A Novel Repertoire of Blood Transcriptome Modules Based on Co-expression Patterns Across Sixteen Disease and Physiological States

Matthew C Altman[1,2,*], Darawan Rinchai[3], Nicole Baldwin[4], Elizabeth Whalen[1], Mathieu Garand[3], Basirudeen Ahamed Kabeer[3], Mohammed Toufiq[3], Scott Presnell[1], Laurent Chiche[5], Noemie Jourde-Chiche[6], J Theodore Phillips[4], Goran Klintmalm[4], Anne O'Garra[7,8], Matthew Berry[9], Chloe Bloom[7], Robert J Wilkinson[10,11,12], Christine M Graham[7], Marc Lipman[13], Ganjana Lertmemongkolchai[14], Davide Bedognetti[3], Farrah Kheradmand[15], Asuncion Mejias[16], Octavio Ramilo[16], Karolina Palucka[4,17], Virginia Pascual[4,18], Jacques Banchereau[4,17], Damien Chaussabel[1,3*]

1 *Systems Immunology, Benaroya Research Institute, Seattle, Washington, USA*

2 *Division of Allergy and Infectious Diseases, University of Washington, Seattle, Washington, USA*

3 *Systems Biology, Sidra Medicine, Doha, Qatar*

4 *Baylor Institute for Immunology Research and Baylor Research Institute, Dallas, Texas, USA*

5 *Department of Internal Medicine, Hospital Europeen, Marseille, France*

6 *Aix-Marseille University, C2VN, INSERM 1263, INRA 1260, Marseille, France*

7 Laboratory of Immunoregulation and Infection, The Francis Crick Institute, London, UK

8 *National Heart and Lung Institute, Imperial College London, London W2 1PG, UK*

9 *Respiratory Medicine, Imperial College Healthcare NHS Trust, London, UK*

10 *The Francis Crick Institute, London, UK*

11 *Department of Medicine, Imperial College, London, UK*

12 *Wellcome Center for Infectious Diseases Research in Africa and Department of Medicine, Institute of Infectious Diseases and Molecular Medicine, University of Cape Town Observatory 7925, Republic of South Africa*

13 *UCL Respiratory, Division of Medicine, University College London, London, UK*

14 *Centre for Research and Development of Medical Diagnostic Laboratories, Faculty of Associated Medical Sciences, Khon Kaen University, Khon Kaen, Thailand*

15 *Baylor College of Medicine, Houston, Texas, USA*

16 *Nationwide Children's Hospital and the Ohio State University School of Medicine, Division of Pediatric Infectious Diseases, Columbus, Ohio, USA*

17 *The Jackson Laboratory for Genomic Medicine, Farmington, Connecticut, USA*

18 *Weill Cornell Medicine, New York, New York, USA*

[*]To whom correspondence may be addressed:

Matthew C Altman, MD, Systems Immunology Division**,** Benaroya Research Institute, 1201 Ninth Avenue, Seattle, WA 98101, USA. Tel. +1 206 287 5648, Fax. 206 287 5682, E-mail: maltman@benaroyaresearch.org

Damien Chaussabel, PhD, Systems Biology Department**,** Sidra Medical and Research Center, PO Box 26999 Al Luqta Street, Doha, Qatar. Tel. +974 4003 7395, E-mail: dchaussabel@sidra.org

Keywords: human immunology; transcriptome; gene expression; systems biology; network analysis; modular repertoire

ABSTRACT

Blood transcriptomics measures the abundance of circulating leukocyte RNA on a genome-wide scale. Dimension reduction is an important analytic step which condenses the number of variables and permits to enhance the robustness of data analyses and functional interpretation. An approach consisting in the construction of modular repertoires based on differential co-expression observed across multiple biological states of a given system has been described before. In this report, a new blood transcriptome modular repertoire is presented based on an expended range of disease and physiological states (16 in total, encompassing 985 unique transcriptome profiles). The input datasets have been deposited in NCBI public repository, GEO. The composition of the set of 382 modules constituting the repertoire is shared, along with extensive functional annotations and a custom fingerprint visualization scheme. Finally, the similarities and differences between the blood transcriptome profiles of this wide range of biological states are presented and discussed.

BACKGROUND:

Blood transcriptomics.

Blood transcriptome profiling approaches have been employed for nearly two decades in a wide range of settings [1-4]. It consists of measuring leukocyte transcript abundance on a genome-wide scale. This application was enabled following the introduction of microarray technologies and more recently of RNA sequencing. Leukocyte transcriptome profiles have more commonly been measured in whole blood or peripheral blood mononuclear cell fractions. But studies have also investigated changes in transcriptome abundance in isolated leukocyte populations as well as in single cells [5, 6]. Global changes in transcript abundance can also be observed upon stimulation of the cells *in vitro* with host or environmental derived immunogenic factors, such as pathogen-associated molecular pattern, antigenic peptides, as well as pro or anti-inflammatory cytokines or chemokines [7, 8].

Dimension reduction approaches.

Approaches that permit organization or reduction of large number of variables being measured are commonly adopted when working with omics datasets. Principal component analysis (PCA) is usually employed as a “first line” dimension reduction strategy. It consists in converting sets of correlated variables into aggregate variables called principal components. Rather than being used to identify gene “signatures” PCA mostly serves to “reveal internal structure of the data in a way that best explains the variance in the data” [9]. For this reason, PCA tends to be performed as one of the first step in the analysis as the information it conveys can direct the design of downstream analyses. For instance, in an experiment where cells from

different donors are exposed to stimuli *in vitro* a PCA plot would permit to determine among sample processing batch, donor and stimulation, which of these factors contributes the most to overall variance observed. If stimulations are especially potent they may "override" the variance associated with donor-donor differences. But in cases where stimulations have subtler effects donor-donor differences may be the predominant source of variation.

Clustering tends to be used in subsequent analysis steps and is by far the most commonly used dimension reduction approach. It consists in grouping transcripts based on similarities in expression levels. Individual genes can thus be reduced into "signatures", each constituted by multiple genes that show some degree of correlation. Clustering can be applied either prior to or following feature selection. Clustering methods which are commonly used include hierarchical clustering and k-means clustering.

Another less commonly employed approach to dimension reduction relies on the construction and mining of correlation networks. A network is constituted of "nodes" and "edges" (the later are the lines connecting the different nodes to indicate a relationship). Nodes can represent one of many different things e.g.: genes, proteins, samples, individuals etc..; likewise, the nature of the relationship depicted by the edges varies, e.g.: physical interaction, regulation, co-occurrence in literature abstracts. Also the first step when "reading" a network consists in determining what its nodes and edges represent. In a correlation network the nodes represent genes and the edges correlation in level of abundance of their product (RNA in the case of transcriptomic data). The use of correlation networks for transcriptome data analysis has been covered extensively in a recent review by Van Dam et al [10]. One of the approaches most commonly employed for correlation-based expression analysis is Weighted Gene Correlation Network Analysis (WGCNA) [11]. Using a transcriptome dataset as input, it consists in building

a weighted correlation network (i.e. edges receive a "connection weight" according to the strength of the correlation). This network is subsequently partitioned into sets of highly correlated genes, referred to as modules, using hierarchical clustering.

Module repertoires.

A first "modular repertoire framework" was developed by our group over 10 years ago, specifically for analysis and interpretation of blood transcriptome data [12] [13]. Such a framework consists of: 1) A collection of transcriptional modules (i.e. the modular repertoire), 2) Functional interpretations for the different modules 3) A fingerprint visualization designed for mapping perturbations of a given modular repertoire (as compared to steady state or appropriate baseline).

The construction of a modular repertoire constitutes a dimension reduction step. The approach is similar to network correlation analyses described above: i.e. a network serves as a basis for module selection, its nodes represent genes, and edges connecting the nodes represent gene co-expression. The fact that differential co-expression across distinct biological states for a given system is taken into account may be the most distinctive feature (**Figure 1**). For instance, in the case of blood transcriptomics, the network would factor in whether co-expression between a pair of genes occurs across all the repertoire of pathological or physiological states or only in some of those states. This information is "encoded" in the network used for module construction via the weights which are attributed to each of the edges. Indeed, if ten different states are covered (i.e. ten different input datasets) then the weight of each edge will vary between 1 and 10 (co-expression observed in one state/dataset or up to 10 states/datasets). As will be further described in the article each module can subsequently be linked to the specific states in which its constitutive genes were co-expressed.

To date two "modular repertoires frameworks" have been constructed and used for analysis and interpretation of whole blood transcritptome profiling data. A first repertoire based on 8 disease states was published in 2008 (**Table 1**) [12]. The total combined number of samples across the 8 input datasets was 239. Transcriptome profiles were generated from purified peripheral blood mononuclear cells using Affymetrix GeneChips. Five years later a second repertoire based on 7 disease states and a total of 410 samples was constructed [14]. Transcriptome profiles were in this case generated from whole blood using Illumina Beadarrays. Input datasets encompass a wide breadth of biological states on which basis the weighted co-expression network was built. This included patients with autoimmune diseases (systemic lupus), inflammatory conditions (systemic onset juvenile arthritis), viral and bacterial infections (e.g. *Staphylococcus* infection, HIV, Influenza, RSV) or cancer (stage IV melanoma). Since perturbations across a wide range of states is factored into the construction of the repertoire it should prove suitable as a generic framework for interpretation of blood transcriptome datasets. This appears to indeed be the case given the extent to which the two repertoire framework which have been previously published have been reused (**Figure 2**).

Attempts were made at assigning functional interpretations to the modules constituting the framework. A common misconception is that function is used as a basis for the construction of modular repertoires. In fact, module construction is entirely data-driven and putative functions are only assigned afterwards based on gene ontology or pathway enrichment analysis the gene sets constituting each of the modules are subjected to.

Visualization is another important element when is comes to interpretation of high dimensional data. Reducing the dimension of datasets from tens of thousands of variables to a few hundred opens new possibilities with that regard. A fingerprint representation was

158 introduced along with the first generation of module repertoires published in 2008. It consists in
159 fixing the position of individual modules on a grid. At each position a spot would indicate
160 compared to a baseline either increase in abundance for transcripts constituting the module (in
161 red) or decrease in abundance (in blue). Functional interpretations can also be mapped to the grid
162 to aid interpretation of the results.

