Bioinformatics, 2018, 1–10
doi: 10.1093/bioinformatics/bty534
Advance Access Publication Date: 28 June 2018
Applications Note

Systems Biology

Supplementary Material for "SLIDE – a web-based tool for interactive visualization of large-scale –omics data"

Soumita Ghosh[1,2], Abhik Datta[3], Kaisen Tan[4] and Hyungwon Choi[1,2,*]

[1]Institute of Molecular and Cell Biology, A*STAR, Singapore 138673, [2]Saw Swee Hock School of Public Health, National University of Singapore, Singapore 117549, [3]Centre for BioImaging Sciences and Department of Biological Sciences, National University of Singapore, Singapore 117551, [4]Department of Otolaryngology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 119228

*To whom correspondence should be addressed.

Associate Editor: Oliver Stegle

Received on December 12, 2017; revised on May 15, 2018; editorial decision on June 26, 2018; accepted on June 27, 2018

Abstract

Summary: In this supplement, we review the literature of visualization tools for –omics data and describe the software implementation of SLIDE in detail. We illustrate the functions of SLIDE using a time course microarray data set, which profiles whole lung mouse cells infected with influenza virus of different strains in varying lethal and sub-lethal doses. The data set consists of 133 microarrays with >45,000 transcripts, which could not be visualized in other open-source tools.

Contact: hwchoi@imcb.a-star.edu.sg

Visualization interfaces in SLIDE can be classified into feature-level modules and group-level modules. These two interfaces are presented in **Figure S1**. The various functionalities are discussed in **Section S3**.

S1 Relevant Work

Currently, many open-source and commercial –omics data visualization tools are already available. These tools offer a wide range of functionalities and different degrees of scalability, some with interactive visualization. The strengths and limitations of commonly used tools are summarized in **Table S1** with direct comparison to SLIDE. For example, one notable limitation of existing alternatives is the scalability of clustering. Typical molecular datasets can reach up to >50,000 features from hundreds of samples (e.g. microarray, RNA-Seq). The scalability of Java TreeView (Saldanha, 2004), MapleTree (Simirenko, 2003) and FTreeView (Freudenberg *et al*., 2009) is partly constrained by their dependency on Cluster 3.0 (De Hoon *et al*., 2002; Eisen *et al*., 1998), which is their main engine for clustering. Heatmapper (Babicki *et al*., 2016), ClustVis (Metsalu and Vilo, 2015) and INVEX (Xia *et al*., 2013) use the computational capabilities of R packages to cluster the data. These tools can cluster at most a few thousand genes in a reasonable amount of time (e.g. minutes).

One of the most popular tools for –omics visualization is Multi-experiment Viewer (MeV) (Saeed *et al*., 2003). MeV provides several statistical analysis and clustering methods. Although MeV also visualizes the data at user-defined resolutions, the tool has an upper limit in the number of features and does not perform separate analysis for subsets of features as SLIDE does. In terms of graphics, it has a lower bound of one pixel per data element. At this resolution, tall or fat matrices cannot be visualized well in one snapshot, a feature that is available in SLIDE irrespective of the size of the dataset.

Many tools listed in **Table S1**, including MeV, offer string matching-based searching of user provided metadata (such as gene symbols, Entrez and sequence accession numbers). However, they do not allow querying, tagging and extraction of individual features into an independent sub-analysis based on additional information such as biological pathways and evolutionarily conserved functions such as Gene Ontology (GO) (Ashburner *et al*. 2000). The ability to search keywords and tag features, simultaneously displaying the relevant feature data on a side panel in real-time, is a key advantage of SLIDE over others.

