NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data

Jianguo Xia[1–3], Erin E Gill[1] & Robert E W Hancock[1,4]

[1]Department of Microbiology and Immunology, University of British Columbia, Vancouver, British Columbia, Canada. [2]Institute of Parasitology, and Department of Animal Science, McGill University, Ste. Ann de Bellevue, Québec, Canada. [3]Department of Microbiology and Immunology, McGill University, Montreal, Québec, Canada. [4]Wellcome Trust Sanger Institute, Hinxton, United Kingdom. Correspondence should be addressed to J.X. (jeff.xia@mcgill.ca) or R.E.W.H. (bob@hancocklab.com).

Published online 7 May 2015; doi:10.1038/nprot.2015.052

Meta-analysis of gene expression data sets is increasingly performed to help identify robust molecular signatures and to gain insights into underlying biological processes. The complicated nature of such analyses requires both advanced statistics and innovative visualization strategies to support efficient data comparison, interpretation and hypothesis generation. NetworkAnalyst (http://www.networkanalyst.ca) is a comprehensive web-based tool designed to allow bench researchers to perform various common and complex meta-analyses of gene expression data via an intuitive web interface. By coupling well-established statistical procedures with state-of-the-art data visualization techniques, NetworkAnalyst allows researchers to easily navigate large complex gene expression data sets to determine important features, patterns, functions and connections, thus leading to the generation of new biological hypotheses. This protocol provides a step-wise description of how to effectively use NetworkAnalyst to perform network analysis and visualization from gene lists; to perform meta-analysis on gene expression data while taking into account multiple metadata parameters; and, finally, to perform a meta-analysis of multiple gene expression data sets. NetworkAnalyst is designed to be accessible to biologists rather than to specialist bioinformaticians. The complete protocol can be executed in ~1.5 h. Compared with other similar web-based tools, NetworkAnalyst offers a unique visual analytics experience that enables data analysis within the context of protein-protein interaction networks, heatmaps or chord diagrams. All of these analysis methods provide the user with supporting statistical and functional evidence.

INTRODUCTION

High-throughput omics studies are rapidly generating a wealth of data sets from many model organisms in various physiological states, disease conditions or treatments. Recent years have seen a growing interest in conducting integrative analysis on these data sets to help reduce study bias, increase statistical power and improve mechanistic understanding[1–4]. Two types of data integration are commonly used. Horizontal data integration aims to integrate data sets measuring the same molecular events—e.g., combining multiple gene expression data sets to identify core signatures of complex diseases or host responses[1,2]. Conversely, vertical data integration involves joint analysis of data sets from different levels in the omics cascade, such as integrating gene expression and metabolomics data to gain insights into the underlying biological processes[3]. The term meta-analysis usually refers to horizontal data integration. It is sometimes also used to describe comprehensive exploratory analyses on a single data set with regard to multiple associated metadata or phenotypic labels.

Bioinformatics for data integration is still in its infancy, and both the theories and practices are still evolving quickly[5–7]. Three general approaches have emerged, namely the semantic web approach, the database approach and the data mining approach. The semantic web strategy uses an ontology-based framework to enable automatic data integration in a meaningful and intelligent manner[8,9]. The database-based approach involves creating a comprehensive knowledge base by collecting and organizing data concerning a particular situation such as a disease or biological system[10–12]. However, the majority of researchers rely on various data mining approaches to analyze their own collection of data sets. Over the past decade, it has been recognized that effective visual exploration by domain experts is the key factor for successful extraction of knowledge from integration of large and complex omics data sets[5,7]. Interactive visualization tightly coupled with robust statistical and machine-learning techniques has proven to be the most effective strategy for facilitating data understanding and hypothesis generation. This type of visual data mining technique is now generally referred to as visual analytics[13,14].

Until recently, to perform such analyses, researchers generally needed to be familiar with several complementary bioinformatics tools and/or master a statistical programming language. To address the growing need for biologist-friendly tools and to enable bench researchers to examine data sets for novel insights, we have developed a series of web-based tools for statistical meta-analysis[15], visual data mining[16] and data integration through the rapid generation of biological networks[17]. Together, they offer a comprehensive tool suite that allows various common and complex meta-analyses to be carried out by biologists via a standard web browser. Based on user requests, we have recently integrated all three tools into NetworkAnalyst, added new features for visual analytics and expanded support for more model organisms. NetworkAnalyst has been implemented using an innovative visual analytics framework that integrates server-side statistical computing based on the R language (http://www.r-project.org/) with client-side visualization using HTML5 and JavaScript. By harnessing the power of modern web browsers, NetworkAnalyst reduces the computational load on the server and provides users with a fast and interactive visual analytics experience. These features will assist biologists in the generation of novel hypotheses regarding the mechanisms underlying or driving differences in complex data sets, including fundamentally new concepts, and the expression differences (biomarkers) that underlie the different conditions analyzed.

Comparison with other available tools for data integration

Many excellent web-based tools exist for analyzing a single omics data set, such as GEPAS[18], GenePattern[19], MetaboAnalyst[20,21] and so on. Similarly, multiple locally installed programs are available for interactive visual exploration of omics data or biological networks such as TM4 (ref. 22), Gitools[23], Cytoscape[24] and so on. Bioinformatics tools that can be used for integrative data analysis are rare and have diverse forms with very different capacities. **Table 1** compares NetworkAnalyst with several tools that allow users to upload their own data sets for data integration analysis.

The most popular web-based tools for integrative analysis of gene list data are probably DAVID[25] and g:Profiler[26]. The key strength of both tools is the comprehensive support of various gene identifiers, microarray platforms and their associated functional annotation. They are also easy to use and fast in response. For locally installed programs, Cytoscape is widely used for network-based data integration[24]. Users usually need to prepare a network file and/or attribute data for layout and visualization with Cytoscape. A popular approach is to first upload the gene list to InnateDB[27] to build the network file and then visualize the result through Cytoscape. Gitools is a stand-alone tool for heatmap-based data visualization and integration[23]. Interactive heatmaps coupled with functional enrichment analysis make it a very powerful visual analytics tool.

In contrast, NetworkAnalyst is a completely web-based application designed to address the statistical and visualization

TABLE 1 | Comparison of tools for data integration.

Tools	NetworkAnalyst	DAVID	g:Profiler	InnateDB and Cytoscape	Gitools
URLs	http://www.networkanalyst.ca	http://david.abcc.ncifcrf.gov	http://biit.cs.ut.ee/gprofiler/	http://www.innatedb.ca http://www.cytoscape.org	http://www.gitools.org
Software type	Web-based	Web-based	Web-based	Mixed	Stand-alone
Supported organisms	Human, mouse, *D. melanogaster*, *C. elegans* and *S. cerevisiae*	>65,000 species	>85 species	Human, mouse and cow	Independent (user supplied)
Approaches to integration	Statistical meta-analysis, visual integration and PPI network	Comprehensive knowledge base	Comprehensive knowledge base	PPI network	Visual integration
Data input	Gene or protein lists and gene expression data	Gene or protein lists	Gene or protein lists	Gene or protein lists	Genomic or gene expression data
Statistical analysis	Meta-analysis (*P* value, effect size, rank, vote counts and direct merge)	NA	NA	NA	*P* value combination, correlation analysis
Enrichment analysis	GO and pathway	>40 categories	GO, pathway, TFBS, microRNA, disease and PPI modules	GO, pathway and TFBS	Any (user supplied)
Network analysis					
Interactive	✓		✓	✓	
Topology	✓			✓	
Customization	✓			✓	
Other visual analytics					
Chord diagrams	✓				
Heatmaps	✓	✓			✓

NA, not applicable. See **Box 1** for definitions of terms.

challenges during meta-analysis of complex gene expression data sets. It supports meta-analysis of gene lists or a gene expression data set labeled with multiple metadata parameters, as well as of multiple independent gene expression data sets. Data integration is achieved through robust statistical procedures and then visually explored within protein-protein interaction (PPI) networks, interactive heatmaps or chord diagrams. The network visualization component of NetworkAnalyst has been developed as a high-performance web-based alternative to Cytoscape, with universal online access and built-in support for statistical meta-analysis. The current implementation supports fast searching, zooming and customization, together with module detection and functional enrichment analysis—the two most common needs during network analysis[28]. By adopting cutting-edge web techniques, NetworkAnalyst offers powerful features that were previously only available with locally installed programs and the advantage of substantially increased analysis speed and centralized management. The integration of statistics, interactive graphics and *in situ* functional analysis allows data to be examined from different perspectives, which not only brings more transparency in data analysis but also provides different ways for data integrity validation.

Applications

NetworkAnalyst contains many features and functions for a variety of purposes and applications. In particular, NetworkAnalyst supports flexible differential expression analysis (DEA, see **Box 1** for glossary) for a wide array of experimental designs including two or multiple group comparisons, time series, common-reference design, nested comparisons, as well as paired or block design. The underlying statistics are based on the well-established limma package[29], which was initially designed for microarray data analysis and has recently been extended to also support RNA-seq data[30]. For microarray data from humans and mice, NetworkAnalyst contains built-in support for popular commercial microarray platforms from Affymetrix GeneChip, Illumina's BeadArray and Agilent one-color microarrays. For data from RNA-seq or other array platforms, users need to first convert the platform-specific feature IDs to common gene or transcript IDs (Entrez, RefSeq, Genbank or Ensembl) before uploading the expression tables to NetworkAnalyst. Details on DEA will not be covered in this protocol. Instead, this protocol focuses on three unique features of NetworkAnalyst, network analysis, visual analytics and meta-analysis, through analysis of three separate example data sets. For more technical details or introductions to other features, users are encouraged to consult our screenshot tutorials and FAQs posted on the NetworkAnalyst website (http://www.networkanalyst.ca).

Limitations of the software and protocol

NetworkAnalyst currently supports only five model organisms (human, mouse, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Saccharomyces cerevisiae*), for which high-quality PPI data are available. The underlying protein interaction data were obtained from InnateDB[11] (for human and mouse), a member of the International Molecular Exchange (IMEx) consortium[31], and iRefIndex[32] (for other three species). Both databases provide comprehensive PPI coverage by integrating data from a number of primary interaction databases including BIND[33], BioGrid[34], MINT[35], IntAct[36] and so on. For human data, we also offer the option to use the high-quality experimentally validated binary interaction data, as reported recently by Rolland *et al.*[37].

Box 1 | Glossary

DAVID—Database for Annotation, Visualization and Integrated Discovery: http://david.abcc.ncifcrf.gov/
DE—Differentially expressed
DEA—Differential expression analysis
FAQ—Frequently asked questions
FC—Fold change
FDR—False discovery rate: a statistical procedure for multiple testing correction
FEM—Fixed effects model: a statistical model to combine effect sizes of a set of studies
GEO—Gene Expression Omnibus: a repository for gene expression information
GO—Gene ontology: formal functional categorization of genes into functional associations
KEGG—Kyoto Encyclopedia of Genes and Genomes: http://www.genome.jp/kegg/
Limma—Linear models for microarray data
LPS—Lipopolysaccharide: bacterial signature molecule that interacts with Toll-like receptor 4 on host cells to trigger innate immune/inflammatory responses. Two consecutive treatments trigger a cellular reprogramming termed endotoxin tolerance
ORA—Over-representation analysis
PBMCs—Peripheral blood mononuclear cells: white blood cells
PCA—Principal component analysis
PNG—Portable network graphics: a graphics file format that supports data compression without loss of information
PPI—Protein-protein interaction: physical, biochemical or regulatory interactions between gene products
REM—Random-effects model: a statistical model to combine effect sizes of a set of studies
SVG—Scalable vector graphics: XML-based vector image format for 2D graphics that supports interactivity and animation
TFBS—Transcription factor binding site: the position at which a transcription factor binds; genes sharing TFBS are likely under the control of the same transcription factor
TLR—Toll-like receptor: a receptor on immune cells that responds to microbial signatures, also called pathogen-associated molecular patterns

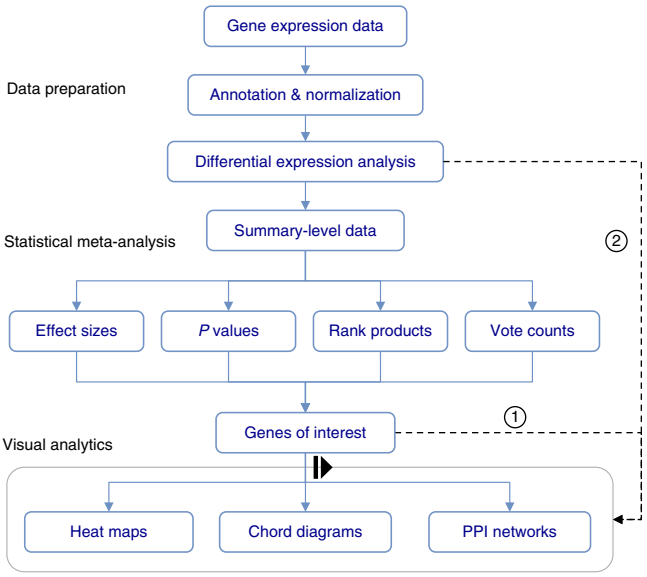

Figure 1 | NetworkAnalyst analysis flowchart. NetworkAnalyst is composed of three main functional modules: data preparation, statistical meta-analysis and visual analytics. The complete workflow is available when users perform meta-analysis on multiple gene expression data sets. Users can also perform meta-analysis on gene lists (dashed line 1) or a single gene expression data set (dashed line 2). The 'resume' icon indicates the point at which users can download a copy of the session and resume the analysis later on.