163 This article serves on one hand as a resource, making available a new blood modular
164 transcriptome repertoire, along with: 1) the algorithm used for its construction, 2) the 16 input
165 datasets, each representing a different disease or physiological state; 3) functional interpretations,
166 along with underlying functional/literature profiles; and 4) a new fingerprint representation. It
167 provides on the other hand a high-level unbiased molecular classification of a wide range of
168 immunological processes involved in health maintenance and pathogenesis.

169

170 **METHODS:**

171 **Study subjects**

172 Module construction: Gene expression datasets from 985 de-identified subjects from
173 distinct cohorts from the Baylor Institute for Immunology Research (BIIR) were used for this
174 study. Each of those studies was approved by the Baylor Institutional Review Board (IRB #'s
175 009-240, 006-177, 002-197, 009-257, H-18029, HE-470506). Gene expression datasets were
176 selected to cover major classes of immune states (**Table 2**), were required to have a minimum of
177 25 total samples, and at least 20% of the total samples were required to be appropriately matched
178 controls.

RNA extraction and processing

Whole blood for all sample sets were collected into Tempus Blood RNA Tubes (Thermo Fisher Scientific). Total RNA was isolated from whole blood lysate using MagMAX for Stabilized Blood Tubes RNA isolation kit for Tempus Blood RNA Tubes (Thermo Fisher Scientific). RNA quality and quantity were assessed using Agilent 2100 Bioanalyzer (Agilent Technologies) and NanoDrop 1000 (NanoDrop Products, Thermo Fisher Scientific). Samples with RNA integrity numbers values >6 were retained for further processing.

Microarray analysis

Gene expression profiles from whole blood samples generated using Illumina HumanHT-12 v3.0 or Illumina HumanHT-12 v4.0 expression BeadChips were obtained for 16 groups of patients selected as above. Sixteen datasets were used as input (Table 1). Each dataset's expression data was preprocessed and clustered independently of the rest. First probes were discarded if they were not present (detection $P < 0.01$) in at least ten samples or in at least ten percent of the samples, whichever was greater. Then, the sample data for each dataset was normalized using the BeadStudio average normalization algorithm. Once normalized, the signal was floored such that all signals less than ten were set to ten. Then, the fold change was calculated relative to the median signal for that probe across all samples. If the difference between a signal and the probe's median signal was less than 30, or the calculated absolute magnitude of the fold change was less than 1.2, the fold change was set to 1 in order to reduce noise from low-level responses. At this stage, probes were filtered again. Probes were retained only if they had a calculated absolute fold change greater than 1 in at least ten samples or in at least ten percent of the samples, whichever was greater. Finally, the data was transformed to the $\log_2$ of the calculated fold changes.

Module construction algorithm

Sets of coordinately regulated genes, or transcriptional modules, were extracted from the whole blood microarray datasets. Each of the preprocessed microarray datasets was clustered in parallel using Euclidean distance and the Hartigan's K-Means clustering algorithm. The 'ideal' number of clusters (k) for each dataset was determined within a range of k=1 to 100 by means of the jump statistic [15] . Taking the sixteen sets of clusters as input (**Table 2**), we constructed a weighted co-cluster graph [12, 16] . To select modules, we employed an iterative algorithm to extract sets of probes that are most frequently clustered together in the same datasets, proceeding from the most stringent requirements to the least as previously described [12] . This iteration differed from previous implementation of this algorithm in that the k was calculated independently for each dataset cluster and the size of the core sub-networks was smaller (10 probes). The algorithm also was changed from previous implementations to ensure that the core sub-networks co-clustered in the same datasets. Further details and an example of the code are included in the supplemental methods (**Supplementary File 1**). The resulting 382 module set constitutes the third generation of modular blood transcriptome repertoire constructed since the initial publication in 2008 [12] of the first generation, and in 2013 of the second [14].

Module annotation

Module gene lists were investigated using Database for Annotation Visualization and Integrated Discovery (DAVID) version 6.7[17, 18] . This database uses a modified Fisher exact test to identify specific biological/functional categories that are overrepresented in gene sets in comparison with a reference set (the human genome was used as the reference set). The top matched DAVID annotation cluster (using default settings), the top matched canonical pathway from Kyoto encyclopedia of genes and genomes (KEGG), the top matched pathway from

BioCarta, and the top matched Gene Ontology biologic process (GO_BP) and molecular function (GO_MF) terms were identified for each module. Each module was also investigated for significant overlap with 2 other established blood transcriptome module repertoires[14, 19] . These findings are summarized in the module annotation spreadsheet (**Supplemental File 2)**.

Literature profiling: Acumenta Biotech Literature Lab™ (LitLab) was used to associate genes within a particular module to terms in PubMed abstracts [20]. Association scores reflecting the strength of the associations were used to calculate the "Product Scores". The top 3 terms that showed the strongest association and highest "Product Scores" were used to create the functional annotation. A similar approach using LitLab has been previously reported [7]. The steps taken to annotate all 382 modules is described briefly here. All statistical analyses were performed using Microsoft Excel (2010) with Visual Basic for Applications (VBA), Linux-based command line in Mac OS, and R statistical software.

The first part into the construction of a Product Scores table consist of listing all the term available in LitLab (over 80,000). Next genes in each module were submitted as a list to LitLab Editor and manually validated using LitLab's built-in validation tool and/or NCBI Gene (https://www.ncbi.nlm.nih.gov/gene) prior to submission for analysis using all domains available. After the analysis was completed the summary result page was exported to an xls file. Using UNIX command line, the exported files were converted to csv files with the filename appended in the last column of each row and vertically appended. The "merged" file was used to populate the table including all available LitLab terms. The top 3 terms with highest Product Scores were selected to represent the module functional annotation and are tabulated in column I of the module annotation table (**Supplementary file 2**).

248 **Module grid visualization**

249 Modules were arranged on a grid based on similarities in patterns of activity across the 16 input
250 datasets, each of them corresponding to a different pathological or physiological state. First,
251 modules were partitioned using K-means clustering, which resulted in the constitution of 38
252 clusters. Given the possibility of collapsing values of the modules constituting each cluster in a
253 single "aggregate" value the term "module aggregate" was used to designate each cluster (A1 to
254 A38). Of these 38 k-means clusters, 27 comprised of more than one module. Modules were next
255 arranged on a grid with each row corresponding to modules belonging to the same aggregate
256 (Figure 4). Therefore, the total number of row on the grid equals 27 and number of columns
257 equals the largest number of modules for a given aggregate, which is 42 (for aggregate A2). For
258 each module the highest of the two values indicating increase or decrease is selected for
259 visualization (e.g. if % increase > % decrease, then a red sport representing % increase is
260 shown).
261
262

RESULTS:

Input datasets

Sixteen datasets were used as input for the construction of a third blood transcriptome module repertoire. This collection encompasses 985 individual whole blood transcriptome profiles. All the samples were processed in the same facility and and run on Illumina HT12 beadarrays. Each dataset corresponds to a different pathological or physiological state. The range has been expended considerably compared to the first and second repertoires which were published previously (**Table 1**).

Some "core" pathologies are again covered, with for instance systemic lupus erythematosus, systemic onset juvenile idiopathic arthritis, liver transplant recipients under immunosuppression and patients with metastatic melanoma, which are represented in all three versions. Infectious diseases are again well represented, including viral respiratory viruses, influenza and RSV, as well as HIV and infections caused by *Mycobacterium tuberculosis*, *Staphylococcus aureus*, or *Burkholderia pseudomallei* (agent of Melioidosis).

New to this third framework are inflammatory conditions of the skin, lung or circulation (COPD, juvenile dermatomyositis, Kawasaki disease, respectively. A neurogenerative disease (MS). Primary immune deficiency (B-cell deficiency) and physiological variant, pregnancy.

Absent from this repertoire, which were represented earlier are type 1 diabetes and *Escherichia coli* infection. The intent while constituting the collection of input dataset was to capture a wide breadth of immunological response or perturbations (e.g. interferon responses, inflammation, autoimmune processes, tolerance, "loss" of a leukocyte population) as to be able to construct a "generic" modular repertoire. Number of samples included in each of the datasets are provided in the methods section and in **Table 2**. The datasets have been deposited in the NCBI Gene Expression Omnibus, GEO (GSE100150).

Module repertoire construction

The algorithm employed for construction of the module repertoire is described in details in the supplementary methods section. Pseudocode is also provided to facilitate implementation in different programming language. The major steps are also described in **Figure 3**. Briefly: 1) input datasets are assembled; 2) transcripts which show no or very little expression across all conditions are filtered out; 3) clustering is performed for each individual dataset; 4) a weighted co-expression network is constructed, where edges between the genes represent at least one co-clustering event in one of the input datasets. Weight is assigned based on the total number of co-clustering events (up to 16, when co-clustering between the pair of genes occurs in all input datasets); 5) The resulting network is mined to identify highly inter-connected sub-networks, which form modules. The approach takes into account weights since the first sub-networks to be "extracted" are those with the highest number of states in which co-clustering is observed.

This approach captures relationships that exist among constitutive elements of our biological system (blood) and the given range of disease states. It is unbiased in that it does not rely on any previous information about interactions among genes or knowledge about the gene

function. Using this technique, from >47,000 total transcripts, 15132 total transcripts passed the expression filter and 382 modules were identified which consist of 14,502 of those transcripts (95.8%).

Blood transcriptome modular repertoire

The output of the module repertoire construction process is a collection of gene sets, aka "modules". The gene composition of each of the 382 module which were identified is provided in a supplemental file (Column D: Number of unique genes; Column E: Illumina probe IDs; and column F: symbols of member genes). Average number of unique constitutive genes per module is 37.1, median is 26.5 and range is 12 – 169. Extensive functional profiling was also carried out with, enrichment results provided for: a) Literature lab abstract keyword profiling (column I) [ref], b) DAVID (columns J-L) [ref], KEGG (columns N-P), Biocarta (columns R-T) [ref], OMIM (columns U-X), GOTERM (columns Y-AF). In addition, extent of overlap with the previous modular repertoire as well as with a set of modules constituted at Emory university [19] is presented in columns AG-AN.

