Check for updates

METHOD ARTICLE

REVISED ANIMA: Association network integration for multiscale analysis [version 3; referees: 2 approved, 1 approved with reservations]

Armin Deffur[1], Robert J. Wilkinson[1-4], Bongani M. Mayosi[1], Nicola M. Mulder[5]

[1]Department of Medicine, University of Cape Town, Cape Town, 7925, South Africa
[2]Wellcome Centre for Infectious Diseases Research in Africa, University of Cape Town, Cape Town, 7925, South Africa
[3]Francis Crick Institute, London, NW1 1AT, UK
[4]Imperial College London, London, W2 1PG, UK
[5]Computational Biology Division, Department Integrative Biomedical Sciences, IDM, University of Cape Town, Cape Town, 7925, South Africa

v3 **First published:** 12 Mar 2018, **3**:27 (https://doi.org/10.12688/wellcomeopenres.14073.1)

Second version: 05 Jun 2018, **3**:27 (https://doi.org/10.12688/wellcomeopenres.14073.2)

Latest published: 14 Nov 2018, **3**:27 (https://doi.org/10.12688/wellcomeopenres.14073.3)

Abstract

Contextual functional interpretation of -omics data derived from clinical samples is a classical and difficult problem in computational systems biology. The measurement of thousands of data points on single samples has become routine but relating 'big data' datasets to the complexities of human pathobiology is an area of ongoing research. Complicating this is the fact that many publicly available datasets use bulk transcriptomics data from complex tissues like blood. The most prevalent analytic approaches derive molecular 'signatures' of disease states or apply modular analysis frameworks to the data. Here we describe ANIMA (association network integration for multiscale analysis), a network-based data integration method using clinical phenotype and microarray data as inputs. ANIMA is implemented in R and Neo4j and runs in Docker containers. In short, the build algorithm iterates over one or more transcriptomics datasets to generate a large, multipartite association network by executing multiple independent analytic steps (differential expression, deconvolution, modular analysis based on co-expression, pathway analysis) and integrating the results. Once the network is built, it can be queried directly using Cypher (a graph query language), or by custom functions that communicate with the graph database via language-specific APIs. We developed a web application using Shiny, which provides fully interactive, multiscale views of the data. Using our approach, we show that we can reconstruct multiple features of disease states at various scales of organization, from transcript abundance patterns of individual genes through co-expression patterns of groups of genes to patterns of cellular behaviour in whole blood samples, both in single experiments as well in meta-analyses of multiple datasets.

Open Peer Review

Referee Status: ✓ ? ✓

	Invited Referees 1	 2	 3
REVISED **version 3** published 14 Nov 2018			✓ report
REVISED **version 2** published 05 Jun 2018	✓ report		? report
version 1 published 12 Mar 2018	? report	? report	

1 **Christopher L. Plaisier**, Arizona State University, USA

2 **James A. Eddy**, Sage Bionetworks, USA

3 **Emre Guney**, Pompeu Fabra University, Spain

Discuss this article

Comments (0)

Keywords

Transcriptomics, complex networks, graph databases, data integration

This article is included in the The Francis Crick Institute gateway.

Corresponding author: Armin Deffur (a.deffur@uct.ac.za)

Author roles: Deffur A: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Wilkinson RJ**: Conceptualization, Funding Acquisition, Methodology, Project Administration, Resources, Supervision, Writing – Review & Editing; **Mayosi BM**: Conceptualization, Funding Acquisition, Project Administration, Supervision, Writing – Review & Editing; **Mulder NM**: Conceptualization, Methodology, Project Administration, Resources, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work is supported by the Wellcome Trust [104803, 203135]. This study was also supported by the Medical Research Council of South Africa, Hasso Plattner Foundation, and Newton Fund [NA150173]. This work was supported by the Francis Crick Institute which receives its core funding from Cancer Research UK (FC001218), the UK Medical Research Council (FC001218), and the Wellcome Trust (FC001218).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Deffur A, Wilkinson RJ, Mayosi BM and Mulder NM. **ANIMA: Association network integration for multiscale analysis [version 3; referees: 2 approved, 1 approved with reservations]** Wellcome Open Research 2018, **3**:27 (https://doi.org/10.12688/wellcomeopenres.14073.3)

First published: 12 Mar 2018, **3**:27 (https://doi.org/10.12688/wellcomeopenres.14073.1)

REVISED **Amendments from Version 2**

Minor changes were made for this version:

1. Support from the Crick Institute is now acknowledged in the "Grant information" section
2. The Introduction was revised to incorporate additional references to related work, and the logical flow has been improved so that it reads more clearly

See referee reports

Introduction

A frequent issue with bioinformatic analysis is the following scenario: a given dataset is analysed using various approaches in a linear workflow, in an attempt to extract biological features of interest from the data, often at different scales. For instance, a list of differentially abundant transcripts provides information on the system that differs from a list of co-expressed genes. While both such approaches are valid and provide independently useful information, these linear workflows do not not expose potentially informative relationships between different outputs; instead, multiple individual output items (plots, tables, etc.) are written to the output directory. However, multiple kinds of relationships, or associations, exist between entities of the same or different class in biological systems. For instance, two clinical variables might be correlated; the relationship is then expressed as correlation, its extent as the Pearson correlation coefficient, and the statistical significance of the relationship by a P-value, which may be corrected for multiple testing.

Such relationships can be discovered systematically in the analytic workflow. Typically, the result of such an analysis is written to disk in tabular form. However, a table of correlation statistics contains many non-significant entries, and therefore has a low signal to noise ratio, especially in large datasets. A more efficient approach would be to discard all results that fail to meet a pre-specified statistical cutoff and retain only the significant entries as a monopartite or bipartite network. Indeed, bipartite graphs are the dominant paradigm used to encode and represent information about *relationships* between entities of different classes. Bipartite graphs in systems biology and medicine have recently been reviewed[1], illustrating numerous use cases and applications as well as describing their mathematical properties. Once the decision has been made to store analytic results and their relationships as bipartite graphs (or other types of network), the next question is where and how to store the data. Cytoscape[2] is extensively used for visualising network data, mainly in biology. Large networks can be searched and subsetted, and basic network analysis can be performed. However, using Cytoscape for a network data store rather than a visualisation tool becomes cumbersome. Recently, graph databases have emerged as a type of noSQL database. Data in a graph database is stored as nodes and edges (relationships between nodes). Both nodes and edges can be assigned multiple properties with values. One such database is Neo4j, which is freely available. Neo4j databases are queried using Cypher, an intuitive query language, and provide computational access to scripting languages via dedicated APIs. In our view, graph databases are the ideal data store for network-type data.

Reproducibility of research is critical for scientific progress. Bioinformatic analyses can be very complex, but usually the results obtained depend strongly on the methods used, software versions, and even operating system. Typical narrative descriptions of analytic methods provide insufficient information to guarantee reproducibility. Workflow tools like Taverna[3] exist but are more suited to performing tasks that link together functionality offered by services (such as webservices). We prefer scripted workflows, where running and re-running the same script on the same data is guaranteed to produce identical results. However, providing the code used in analysis and the raw data does not guarantee reproducibility, as the computational environment in which the analysis is run can also influence the outcome of computations. Typically, this becomes an issue when the functionality of a software package changes between versions. Therefore, in addition to data and source code, one has to provide the exact configuration and package versions of all software involved in the project. In bioinformatics, this can become daunting very quickly, and leads to the well-known problem of "dependency hell" where not only the software packages need the right version, but also their dependencies. A solution for this is to package the entire computational environment in one or more "software containers". The containerization platform Docker[4] is frequently used to fulfil this function and has enjoyed widespread adoption in reproducible research. This is exemplified by Nextflow[5], a workflow management system using Docker.

A recent paper presents GeNNet[6], which describes the rationale for scripted workflows and the use of graph databases in reproducible research. In this paper a scripted workflow in R[7], use of Neo4j to store data and the use of Docker for reproducibility is described. Other recent efforts in the very wide field of -OMICS data integration, research reproducibility and data integration include the Omics Integrator package[8] which takes a variety of -OMIC data (such as transcriptomic and proteomic) as input and identifies possible underlying molecular pathways using network optimization algorithms, NDEx[9], an online commons for sharing and searching biological networks.

Here we present ANIMA, Association Network Integration for Multiscale Analysis, a framework for producing and interrogating a multiscale association network, which allows summary and visualization of different, but simultaneously valid views of the state of the immune system under different conditions and at multiple scales. While ANIMA employs key strategies presented in the GeNNet paper, mainly "dockerisation", scripted workflows in R and storage of relationships in a graph database, it differs in detail in implementation as well as complexity. In particular, we have defined novel functionality that allows the extraction of new information form the network structure in ANIMA, given an increased number of node types and increased complexity of the data model of the Neo4j database. In addition, GeNNet is implemented in a single container containing R and Neo4j, whereas ANIMA uses separate containers for each tool.

ANIMA is targeted both at bioinformaticians and computational biologists who wish to reproduce the results presented or analyse their own data, as well as biologists who would interact with the system through the provided graphical user interface.

In its current form, ANIMA is best suited for investigating human immunity. The human immune system can be regarded as a complex adaptive system[10], even though it is integrated with the more complex system of the whole organism. A true systems view[11] of the immune system needs to account for the various aspects that characterize complex adaptive systems generally. This includes emergence, non-linearity, self-organization, noise, scaling, heterogeneity, a network architecture and preservation of context for individual observations[12]. Whole blood is a "window" into the immune system[13], allowing a reasonably detailed assessment of the overall state of the immune system based on analysis of mRNA transcript abundance patterns from whole blood samples. Here we use the responses to three common infections (acute HIV infection, malaria and respiratory viral infections), as measured by transcript abundance patterns in whole blood, to demonstrate ANIMA functionality

Methods

Method overview

ANIMA generates a multiscale association network (stored as a graph database) from multiple data types (expression data, clinical data and annotation data, e.g. biological pathways databases) by executing a comprehensive analytic workflow, enumerating bipartite graphs from the results, and merging all graphs into a single network. Figure 1 provides a conceptual overview of the multiscale analytics pipeline. As biological systems can be understood at multiple scales, our approach aims to integrate information across multiple scales which range from single genes to clinical phenotypes (Figure 1A). This integration is summarised in the core data model for ANIMA (Figure 1B) which represents relationships between various outputs of the analytic pipeline. Multiple bipartite graphs are generated from the outputs, and finally merged into a single data structure (Figure 1 C–E), which is then accessed to gain novel information about the system. The various steps as implemented in the ANIMA_build script are illustrated in Figure 1F.

Data ingestion and preparation

The first step in constructing an ANIMA database consists of accessing raw data, which consists of non-normalised Illumina BeadArray expression microarray data and clinical/phenotype data. A script (ANIMA_data.R) imports the data to the R workspace as well as the clinical data and creates a LumiBatch object of the imported data. Based on the experimental design, the data may be subsetted at this point. One or more LumiBatch objects are then saved to the project output folder for later re-use.

Scripted workflow execution

The second step in constructing an ANIMA database consists of iterating over all datasets included in the analysis (each saved as an RData file), sequentially performing thirteen analytic tasks (A1–A13 in Table 1 and Figure S7; see Supplementary Data for details). Array normalization, probe filtering and differential expression analysis are frequently performed on transcriptomics datasets, but the other analyses tend to be performed in isolation. To our knowledge, this is the first transcriptomics workflow to combine WGCNA and Chaussabel modular analyses and integrate these with cellular-level approaches. The chosen approaches were selected based on their perceived popularity in the systems immunology literature, and selected annotation sources are biased towards immunology and cell-type specific data sets. This does not exclude the possibility of other approaches or annotation sets to be substituted in their place if the focus of investigation concerns other tissues or species of interest.

Enumeration of bipartite graphs

Following the transcriptional analytic tasks, the algorithm constructs twenty-nine bipartite networks from a multi-scale analytic pipeline (Figure 1A, C), combining the output, and the relationships between different classes of output approaches (Table 1). Each network contains the associations between two distinct data types (Figure 1B, D, Table 2, Table 3). The final association network is a result of graph union, merging all networks (Table 3) on shared node types while retaining all edges (Figure 1E). This results in a data structure that exposes the relationships between modular gene expression and higher-level phenomena, while retaining key probe- and gene-level information. Three types of association (with their respective association indices) are utilised, resulting in three distinct edge types in the final data structure (See Figure 1B, 1E, Table 2 and supplementary methods; Supplementary File 1). Weighted gene co-expression network analysis[14] (WGCNA) is the core analytic method in ANIMA as this is used to discover biological processes in the system of interest. Supplementary Figure S7 indicates the points where the bipartite graphs are enumerated.

Network construction relies on three principles. Firstly, mathematical operations on data are performed, independent of prior knowledge (e.g. WGCNA networks). This aspect of the approach is completely unsupervised. The second process involves the testing of hypotheses, (e.g. determination of differential transcript abundance and differential co-expression of transcripts, requiring knowledge of phenotype/trait classes for the samples). Finally, the results of the first two processes are integrated in various ways with prior biological knowledge. The result is a collection of statistically robust analytic results and various associations between them and known biology, in the form of a multiple bipartite graphs.

