,418 edges between the 79 ribosome genes (Suppl Fig 6.). Interacting proteins are more context dependent and less likely to coexpress, yet we found 1,047 protein interacting pairs in our master network, out of the total 39,174 interacting pairs reported in HPRD database (Keshava et al., 2009).

Due to their excessive connections, ribosomal genes were removed from the master network in subsequent steps. We next generated 77 context-specific subnetworks from the master network, using gene ontology, cell type specificity, interactome and bibliome as biological contexts. Each context is defined by a list of genes, and a subnetwork is built from these genes with respective edges from the master reference network.

The nature of mutual information is statistical dependency between two vectors, which may appear as a positive correlation or a negative correlation. We compared MI with Pearson correlation, and the number of negative correlation among all network edges was estimated to be under 0.1%. The nature of this type of coexpression analysis strongly favors positive correlations, because the transcriptional repression of a gene is hard to observe amongst the default state of most genes being off. An early coexpression network study using Pearson correlation (Lee et al., 2004) reported that 88.8% confirmed links were positively correlated (when confirmed in 3+ studies). Our result here has a yet higher percentage, possibly due to the use of a single type of tissue and more stringent procedures.

Given the rarity of negative correlations, we did not attempt to remove them from the networks. When an activity score is averaged over all member genes of a module, the presence of rare negatively correlated member will not significantly affect the module performance. On the other hand, it can be informative to the investigators if a negatively correlated member is kept in the module, should further research is desired. A clear example is module M240 of chromosome Y linked genes (Suppl Fig 7), where "opposing" members, including XIST, TSIX and KAL1, are present. We believe more exciting biology like this can be discovered from our results.

Gene ontology categories as specific contexts

The following gene ontology categories were used:

GO:0002682 : regulation of immune system process [1267 gene products]
GO:0032879 : regulation of localization [1712 gene products]
GO:0040012 : regulation of locomotion [582 gene products]
GO:0050776 : regulation of immune response [876 gene products]
GO:0001775 : cell activation [1053 gene products]
GO:0007155 : cell adhesion [1545 gene products]
GO:0007267 : cell-cell signaling [1401 gene products]
GO:0007166 : cell surface receptor signaling pathway [3394 gene products]
GO:0009966 : regulation of signal transduction [2590 gene products]
GO:0002764 : immune response-regulating signaling pathway [332 gene products]
GO:0030522 : intracellular receptor mediated signaling pathway [261 gene products]
GO:0007049 : cell cycle [1856 gene products]
GO:0022402 : cell cycle process [1297 gene products]
GO:0008219 : cell death [2367 gene products]
GO:0051301 : cell division [544 gene products]
GO:0016049 : cell growth [504 gene products]
GO:0034330 : cell junction organization [276 gene products]
GO:0008037 : cell recognition [102 gene products]
GO:0006928 : cellular component movement [1706 gene products]
GO:0048468 : cell development [2157 gene products]
GO:0030154 : cell differentiation [3897 gene products]
GO:0045165 : cell fate commitment [349 gene products]
GO:0001709 : cell fate determination [64 gene products]
GO:0001708 : cell fate specification [123 gene products]
GO:0048469 : cell maturation [164 gene products]
GO:0019725 : cellular homeostasis [1054 gene products]
GO:0051641 : cellular localization [3022 gene products]

GO:0016044 : cellular membrane organization [530 gene products]
GO:0055085 : transmembrane transport [1114 gene products]
GO:0002253 : activation of immune response [517 gene products]
GO:0019882 : antigen processing and presentation [551 gene products]
GO:0002252 : immune effector process [675 gene products]
GO:0006955 : immune response [1873 gene products]
GO:0002520 : immune system development [836 gene products]
GO:0045321 : leukocyte activation [816 gene products]
GO:0001776 : leukocyte homeostasis [86 gene products]
GO:0050900 : leukocyte migration [274 gene products]
GO:0016032 : viral reproduction [542 gene products]
GO:0045202 : synapse [874 gene products]
GO:0030054 : cell junction [1090 gene products]
GO:0031012 : extracellular matrix [744 gene products]
GO:0044421 : extracellular region part [1579 gene products]
GO:0005789 : endoplasmic reticulum membrane [960 gene products]
GO:0010008 : endosome membrane [387 gene products]
GO:0000139 : Golgi membrane [707 gene products]
GO:0005774 : vacuolar membrane [247 gene products]
GO:0012506 : vesicle membrane [493 gene products]
GO:0097060 : synaptic membrane [340 gene products]

Genes under each GO category were retrieved via GOOSE interface (http://www.berkeleybop.org/goose/) using SQL, e.g.

```
SELECT
term.acc AS superterm_acc,
gene_product.symbol AS gp_symbol,
species.ncbi_taxa_id
FROM term
INNER JOIN graph_path ON (term.id=graph_path.term1_id)
INNER JOIN association ON (graph_path.term2_id=association.term_id)
INNER JOIN gene_product ON (association.gene_product_id=gene_product.id)
INNER JOIN species ON (gene_product.species_id=species.id)
INNER JOIN dbxref ON (gene_product.dbxref_id=dbxref.id)
WHERE
term.acc = 'GO:0002682'
AND
species.genus = 'Homo';
```

Cell type specific gene expression

We define cell type specific gene expression based on the IRIS dataset (Abbas et al., 2005, 2009), which profiled major immune cell types on Affymetrix chips.