Taken together outputs from this wide range of functional enrichment analyses was employed to assign, when possible, a consensus functional association title for modules (column C). For instance, M16.3 (145 genes) shows the following enrichment pattern (encompassing, literature terms, pathways, diseases, ontologies in columns I through AF): T-Lymphocytes, Lymphocytes / Structure_of_Caps_and_SMACs; Ikaros_and_signaling_inhibitors / Primary immunodeficiency / Lck and Fyn tyrosine kinases in initiation of TCR Activation / lymphocyte activation / phosphatase activity. While some of the terms may appear rather cryptic or lack specificity based on overall convergence "T-cell" was the consensus functional annotation title

assigned for this module. However, for the majority of modules functional annotations did not show sufficient convergence, or were too few for a consensus annotation to be assigned and received instead the TBD label (279 out of 382 modules).

Module-level analyses

For each module the proportion of its constitutive transcripts which abundance levels differ between study groups is determined (e.g. cases vs controls; pre-treatment vs post-treatment). Two values are computed, corresponding to percent of transcripts increased and percent of transcripts decreased. Cutoffs employed to determine change can be adjusted based on study design and level of tolerance for false positives or negatives chosen by the user. For instance, for group comparisons cutoffs can be based both on statistics, fold changes and/or differences with or without multiple testing correction (e.g. p-value <0.01, FC = 1.5, Diff = 50, FDR = 0.1).

Comparisons can also be made at the individual subject level (e.g. one case vs controls). When comparing an individual sample to a control group a combined fold change and difference cutoff can be used (e.g. FC = 1.5, Diff = 50). Alternatively, the cutoff can be adjusted based on variance observed for each individual module among the control group samples (cutoff = means of control ± 2SD).

Fingerprint grid plot visualization

Differential expression at the module level can be displayed as a “fingerprint”, where the percentage of differentially expressed genes for a given module is represented by either a red spot or a blue spot, indicating increased abundance and decreased abundance for the constitutive transcripts, respectively (**Figure 4**). Each module is assigned a fixed position on a grid plot (coordinate on the grid; i.e. rows and columns). The number and intensity of the spots may

denote quantitative differences, sometimes correlated for instance with disease severity. Differences in distribution of the spots on the grid and their color may denote qualitative differences.

Such fingerprint grid plots were generated along with each of the two previous versions of blood transcriptome module repertoires. In those earlier versions the modules were arranged based on the order in which they were identified when running the script for module construction, which is based on weight and module size. However, for the third version presented here modules are arranged on the grid instead based on similarities in transcriptional patterns across the 16 input datasets. This is described in more details in the materials and methods section. A consequence of adopting this approach is that each row on the grid is constituted by modules for which changes in expression levels is often coordinated (**Figure 4**, module grid on the right). It means that, when mapping changes for a given disease, modules on the same row tend to follow the same trend (increase or decrease), which makes the fingerprint easier to read. It also means that a certain degree of functional convergence can be found for a given row of modules. This is for instance the case of row A28, which comprises 6 distinct "interferon modules".

In the example provided in Figure 4 transcriptome profiles of 55 pediatric patients with SLE and 14 control subjects are compared. As was previously reported an interferon signature dominated the response (A28), and was accompanied by modules associated with cell cycle (A27 and A29, including antibody production). Increase in abundance levels of modules associated with inflammation and neutrophils, another hallmark of the lupus transcriptome signature, was also observed (A35). These changes were also accompanied by decrease in transcript abundance, which was more apparent for some modules that belong to A1, A2 and A3. More specifically,

under A1, the most marked decreases were observed for modules which the functional map associates with protein synthesis (dark purple color, at positions 1, 5, 11 and 19 on row A1).

One may even go one step further and "aggregate" changes observed by row, further reducing dimension for a given dataset from 382 modules to 27 "aggregates" (**Figure 4**, module aggregates on the left). While employing the simplest framework possible would generally be best, our earlier work shows that distinct interferon modules are biologically and clinically meaningful [21]. Whether to work at the module-level or aggregate-level would depend then on the desired level of resolution.

Module grid plots are provided for each disease / physiological state in a supplementary file (**Supplementary file 3**). Six such module fingerprints representative of the range of signatures observed are shown in **Figure 5**. Blood transcriptome perturbations were for instance most widespread in the case of both MS and *S. aureus* infection (top panels) but with opposite patterns of changes, which will be even more apparent when profiles are compared directly across all 16 states (**Figure 6**). Changes associated with COPD or stage IV melanoma (middle panels) were most subtle but nonetheless distinct, with differences in abundance vs controls subjects most visible for aggregates A24 through A26 (Oxydative phosphorylation, Monocytes, Inflammation), and A36 through A38 ("erythrocytes", "neutrophil activation"). In the case of SLE and TB (bottom panels) interferon signatures constituted a common trait (A28) but with at the same time opposite patterns observed for an aggregate which was directly adjacent (A29: cell cycles). Other subtler differences in intensity of sets of modules associated with inflammation were also observed between these two diseases (A33-A35).

Heatmap visualization

A more traditional heatmap visualization can also be employed to represent module-level or module aggregate-level data. One configuration has module aggregates set as rows and

disease/physiological states as columns (**Figure 5**). Hierarchical clustering can then be used to arrange states on the heatmap based on similarities in abundance profiles across module aggregates. Such a heatmap permits to explore high-level similarities in blood transcriptome profiles across all 16 states. The first order of separation groups in one cluster a rather unexpected set of diseases, which are: acute HIV infection, Multiple sclerosis, Juvenile Dermatomyositis and COPD. All remaining 14 states are grouped in a second cluster, with RSV figuring as an outlier. The main trend driving the dichotomy between the first four diseases and the rest was an overall suppression of modules associated with inflammation / myeloid cell responses (A34-A38), accompanied by an increase in modules corresponding to aggregates A1 through A8 which are in part associated with lymphocytic responses. The factors underlying these two distinct "overarching" signatures are unclear. Diseases belonging to either group can display marked interferon signatures (e.g. acute HIV infection on one hand and SLE or Influenza infection on the other). The dichotomy does not appear to run along the traditional Th1/Th2 paradigm either, nor does it seem to reflect organ involvement.

Another heatmap configuration consists in arranging disease/physiological states as rows and modules belonging to a given aggregate as columns (**Figure 7**). In the example shown on figure 6, modules constituting the A28 aggregate were used. These modules are annotated functionally as "interferon modules". In this case, as could be expected, the dichotomy obtained separates diseases or states in which interferon signatures are present (all infectious diseases, with the notable exception of *S aureus* sepsis, systemic autoimmune/autoinflmammatory diseases such as SLE, SoJIA and liver transplant recipients under immunosuppressive therapy), from those in which interferon signatures are absent (JDM, Kawasaki disease, B-cell deficiency) and even possibly repressed (COPD, Melanoma, Pregnancy and MS).

DISCUSSION:

Improvements compared to the two earlier version of published blood module repertoires include: 1) the expansion of the number and range of biological states included for module construction to 16, encompassing nearly 1000 individual transcriptome profiles; 2) the grouping modules on the fingerprint grid based on similarities in activity profiles across diseases. This later development allows on one hand accommodating for the larger number of modules identified in this version (384) and on the other adds another level of dimension reduction with the possibility to analyze and visualize blood transcriptome changes at the "module aggregate level" (27 aggregates).

Further improvements may of course already be envisioned for subsequent versions of blood modular repertoire, with for instance the use of blood transcriptome datasets generated via RNAsequencing as input, or the possible introduction of more robust clustering methods as a basis for the construction of weighted co-clustering networks. Another direction to explore next could also include development of more specialized repertoires, focusing for instance on a given spectrum of diseases (e.g. neurogenerative disorders, respiratory illnesses).

Compared to earlier versions of modular repertoires reusability should be facilitated by the availability of R Scripts which have been developed to perform module repertoire analyses and generate grid or heatmap fingerprint visualizations. These scripts will be available on GitHub and described in detail in a separate publication (in preparation, early draft to be deposited shortly in BioRxiv). These bioinformatics tools were recently used as a basis for a weeklong workshop organized at the Sorbonne Universite's Inflammation-Immunopathology-Biotherapy Department. Hands-on training activities included the analyses by participants of

several public blood transcriptome datasets generated using samples obtained from patients with RSV infection and control subjects. This exercise also permitted to compare aggregate-level modular RSV fingerprint obtained across the six different studies. Furthermore, the heterogeneity at the level of individual patients observed with and across RSV dataset was explored, leading to the characterization of several modular RSV signatures "endotypes".

Finally, this third generation of blood modular repertoires also served as a backbone for development of cost-effective assays that can be substituted for genome-wide screens in biomarker discovery and in immune phenotyping or monitoring (Altman et al. manuscript in preparation and to be submitted to BioRxiv). The so-called transcriptome fingerprint assays are based on down-selecting to the most representative genes (i.e. surrogate genes) within each module. This can be achieved via a purely unsupervised data driven methodology. Such assays will be able to recognize changes occurring at the global level (as the original modular repertoire), while ensuring practicality, and cost effectiveness in that it can be performed using sensitive 'meso-scale' profiling assay (interrogating tens or hundreds of transcripts) such as high-throughput PCR, or targeted RNA sequencing.

REFERENCES

1. Banchereau, R., et al., *Understanding Human Autoimmunity and Autoinflammation Through Transcriptomics.* Annu Rev Immunol, 2017. **35**: p. 337-370.
2. Chaussabel, D., *Assessment of immune status using blood transcriptomics and potential implications for global health.* Semin Immunol, 2015. **27**(1): p. 58-66.
3. Devaux, Y., *Transcriptome of blood cells as a reservoir of cardiovascular biomarkers.* Biochim Biophys Acta Mol Cell Res, 2017. **1864**(1): p. 209-216.
4. Sumitomo, S., et al., *Transcriptome analysis of peripheral blood from patients with rheumatoid arthritis: a systematic review.* Inflamm Regen, 2018. **38**: p. 21.

5. Crinier, A., et al., *High-Dimensional Single-Cell Analysis Identifies Organ-Specific Signatures and Conserved NK Cell Subsets in Humans and Mice.* Immunity, 2018. **49**(5): p. 971-986 e5.