ANIMA database

The twenty nine graphs share node types between them, as indicated in Supplemental Figure S7b, which describes the "data model" for the graph database. As each bipartite graph is enumerated, it is added to the Neo4j database instantiated at the start of the build process. Nodes and relationships are added using the "MERGE" command, ensuring that nodes are not added in duplicate. At the end of the build script run, a large graph has been generated, stored as a Neo4j graph database. This

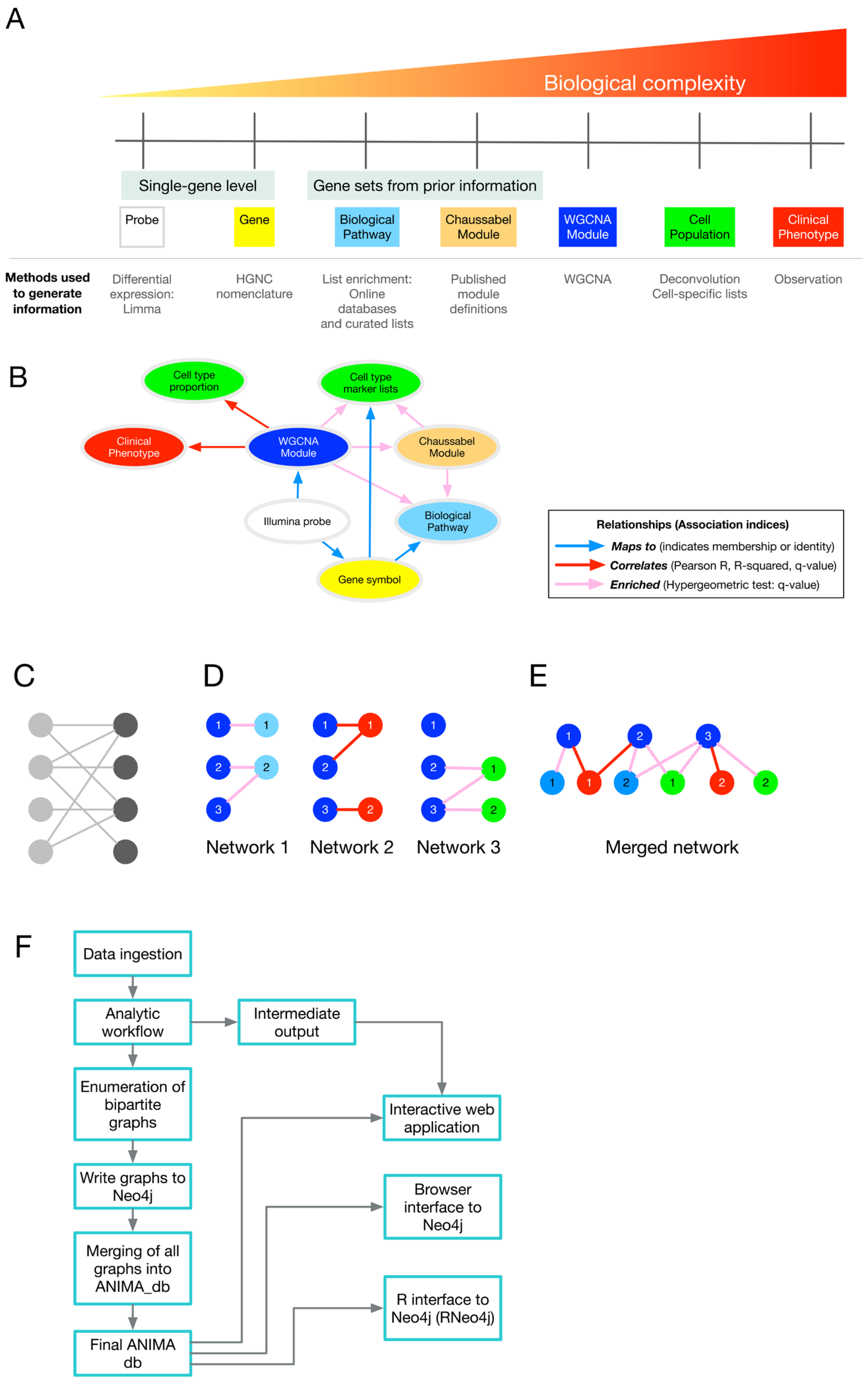
A
Biological complexity
Single-gene level
Gene sets from prior information
Probe
Gene
Biological Pathway
Chaussabel Module
WGCNA Module
Cell Population
Clinical Phenotype
Methods used to generate information
Differential expression: Limma
HGNC nomenclature
List enrichment: Online databases and curated lists
Published module definitions
WGCNA
Deconvolution Cell-specific lists
Observation
B
Cell type proportion
Cell type marker lists
Clinical Phenotype
WGCNA Module
Chaussabel Module
Illumina probe
Biological Pathway
Gene symbol
Relationships (Association indices)
Maps to (indicates membership or identity)
Correlates (Pearson R, R-squared, q-value)
Enriched (Hypergeometric test: q-value)
C
D
Network 1
Network 2
Network 3
E
Merged network
F
Data ingestion
Analytic workflow
Intermediate output
Enumeration of bipartite graphs
Interactive web application
Write graphs to Neo4j
Browser interface to Neo4j
Merging of all graphs into ANIMA_db
R interface to Neo4j (RNeo4j)
Final ANIMA db

Figure 1. Method overview. (**A**) Analytical approaches and biological complexity. This conceptualises the need for understanding biological systems at multiple scales. (**B**) Relationships between output types (**C**) A bipartite graph, with two classes of nodes connected by edges. (**D**) The separate bipartite graphs, with one node type in common. (**E**) Multipartite graph obtained after merging the three graphs in (**D**). (**F**) Outline of different steps in setting up and accessing the ANIMA database. Abbreviations: HGNC, HUGO Gene Nomenclature Committee; WGCNA, weighted gene co-expression network analysis.

Table 1. List of analytic tasks used in constructing the ANIMA database. Tasks are referenced to Supplemental Figure S7, and described in Supplementary Methods under "Nodes in ANIMA".

Task	Approach	Method	R package/implementation	Cutoff P value (BH corrected)	Reference
A1	Array data import and normalisation	variance stabilising transformation, quantile normalisation	lumi		15
A2	Probe filter	quality filter to remove non-informative probes from analysis	ReMoat, ReAnnotator		16,17
A3	Differential expression	linear model/moderated t-test	limma	<0.05	18
A4	Estimate of cell-type proportions	Deconvolution by least-squares fitting	CellMix		19
A5	Chaussabel module expression	Published module definitions and custom R code			20,21
A6	Chaussabel module differential expression	linear model/moderated t-test	limma	<0.05	18
A7	WGCNA module detection	Clustering of topological overlap matrix and dynamic tree cutting	WGCNA		14
A8	Chaussabel and WGCNA module annotation	List enrichment testing by hypergeometric test and multiple testing correction	WGCNA (UserListEnrichment); ReactomePA	<0.05	14,22
A9	WGCNA module metrics	Differential expression, module AUC, signature enrichment	Custom code		
A10	Module eigengene correlations	Pearson or Spearman rank correlation	WGCNA, custom code	<0.05	14
A11	Phenotype data analysis	Univariate analysis (ordinal and numeric data)	Base R	<0.05	7
A12	Construction of bipartite graphs	Bipartite graphs from adjacency lists (based on list enrichment) or incidence matrices (based on correlation)	igraph		23
A13	Merging of all bipartite graphs	Graph union	Neo4j: cypher command *merge*		24

Table 2. Types of associations used in the ANIMA database.

Association type	Association index	Intermediate result	Multiple testing correction
Correlation	Pearson R or Spearman rho	Incidence matrix	Yes
List enrichment	hypergeometric index	Adjacency list	Yes
Mapping	simple mapping	Adjacency list	Not applicable

is referred to as the *ANIMA* database, which makes the overall network structure as well as the individual nodes and edges accessible for further analysis (see supplementary methods; Supplementary File 1).

Accessing ANIMA

After network construction, information in the graph is accessible and utilized to expose new information not present in any of the individual steps. The key to making the ANIMA

Table 3. Bipartite graphs. Data types 1 and 2 (nodes in Neo4j) are defined in Supplementary Table S5, and indices refer to numbered bipartite graphs in Supplemental Figure S7 as well as their narrative description in Supplementary Data p.11–14. Data type names are given as implemented in the Neo4j database.

Index	Data type 1	Data type 2	Statistical method/ association type	Multiple testing correction	Cutoff (corrected P-value)	Implementation	Description and Biological implication	Reference(s)
1	PROBETYPE	SYMBOL	mapping	N/A	N/A	getSYMBOL function in lumi package	Reference mapping of all probes on the chip to genes (platform-specific)	15
2	PROBE	SYMBOL	mapping	N/A	N/A	topTable function in limma	Mapping of experiment-specific probes to genes	18
3	PROBE	wgcna	mapping	N/A	N/A	WGCNA	Indicates membership of specific probes in enumerated WGCNA modules	14
4[x]	PROBE	PROBE	Connectivity (weight in topological overlap matrix)	N/A	Maximum of 2000 per module	Custom function: at least the top 10 % of edges in the module network, capped at 2000	Subset of probes in WGCNA modules with high connectivity	14
5	SYMBOL	reactomePW	mapping	N/A	N/A	lumi, ReactomePA (uses reactome. db)	Mapping of gene symbols to reactome pathways in the current dataset	14,22
6	SYMBOL	PalWangPW	mapping	N/A	N/A	WGCNA (uses PWLists included with WGCNA)	Mapping of gene symbols to a manually curated list of pathways in the current dataset	14
7	SYMBOL	ImmunePW	mapping	N/A	N/A	WGCNA (uses ImmunePathwayLists included with WGCNA)	Mapping of gene symbols to a manually curated list of immune pathways in the current dataset	14
8	SYMBOL	cellEx	mapping	N/A	N/A	WGCNA (uses BloodLists included with WGCNA)	Mapping of gene symbols to a manually curated list of cell-type specific genes in the current dataset	14
9	SYMBOL	cellEx	mapping	N/A	N/A	CellMix (HaemAtlas dataset)	Mapping of gene symbols to the Watkins, *et al.* list of cell-type specific genes in the current dataset	19, Supp. Ref 30
10	SYMBOL	cellEx	mapping	N/A	N/A	CellMix (Abbas dataset)	Mapping of gene symbols to the Abbas, *et al.* list of cell-type specific genes in the current dataset	19, Supp. Ref 11
11	wgcna	baylor (Chaussabel modules)	Hypergeometric test	BH	<0.05	WGCNA (UserListEnrichment function with Chaussabel module genes as input)	Enrichment of WGCNA modules for Chaussabel modules, allowing functional comparison of the two approaches	14
12	baylor (Chaussabel modules)	reactomePW	Hypergeometric test	BH	<0.05[*]	WGCNA (UserListEnrichment function with Reactome genes as input)	Enrichment of Chaussabel modules for Reactome pathways, allowing functional annotation of the module	14,22
13	baylor (Chaussabel modules)	PalWangPW	Hypergeometric test	BH	<0.05[**]	WGCNA (UserListEnrichment function with a custom list of pathway genes as input)	Enrichment of Chaussabel modules for a manually curated list of pathways, allowing functional annotation of the module	14

Index	Data type 1	Data type 2	Statistical method/ association type	Multiple testing correction	Cutoff (corrected P-value)	Implementation	Description and Biological implication	Reference(s)
14	baylor (Chaussabel modules)	ImmunePW	Hypergeometric test	BH	<0.05***	WGCNA (UserListEnrichment function with a custom list of immune pathway genes as input)	Enrichment of Chaussabel modules for a manually curated list of immune pathways, allowing functional annotation of the module	14
15	wgcna	reactomePW	Hypergeometric test	BH	<0.05*	WGCNA (UserListEnrichment function with Reactome genes as input)	Enrichment of WGCNA modules for Reactome pathways, allowing functional annotation of the module	14,22
16	wgcna	PalWangPW	Hypergeometric test	BH	<0.05**	WGCNA (UserListEnrichment function with a custom list of pathway genes as input)	Enrichment of WGCNA modules for a manually curated list of pathways, allowing functional annotation of the module	14
17	wgcna	ImmunePW	Hypergeometric test	BH	<0.05***	WGCNA (UserListEnrichment function with a custom list of immune pathway genes as input)	Enrichment of WGCNA modules for a manually curated list of immune pathways, allowing functional annotation of the module	14
18	baylor (Chaussabel modules)	cellEx	Hypergeometric test	BH	<0.05	WGCNA (UserListEnrichment function, uses **BloodLists** included with WGCNA)	Enrichment of Chaussabel modules for a manually curated list of cell-type specific genes, allowing interpretation of cellular context of the module	14
19	baylor (Chaussabel modules)	cellEx	Hypergeometric test	BH	<0.05	WGCNA (UserListEnrichment function, uses **HaemAtlas** list included with Cellmix)	Enrichment of Chaussabel modules for a manually curated list of cell-type specific genes, allowing interpretation of cellular context of the module	14,19, Supp. Ref 30
20	baylor (Chaussabel modules)	cellEx	Hypergeometric test	BH	<0.05	WGCNA (UserListEnrichment function, uses **Abbas** list included with Cellmix)	Enrichment of Chaussabel modules for a manually curated list of cell-type specific genes, allowing interpretation of cellular context of the module	14,19, Supp. Ref 11
21	wgcna	cellEx	Hypergeometric test	BH	<0.05	WGCNA (UserListEnrichment function, uses **BloodLists** included with WGCNA)	Enrichment of WGCNA modules for a manually curated list of cell-type specific genes, allowing interpretation of cellular context of the module	14
22	wgcna	cellEx	Hypergeometric test	BH	<0.05	WGCNA (UserListEnrichment function, uses **HaemAtlas** list included with Cellmix)	Enrichment of WGCNA modules for a manually curated list of cell-type specific genes, allowing interpretation of cellular context of the module	14,19, Supp. Ref 30
23	wgcna	cellEx	Hypergeometric test	BH	<0.05	WGCNA (UserListEnrichment function, uses **Abbas** list included with Cellmix	Enrichment of WGCNA modules for a manually curated list of cell-type specific genes, allowing interpretation of cellular context of the module	14,19, Supp. Ref 11

Index	Data type 1	Data type 2	Statistical method/ association type	Multiple testing correction	Cutoff (corrected P-value)	Implementation	Description and Biological implication	Reference(s)
24	wgcna	cellprop	Pearson correlation	BH	<0.05***	WGCNA (WGCNA::cor function)	Correlation of WGCNA module eigengenes with estimated cell-type proportions (as output using functions in the CellMix package)	14
25	CELL	cellEx	mapping	N/A	N/A	Mapping of consensus cell type names to cell type names used in the **BloodLists** dataset included with WGCNA)	Convenience mapping that facilitates searching for cell types in the ANIMA database regardless of the name of the cell type in the original gene marker list	14
26	CELL	cellEx	mapping	N/A	N/A	Mapping of consensus cell type names to cell type names used in the **HaemAtlas** dataset included with CellMix)	Convenience mapping that facilitates searching for cell types in the ANIMA database regardless of the name of the cell type in the original gene marker list	19, Supp. Ref 30
27	CELL	cellEx	mapping	N/A	N/A	Mapping of consensus cell type names to cell type names used in the **Abbas** dataset included with CellMix)	Convenience mapping that facilitates searching for cell types in the ANIMA database regardless of the name of the cell type in the original gene marker list	19, Supp. Ref 11
28	CELL	cellprop	mapping	N/A	N/A	Mapping of consensus cell type names to cell type names used in the **Abbas** dataset included with CellMix)	Convenience mapping that facilitates searching for cell types in the ANIMA database regardless of the name of the cell type in the original gene marker list	19, Supp. Ref 11
29	wgcna	pheno	Pearson correlation	BH	<0.05***	WGCNA (WGCNA::cor function)	Correlation of WGCNA module eigengenes with a numeric matrix of clinical and other phenotype data	14

*a minimum of three and a maximum of ten results were included in the ANIMA database

** a minimum of four and a maximum of ten results were included in the ANIMA database

*** results were reviewed algorithmically; in the case of no hits for enrichment at the specified cutoff, less stringent criteria were applied (See Supplementary information). In all cases, the corrected P-value was included as a property of the relationship, allowing filtering of the database entries.