A gene is deemed to be specific to a cell type if its expression values in this cell type are 4 fold higher than the mean expression value in all other cell types, and with p-value < 0.001 in a rank sum test. This selection method does not guarantee a gene is exclusive to a cell type, but the gene is discriminative and informative when applied to whole blood transcriptome data. The cell types used here include B cells, naïve B cells, memory B cells, plasma cells, T cells, Th1 stimulated CD4+ T cells, Th2 stimulated CD4+ T cells, monocytes, neutrophils, resting dendritic cells, activated dendritic cells and NK cells. Immunology has a long history of using cell surface markers, and cell surface genes remodel quickly in biological events, e.g. cell differentiation or migration. We therefore defined a set of cell type specific genes in parallel, solely using cell surface genes. Examples of these cell type specific surface genes are shown in Figure B3, and the member genes are shown in Table B1.

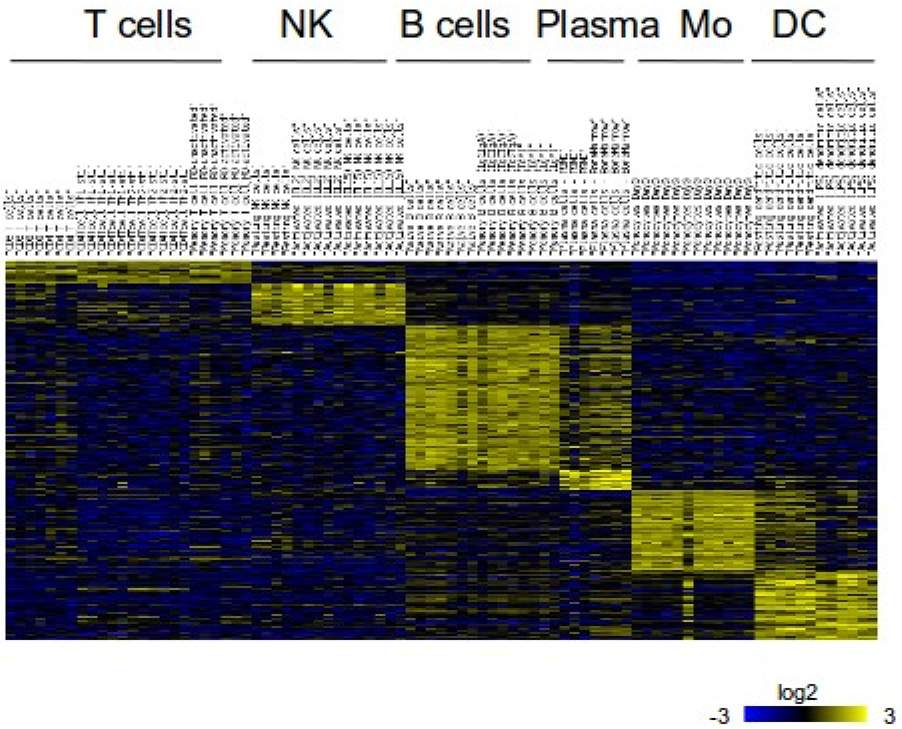


Figure B3: Cell type specific surface genes. NK: natural killer cells. Mo: monocyte. DC: dendritic cells.

Interactome and Bibliome

The interactome data used in this study was retrieved from Pathway Commons (Oct. 26, 2011, Cerami et al., 2011), including data from several databases: HPRD (Keshava et al., 2009), BioGRID (Breitkreutz et al., 2008), IntAct (Aranda et al., 2010), MINT (Ceol et al., 2010), Reactome (Matthews et al., 2009), NCI/Nature PID (Schaefer et al., 2009) and HumanCyc (Romero et al., 2005). The data were represented by a network of 10,426 genes and 141,227 edges. This is a loose definition of interactome, where the bulk of these data are protein-protein interactions.

The bibliome network was built using gene keywords in Pubmed entries (retrieved Aug. 26, 2010), where concurrence of two genes in the same paper will constitute an edge between two genes. Papers with 10 or more keyword genes were excluded, as they are likely to be based on high throughput assays. From 93,436 papers, a network of 16,416 genes and 137,710 edges was constructed.