At this moment, the tool does not allow users to upload their own PPI data for analysis. However, the design of the tool is very flexible and scalable, and we intend to add support for more species, more platforms and more PPI databases. Users are encouraged to give us feedback, as many recent updates were in response to user requests.

NetworkAnalyst has been designed as a high-performance web application for real-time interactive data analysis and visual exploration. When a user begins analysis, a temporary folder is created in which to save user data. The folder is removed after the session completes. However, visual data mining is an iterative process, and many users have requested to be allowed to continue previous analyses instead of starting from the beginning each time. Thus, although the public server does not allow users to permanently store their data to enable them to resume analysis at a later date, we recently added the capability to allow users to download a copy of the session file. When they return, they can simply upload the session file to enter the visual analytics page without having to repeat the time-consuming data preparation steps.

Server performance is another potential limitation when many users upload large data sets at approximately the same time. The public server currently limits each user to upload a maximum of 1,000 samples. Researchers who wish to conduct very large-scale meta-analyses are encouraged to contact the authors to obtain a copy of NetworkAnalyst to install on their local server (Mac or Linux based).

Analysis overview

Meta-analysis can be applied to three different types of input data: gene lists, a single gene expression data set or multiple gene expression data sets. **Figure 1** shows the overall design of NetworkAnalyst. The complete workflow can be applied when users upload multiple gene expression data sets. There are three main analysis stages: data preparation, statistical meta-analysis and visual analytics. In the data preparation stage, an individual data set is uploaded for annotation, normalization and DEA. Subsequently, during statistical meta-analysis, the summary-level data (i.e., *P* values, fold changes (FCs) or effect sizes) are extracted and integrated to identify genes that are significantly altered in expression, based on the overall evidence. Finally, the selected genes are presented within PPI networks, chord diagrams or heatmaps to enable interactive visual data mining. NetworkAnalyst has also been designed to support meta-analysis on a single gene expression data set labeled with multiple metadata parameters, as well as supporting network analysis on gene lists. In such cases, users need to be aware that some functions may become inapplicable for certain data types. In particular, chord diagrams are used to compare multiple analysis results, and they will not be available when only a single analysis has been performed; similarly, heatmaps require gene expression data and cannot be used for gene lists.

During analysis, NetworkAnalyst provides three complementary navigation aids, the history track on the top of the page, the buttons at the bottom and real-time system messages (**Fig. 2**). When users enter a particular analysis page, the system will provide feedback to guide users to complete the necessary steps. Afterward, users need to click the 'Proceed' button to enter the next page. Those completed steps will then become

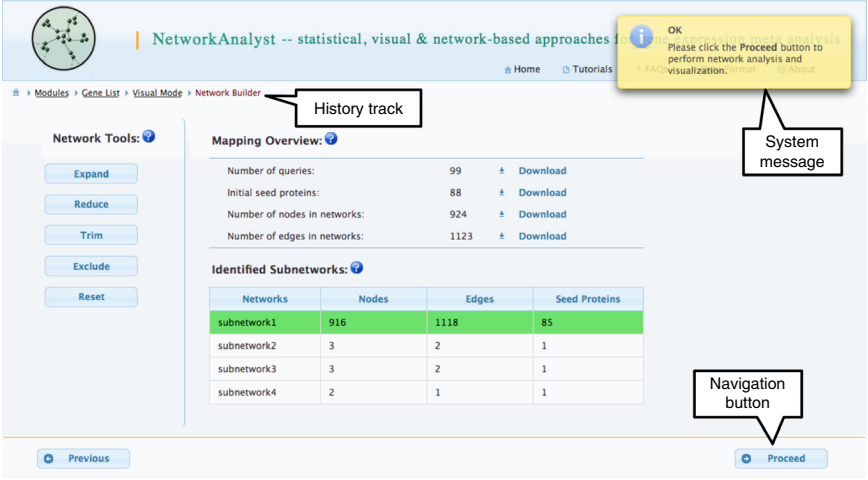

Figure 2 | Navigations in NetworkAnalyst. This screenshot of the 'Network Builder' page illustrates the NetworkAnalyst navigation controls during data analysis. The top right corner shows a real-time system message indicating the current status, and it also provides suggestions for the next step. The top left corner shows the history track with the current page highlighted in orange. Users can click the links to go back to the corresponding pages. The buttons on the bottom allow users to move forward to the next page or return to the previous page. Detailed instructions are available for each function when users hover the mouse over a question mark icon.

part of the history track, with the current step highlighted in orange. Users can click a link on the history track to go directly back to the corresponding page.

The protocol below is organized into three sections with increasing complexity: (i) network analysis for a single gene list (Steps 1–16); (ii) meta-analysis of a single gene expression data set using heatmaps and chord diagrams (Steps 17–31); and (iii) meta-analysis of multiple data sets (Steps 32–50). Each section illustrates one major function of NetworkAnalyst. Three example data sets are provided to demonstrate these procedures. The first data set is a gene list of the 99-gene endotoxin tolerance signature that predicts severe sepsis and organ failure and was identified during a recent study[2]. The second data set is the original gene expression data comparing lipopolysaccharide (LPS)-induced inflammation with endotoxin tolerance using human peripheral blood mononuclear cells (PBMCs)[38]. The third data set includes three colon cancer gene expression studies downloaded from the Gene Expression Omnibus (GEO). All the example data sets can be downloaded from the 'Data Formats' page on the NetworkAnalyst website. Alternatively, users can directly select these data sets from the 'Try Examples' dialog box during analysis.

Network analysis and visualization for a list of genes (Steps 1–16)

The result of gene expression analysis is usually a list of genes that are significantly differentially expressed under the experimental

Box 2 | PPI network analysis

PPI networks are often presented as undirected graphs, with nodes representing proteins and edges indicating known interactions between two connecting proteins. Two types of analyses are often performed. Topology analysis studies the network itself, and the goal is to understand its overall structural properties and organizational principles, whereas the subnetwork or module analysis focuses on parts of the network that have shown significant changes in conditions under study, and the objective is to identify important nodes, their relationships and collective functions. A subnetwork analysis can be executed in three steps: (i) obtaining the PPI data; (ii) building networks based on the user-supplied genes or proteins (called seed nodes in network construction); and (iii) analyzing these networks. The main concepts and common practices are briefly discussed below.

Network types

First-order interaction network: composed only of the seed nodes and the nodes that interact directly with them. This is the default type of network created by NetworkAnalyst.

Zero-order interaction network: composed only of the seed nodes and the edges that interconnect them. This is suitable when there are a large number of seed nodes (>500); it can be created by applying the Reduce function on the default network.

Higher-order interaction network: created by including nodes that are more distant from the seed nodes—e.g., by including second, third and other interactors; it is suitable when there are very few seed genes. It can be created by applying the Expand function on the default network.

Minimum interaction network: created by trimming the first-order network to keep only those nodes that are necessary to connect the seed nodes; this is suitable when the first-order network is too dense, whereas the zero-order network is too sparse. This type of network can be created by applying the Trim function. All these functions are available on the Network Builder page (**Fig. 2**).

Analysis using modules

Sometimes, a network or subnetwork is still too complex to comprehend. In this case, it is recommended to further break it into smaller units (modules) that are more easily interpretable. The main goal in module analysis is to reduce the network complexity while still keeping the most interesting functions and connections. How to identify the most relevant modules based on user input remains an open issue[40]. Mathematically optimal solutions that simultaneously capture most biological changes while maximizing connectivity are computationally intensive, and the resulting modules are often still very difficult to interpret[41,42]. NetworkAnalyst offers two heuristic approaches to enable fast and intuitive module analysis.

In the 'function-first approach', users first identify enriched functions within the network, and then extract the module containing the nodes that are involved in the particular functions of interest; this approach will guarantee that the resulting module is interpretable according to the basic functions or pathways that it reflects.

In the 'connection-first approach', users first perform module detection based on the overall pattern of connectivity, and the resulting modules are then evaluated on the basis of experimental data obtained from user input, as well as pathway or functional enrichment analysis results. This approach will guarantee that the resulting modules will be highly connected. These two approaches are very complementary.

Analysis of node importance

Within a module or subnetwork, not all nodes are equally important. Many studies have shown that 'hub' proteins are more likely to be encoded by pleiotropic genes or related to diseases[43,44]. Hub nodes are potentially key molecules in signaling, as they are highly interconnected with dysregulated genes[45,46]; they receive and integrate multiple signals and pass them on to downstream nodes. The most commonly used measure for node 'hubness' is centrality. In particular, degree centrality is the number of connections that a node has to other nodes, whereas betweenness centrality corresponds to the number of shortest paths passing through the node. These are complementary measures of node centrality and hubness. Hub degree reflects local properties, whereas betweenness is based on the global connectivity pattern. Nodes that occur between two dense clusters will have high betweenness values but low degree values. The node properties table in NetworkAnalyst displays both of these measures adjacent to the networks that are created. Users can simply click on a node name to zoom in to see its position and interconnectivities within the current subnetwork.

Box 3 | Visual analytics for data integration

Owing to their size and complexity, direct visualization of omics data sets is usually not possible or effective. A common practice is to apply analytics to reduce the large amount of data to a point where it can be effectively visualized and explored using powerful visualization strategies. During the past decade, many innovative approaches have been developed for gene expression data presentation and integration[47]. Three main approaches have been implemented in NetworkAnalyst, namely, networks, heatmaps and chord diagrams.

Networks

Biological networks are usually large and complex, without inherent coordinates to enable easy navigation (see **Fig. 3** for an example). Efficient layout, searching and zooming and customization are crucial support functions during network analysis.

Network layout: good layout algorithms arrange nodes in a way that balances both visual perception and aesthetics. NetworkAnalyst offers several state-of-the-art network layout algorithms based on the well-known Gephi software[48].

Searching and zooming: this function allows users to pinpoint their genes of interest within a complex network. NetworkAnalyst will automatically perform a search and zoom in when the user clicks on or enters a node name for search.

Customization: it is very rare that the default layout or style will be satisfactory; further manual adjustments are often needed to improve the quality and to enhance biological insights. In NetworkAnalyst, the size, color and position of nodes can be customized individually or in batch mode (i.e., with a node cluster such as a pathway).

Chord diagrams

Chord diagrams are an elegant and compact way for summarizing and comparing the similarities and differences among different entities. They are very suitable for comparing related results obtained from analyzing multiple data sets or from analyzing a single data set with regard to different metadata parameters. Chord diagrams are a circular graph in which segments of the arc represent different analysis results, and the chords connecting them represent shared characteristics (e.g., differentially expressed genes). An example chord diagram is shown in **Figure 6**. By clicking the arc of a particular result, users can visualize its differentially expressed genes that also appear in any other results. NetworkAnalyst further supports computing genes that are unique to a particular result, or genes that are shared among any selected results. These gene lists then become available for functional enrichment analysis.

