**TITLE: Development and Characterization of a Fixed Repertoire of Blood Transcriptome Modules Based on Co-expression Patterns Across Immunological States**

**AUTHORS/AFFILIATIONS:**

Matthew C Altman[1,2,#*], Darawan Rinchai[3#*], Nicole Baldwin[4], Mohammed Toufiq[3], Elizabeth Whalen[1] , Mathieu Garand[3], Basirudeen Ahamed Kabeer[3], Mohamed Alfaki[3], Scott Presnell[1], Prasong Khaenam[1], Aaron Ayllon Benitez[5], Fleur Mougin[5], Patricia Thébault[6], Laurent Chiche[7], Noemie Jourde-Chiche[8], J Theodore Phillips[4], Goran Klintmalm[4], Anne O'Garra[9,10], Matthew Berry[11], Chloe Bloom[10], Robert J Wilkinson[12,13,14], Christine M Graham[9], Marc Lipman[15], Ganjana Lertmemongkolchai[16], Davide Bedognetti[3], Rodolphe Thiebaut[5], Farrah Kheradmand[17], Asuncion Mejias[18], Octavio Ramilo[18], Karolina Palucka[4,19], Virginia Pascual[4,20], Jacques Banchereau[4,19], 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 *Research Branch, Sidra Medicine, Doha, Qatar*

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

5 *Inserm U1219 Bordeaux Population Health Research Center, Bordeaux University, Bordeaux, France*

6 *LaBRI, CNRS UMR5800, Bordeaux University, Bordeaux, France*

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

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

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

23 10 *National Heart and Lung Institute, Imperial College London, London, UK*

24 11 *Royal Cornwall Hospitals NHS Trust, Truro, UK*

25 12 *The Francis Crick Institute, London, UK*

26 13 *Department of Infectious Disease, Imperial College, London, UK*

27 14 *Wellcome Center for Infectious Diseases Research in Africa and Department of Medicine,*

28 *Institute of Infectious Diseases and Molecular Medicine, University of Cape Town Observatory*

29 *7925, Republic of South Africa*

30 15 *UCL Respiratory, Division of Medicine, University College London, London, UK*

31 16 *Centre for Research and Development of Medical Diagnostic Laboratories, Faculty of*

32 *Associated Medical Sciences, Khon Kaen University, Khon Kaen, Thailand*

33 17 *Baylor College of Medicine, Houston, Texas, USA*

34 18 *Abigail Wexner Research Institute at Nationwide Children's Hospital and the Ohio State*

35 *University School of Medicine, Columbus, Ohio, USA*

36 19 *The Jackson Laboratory for Genomic Medicine, Farmington, Connecticut, USA*

37 20 *Weill Cornell Medicine, New York, New York, USA*

38

39 # Equally contributing authors

40 *To whom correspondence may be addressed:

41

42

43

44

45 **CONTACT INFORMATION:**

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

49 Darawan Rinchai, Cancer Program, Sidra Medicine, PO Box 26999 Al Luqta Street, Doha,
50 Qatar. Tel. +974 4003 7413, E-mail: drinchai@sidra.org

51 Damien Chaussabel, PhD, Immunology Program**,** Sidra Medicine, PO Box 26999 Al Luqta
52 Street, Doha, Qatar. Tel. +974 4003 7611, E-mail: dchaussabel@sidra.org

53

54

55

56

57

58

59

60

61

**SUMMARY**

As the capacity for generating large scale data continues to grow the ability to extract meaningful biological knowledge from it remains a limitation. Here we describe the development of a new fixed repertoire of transcriptional modules. It is meant to serve as a stable reusable framework for the analysis and interpretation of blood transcriptome profiling data. It is supported by customized resources, which include analysis workflows, fingerprint grid plots data visualizations, interactive web applications providing access to a vast number of module-specific functional profiling reports, reference transcriptional profiles and give users the ability to visualize of changes in transcript abundance across the modular repertoire at different granularity levels. A use case focusing on a set of six modules comprising interferon-inducible genes is also provided. Altogether we hope that this resource will also serve as a framework for improving over time our collective understanding of the immunobiology underlying blood transcriptome profiling data.

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

85 **INTRODUCTION:**

86 Technology advancements over the past two decades paired with improvements in cost-
87 effectiveness have greatly increased the capacity to conduct large-scale molecular profiling in
88 biomedical research. In translational settings, these advances have enabled the measurement of
89 molecular phenotypes at very high resolutions, reaching the point where it has become possible to
90 routinely generate whole genome, proteome, metabolome, microbiome and transcriptome profiles
91 on an almost routine basis (1). The current bottleneck is in interpreting the biological significance
92 or potential clinical implications of these molecular signatures.

93 Our group has developed co-expression modules and used them as a framework to analyze
94 and interpret blood transcriptome data (2–4). Importantly, we have used the modules as a fixed
95 framework: transcriptional modules are not generated each time a new dataset needs to be
96 analyzed; instead, the same set of pre-existing modules is reused to analyze each new dataset.
97 Consequently, our team has released only two different module repertoires over a 12-year period,
98 which we and others have used to analyze numerous blood transcriptome datasets [e.g. (5–8)].
99 Other blood transcriptome frameworks have been generated (9), as well as those with a more
100 generic transcriptional module repertoire (10). An inherent benefit of using a collection of gene
101 sets is that it permits the reduction of high-dimensional data. Using a fixed repertoire of modules
102 based on co-expression analyses across a large dataset collection, improves the robustness of
103 analyses of small datasets (10). As such, efforts can rather focus on the functional interpretation
104 of the collection of gene sets constituting the repertoire. Furthermore, using the same fixed
105 framework to analyze multiple independent datasets means that cross-study comparisons and
106 interpretation are easier and more reliable. With the construction of this new repertoire we aimed
107 first at increasing the range of immunological states upon which the repertoire is based, which was

limited to 7 in our earlier attempt. Secondly, considerable efforts were deployed to improve resources to support analysis and interpretation.

Thus, the range of disease and physiological states employed for the construction of this new modular repertoire has been expended considerably compared to the first and second generation repertoires. Specifically, 16 input datasets were employed, comprising 985 unique blood transcriptome profiles from: patients with autoimmune, infectious or inflammatory diseases; cancer patients; liver transplant recipients; and pregnant mothers. We describe here how we developed the analysis workflows for group comparison and individual molecular "fingerprinting". Ad hoc data visualization strategies are presented as well, namely the use of module fingerprint grids and heatmaps to represent the data. We also provide access to extensive functional annotations for each of the 382 modules comprising the repertoire via an interactive web application. Reference blood transcriptome fingerprints are also made available via a web application that enables users to explore the changes observed across diseases/studies and among individuals. A focused effort aiming at the characterization of a set of six interferon modules is presented to illustrate the manner in which such resources may be deployed to help unravel the biology underlying blood transcriptome signatures.

131 **RESULTS**

132 **Constitution of a collection of datasets covering a wide range of immune states.**

133 The development of transcriptional module repertoires relies on identifying gene co-expression
134 events using transcriptome profiling data as a starting point. For this new blood transcriptome
135 module repertoire, we used 16 datasets (GEO ID: GSE100150) that encompassed 985 individual
136 whole blood transcriptome profiles. Each dataset corresponds to a different pathological or
137 physiological state (**Table 1**). These datasets were processed in the same facility and run on
138 Illumina HT12 BeadArrays (details are provided in the method section). Similar to our first two
139 repertoires (**Table 2**), we included data from patients (adult and pediatric) with: systemic lupus
140 erythematosus (SLE), systemic onset juvenile idiopathic arthritis (SoJIA), liver transplants and
141 receiving maintenance immunosuppressive therapy, metastatic melanoma, and infectious diseases
142 [with an expanded range that now includes infections caused by influenza, respiratory syncytial
143 virus (RSV), human immunodeficiency virus (HIV) infections, *Mycobacterium tuberculosis,*
144 *Staphylococcus aureus*, and *Burkholderia pseudomallei (the agent of Melioidosis) and sepsis*
145 *caused by other bacteria (Streptococcus pneumoniae, Salmonella spp, Pseudomonas*
146 *aeruginosa)*]. We also added six new conditions to our framework: inflammatory conditions of the
147 skin (juvenile dermatomyositis), lung [chronic obstructive pulmonary disease (COPD)] and
148 circulation (Kawasaki disease); multiple sclerosis (MS); primary immune (B-cell) deficiency; and
149 pregnancy (serving as a physiological variant). Absent from this repertoire (but represented earlier)
150 are patients with type 1 diabetes or with an *Escherichia coli* infection.

151 In summary, the sixteen datasets that have been assembled capture a wide breadth of
152 immunological responses. This should permit the construction of a transcriptional module
153 repertoire that will prove useful as a generic framework for blood transcriptome data analyses.

154 **Implementation of a step-wise approach to blood transcriptional module repertoire**
155 **construction.**

156 After collating the 16 input datasets, we next followed a stepwise process to construct a module
157 repertoire and identify co-expression networks (**Figure 1**). We used the module construction
158 algorithm that we have previously implemented (the code is provided in **Supplementary File 1**).
159 This approach followed four main steps: 1) clustering for each individual dataset; 2) co-clustering,
160 where the number of instances that two genes were included in the same cluster was recorded, with
161 the values ranging from 0 to 16 (i.e. co-clustering observed in none or all of the datasets,
162 respectively); 3) constructing a weighted co-expression network, where the edges between the
163 genes represent at least one co-clustering event in one of the input datasets and the weight is
164 assigned based on the total number of co-clustering events; (**supplementary Figure 1**); and 4)
165 mining the resulting network to identify highly inter-connected sub-networks that form the
166 modules.

167 Constituting a module repertoire in this manner is thus entirely data-driven: it does not rely
168 on any *a priori* information about gene interactions or functions. In total, we identified 382
169 modules comprising 14,168 transcripts (95.8% of the transcripts detected in this dataset
170 collection).

171

172 **Development of module-level analysis workflows and visualizations**

173 A key characteristic of the gene sets collected via the process described above is that, by
174 construction, changes in abundance of the corresponding transcripts within a given module will
175 tend to be coordinated. As such, it should be possible to use these modules as a "framework" to:
176 1) identify functional convergences among the genes that comprise each set, and 2) summarize

177 changes in overall transcript abundance related to pathological processes or therapeutic
178 interventions.

179 We determined the gene composition of each of the 382 modules (**supplementary file 2)**,
180 The average number of genes per module as 37.1, the median as 26.5 and the range as 12-169.
181 Functional profiling and enrichment results were generated using multiple tools (GSAn, Literature
182 Lab, IPA, DAVID, KEGG, Biocarta, OMIM, and GOTERM). We also determined the extent of
183 overlap with the previous modular repertoire and a set of modules constituted at Emory University
184 and our previous generation of modules (9). For module-level analyses, we determined the
185 proportion of the constitutive transcripts that differ in abundance levels between study groups (e.g.
186 cases vs. controls; pre-treatment vs. post-treatment). By this approach, two values, corresponding
187 to the percent of transcripts that are (i) increased and (ii) decreased, are derived. The cutoff points
188 can be chosen based on user preferences. For example, cutoffs can be based on statistics, fold
189 changes and/or differences with or without multiple testing correction for group comparisons.

190 From here, the extent of differential expression at the module level can be displayed as a
191 "fingerprint", assigning each module to a fixed position on a grid plot and color-coding it according
192 to the level of increased or decreased abundance for the constituting transcripts (**Figure 2**). For
193 this we then performed a second tier of clustering to group the 382 modules into 38 "aggregates",
194 with each row on the grid displaying the modules corresponding to one such aggregate.
195 Segregation into distinct aggregates was based on similarities in abundance levels observed across
196 the collection of 16 datasets. By this approach, we derived two levels of granularity (i.e. module-
197 level vs. module aggregate-level) with the number of variables for interpretation constrained to a
198 more manageable number. The overall result is that changes in expression levels for each row on
199 the fingerprint grid will tend to be coordinated, which was not the case of prior iterations of such

200 fingerprint grids (**Figure 3**). Some degree of functional convergence can thus be observed within
201 a given row of modules. As an example, in our fingerprint we found that row A1 comprised several
202 modules associated with lymphocytes, while row A28 comprised six distinct "interferon modules"
203 and rows A33 and A35 comprised a number of modules functionally associated with inflammation.
204 Overall, fingerprint grid plots can complement traditional heatmap representations. With
205 the positions of modules on the grid being fixed it becomes for instance possible to rapidly identify
206 the key changes associated with a given pathology or immunological state.
207
208 **Illustrative use case of fingerprint grid plot representations**
209 We next illustrate the analysis and visualization approach described above with a use case. It will
210 primarily focus on SLE, a disease which blood transcriptome signature has been well-
211 characterized. Fingerprints of other reference disease cohorts employed for module construction
212 will be included to provide additional context.
213 As mentioned, data interpretation is facilitated by tiered dimension reduction: the first
214 vertical reading of the fingerprint grid permits visualization of changes across the aggregates,
215 while the horizontal reading permits visualization of changes within an aggregate and across
216 modules. As pointed out earlier, in the fingerprint grid, the number and intensity of the spots
217 represents qualitative as well as quantitative differences. As an example, we compared the
218 transcriptome profiles among 55 pediatric patients with SLE and 14 healthy control subjects
219 (**Figure 3**). We identified an interferon-dominated signature (A28) accompanied by modules
220 associated with cell cycle (A27 and A29, including antibody production). An increase in the
221 abundance of modules associated with inflammation and neutrophils (A35) — a hallmark of the
222 SLE transcriptome signature was also observed. These changes were accompanied by a decrease

223 in transcript abundance, which was more apparent for some modules belonging to aggregates A1,
224 A2 and A3 which are arrayed across the first three rows of the fingerprint grid. More specifically,
225 for the module aggregate A1, the most marked decreases were observed for modules associated
226 with protein synthesis (dark purple color at positions 1, 5, 11 and 19 on row A1).

227 It is possible to go a step further and "aggregate" the changes observed by row, thereby
228 reducing the dimensions for a given dataset even further. In this case, we reduced the dataset from
229 382 modules to 27 "aggregates" (**Figure 3A**). Users can decide taking this extra step depends on
230 the desired level of resolution: module-level or aggregate-level. For example, mapping changes at
231 the aggregate level, the simplest framework possible, is optimal when looking at signatures in a
232 broader context. But at the same time our earlier work showed that distinct interferon modules are
233 biologically and clinically meaningful, also it can also be indicated to work instead at the module-
234 level (11).

235 We generated fingerprint grid plots for each of the sixteen diseases or physiological states
236 (**Supplementary file 3**); six module fingerprints are shown in **Figure 3** as an illustration. In brief,
237 among these we found that blood transcriptome perturbations were most widespread in patients
238 with MS and patients with a *Staphylococcus aureus* infection (**Figure 3B**) with opposing patterns
239 of change. Changes associated with COPD or stage IV melanoma (**Figure 3C**) were most subtle
240 but nonetheless distinct, with differences in transcript abundance compared to control subjects
241 most visible for aggregates concerning oxidative phosphorylation, monocytes, inflammation
242 (A24-A26), erythrocytes and neutrophil activation (A36-A38). Interferon signatures were a
243 common trait (A28) in patients with SLE and patients with tuberculosis (TB) caused by *M.*
244 *tuberculosis* infection (**Figure 3D**), while opposite patterns were observed for the cell cycle (A29).

245 We also found differences in the intensities of sets of modules associated with inflammation
246 between these two diseases (A33-A35).
247 Taken together, this use case demonstrates the use of "fingerprint" representations.
248 Notably, fixing positions on the grid permits to overlay functional annotations associated with each
249 modular signature, and the use reference collections of fingerprints for comparative interpretation.
250 However, while this representation can be used as a complement, it altogether does not replace the
251 more traditional heatmap representations.
252
253 **In depth functional annotation of fixed transcriptional module repertoires**
254 Based on earlier iterations, the expectation is for this repertoire to be of use over a period spanning
255 at least 5-6 years. We thus put great effort into functionally annotating the repertoire. We followed
256 two main annotation approaches: (1) concurrently running ontologies, pathways or literature-term
257 profiling analyses; and (2) determining for each module the patterns of gene expression for select
258 reference transcriptome datasets. We compiled the resulting information for all 382 modules and
259 have since made it available via an interactive web application Here, it is possible to zoom in and
260 out, determine spatial relationships and interactively browse the very large compendium of
261 analysis reports and heatmaps generated as part of our annotation efforts. (**Figure 4**; Links for each
262 aggregate are listed in **Table 3**.; A demonstration video is available:
263 https://youtu.be/8ajESET2mqI) Below, we describe briefly how we conducted our annotations
264 (see also Experimental Procedures).

265 *Step 1 – Functional profiling:* We conducted gene ontology profiling for each of the 382 modules
266 using DAVID (12), GOTERM, and GSAn (13). GSAn interactive reports were uploaded to a
267 custom web portal (https://ayllonbe.github.io/modulesV3/index.html). We also performed

268 pathway enrichment analyses using KEGG, Biocarta and Ingenuity Pathway, as well as literature
269 term enrichment with Literature Lab. We synthesized and compiled the analyses (**Supplementary**
270 **File 2**) and used them to identify convergences and attribute functional titles to the different
271 modules. Functional titles could not be attributed in all cases, due to a lack of convergence or poor
272 enrichment in one or several of the analyses.

273 *Step 2 – Expression patterns in reference transcriptome datasets:* We used transcriptome datasets
274 as a reference to improve characterization and biological interpretation of the module framework.
275 Two different datasets were used. The first was contributed by Novershtern *et al.* and comprised
276 the transcriptome profiles of 38 human hematopoietic cell populations (14). The second was
277 contributed by Speake *et al.* and comprised the RNAseq profiles of six circulating leukocyte
278 populations from patients with various immune-associated diseases (15). We generated heatmaps
279 for each module to display the abundance patterns of the constitutive transcripts for each dataset.