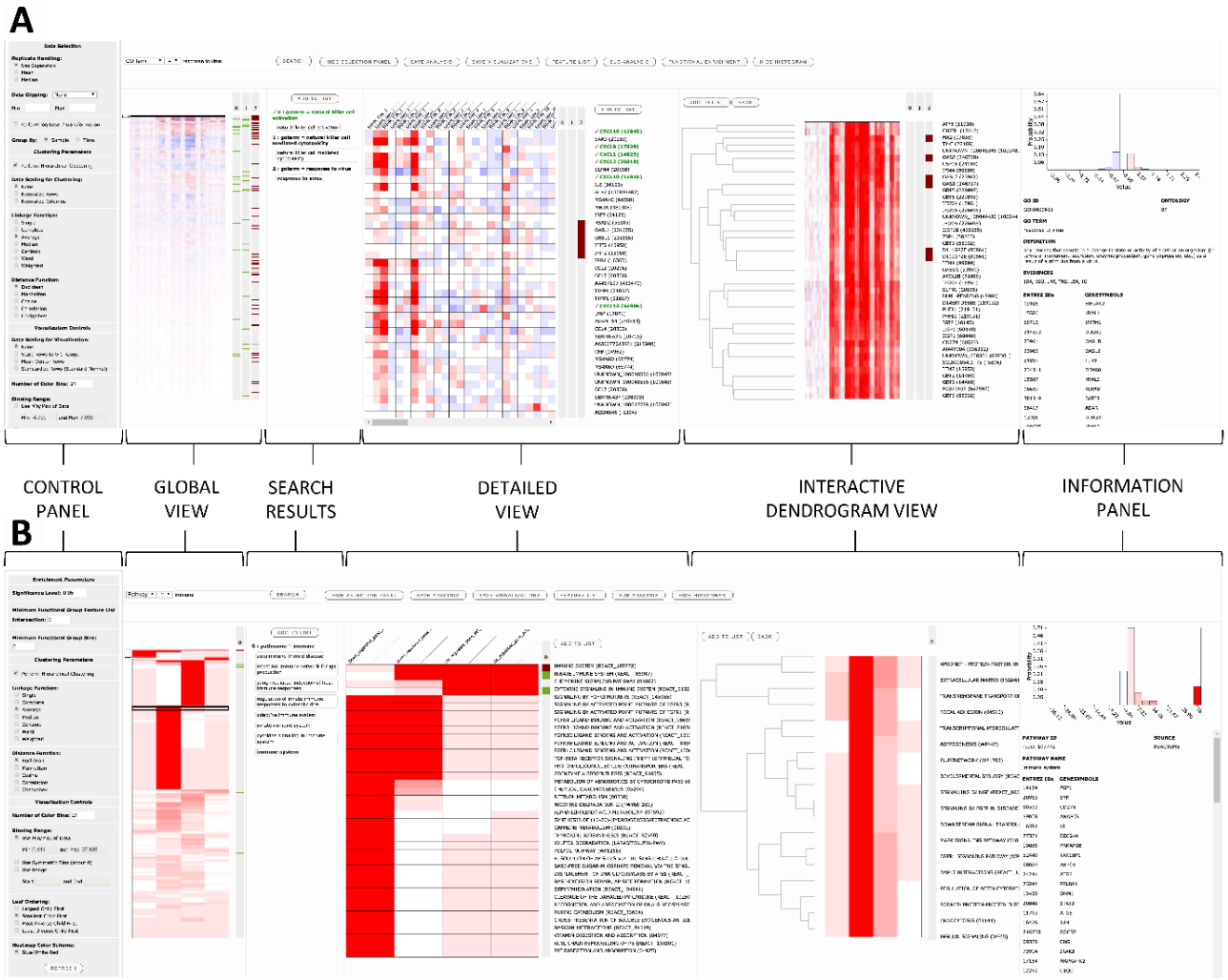

Fig. S1. SLIDE visualization interface. (A) Feature-level visualization interface of SLIDE. The global view heatmap visualizes the entire expression matrix after hierarchical clustering of the features. The search panel on top of the global view allows real-time search and tagging of the data. The search tags highlight features with horizontal (green and brown) stripes on vertical bars alongside the heatmaps, while the search terms are displayed in the search results panel. The detailed view heatmap gives a zoomed-in view of a portion of the entire data, selected using a slider on the left edge of the global view panel. In the interactive dendrogram view, the branches of the tree can be clicked to visualize a subset of the clustered data. Features can be selected and added to the user-created feature lists using the "ADD TO LIST" functionality in the search results, detailed view and interactive dendrogram view panels. Clicking features, pathways or gene ontologies (in the heatmap views and in search results) displays their details in the information panel. (B) Group-level visualization interface has the same components as feature-level visualization. However, the columns of the heatmaps are user-created feature lists and the rows are functional terms. The control panels in the two types of visualizations also have slightly different sets of parameters. See Software manual for details.

Table S1. Summary of features in existing –omics data visualization tools and SLIDE

Features \ Tools	MeV	Gitools	Heatmapper	Java TreeView / MapleTree	Functional TreeView	CIMminer	INVEX (Heatmap Clustering in NetworkAnalyst)	SLIDE
Scalability	Clustering scalability is limited by Java Virtual Machine's memory. Fails to perform hierarchical clustering of 45,281 genes when JVM is allocated 2 GB virtual memory.	Distance matrix computation for hierarchical clustering of 45,281 genes was less than 15% complete after 1 hour of processing	The input data is limited to 2500 features and 300 samples	Execution of hierarchical clustering failed for 45,281 genes with Cluster 3.0 due to insufficient memory space	Execution of hierarchical clustering failed for 45,281 genes with Cluster 3.0 due to insufficient memory space	Limited to 1000 rows	Limited to 5000 rows	Visualizing and clustering 45,281 genes across 133 samples requires <400 MB memory. Hierarchical clustering of genes using average linkage and Euclidean distance requires <7 minutes.
Visualizations	Heatmaps, dendrograms and other plots	Heatmaps, dendrograms, etc	Heatmaps, dendrograms	Heatmaps, dendrograms	Heatmaps, dendrograms	Heatmaps	Heatmaps	Heatmaps, dendrograms
Clustering capabilities	Multiple clustering algorithms available	Hierarchical and kmeans++ clusterings available	Uses R packages for clustering	Depends on packages such as Cluster 3.0. Clustering output is the input for this tool.	Depends on packages such as Cluster 3.0. Clustering output is the input for this tool.	Uses R packages for clustering	Only hierarchical clustering is available	Only hierarchical clustering is available
Search Capabilities	Approximate (wildcard) and exact string matching based search of feature metadata	Approximate and exact string matching based search of feature and sample metadata	None	Approximate and exact string matching based search of feature metadata	Approximate and exact string matching based search of feature metadata	None	None	Approximate and exact search of feature metadata (entrez, gene symbol, ensembl, uniprot and refseq identifiers) as well as functional information
Functional Tagging	None	None	None	None	None	None	None	User input independent functional tagging
Enrichment Analysis	Multiple methods are available for biological interpretation of data	Requires genes to pathway mapping to be provided by user	None	None	Provides interface to DAVID and Enrichr for functional analysis	None	Gene ontology and pathway enrichment can be performed within the tool	Hypergeometric test based enrichment analysis and heatmap based visualization of enrichment. Gene to functional group mapping is available within the tool.