Different from GO categories and cell type gene expression, interactome and bibliome have their own network edges. The intersection with the master network is rather to define the context of blood transcriptome for interactome and bibliome. The intersection was done on network edges instead of genes for both cases. This produced an interactome subnetwork of 3,949 edges and a bibliome subnetwork of 4,837 edges.

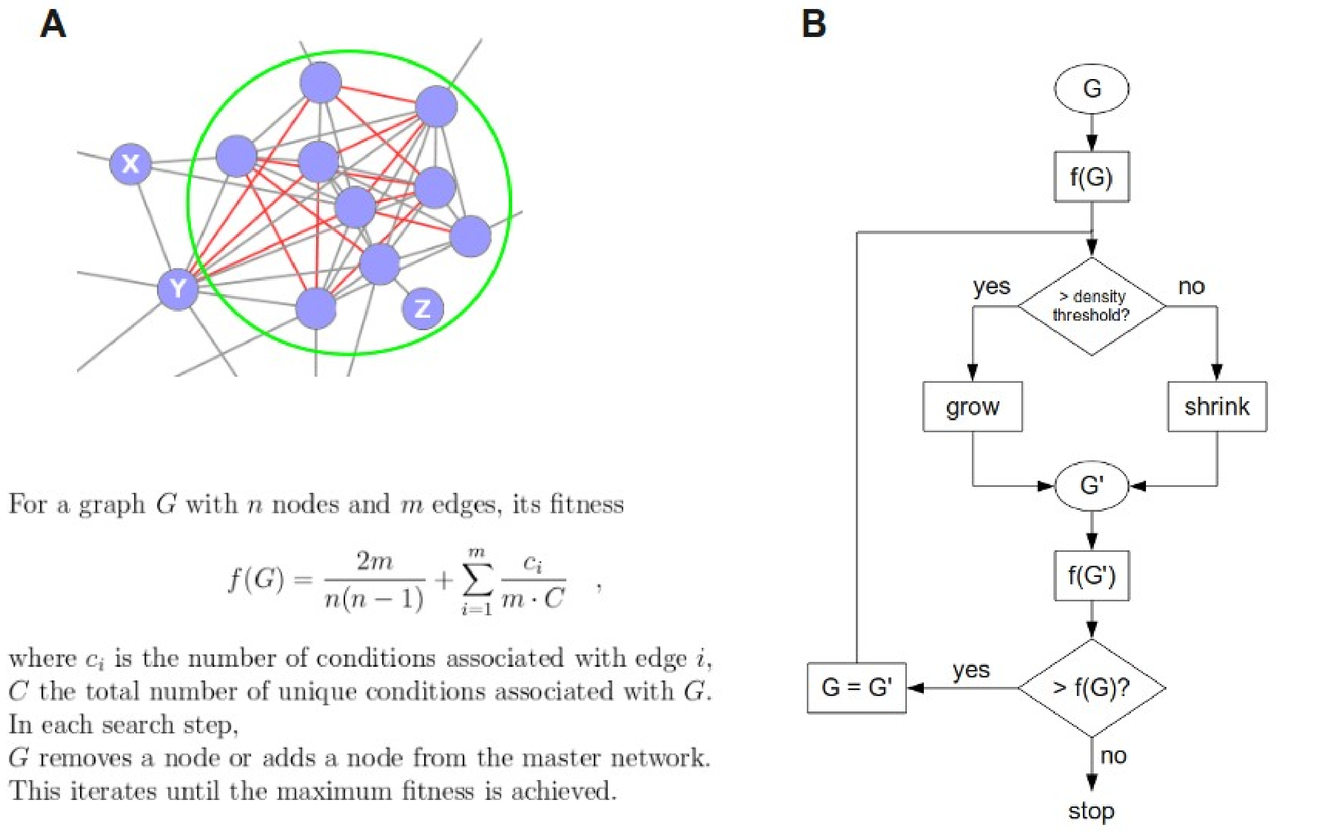
Cell types		Genes
T	27	SIRPG, CLEC2D, MAL, NDFIP2, CD3E, CD28, CD320, ITM2A, SLC38A1, SLC37A3, PTPRCAP, FLT3LG, LRRN3, CD6, ICOS, TMEM106C, ANKH, CXCR6, LAG3, GIMAP2, GPR171, C6orf129, NPDC1, SIT1, C12orf23, CD3G, CD2
NK	48	KIR2DS4, KIR2DS5, KIR2DS2, KIR2DS3, KIR2DS1, ADRB2, KIR2DL3, KIR2DL1, HAVCR2, TMEM64, KLRF1, PTGER2, PTGDR, IL2RB, NCR1, KLRK1, KIT, S1PR5, IL18RAP, SLCO4C1, GPR82, ATP9A, SLC7A5, ELOVL6, PDGFRB, SYNGR3, GPR114, KIR2DL2, KIR2DL5A, DLL1, SLAMF7, IL12RB2, KIR3DL3, TIE1, CX3CR1, ATP8B4, IL18R1, PVRIG, TNFSF11, TNFSF14, TGFBR3, NCAM1, CD97, KLRC3, RARRES3, KIR3DL1, ENPP5, FASLG
B	169	SSPN, HRH4, MOGAT2, TACR3, TMEM156, HLA-DMA, PTH2R, NRXN3, TMEM47, LHCGR, BTC, UGT2A3, CSMD1, NLGN1, CELSR1, CXADR, HLA-DPA1, EDNRB, FCGR2B, CD37, FLRT2, NPY1R, CD200, PLD4, CD79B, CD79A, SLC6A16, SLC6A15, OR2J2, TRHDE, CD22, ABCB4, CD24, ADAM7, GPR6, CCR6, LPPR4, ITM2C, OR11A1, GJC1, P2RX5, PCDH9, GRM5, CD19, SLC17A6, NLGN4Y, SLC24A1, SCN1A, TMEM100, LRRTM4, PCDH10, GPR98, ITGB6, PPP1R3A, FREM2, PDZK1IP1, NT5E, PDPN, GPR85, GPR87, IMPG2, CD180, ADAM28, KCNE4, SLC44A5, USH2A, MS4A1, LY9, ROBO2, NOX4, SLC26A7, CNR1, SGCE, PTPRR, SEMA4B, OR12D3, SLITRK6, GPR116, OR2W1, IL5RA, HLA-DOB, SLC30A10, HLA-DRA, CXCR5, PCDH20, TAS2R13, GHR, CDH5, TPTE, ASTN1, TRPC1, TRPC7, C11orf87, HHLA2, CADM2, SLC5A7, LPHN3, LPHN2, FCRL1, FCRL2, CHRM2, HTR1E, GPR37, IL28RA, F2RL3, IL13RA2, OR1E1, HLA-DOA, PCDHA3, KCNMB3, FCRL3, FCRL5, TMEFF2, EPHA3, TLR10, PTPRK, GYPA, GYPE, CHL1, GABBR1, MC3R, DCC, PLP1, CLCA4, PCDHB10, PCDHB16, PCDHB15, PCDHB14, FAIM3, ABCA8, FAM26F, TSPAN13, HEPACAM2, EGF, POPDC3, SLC12A1, CLDN8, NCAM2, JAM2, SCN3A, CD72, GABRA5, BTLA, TAAR2, KCND2, BAI3, AVPR1A, GP5, HLA-DPB1, TM4SF20, NMBR, PCDHB8, PCDHB4, TMEM133, CR2, AGTR1, CCR9, CHRNB4, GABRB1, CNTNAP3, CSPG4, SYT2, SLC13A1, LRRC19, CD52, LRP1B, JAM3, HEPH, SELE
Plasma	24	KCNH1, KCNG2, DRD4, SLC16A14, CCR10, TEK, CLPTM1L, PPAP2C, TMEM37, SLC44A1, KCNN3, SDC1, GP1BB, SCARB2, SLC5A4, CHRM1, CAV1, TRAM2, GPR25, HM13, AMIGO3, ICAM2, KRTCAP2, TXNDC15
Monocytes	94	LILRB3, LILRB2, LILRB1, ITGAM, TMEM154, SPNS1, C10orf54, MCTP1, ASGR1, ASGR2, LRRC33, FCGR2A, CD33, CD36, APCDD1, FCAR, FXYD6, LPPR2, IL1R2, SLC16A3, CD302, FCER1A, CSF3R, TMEM55A, FCGRT, P2RY2, FCGR1B, HLA-DRB4, MARCO, SLC46A2, S1PR3, IL6R, MS4A6A, GPR133, LTBR, FPR1, FPR2, PTAFR, ANO10, TLR2, TLR1, TLR4, TLR5, PYCARD, PTGIR, KCNQ1, SIRPB1, MBOAT7, P2RY13, GPR109B, C5AR1, PLXND1, TREM1, EMR1, EMR2, MGAM, VNN2, PECAM1, P2RX1, CYBRD1, CLEC12B, CLEC12A, SLC7A7, AGTRAP, STEAP4, GLIPR1, SLC40A1, PTPRE, DYSF, C19orf59, TNFSF12, TNFRSF10B, ABCA1, SLC24A4, CD93, GPER, VSTM1, MFSD1, TMEM71, TNFRSF1B, CD4, CCR1, CCR2, LRP1, C1orf162, TMTC2, BRI3, NFAM1, KCNE3, AMICA1, CECR6, CD163, SLC11A1, SIGLEC9
DC	82	TMEM51, GRINA, PRRG4, SLC36A1, CALCRL, COLEC12, ALCAM, C19orf28, LAMP2, CSF2RB, HLA-DQB1, KMO, GPNMB, AGPAT3, TMEM158, TFRC, OLR1, ADAM9, CD151, IL1R1, ABHD12, TNFRSF11A, FZD5, DENND1B, PPAP2B, CXCL16, CD1E, CD1B, CD1A, GPR137B, FAM70A, LHFPL2, GJB2, P2RY6, SLC1A3, JAG1, TACSTD2, SRD5A3, FPR3, CCRL2, CRIM1, SLC7A8, SDC2, CD83, ADAM12, SLC38A6, NRP1, NRP2, RAMP1, SLC41A2, PDCD1LG2, SLAMF8, CD274, ITPRIPL2, LRFN4, PTGFRN, ACE, SUCNR1, EMP1, SLC6A6, TREM2, TM2D2, ABCA6, GPR157, ATP1B2, ATP1B1, PSEN2, DIRC2, TM7SF4, SLCO2B1, CD9, CLDN23, HLA-DQA1, ABCC3, MSR1, SPINT2, TGFA, TSPAN33, CD58, LRP11, SLC7A11, SIGLEC1