280 Overall, this resource serves two purposes. First, it provides access to information that is
281 necessary for interpreting transcriptome fingerprints generated via module-level analyses. Second,
282 it helps us to improve the attribution of functional titles and roles to the different modules and
283 aggregates. Indeed, although the transcriptional module repertoire is fixed over time, we anticipate
284 that the functional annotations will continue to evolve over its lifespan.

285

286 **Measuring inter-individual variability for the molecular stratification of patient cohorts**

287 The analysis and visualization steps presented so far focused on characterizing differences between
288 groups of subjects (e.g. cases and controls). However, it is also important to characterize
289 heterogeneity among groups of patients since inter-individual variability can serve as a basis for
290 the definition of molecular phenotypes and patient stratification.

291 Within each module, and for each individual subject, we used fixed cutoffs to count the
292 number of transcripts that increase, decrease or do not change in abundance compared to a baseline
293 value (e.g. absolute fold change in expression and absolute difference in expression vs. average of
294 control samples). The percentage of differentially expressed genes for each module are then
295 computed. These percentages are equivalent to values derived from group comparisons, except
296 that they are derived for each individual sample.

297 The sepsis cohort comprised in the reference dataset collection was used to illustrate how
298 this approach can be employed for assessment of inter-individual variability for a given pathology
299 (**Figure 5**). Changes in transcript abundance across was found to be highly consistent across
300 patients for some module aggregates. This was for instance the case of aggregate A1 (broadly
301 associated with lymphocytic cells/responses), with consistent decreases in transcript abundance
302 observed across patients. Conversely, consistent increases were observed for modules comprising
303 aggregate A35 (broadly associated with inflammatory neutrophil responses). In this case
304 differences were observed in the intensity of the response. Other module aggregates, however,
305 showed more mixed responses, which was the case of A37 (erythroid cells), A33 (functional
306 association to be defined) or A28 (interferon response).

307 A web application was developed as a resource to explore the inter-individual differences
308 for a given disease, module aggregates or a combination of aggregates
309 (https://drinchai.shinyapps.io/dc_gen3_module_analysis/#; video: https://youtu.be/y__7xKJo5e4
310 ). This application permits the generation of fingerprint grid plots and heatmaps representing
311 module aggregates activity across the 16 reference datasets (**Figure 6**).

312 Altogether, as illustrated by the sepsis use case provided here, deriving individual modular
313 fingerprints can provide a means to achieve molecular stratification of patient cohorts. However,
314 the biological and clinical relevance of such distinct molecular phenotypes would still remain to
315 be determined in follow on analysis steps.

316 **Profiling the abundance of A28 interferon-inducible genes at the aggregate-level across**
317 **reference patient cohorts**

318 The earlier sections dealt first with the approach implemented for the construction and
319 characterization of the fixed transcriptional module repertoire. Analysis and visualization
320 strategies for both group-level and individual-level comparisons were presented next. Now we
321 would like to go over a use case focusing on the dissection of changes in abundance for module
322 aggregate A28 (interferon responses). We first start here from the highest possible perspective,
323 examining changes in abundance for the A28 aggregate across reference disease cohorts.

324 A heatmap was derived showing patterns of abundance of a subset of 27 module aggregates
325 comprising two or more modules across 16 health states (**Figure 7A**). The first order of separation
326 grouped acute HIV infection, MS, juvenile dermatomyositis and COPD in one cluster. The
327 remaining 14 states grouped into a second cluster, with RSV infection presenting as an outlier. The
328 main trend driving the dichotomy between the first four diseases and the rest was an overall
329 suppression of modules associated with inflammation and/or myeloid cell responses (A34-A38),
330 accompanied by an increase in modules corresponding to aggregates associated, in part, with
331 lymphocytic responses (A1-A8). The factors underlying these two distinct, "overarching"
332 signatures are unclear; diseases belonging to either group can exhibit marked interferon signatures
333 (e.g. acute HIV infection in one cluster, and SLE or influenza infection in the other cluster). The
334 circle plots shown on this figure illustrate at a more granular module and gene-level the changes

335 that are being represented by spots on this heatmap. The gene composition for each of the six
336 module is shown in one of those circle plots (**Figure 7B**). The other represent which of the genes
337 among those modules are changed when comparing patients with infection to their respective
338 control groups (**Figure 7C**). Finally, we added for reference circle plots showing patterns of *in*
339 *vivo* response of A28 genes to administration of IFNα to patients with hepatitis C infection or of
340 IFNβ to patients with multiple sclerosis [transcriptome profiling data were made publicly available
341 by the authors (16,17)] (**Figure 7D**). The plots on this figure show that changes observed at the
342 aggregate level are not always distributed evenly across the six modules constituting aggregate
343 A28, and in turn of genes constituting each of the modules. The response to type I interferons was
344 dominated by a disproportionate increase in abundance of transcripts constituting M8.3 and M10.1.
345 Transcripts forming M15.86, that showed very little changes in response to those treatments, were
346 on the other hand markedly increased during acute HIV and influenza infection (**Figure 7C**). It is
347 thus possible that this gene set be more specifically induced by interferon gamma. Interferon
348 responses were weaker among RSV patients compared to these two other viral illnesses. Similarly,
349 in the context of bacterial infection the interferon response was most marked in response to TB
350 infection, which is consistent with the literature (8). A large proportion of adult patients comprising
351 the sepsis cohorts were infected with *Burkholderia pseudomallei*, the intracellular bacteria
352 responsible for melioidosis that also tends to induce higher interferon responses (8). Changes in
353 abundance for transcripts constituting A28 in the context of autoimmune or inflammatory diseases
354 is shown in **Supplementary Figure 2**.

355 The visualization schemes employed here already affords some new perspectives about the
356 contribution of different modules and genes to the overall aggregate-level interferon responses.

357 However other variation in the selection of variables and granularity levels are possible and will
358 be explored next.

359

360 **Profiling the abundance of A28 interferon-inducible genes at the module-level across**
361 **reference patient cohorts.**

362 Another perspective was gained next by plotting changes of A28 genes across the same reference
363 cohorts but at the more granular module-level, rather than at the aggregate level. It was also at
364 this level that we chose to compare literature keyword enrichment profiles for each of the A28
365 module.

366 Functional enrichment analyses showed that all six modules in this aggregate were
367 associated with the interferon response (see https://prezi.com/view/E34MhxE5uKoZLWZ3KXjG/
368 ). The heatmap (**Figure 8A**) showing literature enrichment profiles highlighted keywords
369 associated with viral pathogens ("hepatitis", "herpes" or "influenza"), as well as host-derived and
370 pathogen-derived molecules ("RIG-I", "Interferon", "Interferon, "double stranded RNA"). From
371 the six modules, four seemed to be "core" interferon modules (M15.86, M10.1, M8.3, M15.127),
372 while the remaining two (M13.17, M15.64) were associated to the interferon literature to a lesser
373 degree. These latter two modules were more strongly associated with the Herpes simplex virus
374 than the other four modules while the four core modules were preferentially associated with
375 hepatitis. The heatmaps available via the Prezi interface, which depict the gene expression profiles
376 for the A28 modules provide additional perspectives. For example, the dataset from Speake *et al.*
377 comprised samples of patients with MS collected immediately before and 24 h after the
378 administration of their first dose of interferon-β (**Supplementary Figure 3** [GEO ID GSE60424);

379 (15)]. Despite the small number of subjects in this category, we observed a clear pattern of response
380 to interferon *in vivo* across all six modules. This observation confirms the functional associations
381 obtained from enrichment analyses.

382 Next, we examined the degree of changes for the six interferon modules across the 16 input
383 datasets (**Figure 8B**). The first cluster showing the highest induction levels comprised SLE,
384 influenza infection, HIV infection and active *M. tuberculosis* infection (noted as C1 on the figure).
385 Interferon has antiviral properties; therefore, it is no surprise to see viral infections included in this
386 set. Blood transcriptome profiling studies conducted nearly 20 years ago identified interferon
387 responses in SLE pathogenesis (18,19). More recent profiling also revealed the prominence of this
388 signature in patients with tuberculosis, which contrasts with findings made in other bacterial
389 infections (6). The second cluster (noted as C2) comprised diseases with an "intermediate" level
390 of interferon responses, including RSV infection, sepsis caused by *Staphylococcus aureus* in
391 pediatric patients and a range of bacterial pathogens in adults, SoJIA and liver transplant recipients
392 receiving maintenance immunosuppressive therapy. The illness conferred by RSV infection has a
393 clinical presentation very similar to that of influenza in pediatric patients. However, studies suggest
394 that interferon responses may be partially inhibited by RSV (20,21).

395 The final, third cluster was formed by pathologic and physiological states where an
396 increase in abundance of interferon-inducible genes was either modest or inexistent. This cluster
397 included patients with juvenile dermatomyositis and B-cell deficiency (low levels), Kawasaki
398 disease, melanoma, COPD, pregnancy (no increase), and individuals with MS (apparent decrease).
399 The latter observation presents some interest since one line of treatment for multiple sclerosis is
400 consists in administering interferon beta. This should as a result compensate the defect observed
401 here in treatment-naïve individuals.

402 Taken together, this closer examination of annotations for A28 modules and patterns of
403 changes in abundance in different reference datasets provided a clearer picture of its biological
404 significance.

405

406 **Profiling the abundance of A28 interferon-inducible genes at the module-level across**
407 **individual subjects.**

408 As described earlier analysis workflows have been developed to determine changes in transcript
409 abundance at the level of individual subjects. This approach, which offers an even more granular
410 perspective, was applied next for molecular stratification of patient cohorts using the six A28
411 interferon modules.

412 We show, for instance, that in cohorts where no changes are detectable via comparisons at
413 the overall group level, a minority of patients in fact present the signature. This scenario indeed
414 applies for the cohort of melanoma patients with 3 out of 22 subjects showing some degree of
415 increase in abundance for the six A28 modules, while the majority showed little changes or
416 decrease in abundance. At least four of the patients showed a marked decrease in abundance
417 (**Figure 9A**). The observation carries potential biological and clinical significance, as interferon
418 activity in patients with melanoma has been previously associated with disease outcomes (22,23).
419 Pathologies with an intermediate A28 signature might induce responses in a somewhat higher
420 proportion of subjects. This is the case of JDM patient cohort that comprises a cluster of 11 patients
421 our of 40 presenting with modular interferon signatures (**Figure 9A**). Of the 793 articles in
422 PubMed mentioning "Juvenile Dermatomyositis" in their titles, 8 also mentioned "interferon",
423 which indicates that a role for interferon in this disease while not widely acknowledged has

424 nonetheless been described (24). In diseases where the role for interferon is well described, such
425 as influenza infection or SLE, increase in A28 modules is widespread. But it is nevertheless
426 possible to find at least a few subjects in each cohort for whom an interferon signature is not
427 present (**Figure 9A**). Notably in the case of SLE the proportion of interferon negative subjects
428 tends to be higher in adult patient cohorts in comparison to pediatric patient cohorts as the one that
429 is being used for illustrative purposes here. Stratification of an adult SLE cohort based on patterns
430 of abundance of A28 modules is presented next.

431 We previously showed using our second generation repertoire that the interferon signature
432 characterizing SLE comprises distinct "sub-signatures" at the module level (11). In this earlier
433 work we observed a sequential increase for a set of three interferon modules (including M1.2,
434 M3.4 & M5.12). M1.2 showed a higher degree of sensitivity, followed by M3.4 and then M5.12.
435 Based on this finding, we could stratify SLE patients based on whether one, two or all three of
436 those interferon modules were activated. By combining functional profiling with a reference
437 dataset, we concluded that the modules responded differently to each interferon type: interferon-α
438 induced an increase in the abundance of genes belonging to M1.2; interferon-β induced an increase
439 in the abundance of genes belonging to M1.2 and M3.4; and interferon-γ induced an increase in
440 the abundance of genes belonging to M5.12. Next, we sought to determine the equivalence
441 between these three, second generation interferon modules and the six new, third generation
442 interferon modules regrouped in aggregate A28. Based on gene composition, M8.3 and M15.127
443 mapped to M1.2 (inducible by both interferon-α and β), M10.1 and M15.86 mapped to M3.4
444 (inducible by interferon-β), and M13.17 and M15.64 mapped to M5.12 (inducible by interferon-
445 γ). Notably, the latter two modules did indeed segregate from the other four on literature the
446 enrichment profiling heatmap presented earlier (**Figure 8A**). We also used the six interferon

447 modules to re-classify the adult SLE dataset profiled in our earlier study [GEO ID GSE49454
448 (11)]. The resulting clustering and stratification mirrored our earlier findings made using the three
449 interferon modules (**Supplementary Figure 4**). Overall these observations confirm that interferon
450 "sub-signatures" may be employed for patient stratification. This may be relevant in the tailoring
451 of biologics targeting interferon which are under development for the treatment of SLE. However,
452 whether utilizing six modules from the new repertoire would necessarily be an improvement over
453 the three from the second-generation repertoire that were employed previously remains to be
454 determined.

455 Finally, we wanted to illustrate how further insights could be obtained by examining
456 changes in A28 modules abundance at the individual level concurrently with that of other module
457 aggregates. As an example, we present the changes for both A28 (interferon) and A35
458 (inflammation) modules (**Figure 8B**). By combining the two modular signatures, we could assess
459 their relative contributions in a given cohort and identify distinct phenotypes: i.e. interferon or
460 inflammation "positive", "double positive" or "double negative". We found the most contrasting
461 patterns of change in transcript abundance in the sepsis and melanoma cohorts. Here, the
462 abundance of interferon (A28) and inflammation (A35) modules almost uniformly increased in
463 patients with sepsis. By contrast, we only observed increases in these modules in a minority of
464 melanoma patients, with increases of inflammation modules being more widespread than that of
465 interferon modules (~50% compared to ~10% of subjects, respectively). This finding was
466 paralleled in pregnant women, who represent another immunosuppressive state, although increases
467 in abundance of a subset of A35 modules appeared to be more prevalent in this group than in
468 melanoma patients. Influenza infection was more similar to sepsis but with higher levels of
469 interferon induction and a lower level of inflammation.

470 Taken together, the use case presented here illustrates the preliminary stepwise dissection
471 of a given module aggregate and investigation of the underlying biological relevance. Such a
472 process would be to be repeated for other module aggregates. While this is beyond the scope of
473 the present work it is a process that the scalable annotation infrastructure that has been developed
474 here will support. We indeed expect the range of reference datasets and functional profiling
475 approaches available for interpretation to continue to expand. This work should also help
476 determine to what extent subdivision of signatures in distinct modules is warranted.

477

478 **DISCUSSION:**

479 We have developed a new resource for the analysis and interpretation of blood transcriptome data.
480 At its core is a fixed, transcriptional module framework. This framework has been constructed via
481 co-expression analyses carried out across a wide range of diseases and physiological states. It
482 constitutes the third iteration released by our group since 2008. Another blood transcriptome
483 repertoire was developed and made available by our collaborators from Emory University [Li et
484 al.: (9)]. While methodological differences exist for the construction of such frameworks the
485 principle remains the same: i.e. utilization of reference datasets for constitution of reusable sets of
486 transcriptional modules.

487 Collections of gene sets are commonly used for the interpretation of transcriptome profiling
488 data [Gene Set Enrichment Analyses (GSEA): (25) (26)]. Such reference collections are very large,
489 typically numbering tens of thousands of signatures. While they prove useful as a referential for
490 functional interpretation, the majority of the sets are not derived via co-expression analysis. They
491 would therefore not be sufficiently coherent to serve as a basis for dimension reduction, nor were
492 they meant to be used in this fashion. Conversely, the construction of the set of transcriptional

493 modules that is presented here is entirely driven by co-expression, and is suitable for reducing
494 dimensions, but is intended for a much narrower use (i.e. the analysis and interpretation of blood
495 transcriptome data). Zhou and Altman recently described the development of another fixed
496 transcriptional module repertoire via co-expression analysis and discuss the advantages of such
497 approaches (27). In brief, they assembled a collection of 2753 public datasets that were deposited
498 in the NCBI GEO that encompassed 97,049 unique transcriptome profiles. They used independent
499 component analysis for resolving co-expression relationships which led to the identification of a
500 set of 139 transcriptional modules. They went on to demonstrate the biological relevance of these
501 sets and advantages their use presents for improving robustness of analyses, especially for smaller
502 datasets. Altogether this article provides additional arguments in favor of the reuse of fixed
503 transcriptional module repertoires, an approach that to date has not seen widespread use so far.
504 The authors chose a very wide breadth of datasets as input and the resulting framework may prove
505 rather generic since transcriptional regulation is largely cell and tissue-specific.

506 A key characteristic of fixed transcriptional module frameworks is that they are intended
507 for reuse, often over long periods of time. One of the major implications is that greater amounts of
508 effort can be dedicated to the annotation and development of *ad hoc* visualization tools. This is
509 obviously not possible when new module repertoires are constituted each time a new dataset needs
510 to be analyzed. And indeed, considerable annotation and development efforts have been
511 undertaken to support this latest generation of blood transcriptional modules. This permits us in
512 turn to make available an ecosystem of custom tools and resources that were not available for
513 transcriptional frameworks which have been released before. This includes a compendium of
514 functional profiling reports and reference transcriptional profiles that can be accessed via an
515 interactive web application. In addition, we are releasing another web application for data

516 browsing and multi-tiered visualization. And it should also be noted that this ecosystem is scalable
517 and will continue to grow over time as more analyses are being conducted, custom scripts
518 developed, and reference dataset profiles added to the annotation scaffold. It is illustrated by
519 ongoing efforts that consist in compiling relevant aggregate-centric information for A28. This is
520 with the intent of improving over time understanding of the clinical relevance and biological
521 significance of this transcriptional signature (accessible by clicking the information icon on the
522 A28 interactive presentation: https://prezi.com/view/E34MhxE5uKoZLWZ3KXjG/ and video
523 demonstration: https://youtu.be/0j-kcE1tlXA). A bioinformatics analysis toolkit that comprises a
524 collection of custom R scripts will also follow the publication of this report. Notably, earlier
525 attempts were made by our group to develop such a support infrastructure in the past, albeit on a
526 smaller scale. But maintaining such resources over long periods of time proved a challenge and
527 these by now have gone offline through a combination of hardware failures, hacking or
528 discontinued institutional support. Such issues should be addressed here in part through adoption
529 of zero cost infrastructure (e.g. hosting of Shiny apps, Prezi presentations or YouTube videos is
530 free as long as public access is maintained).