6. Villani, A.C., et al., *Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors.* Science, 2017. **356**(6335).

7. Alsina, L., et al., *A narrow repertoire of transcriptional modules responsive to pyogenic bacteria is impaired in patients carrying loss-of-function mutations in MYD88 or IRAK4.* Nat Immunol, 2014. **15**(12): p. 1134-42.

8. Cepika, A.M., et al., *A multidimensional blood stimulation assay reveals immune alterations underlying systemic juvenile idiopathic arthritis.* J Exp Med, 2017. **214**(11): p. 3449-3466.

9. Narayan, R., *Encyclopedia of Biomedical Engineering*. 2018: Elsevier.

10. van Dam, S., et al., *Gene co-expression analysis for functional classification and gene-disease predictions.* Brief Bioinform, 2018. **19**(4): p. 575-592.

11. Zhang, B. and S. Horvath, *A general framework for weighted gene co-expression network analysis.* Stat Appl Genet Mol Biol, 2005. **4**: p. Article17.

12. Chaussabel, D., et al., *A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus.* Immunity, 2008. **29**(1): p. 150-64.

13. Chaussabel, D. and N. Baldwin, *Democratizing systems immunology with modular transcriptional repertoire analyses.* Nat Rev Immunol, 2014. **14**(4): p. 271-80.

14. Obermoser, G., et al., *Systems scale interactive exploration reveals quantitative and qualitative differences in response to influenza and pneumococcal vaccines.* Immunity, 2013. **38**(4): p. 831-44.

15. Sugar, C.A. and G.M. James, *Finding the Number of Clusters in a Dataset.* Journal of the American Statistical Association, 2003. **98**(463): p. 750-763.

16. Chaussabel, D. and N. Baldwin, *Democratizing systems immunology with modular transcriptional repertoire analyses.* Nat Rev Immunol, 2014. **14**(4): p. 271-280.

17. Huang, D.W., B.T. Sherman, and R.A. Lempicki, *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.* Nucleic Acids Research, 2008. **37**(1): p. 1-13.

18. Huang, D.W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nat Protoc, 2008. **4**(1): p. 44-57.

19. Li, S., et al., *Molecular signatures of antibody responses derived from a systems biology study of five human vaccines.* Nat Immunol, 2014. **15**(2): p. 195-204.

20. Febbo, P.G., et al., *Literature Lab: a method of automated literature interrogation to infer biology from microarray analysis.* BMC Genomics, 2007. **8**: p. 461.

21. Chiche, L., et al., *Modular transcriptional repertoire analyses of adults with systemic lupus erythematosus reveal distinct type I and type II interferon signatures.* Arthritis Rheumatol, 2014. **66**(6): p. 1583-95.

ABBREVIATIONS

BIIR: Baylor Institute for Immunology Research

COPD: Chronic Obstructive Pulmonary Disease

DAVID: Database for Annotation Visualization and Integrated Discovery

FC: Fold Change

FDR: False Discovery Rate

GO_BP: Gene Ontology Biologic Process

GO_MF: Gene Ontology Molecular Function

HIV: Human Immunodeficiency Virus

IFN: Interferon

IRB: Institutional Review Board

KEGG: Kyoto Encyclopedia of Genes and Genomes

MS: Multiple Sclerosis

PCR: Polymerase Chain Reaction

PID: Primary Immune Deficiency

ROC: Receiver Operating Characteristic

RSV: Respiratory Syncytial Virus

SLE: Systemic Lupus Erythematosus

SOJIA: Systemic onset Juvenile Idiopathic Arthritis

TB: Tuberculosis

TBD: To Be Determined

DECLARATIONS

Ethics approval and consent to participate

Each of the studies contributing samples for this manuscript was independently approved by the BIIR IRB (IRB #'s 009-240, 006-177, 002-197, 009-257, H-18029, HE-470506, 011-173).

AVAILABILITY OF DATA AND MATERIAL

Raw gene expression data was deposited in the Gene Expression Omnibus, https://www.ncbi.nlm.nih.gov/geo/, under the accession GSE100150.

COMPETING INTERESTS

The authors declare no conflicts of interest.

FUNDING

This project has been funded in part with Federal funds from the National Institutes of Health under contract number U01AI082110. Authors affiliated with Sidra Medicine, a member of the Qatar foundation for Education, Science and Community development are supported entirely by institutional funding. RJW is supported by wellcome (203135, 104803), NIH (U01AI115940) and Francis Crick Institute which receives funding from wellcome (FC10218), CR UK (FC10218) and UKRI (FC10218).

AUTHOR's CONTRIBUTIONS

Conceptualization: MA, DR, NB, DC. Data curation: NB. Visualization: MA, DR, DC. Analysis and interpretation: MA, DR, NB, EW, MG, BK, MT, DC. Resources: CQ, SP, PK, VG, LC, NJ, PL, TP, GK, ML, HR, AG, MB, CB, ML, RW, CG, GL, MC, JS, FK, AM, OR, KP, VP, JB.Writing of the first draft: DC. Funding acquisition: GK, KP, VP, OR, JB, DC. Methodology development: MA, DR, NB, EW. Writing – review & editing: MA, DR, NB, EW, MG, BK, MT, CQ, SP, PK, VG, LC, NJ, PL, TP, GK, ML, HR, AG, MB, CB, ML, RW, CG, GL, MC, JS, FK, AM, OR, KP, VP, JB, DC. The contributor's roles listed above follow the Contributor Roles

552 Taxonomy (CRediT) managed by The Consortia Advancing Standards in Research

553 Administration Information (CASRAI) (https://casrai.org/credit/).

554 **ACKNOWLEDGEMENTS**

555 We thank Quynh-Anh Nguyen, Kimberly O'Brien, Dimitry Popov, Michael Mason, and Cate

556 Speake for technical assistance.

FIGURE LEGENDS

Figure 1: Construction of weighted co-clustering networks. Weighted co-clustering networks are used as a basis for the construction of modular repertoires. A distinctive characteristic of such networks is that they factor in differences in co-expression across different "states" of the biological system. For the blood transcriptome these states would be different diseases or physiological phenotyes. The weighting of the network is illustrated on this figure. Under scenario A the genes are co-expressed in all three disease states. The weight attributed to the edges of the network on the right is three. Under scenario B and C, co-clustering only occurs in two or one of the disease states, resulting in attributions of weights or 2 and 1, respectively.

Figure 2: Reuse of successive generations of blood modular repertoires. Two successive sets of blood modular repertoires were published previously, the first one in 2008 and the second in 2013. Citations were surveyed to ascertain the extent to which such "frameworks" can be subsequently reused. On these circular plot distinctions are made between self and third party reuse. Results are further broken down by disease areas.

Figure 3: Overview of the module repertoire construction process. Briefly, the starting point for the construction of blood transcriptional module repertoire is a collection of transcriptome datasets. In this case 16 datasets spanning a wide range of immunological and physiological states were employed. First of all each dataset is independently clustered via k-means clustering. Next, gene co-clustering events are recorded in a table, where the entries indicate the number of datasets in which co-clustering was observed for a given gene pair. Subsequently the co-clustering table serves as input for the generation of a weighted co-clustering graph (as illustrated in Figure 1), where nodes represent genes and edges represent co-clustering events. The largest, most highly weighted subnetworks among a large network constituted of 15,132 nodes are identified mathematically and assigned a module ID. The genes constituting this module are removed from the selection pool and the process is repeated resulting in the selection of 382 modules constituted by 14,502 transcripts.

Figure 4: Fingerprint grid plot. Modules are attributed a fixed position on a grid. Increase in abundance of the transcripts constitutive of a given module is represented by a red spot. Decrease in abundance is represented by a blue spot. Modules arranged on a given row belong to a module aggregate (noted A1 to A38). Changes at the “aggregate-level” are represented by spots to the left of the grid next to the denomination for the corresponding aggregate. In addition a module annotation grid is provided below where a color key indicates functional associations attributed to some of the modules on the grid.

Figure 5: Fingerprint grid plots mapping transcriptome repertoire perturbations across six representative disease states. Each grid represents changes in abundance observed at the module-level in subject from different disease groups compared to their respective controls. Modules occupy a fixed position on the grid, with red spots indicating the proportion of transcripts constitutive of a given module for which abundance is significantly increased and blue spots indicating conversely the proportion of constitutive transcripts for which abundance is decreased.

Figure 6: Patterns of abundance of module aggregates across 16 disease or physiological states. Each column on this heat map corresponds to a "module aggregate", numbered A1 to A38 (minus A9-A14 and A19-A24 which each included only one module). Each row on the heatmap corresponds to one of the 16 datasets used for construction of the module framework. A red spot on the heatmap indicates an increase in abundance of transcripts comprising a given module cluster for a given disease or physiologic state. A blue spot indicates a decrease in abundance of transcripts. No color indicates no changes. Disease or physiological states were arranged based on similarity in patterns of aggregate activity via hierarchical clustering.

Figure 7: Patterns of abundance of the six interferon modules constituting aggregate A28 across 16 disease or physiological states. Each column on this heat map corresponds to one of 6 interferon modules constituting the module aggregate A28. Each row corresponds to one of the 16 datasets used for construction of the module framework. A red spot on the heatmap indicates an increase in abundance of transcripts comprising a given module cluster for a given disease or physiologic state. A blue spot indicates a decrease in abundance of transcripts. No color indicates no changes. Disease or physiological states and modules were arranged via hierarchical clustering` based on similarity in patterns of aggregate activity.

TABLES

Table 1: Comparison of the characteristics of the input datasets used for the construction of three consecutive generations of blood transcriptome module repertoires.