^{x} the PROBE-PROBE network is a monopartite correlation network and not a bipartite graph

Abbreviations: BH, Benjamini-Hochberg multiple testing correction database useful lies in the use of functions and web applications (see supplementary methods; Supplementary File 1) that query this large multipartite graph and return visualization of relationships or tables of nodes and/or links (associations between nodes). This has been implemented in several R functions which underpin a Shiny web application called *ANIMA REGO*.

Constructing an ANIMA database from user data

This paper and associated files provided as supplementary data allow the reconstruction of the particular ANIMA database presented here. However, it warrants emphasis that we are describing a *method* here, and that the provided datasets can be exchanged for a user's own data. The scripts at present only support Illumina microarray data, although future work aims to add support for other microarray platforms and RNAseq data.

In order to construct a new ANIMA database, the following is required in the *source_data* folder:

1. Non-normalised expression data (.txt format)
2. Clinical/phenotype data (csv format)
3. A file named "matrixPD" in csv format or similar that contains the names of the datasets to be analysed, as well as the classes of each of the variables included in the phenotype data, distinguishing numeric and categorical data types, and flagging certain variables as identifiers.
4. A file named "questions.R" which is an R script specifying a list which contains, for each dataset, multiple variable assignments required in the running of the analysis. Essentially, this file encapsulates the information required to execute the experimental design.
5. A file named "setlist.R", another R script, providing high level project information, the url of the graph database, the various studies the datasets are drawn from, as well as the identifiers of data sets that contain matched samples, which enable additional analytic functionality.

With the required files in place, the scripts ANIMA_data.R and ANIMA_build.R are run in sequence, and the end state should be a new ANIMA database based on the user data. The supplementary information provides full replication instructions.

Using ANIMA

The novelty and value in ANIMA lies in its three broad approaches to data interpretation: Firstly, it allows detailed, multiscale investigation of a single dataset with a focused research question where two phenotype classes are compared **(multiscale class comparison)**. Secondly, as datasets are stratified by default using two variables (typically one identifying two disease classes, and the other either a potential confounder (e.g. sex), or a second biologic variable of interest, (like co-infection with a second pathogen), we can examine the interaction of the second variable with the first using a factorial study design **(factorial analysis)**. Finally, multiple datasets, and multiple conditions can be meaningfully compared and contrasted to identify similarities and differences **(meta-analysis)**, both at cellular and modular level.

Application areas

In its current state, the ANIMA approach is optimised to address questions in immunology, but other application areas that may be tissue type specific (neuroscience, oncology) are also possible, but will require addition of annotation libraries and lists of cell-type specific genes.

Results

Example study: data sets

We analysed three publicly available microarray datasets using the ANIMA pipeline and toolset (ArrayExpress/GEO identifiers: E-GEOD-29429, E-GEOD-34404/GSE34404, E-GEOD-68310/GSE68310). The first compares a cohort of subjects with acute HIV infection to healthy controls[25], the second compares symptomatic malaria to asymptomatic controls in children from a malaria-endemic region[26], and the last compares the host response in early symptomatic viral respiratory infections[27]. Where the datasets contained samples from multiple timepoints, we restricted the analysis to healthy controls and the first disease timepoint. The respiratory virus infection dataset contained samples from subjects with infections other than influenza or rhinovirus; we excluded those from this analysis. Figure S1 (Supplementary File 1) shows the experimental design for the factorial analysis in *limma* for each of the datasets, together with the numbers of samples in each of the individual groups. Each of the datasets also included clinical data, which was integrated in the analysis.

The ANIMA network

The result of the script that builds the ANIMA network is a large, mostly connected, multipartite graph. Figure S2 (Supplementary File 1) shows a subgraph of this network (for dataset **HIVsetB**, edge 5).

Searching the ANIMA network using network paths and filtering on node and edge properties

In the most direct approach, the large ANIMA graph in the Neo4j database can be searched directly from a web browser using the Cypher Query Language (CQL).

Consider the following query:

```
MATCH (ph:pheno)-[r1]-(n:wgcna {square:'HIVs
etB',edge:5})-[r2]-(p:PROBE)-[r3]-(s:SYMBOL)
WHERE r1.weight > 0.6 AND p.logfc > 2 RETURN *
```

The query language combines commands and functions (e.g. "MATCH", "WHERE") with a typographic representation of nodes and relationships, where a node is indicated with round brackets ( ) and a relationship is indicated by dashes either side of a set of square brackets **-[ ]-**. Taken together, the query above conducts a search (or graph traversal) for a specific graph pattern matching `MATCH` WGCNA modules for the HIV positive vs control comparison in the **HIVsetB** dataset `(n:wgcna {square:'HIVsetB',edge:5})` whose module eigengene (ME) is strongly correlated with any clinical variable (Pearson R > .6) `(ph:pheno)-[r1]-; WHERE r1.weight > 0.6`, and also returning the genes mapping to probes within those modules `-[r2]-(p:PROBE)` that have high levels of over-abundance in HIV-1 infected individuals `AND p.logfc > 2`.

In addition to the standard web browser interface for Neo4j, into which the above query can be directly entered and returned in the browser window (Figure 2A), we also provide a function in R (`igraph_plotter`) that returns the network found, and plots this within R (Figure 2B), exports the network as node and edge lists for import in other software like Cytoscape (Figure 2C), or returns the result as an *igraph*[23] object for further manipulation within R. (A file ANIMA_styles.xml is included in the common folder supplied with the source code; this is a Cytoscape stylesheet that reproduces the colouring shown in the Figure 2C)

More sophisticated methods of using ANIMA are described next. These utilise both bespoke R functions as well dynamic interactivity provided in the Shiny web interface (see below).

Approach 1: Multiscale class comparison

Accessing individual transcript abundance levels in multiple conditions. It is useful to view transcript abundance patterns of specified probesets, for instance to compare microarray data to RT-PCR validation data, or to investigate the behaviour of groups of biologically related genes in various conditions. We provide an interface for this in ANIMA. The user can submit a search string (in the form of a regular expression) containing gene names (HUGO Gene Nomenclature Committee (HGNC) symbols), and box-and whisker plots for the results are returned. In the original paper on acute HIV infection the authors discuss a gene set of six conserved genes that appear at multiple timepoints in an inferred regulatory network of viral set point[25]. We show the normalized expression data stratified by HIV status and sex in the two datasets included in the HIV analysis for these six genes (Figure 3). In the paper on acute viral respiratory infection, IFI27 and PI3 are identified to differ between acute influenza A and human rhinovirus infections. In influenza, IFI27 is upregulated and PI3 downregulated relative to human rhinovirus. The malaria study replicated prior knowledge of differential transcript abundance for C1QB, MMP9, C3AR1, IL18R and HMOX1; we show similar results for these transcripts. Supplementary Table 1 lists the results of differential expression analysis for the above transcripts in the three conditions, providing validation of data at individual transcript level.

Functional annotation of WGCNA modules. An important question is what do WGCNA modules represent, given that these are groups of genes that co-vary across samples, but that can differ dramatically in size. Instead of searching only for evidence of co-regulation of expression by transcription factors, one should consider other causes of this co-variance. We propose the notion of WGCNA modules representing biological processes that may be regulated at different scales. For instance, a group of transcripts will co-vary across samples if these transcripts are expressed (predominantly) in a single cell type, and the proportions of this cell type varies between samples (Figure 4). Another group of transcripts may be expressed in multiple cell types, but represents a concerted transcriptional program executed in response to a specific stimulus, such as interferon-alpha stimulating the expression of a specific group of interferon-regulated genes (see below). Generally, we determine the function of WGCNA modules based on all statistically significant associations with pathways and cells.

Relationship of modules to clinical variables. We performed Pearson correlation of WGCNA module eigengenes (ME) and clinical variables (described in supplementary methods). Figure 5 shows correlation of the *pink* ME with age and CD4 count in the **HIVsetA** data set. Clearly, the *pink* module is significantly associated with CD4 count, an association that is independent of age. Further investigation shows that this module is linked to interferon signalling, and probably expressed in a variety of cells. This agrees with our understanding of acute HIV infection, which is associated with a robust type-I interferon response and an acute drop in CD4 count[28].

Investigating the structure of WGCNA modules. WGCNA modules are groups of co-expressed transcripts. Demonstrating the extent and direction of correlation of their constituent probes, and their relationships to biological pathways requires sophisticated visualization (Figure 6). We showcase the example using the **HIVsetB** dataset and focus on two modules. The *turquoise* module shows the coordinated action of genes involved in cell division, relating to the cell proliferation in the lymphoid compartment. The *yellow* module is clearly related to interferon signalling; of interest here is that within this module there is a group of transcripts highly correlated with each other, all related to interferon signalling, suggesting that these transcripts may all be downstream of a single regulatory factor. Also clear from the figure is that this module is not limited to a single cell type, but rather to innate immune cells in general.

Relationships between module-based approaches. Both WGCNA and the modular approach pioneered by Chaussabel, *et al.*[17] rely on clustering of similarity matrices to derive modules. A key difference in these methods in ANIMA is that the Chaussabel modules are pre-defined, whereas the WGCNA modules are derived from the transcriptional data under study. It is therefore interesting to discover the relationships between these two approaches. Figure 7A shows the bipartite network of WGCNA and Chaussabel modules derived from the **HIVsetA** dataset, as defined by the hypergeometric index. Figures 7B and 7C show the two projections of the bipartite graph. The hypergeometric index is not the only way associations between the two module types can be demonstrated; for instance, a more indirect association can be inferred when a WGCNA and Chaussabel module map to the same biological pathway. It is clear from the plots that only a subset of the list of Chaussabel modules associates with WGCNA modules in any given condition.

Deconvolution. An important question in analysing transcription data from complex tissues is whether differences in transcript abundance are attributable to transcriptional regulation

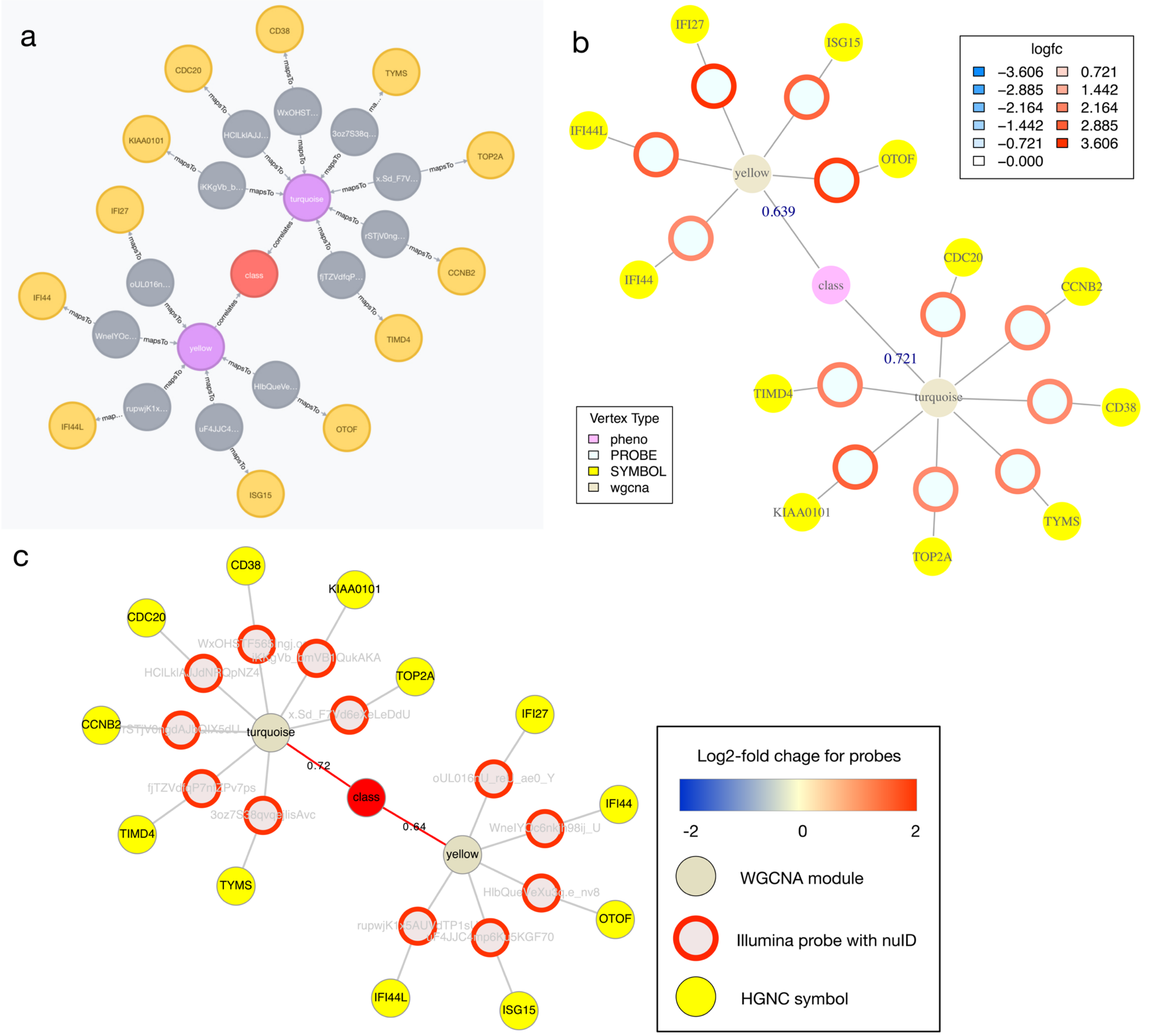

Figure 2. Visualising Cypher query results. Relationships between nodes extracted from the ANIMA database using a Cypher query applied to the **HIVsetB** data (N_{HIV}=30, $N_{Controls}$=17, see Figure S1; Supplementary File 1). Shown are two WGCNA modules that contain probes with increased transcript abundance in acute HIV infection and whose module eigengene is positively correlated with disease class (an ordinal variable). (**A**) Result from native browser interface for Neo4j. (**B**) Result plotted from within an R session connected to the ANIMA database, using the *igraph_plotter* function. Log $_2$-fold change values for the individual probes are shown by coloured rings; values are shown in the legend. (**C**) The same result, visualized in Cytoscape, taking advantage of the *igraph_plotter* function to export node and edge lists for easy import into Cytoscape. Links/edges are annotated with Pearson correlation coefficients where applicable.

in one or more cell types, or to changes in the composition of the overall leukocyte populations.