531 Finally, the resource that we are making available here has some limitations which are also
532 worth noting. First, our framework is specifically designed to analyze and interpret human blood
533 transcriptome profiling data. Analyses of other tissues or in the context of other species would
534 require the development and use of separate frameworks. Indeed, we also developed fixed modular
535 frameworks for various mouse tissues (28) and *in vitro* culture systems, including human whole
536 blood (29) and human dendritic cells (30). Second, we generated our transcriptome datasets for
537 module construction using Illumina BeadArrays, a technology that pre-dates RNA sequencing.
538 Our choice to use this technology was based on data availability at our institute, and the decision

539 to use data that was generated at a single facility over a relatively short period of time. An
540 alternative approach would be to compile available blood RNAseq data that is publicly available
541 and that also comprises appropriately matched controls. The coverage afforded by RNAseq data
542 would likely result in somewhat larger modular gene sets. However, we consider it unlikely that a
543 fundamentally different repertoire would be produced using RNAseq data, as this would assume
544 that entire co-expressed gene sets were missed by microarrays. Third, our modular repertoire is
545 not meant to constitute a ground truth. Indeed, some aggregates are clearly not homogeneous in
546 terms of functionality; and heterogeneity also exists within modules. The attribution of functional
547 annotation titles to modules also confers some degree of subjectivity. However, the repertoire does
548 structure the data so that insights about the biological significance of such a modular signature can
549 be obtained. As a result, while the framework being presented will remain fixed for at least the
550 coming few years, it is likely that the functional annotation map will continue to evolve for the
551 foreseeable future.

552 Going forward, future developments include, firstly, the design and implementation of
553 targeted transcriptional profiling assays based on the current repertoire. Indeed, such repertoires
554 constitute a useful framework for module-based target selection and the development of
555 "Transcriptome Fingerprinting Assays" [a preprint is available: (31)]. Such assays are cost-
556 effective, robust and scalable. Secondly, fixed transcriptional module repertoires may be leveraged
557 for conducting meta-analysis. It was for instance employed as a basis to perform such an
558 integrative analysis encompassing six independent RSV datasets encompassing 490 unique
559 profiles [a preprint is available: (32)]. More recently, this resource has also been used in a different
560 context to identify signatures associated with CTLA4 and combined CTLA4-PD1 blockade in
561 peripheral blood of ovarian cancer patients [P249: (33)]. It may also be possible to extend the

562 overall approach described here to develop similar interpretive transcriptomic frameworks for
563 other biological systems.
564

565 **ACKNOWLEDGEMENTS**

566 The authors would like to thank Quynh-Anh Nguyen, Kimberly O'Brien, Dimitry Popov,
567 Michael Mason, and Cate Speake for technical assistance. We would also like to acknowledge
568 Insight Editing London for assistance in editing the manuscript prior to submission.
569

570 This project has been funded in part with Federal funds from the National Institutes of Health
571 under contract number U01AI082110. Authors affiliated with Sidra Medicine, a member of the
572 Qatar foundation for Education, Science and Community Development, are fully supported by
573 institutional funding. RJW is supported by the Wellcome Trust (203135, 104803), NIH
574 (U01AI115940) and the Francis Crick Institute, which receives funding from the Wellcome Trust
575 (FC10218), CR UK (FC10218) and UKRI (FC10218).
576

577 **AUTHOR'S CONTRIBUTIONS**

578 Conceptualization: MCA, DR, NB, DC. Data curation: NB, MA. Visualization: DR, MT, DC.
579 Analysis and interpretation: MCA, DR, NB, AAB, EW, MG, BAK, MT, DC. Resources: MT,
580 MA, SP, PK, LC, NJC, AAB, FM, PT, TP, GK, ML, HR, AOG, MB, CB, ML, RJW, CG, GL,
581 MC, JS, RT, FK, AM, OR, KP, VP, JB. Writing of the first draft: DC. Funding acquisition: GK,
582 KP, VP, OR, JB, DC. Methodology development: MCA, DR, NB, EW. Writing – review &
583 editing: MCA, DR, NB, MT, EW, MG, BAK, MA, SP, PK AAB, FM, PT, LC, NJC, JTP, GK,
584 AOG, MB, CB, RJW, CMG, ML, GL, DB, RT, FK, AM, OR, KP, VP, JB, DC. The

585 contributor's roles listed above follow the Contributor Roles Taxonomy (CRediT) managed by
586 The Consortia Advancing Standards in Research Administration Information (CASRAI).
587

588 **DECLARATION OF INTERESTS**

589 The authors declare no competing interests
590

591 **FIGURE LEGENDS**

592 **Figure 1: The module repertoire construction process.** A. A collection of 16 transcriptome
593 datasets spanning a wide range of immunological and physiological states were used as input. B.
594 Each dataset was independently clustered via k-means clustering. C. Gene co-clustering events
595 were recorded in a table, where the entries indicate the number of datasets in which co-clustering
596 was observed for a given gene pair. D. The co-clustering table served as the input to a weighted
597 co-clustering graph (see also **Supplementary Figure 1**), where the nodes represent genes and the
598 edges represent co-clustering events. E. The largest, most highly weighted sub-networks among a
599 large network (here constituting 15,132 nodes) were identified mathematically and assigned a
600 module ID. The genes constituting this module were removed from the selection pool and the
601 process was repeated, resulting in the selection of 382 modules constituted by 14,168 transcripts.

602

603 **Figure 2: The development of module fingerprint grids.** A. The modules were arranged onto
604 the grid as follows: the master set of 382 modules was partitioned into 38 clusters (or aggregates)
605 based on similarities among their module activity profiles across the sixteen reference datasets
606 (A1-A38). B. A subset of 27 aggregates comprising 2 modules or more in turn occupied a line on

607 the grid. The length of each line was adapted to accommodate the number of modules assigned to
608 each cluster. The format of the grid was fixed for all analyses carried out using this modular
609 framework. C. Changes in transcript abundance at the module level were mapped onto this grid
610 and represented by color spots of varying intensity.

611

612 **Figure 3: Fingerprint grid plots.** A. Prototypical fingerprint grid plot: Changes in blood
613 transcript abundance for patients with SLE compared to healthy controls are represented on a
614 fingerprint grid plot for this illustrative use case. The modules occupy a fixed position on the
615 fingerprint grid plots (see Figure 2). An increase in transcript abundance for a given module is
616 represented by a red spot; a decrease in abundance is represented by a blue spot. Modules arranged
617 on a given row belong to a module aggregate (here denoted as A1 to A38). Changes measured at
618 the "aggregate-level" are represented by spots to the left of the grid next to the denomination for
619 the corresponding aggregate. The colors and intensities of the spots are based on the average across
620 each given row of modules. A module annotation grid is provided where a color key indicates the
621 functional associations attributed to some of the modules on the grid (top right). Positions on the
622 annotation grid occupied by modules for which no consensus annotation was attributed are colored
623 white. Positions on the gird for which no modules have been assigned are colored grey. B-D.
624 Fingerprint grid plots for additional reference datasets.

625

626 **Figure 4: Functional annotation of the transcriptional module repertoire.** An interactive
627 application is available to explore the 382 modules comprising the blood transcriptome repertoire.
628 A gene list, along with the ontology, pathway, literature term enrichment and transcriptional

629 profiling data for reference transcriptome datasets (circulating leukocyte populations,
630 hematopoiesis) is provided for each module. Zoom in and out functionalities for close-up
631 examinations of the text and figures embedded in the presentation is possible. Web links providing
632 access to modules within a given aggregate are listed in Table 3. For a demonstration video, please
633 visit: https://youtu.be/8ajESET2mqI.

634

635 **Figure 5: Individual-level module heatmap.** Changes in transcript abundance were determined
636 at the individual level across all modules constituting the repertoire. These changes are represented
637 on a heatmap, where an increase in abundance of a given module is represented in red, and a
638 decrease in abundance is represented in blue. The subjects are organized as columns and the
639 modules as rows. The ordering on the heat map was determined by hierarchical clustering.

640

641 **Figure 6: Web application to visualize multi-tiered module fingerprinting.** An application was
642 developed to explore the changes in transcript abundance at the module level across the 16
643 reference datasets used to construct the repertoire. Three types of plot can be displayed and
644 exported: 1) fingerprint grids; 2) module heatmaps displaying changes in abundance in modules
645 comprising a given aggregate across the 16 reference datasets; and 3) module heatmaps displaying
646 changes in abundance in modules comprising a given aggregate across individuals constituting a
647 given dataset. For accessing the application, please visit:
648 https://drinchai.shinyapps.io/dc_gen3_module_analysis/#. For a demonstration video, please visit:
649 https://youtu.be/y__7xKJo5e4.

650

651 **Figure 7: Module aggregate abundance patterns across the 16 disease or physiological states.**
652 **A. Patterns of changes in transcript abundance at the aggregate and cohort levels.** Each
653 column on the heat map corresponds to a "module aggregate", numbered A1 to A38. Modules A9-
654 A14 and A19-A24 were excluded as they each only included one module. Each row on the heatmap
655 corresponds to one of the 16 datasets used to construct the module framework. A red spot on the
656 heatmap indicates an increase in abundance of transcripts comprising a given module cluster for a
657 given disease or physiologic state. A blue spot indicates a decrease in abundance of transcripts.
658 No color indicates no change. Disease or physiological states were arranged based on the level of
659 similarity in the patterns of aggregate activity, determined via hierarchical clustering. **B.**
660 **Representation of the modules and genes constituting aggregate A28.** The circle plot represents
661 the six modules constituting aggregate 28, and the transcripts constituting each of the modules.
662 Some genes on the Illumina BeadArrays can map to multiple probes, which explains the few
663 instances where the same gene can be found in different modules. **C. Patterns of changes in**
664 **transcript abundance at the module-level and gene-level for aggregate A28.** The circle plots
665 illustrate the changes at the gene-level for this aggregate for 6/16 datasets. The position of the
666 genes on each of these plots is the same as shown in panel B. Genes for which transcript abundance
667 is changed are shown in red (increase) or in blue (decrease). **D. Patterns of changes in transcript**
668 **abundance at the module and gene level for aggregate A28 in subjects treated with IFNα or**
669 **IFNβ treatment** The circle plots show changes in abundance for A28 transcripts in patients with
670 hepatitis C infection treated with IFNα [GSE11342 (16)] or patients with multiple sclerosis treated
671 with IFNβ [GSE26104 (17)].

672

673 **Figure 8: Literature profiles and patterns of changes in abundance across reference datasets**
674 **for the modules comprising aggregate A28. A. Functional annotation by literature profiling:**
675 A portion of a heatmap comprising 382 modules organized as columns, and literature terms
676 organized as rows. The six modules shown are associated with the consensus annotation
677 "Interferon" or "Type 1 Interferon". The clusters of keywords associated with those modules are
678 consistent with this annotation and provide added granularity to the module repertoire for
679 functional profiling and interpretation. **B. Changes in abundance across 16 reference datasets:**
680 the heatmap (top) represents the changes in abundance of transcripts constituting the six modules
681 comprising aggregate A28 (columns). The modules are functionally associated with interferon
682 responses. The 16 reference datasets are arranged as rows corresponding to different health states.
683 The columns and rows were arranged by hierarchical clustering. Such heatmaps can be accessed
684 and exported for all 16 datasets and 38 module aggregates using the web application:
685 https://drinchai.shinyapps.io/dc_gen3_module_analysis/# (under the "MODULES x STUDIES"
686 tab).

687

688 **Figure 9: Abundance patterns across individuals. A. Changes in abundance for A28 modules.**
689 The heatmaps display the changes in abundance for the same six modules (rows) across individuals
690 (columns) in four reference cohorts. The rows and columns on the heatmap are arranged based on
691 similarities in abundance patterns. **B. Changes in abundance for A28 and A35 modules.** The
692 heatmaps display the changes in abundance of six modules constituting aggregate A28 and 21
693 modules constituting aggregate A35 (rows) across individuals (columns) in four reference datasets.
694 Functional annotations associated with different modules is indicated by a color code and
695 corresponding legend. Such heatmaps can be accessed and exported for all 16 datasets and 38

696 module aggregates using the web application:
697 https://drinchai.shinyapps.io/dc_gen3_module_analysis/# (under the "MODULES x
698 INDIVIDUALS" tab).

699

700 **Figure S1: Construction of weighted co-clustering networks.** Weighted co-clustering networks
701 were used to construct the modular repertoires. These networks factor in differences in co-
702 expression across different "states" of the biological system. For the blood transcriptome, these
703 states would be different diseases or physiological phenotypes. Under Scenario A, the genes are
704 co-expressed in all three disease states, so the weight attributed to the edges of the network is three.
705 Under Scenarios B and C, co-clustering only occurs in two or one of the disease states, resulting
706 in weights of 2 and 1, respectively.

707

708 **Figure S2: Patterns of changes in transcript abundance at the module and gene level for**
709 **aggregate A28.** The large circle plot represents the six modules constituting aggregate 28, and the
710 transcripts constituting each of the modules. Some genes on the Illumina BeadArrays can map to
711 multiple probes, which explains the few instances where the same gene can be found in different
712 modules. Patterns of changes in transcript abundance at the module-level and gene-level for
713 aggregate A28 are shown for all 16 reference datasets. The position of the genes on each of these
714 plots is fixed. Genes for which transcript abundance is changed are shown in red (increase) or in
715 blue (decrease).

716

717 **Figure S3: Expression patterns of genes constituting the six A28 modules (rows) across whole**
718 **blood and purified circulating leukocyte populations (columns).** This dataset was contributed
719 by our group [Speake *et al.* (GSE60424)], and compared the RNAseq transcriptome signatures of
720 six immune-cell subsets and whole blood from patients with various immune-associated diseases
721 (color coding is provided to indicate cell populations and health status). The columns highlighted
722 by the green boxes show the expression levels in patients with MS before and 24 h after the first
723 dose of IFN-β.

724

725 **Figure S4: Abundance patterns of six interferon modules (A28) across individuals from an**
726 **adult SLE cohort.** The abundance levels for six interferon modules belonging to aggregate A28
727 (columns) is shown across a cohort of adult subjects diagnosed with SLE (rows). These subjects
728 were part of the LUPUCE study: the dataset was contributed by our group [Chiche *et al.* (GEO ID
729 GSE49454)] (11). Changes in transcript abundance were measured in reference to a cohort of
730 healthy subjects included in the study. The colored label with interferon grouping information
731 indicates the classification obtained by Chiche *et al.* based on three "second generation" modules.

732

733

734

735

736

737

738 **TABLES**

739 **Table 1: Datasets used for module construction.**

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

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

745

746

747

748 **Table 2: The characteristics of the input datasets used to construct three consecutive**

749 **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 |
| Systemic lupus erythematosus | X | X | X |
| Juvenile dermatomyositis | | | X |
| Type 1 Diabetes | X | | |
| Multiple sclerosis | | | X |
| Kawasaki disease | | | X |
| COPD | | | X |
| Tuberculosis | | X | X |
| Sepsis | | 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 (4) | 2013 (3) | Current work |

750

751

752

753

754

755

756

757

758

759

760 **Table 3: Links to module aggregates annotation pages**

| Aggregate | Function | Links |
|---|---|---|
| **A1** | Lymphocytic | https://prezi.com/view/sxap39tKxkmCNTTNIlVO/ |
| **A2** | TBD | https://prezi.com/view/96GWajx5mZjuRS4B6gjA/ |
| **A3** | TBD | https://prezi.com/view/OWFVI51FND0WWwNgsgJZ/ |
| **A4** | TBD | https://prezi.com/view/2Zbq8ZDYbO4hbUd4r2KF/ |
| **A5** | Lymphocytic | https://prezi.com/view/62tgA5E6roOlk5DRNvS1/ |
| **A6** | Lymphocytic | https://prezi.com/view/Uks2Nd4lvizNNFVPtBEy/ |
| **A7** | TBD | https://prezi.com/view/kKfergNj0SkLXyFtm0Dg/ |
| **A8** | TBD | https://prezi.com/view/Y4uk1RPJyNcSndJYnFX6/ |
| **A9** | TBD | https://prezi.com/view/jgYehQ9QhyADAttEsdoI/ |
| **A15** | TBD | https://prezi.com/view/jgYehQ9QhyADAttEsdoI/ |
| **A16** | TBD | https://prezi.com/view/SKzHeA0XYdLYvy2sY8gP/ |
| **A17** | TBD | https://prezi.com/view/FS7sE1Vqew5g8EKOM1AM/ |
| **A18** | TBD | https://prezi.com/view/aZMLflMNVrV7JnVaIILm/ |
| **A24** | Oxidative phosphorylation | https://prezi.com/view/eiXvf2LNBLFRgrtaeCuM/ |
| **A25** | TBD | https://prezi.com/view/pwyojaU62Z7GT102ZYwM/ |
| **A26** | TBD | https://prezi.com/view/9CErpW3NwpN2HgRS3Hzf/ |
| **A27** | Cell cycle | https://prezi.com/view/GgIiA0K9kSFHbpVj2I85/ |
| **A28** | Interferon | https://prezi.com/view/E34MhxE5uKoZLWZ3KXjG/ |
| **A29** | TBD | https://prezi.com/view/W4TShTd32dEJx0XPOF1U/ |
| **A30** | TBD | https://prezi.com/view/kl7VHoJTWug0sn7TgXut/ |
| **A31** | TBD | https://prezi.com/view/GqtUO22JJlSf16zMJKbB/ |
| **A32** | TBD | https://prezi.com/view/qlbG9VFzegOndQKD8aiy/ |
| **A33** | Inflammation | https://prezi.com/view/VBqKqHuLWCra3OJOIZRR/ |
| **A34** | TBD | https://prezi.com/view/HcSgIEGP3TJjTSpaPCxv/ |
| **A35** | Inflammation | https://prezi.com/view/7Q20FyW6Hrs5NjMaTUyW/ |
| **A36** | Erythroid | https://prezi.com/view/M7dnztl2h61gXrKFQeB2/ |
| **A37** | Erythroid | https://prezi.com/view/YyQs4WiXSNf0YXE79lfS/ |
| **A38** | Erythroid | https://prezi.com/view/0KUlPICKUZGeUjb206R5/ |

761

762 **METHODS:**

763 **Study subjects**

764 Gene expression datasets from 985 de-identified subjects from distinct cohorts were used for this
765 work. Written informed consent was obtained from all participants. Studies were approved by
766 Institutional Review Boards of the Baylor College of Medicine (COPD dataset: H-18029), the
767 University of Texas Southwestern Medical Center and Baylor Health Care System (Influenza,
768 RSV, S. aureus and Kawasaki disease datasets: UTSW #0802-447 / BIIR #002-141), Saint Jude's
769 Research Hospital (B-cell deficiency), the Baylor Health Care System (Liver transplant: 002-197,
770 Pregnancy: 009-257, Multiple sclerosis: 009-240, Melanoma: 006-025 & 097-027), Khon Kaen
771 University (Sepsis), the University of Texas Southwestern Medical Center (SoJIA,
772 Dermatomyositis, SLE), Duke University and the Baylor Health Care System (HIV: Duke 8485-
773 06-4R0 / Baylor 006-177), St. Mary's Hospital London, UK and University of Cape Town, Cape
774 Town, Republic of South Africa (Tuberculosis: St Mary's REC 06/Q0403/128, U of Cape Town
775 REC 012/2007). The gene expression datasets were selected to cover major classes of immune
776 states (**Table 2**), were required to have at least 25 samples in total, and at least 20% of the total
777 samples were required to be appropriately matched controls.