Table B1: cell type specific surface genes.

Module search algorithms

Modules in a network are usually defined by degree distributions, i.e., members in a module have more connections (edges) to each other than to outsider nodes. Finding modules in a large network is a computationally intensive task. It is impractical to sample all possible subnetworks, thus all algorithms are heuristic in some way. Two algorithms were used for all networks in this study, the MCODE algorithm for de novo search, and a novel algorithm for seeded search that incorporated further condition specificity.

MCODE (Bader and Hogue, 2003) is a widely used algorithm, which was originally designed to find protein complexes in protein-protein interaction networks. This algorithm starts by finding nodes surrounded by the most dense connections, and grow the module outwards until the density drops to a certain threshold. MCODE fits in our situation and is adequately fast. We thus used MCODE (as Cytoscape plugin, default parameters) to do *de novo* search in our networks.



For a graph G with n nodes and m edges, its fitness

$$f(G) = \frac{2m}{n(n-1)} + \sum_{i=1}^{m} \frac{c_i}{m \cdot C} \quad ,$$

where c_i is the number of conditions associated with edge i,
C the total number of unique conditions associated with G.
In each search step,
G removes a node or adds a node from the master network.
This iterates until the maximum fitness is achieved.

Figure B4: The seeded search algorithm. A) The fitness function consists of two parts, degree density and condition density. An edge can be associated with multiple conditions (red is shown for illustration). The condition density is analogous to degree density, counting the fraction of condition numbers over the numbers of uniform conditions (all edges from the same conditions). In the next search step, gene Y will be added to the module because that increases module fitness. B) Flowchart of the seeded search algorithm. Thresholds are predefined for both degree density and condition density. The module will grow (adding genes) if the density thresholds are satisfied, or shrink (removing genes) otherwise. This iterates until maximum fitness is achieved.

A distinctive feature in our study is context specificity, which is explicit in the 77 subnetworks. However, the master network also contains implicit specific contexts. That is, the master network is based on 540 studies – a module may be present in 10 studies but not in others. This is similar to the concept of biclustering, where a transcriptional program is active only under a subset of conditions. We thus accommodate this scenario in a concept similar to degree density (Figure B4, A). A degree density is the fraction of present edges over all possible edges, while our “condition density” is the fraction of conditions (studies) over the ideal scenario of all edges appear under identical conditions. Our seeded search algorithm uses this fitness function as the combination of degree density and condition density.

This seeded algorithm starts with predefined seeds, then examines their neighbors in the network. A neighbor is added to the module if so increases the fitness function. The process iterates until no fitness is gained (Figure B4, B). A seed gene may also be dropped out of module if that benefits the fitness. When this search algorithm was applied to the 77 subnetworks, the “condition density” is set to constant since they already carry specific biological contexts. The predefined seeds provide an opportunity to integrate preexisting biological knowledge, i.e., a pathway (a set of genes) can be just integrated as a seed. We used 1,670 seeds from pathway databases (KEGG, Biocarta, Reactome, NCI/Nature PID) and collection of known transcription factor targets (from MSigDB, Subramanian et al., 2005).

With both algorithms running over 78 networks, the search results have a lot of redundancy by design, which are reconsolidated in the post-processing step.