Visualizing at Multiple Resolutions	Fixed resolution. Requires scrolling to view the entire data. Ability to save heatmaps of tall or fat matrices as images is limited due to fixed resolution.	Allows visualizing at multiple resolutions, but the resolution has a lower bound of 1 pixel per data matrix cell, as a result large matrices cannot be viewed in their entirety in a single frame.	Fixed resolution	Allows visualizing at multiple resolutions, but the resolution has a lower bound of 1 pixel per data matrix cell, as a result large matrices cannot be viewed in their entirety in a single frame.	The data can be viewed at any desired resolution	Fixed resolution	Three fixed resolutions: low, medium and high	Data can be visualized in its entirety (i.e. at the lowest resolution) as well as at high-resolutions where individual features are clearly visible.
Sublist generation	Subsets of features can be selected by selecting clusters in the dendrogram and saved	None	None	Subsets of features can be selected and saved through text based search of metadata, using the dendrogram and by directly selecting data rows.	Subsets of features can be selected and saved through text based search of metadata, using the dendrogram and by directly selecting data rows.	None	List of selected features can be generated through manual selection. List also has to be manually saved in file	Subsets of features can be selected (through enhanced search of metadata, functional search and using the dendrogram) and saved

Test for enrichment of biological functions in selected sets of genes, often called gene set enrichment analysis, can also be performed in most existing tools. However, some tools (with the exception of INVEX) require that the user provide the mapping between genes and pathways. For instance, Gitools (Perez-Llamas and Lopez-Bigas, 2011) requires a gene-to-pathway mapping as input for enrichment analysis. FTreeView provides interfaces to external tools such as DAVID (Huang, D. W. *et al*., 2007) and Enrichr (Chen *et al*., 2013). Similar to INVEX, analysis results can be immediately visualized in SLIDE in the form of a heatmap, a feature that is lacking in most alternatives.

Many existing tools such as CIMminer (Weinstein *et al*., 1994) and GenePattern (Reich *et al*., 2006) were developed without the software architecture necessary to support high levels of user interactivity. CIMminer, for instance, requires the data to be uploaded to a server, which performs clustering and emails back a link to the user for visualizing the result. Several other tools such as HeatmapGenerator (Khomtchouk *et al*., 2014) and matrix2png (Pavlidis and Noble, 2003) are also available for specifically creating heatmaps without any analysis capabilities. In contrast, SLIDE offers a complimentary set of functionalities, such as the ability to interactively explore very large datasets at the level of gene sets or biological functions, along with the ability to customize and save the visualizations in a resolution independent format.

S2 Implementation

S2.1 Software Architecture

SLIDE has been designed as a Java and Python-driven web application with a rich interface that can be accessed using any modern web browser. It uses a well-established web server, the Glassfish server, and the widely used MongoDB database. Both Glassfish and MongoDB are open source. Since the web server, the database, Java and Python are all freely available and can be easily installed on most modern operating systems, SLIDE can also be used as a standalone application. HTML is used to create the backbone of the graphical user interface, within which the visualizations are rendered using Scalable Vector Graphics (SVG). CSS is used for styling and JavaScript is used for client-side user interactivity.

To achieve scalability, a key design philosophy in SLIDE is to maintain a constant amount of data in the web browser, at all times. This is achieved using Asynchronous JavaScript and XML (AJAX) based updates that dynamically load only the portion of the data needed. The heatmaps in global view and interactive dendrogram view are rendered as images of fixed dimensions. The heatmap data for detailed view is asynchronously requested from the server as needed. The amount of data that can be loaded and visualized using SLIDE is therefore only limited by the amount of main memory (random-access memory) available. Loading and visualizing a dataset with 50,000 features and 200 rows requires ~4GB RAM, while running hierarchical clustering on it requires ~16GB RAM.

The server-side contains the GlassFish web server as the visualization engine. The application logic has been implemented using Java Server Pages (JSPs) and Servlets. The data processing logic has been implemented using Java and Python. Combining Java and Python provides a balance between software development time and software performance. For instance, SLIDE uses the Python package fastcluster (Müllner, 2013), which provides efficient implementations of agglomerative hierarchical clustering. The functional information is maintained as a repository in a MongoDB server. MongoDB allows storing non-flat, document like structures that can be queried with close to in-memory like performance. The SLIDE repository comprises of gene information, gene ontology information and biological pathway information. The database schemas have been optimized and indexed such that near real-time performance is possible for the most common