778

779 General descriptions of the study cohorts are as follows: ***S. aureus* cohort:** Children with
780 community-acquired S. aureus infection were enrolled. The clinical syndromes of these patients
781 included skin and soft tissue infection, bacteremia, osteomyelitis, suppurative arthritis,
782 pyomyositis, pneumonia, and disseminated disease. Patients diagnosed with toxic shock
783 syndrome, polymicrobial infections, or treated with corticosteroids in the preceding four weeks
784 were excluded. **Sepsis cohort:** Diagnosis of sepsis was based on accepted international guidelines

785 and defined as presentation with two or more of the following criteria for the systemic
786 inflammatory response syndrome: fever (temperature >38°C or <36°C), tachycardia (heart rate
787 >90 beats/minute), leukocytosis or leukocytopenia (white blood cell count ≥12 × 109/l or ≤4 ×
788 109/l). Blood was collected within 24 hours following the diagnosis of sepsis. Samples were
789 selected for microarray analysis from subjects who had the diagnosis of bacteremic sepsis
790 retrospectively confirmed by the isolation of a causative organism on blood culture. **TB cohort:**
791 Patients were prospectively recruited and sampled, before any anti-mycobacterial treatment was
792 initiated. Active TB disease was confirmed by laboratory isolation of M. tuberculosis on
793 mycobacterial culture of a respiratory specimen (either sputum or bronchoalvelolar lavage fluid).
794 **Influenza cohort:** Children with confirmed influenza infection were recruited. Those with
795 documented bacterial co-infections, children with chronic conditions and systemic steroid
796 treatment within 2 weeks before enrollment were excluded. **RSV cohort:** Patients with confirmed
797 RSV infection were recruited. Children with documented bacterial co-infections, congenital heart
798 disease, chronic lung disease, immunodeficiency, prematurity (<36 wk), systemic steroid
799 treatment within 2 weeks before presentation or additional chronic comorbidities were excluded.
800 **HIV cohort:** Blood samples were obtained from adult patients diagnosed with HIV infection. At
801 enrollment patients were verified as acute Fiebig stages 4 to 6 (plasma RNA+, third generation
802 EIA+, Western blot indeterminant or +). **SLE cohort:** Blood samples were obtained from pediatric
803 patients diagnosed with Systemic Lupus Erythematosus and healthy controls matched for
804 demographic characteristics. MS cohort: Subjects enrolled in the MS cohort had an established
805 diagnosis of relapsing-remitting multiple sclerosis (MS), separately confirmed by an experienced
806 MS neurologist (JTP), exhibited no other health conditions, and had received no treatment(s) for
807 MS, including corticosteroids, for at least 3 months prior to blood draw. **Juvenile**

808 **Dermatomyositis cohort:** Blood samples were obtained from pediatric patients diagnosed with
809 juvenile dermatomyositis and healthy controls matched for demographic characteristics.
810 **Kawasaki Disease cohort:** Subjects <18 years of age who met the definition of Kawasaki Disease
811 based on the American Heart Association (AHA) criteria (34) were enrolled alongside age and
812 gender matched healthy controls. **Systemic onset Juvenile Arthritis cohort:** Blood samples were
813 obtained from SoJIA patients displaying systemic symptoms only (fever, rash, and/or pericarditis)
814 or displaying systemic symptoms accompanied by arthritis. **COPD cohort:** Enrollment criteria
815 included age over 40 years, no history of concurrent lung cancer, chest surgery, or chronic lung
816 diseases other than COPD (e.g., sarcoidosis, fibrosis, etc.). Participants had no history of allergies
817 or asthma and at the time of initial recruitment had not received oral or systemic corticosteroids
818 during the previous 6 weeks; volunteers were enrolled from three clinics within the Texas Medical
819 Center in Houston, Texas. **B-cell deficiency cohort:** This cohort comprised adults with diagnosis
820 of XLA as documented by markedly reduced numbers of peripheral blood B cells. **Pregnancy**
821 **cohort:** Pregnant women were recruited at the Baylor Institute for Immunology Research (Dallas,
822 TX) for a study of immunological signatures of pregnancy. **Melanoma cohort:** Enrollment criteria
823 included ages of 21-75 years stage M1a, M1b, M1c biopsy proven metastatic melanoma patients
824 with measurable metastatic lesions by physical exam or scans, acceptable CBC and blood
825 chemistry results, adequate hepatic and renal function, no active CNS metastatic disease. **Liver**
826 **transplant cohort:** Enrollment criteria included age 17-65 years, having received a liver transplant
827 under maintenance immunosuppression therapy. Subjects in this cohort had not received an acute
828 or chronic rejection diagnosis at the time of sampling.
829

830 **RNA extraction and processing**

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

837 **Microarray analysis and data preprocessing**

838 Gene expression profiles from whole blood samples generated using Illumina HumanHT-12 v3.0
839 or Illumina HumanHT-12 v4.0 expression BeadChips were obtained from 16 groups of patients
840 and controls selected as above. Thus, 16 datasets were used as the input (Table 1). The expression
841 data for each dataset were preprocessed and independently clustered. First, the probes were
842 discarded if they were not present (detection $P < 0.01$) in at least 10 samples or in at least 10% of
843 the samples, whichever was greater. Then, the sample data for each dataset were normalized using
844 the BeadStudio average normalization algorithm. Once normalized, the signal was floored such
845 that all signals <10 were set to 10. Then, the fold change was calculated relative to the median
846 signal for that probe across all samples. If the difference between a signal and the probe's median
847 signal was <30, or the calculated absolute magnitude of the fold change was <1.2, the fold change
848 was set to 1 to reduce the noise from low-level responses. At this stage, the probes were filtered
849 again. Probes were only retained if they had a calculated absolute fold change >1 in at least 10
850 samples or in at least 10% of the samples, whichever was greater. Finally, the data was transformed
851 to the $\log_2$ of the calculated fold changes.

852 **Module construction algorithm**

853 Sets of coordinately regulated genes, or transcriptional modules, were extracted from the patient's
854 whole blood microarray datasets. Each of the preprocessed microarray datasets were clustered in
855 parallel using Euclidean distance and the Hartigan's K-Means clustering algorithm. The 'ideal'
856 number of clusters (k) for each dataset was determined within a range of k=1-100 by means of the
857 jump statistic (35). Taking the 16 sets of clusters as the input (**Table 2**), a weighted co-cluster
858 graph was constructed (Chaussabel, Quinn et al. 2008, Chaussabel and Baldwin 2014) . To select
859 modules, an iterative algorithm was used to extract the sets of probes that are most frequently
860 clustered together in the same datasets, proceeding from the most stringent requirements to the
861 least, as previously described (Chaussabel, Quinn et al. 2008) . This iteration differed from the
862 previous implementation of this algorithm in that the k value was calculated independently for
863 each dataset cluster and the size of the core sub-networks was smaller (10 probes). The algorithm
864 also differed from previous implementations to ensure that the core sub-networks co-clustered in
865 the same datasets. Further details and an example of the code are included in the supplemental
866 methods (**Supplementary File 1**). The resulting 382 module set constitutes the third generation of
867 modular blood transcriptome repertoire constructed since the development of the first generation
868 published in 2008 (4), and the second generation published in 2013 (3).

869 **Module annotation**

870 <u>Gene Ontology / Pathway enrichment:</u> Module gene lists were investigated using Database for
871 Annotation Visualization and Integrated Discovery (DAVID) version 6.7 (12) . This database uses
872 a modified Fisher exact test to identify specific biological/functional categories that are
873 overrepresented in gene sets in comparison with a reference set. The top matched DAVID
874 annotation cluster (using default settings), the top matched canonical pathway from Kyoto

875 encyclopedia of genes and genomes (KEGG), the top matched pathway from BioCarta, and the
876 top matched Gene Ontology biologic process (GO_BP) and molecular function (GO_MF) terms
877 were identified for each module. Each module was also investigated for significant overlap with
878 two other established blood transcriptome module repertoires (3,9). The findings are summarized
879 in the module annotation spreadsheet (**Supplemental File 2**).
880
881 Gene Set Annotation (GSAn): To further annotate the modules, a new alternative to statistical
882 enrichment analysis tool called GSAn was applied (13). Statistical enrichment methods may have
883 limitations (36–38), as these methods tend to focus on the subpart of the most studied genes and
884 to provide gene set annotation results that cover a limited number of the well-annotated genes. To
885 address these issues, GSAn offers: (i) an original method that combines semantic similarity
886 measures and data mining approaches to perform a unified and synthetic annotation for a gene set
887 of interest, and (ii) a visualization approach to facilitate the interactive exploration of the gene set
888 annotation results according to the hierarchical structure of Gene Ontology (13). The tool is
889 available online: https://gsan.labri.fr/. A page listing analysis results for all 382 generation 3 blood
890 transcriptome modules can be accessed here: https://ayllonbe.github.io/modulesV3/index.html.
891
892 Pathway enrichment analyses: These analyses were conducted using Ingenuity Pathway Analysis
893 (Qiagen, Valencia, CA).
894 Literature profiling: Literature Lab™ (LitLab; from Acumenta Biotech, Boston, MA) was used
895 to associate genes within a particular module to terms used in PubMed abstracts (Febbo,
896 Mulligan et al. 2007). Association scores reflecting the strength of the associations were used to
897 calculate the "Product Scores". The top three terms that showed the strongest association and

898 highest "Product Scores" were used to create the functional annotation. A similar approach using
899 LitLab has been previously reported (Alsina, Israelsson et al. 2014). The steps taken to annotate
900 all 382 modules is described here in brief. All statistical analyses were performed using
901 Microsoft Excel (2010) with Visual Basic for Applications (VBA), Linux-based command line
902 in Mac OS, and R statistical software.

903 To construct a Product Scores Table, all the terms available in LitLab (>80,000) were
904 listed. Then, the genes in each module were submitted as a list to LitLab Editor and manually
905 validated using LitLab's built-in validation tool and/or NCBI Gene
906 (https://www.ncbi.nlm.nih.gov/gene) prior to submission for analysis using all domains available.
907 After completing the analysis, the summary result page was exported to an xls file. Using the
908 UNIX command line, the exported files were converted to csv files with the filename appended in
909 the last column of each row and vertically appended. The "merged" file was used to populate the
910 table that included all the available LitLab terms. The top three terms with the highest Product
911 Scores were selected to represent the module functional annotation and are tabulated in column I
912 of the module annotation table (**Supplementary file 2**).

913

914 **Fingerprint grid plot visualization**

915 Modules were arranged on a grid based on their similarities in patterns of activity across the 16
916 input datasets, each of them corresponding to a different pathological or physiological state. First,
917 the modules were partitioned using k-means clustering, which resulted in the constitution of 38
918 clusters. Given the possibility of collapsing values of the modules constituting each cluster in a
919 single "aggregate" value, the term "module aggregate" was used to designate each cluster (A1 to
920 A38). Of these 38 k-means clusters, 27 comprised >1 module. The modules were then arranged

921 on a grid with each row corresponding to modules belonging to the same aggregate (Figure 4). The
922 total number of rows on the grid equaled 27 and number of columns equaled the largest number
923 of modules for a given aggregate (42 for aggregate A2). For each module, the highest of the two
924 values indicating an increase or a decrease was selected for visualization (e.g. if % increase > %
925 decrease, then a red sport representing % increase is shown).
926
927 **Module Fingerprint analysis and visualization**
928 The modular analysis was performed by using 14,168 transcripts. Fold change was computed using
929 gene expression data prior log2 transformation. For group comparisons, paired t-test was
930 performed on log2-transformed data (Fold change (FC) cut off = 1.5; FDR cut off = 0.1). For
931 individual-patients analysis, each sample was compared to mean of control samples in each
932 dataset. Cut off comprised a absolute FC > 1.5 and a difference in gene expression level > 10. A
933 module was considered to be "responsive" when the proportion of differentially expressed
934 transcripts (as defined above) was greater than 15%. Data visualization were performed using
935 "ComplexHeatmap" (36).
936
937 **Generation of Circular Plots**
938 Circular plots were generated to represent module expression at the gene-level (**Figure 7**). Fold
939 change in case vs control group in each dataset were shown. Probes that passed FDR < 0.1 and FC
940 > 1.5 were present in gradated red colors and those with FC < -1.5 were represented in gradated
941 blue colors. The position of the genes on each of the circular plots is fixed.
942
943

944 **Statistical analyses**

945 Numerical data were processed and analyzed using R statistical software (version1.1.463-©2009-
946 2018 RStudio, Inc). Student's t-test was used to determine which group was different. Values of
947 $p < 0.05$ were considered to indicate statistical significance, with adjustment by multiple-testing
948 correction when needed (FDR, Benjamini–Hochberg procedure).

949

950 **Availability of data and material**

951 Transcriptome profiling data are deposited in the Gene Expression Omnibus,
952 https://www.ncbi.nlm.nih.gov/geo/, under the accession GSE100150.

953

954 **SUPPLEMENTAL INFORMATION**

955 **Supplementary File 1: Module generation method and pseudocode**

956 This Word document describes the algorithm used to construct this modular repertoire framework
957 along with pseudocode, which may be used as a basis for implementation in programming
958 languages such as R or Python.

959 **Supplementary File 2: Annotated module repertoire**

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

963 **Supplementary File 3: Module fingerprints of all 16 pathological and physiological states**

964 This PDF file contains the modular fingerprints generated for each of the 16 input datasets.

965

966 **REFERENCES**

967


968 1. Hulsen T, Jamuar SS, Moody AR, Karnes JH, Varga O, Hedensted S, et al. From Big Data to Precision
969 Medicine. Front Med. 2019;6:34.

970 2. Chaussabel D, Baldwin N. Democratizing systems immunology with modular transcriptional
971 repertoire analyses. Nat Rev Immunol. 2014;14(4):271–80.

972 3. Obermoser G, Presnell S, Domico K, Xu H, Wang Y, Anguiano E, et al. Systems scale interactive
973 exploration reveals quantitative and qualitative differences in response to influenza and
974 pneumococcal vaccines. Immunity. 2013 Apr 18;38(4):831–44.

975 4. Chaussabel D, Quinn C, Shen J, Patel P, Glaser C, Baldwin N, et al. A modular analysis framework
976 for blood genomics studies: application to systemic lupus erythematosus. Immunity. 2008 Jul
977 18;29(1):150–64.

978 5. Oswald M, Curran ME, Lamberth SL, Townsend RM, Hamilton JD, Chernoff DN, et al. Modular
979 analysis of peripheral blood gene expression in rheumatoid arthritis captures reproducible gene
980 expression changes in tumor necrosis factor responders. Arthritis Rheumatol Hoboken NJ. 2015
981 Feb;67(2):344–51.

982 6. Berry MPR, Graham CM, McNab FW, Xu Z, Bloch SAA, Oni T, et al. An interferon-inducible
983 neutrophil-driven blood transcriptional signature in human tuberculosis. Nature. 2010 Aug
984 19;466(7309):973–7.

985 7. Banchereau R, Hong S, Cantarel B, Baldwin N, Baisch J, Edens M, et al. Personalized
986 Immunomonitoring Uncovers Molecular Networks that Stratify Lupus Patients. Cell. 2016 Apr
987 21;165(3):551–65.

988 8. Pankla R, Buddhisa S, Berry M, Blankenship DM, Bancroft GJ, Banchereau J, et al. Genomic
989 transcriptional profiling identifies a candidate blood biomarker signature for the diagnosis of
990 septicemic melioidosis. Genome Biol. 2009;10(11):R127.

991 9. Li S, Rouphael N, Duraisingham S, Romero-Steiner S, Presnell S, Davis C, et al. Molecular signatures
992 of antibody responses derived from a systems biology study of five human vaccines. Nat Immunol.
993 2014 Feb;15(2):195–204.