	Generation 1	Generation 2	Generation 3
Number of States	8	7	16
States			
Systemic onset juvenile idiopathic arthritis	X	X	X
Pediatric systemic lupus erythematosus	X	X	X
Juvenile dermatomyositis			X
Type 1 Diabetes	X		
Multiple sclerosis			X
Kawasaki disease			X
COPD			X
Tuberculosis		X	X
Burkholderia pseudomallei infection		X	X
Respiratory Syncytial Virus infection			X
Influenza virus infection	X		X
Human Immunodeficiency Virus infection		X	X
Escherichia coli infection	X		
Staphylococcus aureus infection	X		X
B-cell deficiency			X
Liver transplant recipients	X	X	X
Metastatic melanoma	X	X	X
Pregnancy			X
Number of Input Datasets	8	9	16
Number of individual Profiles	239	410	985
Sample source	PBMCs	Whole Blood	Whole Blood
Platform	Affymetrix U133A&B	Illumina Hu6 v2	Illumina HT12 v3.0
Rounds of module selection	3	8	15
Number of modules	28	260	382
Year published & reference	2008 [ref]	2013 [ref]	Current work

Table 2: Datasets used for module construction

Sixteen distinct datasets were used as input for module repertoire construction. Each dataset corresponds to a different condition or physiological state and comprises both cases and matched controls. Each dataset was processed as a single batch at the same facility with the data generated using Illumina HumanHT-12 v3.0 Gene Expression BeadChips. In total the collection comprises a total of 985 individual transcriptome profiles.

Dataset	Category	# Samples (Cases)	# Samples (Control)	# Samples (Total)

1 Staphylococcus aureus	Bacterial Infection	99	44	143
2 Burkholderia pseudomallei	Bacterial Infection	35	12	47
3 Tuberculosis	Bacterial Infection	23	11	34
4 Influenza	Viral Infection	25	14	39
5 RSV	Viral Infection	70	14	84
6 HIV	Viral Infection	28	35	63
7 Pediatric SLE	Autoimmune	55	14	69
8 Multiple Sclerosis	Autoimmune	34	22	56
9 Juvenile Dermatomyositis	Autoimmune	40	9	49
10 Kawasaki disease	Autoinflammatory	21	23	44
11 Systemic Onset Idiopathic Arthritis	Autoinflammatory	62	23	85
12 COPD	Inflammatory	19	24	43
13 Melanoma	Malignancy	22	5	27
14 Pregnancy	Physiologic variant	25	20	45
15 Liver transplant recipients	Immunosuppressed	94	30	124
16 B-cell deficiency	Immunodeficiency	20	13	33

Supplementary File 1: Module generation and pseudocode

This word document describes the algorithm employed for the construction of this modular repertoire framework along with pseudocode which may be used as a basis for implementation in programming languages such as R or Python.

Supplementary File 2: Annotated module repertoire framework

This excel spreadsheet lists the 382 modules constituting this third generation of blood transcriptome module repertoires. Included are number of genes, list of member genes by symbol and probe ID, summarized functional annotations.

635 **Supplementary File 3: Module fingerprint grids of all 16 pathological and physiological**

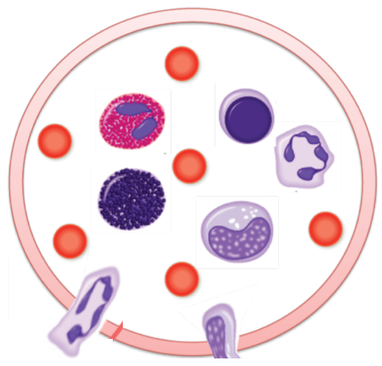
636 **states**

637 This PDF documents contains the modular fingerprints generated for each of the 16 input

638 datasets.

639

SYSTEM: **Blood transcriptome**



Multiple unweighted co-clustering networks

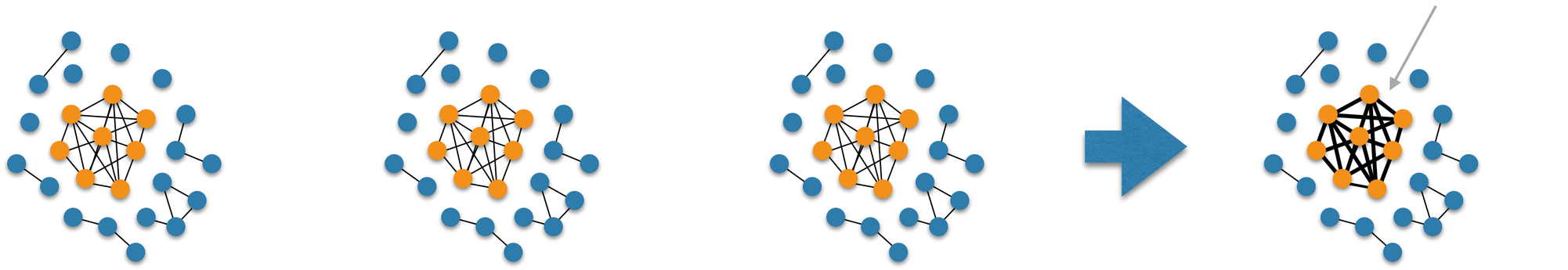
One weighted co-clustering network

STATES:

Disease 1	*Disease 2*	Disease 3

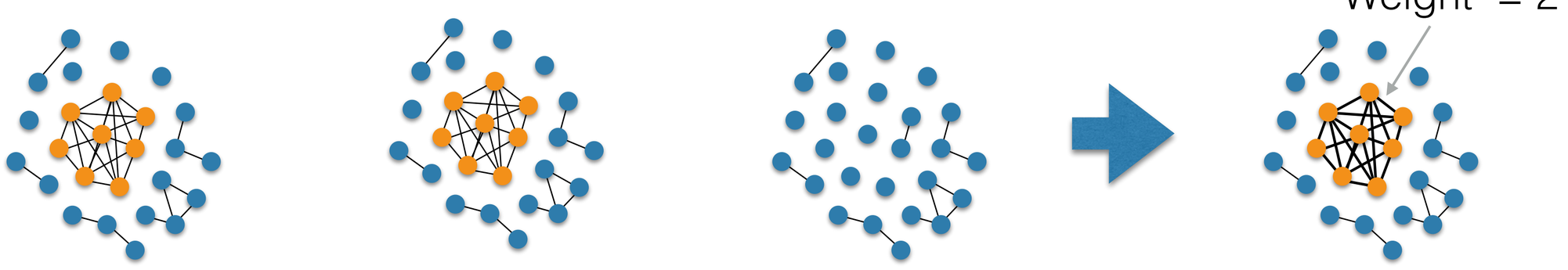
Scenario A

Co-expression in three “states”



Scenario B

Co-expression in two states



Scenario C

Co-expression in one state

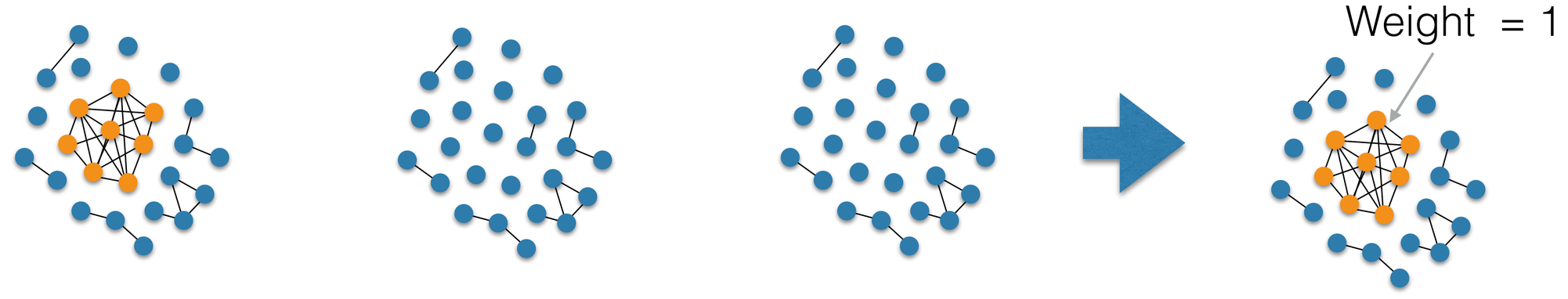
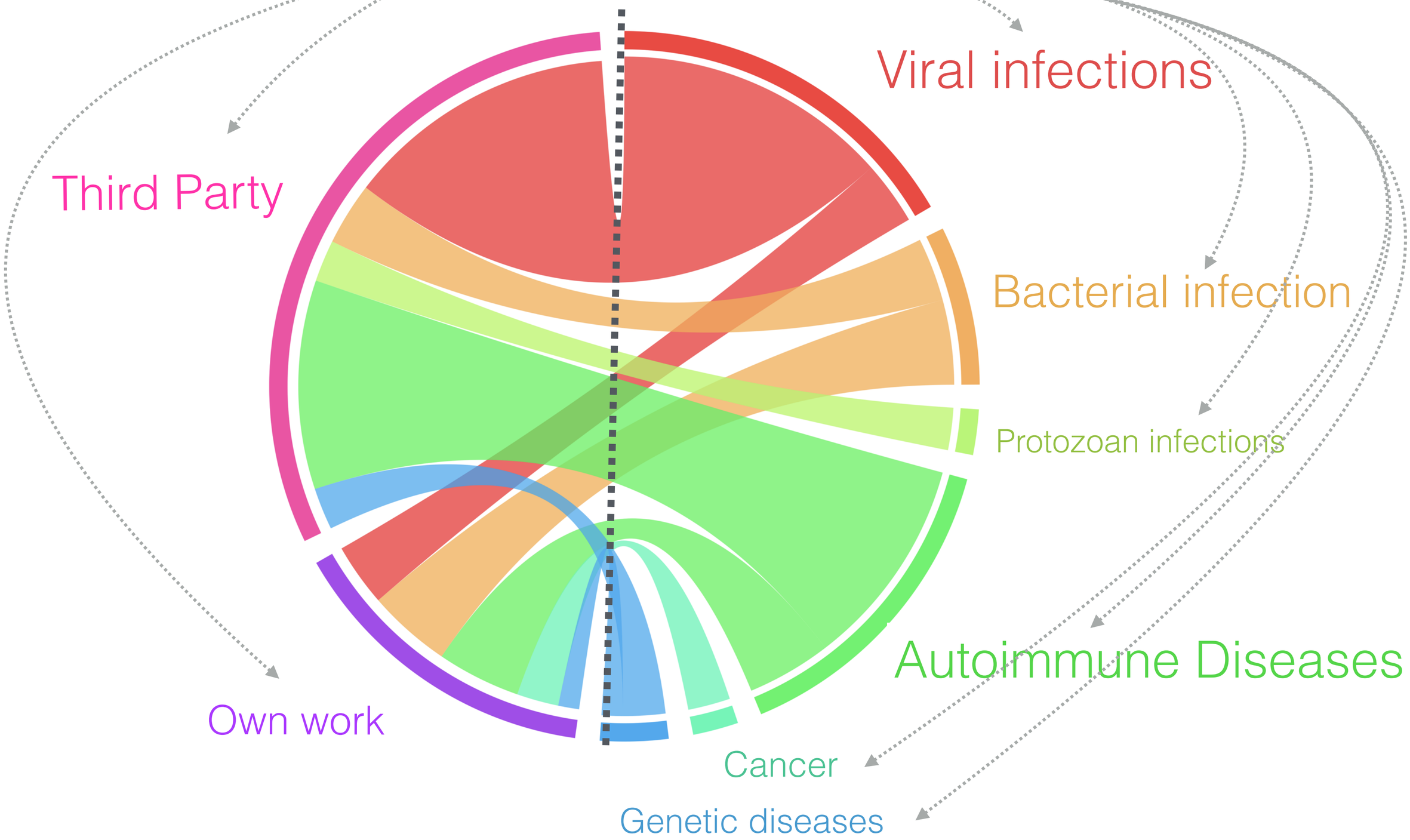