Figure S3 (Supplementary File 1) shows the results for the **HIVsetB** dataset, and Table S2 (Supplementary File 1) shows the results of non-parametric statistical testing for differences in median cell-type proportions, per cell-type and class comparison with uncorrected P values as well as P-values corrected for multiple testing using the Benjamini-Hochberg procedure; these were encoded as parameters *diffP* and *diffQ* of the *cellprop* node type in the ANIMA database. Neutrophils were the most abundant cell-type estimated from the array data. The proportions of activated NK cells and CD8+ T cells were significantly elevated in acute HIV infection, and proportions of B cells and CD4+ T cells were reduced, in line with published observations[29].

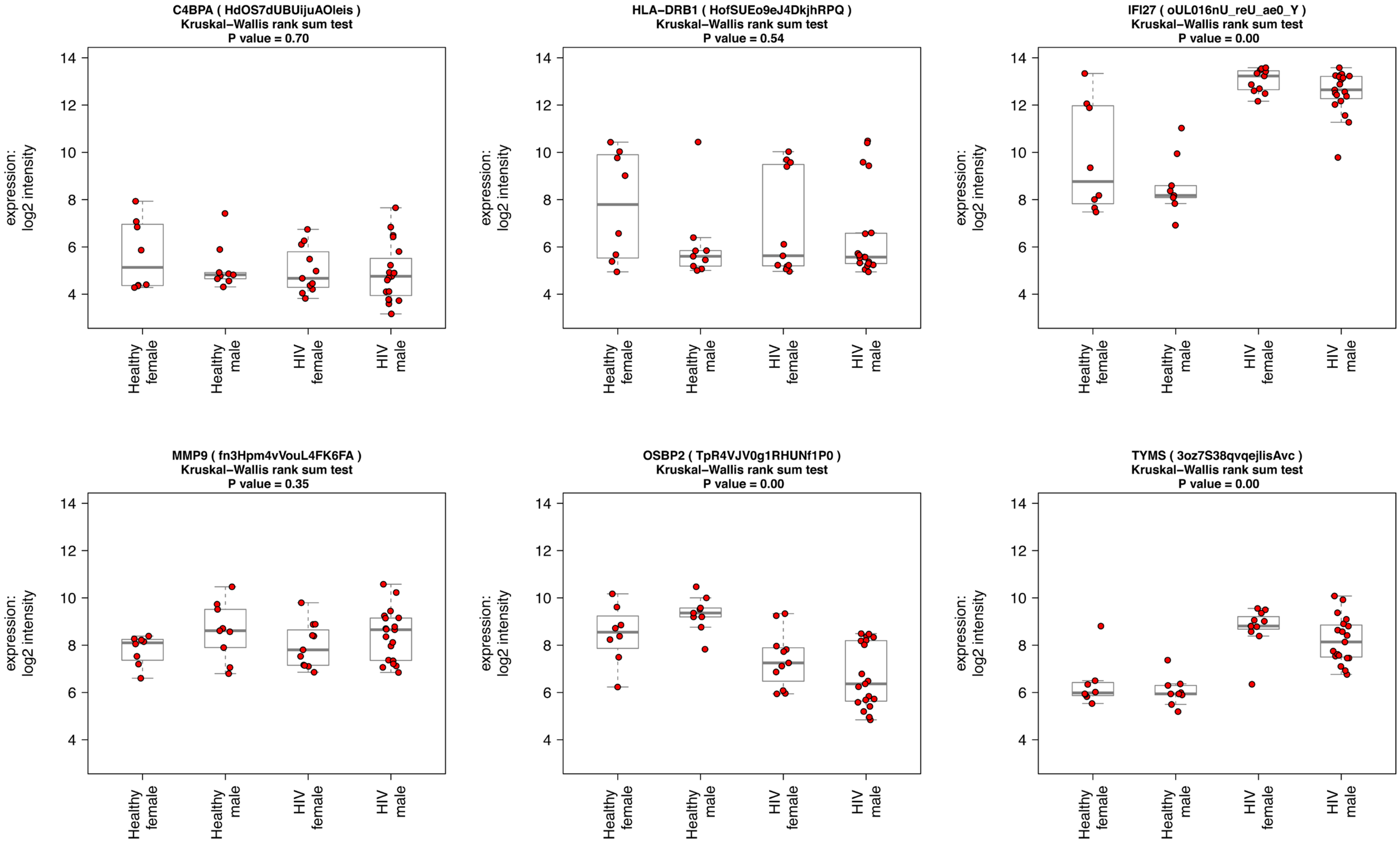

Figure 3. Visualising individual probe-level expression data. Box-and-whisker plots showing normalized, log $_2$-transformed probe-level expression data for six selected genes, obtained by a custom function in R in four groups: Healthy female, N = 8, Healthy male, N = 9, acute HIV female, N = 11, acute HIV male, N = 19; data from **HIVsetB** dataset. Gene (and probe nuIDs for disambiguation) are given for reference; the y-axis shows log2 scale normalized intensity values. Box and whisker plots show median, interquartile range, and range. Outliers are defined as values that lie beyond the whiskers, which extend to maximally 1.5 X the length of the box. Individual datapoints are superimposed in red on the box-and-whisker plots. The four groups are compared using Kruskal Wallis rank sum test, and the P-value for the comparison is shown in the plot title. Results for individual pairwise comparisons are not shown.

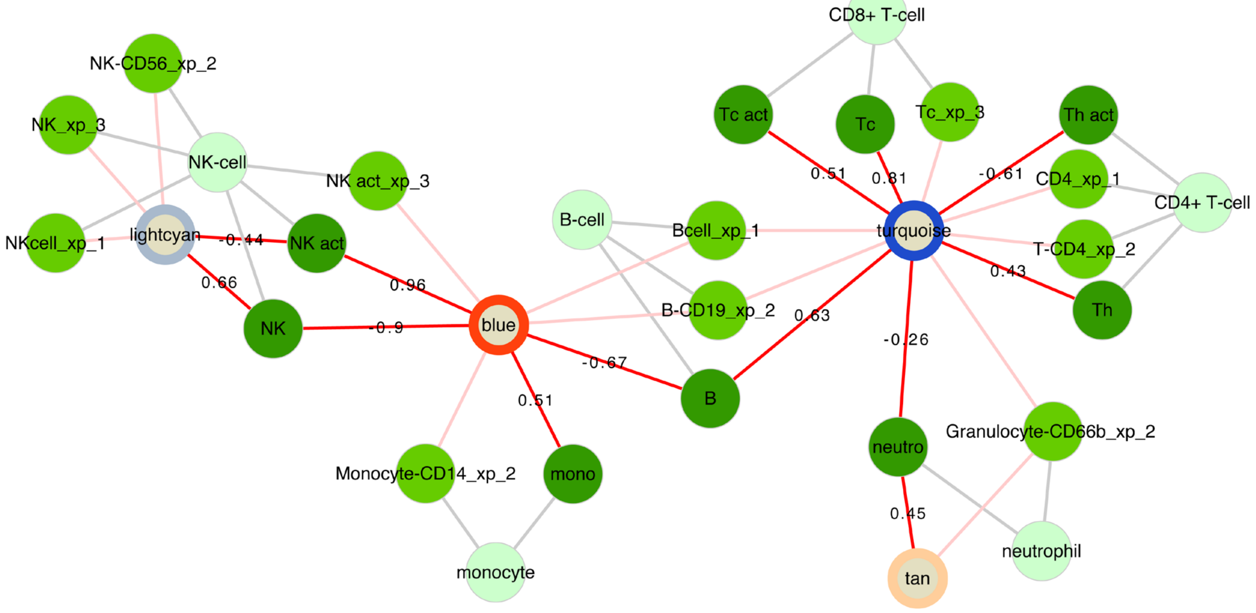

Figure 4. Cell associations of WGCNA modules. Relationships of WGCNA modules and different cell types in the **respInf** dataset (Day 0 acute influenza, N = 46 vs baseline healthy samples, N = 48, see Figure S1; Supplementary File 1). Shown are WGCNA modules whose expression correlates with specific cell-type proportions (dark green, edges annotated with Pearson correlation coefficient *R*) *and* that are enriched for the genes specific to that cell type (medium green, suffixes xp_1-3 indicate the respective gene list on which the cell assignments were based, see Supplementary methods). The classes of cells are indicated in light green. The modules are annotated with coloured rings representing the difference in median eigengene values between cases and controls (diffME, see Supplementary methods); blue indicates modules which are under-expressed, and red indicates modules that are over-expressed in cases relative to controls. WGCNA module names are (arbitrarily) based on colours as per the convention of the WGCNA package, and modules were not renamed manually.

Virtual cells: an estimate of functional phenotypes of different immune cell types. Given that the immune response is mediated by different types of cell, we attempted to re-create "virtual cells" based on the assumption the genes that co-vary in terms of transcript abundance across samples with that of cell-type specific genes are expressed in that particular cell type. With these relationships, we generated virtual cells (probe co-expression matrices annotated with biological pathways and probe differential expression data). Figure S4 (Supplementary File 1) shows two cells (B-cells and neutrophils) in acute HIV infection; both characterized by interferon signalling. Given the "virtual cells" and their functions we can compare pathway-level transcript abundance in different cell types by creating a matrix of cell types and pathways, with each entry representing the "pathway activity" for a given cell and pathway combination. To illustrate this, we show replication of the finding of NK-cell activation in acute influenza infection (**respInf** dataset, Figure 8 A,B), and in addition provide more detail on which pathways are probably up- or downregulated in these and multiple other cells.

Approach 2: Factorial analysis

Modules driven by sample class or sex. Our main interest in investigating transcriptomic datasets is to identify molecular and cellular processes that drive, or at least are associated with, specific phenotypic traits of the samples. Therefore, the experimental design for the differential abundance analysis and the probe filtering steps prior to WGCNA module discovery were both designed to highlight processes associated with two factors: disease class and sex; the former because this is the basis of the research question for the three studies, and the latter because sex has a distinct influence on immune system function[30]. We developed a novel approach to quantify the association of the WGCNA module expression with these two factors (See supplementary methods; Supplementary File 1). Figure 9 shows that, in the case of acute HIV infection, most modules are associated with the disease process (**HIVsetB** dataset), but that one module (*purple*) is strongly associated with sex. Figure S5 (Supplementary File 1) shows the module eigengenes for the *yellow* and *purple* modules, demonstrating differential class associations for WGCNA modules; this finding would have been missed had the data not been stratified by both HIV infection status and sex. Table S4 shows the module statistics for this dataset.

Additional modules were identified that associated with neither sample class nor sex. These represent biologic processes that manifest in heterogeneity of the sampled population. Figure S6 (Supplementary File 1) plots the study subjects in the HIV infection data set (**HIVsetA**) on two axes represented by two module eigengenes.