994 10. Zhou W, Altman RB. Data-driven human transcriptomic modules determined by independent
995 component analysis. BMC Bioinformatics. 2018 Sep 17;19(1):327.

996 11. Chiche L, Jourde-Chiche N, Whalen E, Presnell S, Gersuk V, Dang K, et al. Modular transcriptional
997 repertoire analyses of adults with systemic lupus erythematosus reveal distinct type I and type II
998 interferon signatures. Arthritis Rheumatol Hoboken NJ. 2014 Jun;66(6):1583–95.

999 12. Jiao X, Sherman BT, Huang DW, Stephens R, Baseler MW, Lane HC, et al. DAVID-WS: a stateful web
1000 service to facilitate gene/protein list analysis. Bioinforma Oxf Engl. 2012 Jul 1;28(13):1805–6.

1001 13. Ayllon-Benitez A, Bourqui R, Thébault P, Mougin F. GSAn: an alternative to enrichment analysis for
1002 annotating gene sets. NAR Genomics Bioinforma [Internet]. 2020 Jun 1 [cited 2020 Apr 1];2(2).
1003 Available from: https://academic.oup.com/nargab/article/2/2/lqaa017/5805305

1004 14. Novershtern N, Subramanian A, Lawton LN, Mak RH, Haining WN, McConkey ME, et al. Densely
1005 interconnected transcriptional circuits control cell states in human hematopoiesis. Cell. 2011 Jan
1006 21;144(2):296–309.

1007 15. Linsley PS, Speake C, Whalen E, Chaussabel D. Copy number loss of the interferon gene cluster in
1008 melanomas is linked to reduced T cell infiltrate and poor patient prognosis. PloS One.
1009 2014;9(10):e109760.

1010 16. Taylor MW, Tsukahara T, McClintick JN, Edenberg HJ, Kwo P. Cyclic changes in gene expression
1011 induced by Peg-interferon alfa-2b plus ribavirin in peripheral blood monocytes (PBMC) of hepatitis
1012 C patients during the first 10 weeks of treatment. J Transl Med. 2008 Nov 5;6:66.

1013 17. Malhotra S, Bustamante MF, Pérez-Miralles F, Rio J, Ruiz de Villa MC, Vegas E, et al. Search for
1014 specific biomarkers of IFNβ bioactivity in patients with multiple sclerosis. PloS One.
1015 2011;6(8):e23634.

1016 18. Bennett L, Palucka AK, Arce E, Cantrell V, Borvak J, Banchereau J, et al. Interferon and
1017 granulopoiesis signatures in systemic lupus erythematosus blood. J Exp Med. 2003 Mar
1018 17;197(6):711–23.

1019 19. Baechler EC, Batliwalla FM, Karypis G, Gaffney PM, Ortmann WA, Espe KJ, et al. Interferon-
1020 inducible gene expression signature in peripheral blood cells of patients with severe lupus. Proc
1021 Natl Acad Sci U S A. 2003 Mar 4;100(5):2610–5.

1022 20. Spann KM, Tran K-C, Chi B, Rabin RL, Collins PL. Suppression of the induction of alpha, beta, and
1023 lambda interferons by the NS1 and NS2 proteins of human respiratory syncytial virus in human
1024 epithelial cells and macrophages [corrected]. J Virol. 2004 Apr;78(8):4363–9.

1025 21. Kalinowski A, Galen BT, Ueki IF, Sun Y, Mulenos A, Osafo-Addo A, et al. Respiratory syncytial virus
1026 activates epidermal growth factor receptor to suppress interferon regulatory factor 1-dependent
1027 interferon-lambda and antiviral defense in airway epithelium. Mucosal Immunol. 2018;11(3):958–
1028 67.

1029 22. Carretero R, Wang E, Rodriguez AI, Reinboth J, Ascierto ML, Engle AM, et al. Regression of
1030 melanoma metastases after immunotherapy is associated with activation of antigen presentation
1031 and interferon-mediated rejection genes. Int J Cancer J Int Cancer. 2012 Jul 15;131(2):387–95.

1032 23. Critchley-Thorne RJ, Simons DL, Yan N, Miyahira AK, Dirbas FM, Johnson DL, et al. Impaired
1033 interferon signaling is a common immune defect in human cancer. Proc Natl Acad Sci U S A. 2009
1034 Jun 2;106(22):9010–5.

1035 24. Baechler EC, Bilgic H, Reed AM. Type I interferon pathway in adult and juvenile dermatomyositis.
1036 Arthritis Res Ther. 2011;13(6):249.

1037 25. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set
1038 enrichment analysis: a knowledge-based approach for interpreting genome-wide expression
1039 profiles. Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15545–50.

1040 26. Godec J, Tan Y, Liberzon A, Tamayo P, Bhattacharya S, Butte AJ, et al. Compendium of Immune
1041 Signatures Identifies Conserved and Species-Specific Biology in Response to Inflammation.
1042 Immunity. 2016 Jan 19;44(1):194–206.

1043 27. Altman MC, Gill MA, Whalen E, Babineau DC, Shao B, Liu AH, et al. Transcriptome networks
1044 identify mechanisms of viral and nonviral asthma exacerbations in children. Nat Immunol.
1045 2019;20(5):637–51.

1046 28. Singhania A, Graham CM, Gabryšová L, Moreira-Teixeira L, Stavropoulos E, Pitt JM, et al.
1047 Transcriptional profiling unveils type I and II interferon networks in blood and tissues across
1048 diseases. Nat Commun. 2019 28;10(1):2887.

1049 29. Alsina L, Israelsson E, Altman MC, Dang KK, Ghandil P, Israel L, et al. A narrow repertoire of
1050 transcriptional modules responsive to pyogenic bacteria is impaired in patients carrying loss-of-
1051 function mutations in MYD88 or IRAK4. Nat Immunol. 2014 Dec;15(12):1134–42.

1052 30. Banchereau R, Baldwin N, Cepika A-M, Athale S, Xue Y, Yu CI, et al. Transcriptional specialization of
1053 human dendritic cell subsets in response to microbial vaccines. Nat Commun. 2014 Oct 22;5:5283.

1054 31. Altman MC, Baldwin N, Whalen E, al-shaikhly, taha, Presnell S, Khaenam P, et al. A Transcriptome
1055 Fingerprinting Assay for Clinical Immune Monitoring. bioRxiv [Internet]. Available from:
1056 https://www.biorxiv.org/content/10.1101/587295v1

1057 32. Rinchai D, Konza O, Hassler S, Martina F, Mejias A, Ramilo O, et al. Characterizing blood modular
1058 transcriptional repertoire perturbations in patients with RSV infection: a hands-on workshop using
1059 public datasets as a source of training material. bioRxiv. 2019 Jan 22;527812.

1060 33. 34th Annual Meeting & Pre-Conference Programs of the Society for Immunotherapy of Cancer
1061 (SITC 2019): part 1 : National Harbor, MD, USA. 6-10 November 2019. J Immunother Cancer. 2019
1062 Nov 6;7(Suppl 1):282.

1063 34. Newburger JW, Takahashi M, Gerber MA, Gewitz MH, Tani LY, Burns JC, et al. Diagnosis,
1064 Treatment, and Long-Term Management of Kawasaki Disease: A Statement for Health
1065 Professionals From the Committee on Rheumatic Fever, Endocarditis and Kawasaki Disease,
1066 Council on Cardiovascular Disease in the Young, American Heart Association. Circulation. 2004 Oct
1067 26;110(17):2747–71.

1068 35. Sugar CA, James GM. Finding the Number of Clusters in a Dataset: An Information-Theoretic
1069 Approach. J Am Stat Assoc. 2003;98(463):750–63.

1070 36. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional
1071 genomic data. Bioinforma Oxf Engl. 2016 15;32(18):2847–9.

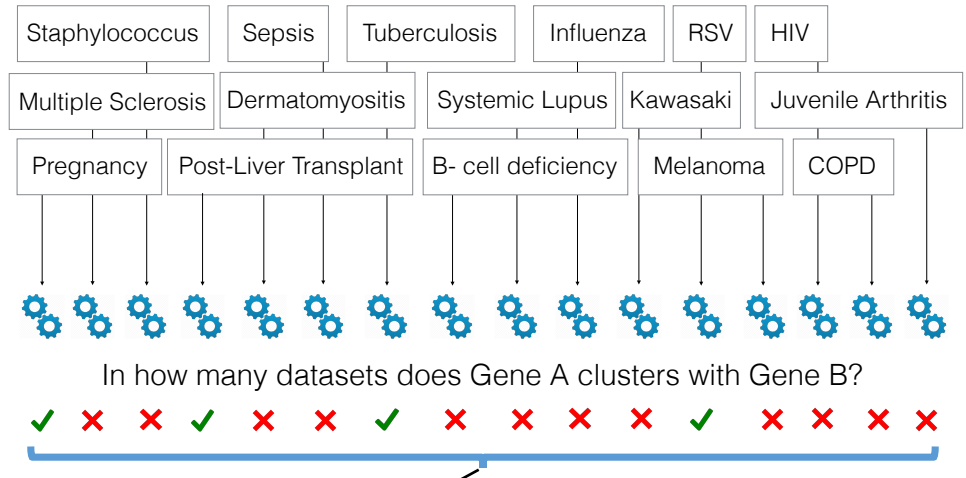
1072

# Module repertoire construction

**A. Input datasets**

16 diseases / physiological states

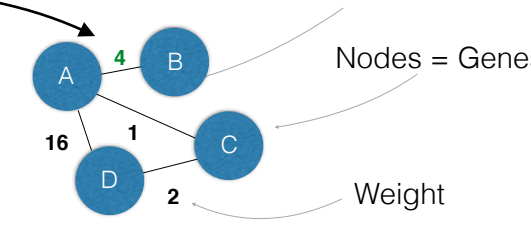
985 Samples



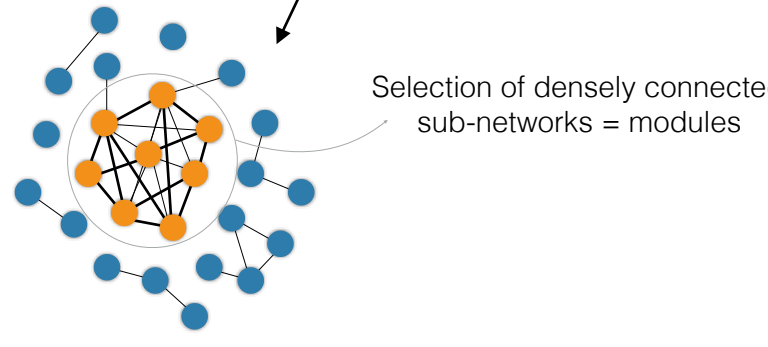

**B. K-means Clustering x 16**

**C. Co-Clustering Table**

| Gene | A | B | C | D | ... |
|---|---|---|---|---|---|
| A | | 4 | 1 | 16 | |
| B | 4 | | 0 | 0 | |
| C | 1 | 0 | | 2 | |
| D | 16 | 0 | 2 | | |
| ... | | | | | |




**D. Weighted Co-Clustering Network**




**E. Module repertoire**

(382 modules comprising 14,168 transcripts)

## Fingerprint visualization

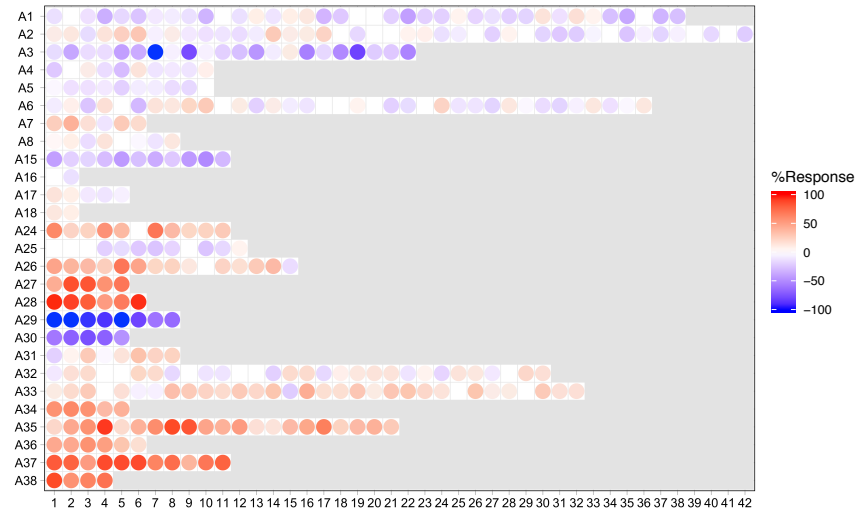
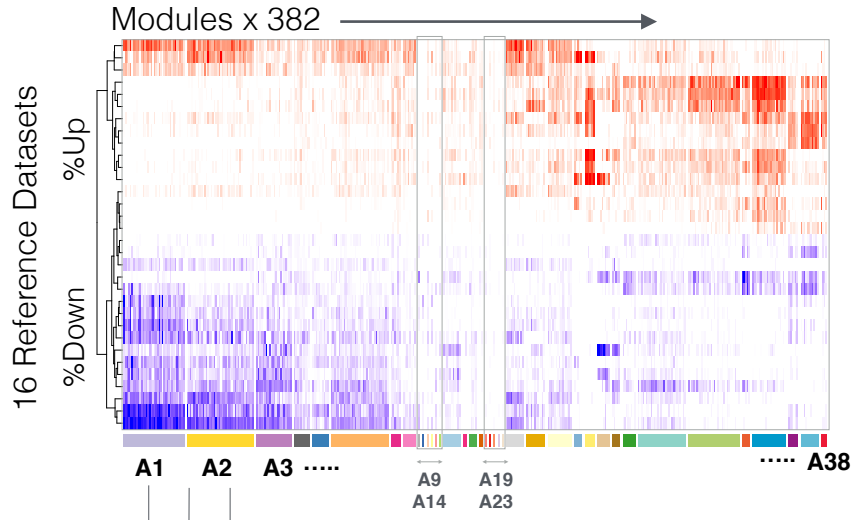


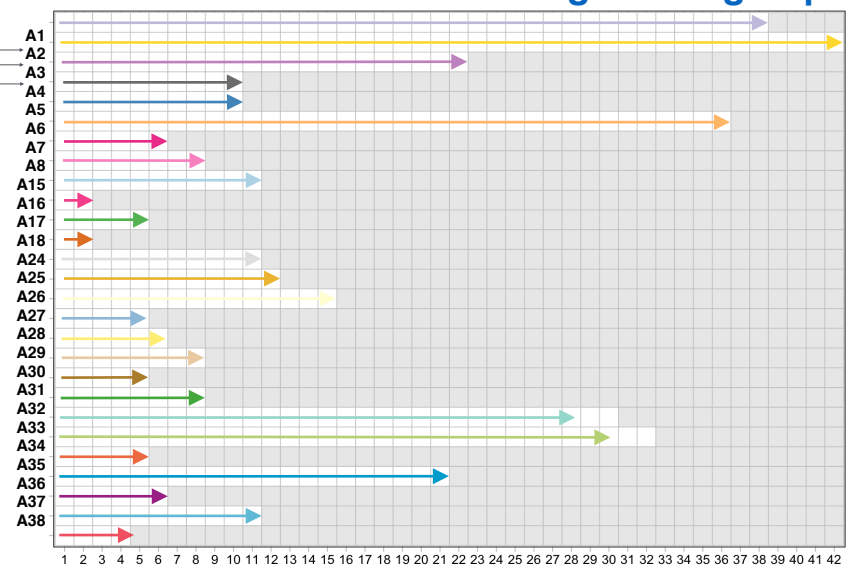

Figure 1

**Fingerprint Grid Visualization**

## A. Modules organized into "aggregates" (A1-A38)




## B. Position of each module assigned on grid plot




*Framework development*

---

*Routine data analyses*

Controls Cases

vs.