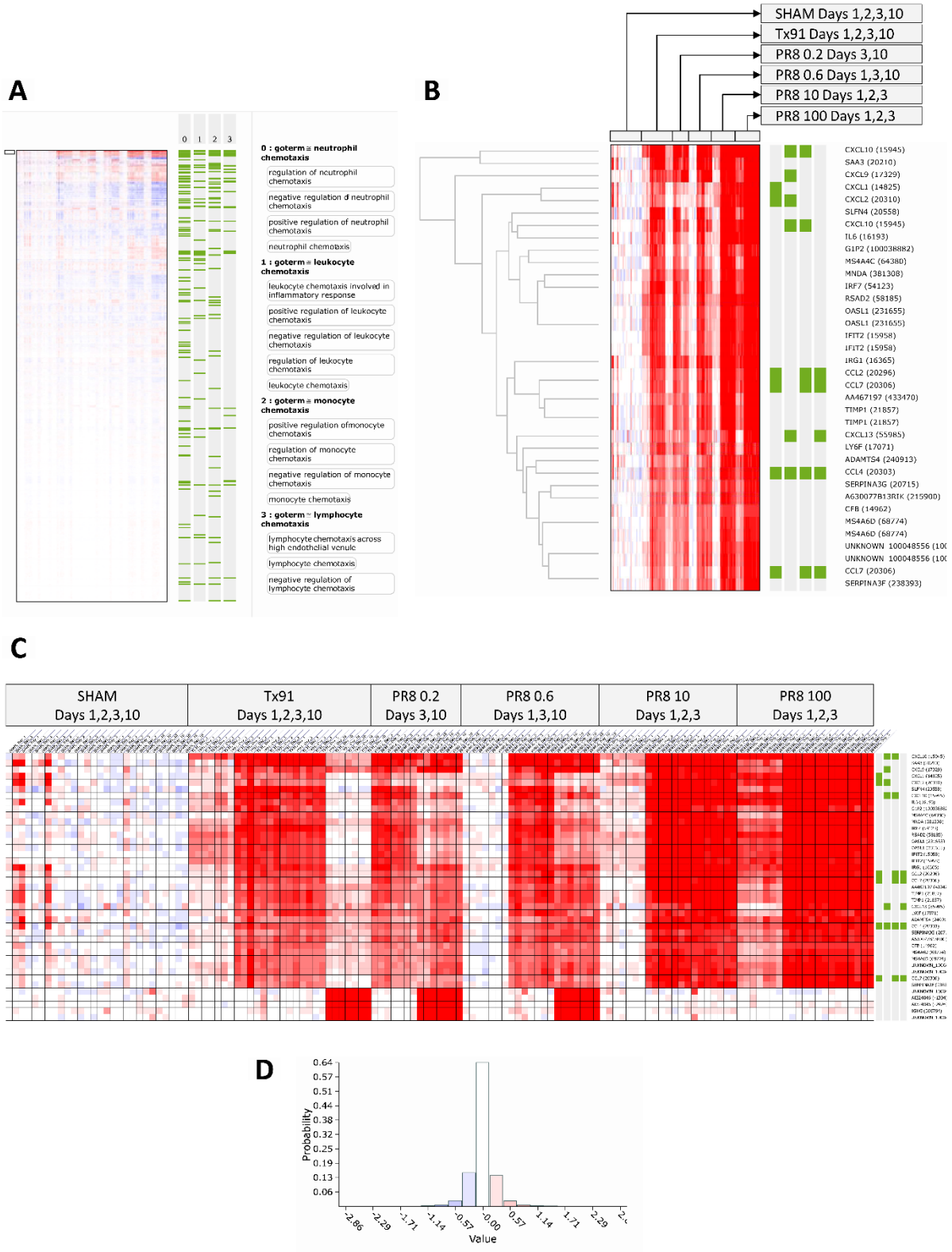

Fig. S2. Search-based feature tagging and sub-analysis creation functionalities applied to mRNA expression data of mouse infection model. (A) A global view visualizing the mRNA expression matrix, comprising of 45,281 genes across 133 experiments, after hierarchical clustering of the genes using average linkage and Euclidean distance. The search results panel displays the results of wildcard search for four GO terms. The genes associated with the matched GO terms are tagged (with green stripes) in real-time. **(B)** The interactive dendrogram view of a sub-analysis created using the features tagged in the search result "neutrophil chemotaxis". **(C)** The detailed view of the same sub-analysis, showing the features at a higher resolution. **(D)** the histogram of the expression data visualized in **(A)**. In **(B)** and **(C)** chemokine and chemokine receptors show distinctive dose dependent expression patterns across the time-course.

queries. The gene information, GO terms are extracted from the R package Bioconductor (Gentleman *et al*. 2004). A comprehensive list of biological pathway information for the human and mouse species are also extracted from ConsensusPathDB (CPDB) (Kamburov *et al*., 2010). In the current release, SLIDE supports functional annotation for human and mouse data. The MongoDB repository is automatically updated periodically with data re-extracted from Bioconductor and CPDB.

A user manual outlining SLIDE's interactive features is available at https://github.com/soumitag/SLIDE/raw/master/application/slide/SLIDE_Users_Manual.pdf. The "HELP" menu in SLIDE's user interface provides quick access to the user manual, example input data and sample information files and GitHub's issue reporting site.