Approach 3: Meta-analysis of multiple datasets

Module-level meta-analysis. Meta-analysis of multiple related expression datasets can lead to insights not available from analysis of any single datasets, and can highlight common

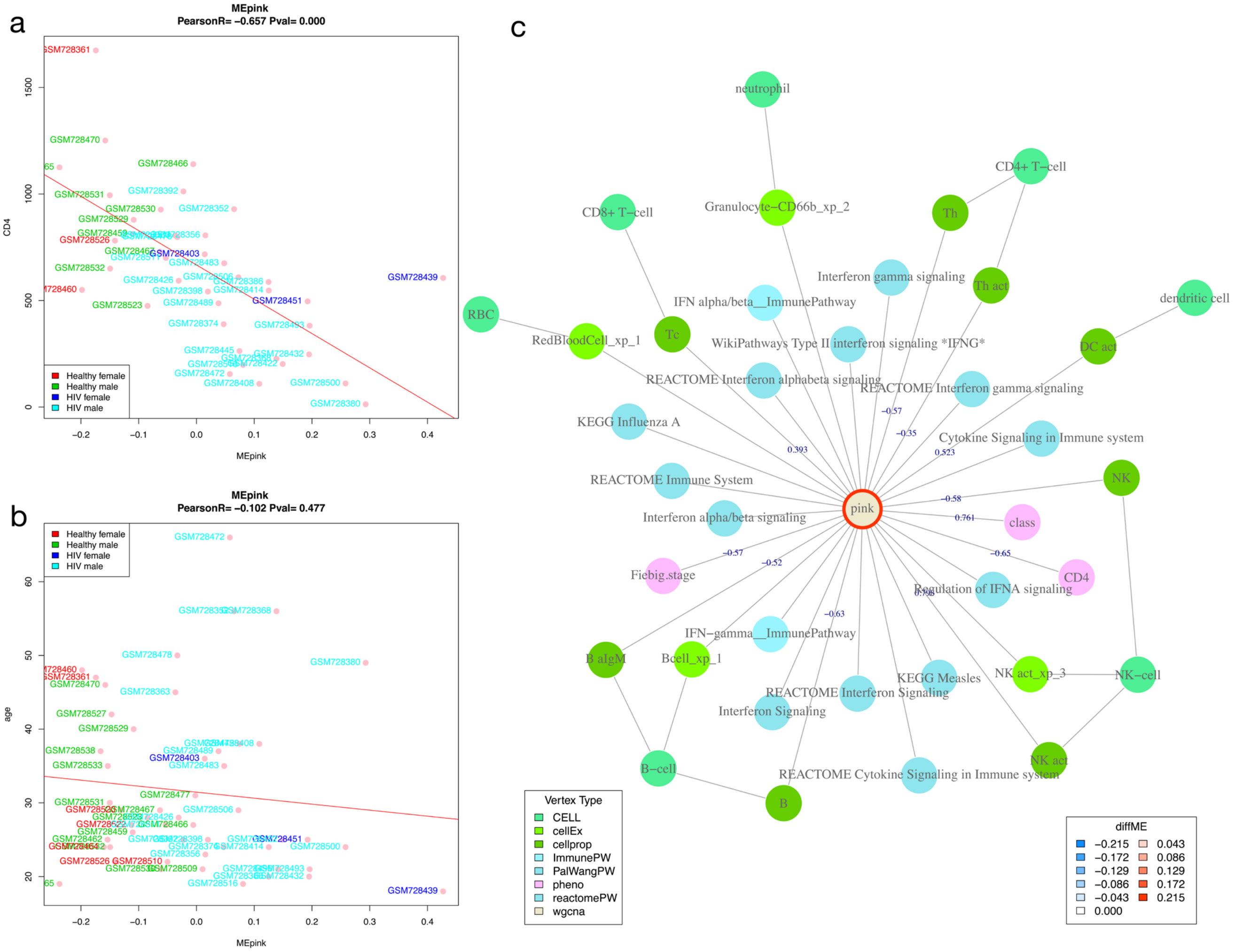

Figure 5. Correlation of module eigengenes with clinical variables. Shown is the Pearson correlation of the *pink* module eigengene with CD4 count (cells/microlitre) (**A**) and with age (years) (**B**) in acute HIV (N = 28) vs healthy controls (N = 23) in the **HIVsetA** dataset. Study subject IDs are used as point labels, and coloured as indicated in the legend. Plot titles show the Pearson coefficient R and the associated P-value. (**C**) WGCNA module annotation obtained from the Neo4j database for the *pink* module. Edges are labelled with the correlation coefficient (R) where applicable. Note that the same coefficient is obtained for CD4 count as in panel **A**. Legends are shown for vertex type and diffME (a measure of differential co-expression (see Supplementary methods), i.e. the extent that the module eigengene median varies between two classes). Abbreviations: **diffME**, differential module eigengene.

patterns of transcript abundance across different conditions, or meaningful differences across highly common conditions. We implemented the approach pioneered[20] and refined[21] by Chaussabel, *et al.* to perform modular transcriptional repertoire analysis[31] on the six datasets. This approach is particularly suited to meta-analysis, as the composition of the modules is always identical. Figure 10 shows modular patterns for the six datasets in a clustered heatmap as well as the subset of modules with similar expression patterns across the six data sets, demonstrating universal patterns in the immune response to infection. Supplementary Table S3 (Supplementary File 2) lists module functions based on all significant enrichment associations for the modules in Figure 10B. Universal upregulation of interferon-related modules particularly stand out, as does the suppression of modules associated with CD4+ T cells, CD8 T cells and B cells.

Meta-analysis at the cell-type level. A second approach to meta-analysis was implemented using virtual cells based on WGCNA modules. Here we compared the pathway scores in a single or several cell types across multiple conditions. For instance, comparing the CD8 T-cell response in acute HIV, acute viral respiratory infection, and symptomatic malaria, we find that proliferation of activated CD8+ T cells characterizes acute HIV infection, and to a lesser extent symptomatic malaria infection. In contrast, there is a suppression of CD8 T cell activity in blood in other conditions, due in part to a reduction in CD8+ T-cell proportions in whole blood (Figure 10C).

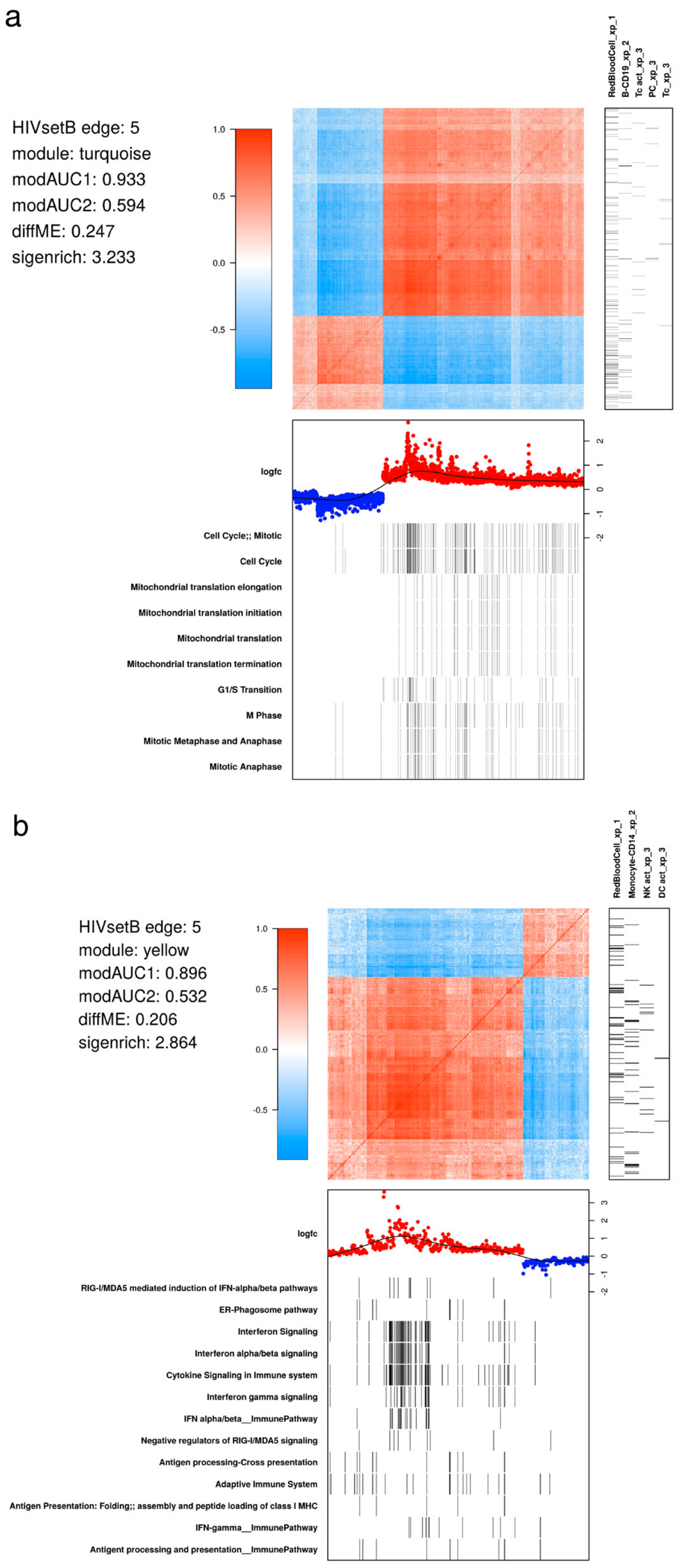

Figure 6. WGCNA module structure. (**A**) Correlation matrix of all probes in the *turquoise* module in the HIVsetB dataset (N_{HIV}=30, $N_{Controls}$=17, see Figure S1; Supplementary File 1). Colours in the heatmap represent Pearson correlation coefficients, ranging from -1 to 1, as indicated by the legend. The module is enriched for lymphocyte-specific genes (right annotation panel) as well as cell cycle/mitosis associated genes, suggesting that various lymphocyte subsets in acute HIV infection are actively proliferating. (bottom annotation panel). Log $_2$-fold change values refer to differential transcript abundance in acute HIV relative to healthy controls. (**B**) Correlation matrix of all probes in the *yellow* module in the HIVsetB dataset. It is enriched for innate cell genes as well as interferon signaling, suggesting that innate immune cells are in an interferon-induced state. Additional annotation information is provided to the left of the heatmap. The parameters *modAUC1*, *modAUC2*, *diffME* and *sigenrich* are defined in Supplementary methods. The plot is generated using a custom R function (*mwat*).

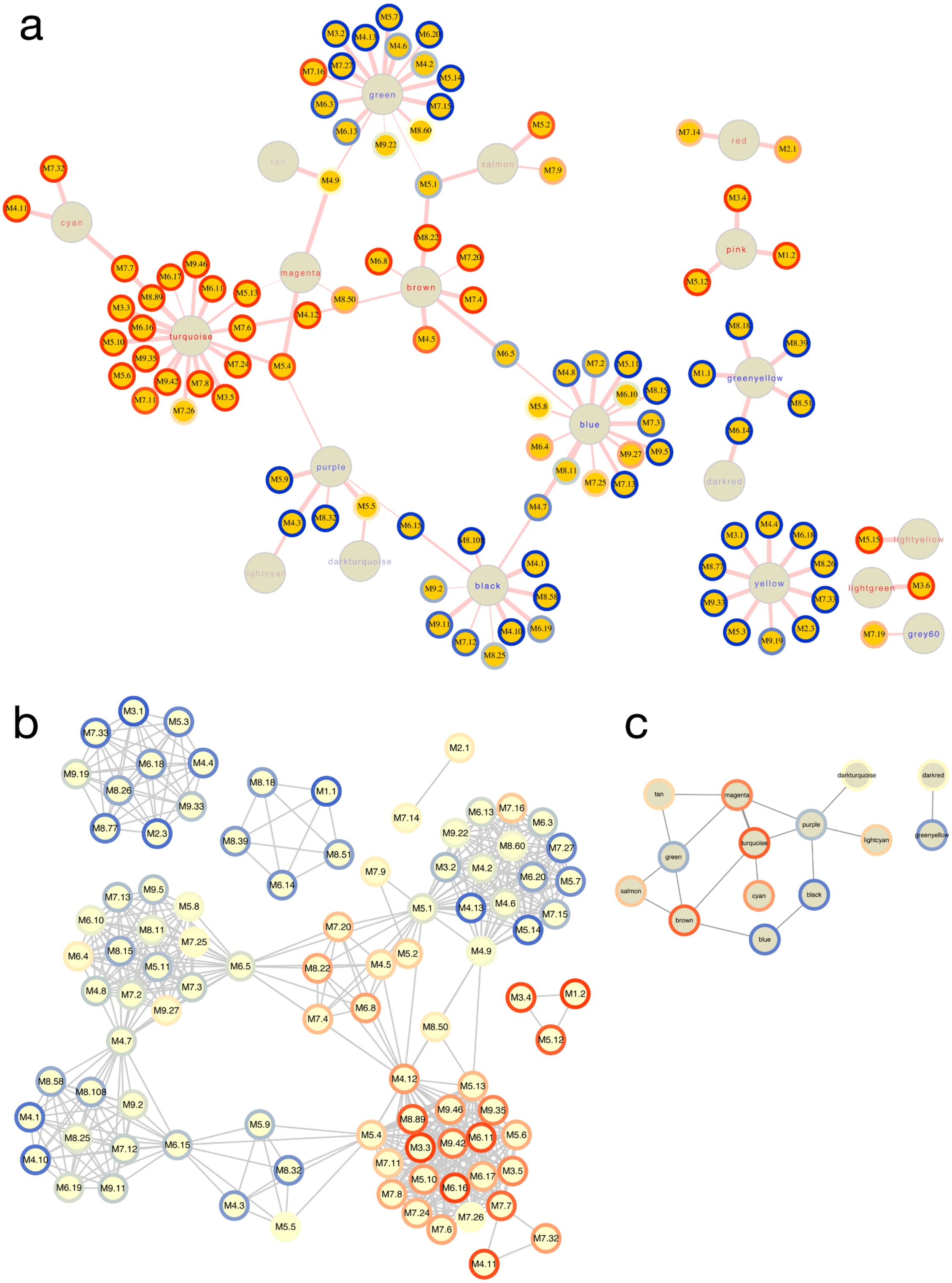

Figure 7. Relationships between WGCNA and Chaussabel modules. (**A**) Bipartite graph of the two module types based on the hypergeometric association index in the **HIVsetA** dataset (acute HIV, N = 28 vs healthy controls, N = 23). Strikingly, Chaussabel modules tend to have the same direction of differential expression (indicated by the rim colour of the Chaussabel modules, red indicating up-regulation in acute HIV, and blue indicating downregulation) as WGCNA modules they map to, indicated by the label colour of the module. (**B**) Projection 1 of (**A**), showing relationships between Chaussabel modules based on shared WGCNA modules; dense cliques of modules are observed. (**C**) Projection 2 of (**A**), showing relationships between WGCNA modules based on shared Chaussabel modules. All associations (hypergeometric test) shown are corrected for multiple testing, BH-corrected P-value < 0.05. All outputs were generated using the *igraph_plotter* function, exporting vertex and edge tables of the bipartite graph and the two projections and importing these into Cytoscape.