Post-processing of gene modules

The post-processing consists of the following steps:

1. Remove redundant modules
2. Filter for gene numbers ≥ 10
3. Filter for degree density > 0.3
4. Filter out modules of over-represented genes
5. Group overlap modules
6. Annotation

The module search algorithms pulled out 5159 raw modules from 78 networks. Many modules were discovered many times, as the search algorithms will converge from different seeds, and the networks have redundancy. E.g. module M54 BCR signaling was discovered by the seeded algorithm in two subnetworks, GO:0001775 (cell activation) and GO:0045321 (leukocyte activation). Module M61.1 (Figure B5) was discovered 3 times by both algorithms in the Bibliome subnetwork, 3 times by both algorithms in the NK specific subnetwork, 4 times by the seeded algorithm in the NK specific surface gene subnetwork:

denovo_bibliome_premodule_8
sub_bibliome_KEGG_ANTIGEN_PROCESSING_AND_PRESENTATION
sub_bibliome_KEGG_NATURAL_KILLER_CELL_MEDIATED_CYTOTOXICITY

denovo_signature_NK_premodule_3
sub_signature_NK_KEGG_ANTIGEN_PROCESSING_AND_PRESENTATION
sub_signature_NK_KEGG_NATURAL_KILLER_CELL_MEDIATED_CYTOTOXICITY

sub_surfaceome_NK_KEGG_ANTIGEN_PROCESSING_AND_PRESENTATION
sub_surfaceome_NK_KEGG_NATURAL_KILLER_CELL_MEDIATED_CYTOTOXICITY
sub_surfaceome_Activated_NK_cells_KEGG_ANTIGEN_PROCESSING_AND_PRESENTATION
sub_surfaceome_Activated_NK_cells_KEGG_NATURAL_KILLER_CELL_MEDIATED_CYTOTOXICITY

Thus, we first filtered out redundant modules. Two modules were deemed redundant if their member genes have Jaccard index > 0.8. As small modules are susceptible to noises in transcriptomics data. only modules of ten or more genes were kept. Additional filtering was on a degree density of 0.3, as higher degree density increases the likelihood of coexpression of member genes. At this stage, the result was 811 modules, still containing many similar modules. A complete linkage clustering was performed on these modules, using Jaccard index > 0.5, yielding 586 clusters. One representative module was chosen from each cluster (highest degree density by number of genes).

We note that some genes appear more frequently than others in these 586 pre-modules. It is expected as some genes have more visible roles and the module membership is not exclusive. However in the application of data analysis, such bias is not desired. A "normalization" procedure was performed to remove modules of overly represented genes. A uniqueness function is defined per module *M* as:

$$F(M) = \frac{1}{N} \sum_{i}^{N} \frac{1}{\sqrt{f_i}} \quad , \quad i \in M$$

where f_i is the number of appearance of gene *i* in all modules. We iteratively remove the worst performing module, each time with the *F(M)* recalculated, until all modules have *F(M)* > 0.5. That is, in the worst case, a module has member genes that appear on average 4 times. The final result was 334 modules.

Each module inherits edges from the subnetwork where it was extracted. The post-processing only filtered out modules, without changing module memberships.

BTM modules annotation

Modules with overlapping genes will come up together in data analysis. The final 334 modules were therefore clustered into 250 groups (by over 50% overlap), and numbered according to the groups. For example, modules M61.1 and M61.2 are separate modules. Their similarity in numbering reflects their overlap in member genes. Each module contains a list of genes, and their connecting edges inherited from the source network. We visualized all modules in Cytoscape (Smoot et al., 2011). An edge, according to the source network, represents a context-specific coexpression relationship learned from the human blood transcriptome compendium.

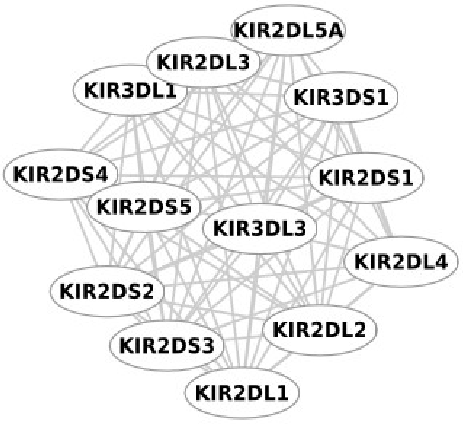


Figure B5. Module M61.1, 13 genes, NK cell signature (KIR cluster)

The biological meaning of many modules is self-evident. For the example in Figure B5, all genes are clearly killer cell immunoglobulin-like receptors. Many others become clear after gathering the information from literature. We have assigned a title to about 80% of the modules. The module annotations are supplied as HTML web pages. The detailed annotation contains origin of module construction, enriched pathways, enriched transcription factor binding sites, most relevant PubMed papers, Gene Ontology terms, cell type bias, and detailed description of member genes from NCBI and iHOP (Hoffmann and Valencia, 2004) databases. The cell type bias was based on Kolmogorov-Smirnov statistic over all genes in a cell type ranked by abundance.

By design, these BTM modules provide a fresh view of the knowledge landscape of biological events in human blood. When compared to KEGG, NCI/Nature PID and BioCarta pathway databases, 25 modules were matched to a known pathway with Jaccard index above 0.2, while about 1/3 of the modules are not matched to a pathway with more than one overlap genes.