Heatmaps

A heatmap is a way of visualizing a table of numbers, in which carefully chosen color gradients substitute for numerical variations to facilitate pattern discovery by visual inspection. In gene expression heatmaps, rows represent genes, columns represent samples and cell colors correspond to gene expression abundances. An example heatmap is shown in **Figure 7**. Users can easily view gene expression changes across samples, conditions and/or studies. Heatmaps are usually coupled with clustering approaches to further enhance support for pattern discovery. The heatmaps implemented in NetworkAnalyst allow flexible sorting according to experimental conditions or metadata values, as well as clustering based on various distance measures between samples or genes. Users can directly drag to select a region of interest for zoom-in visualization and functional analysis. Conversely, they can perform enrichment analysis and then click the name of an enriched function to view the expression patterns of the contributing genes. Visualization and analytics are seamlessly integrated here to maximize data understanding.

or biological conditions. Usually, FCs are also provided to indicate the directions and magnitudes of change. Researchers are most interested in understanding the individual and the collective functions of these genes, as well as the relationships among them, particularly any nonobvious relationships that will provide new biological insights. The most widely used approach for this task is functional enrichment analysis and network analysis. The basic concepts for conducting network analysis are summarized in **Box 2**. These approaches have been seamlessly integrated into the network visual analytics system in NetworkAnalyst. The system is ideal for exploring networks with hundreds to thousands of nodes. During analysis, NetworkAnalyst will build networks from user-supplied genes of interest. Several approaches have been implemented to allow users to further customize the networks to control their sizes and complexities (as accessed from the Network Builder page; **Fig. 2**). The network analysis is then moved to a visual analytics environment for more in-depth exploratory analysis.

Using chord diagrams and heatmaps for meta-analysis of single gene expression data (Steps 17–31)

Visualization is an essential approach to data integration and understanding, as it provides the biologist with the ability to discern patterns created through high-end statistical analyses and through the use of enhancements such as color coding. Over the past decade, many different visualization methods have been explored in the bioinformatics community. Some techniques have proven to be very effective in displaying a large amount of contextual information to address particular needs, such as defining the relationship between dysregulated nodes. **Box 3** describes the three main approaches that are commonly used for visualization and integration of gene expression data. The previous section describes how to visualize genes of interest within the context of biological networks. In this section, we describe two other complementary approaches, namely using chord diagrams for comparing the similarities and differences between different analysis results and using heatmaps for examining the expression patterns across different experimental conditions and phenotypes. We demonstrate the main features of these techniques using a single gene expression data set associated with two metadata parameters. Metadata are variables that are potentially important to the interpretation of the results. They include obvious aspects such as control versus treatment used in challenge studies, but when applied to a complex data set (e.g., humans with a specific disease) they can include demographic information about sex, weight,

Box 4 | Statistical meta-analysis

It is usually not advisable to directly combine or compare the gene expression values from different gene expression data sets owing to their inherent heterogeneity (i.e., owing to different platforms, protocols and so on). Instead, integration is performed on the summary-level data such as *P* values, effect sizes or gene ranks and so on, as defined below. Many sophisticated algorithms have been described in recent years[6]. Some well-established procedures have been implemented in NetworkAnalyst.

Combining *P* values

This is a very simple approach in which *P* values are log-transformed and summed together. Larger scores reflect greater overall differential expression. Fisher's method and Stouffer's method are two popular approaches for combining *P* values. Stouffer's approach weights *P* values based on the number of samples in the contributing experiment and their fraction of the entire sample collection, whereas Fisher's method is a weight-free method. When all studies are of similar quality, Stouffer's method is more appropriate.

Combining effect sizes

Effect size is the difference between two group means after normalization by the s.d. for the data set (Cohen's *d*). There are two popular methods: the fixed effects model (FEM) and the random effects model (REM). The FEM assumes that there is a true effect size constant across experiments, whereas the REM treats the effect size as a random variable to accommodate heterogeneity between experiments. Cochran's Q test is often used to evaluate the homogeneity of the data sets. It is the weighted sum of the squared differences between individual study effects with the effects pooled across studies. NetworkAnalyst implements a Q-Q plot to help users choose the appropriate model. When the estimated *Q* values have an approximately chi-squared distribution, the FEM assumption is most appropriate; otherwise, REM should be used.

Combining rank orders

This is a nonparametric approach based on ranks of FCs. In this approach, the FC ratios are computed for all possible pair-wise comparisons for each data set. The ranks of the ratios are then used to calculate the rank product for each gene. Permutation tests are then performed to assess the null distributions of the rank products within each data set. The whole process is repeated multiple times to compute a *P* value and false discovery rate associated with each gene. It is a very computationally intensive procedure, and it requires powerful computers for a very large data sets.

Other procedures

NetworkAnalyst also supports two other approaches for data integration. The 'vote counting' approach is based on the total number of times a gene is considered as differentially expressed during individual data analyses. The 'direct merging' approach uses the combined gene expression data instead of summary-level data for DEA. Both approaches are statistically inefficient or flawed. They are intended for comparison purposes, not as the primary approach for meta-analysis.

height, body mass, age and so on, as well as clinical information such as diagnosis, temperature, heart rate, treatment, or underlying variables such as genome-wide association data.

Meta-analysis and visualization of multiple gene expression data sets (Steps 32–50)

The fundamental prerequisite for meta-analysis is that all of the included studies have been performed under comparable conditions and/or were testing the same underlying hypothesis. Several general steps are involved in conducting gene expression meta-analysis: (i) data collection involving identification of suitable studies and acquisition of data sets (this can include data from the literature or gene expression repositories); (ii) data cleaning involving quality control and annotation of each data set (including mapping different probe IDs to common gene IDs, and ensuring that metadata or phenotypic labels are accurate and consistent); (iii) data analysis including DEA for individual data sets to compute summary-level statistics for each gene; (iv) performing meta-analysis using suitable statistical methods; and (v) interpreting the results. Detailed reviews and discussions of various issues and approaches to meta-analysis of gene expression data are available[6,39]. NetworkAnalyst has incorporated these steps and good practices to allow users to perform meta-analysis on their collection of gene expression data sets. The main characteristics of several common meta-analysis approaches are discussed in **Box 4**.

MATERIALS

EQUIPMENT

Personal computer (including laptop) with an Internet connection

- Browser requirements: NetworkAnalyst has been tested on major modern web browsers that support HTML5 and JavaScript. We strongly recommend the latest Google Chrome (v39+) for best performance during visual analytics
- Hardware requirements: This is dependent on the size of the data. We recommend a ≥2 GHz CPU (Intel Core i5/i7 or equivalent), 4 GB physical RAM with at least 2 GB free and a minimum of a 15-inch screen with a screen resolution of 1,280 × 800 or higher. A mouse with scrolling support is required for visualization

Data files

- NetworkAnalyst has a number of example data sets for testing purposes. In the data upload page for each module, users can directly select a test data set from the dialog box by clicking the 'Try Examples' button. Alternatively, users can download the examples from the Data Formats page. To do this, first go to the NetworkAnalyst home page (http://www.networkanalyst.ca), and then click the 'Data Format' link on the top menu bar to enter the

Data Formats page. Under the 'Example Datasets', right-click each item and choose 'save link as …' to save these files into a folder with the default names. In particular, the three example data sets used in this protocol for illustrating different procedures are described in Equipment Setup

EQUIPMENT SETUP

endotoxin_list.txt (1 on the Example Datasets page) This is a gene list example data set that contains 99 Entrez gene IDs together with their FC values. These genes and their fold changes comprise the endotoxin tolerance signature that predicts endotoxin tolerance, as identified in a recent study[2]. We want to explore these genes with the context of PPI networks to generate hypotheses regarding the inter-relationships and functions of these signature genes.

endotoxin_data.txt (4 on the Example Datasets page) This data set is from a microarray gene expression study designed to measure gene expression changes in human PBMCs collected from four healthy donors[38]. The platform is Illumina's BeadArray, with genes annotated in RefSeq ID. There are three experimental conditions (Treatment): control, LPS (pro-inflammatory) and LPS-LPS, indicating two doses of LPS treatments within a day (leading to endotoxin tolerance). The two metadata parameters under consideration are Treatment and Donor. We want to explore the influence of the donor effect (basically human genetic variability) on gene expression changes caused by experimental conditions.

colon_cancer.zip (5 on the Example Datasets page) The zip file contains three microarray gene expression data sets—dataset1.txt, dataset2.txt and dataset3.txt. They contain 500 probes that are randomly selected from three GEO data sets—GSE13067, GSE13294 and GSE4554—from three independent studies on expression patterns of primary colorectal cancer. The platform is Affymetrix Human Genome U133 Plus 2.0 array. As we only use a small proportion of the data for illustrating meta-analysis procedures, biological interpretation should be treated with caution.

PROCEDURE

Network analysis and visualization for a list of genes

1| *Starting up*. Go to the NetworkAnalyst home page (http://www.networkanalyst.ca) and click the 'click here to start' link to enter the modules overview page.

? TROUBLESHOOTING

2| Locate the 'Starting with gene or protein lists' and click the 'Proceed' button within the panel to enter the data upload page.

3| *Data upload*. The required format is a list of gene names with optional FC values. We will use the example gene list data file 'endotoxin_list.txt' described in Equipment Setup. Specify 'H. sapiens (human)' as the organism type and set ID type to 'Entrez ID'. Leave the Data Label as the default.

4| Open the downloaded file 'endotoxin_list.txt' in any text editor, e.g., Notepad; select all, copy and paste the content into the text area of NetworkAnalyst. The gene list is entered either as a single column of gene IDs or in a two-column format with each line containing a gene ID and its FC value separated by a space or a tab.

5| Click the 'Upload' button. A message will appear in the top right corner summarizing the ID mapping results. If successful, the data name (Data1) will appear on the 'Data Uploaded' panel on the right, signaling that it is ready for the next step.

▲ **CRITICAL STEP** The user must specify the correct parameters in order to proceed. The procedure expects the majority (>50%) of IDs to be matched. Duplicated entries will be replaced by their averages.

6| (Optional) Users may upload multiple gene lists (i.e., differentially expressed genes from a similar study) at this stage by repeating Steps 3–5. NetworkAnalyst allows gene lists to be compared using network analysis or chord diagrams in subsequent analyses. When you upload multiple gene lists, it is better to give a meaningful name to your data, such as 'endotoxin_data'. Otherwise, the system will give default names in the form of Data1, Data2 … and so on.

7| Make sure that the 'Data1' checkbox is checked in the 'Data Uploaded' panel and click the 'Proceed' button at the bottom of the page to enter the next page. If you have uploaded multiple data sets, you can choose which one(s) you would like to analyze by selecting or deselecting their checkboxes. The current data type 'Single Gene List' is highlighted in orange, and only 'Network Analysis' is applicable for this data type.

■ **PAUSE POINT** You can save a session file from this page by clicking the 'Save Current Session' button. This is usually unnecessary for simple gene list data.

8| *Network creation*. Click the 'Proceed' button in the corresponding row. A dialog box will pop up that shows the available PPI database(s) and the available data generated from previous analysis. Make sure that the 'InnateDB Interactome' and 'Data1' are selected, and click the 'OK' button to proceed to the next page.

9| The 'Network Builder' page (**Fig. 2**) is the place where networks are created and customized. On the left panel, five menu items are available. Hover the mouse cursor over the help icon beside the 'Network Tools' text to view the introductions to

each function. The basic idea behind each function is described in **Box 2**. The main panel at the top summarizes the overall statistics obtained during network construction. Users can download these files by clicking the corresponding 'Download' link. The bottom panel shows a list of networks that have been identified. Accessing the first-order interaction network will typically return one large network ('continent') comprising most differentially expressed genes (seeds) and their interacting neighbors, with several smaller ones ('islands'). Most subsequent analyses are performed on the continent.
▲ **CRITICAL STEP** We recommend that users adjust the number of nodes so that they are in the range of 200–2,000 for practical (visual, biological and computational) reasons, as larger networks will lead to a 'hairball' effect, which rarely produces any informative outcome, whereas smaller networks will not enable a systems-level understanding.

10| (Optional) Experiment with the use of different functions to see the results. When each function is completed, make sure to click the 'Reset' button to return to the default state.

11| The default first-order network in the example data set contains ~1,000 nodes, which is a good size for visual analytics. Accept the default first-order interaction network and click the 'Proceed' button.

12| *Visual exploration of the network*. Depending on the speed of your Internet connection, it may take up to 30 s to load the Network Viewer page. The network viewer (**Fig. 3**) is on the center of the page showing the largest network, and it is surrounded by various supporting tables and tool menus. Mouse events are defined in **Table 2**. Experiment with these functions using the current network, using option A to zoom, option B to select or highlight nodes, and to drag and drop nodes, or option C to extract the highlighted nodes.