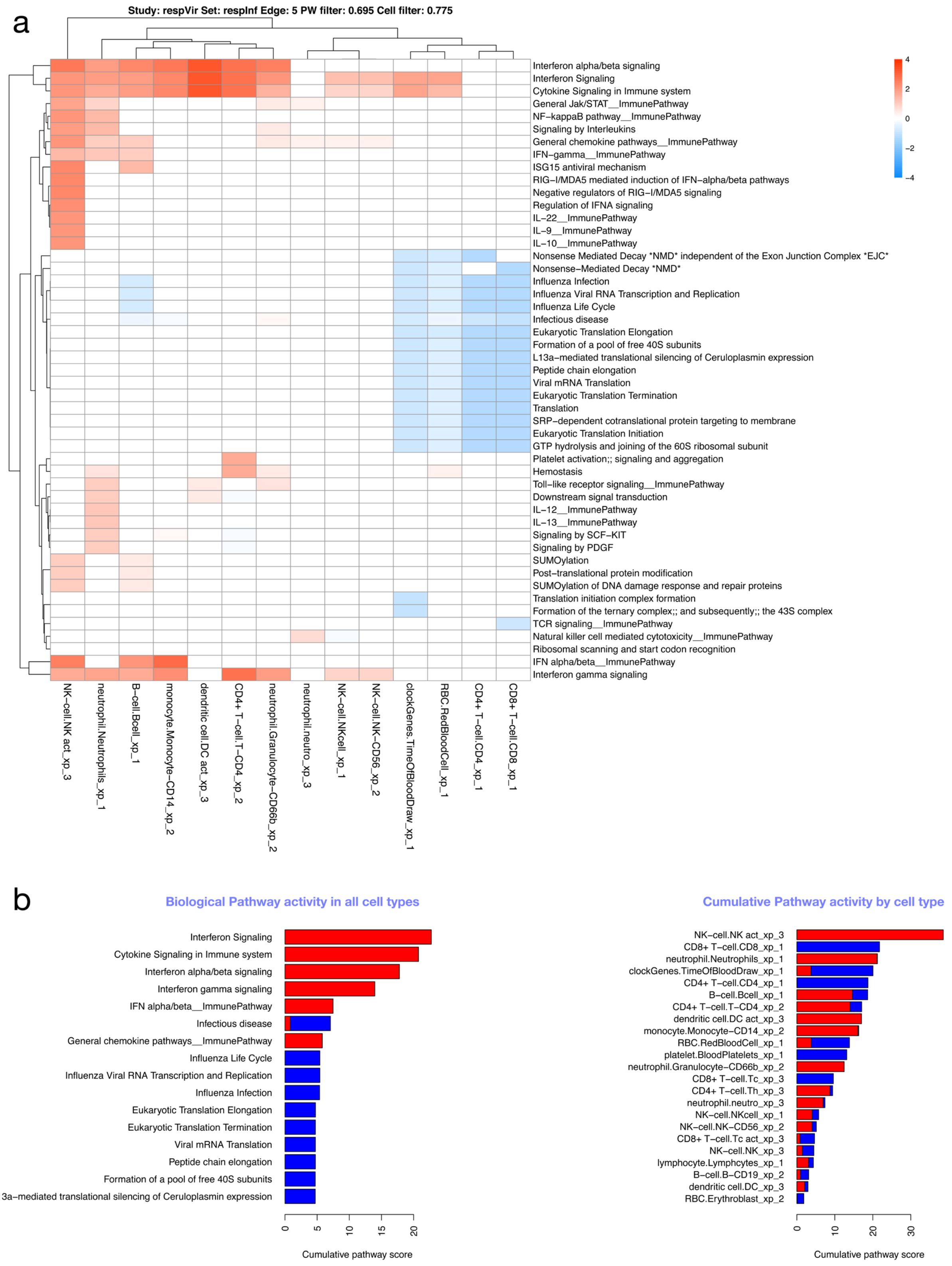

Figure 8. Cell/pathway activity matrix. (**A**) Cell/pathway activity matrix for all cell-types for the **respInf** dataset (Day 0 acute influenza, N = 46 vs baseline healthy samples, N = 48, see Figure S1). The clustered heatmap shows pathway activity scores representing the mean log-2 fold change for all probes in the pathway for a particular cell type (see Supplementary methods). There is a clear interferon response in multiple cell types, as well as down-regulation of other pathways associated with translation. (**B**) Barplots highlighting the most highly differentially regulated pathways (left panel, determined by row sums of matrix in **A**), and cells with highest levels of differential expression (right panel, determined by column sums of matrix in **A**). In all cases, up- and downregulated pathway scores are kept separate.

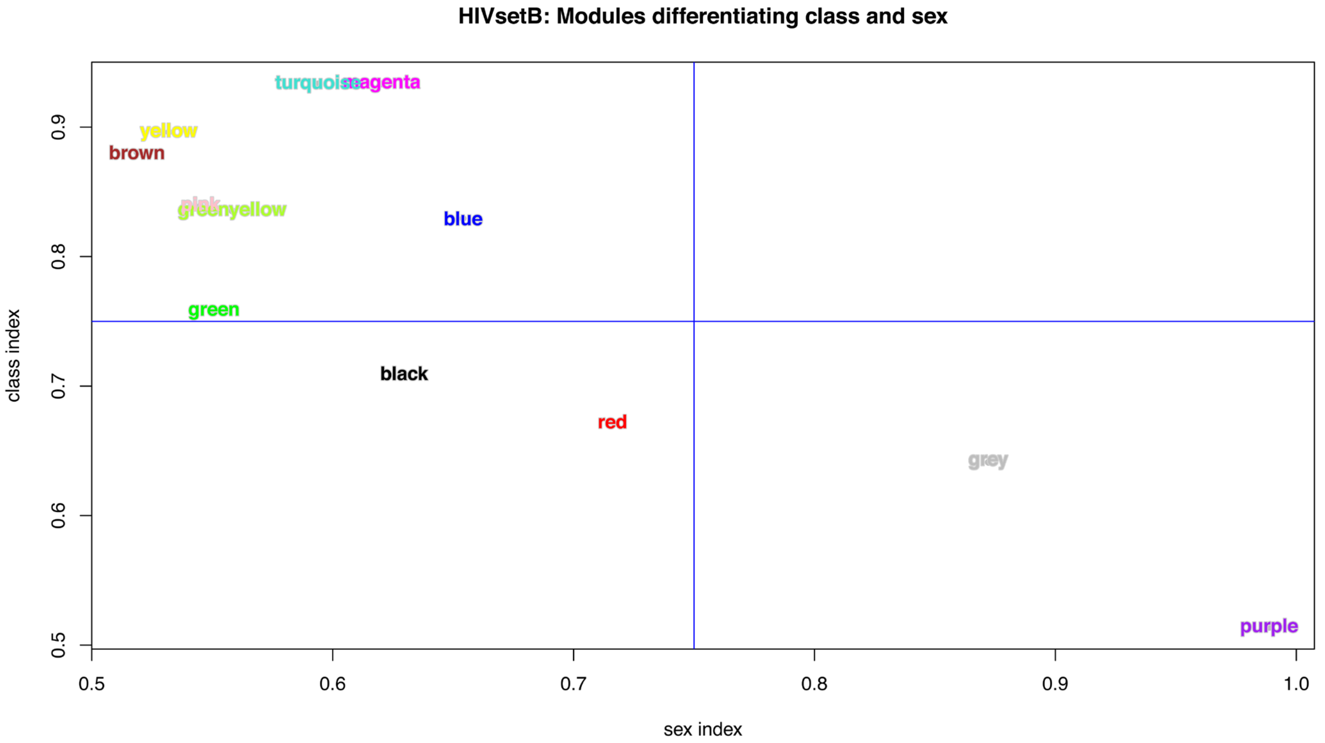

Figure 9. WGCNA module indices. Plot of module indices representing the area-under-the-ROC curve for the two classes for all WGCNA modules in the **HIVsetB** dataset (N_{HIV}=30, $N_{Controls}$=17, see Figure S1; Supplementary File 1). The indices are named per the variable they aim to differentiate (disease class or sex). The class index corresponds to the modAUC1 variable and the sex index corresponds to modAUC2. These indices are calculated form the module eigengenes and given class assignments using functions from the *rocr* package. See text and Supplementary methods for details.

Runtime statistics and output

The build script produces a file called session_output.txt. This file captures various runtime statistics, as well as some textual output. This specific ANIMA database was created on a 2013 Mac Pro, 8 cores, 64 GB RAM, with 7 cores and 55 GB RAM allocated to Docker. The build phase took 34.3 hours to complete. Once the database is built, the various components of the Shiny web interface compute and render within seconds, depending on resources available.

Discussion

Systems immunology aims to understand the complex web of relationships between immune system components (cells, cytokines, effector molecules and other mediators) in immune-mediated disease states. Much progress has been made in single-cell techniques that yield large amounts of information, e.g. single-cell RNAseq and mass cytometry. These approaches however are expensive. In contrast, RNAseq or microarray analysis of complex tissue samples (like blood) in principle contain information on the transcriptomic state of all cells present in the sample and is thus an unbiased approach. The main difficulty with this lies in the interpretation of the data, and in many cases a complexity-reducing analysis approach is employed, where the focus is placed on differentially expressed genes. Other approaches based on co-expression analysis often fail to explain the drivers of the co-expression patterns.

We demonstrate, using a novel method of aggregating information obtained from clinical and microarray data, an ability to reconstruct many aspects of the immune response, and to discuss this not in the language of probes, genes and signatures, but rather as coordinated biological processes and the cellular context for these processes, allowing the generation of hypotheses at multiple scales. Our multiscale class comparison approach can be used to validate findings from individual papers, for instance the intense NK-cell activation described in influenza[27]. In contrast to other approaches, e.g. PARADIGM[32], which rank pathways in specified conditions, we utilise the multiscale association information to infer states of immune cells (virtual cells) and providing an interface to compare such cell states within and between datasets. Pathway and cell-activity ranking are provided as a summary of multiple cell states.

Using our factorial approach, we can begin to dissect inter individual heterogeneity in transcriptional patterns from transcriptional patterns that are in a causal relationship with defined factors. Application of the two meta-analysis approaches allows comparison of arbitrary datasets to detect similarities and differences at modular and cellular levels. An interesting and somewhat unexpected finding is that acute symptomatic malaria and acute respiratory viral illnesses are more similar to each other than to acute HIV infection, another viral illness. Despite these differences, we demonstrate that, at least for

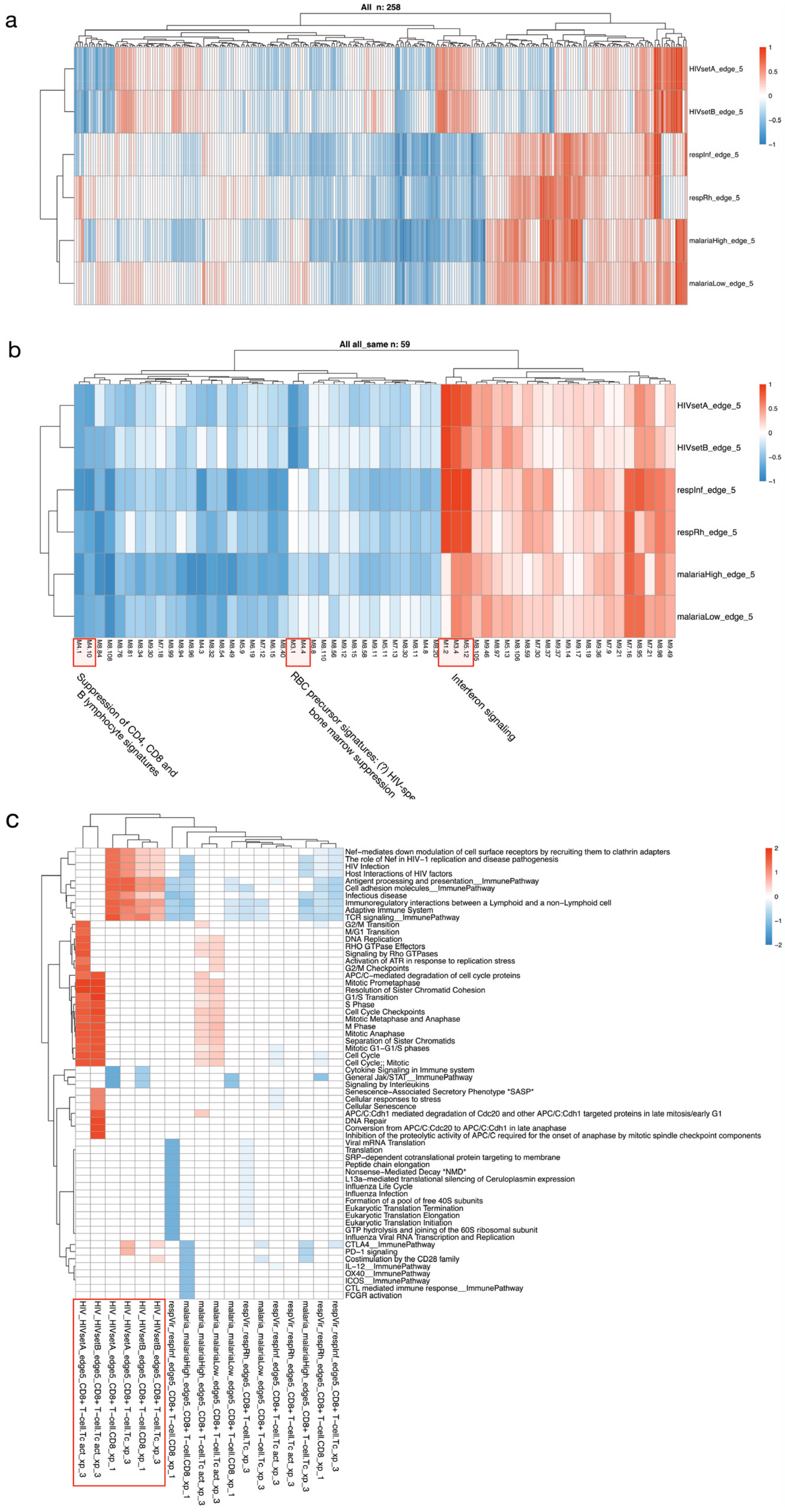

Figure 10. Meta-analysis of transcriptional and cellular patterns. (**A**) All 258 Chaussabel modules plotted as a heatmap in all six datasets. (**B**) The subset of modules all expressed in the same direction. Three module groups of interest are identified. (**C**) Cell/pathway activity matrix for a single cell type (CD8 + T cell) based on three celltype-gene lists (xp1, xp2, xp3, see Supplementary methods) in all three conditions. Activity in CD8+ T cells in HIV all cluster together, and differ from both malaria and respiratory infections. Cell labels are constructed by [condition]_[dataset]_[comparison]_[cell class]_[cell type]_[gene list].

these three rather different infections, a broadly similar pattern of transcriptional module activity can be described.