S2.1 Input and Output Data Formats

Feature-level visualization in SLIDE (e.g. expression data) requires an input data file containing the matrix of quantitative values in a delimited text format (comma, tab, space, semi-colon or pipe delimited). Additionally, a sample information file is required if the data contains sample group information (such as replicates, experimental conditions, time points). The sample information file should also be a delimited text file. The user will be prompted to specify delimiters for both input files. Sample group names can be used to indicate experimental conditions such as vital status, control and test conditions. Each such grouping is referred to as a factor and SLIDE allows grouping by up to two factors (e.g. disease status as having a tumor or not can be two sample groups of one factor and treatment groups such as chemotherapy, immunotherapy can be sample groups of the second factor). The detailed format of the sample information file is discussed in **Section II.2** of SLIDE user manual, available at available at https://github.com/soumitag/SLIDE/raw/master/application/slide/SLIDE_Users_Manual.pdf.

The sample names in the sample information file must be identical to the column headers in the input data file. Comments can be added in the sample information file by starting the comment line with a '#' symbol. Any line that does not begin with a '#' symbol and is not empty must contain a valid sample mapping. Rows with the same sample group name and time point are automatically detected as replicates. Only the sample names mentioned in the sample information file are loaded from the input data file.

An example sample information file for the mouse data set used here is available at https://github.com/soumitag/SLIDE/blob/master/data/Brandes_et_al_GSE42638_Sample_Information.txt. The input data file should contain one or more feature meta-data such as Entrez, official gene symbol, RefSeq, Ensemble and Uniprot identifiers. SLIDE can automatically map missing meta-data information if any one of these identifiers is available.

The feature lists created within SLIDE can be saved in a text file format. Any analysis or sub-analysis workspace can be saved in a ".slide" file, which can be loaded back onto SLIDE for continued analysis. All visualizations within SLIDE can be saved as Scalable Vector Graphics (SVG) files, which are resolution-independent and can be used to create high quality images. When saving visualizations from SLIDE, several customization options are available. For instance, one can choose to include the search tags as well as the histogram of values.

S3 Illustration of SLIDE in a time course microarray data with complex experimental design

We demonstrate the capabilities of SLIDE using mRNA expression data from an influenza infection study of murine lung epithelial cells. The study profiled the transcriptomic changes in the whole lung epithelial cells after non-lethal, sub-lethal and lethal influenza infection (Brandes *et al*., 2013). Specifically, the mouse lung epithelial cells were infected with non-lethal H1N1 virus strain A/Texas/36/91 (Tx91) strain at 10^6 plaque-forming units (PFU), sub-lethal and lethal doses of H1N1 A/Puerto Rico/8/34 (PR8) strain and sham infection on 1, 2, 3 and 10 days post infection. The sub-lethal doses are 0.2 and 0.6 times the median lethal dose (LD50) and the lethal doses are 10 and 100 times the median lethal dose. The microarray experiment consists of 19 experimental conditions (see **Figures S2B** and **S2C** for details) with 7 biological replicates per condition. The expression matrix consists of 45,281 genes and 133 (19x7) gene expression microarray experiments. The input data was quantile normalized, log base 2 transformed ($\log_2$) and baseline transformed by the median of sham-infected. For a demo analysis of this data using SLIDE go to http://137.132.97.109/VTBox/ and click the 'Load Demo' button.

Brandes *et al*. applied systematic statistical filtering to first identify 8,291 differential genes between the 19 experimental conditions and formed 50 clusters of genes (referred to as modules) that are highly correlated across multiple conditions through an optimization process. Their analysis identified condition specific inflammatory responses. For instance, excess neutrophil-mediated inflammatory damage was identified as characteristic of fatal infection. While formal statistical procedures based inference is indispensable, here we show that user-guided graphical navigation of the data with SLIDE, can also recover similar gene modules solely based on visualization, and can even confirm the conclusions from statistical filtering graphically.

S3.1 Sample (Meta) information file

Each row in the sample information file contains three entries in this example: sample name, sample group name, and time point. For this dataset, it is possible to group samples based on their characteristics or experimental conditions, such as dose, strain, sample identifier, time point. Here, we use two sample grouping factors: (i) a combination of H1N1 strains and doses, with sample group names such as Sham, Tx91, 0.2PR8LD50, and so on; and (ii) the number of days post infection (time point). The column order can be modified within SLIDE in real-time, to sort the columns either by sample group name or by time point. When sorted by sample group name, the samples within each sample group are sorted by time points, and vice versa. Thus, input data need not be sorted before loading into SLIDE; a feature that can be particularly useful for data sets with many columns.

S3.2 Feature-level visualization of the entire data

The global heatmap in **Figure S2A** shows the result of hierarchical clustering of the 45,281 genes in the whole lung RNA expression data using Euclidean distance and average linkage. The clustering trees in **Figure S1A** are sorted using the "smallest child first" ordering scheme (Bar-Joseph, Z. *et al*., 2001). A small region of the top-most upregulated cluster in **Figure S2A** is shown at a greater depth in the interactive dendrogram view in **Figure S2B**. The dose dependent immune response can be visually identified in this figure. For instance, CXCL1 and IL6, two