(A) Zoom

(i) Zoom in and out on the network with the mouse scroll function; move the network by clicking on an empty area within the network and dragging it to a new position; click the 'Auto fit' icon from the left tool bar to fit the entire network into the current window size.

(B) Node selection/highlighting, and dragging and dropping nodes

(i) In NetworkAnalyst, node selection is the same as highlighting. Click on the color palette that is located on the top left corner of the viewer. Select a color (e.g., light blue), and then click the 'Choose' button. The color palette now shows the new color. Point the mouse cursor over a node. When its label becomes visible, double-click on a node (i.e., transglutaminase 2 (TGM2)). The node is colored with the new color. If you change the Scope option on the top menu bar to 'Node and dependents', both the node and any of its interacting nodes will be highlighted.

(ii) Change the Scope option to 'Current node.' Point the mouse cursor over a node. When its label becomes visible, click and drag the node to a new position. Change the 'Scope' option to 'Node and dependents', and then drag a node with many connections (e.g., TGM2) to a new position.

(iii) NetworkAnalyst has built-in node selection/highlighting functions for results displayed in all major panels including Node Explorer, Function Explorer, Module Explorer and Path Explorer. Users can simply click an item of interest in the result panel to highlight the corresponding nodes in the network. To manually reposition those nodes, make sure that the Scope option is set to 'Highlighted nodes', and then drag any highlighted node to a new position.

(C) Extract the highlighted nodes as a connected subnetwork

(i) Click the Reset icon on the left to return to the default view. Make sure that the Scope option is set to 'Node and dependents;' highlight the TGM2 node and its interactors. Click the 'Extract' icon on the tool panel on the left. After ~20 s, the TGM2 and its interacting nodes are now extracted as a new subnetwork module. Note that the TGM2 node is not deleted from the original network. To do so, you need to use the Node Explorer (Step 7).

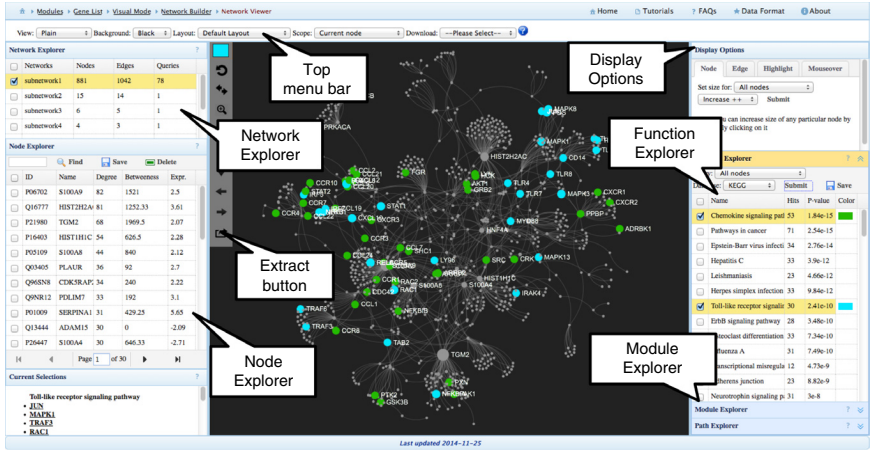

Figure 3 | Organization of components in Network Viewer. A screenshot of the 'Network Viewer' illustrates the main organization of various components. The current network is displayed in 'plain view', with two enriched pathways (chemokine signaling pathway and Toll-like receptor signaling pathway) highlighted in different colors. Clicking on the 'Extract' icon will extract a module that shows the interactions among all those highlighted nodes. The screenshot is generated in Steps 15–17 using example data set 1.

TABLE 2 | Mouse actions during network visualization and customization.

Purpose		Mouse event
Single node[a]	Display node name	Mouse over the node
	Show more details	Single-click the node
	Change color	Set the highlight color, and then double-click the node
	Increase size	Repeatedly click the node
	Change position	Mouse over the node until its label shows up, and then drag the node
Node and its dependents[b]	Change positions	Mouse over the central node until its label appears, and then drag the node
	Change colors	Set the highlight color, and then double-click the central node
	Change sizes	Go to Node tab under Display Options and change the node size for 'highlighted nodes'
Highlighted nodes[c]	Change positions	Mouse over any highlight node until its label appears, and then drag the node
	Change colors	Set the highlight color and then re-perform highlighting
	Change sizes	Go to Node tab under Display Options and change the node size for 'highlighted nodes'
Network	Zoom	Mouse over an empty area, and then scroll
	Change positions	Mouse over an empty area, and then drag
	Change sizes	Go to Node tab under Display Options and change the node size for 'all nodes'

[a]Users need to set the Scope option to 'Current node'. [b]Users need to set the Scope option to 'Node and dependents'. The dependent nodes are those leaf nodes connected only to the central node, without any other connections. [c]Users need to set the Scope option to 'Current highlights'. Node highlighting is described in Steps 12B(iii), 16 and 23.

(ii) (Optional) The extracted node cluster is now listed as 'module0' in the 'Network Explorer' table on the top left panel together with other available networks identified from the previous step. Users can click to load another network to the viewer. When the steps are complete, make sure to click 'subnetwork1' in the Network Explorer table to reload it to the current view.
? TROUBLESHOOTING

13| *Customizing the network*. The top menu bar under the history track (**Fig. 3**) provides tools to control the overall style of the network. The 'Display Options' on the top right panel (**Fig. 3**) provides more functions to allow finer control over the styles for nodes and edges. Experiment with these functions using the current network, using option A to change view, option B to change background, option C to change layout, option D to change node styles or option E to change edge style.

(A) Change view

(i) The default view is 'Topology', which uses color gradients to emphasize hub nodes. Change the option to 'Expression view' to see the expression patterns of the seed nodes within the network. The interacting non-differentially expressed nodes are colored gray.

(B) Change background

(i) Switch to the 'Topology view' and change the current background to 'white'. This is usually preferred for publication or presentation. The default black background provides better contrast during interactive visual exploration.

(C) Change layout

(i) Click the layout dropdown menu and choose the 'ForceAtlas' option to change the layout. In the updated view (**Fig. 4a**), hub nodes are moved toward the central area of the network window to form a tightly interconnected 'core' of the network. Click to apply the 'YuFan Fu' layout—observe that all hub nodes are 'spun out' to produce a less clustered view. In general, the ForceAtlas layout gives a more unbiased view of the relationships between nodes in that those that are more connected will be closer together. However, this will produce an indistinct, tight cluster for large networks. The default layout uses a combined approach based on ForceAtlas and YuFan Fu, which usually provides a decent arrangement. Note that the layout procedures usually use some random values to initialize the node positions and then iteratively improve upon the initial layout. The result may be slightly different each time.

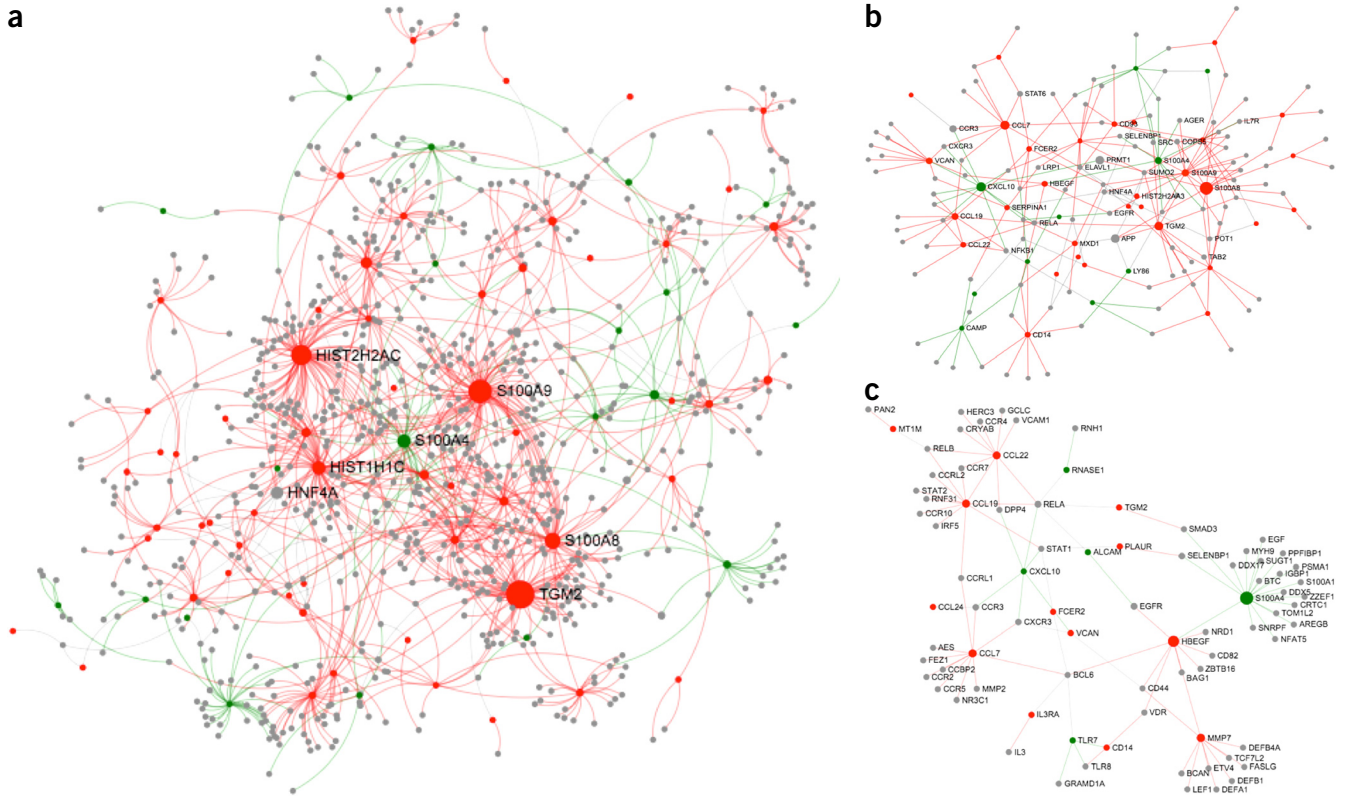

Figure 4 | Network customization and module extraction. (**a**) The same network as in **Figure 3**, but using the ForceAtlas layout (Step 13C) with Expression view (red for upregulated and green for downregulated genes). In this layout, it is very easy to see the core of the subnetwork formed by a few key hub nodes. (**b**) The innate immune module constructed using the function-first approach (Steps 15–19). (**c**) The module identified by the connection-first approach that is enriched for genes involved in innate immunity (Steps 21–28).

(D) Change node styles

(i) Click the 'Node' tab under 'Display Options', and then click the 'Submit' button with the default parameters to increase the size of all nodes in the network. Node labels will appear automatically after node sizes reach a certain threshold. Alternatively, you can choose to adjust the sizes only for highlighted nodes. These two approaches are usually used together (e.g., first apply a new color and then increase the node sizes) to achieve better highlighting effects.

(E) Change the edge style

(i) The edges are interconnecting lines that reflect known interactions between nodes. Both the edge shape and edge thickness can be adjusted using functions within the 'Edge' tab under 'Display Options'. Curved edges are better for visualization of large networks (**Fig. 4a**). Change the edge shape to 'Line' and edge width to 'Thick', and then click Submit. These settings usually work better for smaller networks (**Fig. 4b**,**c**).

14| *Node visualization and manipulation*. The 'Node Explorer' table in the middle-left panel (**Fig. 3**) shows all the nodes in the current network. These nodes are identified by their IDs and name, together with hub degree and betweenness values. For seed nodes, their FC values are given under the 'Expr' column. Users can sort the node table by clicking on column headers. Manipulate and explore the nodes using option A to view nodes, option B to search nodes, option C to highlight hub nodes or option D to delete nodes.

(A) Node viewing

(i) Click a node name (i.e., SA1004-S100 calcium-binding protein A4) to view it in the network.

(B) Node search

(i) Enter a node name (e.g., CCL7) and click 'Find' to pinpoint it in the network.

(C) Hub node highlighting

(i) Sort the nodes by their degree values in descending order. Use the checkbox to select all nodes with a degree value >30. Click the 'Auto fit' icon to view all of the highlighted nodes.