In summary, ANIMA is both a robust implementation of various well-regarded analytic paradigms in microarray analysis, as well as a framework for integrating these various methods to expose relationships at multiple scales and render these computationally accessible.

Future work

ANIMA is an open-ended project. Future work may include a demonstration that the associations in the final network are robust to choice of package. This requires a technical comparison study incorporating multiple packages with the same goals. A priority of ongoing development is improving the ease with which users can import their own data, as well as adding support for more microarray platforms and RNAseq data. Customisation of ANIMA at this stage requires modification of the source code to explicitly add additional functionality. With modularization of code, we hope to make this process easier. The scope of ANIMA can be widened by including support for more tissue types and organisms. Integration of protein-protein interaction data using existing databases is planned. Finally, we envision a repository where completed ANIMA projects may be archived as well as hosted.

Software availability

Source code (R code, dockerfiles) are available at GitHub at https://github.com/adeffur/anima under a CC BY-NC 4.0 license

An archive of source code releases linked to github is available on Zenodo (http://doi.org/10.5281/zenodo.1163398[33])

The docker image used to implement ANIMA is available at Docker hub (https://hub.docker.com/r/animatest/anima/) with the tag v3.3.3

The same image has been archived and is available on Zenodo (https://doi.org/10.5281/zenodo.1161475[34])

System requirements

1. Minimum system requirements for installing and running Docker must be met. Docker is freely available for macOS, Microsoft Windows and various Linux distributions. See https://docs.docker.com/install/ for further information
2. To run the full ANIMA pipeline, including the build phase, Docker must be provisioned with at least 50GB RAM, and 6–8 cores
3. To run the ANIMA web application, at least 8GB of system RAM is required, but 16GB is preferable.

Data availability

Array Express: **Whole Blood Transcriptional Response to Early Acute HIV**, E-GEOD-29429

Gene Expression Omnibus: **The genomic architecture of host whole blood transcriptional response to malaria infection**, GSE34404

Gene Expression Omnibus: **Host transcriptional response to influenza and other acute respiratory viral infections – a prospective cohort study**, GSE68310

A single zip archive will all data, and clinical and other metadata has been archived on Zenodo; for reproducibility purposes it is advised that this archive is used as described in the manual (Supplementary File 3) (https://doi.org/10.5281/zenodo.1161380[35]).

Grant information

This work is supported by the Wellcome Trust [104803, 203135].

This study was also supported by the Medical Research Council of South Africa, Hasso Plattner Foundation, and Newton Fund [NA150173].

This work was supported by the Francis Crick Institute which receives its core funding from Cancer Research UK (FC001218), the UK Medical Research Council (FC001218), and the Wellcome Trust (FC001218).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplementary material

1. **Supplementary File 1: ANIMA_method_SUPP_submission_WOR.7z**

   This document contains supplementary methods, additional figures and tables, implementation and replication information.

   Click here to access the data.

2. **Supplementary File 2: Table S3 New.xlsx**

   Supplementary Table S3 in Excel format.

   Click here to access the data.

3. **Supplementary File 3: ANIMA_manual_v1.0.0.pdf**

   This document contains full instructions to replicate the ANIMA pipeline and web application, as well as screenshots describing the functionality of the web application.

   Click here to access the data.

Pre-print

A preprint of this work has been submitted to biorxiv (doi: https://doi.org/10.1101/257642[36])

References

1. Pavlopoulos GA, Kontou PI, Pavlopoulou A, *et al.*: **Bipartite graphs in systems biology and medicine: a survey of methods and applications.** *GigaScience.* 2018; **7**(4): 1–31.
   PubMed Abstract | Publisher Full Text
2. Shannon P, Markiel A, Ozier O, *et al.*: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res.* 2003; **13**(11): 2498–2504.
   PubMed Abstract | Publisher Full Text | Free Full Text
3. Li P, Castrillo JI, Velarde G, *et al.*: **Performing statistical analyses on quantitative data in Taverna workflows: an example using R and maxdBrowse to identify differentially-expressed genes from microarray data.** *BMC Bioinformatics.* 2008; **9**: 334.
   PubMed Abstract | Publisher Full Text | Free Full Text
4. **https://www.docker.com**
5. Di Tommaso P, Chatzou M, Floden EW, *et al.*: **Nextflow enables reproducible computational workflows.** *Nat Biotechnol.* 2017; **35**(4): 316–319.
   PubMed Abstract | Publisher Full Text
6. Costa RL, Gadelha L, Ribeiro-Alves M, *et al.*: **GeNNet: an integrated platform for unifying scientific workflows and graph databases for transcriptome data analysis.** *PeerJ.* 2017; **5**: e3509.
   PubMed Abstract | Publisher Full Text | Free Full Text
7. R Core Team: **R: A language and environment for statistical computing.** R Foundation for Statistical Computing, Vienna, Austria. 2018.
   Reference Source
8. Tuncbag N, Gosline SJ, Kedaigle A, *et al.*: **Network-Based Interpretation of Diverse High-Throughput Datasets through the Omics Integrator Software Package.** *PLoS Comput Biol.* 2016; **12**(4): e1004879.
   PubMed Abstract | Publisher Full Text | Free Full Text
9. Pratt D, Chen J, Welker D, *et al.*: **NDEx, the Network Data Exchange.** *Cell Syst.* 2015; **1**(4): 302–305.
   PubMed Abstract | Publisher Full Text | Free Full Text
10. Ahmed E, Hashish AH: **On modelling the immune system as a complex system.** *Theory Biosci.* 2006; **124**(3–4): 413–418.
    PubMed Abstract | Publisher Full Text
11. Benoist C, Germain RN, Mathis D: **A *plaidoyer* for 'systems immunology'.** *Immunol Rev.* 2006; **210**: 229–234.
    PubMed Abstract | Publisher Full Text
12. Mazzocchi F: **Complexity in biology. Exceeding the limits of reductionism and determinism using complexity theory.** *EMBO Rep.* 2008; **9**(1): 10–14.
    PubMed Abstract | Publisher Full Text | Free Full Text
13. Chaussabel D, Pascual V, Banchereau J: **Assessing the human immune system through blood transcriptomics.** *BMC Biol.* 2010; **8**: 84.
    PubMed Abstract | Publisher Full Text | Free Full Text
14. Langfelder P, Horvath S: **WGCNA: an R package for weighted correlation network analysis.** *BMC Bioinformatics.* 2008; **9**: 559.
    PubMed Abstract | Publisher Full Text | Free Full Text
15. Du P, Kibbe WA, Lin SM: ***lumi*: a pipeline for processing Illumina microarray.** *Bioinformatics.* 2008; **24**(13): 1547–1548.
    PubMed Abstract | Publisher Full Text
16. Barbosa-Morais NL, Dunning MJ, Samarajiwa SA, *et al.*: **A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data.** *Nucleic Acids Res.* 2010; **38**(3): e17.
    PubMed Abstract | Publisher Full Text | Free Full Text
17. Arloth J, Bader DM, Röh S, *et al.*: **Re-Annotator: Annotation Pipeline for Microarray Probe Sequences.** *PLoS One.* 2015; **10**(10): e0139516.
    PubMed Abstract | Publisher Full Text | Free Full Text
18. Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments.** *Stat Appl Genet Mol Biol.* 2004; **3**(1): Article3, 1–25.
    PubMed Abstract | Publisher Full Text
19. Gaujoux R, Seoighe C: **CellMix: a comprehensive toolbox for gene expression deconvolution.** *Bioinformatics.* 2013; **29**(17): 2211–2212.
    PubMed Abstract | Publisher Full Text
20. Chaussabel D, Quinn C, Shen J, *et al.*: **A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus.** *Immunity.* 2008; **29**(1): 150–164.
    PubMed Abstract | Publisher Full Text | Free Full Text
21. Obermoser G, Presnell S, Domico K, *et al.*: **Systems scale interactive exploration reveals quantitative and qualitative differences in response to influenza and pneumococcal vaccines.** *Immunity.* 2013; **38**(4): 831–844.
    PubMed Abstract | Publisher Full Text | Free Full Text
22. Yu G, He QY: **ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization.** *Mol Biosyst.* 2015; **12**(2): 477–479.
    PubMed Abstract | Publisher Full Text
23. Csardi G, Nepusz T: **The igraph software package for complex network research.** *InterJournal.* 2006.
    Reference Source
24. **Neo4j's Graph Query Language: An Introduction to Cypher**. (Accessed: 3rd January 2017).
    Reference Source
25. Chang HH, Soderberg K, Skinner JA, *et al.*: **Transcriptional network predicts viral set point during acute HIV-1 infection.** *J Am Med Inform Assoc.* 2012; **19**(6): 1103–1109.
    PubMed Abstract | Publisher Full Text | Free Full Text
26. Idaghdour Y, Quinlan J, Goulet JP, *et al.*: **Evidence for additive and interaction effects of host genotype and infection in malaria.** *Proc Natl Acad Sci U S A.* 2012; **109**(42): 16786–16793.
    PubMed Abstract | Publisher Full Text | Free Full Text
27. Zhai Y, Franco LM, Atmar RL, *et al.*: **Host Transcriptional Response to Influenza and Other Acute Respiratory Viral Infections--A Prospective Cohort Study.** *PLoS Pathog.* 2015; **11**(6): e1004869.
    PubMed Abstract | Publisher Full Text | Free Full Text
28. Hardy GA, Sieg S, Rodriguez B, *et al.*: **Interferon-α is the primary plasma type-I IFN in HIV-1 infection and correlates with immune activation and disease markers.** *PLoS One.* 2013; **8**(2): e56527.
    PubMed Abstract | Publisher Full Text | Free Full Text
29. McMichael AJ, Borrow P, Tomaras GD, *et al.*: **The immune response during acute HIV-1 infection: clues for vaccine development.** *Nat Rev Immunol.* 2010; **10**(1): 11–23.
    PubMed Abstract | Publisher Full Text | Free Full Text
30. Klein SL, Flanagan KL: **Sex differences in immune responses.** *Nat Rev Immunol.* 2016; **16**(10): 626–638.
    PubMed Abstract | Publisher Full Text
31. Chaussabel D, Baldwin N: **Democratizing systems immunology with modular transcriptional repertoire analyses.** *Nat Rev Immunol.* 2014; **14**(4): 271–280.
    PubMed Abstract | Publisher Full Text | Free Full Text

32. Vaske CJ, Benz SC, Sanborn JZ, *et al.*: **Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM.** *Bioinformatics.* 2010; **26**(12): i237–45.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

33. Deffur A: **adeffur/ANIMA: ANIMA Source code (Version v1.0.0).** *Zenodo.* 2018.
http://www.doi.org/10.5281/zenodo.1163398

34. Deffur A, Wilkinson RJ, Mayosi BM, *et al.*: **ANIMA: Association Network Integration for Multiscale Analysis (TAR archive of anima Docker image used in the publication) (Version 3.3.3).** *Zenodo.* 2018.
http://www.doi.org/10.5281/zenodo.1161476

35. Deffur A, Wilkinson RJ, Mayosi BM, *et al.*: **ANIMA: Association Network Integration for Multiscale Analysis (source data, metadata and scripts) (Version 1.0.0).** *Zenodo.* 2018.
http://www.doi.org/10.5281/zenodo.1161381

36. Deffur A, Wilkinson RJ, Mayosi BM, *et al.*: **ANIMA: Association Network Integration for Multiscale Analysis.** *bioRxiv.* 257642.
Publisher Full Text

Open Peer Review

Current Referee Status: ✓ ? ✓

Version 3

Referee Report 10 December 2018

https://doi.org/10.21956/wellcomeopenres.16261.r34252

✓

Emre Guney
Research Programme on Biomedical Informatics, the Hospital del Mar Medical Research Institute, Pompeu Fabra University, Barcelona, Spain

The authors have appropriately addressed my previous comments.

Competing Interests: No competing interests were disclosed.

Referee Expertise: Systems biology, disease bioinformatics, network pharmacology, personalized medicine

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 2

Referee Report 10 September 2018

https://doi.org/10.21956/wellcomeopenres.15932.r33740

?

Emre Guney
Research Programme on Biomedical Informatics, the Hospital del Mar Medical Research Institute, Pompeu Fabra University, Barcelona, Spain

Deffur and colleagues describes a transcriptomics data integration and analysis pipeline named ANIMA. ANIMA starts from raw microarray data sets, processes them using R and then applies WGCNA and another previously published module discovery method. The resulting networks (covering different relationships extracted from the data sets such as phenotype-differentially expressed genes, gene-gene co-expression, etc.) are stored using Neo4j graph database, allowing fast querying of relationships that satisfy certain criteria. In addition to the pipeline, the authors build a web application using Shiny. They also showcase the use of the pipeline by analyzing three publicly available data sets. The article explains the technical details of the pipeline clearly. Although ANIMA currently supports only Illumina BeadArray other platforms such as Affymetrix and sequencing data is planned to included in the future. Indeed, addition of other platforms and RNAseq data analysis capabilities would be a important improvement.