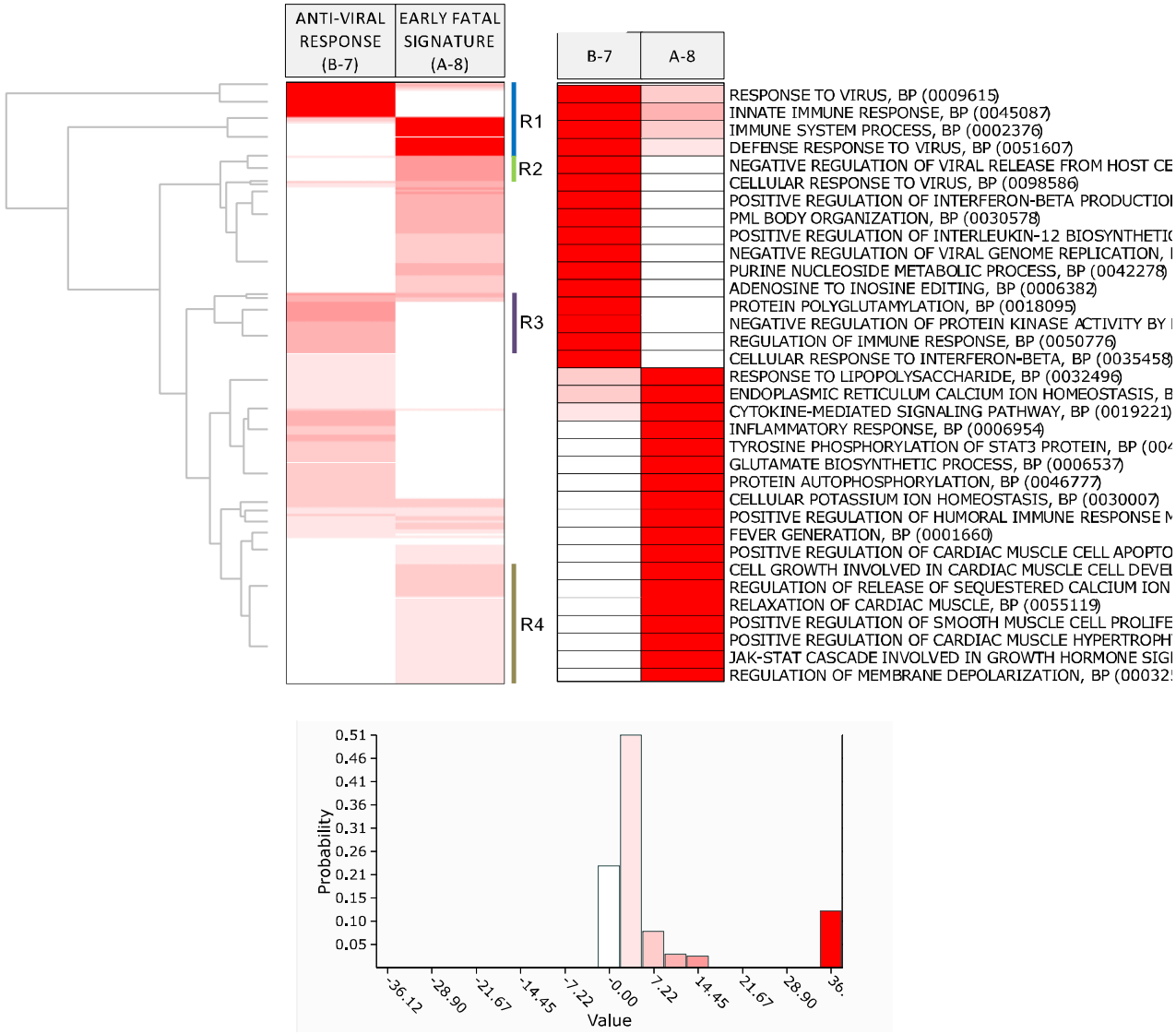

Fig. S3. Enrichment Analysis and Group-level visualization using feature lists. Enrichment analysis results in the interactive dendrogram view. Brandes *et al*. identified two distinct biological processes through their modular analysis of the mouse infection model. The first was a common antiviral response pattern of host across all infection conditions, the second was a condition specific innate immune response, which was pro-inflammatory and constituted a fatal molecular signature with increased lethality. **(A)** The enrichment level of significant functional terms in the gene lists associated with these two modules, referred to as B-7 (anti-viral response) and A-8 (early fatal signature). **(B)** The region in **(A)** highlighted by the vertical bar R1 in detailed view. The regions show several highly enriched biological processes in both feature lists. The regions highlighted by the vertical bars R2, R3 and R4 are presented in **Figure S4**. **(C)** The histogram of the data visualized in **(A).**

important markers of inflammatory response, show large increase in expression levels with increasing dose. Figure S2C shows the same features in a detailed view, the non-lethal strain Tx91 and sub-lethal doses of PR8 strain, CXCL1 and IL6 are over expressed in Days 2 and 3 compared to baseline (Day 1). While the expression levels of these genes return to the baseline levels by Day 10 for non-lethal Tx91 and sub-lethal doses of PR8, the same genes remain at relatively higher expression levels at Day 10 for the lethal PR8 doses.