Decrease | No Change | Increase

## C. Changes in cases vs controls mapped

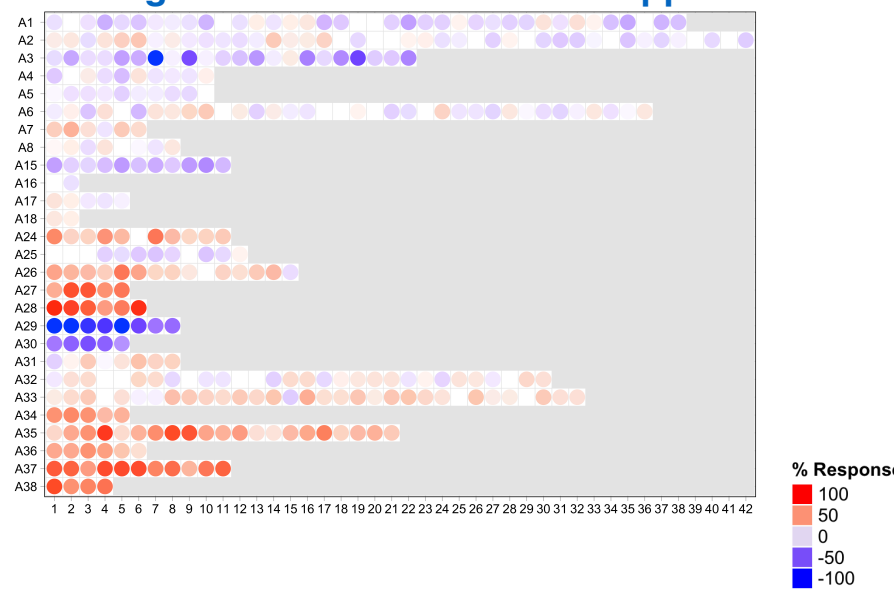
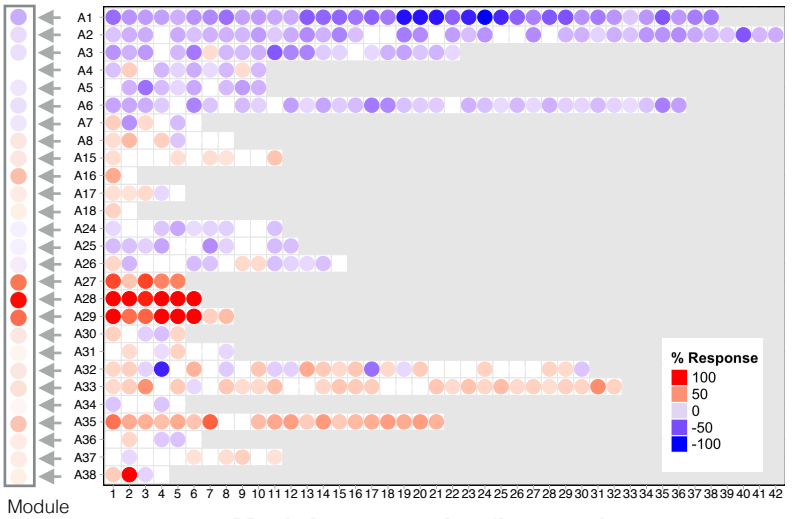
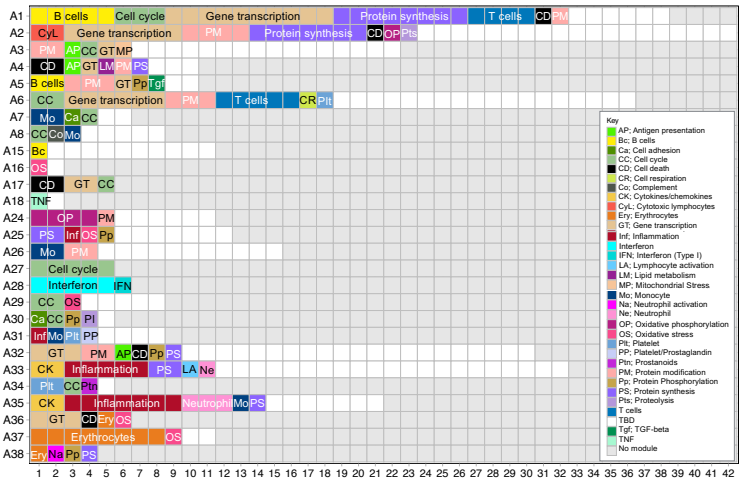



Figure 2

A. **SLE patients (N=55) / Control subjects (N=14)**

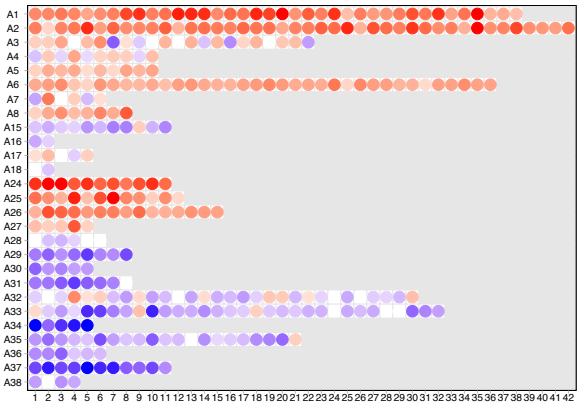



Module repertoire fingerprint

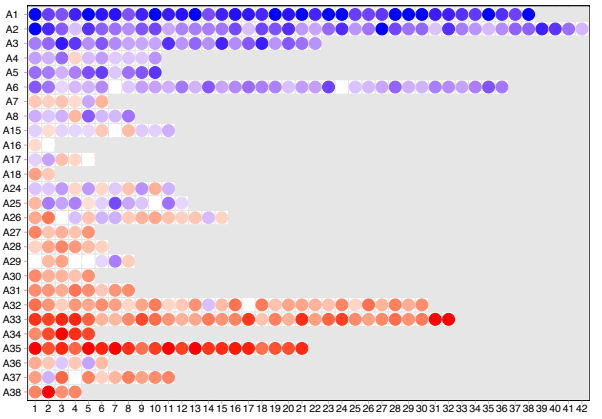



Module annotation grid

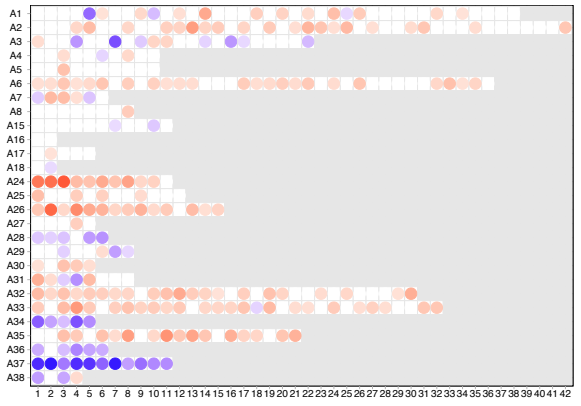
B. **MS**



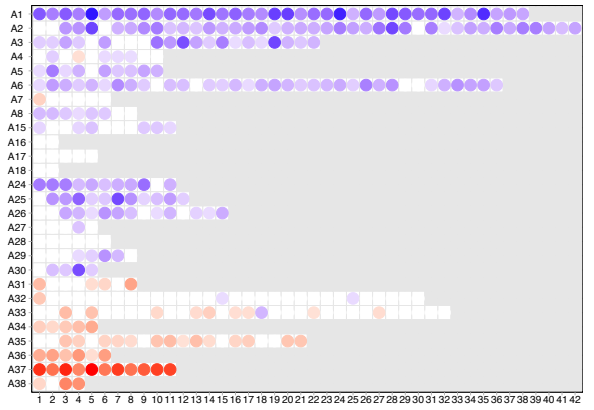
**S. aureus**



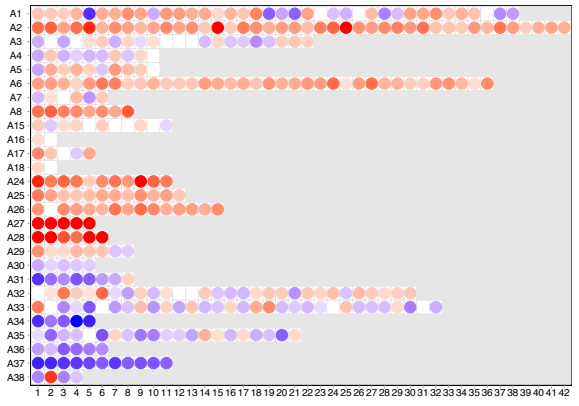
C. **COPD**



**Melanoma**



D. **HIV**



**TB**

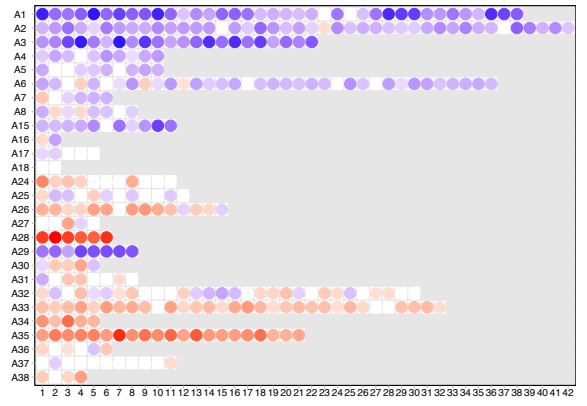
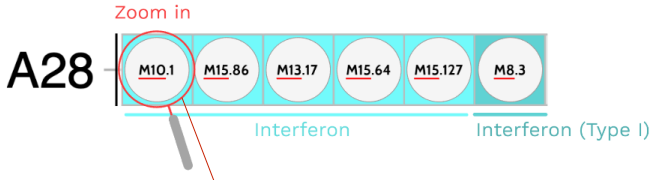
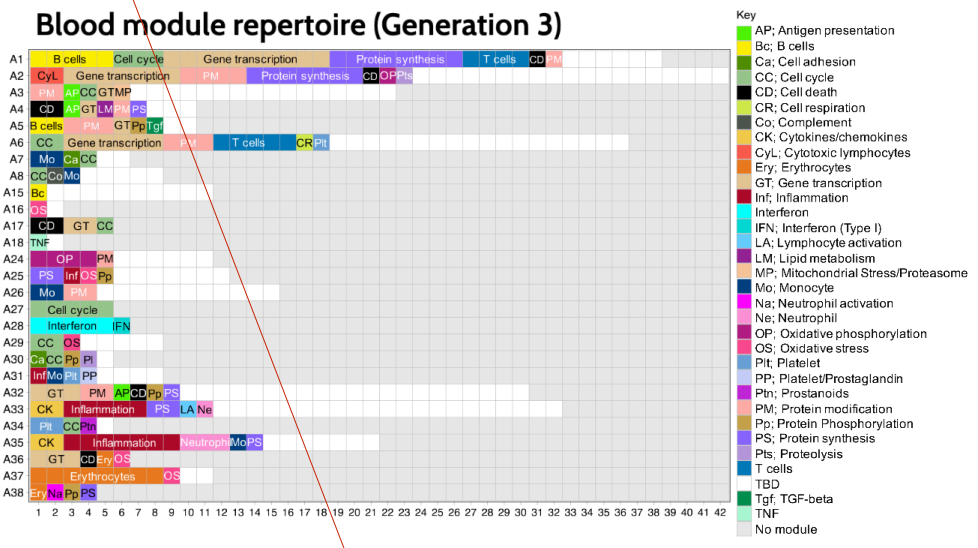


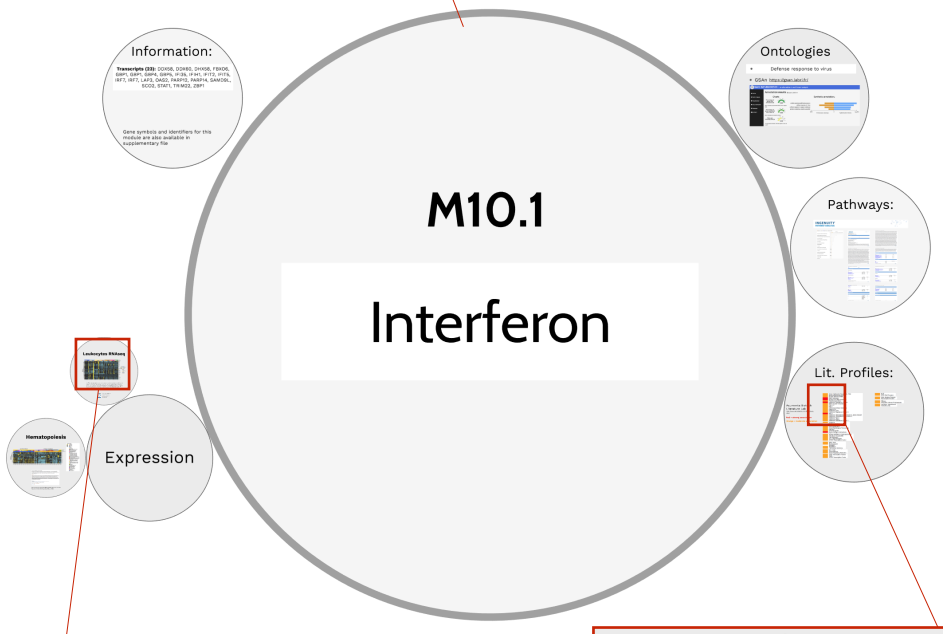
Figure 3

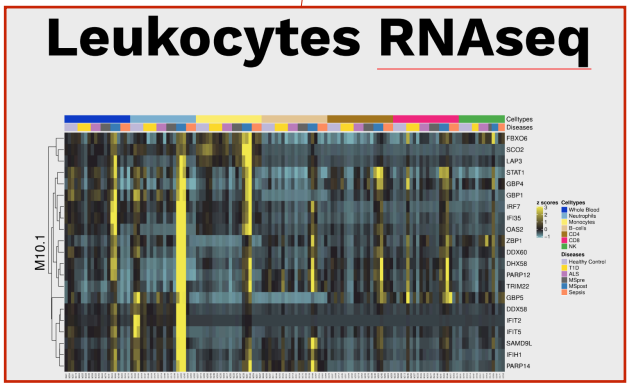
Web links: Table 3

Demo video: https://youtu.be/8ajESET2mqI

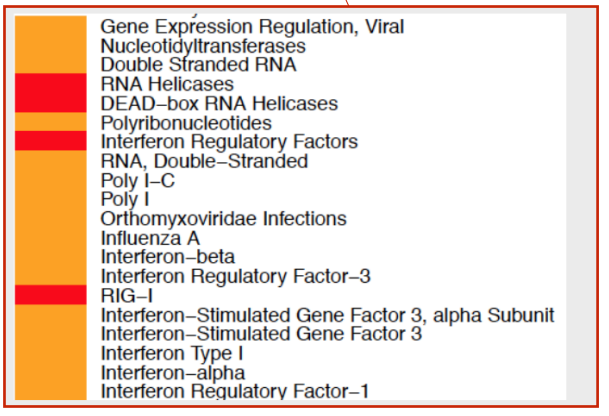



Reference transcriptional profiling data




Functional enrichment

Figure 4

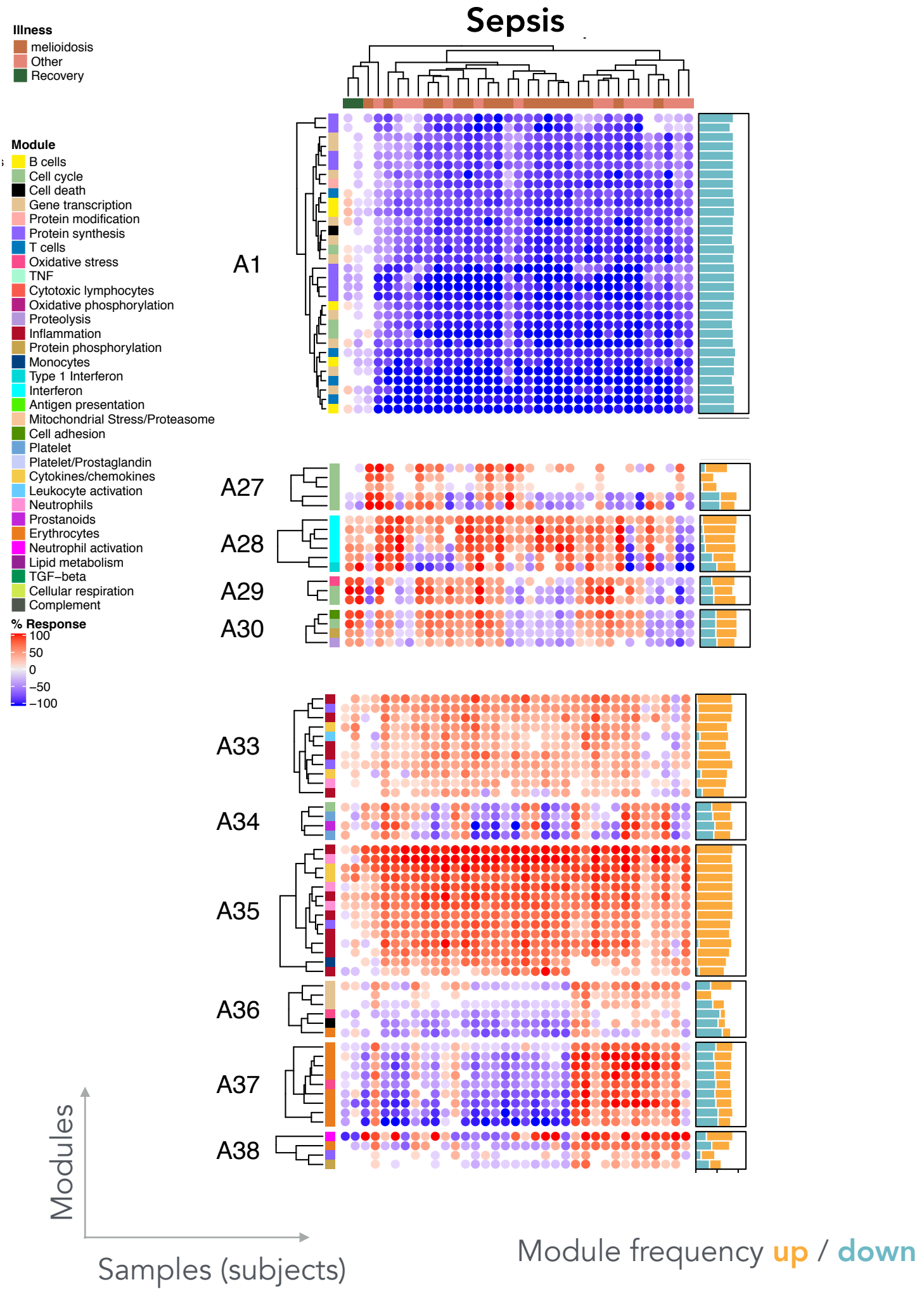

Sepsis
Illness
melioidosis
Other
Recovery
Module
B cells
Cell cycle
Cell death
Gene transcription
Protein modification
Protein synthesis
T cells
Oxidative stress
TNF
Cytotoxic lymphocytes
Oxidative phosphorylation
Proteolysis
Inflammation
Protein phosphorylation
Monocytes
Type 1 Interferon
Interferon
Antigen presentation
Mitochondrial Stress/Proteasome
Cell adhesion
Platelet
Platelet/Prostaglandin
Cytokines/chemokines
Leukocyte activation
Neutrophils
Prostanoids
Erythrocytes
Neutrophil activation
Lipid metabolism
TGF–beta
Cellular respiration
Complement
% Response
100
50
0
−50
−100
A1
A27
A28
A29
A30
A33
A34
A35
A36
A37
A38
Modules
Samples (subjects)
Module frequency up / down