(D) Node deletion

(i) This is a form of *in silico* deletion mutation that can be used to remove promiscuous proteins that might skew the results. In this case, we want to remove the UBC node (ubiquitin C, ID: P0CG48), which is known to be involved in many nonspecific interactions. After resetting to the default view, select the UBC node, and then click the 'Delete' button on the tool bar of the node table.

! CAUTION Network deletion is a computationally expensive task. It involves not only the node itself but also the nodes that it connects with. Nodes that connected only to the deleted node will be removed from the resulting network. Deleting an important hub node may also break the current network into several pieces. After node deletion, NetworkAnalyst will perform network construction on the new node list, update the network layout and recompute the centrality values for the remaining nodes.

15| *Identification and extraction of functional modules (function-first approach)*. This process (Steps 15–18) allows the identification of important functional modules by using enrichment analysis. The 'Function Explorer' on the middle right panel (**Fig. 3**) supports functional enrichment analysis based on gene ontology (GO) terms or pathways according to the Kyoto Encyclopedia of Genes and Genomes (KEGG) or Reactome databases. Click the 'Submit' button to perform enrichment analysis on all nodes using KEGG as the source for annotated pathways. The resulting table shows a list of the pathways that are enriched. The top pathway is 'Chemokine signaling pathway' with 53 hits within the network.

▲ **CRITICAL STEP** The pathways are ranked by their raw *P* values from over-representation analysis (ORA) based on hypergeometric tests. Note that owing to the overlap and interdependence of the genes within the pathways or GO categories, multiple testing adjustments such as false discovery rate (FDR) become inappropriate, and the raw *P* values are only used for ranking the results.

16| Use the color palette to set a new highlight color (i.e., green) as per Step 12B(i). Select the 'Chemokine signaling pathway'. The corresponding nodes will be highlighted in green and made larger within the current network (**Fig. 3**).

17| Set a new highlight color (i.e., light blue) and then select 'Toll-like receptor signaling pathway', which is another significant pathway that is involved in innate immunity (**Fig. 3**).

18| Now, to extract a module that will contain all nodes that are related to these two pathways, click the 'Extract' button on the viewer's tool bar (**Fig. 3**). This process may take ~20 s to finish. This module is now listed as 'module1' in the 'Network Explorer' panel (**Fig 3**). It contains 154 nodes and among them are our 41 differentially expressed genes (queries). This is much more concentrated for differentially expressed genes compared with the parent network containing 881 nodes and 78 queries. This then provides us with new biological insights regarding the endotoxin tolerance signature and its likely relationships with innate immune signaling. We can perform enrichment analysis on this new module by repeating Step 15 to confirm that the two pathways are indeed significantly enriched in this module. Please note that the total number of nodes may be slightly different owing to certain randomness, as well as the dynamic nature of module extraction.

19| *Module display*. The module is displayed using a default style. We can use the previously described functions (Step 13) to perform further customization. Given the smaller size of this subnetwork, it is very amenable to manually dragging the node(s) to new positions for discovery and highlight purposes. **Figure 4b** shows an example of the module produced using the automatic layout followed by manual adjustment.

▲ **CRITICAL STEP** The order of these steps is important. Make sure that manual layout is performed after automatic layout. The server is not aware of the new node positions on the users' browser. Re-applying the automatic layout later will overwrite any previous manual efforts. For this reason, it is essential to save any interesting results after manual adjustment.

20| *Save the results*. Use the Download menu to save the current module as a portable network graphics (PNG) image, and name it 'innate_immunity_module.png'. The network is also available in the scalable vector graphics (SVG) or GraphML format. The latter format can be viewed in other popular network visualization tools such as Cytoscape. You can also download the node property tables and the enrichment analysis results by clicking the Save icons at the top of the corresponding tables.

▲ **CRITICAL STEP** Note that the node property table and enrichment analysis results are generated dynamically on the basis of the current network and selections. It is important to save the results at the time of the analysis and to give them meaningful names.

21| *Module detection and extraction (connection-first approach)*. Click on 'subnetwork1' in the 'Network Explorer' panel on the left (**Fig. 3**) to reload the main network into the network viewer.

22| Click 'Module Explorer' on the bottom right (**Fig. 3**) to bring up the panel, and then click the 'Perform Module Detection' button. A list of modules will be returned together with summary statistics about their sizes, *P* values and how many seed nodes they contain.

23| Click any module to view its nodes highlighted in the current network. You can set new colors for different modules. Some of the significant modules detected with this approach mainly reflect direct PPI. This is anticipated as we can clearly see this modular structure in the network (**Fig. 3**).

24| The *P* values are computed by the Mann-Whitney-Wilcoxon tests to compare the number of connections of each node with that of other nodes within the module and with that for nodes outside the module. We also need to consider the biological content (i.e., number of seeds and FCs) captured in each module. In this example, select the top four modules with *P* values of <0.05 and that contain at least three seed nodes.

25| Extract these four modules of interest by clicking the 'Extract' button in the network viewer.

26| Perform pathway enrichment analysis on the nodes within this extracted module using KEGG by repeating Step 15. It is reassuring that the innate immune pathways 'Chemokine signaling pathway' and 'Toll-like signaling pathway' are among the top pathways in the list. It is thus clear that this is a major theme in our uploaded gene list.

27| (Optional) At this stage, the expert biologist might like to explore the pathway list for the pathways that are not anticipated as a method for discovering new mechanisms underlying these data sets.

28| Further customize the module as described in Step 13, and save the results as a PNG or SVG image. An example output is shown in **Figure 4c**.

29| Download the node property table from the 'Node Explorer' on the middle left panel and the enrichment analysis result table from the 'Function Explorer' on the middle right panel and save them to files with suitable descriptive names.
▲ **CRITICAL STEP** It is easy to collect massive amounts of data, so we recommend that when saving results you should keep in mind the goal of developing new biological insights. Remember that the purpose of the above exercise was to obtain new biological insights regarding these signature genes and how they might be related to the development of sepsis and organ failure. Thus, results that shed light on the mechanisms and functions associated with this signature are important to save.

Meta-analysis of a gene expression data set with multiple metadata parameters
▲ **CRITICAL** The following procedures are described for microarray data sets, but NetworkAnalyst works equally well for count data from RNA-seq experiments. For RNA-seq data, users need to first annotate those platform-specific feature IDs by transforming them into common gene or transcript IDs (Entrez, RefSeq, Genbank or Ensembl) before uploading the count tables to NetworkAnalyst.

30| *Starting up*. Click the 'Home' icon to return to the home page, and then click on 'Click here to start analysis'. Locate 'Starting with a microarray or RNA-seq data' and click the 'Proceed' button within the panel to enter this module.
! **CAUTION** Although most browsers support multiple tabs, do not access NetworkAnalyst from more than one tab during the analysis. Opening up multiple connections to NetworkAnalyst within the same browser will cause unpredictable behavior of the application and the results.

31| *Data Upload*. The 'Data Analysis' page provides step-by-step procedures for data preparation for a single gene expression data set. Click the 'Browse' button to locate 'endotoxin_data.txt' file, and then click 'Submit' to upload the data. Alternatively, click the 'Try Examples' button at the bottom, and then choose the 'Endotoxin' data set. Depending on the speed of the Internet, the process may take a while. A message will appear indicating that the data have been successfully uploaded.
▲ **CRITICAL STEP** The required format is a tab-delimited text (.txt) file with genes or probes in rows and samples in columns. The sample names must be in the first line, followed by the metadata labels. Multiple metadata parameters can be provided (for example, treatment type, demographic information such as age or sex, response to treatment and so on). Each metadata parameter should start with a new line beginning with #CLASS:' (e.g., #CLASS:Treatment to indicate the Treatment metadata). Users can click on and visit the 'Data Format' page for more instructions. Large data sets can be uploaded as a .zip file.

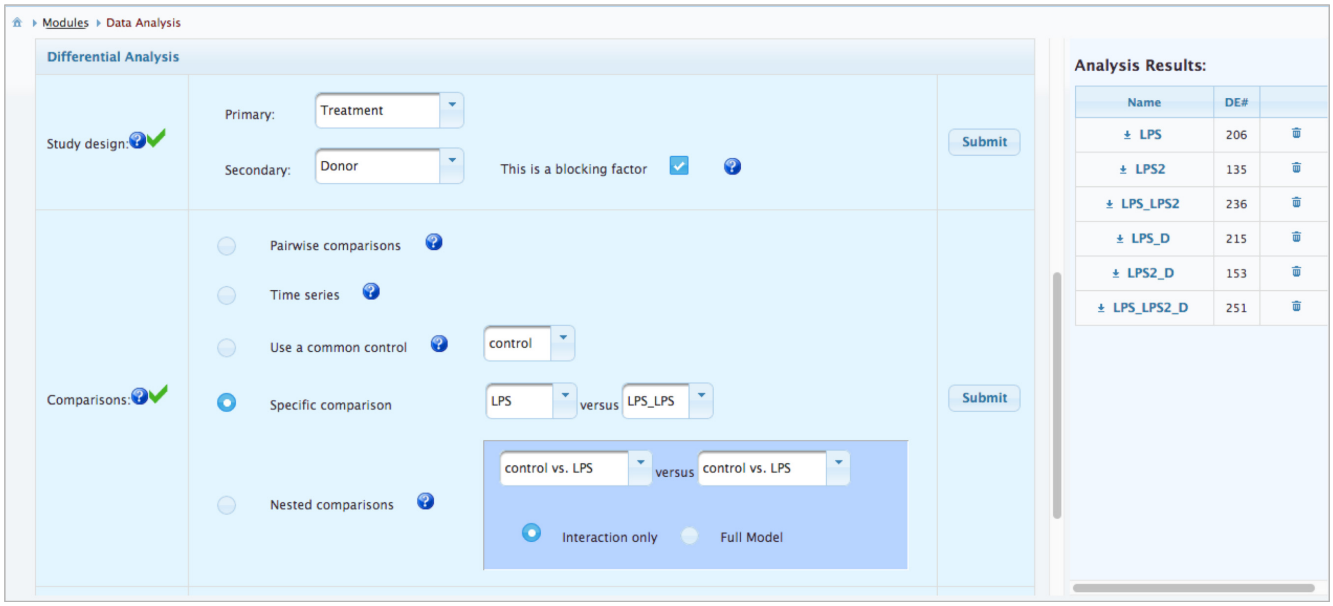

Figure 5 | Screenshot of the interface for differential expression analysis. NetworkAnalyst provides an interface that allows complex study designs and flexible comparisons. The results of each analysis (shown at the top right) will be available for visual analysis further downstream. The screenshot shows the result generated in Step 34 using example data set 2.

32| *Data set annotation*. Annotations of data sets are required in order to perform functional analysis in later stages. For the example data set, set Organism to 'H. sapiens (human)', 'Microarray data (intensities)' as the Data Type, 'RefSeq ID' as the ID type and 'Mean' for Gene-level Summarization. For RNA-seq data, 'Sum' is more correct for read counts. Click 'Submit' to perform the annotation. After ~20 s, a message will display indicating the status and summary of the annotation results.

33| *Data normalization*. The example data are already normalized, so keep the default option 'No normalization' and click 'Submit'. Users can use the 'View data' function to generate two useful summary plots. Box plots can be used to check the normalization status. If all data values are below 20, then the data are already on a log scale; if all samples have identical distributions, then the data are already quantile normalized. Principal component analysis (PCA) plots can be used for checking the overall data quality and discovering unusual patterns. For RNA-seq data, users should choose the 'log counts per million' or 'log counts per million followed by quantile normalization' method, which uses the 'voom' function to transform raw read counts to log counts per million with associated precision weights. After this transformation, the data can be analyzed using the same functions as would be used for microarray data.

34| *DEA*. NetworkAnalyst allows flexible comparisons for a variety of complex study designs (**Fig. 5**). For illustration purposes, we will compare the treatment effects (relative to control) of LPS (inflammatory) with those of double LPS (endotoxin tolerance inducing) without (option A) and with (option B) considering any donor effect (effect of human genetic variation in individual donors that provided PBMCs for analysis). The approach is very basic: performing DEA separately (LPS versus control, LPS_LPS versus control, LPS versus LPS_LPS) and then visually comparing the results.

(A) Performing comparisons without considering the donor effect

(i) Under 'Study design', for metadata of interest, select 'Treatment' as the primary (condition of interest) and leave the secondary as 'Not Available'. Click 'Submit'.