S3.3 Query-based feature tagging and sub-analysis

Brandes *et al*. found that excess inflammatory host response is caused by chemokine-driven neutrophil infiltration at the site of infection. Chemokines orchestrate cell migration and play a crucial role as mediators of acute inflammation. CXCL1, the main trigger of neutrophil recruitment, exhibits a distinguishable expression pattern in lethal dose levels of infection by pathogenic strain PR8 in **Figure S2B**. To visualize the distribution of genes related to cell chemotaxis across various gene clusters, GO terms related to neutrophil chemotaxis, leukocyte chemotaxis, monocyte chemotaxis, and lymphocyte chemotaxis were searched by user query in SLIDE. The result of this wildcard search-based tagging pulled

the subset of expression data for relevant genes, as shown in **Figure S2A**. A wildcard search in SLIDE returns all searched entities that contain the searched literal. For instance, the wildcard search for GO terms containing "neutrophil chemotaxis" found four matches (**Figure S2A**).

A sub-analysis was created with the genes associated with the four GO terms returned by the wildcard search for the GO term "neutrophil chemotaxis". A portion of this subset of genes is shown in detailed view in **Figure S2C**, where the expression levels of neutrophil chemokines such as CCL4, CCL7, CXCL1, CXCL2 increased rapidly with increased lethality. For 10 times LD50 and 100 times LD50 doses of PR8, these genes were already highly expressed in day 2, whereas for 0.6 times LD50 dose of PR8 their expression change is visible only from day 3. This visualization indicates there is an early neutrophil infiltration at the infection site for higher lethal doses compared to lower lethal doses, consistent with the findings of Brandes *et al*.

S3.4 Group-level visualization: enrichment analysis

As discussed above, Brandes *et al*. took a modular approach to analyze the biological processes that distinguish influenza infections based on their lethality and identified 50 exclusive modules. We use SLIDE's group-level visualization to determine biological processes enriched in two modules, B-7 and A-8, which they investigated in depth and found to be associated with distinct molecular signatures of host response to infection. The first module, B-7, is linked to antiviral response and the second module, A-8, is linked to the host inflammatory response pattern that constitute early fatal signature. The result of GO enrichment analysis on these two modules is presented in **Figure S3A**, which shows the interactive dendrogram view containing all biological processes present in the Gene Ontology database, that were found to be significant. The colors in the heatmap represent magnitude of statistical significance ($-\log 10\ p-\text{value}$). Darker colors therefore indicate greater statistical significance of enrichment in a feature list. As can be seen in **Figure S3A**, there are several significant pathways with *p*-value less than 0.05 in the hypergeometric test. Some interesting clusters of biological processes are presented in detailed view in **Figure S3B** and in the **Figures S4A**, **S4B** and **S4C**. The heatmap in **Figure S3B** shows the biological processes in Region R1 of **Figure S3A**, and contains the most enriched pathways. The biological processes in Regions R2, R3 and R4 of **Figure S2A** are presented in **Figures S4A**, **S4B** and **S4C**, respectively.

In **Figure S3B**, the GO terms "inflammatory response" and "response to virus", as expected, are enriched in modules A-8 and B-7, respectively. In addition, a number of other enriched biological processes were also found. For instance, several immune system related GO terms such as "Innate Immune Response", "Immune System Process", "Regulation of Immune Response" are also highly enriched in module B-7. JAK-STAT signaling related pathways, which Brandes *et al*. identified as distinguishing host response patterns associated with early fatality, are also enriched in module A-8 (**Figure S4A** and **Figure S4B**).