Evaluation

Our master gene network before module extraction already showed excellent quality. The primary concern in module evaluation is their sensitivity in application to analyzing immunological blood transcriptome data. From a set of transcriptome data, we compute a single activity score per module. If the module carries biological meaning under the respective study, its member genes will show concerted expression and favor a good activity score. Otherwise, the genes will cancel out each other and yield a low score. Thus, this approach has a built-in mechanism to evaluate both module quality and sensitivity. We use the mean expression value of member genes as the activity score of a module. More elaborate methods were evaluated (e.g. Chuang et al., 2007), but did not offer clear advantage in our test data.

All test data were excluded from BTM construction process to ensure a fair evaluation. We use t-score in a student t-test between two sample classes as a metric for module or pathway sensitivity (Lee et al., 2008). As shown in Figure 4B and Suppl Figure 8, our BTM modules and Chaussabel modules in general outperform all canonical pathways.

The superior sensitivity of BTM modules helps the antibody correlation analysis, where single gene correlations are difficult to distinguish from random data, due to the large number of genes and small cohorts in vaccine studies. When modules are used, the random data from permutations always display a perfect normal distribution. How well the BTM data are separated from random background depends on the size of cohort, the magnitude of transcriptomic response and the spread of antibody response. In the examples like influenza TIV (24 people), the distribution of BTM correlation differs clearly from the random data, while the canonical pathways offer limited resolution (Figure B6). The significance of BTM correlation in MPSV4 antibody response (13 people, limited antibody range) was more limited, but further confidence was gained in the comparison between vaccines.

Many BTM modules carry cell type specific information. To add a redundant control mechanism, we append to BTMs a set of 12 cell specific surface modules. These are the gene sets selected by rank-sum test as an input for context specific subnetwork, before going through the module search process. These additional surface modules (IDs starting with letter “S”) function more like the conventional surface markers.

Discussion

The construction of blood transcription modules shares the same motivation as Chaussabel et al. (2008), while benefits from the approach of large-scale data integration. Our BTM modules provide a sensitive and robust framework for vaccine antibody analysis, as presented in this study. We will continue to explore other applications. For example, the plasma cell/immunoglobulin module (Figure 4D) successfully predicts four regulators of B cell response, TNFRSF17, POU2AF1, CD27 and MZB1, all backed by substantial literature. This shows that BTM modules and the underlying gene network have great potential of generating high-quality gene regulatory hypotheses.

The resolution of what can be learned from whole blood samples can be limiting. There are programs that can only be learned in pure cell populations and they may never show up in whole blood analysis. This also poses a challenge to the annotation of BTM modules. Furthermore, the BTM modules were computationally learned from the data in an unbiased manner. It is no surprise that many modules go beyond the text in current databases and literature. As typically for large-scale data integration approaches, the project will continue to benefit from the newly emerging data.

References

Abbas, A.R. et al. Immune response in silico (IRIS): immune-specific genes identified from a compendium of microarray expression data. Genes and Immunity 6, 319-331(2005).

Abbas, A.R. et al. Deconvolution of Blood Microarray Data Identifies Cellular Activation Patterns in Systemic Lupus Erythematosus. PloS One 4, e6098(2009).

Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, Derow C, Feuermann M, Ghanbarian AT, Kerrien S, Khadake J, et al. The IntAct molecular interaction database in 2010. Nucleic Acids Res. 2010;38:D525-D531.

Auffray, C., Chen, Z. & Hood, L. Systems medicine: the future of medical genomics and healthcare. Genome Medicine 1, 2(2009).

Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics. 2003 Jan 13;4:2.

Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A. Reverse engineering of regulatory networks in human B cells. Nat Genet. 2005 Apr;37(4):382-90.

Ben-Dor,A., Chor,B., Karp,R. and Yakhini,Z. (2002) Discovering local structure in gene expression data: the order-preserving sub-matrix problem. In Proceedings of the 6th Annual International Conference on Computational Biology, ACM Press, New York, NY, USA, pp. 49–57.

Berry, M.P.R. et al. An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. Nature 466, 973-977(2010).

Breitkreutz BJ, Stark C, Reguly T, Boucher L, Breitkreutz A, Livstone M, Oughtred R, Lackner DH, Bahler J, Wood V, et al. The BioGRID interaction database: 2008 update. Nucleic Acids Res. 2008;36:D637-D640.

Ceol A, Chatr Aryamontri A, Licata L, Peluso D, Briganti L, Perfetto L, Castagnoli L, Cesareni G. MINT, the molecular interaction database: 2009 update. Nucleic Acids Res. 2010;38:D532-D539.

Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, Anwar N, Schultz N, Bader GD, Sander C. Pathway Commons, a web resource for biological pathway data. Nucleic Acids Res. 2011 Jan;39(Database issue):D685-90.

Chaussabel, D. et al. A Modular Analysis Framework for Blood Genomics Studies : Application to Systemic Lupus Erythematosus. Immunity 29, 150(2008).