(ii) Under 'Comparisons', choose 'Specific comparison' and make sure 'control' versus 'LPS' are selected. Click the 'Submit' button. The process may take ~10 s to complete.

(iii) Set the cutoffs for differentially expressed gene selection. Set the adjusted *P* value cutoff to 0.05 and the $\log_2$ FC cutoff to 1.0, enter the result name 'LPS' and then click 'Submit'. The message indicates that 206 differentially expressed genes were identified.

(iv) Retain the same study design, and change the comparison to 'control' versus 'LPS_LPS'. Click 'Submit'.

(v) Use the same cutoff and 'LPS2' for the result name, and then click 'Submit'. The message indicates that 135 DE genes were identified.

(vi) Retain the same study design, and update the comparison to 'LPS' versus 'LPS_LPS'. Click 'Submit'.

(vii) Use the same cutoff and 'LPS_LPS2' for the result name, and then click 'Submit'. The message indicates that 236 differentially expressed genes were identified.

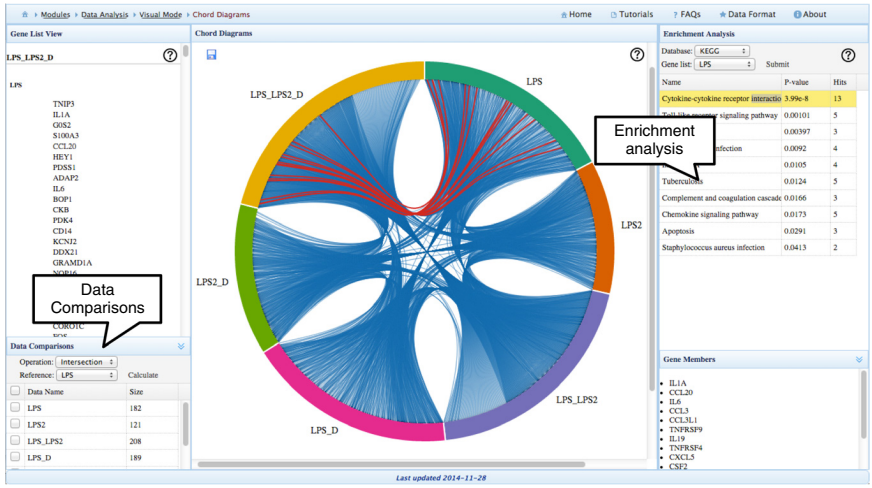

Figure 6 | A chord diagram comparing six analysis results. The chord diagram displays the shared differentially expressed genes (chords) among six related analysis results (arcs). Clicking on the arcs will display shared genes on the top left. The data comparison on the bottom left allows more advanced and flexible comparisons. The resulting gene lists can be used for enrichment analysis on the right. The screenshot is generated in Step 38 using example data set 2.

(B) Performing comparisons considering the donor effect

(i) Under 'Study design', for metadata of interest, select 'Treatment' as the primary condition of interest, and 'Donor' for the secondary. Make sure that 'this is a blocking factor' is checked because we do not want to analyze the donor effect as an independent experimental factor but to account for the effect of donor variability in the experimental results. Click the 'Submit' button. This function estimates the correlation between repeated observations on the blocking variable. The process may take ~1 min to complete.

(ii) Repeat Steps 34A(ii–vii), making sure that for each result you assign a unique name (e.g., add a suffix '_D').

35| When complete, the view should look similar to **Figure 5**. Detailed result tables are available on the right. We can clearly see that when considering the donor effect more differentially expressed genes are identified. This is because the model fits the study design better and leads to greater statistical power.

36| (Optional) Click each result name to download the corresponding analysis results for future reference.

37| Click the 'Proceed' button on the bottom of the page.

38| *Using chord diagrams to compare the results.* On the following page, locate the 'Chord diagrams' row and click the adjacent 'Proceed' button. A dialog box will appear asking which results you want to include. Keep all results from the model considering the donor effect (names that end with '_D') checked, and click 'OK'.

On the page that appears, the chord diagram is located at the center, surrounded by supporting tools and tables. An example screenshot is shown in **Figure 6**. The arcs of the diagram circle represent the individual results, and the chords connecting them represent their shared genes. Refer to **Box 3** for more details on how to interpret chord diagrams. Use the supporting tools and tables to explore the chord diagram, using option A to see an overview of shared genes, option B to obtain the shared gene lists, option C for enrichment analysis, option D to view enriched functions or option E to perform flexible comparisons.

(A) Overview of shared genes

(i) Mouse over the LPS_D arc. Those chords representing differentially expressed genes that are shared between LPS_D and any other data sets will be highlighted in dark green. By moving the mouse over different arcs, we can clearly see that LPS_LPS2_D contains most of the genes from the other two data sets. It also has a significant proportion of unique differentially expressed genes.

(B) Obtaining the shared gene lists

(i) Click the LPS_D arc to compare its differentially expressed genes with those from the other data sets. The result is shown in the left panel. Genes that are shared with other data sets are listed under the corresponding name of the data set. Genes that are unique to the LPS_D data set are listed under the LPS_D header. Clicking the name of a shared gene will highlight the corresponding chord on the diagram.

(C) Enrichment analysis

(i) The gene lists are available for enrichment analysis in the right panel. Choose 'LPS_LPS2_D', retain KEGG as the database and then click 'Submit.' The result shows the enriched pathways in the genes shared between LPS_D and LPS_LPS2_D.

(D) View enriched functions

(i) Click the top enriched pathway 'Cytokine-cytokine receptor interaction'. The corresponding chords will be highlighted on the chord diagram between the two arcs. In addition, all of the contributing genes are listed in the bottom right panel.

(E) Performing flexible comparisons using the tools on the bottom-left panel

(i) *Compute intersection*. To search for overlapping genes among all the data sets, set operation to 'Intersection'. Make sure all data sets are selected, and click 'Calculate'. The returned gene list contains the genes shared among all three data sets. Mouse over the gene names to view the highlighted chords on the diagram. Select the gene list, copy and paste it to a spreadsheet and save it as a text file ('Overlapped_genes.txt'). These are the core genes whose expressions change significantly in each stage of the treatments. These genes are potential biomarkers. Note that this gene list is also available for enrichment analysis under the name 'Current result.'

(ii) *Compute differences*. We can easily compare the differences between differentially expressed genes obtained with and without accounting for donor effect in our design model. Previously, we saw that when donor variability is considered we obtained a few more differentially expressed genes. We can further investigate that here. Click 'Visual Mode' from the history track to go to the previous page, and click the 'Proceed' button on the 'Chord Diagrams' row again. This time, we will keep all available results checked, and click 'OK'. In the new diagram (**Fig. 6**), select the 'LPS' and 'LPS_D' data sets from the 'Data Comparisons' panel on the bottom left. Choose 'Difference' as the operation, make sure that 'LPS' is set as the reference and then click 'Calculate'. The results pane shows 'null', which means that all of the genes in the LPS data set are also in the LPS_D data set (that LPS is a subset of LPS_D). Now, set LPS_D as the reference, and click 'Calculate'. A list of seven genes is returned. Select, copy and paste these into a spreadsheet and save the gene list as 'LPSD_LPS.txt'.

▲ **CRITICAL STEP** To calculate the unique genes in a particular list, you need to specify it as the 'Reference'. For other operations, this option is ignored.

39| *Using heatmaps to visualize expression profiles*. Click on 'Visual Mode' on the history track again to return to the previous page. Click the 'Proceed' button corresponding to the 'Interactive Heatmaps' row. A dialog box will appear asking which data set you would like to view. In this case, choose 'LPS_D' and click 'OK' to proceed. On the new page, there are three columns. The left one shows the heatmap with low resolution for overview purpose. The middle column is the Focus View, which displays the expression profile for genes of interest in higher resolution. The right column contains tools and tables for enrichment analysis. The top menu bar contains various supporting functions (**Fig. 7**). Use the supporting tools and tables to explore the heatmap, using option A to browse, option B to cluster genes, option C to cluster samples, option D to select a region of interest, option E for functional profiling, option F to create a custom signature or option G to build custom heatmaps.

(A) Browsing heatmaps

(i) Move the mouse over the heatmap. The panel in the top right corner will display gene annotation information corresponding to the gene to which the mouse is pointing, as well as the *P* values from the DEA. The two metadata labels (Treatment and Donor) are also provided at the top of the heatmap. Move the mouse cursor over each row to see the labels for different colors. The bottom of the middle Focus View shows sample labels for each column.

(B) Clustering genes

(i) By default, genes are ranked by their *P* values. Click the drop-down menu to select 'Euclidean distance' to cluster the genes on the basis of their Euclidean distance to each other. Note that genes will be reorganized in both the Overview and Focus View.

(C) Clustering samples

(i) The samples are ordered from left to right according to the Donor by default. Double-click on 'Treatment' row to sort the samples accordingly. Alternatively, users can use the 'Cluster samples' option on the top menu bar.

(D) Selecting a region of interest

(i) There are many distinctive patterns in the overview pane. For instance, the top one-third of the overview heatmap shows the cluster of genes that are suppressed after LPS treatment (as revealed by the dark blue color corresponding to the LPS group). Drag to select this region. The genes will show up in the central Focus View.

(E) Functional profiling

(i) To identify whether there is a primary functional theme among the selected differentially expressed genes, we can apply enrichment analysis. In the Enrichment Analysis pane on the right side, set the query to 'Genes in Focus View', choose KEGG and then click Submit. The results do not show any pathway that explains a large number of the genes in the Focus View. Set the database to 'GO:BP', and click 'Submit'. The results show some enriched functions containing over ten genes. The top one is 'inflammatory response', which contains 11 genes. Double-click the row to display only the 11 genes in the Focus View. Click 'Reset view' to go back to the default.

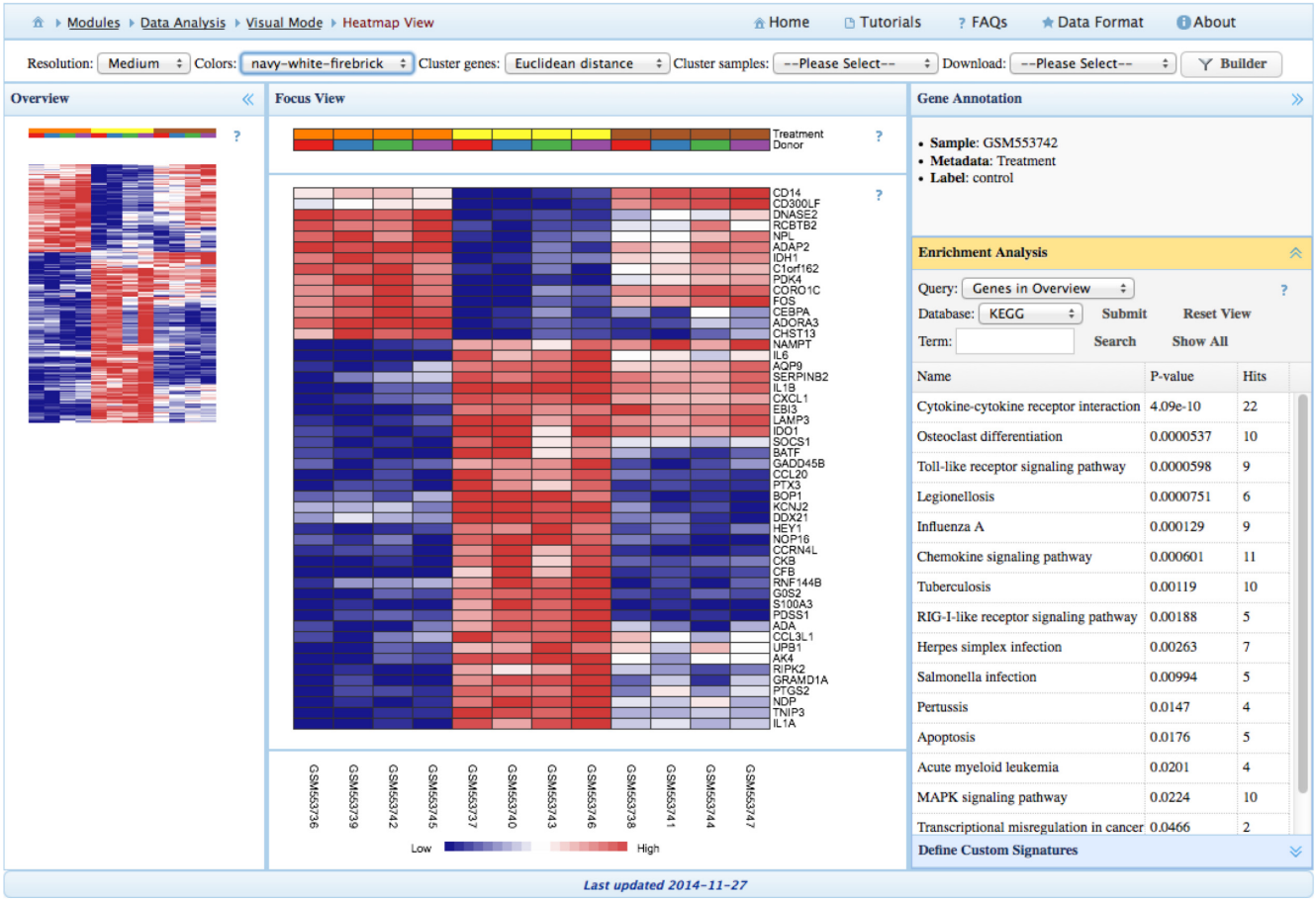

Figure 7 | Visual analytics with heatmaps. This screenshot shows the overall arrangement of the heatmap visualization interface. The Overview on the left-hand side displays the overall gene expression pattern. Users can drag to select genes from any region of interest. The Focus View in the center displays the current genes of interest. The Enrichment analysis panel on the right side allows users to perform enrichment analysis on genes displayed in either the Overview or Focus View panel. The toolbar at the top of the screen provides functions that are commonly used during heatmap visualization. This screenshot is generated in Step 39 using example data set 2.