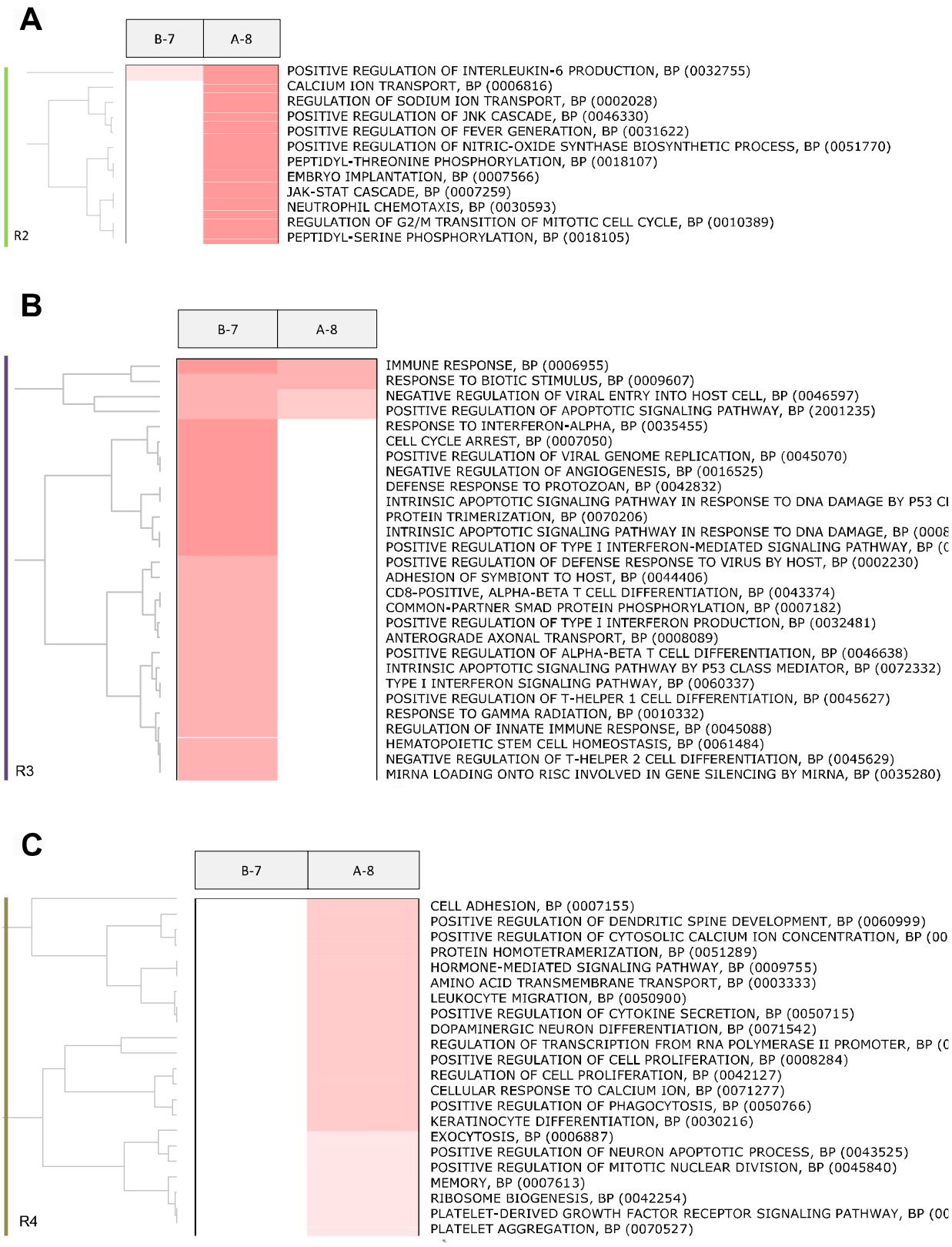

Fig. S4. Enrichment Analysis and Group-level visualization using feature lists (continued). (A) Shows the region of Figure S3A highlighted by the vertical bar R2, in detailed view. (B) Shows the region of Figure S3A highlighted by the vertical bar R3, in detailed view. (C) Shows the region of Figure S3A highlighted by the vertical bar R4, in detailed view.

References

Ashburner, M. *et al*. (2000) Gene Ontology: tool for the unification of biology. Nature genetics, **25(1)**, 25-29.

Babicki, S. *et al*. (2016) Heatmapper: web-enabled heat mapping for all. *Nucleic acids research*, **44(W1)**, W147-W153.

Bar-Joseph, Z. *et al*. (2001) Fast optimal leaf ordering for hierarchical clustering. Bioinformatics, **17(suppl_1)**, S22-S29.

Brandes, M. *et al*. (2013) A systems analysis identifies a feedforward inflammatory circuit leading to lethal influenza infection. *Cell*, **154(1)**, 197-212.

Chen, E. Y. *et al*. (2013) Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC bioinformatics*, **14(1)**, 128.

De Hoon, M. J. *et al*. (2002) Open source clustering software. *Genome Informatics*, **13**, 250-251.

Eisen, M. B. *et al*. (1998) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, **95(25)**, 14863-14868.

Freudenberg, J. M. *et al*. (2009) Clean: Clustering enrichment analysis. *BMC bioinformatics*, **10(1)**, 234.

Gentleman, R. C. *et al*. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, **5(10)**, R80.

Huang, D. W. *et al*. (2007) The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome biology*, **8(9)**, R183.

Kamburov, A. *et al*. (2010) ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic acids research*, **39(suppl_1)**, D712-D717.

Khomtchouk, B. B. *et al*. (2014) HeatmapGenerator: high performance RNAseq and microarray visualization software suite to examine differential gene expression levels using an R and C++ hybrid computational pipeline. *Source code for biology and medicine*, **9(1)**, 30.

Metsalu, T. and Vilo, J. (2015) ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. *Nucleic acids research*, **43(W1)**, W566-W570.

Müllner, D. (2013) fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python. *Journal of Statistical Software*, **53(9)**, 1-18.

Pavlidis, P. and Noble, W. S. (2003) Matrix2png: a utility for visualizing matrix data. *Bioinformatics*, **19(2)**, 295-296.

Perez-Llamas, C. and Lopez-Bigas, N. (2011) Gitools: analysis and visualisation of genomic data using interactive heat-maps. *PloS one*, **6(5)**, e19541.

Reich, M. *et al*. (2006) GenePattern 2.0. *Nature genetics*, **38(5)**, 500-501.

Saldanha, A. J. (2004) Java Treeview—extensible visualization of microarray data. *Bioinformatics*, **20(17)**, 3246-3248.

Saeed, A. I. *et al*. (2003) TM4: a free, open-source system for microarray data management and analysis. *Biotechniques*, **34(2)**, 374-378.

Simirenko, L. (2003) Maple Tree. http://mapletree.sourceforge.net/

Weinstein, J. N. *et al*. (1994) Predictive statistics and artificial intelligence in the US National Cancer Institute's Drug Discovery Program for Cancer and AIDS. Stem Cells, 12(1), 13-22. Weinstein, J. N. et al. (1994) Predictive statistics and artificial intelligence in the US National Cancer Institute's Drug Discovery Program for Cancer and AIDS. *Stem Cells*, **12(1)**, 13-22.

Xia, J. et al. (2013) INVEX—a web-based tool for integrative visualization of expression data. *Bioinformatics*, **29(24)**, 3232-3234.