1. The intro falls short on existing related work, several recent efforts aim to integrate, analyze and store omics data as well as provide a reproducible work flow such as

    - Omics Integrator (Tuncbag et al., 2016, Plos Comp Bio)
    - NDEx (Pratt et al., 2015, Cell Systems)
    - nextflow (Di Tomasso et al., 2017, Nat Biotech)

2. How does ANIMA differs from GeNNet?

3. Are there any plans for including protein interaction network analysis tools such as BIANA (Garcia-Garcia et al., 2010, BMC Bioinformatics) into the pipeline? Also, how easy is to customize the ANIMA, say how easy to choose another co-expression network analysis software?

4. What do the P-values on the title of the Figure 3 correspond to? Which groups were compared?

5. I find using color names to represent modules in Figure 4 highly confusing, especially given that blue is also used to denote underexpression. They can be renamed and/or described better in the legend.

Minor:

- The focus on immune system in the beginning of the intro constitutes a slight disconnection with respect to the abstract / rest of the text. I understand the focus due to the case studies but still reads isolated from the main theme of the paper, which is introducing ANIMA as a framework. The authors might as well redirect the focus to inference of cell state for immune response but then they would need to provide a more stringent validation of the obtained results in terms of the biological findings.

- The use of subsections in the introduction feels rather unconventional for scientific writing and somewhat confusing given that the current approaches only mention one existing approach followed by the proposed approach.

- Question 2 in problem definition fails to pose a general question, going to the technical details before having explained the graph under concern, the triples, P-value correction etc.

- In fact, to me it seems that the authors describe the problem in the latter “rationale” sections, where they highlight the problem on the storage of research-related results and network data as well as the difficulty of reproducibility due to software / platform version changes.

- “We prefer scripted workflows ...” We? Scientists?

- “high order, high complexity and low dimensional state space vectors from high dimensional low order, low complexity data (i.e. non-normalised microarray data and clinical/sample phenotype” not clear what the authors mean by the order and complexity of the data (I see later in the figure 1 that it is biological complexity). Moreover, this sentence is very long and can use a revision.

- Some references are not in right order (e.g., Chaussabel refered as 14 while being 17).

- suffixes => suffices

Is the rationale for developing the new method (or application) clearly explained?

Partly

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

Referee Expertise: Systems biology, disease bioinformatics, network pharmacology, personalized medicine

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 30 Oct 2018

Armin Deffur, University of Cape Town, South Africa

Many thanks for the thorough review and comments.

The introduction has been reworked to address most of the comments made in the referee report. It now reads more clearly. Suggested references have been incorporated.
With respect to comment (4), the legend of figure 3 addresses the question. The introduction now reduces the focus on the immune system, first discussing the ideas underlying the ANIMA framework, before showcasing ANIMA by example data derived from immune responses to infectious diseases.

Subsections in the introduction have been removed.

Vague allusions to the theory of complex systems have been either removed or made clearer.

While I agree that colour names for WGCNA modules can be confusing, I avoided relabelling the modules manually to remain true to the original output. Replacing colours with numbers would not offer a significant benefit, and coming up with function-related names would be to some extent subjective. Indeed, the annotation of the WGCNA modules suggests that assigning singular descriptions to them is not straightforward.

Competing Interests: No competing interests were disclosed.

Referee Report 26 June 2018

https://doi.org/10.21956/wellcomeopenres.15932.r33262

✓ **Christopher L. Plaisier**
Arizona State University, Tempe, AZ, USA

Overall the revisions have improved the work substantially. The addition of Table 3 was helpful, but it left additional questions.

Critiques:

1. Table 3 is helpful but the data types are named in ways that are not described anywhere in the main text. For instance baylor is Chaussabel modules. But it takes a lot of effort to figure this out. It would be very helpful to have an additional table that describes each data type. Without this the information in Table 3 is less useful.

2. The data type baylor in table 3 ought to be changed to Chaussabel for the sake of consistency.

Competing Interests: No competing interests were disclosed.

Referee Expertise: Computational biology, systems biology, networks, statistics, infectious disease, immunology

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 30 Oct 2018

Armin Deffur, University of Cape Town, South Africa

Many thanks for reviewing the second version of the paper.

Table 3 has been updated to indicate the "Chaussabel modules" are equivalent to the "baylor" datatype. In future versions of ANIMA we will change the datatype in the source code; as an aside it would be ideal to avoid eponyms for this, as in the case of WGCNA.

We did not include a table describing each individual datatype in the main text, but it is available in the Supplementary information as Table S5, and detailed descriptions are given on page 9-10 of the Supplementary file.

Competing Interests: No competing interests were disclosed.

Version 1

Referee Report 17 April 2018

https://doi.org/10.21956/wellcomeopenres.15308.r32374

?

James A. Eddy
Sage Bionetworks, Seattle, WA, USA

Deffur et al. present ANIMA, a network-based method for integrating gene expression (microarray), annotation data, and clinical data. ANIMA utilizes a merged graph database to represent and allow exploration of associations/relationships between biological processes and cellular contexts at multiple scales. The authors provide software (R, Docker) and extensive documentation for building and interacting with ANIMA networks for several example infectious disease datasets; they also present a web interface (R/Shiny application) for interactive visualization of results. The example datasets are used to demonstrate three central approaches for interpreting data with ANIMA, highlighting selected results and features of the method and tools.

I would be interested in using this method for some of my own projects related to systems immunology and immuno-oncology, and the web tool seems like it would provide a nice tool for biologists who wish to explore a particular dataset. As a “framework,” ANIMA would be even more appealing to computational biologists as a more modular/extensible system — but still offers a healthy breadth of analyses in its current state. While this manuscript is technically sound, there are issues with structure and organization that I would recommend addressing. In particular, the authors should (1) more clearly describe the rationale for their methods; and (2) present method details more prominently in the main body of the manuscript.

Major comments/critiques

Regarding rationale:

1. The authors mention the ability to “discuss [immune response] not in the language of probes, genes and signatures, but rather as coordinated biological processes and the cellular context for these processes.” I like this idea, and I feel that it’s a key distinction of the ANIMA method. However, the authors should explain why they feel this approach is better or complementary to other, more feature-based multi-scale association networks (e.g., Vaske et al., Bioinformatics. 2010. 26(12)).
2. The authors do not indicate whether any similar methods to their own (e.g., Costa et al., PeerJ. 2017. 5:e3509 — perhaps) exist or, if so, describe how ANIMA compares to or improves upon such methods.
3. The authors should provide at least some justification for the technical details of their method and tools: Why use a graph database; why neo4j? How did they select specific packages/implementations for analytic approaches (Table 1)? Alternatively, the authors could demonstrate that associations in the final network are “robust” to choices of packages and parameters for individual analyses — however, this might be more appropriate for a follow-up study.

Regarding method details:

1. Given that ANIMA, the subject of this Methods article, is described as “a network-based data integration method,” the authors should present more details about the network building process — beyond the relatively brief overview in the “Method outline” section. The Supplementary Methods and Figures are immensely helpful for understanding the manuscript; suggestions for pieces of information to move or at least duplicate in the main text include:

- Enumeration of full set of bipartite graphs (i.e., from Supplementary Methods ~ “Bipartite graphs based on association indices”, Figure S7) — a condensed version of the Supp. Methods section in table form would likely be sufficient; however, references to Figure 1 should clearly indicate that this figure is illustrative and does not include the full scope of output types and relationships used.
- Reference to detailed methods for analytic approaches (i.e., from Supplementary Methods) — this could be included as an additional column in Table 2, referencing a specific page or section number in the Supplementary Methods. Additionally, it would be helpful to list the bipartite graphs for which each analytic approach serves as input or output (referencing the graph number 1-29 and/or the suggested table above).

2. Somewhere in the methods section, it would also be helpful to include a brief overview of how a user would approach and use ANIMA to build a new network for their data:
    - This point might be more obvious to some, but was lost on me at first: clarify that ANIMA networks produced by the method is specific to the input dataset(s), and there’s not a central, global “ANIMA” graph database (though providing a repository to share and explore user-generated ANIMA networks could be an exciting future direction).
    - A simplified, more process-oriented version of Figure S9 (don’t need to list specific functions and scripts, but diagram and summary of inputs, outputs, and steps) would be great.
    - Describe computational requirements (operating system, CPUs, RAM, time, etc.) to perform various steps.
    - Describe which microarray platforms are supported and any expected format requirements for the input data.

3. The three approaches presented in the “Results” section give helpful examples of how a user can interact with an ANIMA network, once constructed, and what can be learned from the method*. Additional results of interest include:
    - Performance (related to requirements, but actual runtime statistics for examples presented);
    - Comparison to other methods (if applicable).

*Note: I really think the authors could consider presenting the Shiny-based web tool (and corresponding “use cases”) as a separate “Software Tool” article, as the target audience and key results are subtly different than the ANIMA method; however, I leave this to the discretion of the authors.

Minor/discretionary comments

1. For the Cypher query example, a corresponding “translation” of the syntax would be helpful — either as something more SQL-like or a verbose, literal explanation of what the query is doing.
2. There’s a typo in Figure S10 (“F[i]lesystem”).
3. In the “Discussion” section, the reference to “selection of one or a few cell types” as a limitation of single-cell RNAseq is increasingly less true with technologies like 10x — might want to revise.
4. For differential transcript abundance prior to WGCNA, genes are “ranked” and the top 4000 are selected; similarly, the 500 are used in pathway enrichment. For Chaussabel modules, was any cutoff used to determine which genes were up- or down-regulated?
5. Authors should clarify that for this particular iteration (due to the choices of cell types from Abbas et al. and the Chaussabel modules), the method is most applicable to datasets with an immune component or context.

Is the rationale for developing the new method (or application) clearly explained?

Partly

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

Referee Expertise: Computational biology, systems biology, systems immunology, bioinformatics

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 20 May 2018

Armin Deffur, University of Cape Town, South Africa

Many thanks for the thorough and comprehensive review. The following changes were made in order to address the major concerns:

1. We discuss the key distinction of ANIMA compared to other methods as the ability to generate virtual cells from the expression data and annotation libraries; the specific example referenced (PARADIGM) outputs rankings of pathways
2. We thank the reviewer for pointing out the Costa et al paper, which was published after our own development of ANIMA. While both approaches rely on scripted workflows, graph databases and reproducible computational environments, our approach is sufficiently novel in its ability to generate new information from the multipartite graph.
3. A more detailed introduction is now provided, providing extended justification for the approaches used
4. The Methods section was reworked by moving some material from the supplementary information to the main manuscript, and adding a new Table 3, which lists the individual bipartite graphs. Figure S7 is now cross-referenced to Table 1, and the various workflow components indicated in the Table and Figure.
5. A section has been added for the requirements (hardware, data format, etc.) of using ANIMA for a user's own data, and the outline for the procedure is listed.
6. A simpler version of Figure S9 has been added as a subfigure to Figure 1 in the main manuscript.

7. We feel that the Shiny web application is key to delivering the functionality of ANIMA from a user perspective, and have retained the relevant sections. In future, once technical issues regarding hosting a multi-container based web-app have been solved, and we start hosting the web interface, we may publish an application note describing this.
8. We have translated the Cypher syntax for the example query as suggested.
9. The typo in Figure S10 has been corrected.
10. The discussion of single-cell RNAseq has been revised.
11. The information on the cutoff for significant up- down regulation of genes in Chaussabel modules has been added to the relevant section in Supplementary information.
12. We emphasise the immunology use case for ANIMA in its present form, but leave open the possibility for additional use cases with appropriate code changes (future work section)

Competing Interests: None

Referee Report 27 March 2018

https://doi.org/10.21956/wellcomeopenres.15308.r31798

Christopher L. Plaisier
Arizona State University, Tempe, AZ, USA

This manuscript provides a description of a method to infer many useful computational relationships from patient data in one comprehensive resource. The method is tested in proof-of-principle studies of human patient data from infectious diseases. The examples are fairly thorough, but cannot possibly cover the complex nature of the relationships that are potentially computed in a cohesive manner. Nor are the statistical filters that should be or are applied described in detail.

Major critiques:

1. All the potential cutoffs and alpha values should be described and how to correct for multiple hypothesis corrections for each analysis, and overall. One good place for this might be Table 2.

2. A better table must be provided with the following column headers: "Data Type 1", "Data Type 2", "Statistical Method", "Implementation", "Biological Implication", "Description", "Reference(s)"

3. Figure 1a why is there a distinction between Biological Pathway and Chaussabel Module? They are both forms of prior information that is overlaid onto the network via the same method. A more accurate description would be "Gene Sets from Prior Information" or simply "Gene Sets".

4. The first section of the results is Validation study. This would be better described as an Example study as validation means studies done to confirm prior observations.

5. This paper reads more like a manual than a paper. One thing that would be very helpful would be a demonstration of how the network-based approach provides a synthesis of information and describes something novel or at least puts together things that weren't previously associated.

6. From a computational perspective, the motivation is clear. But from the perspective of a potential biologist why would they use this over something else.

7. What tools does this compare to and how is your method an improvement?

Is the rationale for developing the new method (or application) clearly explained?
Partly

Is the description of the method technically sound?
Partly

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

Referee Expertise: Computational biology, systems biology, networks, statistics, infectious disease, immunology

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 20 May 2018

Armin Deffur, University of Cape Town, South Africa

Many thanks for a thorough review. We made the following changes to address the stated concerns.

In response to points 1 and 2:
We added adjusted P-values used as cutoffs in the analytic pipeline in Table 1, and have added a new Table 3, providing detail on the bipartite graphs, based on the recommendation.

3. Figure 1 has been edited to clarify that both pathways and Chaussabel modules are gene sets forming prior information.

4. The term "validation study" has been replaced with "example study"

5. We emphasise in the Results section that the "virtual cells" and the cell matrices are new

information we were able to generate based on the shared relationships found in ANIMA

6. We have added a short section on target audience, in order to clarify use cases for both computational biologists and immunologists

7. Additional material on comparable tools is provided

Competing Interests: None