(F) Creating a custom signature

(i) The panel on the bottom right allows you to easily create a custom heatmap from a given gene list. Click the 'Define Custom Signatures' tab to bring up the panel. Open the file 'Overlapped_genes.txt' that we saved before. Copy and paste the gene names into the text area. Click 'Submit'. The expression profile of these genes is now shown in the Focus View. These genes show distinctive patterns in each of the treatment stages.

(G) Building custom heatmaps

(i) This enables the creation of a heatmap showing the expression patterns of the differentially expressed genes involved in 'Cytokine-cytokine receptor interaction'. In the enrichment analysis panel to the right, set the database to KEGG and query to 'Genes in Overview'. Next, Click 'Submit'.

(ii) Double-click on the 'Cytokine-cytokine receptor interaction pathway' row. The corresponding 22 genes now appear in the Focus View.

(iii) Click the 'Builder' icon on the top menu bar. The Builder view will show up below the Focus View.

(iv) Drag and select all genes in the Focus View to copy these genes down to the Builder view. Within the Builder view, you can manually drag a row to a new position, double-click to delete a row or insert blank rows for further customization.

(v) Download the 'Custom heatmap' using the download option on the top menu bar.

40| (Optional) Click 'Visual Mode' to go back, and choose other data sets to explore their expression profiles using heatmaps.

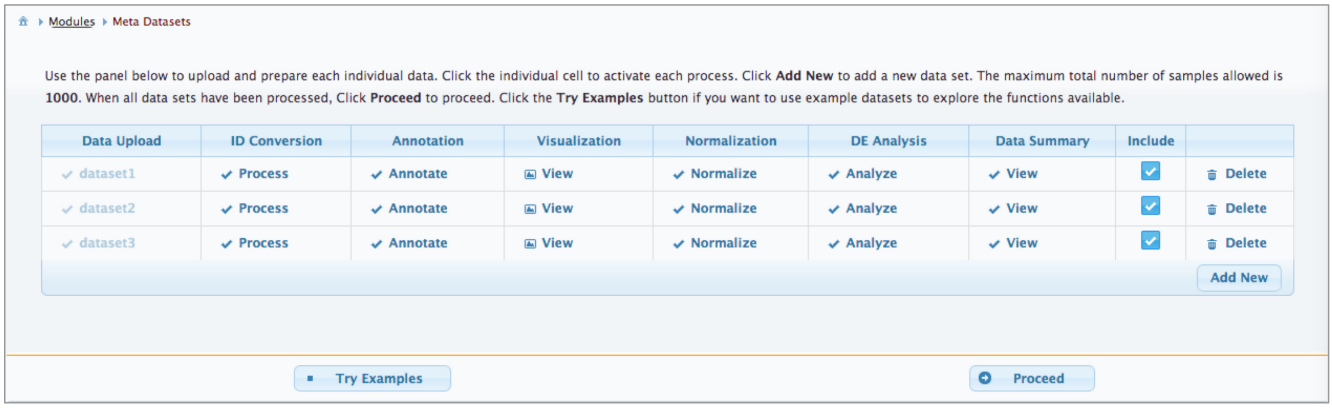

Figure 8 | Upload and process multiple data sets for meta-analysis. This screenshot shows the interface with a table-based navigation panel to allow users to upload and process multiple data sets for meta-analysis, with table columns corresponding to data processing procedures and table rows corresponding to data sets. Clicking on a table cell will trigger a dialog box to guide users through each step. This screenshot is generated in Step 53 using example data set 3.

Meta-analysis of multiple gene expression data sets

41| *Starting up*. Click the 'Home' icon to return to the home page, and then click 'Click here to start analysis'. Locate 'Multiple gene expression datasets' and click the 'Proceed' button within the panel.

42| *Data preparation*. The data preparation page allows users to upload and prepare multiple data sets. NetworkAnalyst provides a table-based navigation approach with columns corresponding to data processing steps including Data Upload, ID Conversion, Annotation, Visualization, Normalization, DE Analysis and Data Summary (**Fig. 8**). Adding a new row to the table will allow the user to upload a new data set. Clicking on a cell within each row will bring up a dialog box that will guide users to complete the processing steps on the corresponding data set. The procedures for data preparation on individual data sets are almost identical to the Steps 31–34. The main difference is that, when dealing with multiple data sets, the DEA is restricted to two-group comparisons for a single metadata parameter. This is because the results will become extremely difficult to interpret when multiple metadata parameters and multiple data sets are considered simultaneously. More importantly, when combining more data, the bias (such as the donor effect owing to small sample size) will be reduced or eliminated. Unzip the 'colon_cancer.zip' file downloaded during the Equipment Setup. There are three text files—dataset1.txt, dataset2.txt and dataset3.txt. In this example, we will only show how to process one of these data sets, and then use the three example data sets that have already been prepared, as described in Steps 43–51.

43| Click the 'Upload' cell in the first row; in the data upload dialog box, click 'Choose File' to locate the first example file for meta-analysis 'dataset1.txt' and click 'Submit'.

44| When the data are uploaded, click the 'Process' cell under the 'Data Conversion' header. In the new dialog box, select 'H. sapiens (human)' as the organism type and 'Affymetrix Human Genome U133plus2 (hgu133plus2)' as the platform. Click 'Process' to perform ID conversion from probe ID to Entrez gene ID. Click 'Done' to close the dialog box when the conversion is complete.

45| Click the 'Annotate' cell to bring up the dialog box. Users can edit metadata labels to make sure that they are consistent across different data sets. In this case, click 'Submit' to accept the default.

46| Click the 'View' cell under the 'Visualization' header to see summary plots for this data set. From the box plots, we can see that the data are already on a log scale. From the PCA plots, we can see that all except two samples are clustered together.

47| Click the 'Normalize' cell under the 'Normalization' header to bring up the dialog box. Set the data type to 'Microarray data (intensities)' and choose 'quantile normalization only', and then click 'Submit'.

48| (Optional) Click 'View' cell again to check the normalization results. We can see that all samples now have identical distributions in the box plot and that all samples are well clustered together in the PCA plot.

49| Click the 'Analyze' cell under the 'DE Analysis' header. The dialog box allows users to perform two-group DEA within a single metadata parameter. As our data only contain two groups and a single metadata parameter, click 'Submit' to use the default values. The result shows 14 differentially expressed genes and 372 non–differentially expressed genes based on the 0.05 *P* value cutoff. Click 'Done' to close the dialog box.

50| Click 'View' cell under the 'Data Summary' header to see a summary of processed data so far.
▲ **CRITICAL STEP** This dialog contains an option to 'Set conditions in meta-analysis.' When there are multiple groups present in the data, users must specify which two groups will be compared in the meta-analysis. Users must also make sure that these two groups are labeled consistently across all data sets included.

51| Click 'Done' to close the dialog. The first data set is now ready.

52| (Optional) Repeat Steps 43–51 to prepare dataset2.txt and dataset3.txt.

53| Instead of executing Step 52, however, we can use the preloaded example data sets. Click 'Try Examples' to bring up the dialog, and click 'Yes' to perform the default analysis procedures. The process will take ~30 s. A message will appear indicating that the three data sets were processed successfully. A corresponding screenshot is shown in **Figure 8**.

54| Click the 'Proceed' button on the bottom of the page. A dialog box will show up indicating that the uploaded data sets have passed the integrity check. Click 'Next'.
▲ **CRITICAL STEP** During the integrity check, NetworkAnalyst will inspect each individual data set to make sure that the metadata parameters are consistent across all data sets and a significant number of overlapping genes can be identified.
? TROUBLESHOOTING

55| *Statistical meta-analysis.* The meta-analysis page shows five common approaches for meta-analysis. Please refer to **Box 4** for more details about these methods. In this case, we will choose the widely used *P* value combination method. As all data sets are from the same platform with no missing values, we can use Stouffer's method to give more weight to *P* values based on a larger sample size. Keep the *P* value cutoff at the default (0.05). Click 'Next'.

56| (Optional) When data sets are from different platforms or of different qualities, Fisher's method for *P* value combination should be preferred. The effect size combination is also a valid option in this case. When data sets are very heterogeneous, the nonparametric method (i.e., rank-based combination) should also be considered. Users need to be aware that nonparametric approaches generally have less statistical power.
? TROUBLESHOOTING

57| The differentially expressed genes identified from the meta-analysis are displayed in a table on the new page. You can click the 'View' button (right-hand column) for genes of interest to get a box plot summary of the gene expression patterns within each individual data set.

58| Click the 'Visual Exploration' button at the bottom of the page. On the 'Visual Mode' page, all three visual analytics tools are enabled. Before we proceed, it is important to save a copy of the current session file. Click 'Save Current Session' to download the session file.
! CAUTION The session file contains only the information that is necessary for visual analytics. It is assumed that the data preparation and meta-analysis stages are complete. Thus, when resuming from the saved session file, users will not be able to return to the previous data preparation steps. The session file is stored in a special format for NetworkAnalyst to understand. Users should not modify this file.

59| *Subnetwork analysis.* On the 'Visual Mode' page, click the 'Proceed' button in the Network Analysis row. In the popup dialog box, make sure that the 'InnateDB Interactome' and 'meta_dat' are selected. Click 'OK' to enter the Network Builder page.

60| The top network created by default procedures contains 1,665 nodes and 2,252 edges. Click 'Proceed' to view this network.

61| *Network overview and module analysis*. In the network visualization page, we can see a very dense network displayed in the center. Direct visual exploration is thus not a good idea, as it will yield a 'hairball' of overlapping edges and nodes in which it is very difficult to visualize any individual features or interactions. At this stage, we can perform network or module analysis using either a connection-first approach (option A) or a function-first approach (option B) to focus on small modules of interest.

(A) Connection-first approach

(i) Perform a connection-first approach for module analysis, as described in Steps 21–29. The top module is very significant (*P* value, 2.4×10^{-14}), and it contains 23 seed nodes. Extract the top module and perform enrichment analysis using KEGG. We can clearly see that this module contains many genes involved in various cancer pathways.

(B) Function-first approach

(i) Reload subnetwork1 and apply the function-first module analysis described in Steps 15–20 to create a module consisting of genes that are involved in cancer pathways. The above two modules are of similar size and biological contents, suggesting that the main functional theme of this network is captured.

62| *Using 'Trim' to explore minimum interaction network*. This approach will reduce both the size and complexity of the resulting network, while still maintaining the connectivity among the seed genes (see **Box 2** for more details). Click 'Network Builder' to go back to the previous page, and click the 'Trim' button. The top subnetwork now contains 342 nodes and 790 edges, about one-third the size of the default network. Next, click 'Proceed'.

63| The new subnetwork is much more amenable to visual exploration. Now, use the 'Function Explorer' to perform KEGG pathway enrichment analysis on this trimmed network. Indeed, the 'Pathways in cancer' is at the top of the returned list of enriched pathways.