Figure 1

Blood transcriptome repertoire - v1 (2008)

N=55 publications



Blood transcriptome repertoire - v2 (2013)

N=56 publications

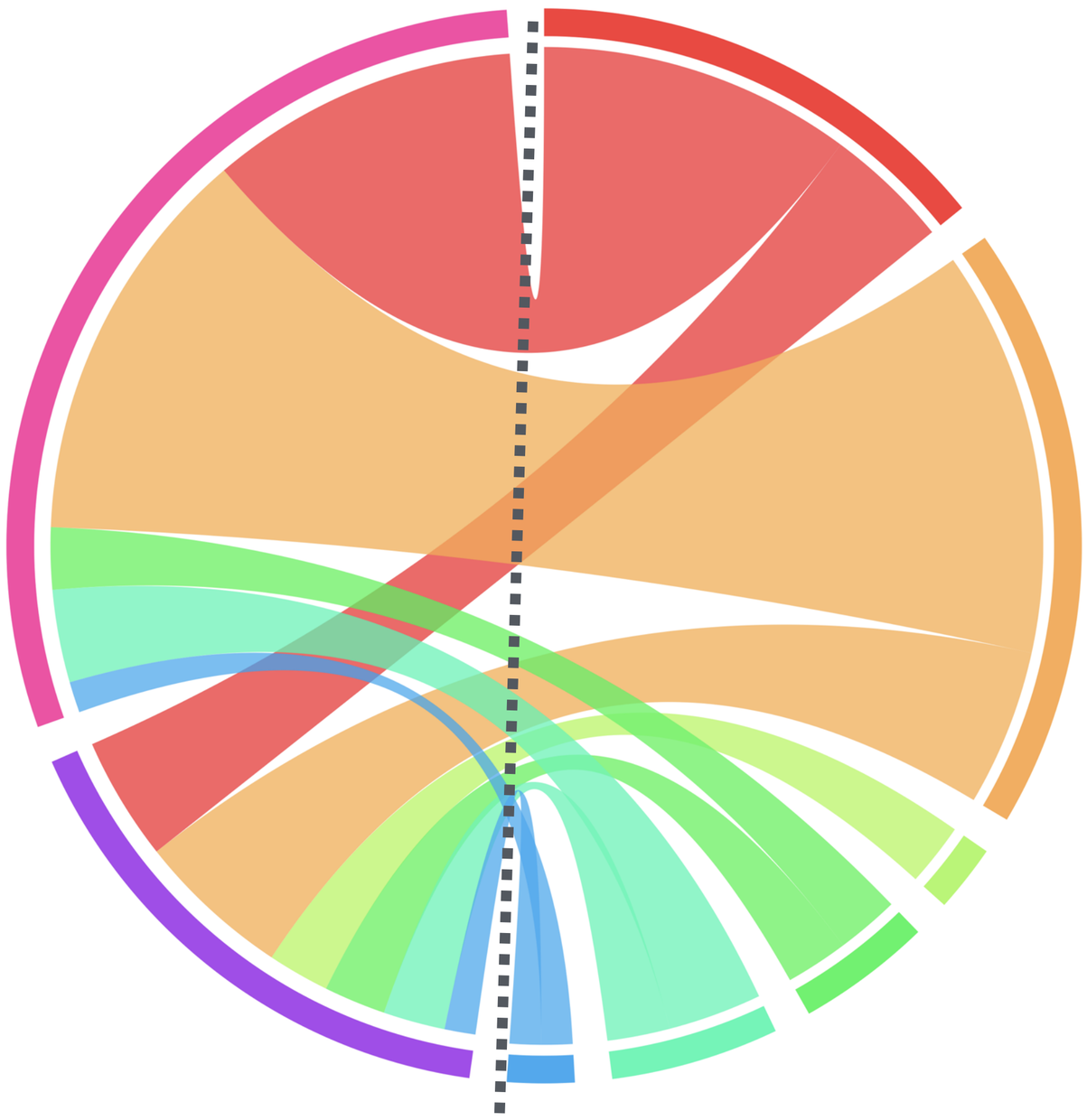
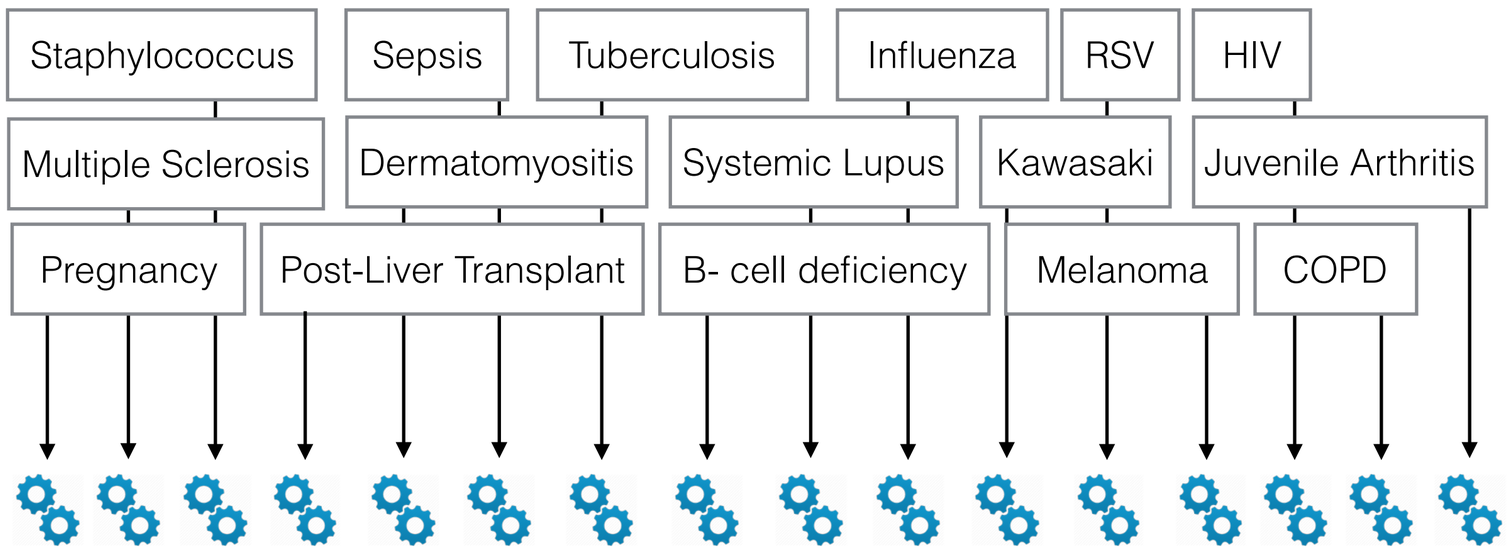


Figure 2

Input datasets

16 diseases / physiological states

985 Samples



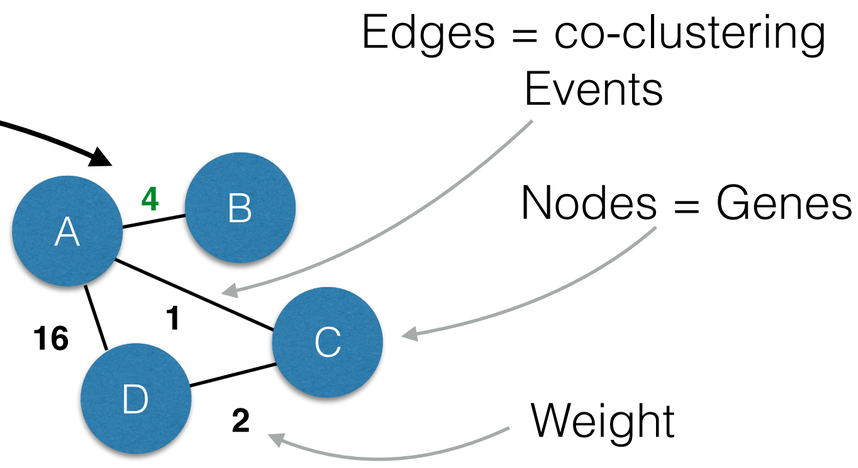
K-means Clustering x 16

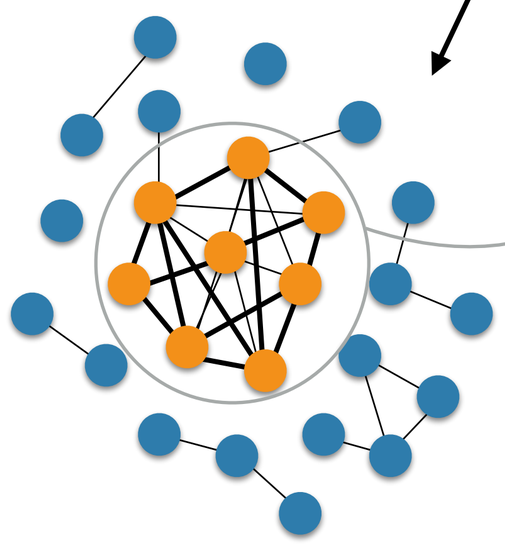
In how many datasets does Gene A clusters with Gene B?