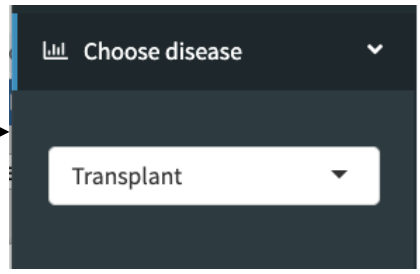
Figure 5

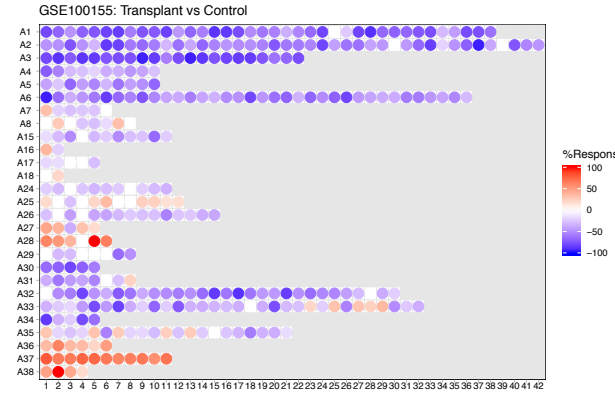
https://drinchai.shinyapps.io/dc_gen3_module_analysis/#

Demo video: https://youtu.be/y__7xKJo5e4

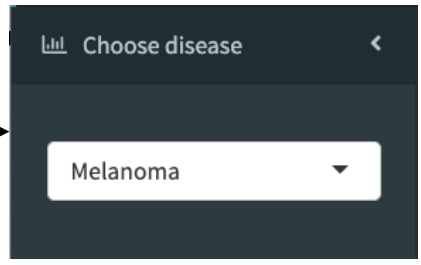
# Modular fingerprinting assay

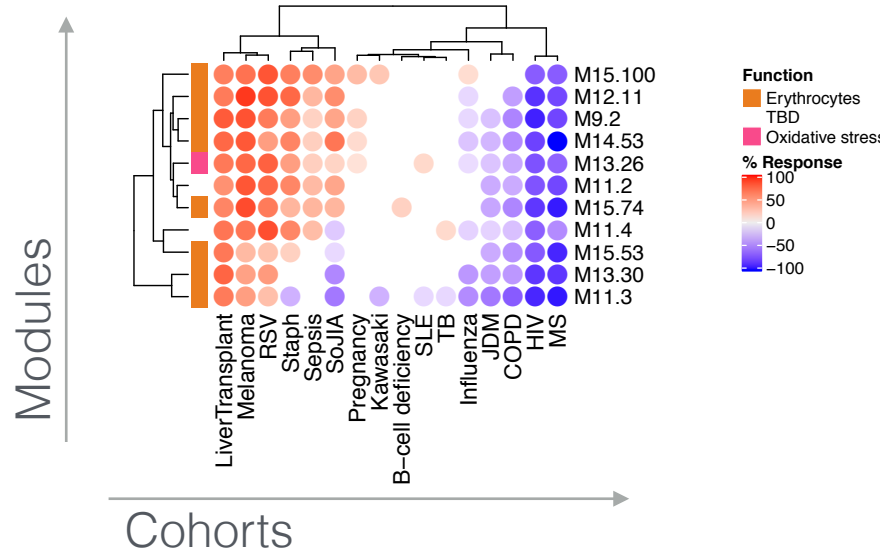
## Fingerprint grids



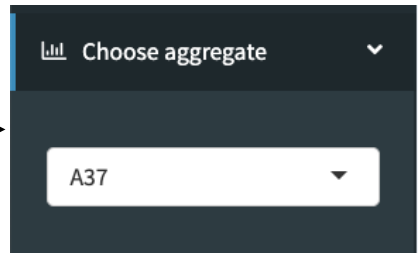




## Modules x Cohorts




## Modules x Individuals



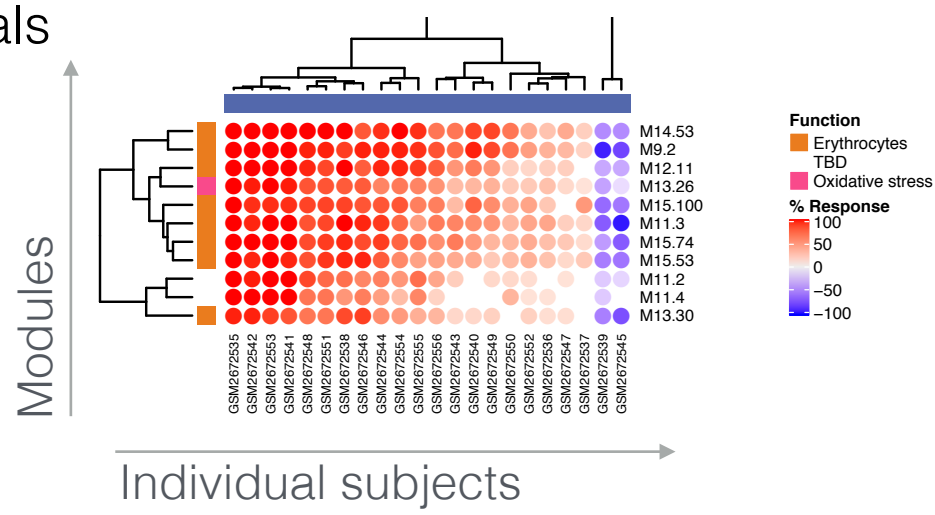




Figure 6

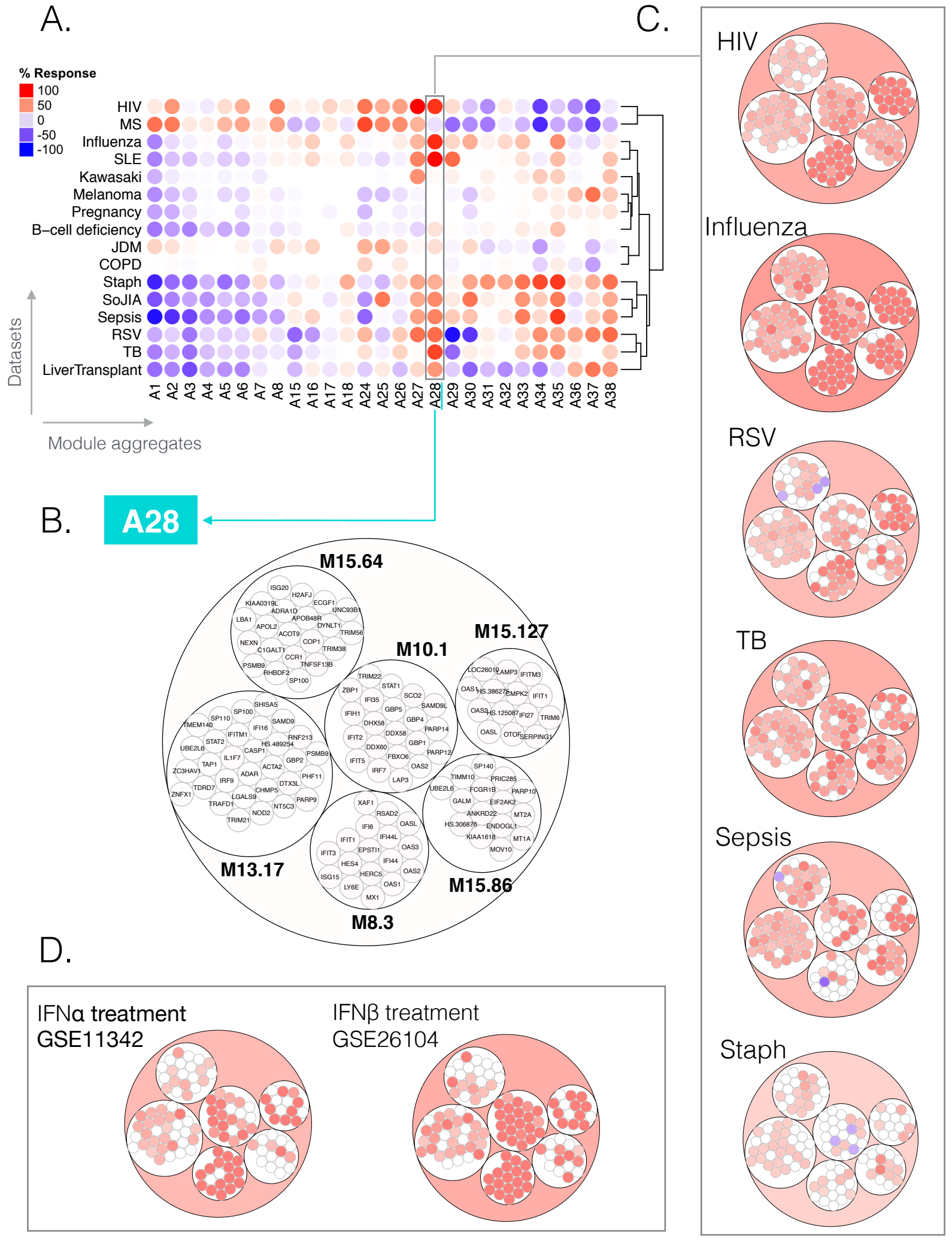

A.
% Response
100
50
0
-50
-100
HIV
MS
Influenza
SLE
Kawasaki
Melanoma
Pregnancy
B–cell deficiency
JDM
COPD
Staph
SoJIA
Sepsis
RSV
TB
LiverTransplant
Datasets
A1 A2 A3 A4 A5 A6 A7 A8 A15 A16 A17 A18 A24 A25 A26 A27 A28 A29 A30 A31 A32 A33 A34 A35 A36 A37 A38
Module aggregates
B.
A28
M15.64
M15.127
M10.1
M13.17
M8.3
M15.86
C.
HIV
Influenza
RSV
TB
Sepsis
Staph
D.
IFNα treatment
GSE11342
IFNβ treatment
GSE26104

Figure 7

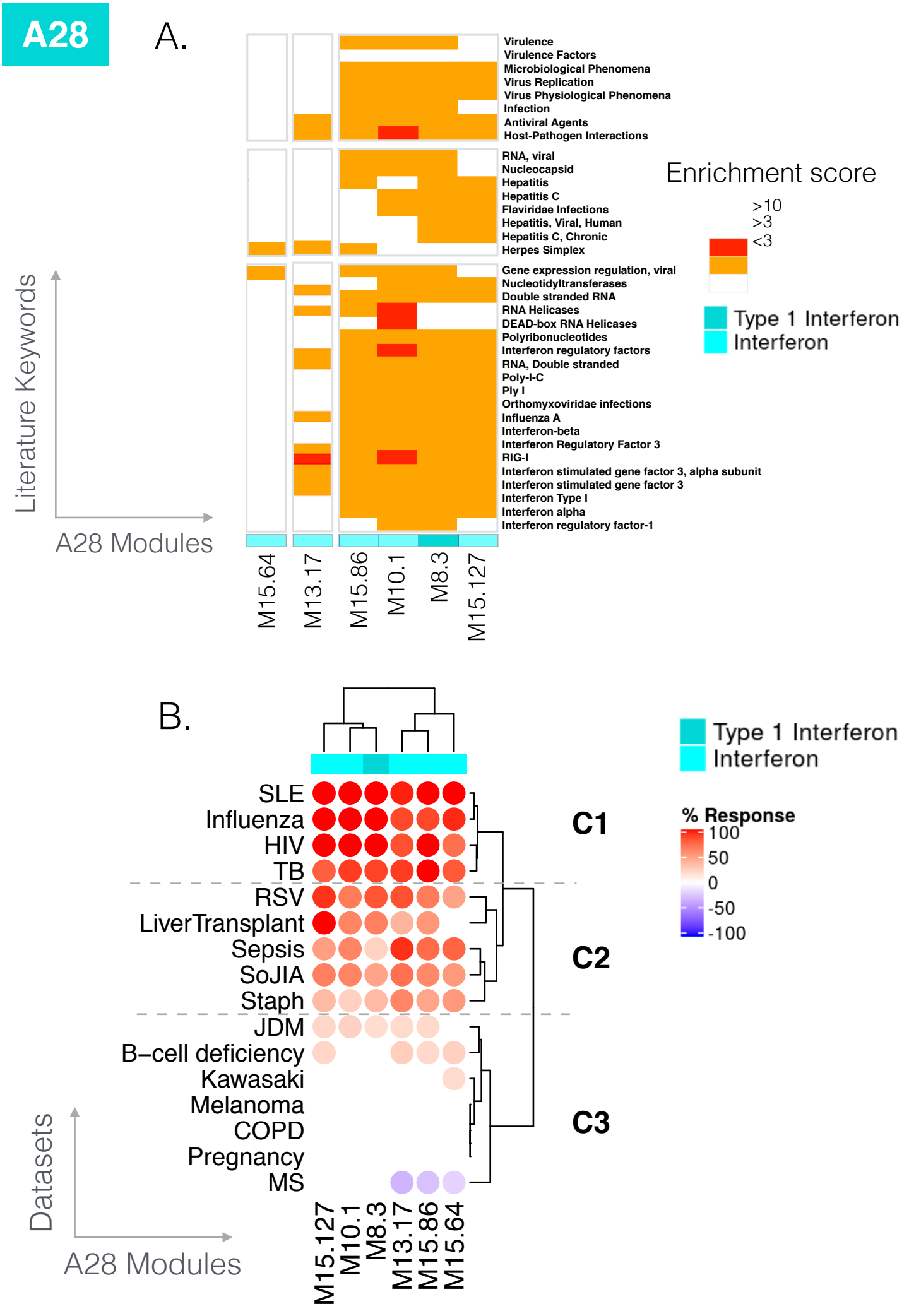

A28
A.
Virulence
Virulence Factors
Microbiological Phenomena
Virus Replication
Virus Physiological Phenomena
Infection
Antiviral Agents
Host-Pathogen Interactions
RNA, viral
Nucleocapsid
Hepatitis
Hepatitis C
Flaviridae Infections
Hepatitis, Viral, Human
Hepatitis C, Chronic
Herpes Simplex
Gene expression regulation, viral
Nucleotidyltransferases
Double stranded RNA
RNA Helicases
DEAD-box RNA Helicases
Polyribonucleotides
Interferon regulatory factors
RNA, Double stranded
Poly-I-C
Ply I
Orthomyxoviridae infections
Influenza A
Interferon-beta
Interferon Regulatory Factor 3
RIG-I
Interferon stimulated gene factor 3, alpha subunit
Interferon stimulated gene factor 3
Interferon Type I
Interferon alpha
Interferon regulatory factor-1
Enrichment score
>10
>3
<3
Type 1 Interferon
Interferon
Literature Keywords
A28 Modules
M15.64
M13.17
M15.86
M10.1
M8.3
M15.127
B.
Type 1 Interferon
Interferon
% Response
100
50
0
-50
-100
SLE
Influenza
HIV
TB
RSV
LiverTransplant
Sepsis
SoJIA
Staph
JDM
B–cell deficiency
Kawasaki
Melanoma
COPD
Pregnancy
MS
C1
C2
C3
Datasets
A28 Modules
M15.127
M10.1
M8.3
M13.17
M15.86
M15.64

Figure 8

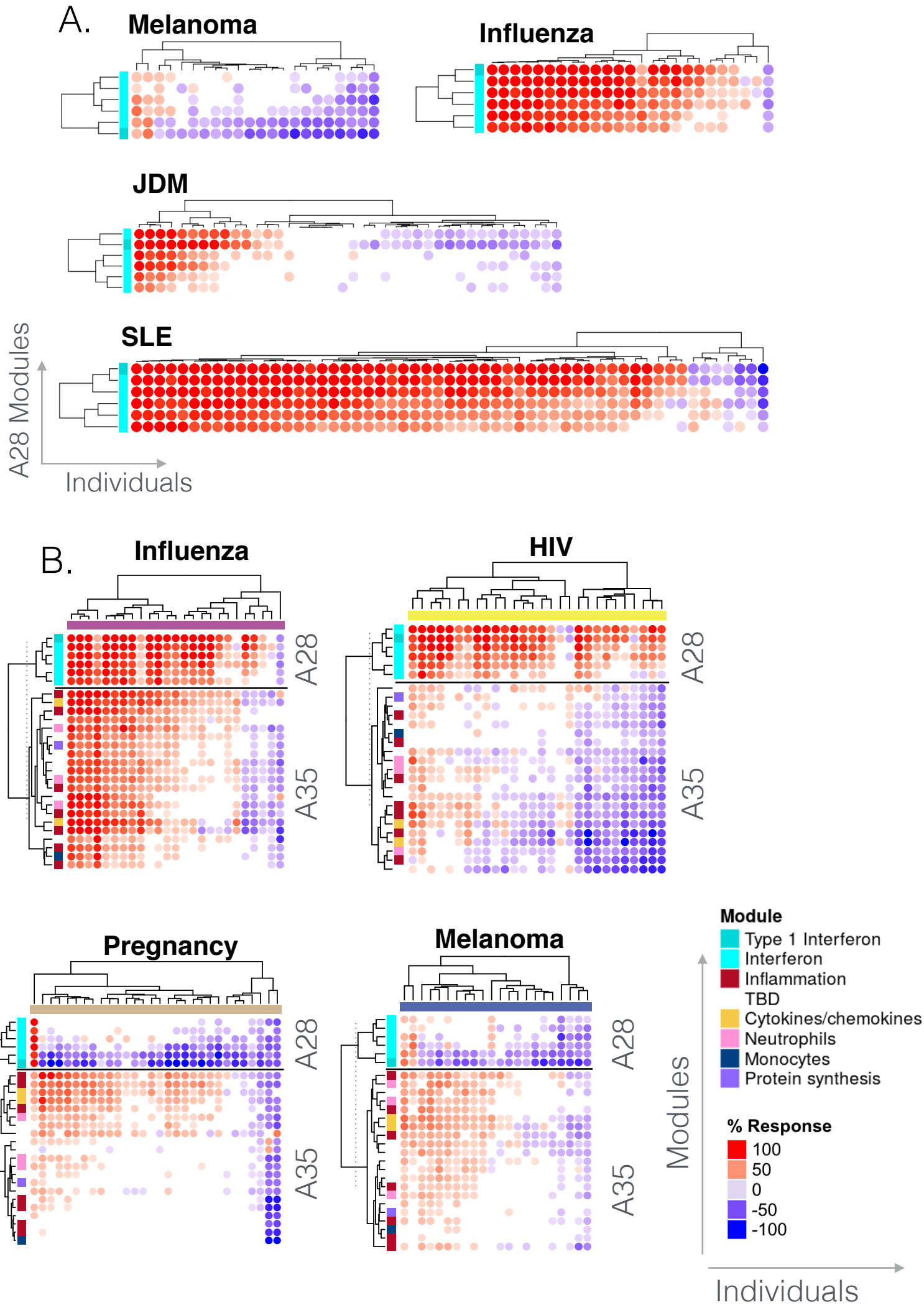

A.
Melanoma
Influenza
JDM
SLE
A28 Modules
Individuals
B.
Influenza
HIV
Pregnancy
Melanoma
A28
A35
Modules
Individuals
Module
Type 1 Interferon
Interferon
Inflammation
TBD
Cytokines/chemokines
Neutrophils
Monocytes
Protein synthesis
% Response
100
50
0
-50
-100