64| Customize the subnetwork as described in Step 13 and save the result as a PNG or SVG image.

65| *Visual analytics with a chord diagram*. Click 'Visual mode' to go back to the corresponding page. Click 'Proceed' in the Chord Diagrams row. You can choose any combination (at least two) of the available data sets for comparative visual analysis using a chord diagram. Keep all data sets checked and click 'OK'.

The chord diagram shows the overall pattern of differentially expressed genes identified from selected data sets and from the meta-analysis. Note to ensure that they are comparable; all differentially expressed genes are selected on the basis of the universal *P* value cutoff used during the meta-analysis. The *P* value cutoffs used during individual analyses are not applicable here.

66| Click on the arcs corresponding to any of the data sets to view their relationships.

67| Explore the enriched functions within a gene list of interest using the Enrichment Analysis panel, as described in Step 38C and Step 38D.

68| Use the functions in the bottom-left 'Data Comparisons' panel to calculate the differences between any data sets, or to compute the shared 'core' among all the data sets using the 'Intersection' option (Step 38E(i)).

69| *Visual analytics with heatmaps*. Click the 'Visual Mode' link on the history track to go back to the corresponding page.

70| Click the 'Proceed' button corresponding to the 'Interactive Heatmaps' view. A new dialog box will appear that will allow you to choose which data sets you want to include in your heatmap visualization. Keep all data sets checked. Click 'OK' to proceed. Note that the heatmaps are created from the differentially expressed genes selected during meta-analysis. When the heatmap is too large (owing to too many samples), you can use this dialog box to display only a subset of the data at a time to save space and for better performance.

71| On the new page, the heatmaps are displayed showing the expression patterns of genes across all three data sets. Follow the procedures described in Step 39A–G to explore the results.

? TROUBLESHOOTING

? Troubleshooting advice can be found in **Table 3**.

TABLE 3 | Troubleshooting table.

Step	Problem	Possible reason	Solution
1	The content of the home page does not show up	JavaScript is disabled in your browser	For Google Chrome, go to the Chrome menu > Preferences > Privacy > Content Settings > JavaScript tab and choose the option 'Allow all sites to run JavaScript'. For Mozilla Firefox 3.0+, go to the Tools > Options > Content, and then select the checkbox beside 'Enable JavaScript'. For Safari 4.0+, go to the Edit > Preferences > Security, and then select the checkbox beside 'Enable JavaScript'. Please check the documentation for other browsers on how to enable JavaScript
12	Browser is slow or freezes	CPU or memory is not sufficient	Use a computer equipped with better resources. We recommend 4 GB of RAM and an Intel Core i5/i7 processor or equivalent. If this is not possible, you can try the following: try to close other programs that are competing for resources; update the browser to the latest version (we recommend Google Chrome); and reduce the amount of the data to be visualized
54	Data integrity check failed	Class labels are not consistent. There are very few overlapping genes	Edit the class labels using the 'Annotate' functions; make sure that the 'DE Analysis' and 'Data Summary' show the same class comparison. Exclude the data set(s) (in the case of very few overlapping features or with too many missing values)
56	Combining ranks failed	Data sets are too large for this procedure	Remove the data set that is most heterogeneous compared with others. Try a parametric approach that can deal with a reasonable amount of heterogeneity (i.e., REM for combining effect size). Install a local copy of NetworkAnalyst on a more powerful server

● **TIMING**

The timing required for the steps of the protocol depends on the size of the data, the speed of Internet access, as well as the number of active users currently connected to the public web server. For the example data sets used for the protocols, most results should be returned in a few seconds when a user adjusts parameters. The most time-consuming part is visual analytics, where users can spend a long time interacting with the graphics and the server. An experienced user can execute this list of protocols in ~30 min. A novice user should be able to complete these protocols within ~90 min.

ANTICIPATED RESULTS

Graphical output

The graphical outputs produced during the analysis procedures are given in **Figures 3**–**7**. Please note that some NetworkAnalyst algorithms including network layout algorithms use time-dependent random number generators, and thus results may vary slightly among runs.

Network builder results generated in Step 9 using example data set 1

Four subnetworks are identified as the default first-order interaction subnetworks. The top one (subnetwork1) contains 916 nodes (including 85 seeds) and 1,118 edges.

Meta-analysis of a single gene expression data set generated in Step 34 using example data set 2

The number of differentially expressed genes for analysis without considering donor effect are as follows: LPS (206), LPS2 (135) and LPS_LPS2 (236), and the number of genes for analysis with donor effect are: LPS_D (215), LPS2_D (153) and LPS_LPS2_D (251).

Meta-analysis of multiple gene expression data sets generated in Step 57 using example data set 3

A total of 154 differentially expressed genes should be identified using the Stouffer's method for combining *P* values using the default 0.05 *P* value cutoff.

ACKNOWLEDGMENTS The authors thank the Canadian Institutes for Health Research (CIHR) for financial support.

AUTHOR CONTRIBUTIONS J.X. developed NetworkAnalyst and prepared the protocol, E.E.G. tested the tool and the protocol and R.E.W.H. participated in all processes. All authors have read and approved the paper.

COMPETING FINANCIAL INTERESTS The authors declare no competing financial interests.

1. Li, S. *et al.* Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. *Nat. Immunol.* **15**, 195–204 (2014).
2. Pena, O.M. *et al.* An endotoxin tolerance signature predicts sepsis and organ dysfunction at initial clinical presentation. *EBioMedicine* **1**, 64–71 (2014).
3. Zhang, G. *et al.* Integration of metabolomics and transcriptomics revealed a fatty acid network exerting growth inhibitory effects in human pancreatic cancer. *Clin. Cancer Res.* **19**, 4983–4993 (2013).
4. Gieger, C. *et al.* Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet.* **4**, e1000282 (2008).
5. Gomez-Cabrero, D. *et al.* Data integration in the era of omics: current and future challenges. *BMC Syst. Biol.* **8** (suppl. 2), I1 (2014).
6. Tseng, G.C., Ghosh, D. & Feingold, E. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res.* **40**, 3785–3799 (2012).
7. O'Donoghue, S.I. *et al.* Visualizing biological data-now and in the future. *Nat. Methods* **7**, S2–S4 (2010).
8. Goble, C. & Stevens, R. State of the nation in data integration for bioinformatics. *J. Biomed. Inform.* **41**, 687–693 (2008).
9. Smith, B. *et al.* The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* **25**, 1251–1255 (2007).
10. Wodke, J.A. *et al.* MyMpn: a database for the systems biology model organism *Mycoplasma pneumoniae*. *Nucleic Acids Res.* **43** (Database issue): D618–D623 (2014).
11. Breuer, K. *et al.* InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation. *Nucleic Acids Res.* **41**, D1228–D1233 (2013).
12. Rhodes, D.R. *et al.* Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia* **9**, 166–180 (2007).
13. Chelaru, F., Smith, L., Goldstein, N. & Bravo, H.C Epiviz: interactive visual analytics for functional genomics data. *Nat. Methods* **11**, 938–940 (2014).
14. Nekrutenko, A. & Taylor, J. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nat. Rev. Genet.* **13**, 667–672 (2012).
15. Xia, J. *et al.* INMEX—a web-based tool for integrative meta-analysis of expression data. *Nucleic Acids Res.* **41**, W63–W70 (2013).
16. Xia, J., Lyle, N.H., Mayer, M.L., Pena, O.M. & Hancock, R.E.W. INVEX—a web-based tool for integrative visualization of expression data. *Bioinformatics* **29**, 3232–3234 (2013).
17. Xia, J., Benner, M.J. & Hancock, R.E.W. NetworkAnalyst—integrative approaches for protein-protein interaction network analysis and visual exploration. *Nucleic Acids Res.* **42**, W167–W174 (2014).
18. Tarraga, J. *et al.* GEPAS, a web-based tool for microarray data analysis and interpretation. *Nucleic Acids Res.* **36**, W308–W314 (2008).
19. Reich, M. *et al.* GenePattern 2.0. *Nat. Genet.* **38**, 500–501 (2006).
20. Xia, J., Mandal, R., Sinelnikov, I.V., Broadhurst, D. & Wishart, D.S. MetaboAnalyst 2.0—a comprehensive server for metabolomic data analysis. *Nucleic Acids Res.* **40**, W127–W133 (2012).
21. Xia, J. & Wishart, D.S. Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst. *Nat. Protoc.* **6**, 743–760 (2011).
22. Saeed, A.I. *et al.* TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* **34**, 374–378 (2003).
23. Perez-Llamas, C. & Lopez-Bigas, N. Gitools: analysis and visualisation of genomic data using interactive heat-maps. *PLoS ONE* **6**, e19541 (2011).
24. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
25. Huang, D.W. *et al.* DAVID bioinformatics resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res.* **35**, W169–W175 (2007).
26. Reimand, J., Arak, T. & Vilo, J. g:Profiler—a web server for functional interpretation of gene lists (2011 update). *Nucleic Acids Res.* **39**, W307–W315 (2011).
27. Lynn, D.J. *et al.* InnateDB: facilitating systems-level analyses of the mammalian innate immune response. *Mol. Syst. Biol.* **4**, 218 (2008).
28. Saito, R. *et al.* A travel guide to Cytoscape plugins. *Nat. Methods* **9**, 1069–1076 (2012).
29. Smyth, G.K. Limma: linear models for microarray data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (eds. Gentleman, R. *et al.)* 397–420 (Springer, 2005).
30. Law, C.W., Chen, Y., Shi, W. & Smyth, G.K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
31. Orchard, S. *et al.* Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat. Methods* **9**, 345–350 (2012).
32. Turinsky, A.L., Razick, S., Turner, B., Donaldson, I.M. & Wodak, S.J. Interaction databases on the same page. *Nat. Biotechnol.* **29**, 391–393 (2011).
33. Bader, G.D., Betel, D. & Hogue, C.W. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.* **31**, 248–250 (2003).
34. Stark, C. *et al.* BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* **34**, D535–D539 (2006).
35. Licata, L. *et al.* MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.* **40**, D857–D861 (2012).
36. Hermjakob, H. *et al.* IntAct: an open source molecular interaction database. *Nucleic Acids Res.* **32**, D452–D455 (2004).
37. Rolland, T. *et al.* A proteome-scale map of the human interactome network. *Cell* **159**, 1212–1226 (2014).
38. Pena, O.M., Pistolic, J., Raj, D., Fjell, C.D. & Hancock, R.E.W. Endotoxin tolerance represents a distinctive state of alternative polarization (M2) in human mononuclear cells. *J. Immunol.* **186**, 7243–7254 (2011).
39. Ramasamy, A., Mondry, A., Holmes, C.C. & Altman, D.G. Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med.* **5**, e184 (2008).
40. Mitra, K., Carvunis, A.R., Ramesh, S.K. & Ideker, T. Integrative approaches for finding modular structure in biological networks. *Nat. Rev. Genet.* **14**, 719–732 (2013).
41. Ideker, T., Ozier, O., Schwikowski, B. & Siegel, A.F. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* **18** (suppl. 1), S233–S240 (2002).
42. Beisser, D., Klau, G.W., Dandekar, T., Muller, T. & Dittrich, M.T. BioNet: an R package for the functional analysis of biological networks. *Bioinformatics* **26**, 1129–1130 (2010).
43. Vidal, M., Cusick, M.E. & Barabasi, A.L. Interactome networks and human disease. *Cell* **144**, 986–998 (2011).
44. Yu, H. *et al.* High-quality binary protein interaction map of the yeast interactome network. *Science* **322**, 104–110 (2008).
45. Schramm, S.J. *et al.* Disturbed protein-protein interaction networks in metastatic melanoma are associated with worse prognosis and increased functional mutation burden. *Pigment Cell Melanoma Res.* **26**, 708–722 (2013).
46. Liu, Y., Koyuturk, M., Barnholtz-Sloan, J.S. & Chance, M.R. Gene interaction enrichment and network analysis to identify dysregulated pathways and their interactions in complex diseases. *BMC Syst. Biol.* **6**, 65 (2012).
47. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
48. Bastian, M., Heymann, S. & Jacomy, M. Gephi: an open source software for exploring and manipulating networks. in *International AAAI Conference on Weblogs and Social Media* http://www.aaai.org/ocs/index.php/ICWSM/09/paper/viewFile/154Forum/1009 (2009).