Chaussabel, D., Pascual, V. & Banchereau, J. Assessing the human immune system through blood transcriptomics. BMC Biology 8, 84(2010).

Cheng,Y. and Church,G. (2000) Biclustering of expression data. Proc. Int. Conf. Intell. Syst. Mol. Biol. pp. 93–103.

Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. Mol Syst Biol. 2007;3:140.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999 Oct 15;286(5439):531-7.

Hoffmann R, Valencia A. A gene network for navigating the literature. Nat Genet. 2004 Jul;36(7):664.

Ihmels,J. et al. (2004) Defining transcription modules using large-scale gene expression data. Bioinformatics, 20, 1993–2003.

Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, et al. Human protein reference database-2009 update. Nucleic Acids Res. 2009;37:D767-D772.

Kraskov,A. et al. (2004) Estimating mutual information. Phys. Rev. E, 69, 066138.

Lee, H. K., Hsu, A. K., Sajdak, J., Qin, J., & Pavlidis, P. (2004). Coexpression analysis of human genes across many microarray data sets. Genome research, 14(6), 1085-1094.

Lee, E., Chuang, H. Y., Kim, J. W., Ideker, T., & Lee, D. (2008). Inferring pathway activity toward precise disease classification. PLoS computational biology, 4(11), e1000217.

Lefebvre C, Rajbhandari P, Alvarez MJ, Bandaru P, Lim WK, Sato M, Wang K, Sumazin P, Kustagi M, Bisikirska BC, Basso K, Beltrao P, Krogan N, Gautier J, Dalla-Favera R, Califano A. A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. Mol Syst Biol. 2010 Jun 8;6:377. PubMed PMID: 20531406; PubMed Central PMCID: PMC2913282.

Lefebvre C, Rieckhof G, Califano A. Reverse-engineering human regulatory networks. Wiley Interdiscip Rev Syst Biol Med. 2012 Jul-Aug;4(4):311-25. Doi: 10.1002/wsbm.1159.

Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, de Bono B, Garapati P, Hemish J, Hermjakob H, Jassal B, et al. Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res. 2009;37:D619-D622.

Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics. 2006 Mar 20;7 Suppl 1:S7.

Prelić A, Bleuler S, Zimmermann P, Wille A, Bühlmann P, Gruissem W, Hennig L, Thiele L, Zitzler E. A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics. 2006 May 1;22(9):1122-9.

Pulendran, B., Li, S. & Nakaya, H.I. Systems vaccinology. Immunity 33, 516-29(2010).

Quartier P, Allantaz F, Cimaz R, Pillet P, Messiaen C, Bardin C, Bossuyt X, Boutten A, Bienvenu J, Duquesne A, Richer O, Chaussabel D, Mogenet A, Banchereau J, Treluyer JM, Landais P, Pascual V. A multicentre, randomised, double-blind, placebo-controlled trial with the interleukin-1 receptor antagonist anakinra in patients with systemic-onset juvenile idiopathic arthritis (ANAJIS trial). Ann Rheum Dis. 2011 May;70(5):747-54.

Romero P, Wagg J, Green ML, Kaiser D, Krummenacker M, Karp PD. Computational prediction of human metabolic pathways from the complete human genome. Genome Biol. 2005;6:R2.

Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Mewes HW. CORUM: the comprehensive resource of mammalian protein complexes--2009. Nucleic Acids Res. 2010 Jan;38(Database issue):D497-501.

Sales G, Romualdi C. parmigene--a parallel R package for mutual information estimation and gene network reconstruction. Bioinformatics. 2011 Jul 1;27(13):1876-7.

Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH. PID: the Pathway Interaction Database. Nucleic Acids Res. 2009;37:D674-D679.

Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet. 2003 Jun;34(2):166-76.

Segal E, Friedman N, Koller D, Regev A. A module map showing conditional activity of expression modules in cancer. Nat Genet. 2004 Oct;36(10):1090-8.

Segal et al. (2005a) Learning Module Networks. Journal of Machine Learning Research 6, 557-588.

Segal E, Friedman N, Kaminski N, Regev A, Koller D. (2005b) From signatures to models: understanding cancer using microarrays. Nat Genet. 37 Suppl:S38-45.

Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics. 2011 Feb 1;27(3):431-2.

Subramanian, A. et al. Gene set enrichment analysis : A knowledge-based approach for interpreting genome-wide. PNAS 102, 15545(2005).

Tanay,A. et al. (2002) Discovering statistically significant biclusters in gene expression data. Bioinformatics, 18 (Suppl. 1), S136–S144.

Tanay,A. et al. (2004) Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. Proc.Natl Acad. Sci. USA, 101, 2981–2986.

Tattermusch, S. et al. Systems Biology Approaches Reveal a Specific Interferon-Inducible Signature in HTLV-1 Associated Myelopathy. PLoS pathogens 8, e1002480(2012).