✓ ✗ ✗ ✓ ✗ ✗ ✓ ✗ ✗ ✗ ✗ ✓ ✗ ✗ ✗ ✗

Co-Clustering Table

Genes	A	B	C	D	…
A		4	1	16	
B	4		0	0	
C	1	0		2	
D	16	0	2		
…					





Selection of densely connected sub-networks = modules

Weighted Co-Clustering Network

382 Module repertoire (14,502 transcripts)

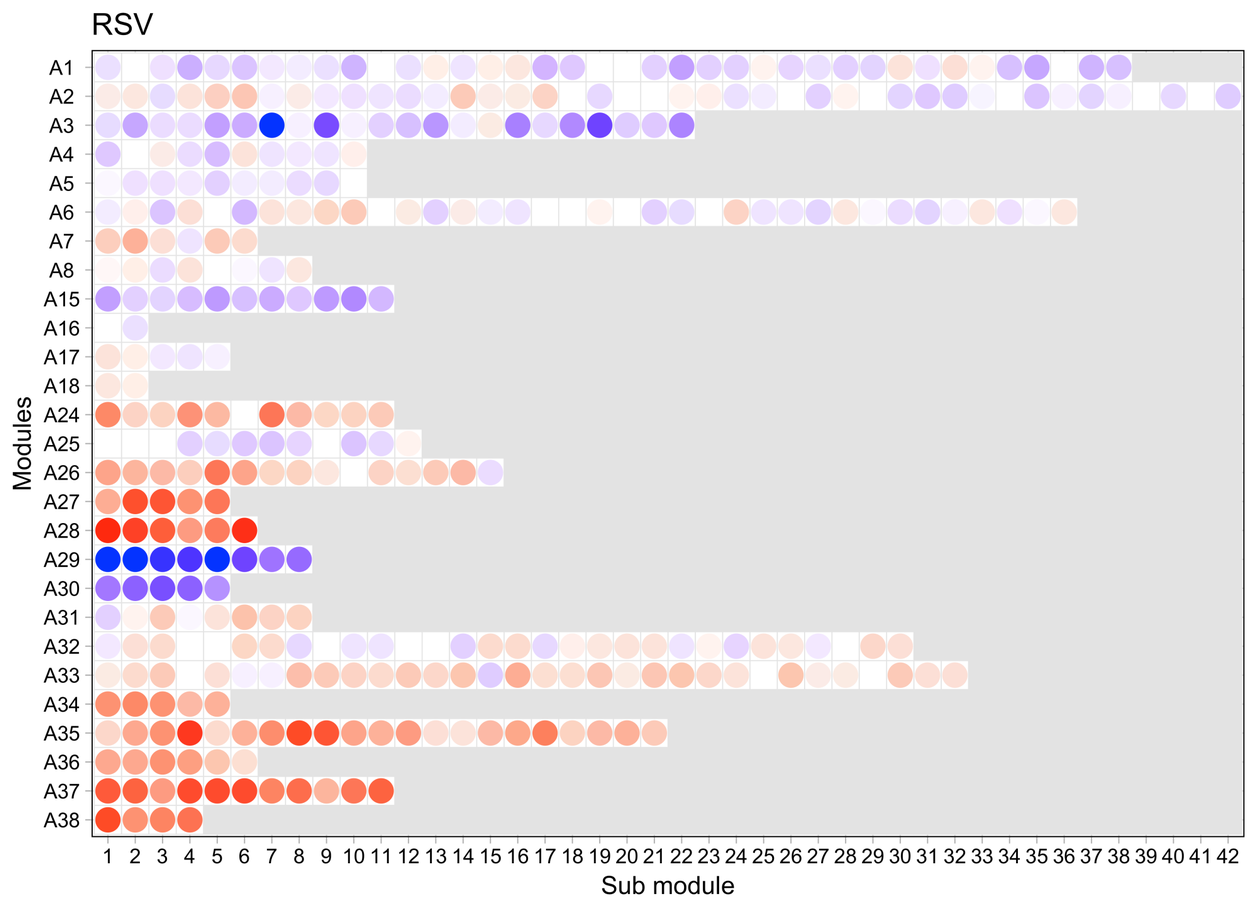
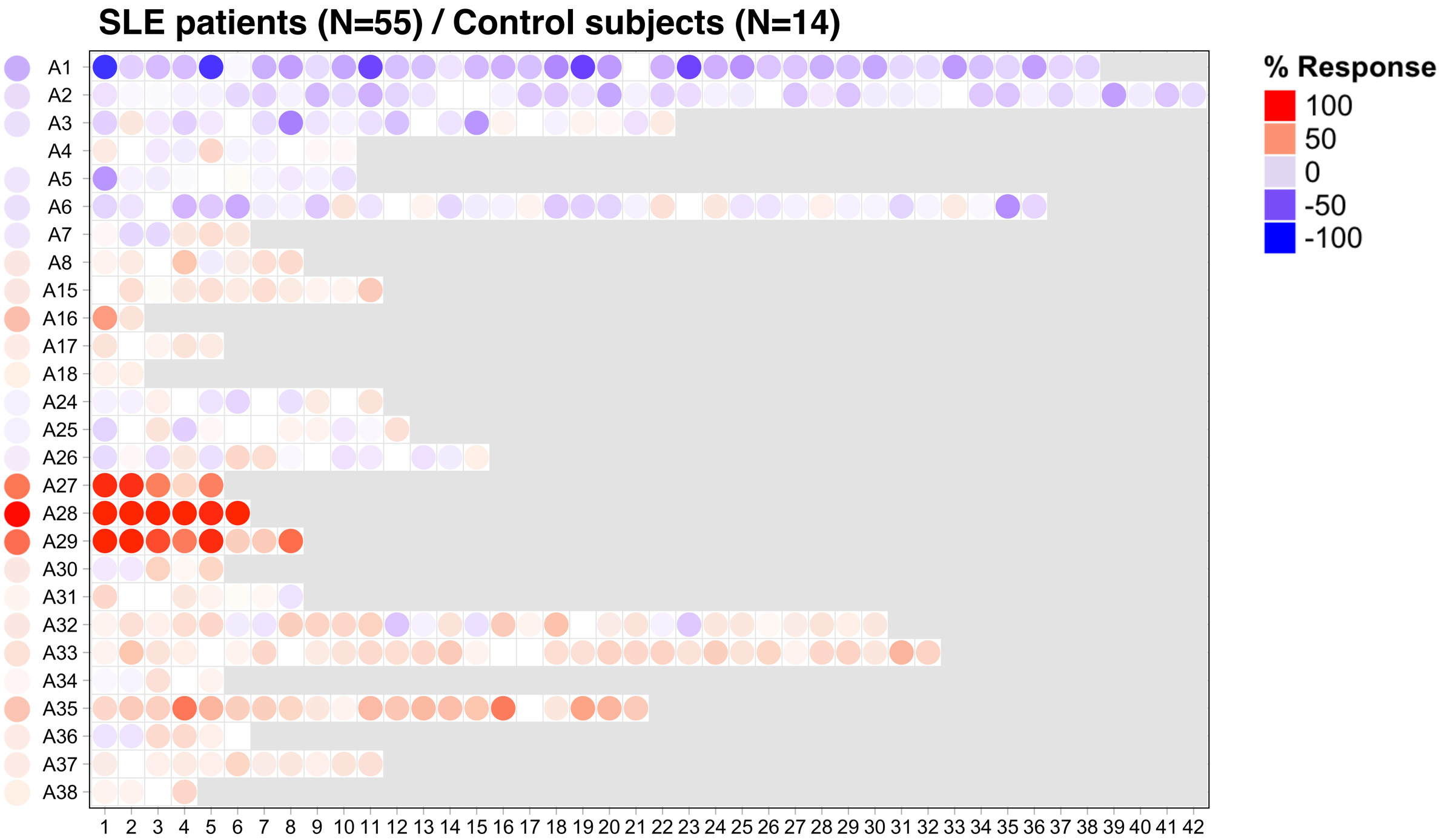