Figure 9

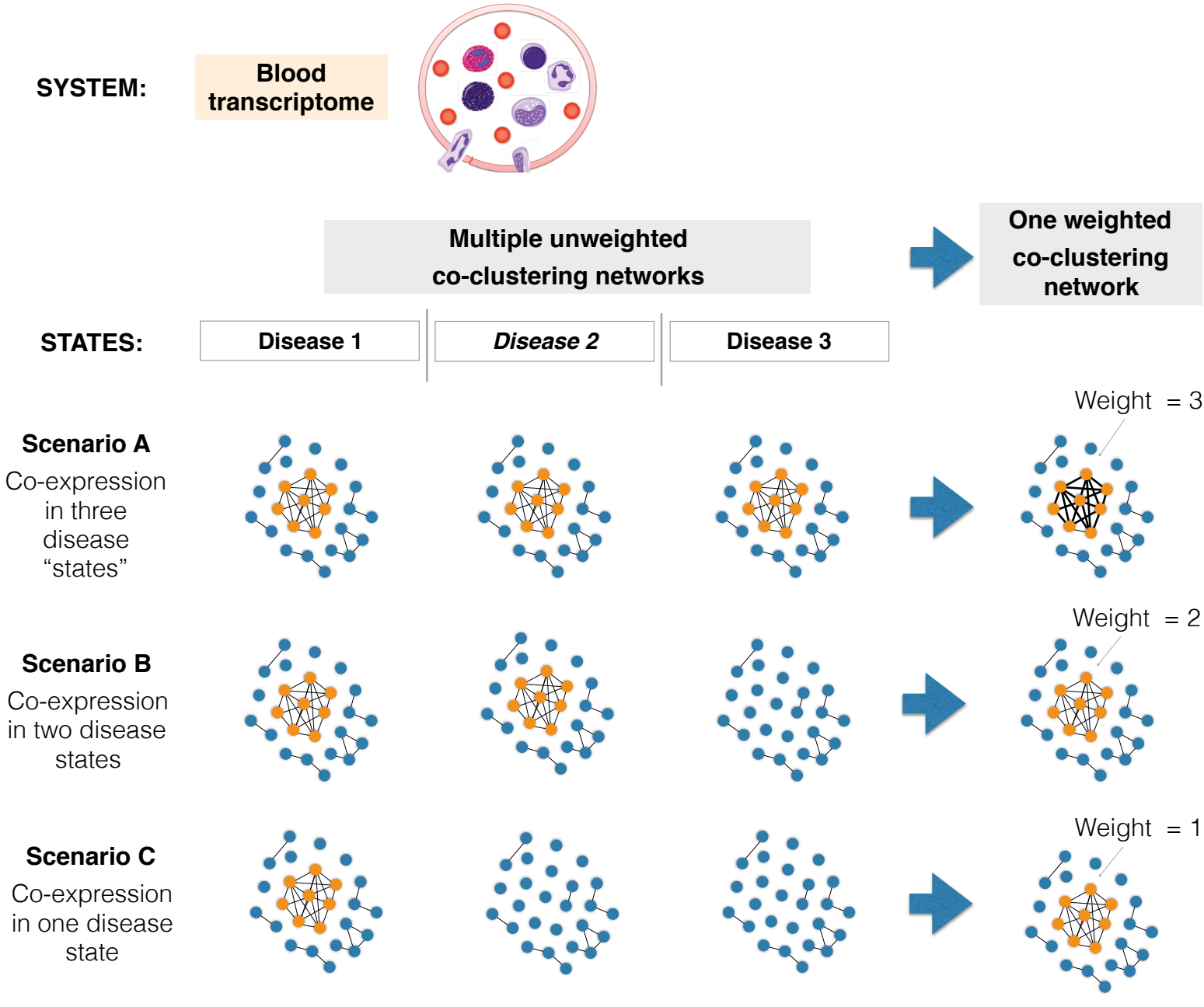

SYSTEM:
Blood transcriptome
Multiple unweighted co-clustering networks
One weighted co-clustering network
STATES:
Disease 1
Disease 2
Disease 3
Scenario A
Co-expression in three disease "states"
Weight = 3
Scenario B
Co-expression in two disease states
Weight = 2
Scenario C
Co-expression in one disease state
Weight = 1

Figure S1

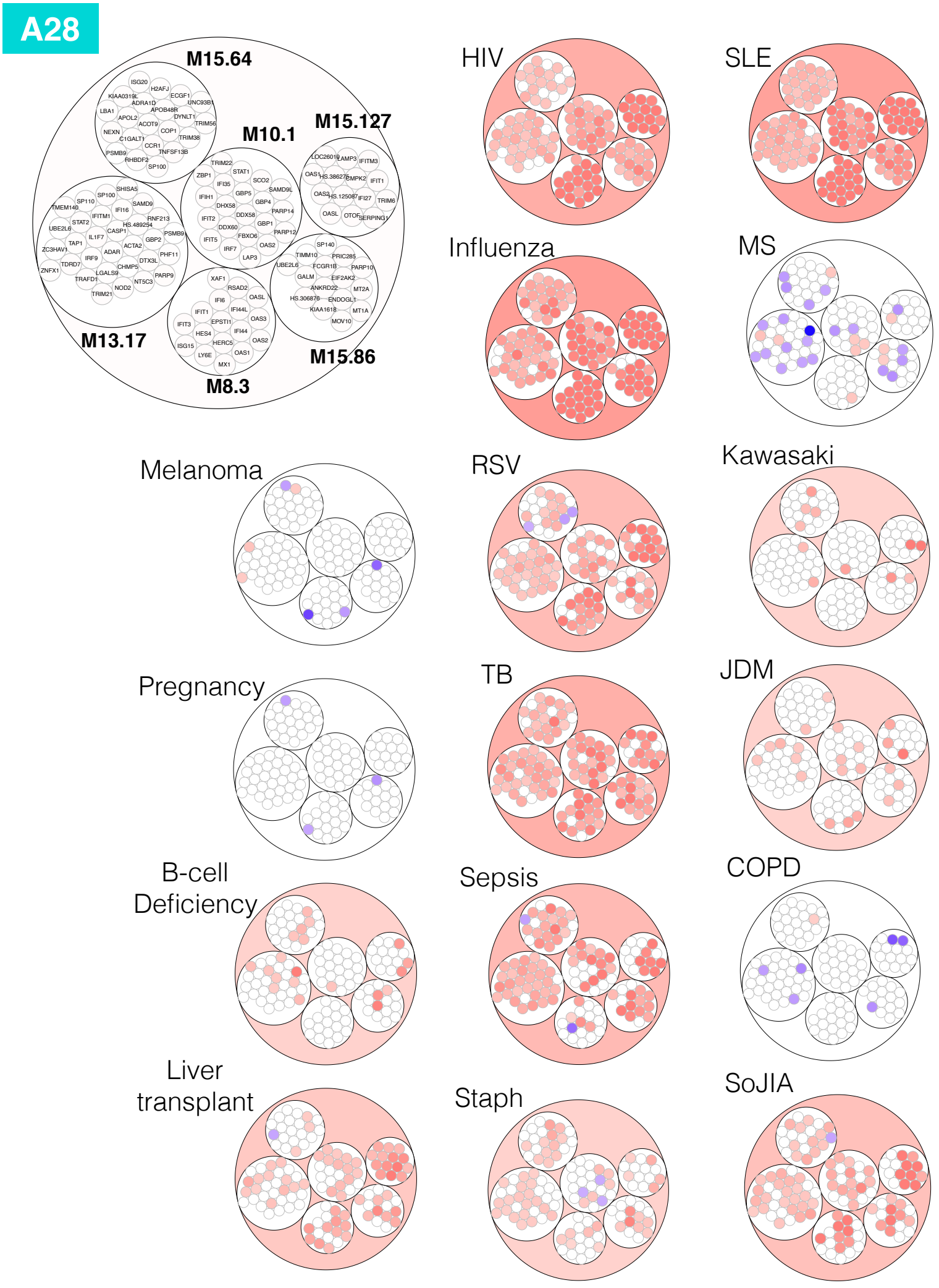

A28
M15.64
M10.1
M15.127
M13.17
M8.3
M15.86
HIV
SLE
Influenza
MS
Melanoma
RSV
Kawasaki
Pregnancy
TB
JDM
B-cell Deficiency
Sepsis
COPD
Liver transplant
Staph
SoJIA

Figure S2

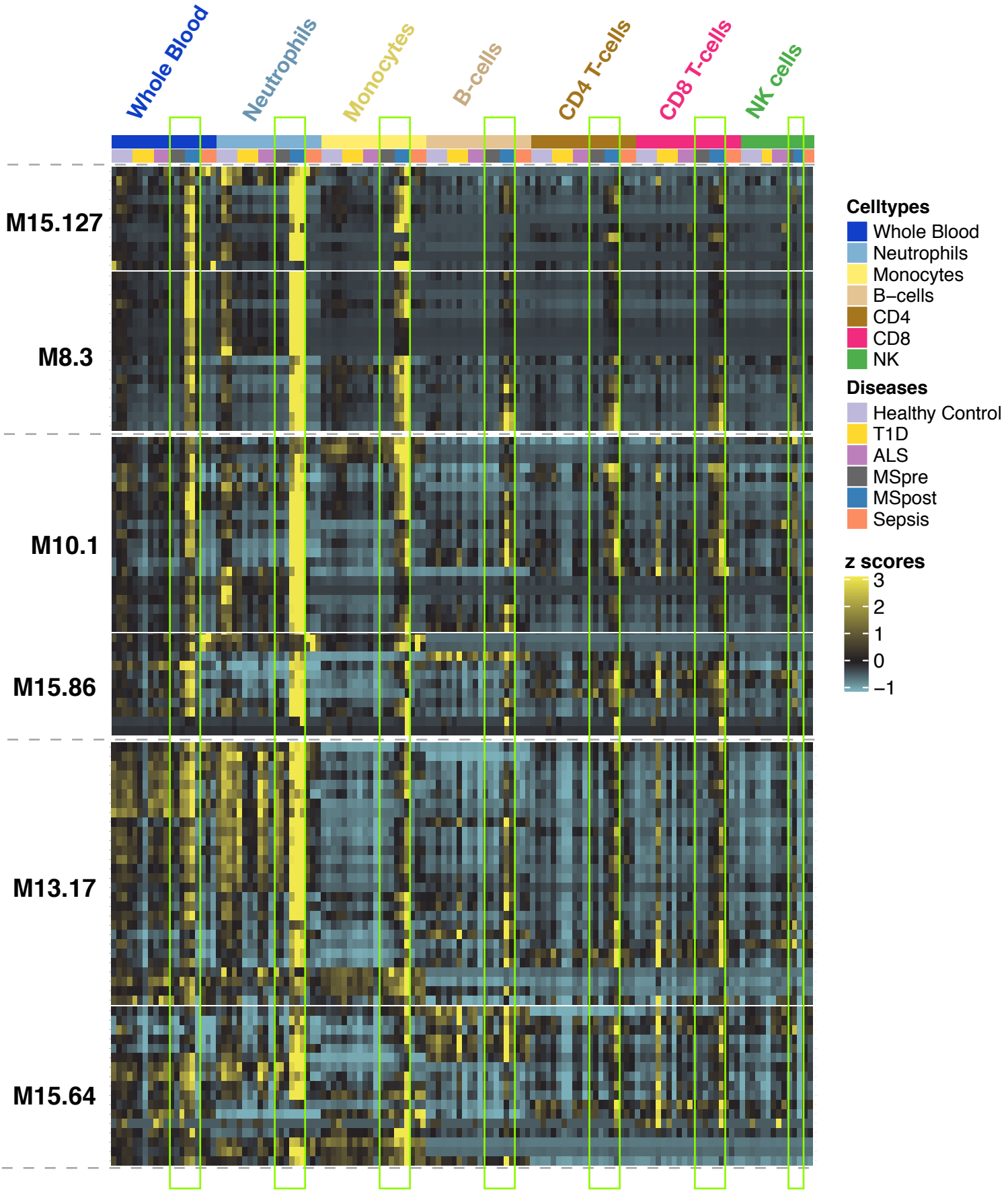

Whole Blood
Neutrophils
Monocytes
B-cells
CD4 T-cells
CD8 T-cells
NK cells
M15.127
M8.3
M10.1
M15.86
M13.17
M15.64
Celltypes
Whole Blood
Neutrophils
Monocytes
B–cells
CD4
CD8
NK
Diseases
Healthy Control
T1D
ALS
MSpre
MSpost
Sepsis
z scores
3
2
1
0
−1

Figure S3

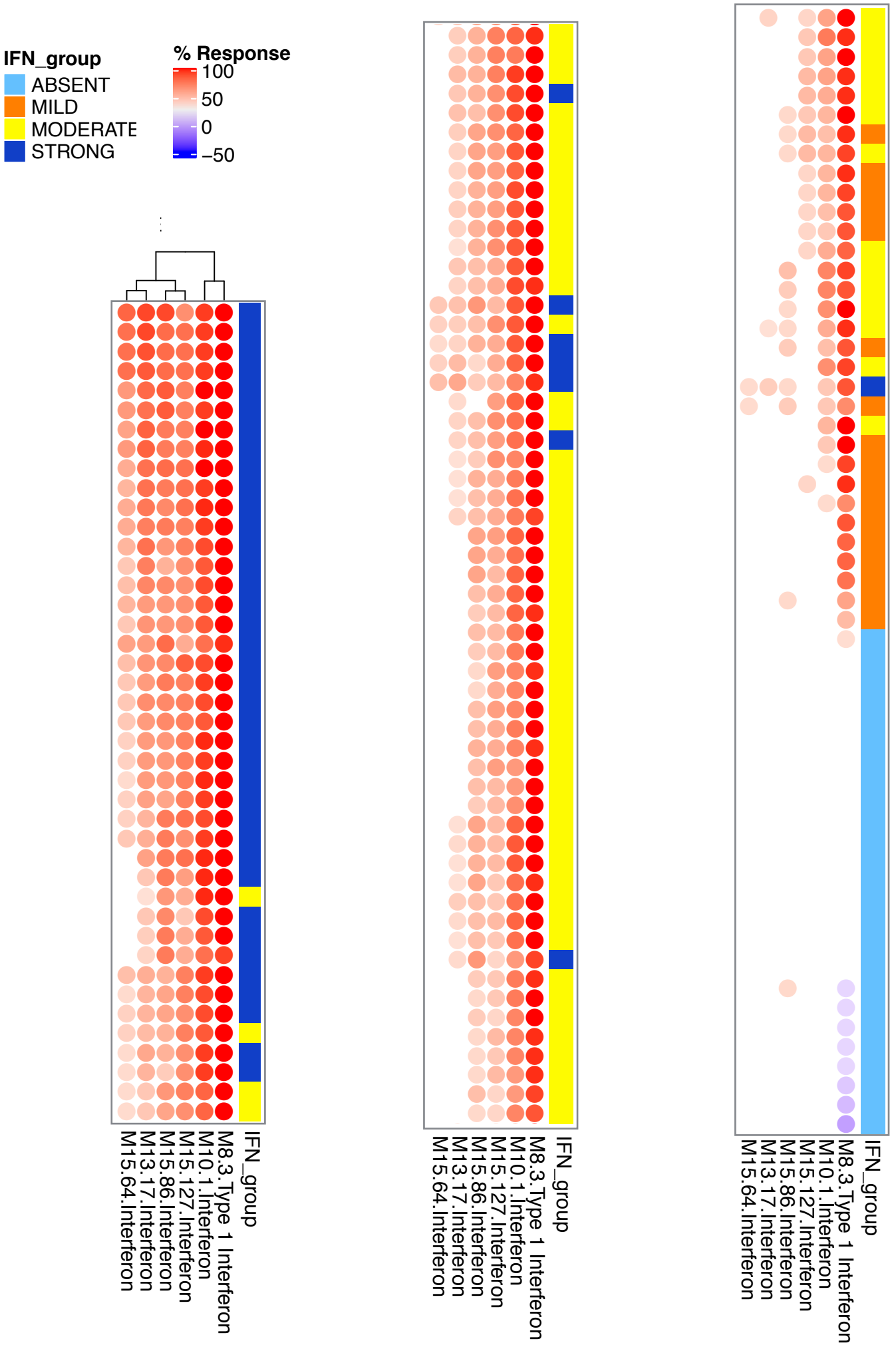

IFN_group
ABSENT
MILD
MODERATE
STRONG
% Response
100
50
0
−50
IFN_group
M8.3.Type 1 Interferon
M10.1.Interferon
M15.127.Interferon
M15.86.Interferon
M13.17.Interferon
M15.64.Interferon

Figure S4