Figure 3



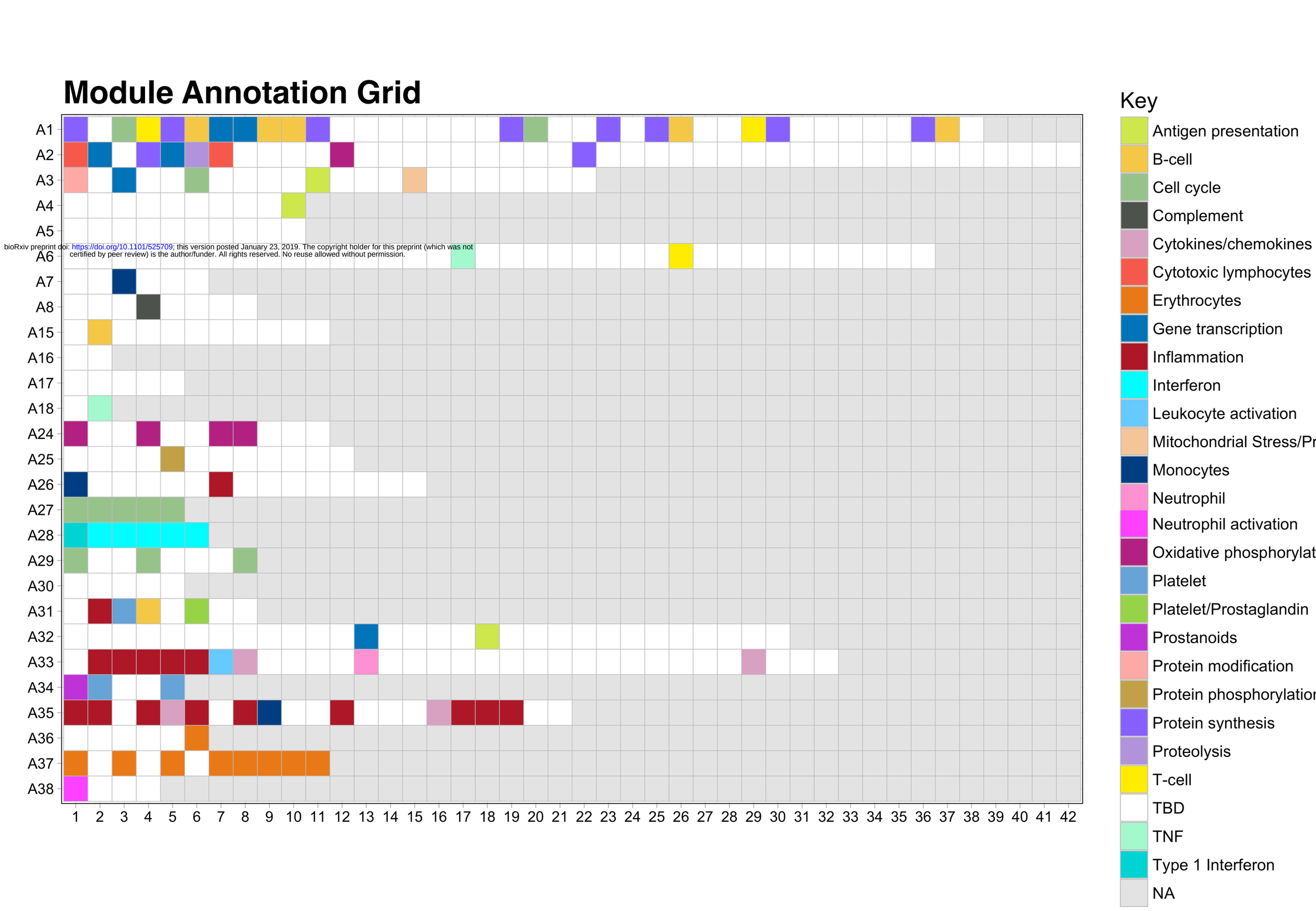
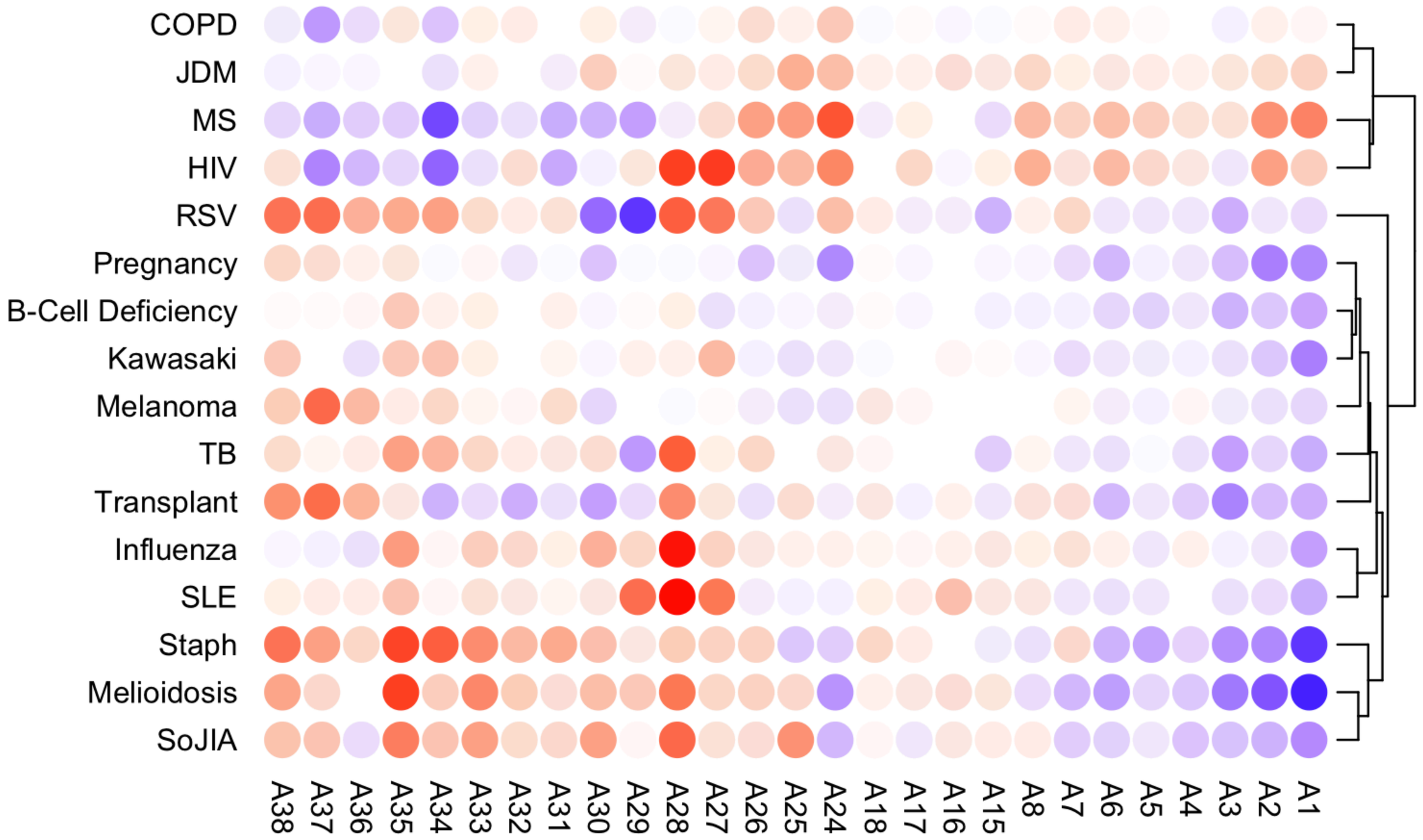


Figure 4

Figure 5



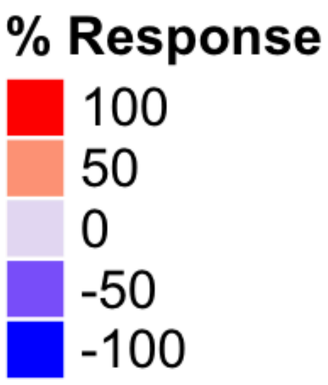
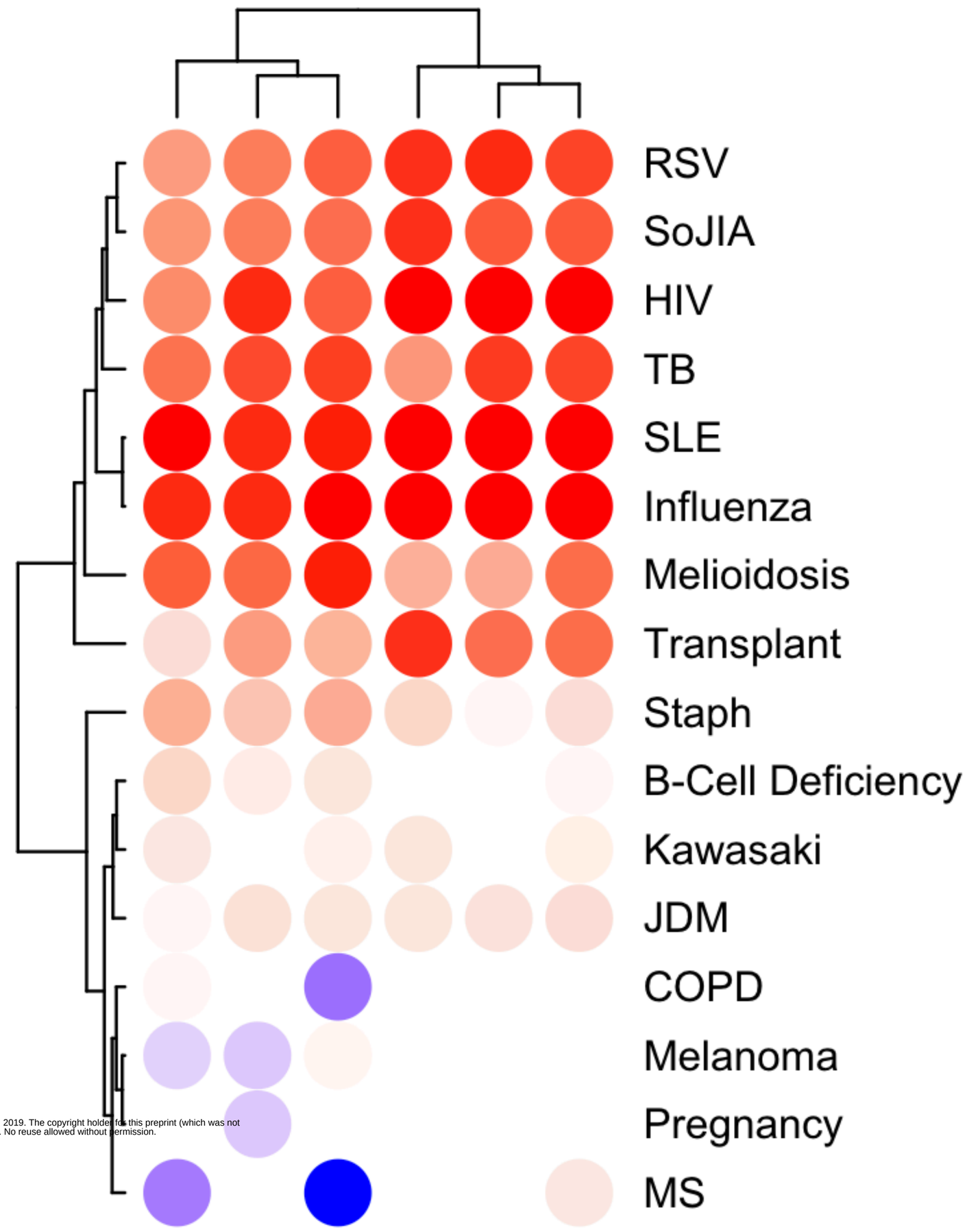


Figure 6



A28 modules

Figure 7