multiSLIDE: a web server for exploring connected elements of biological pathways in multi-omics data

Soumita Ghosh[1], Abhik Datta[2,3], Hyungwon Choi[1,*]

1. Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
2. Centre for BioImaging Sciences, National University of Singapore, Singapore
3. Department of Biological Sciences, National University of Singapore, Singapore

*Corresponding author: Hyungwon Choi (hyung_won_choi@nus.edu.sg)

Abstract

Emerging multi-omics experiments pose new challenges for exploration of quantitative data sets. We present multiSLIDE, a web-based interactive tool for simultaneous heatmap visualization of interconnected molecular features in multi-omics data sets. multiSLIDE operates by keyword search for visualizing biologically connected molecular features, such as genes in pathways and Gene Ontologies, offering convenient functionalities to rearrange, filter, and cluster data sets on a web browser in a real time basis. Various built-in querying mechanisms make it adaptable to diverse omics types, and visualizations are fully customizable. We demonstrate the versatility of the tool through three example studies, each of which showcases its applicability to a wide range of multi-omics data sets, ability to visualize the links between molecules at different granularities of measurement units, and the interface to incorporate inter-molecular relationship from external data sources into the visualization. Online and standalone versions of multiSLIDE are available at https://github.com/soumitag/multiSLIDE.

Introduction

Constantly evolving massively parallel sequencing, mass spectrometry, and other omics technologies have made the joint use of omics modalities for molecular profiling of the same biological samples in both cell biology studies as well as in clinical applications (1). However, multi-omics data sets are extremely challenging to analyze not only because of the expanding dimensionality of data, but also because of the complexity that comes from the interconnected nature of multiple high-dimensional data sets. In traditional omics data analysis workflow, the data analysis often depends on considerable reduction of the data using statistical filters or abstraction via a projection of data into a low-dimensional space for visualization and interpretation. Although data reduction is unavoidable for effective presentation of the high-dimensional data, the dependence on reduction and abstraction inevitably causes scientists to miss meaningful fraction of the data features that fail to pass such filters. Therefore, there is a need to explore the unfiltered data in detail prior to any statistical analysis, and easy-to-use, interactive visualization tools can play an essential part in facilitating the unbiased exploration of the data.

There are a handful of published bioinformatics tools for multi-omics data visualization in the current literature. Open-source, data-rich web resources such as cBioPortal (2), UCSC Xena (3), and LinkedOmics (4) provide web-interfaces for query-based exploration and visualization of oncogenes from fixed data sources such as large-scale cancer cohorts of TCGA (5) and METABRIC (6). Pathway-based visualizations, such as PaintOmics3 (7), Escher (8), PathVisio3 (9), and network-based visualizations such as Cytoscape (10), 3Omics (11), MONGKIE (12), are popular options for summarizing the complex interconnectivities and dynamics between biomolecules in a single snapshot. These methods are focused on either displaying quantitative values for a small number of select markers or visualizing overall trends at an abstract level such as pathways or networks.

However, few existing tools directly visualize the quantitative data across all omics modalities in an intuitive manner and at a legible scale (13). The ability to inspect the trends of individual molecular features at multiple molecular levels at once is critical for the understanding of multiple data sets, especially when the molecular changes are discordant at different molecular levels. In large-scale multi-omics studies, we often find that the biological samples with the same phenotype can have a large variation in their molecular profiles between different omics modalities. These intra-sample variations should be accounted for before drawing biological inference using statistical models for multi-omics data (14–18).

To fill this crucial gap, we have developed multiSLIDE, an interactive heatmap visualization tool for easy exploration of multi-omics data. Through multiSLIDE, we provide an

interactive interface to explore multi-omics data through keyword-based queries. Quantitative molecular datasets are best represented using heatmaps (19, 20), and multiSLIDE visualizes the queried fraction of the omics data sets in separate heatmaps in one screen, with multiple panels synchronized with one another and with lines connecting related measurements to highlight the inter-connectivity. In summary, the tool visualizes raw multi-omics data, prioritized to a reasonable scale by keyword searches according to the user's hypothesis on the biological relevance of molecular features.

We demonstrate the visualization functionalities using three example studies with multi-omics data sets. The first dataset comes from a study profiling the time course mRNA and protein expression in HeLa cells undergoing unfolded protein response in the endoplasmic reticulum (ER) (21). The second data set presents connected visualization of phosphoproteome and proteome data, hierarchically linking phosphorylation sites (phosphosites) to the parent proteins, or linking kinase proteins with phosphorylation sites on substrate proteins (22). In the last example, we visualize the measurements of circulating microRNAs and proteins in human plasma samples from obese insulin resistant subjects and lean insulin sensitive subjects, connecting the 3' UTR sequence scan-based map of proteins and microRNAs (23). In combination, these examples demonstrate the versatility of multiSLIDE for visualizing key segments of multi-omics data with the guidance of keyword searches.

Methods

Functionalities of multiSLIDE web server

multiSLIDE visualizes the quantitative data relevant to the keywords in heatmaps, with molecular relationship or interactions indicated by connecting lines across omics modalities. Keywords can be pathways, gene ontology (GO) terms, and gene identifiers, e.g., immune response, insulin signaling, or AKT1. Given keywords, multiSLIDE retrieves all matching pathways, gene ontologies, and genes (or other individual molecules), and the user selects one or more search results that are of interest to them. multiSLIDE then visualizes the quantitative data for the retrieved molecules in heatmaps, simultaneously at all molecular levels. Users can apply additional statistical filtering on the retrieved data to narrow down to differentially expressed molecules between phenotypic groups using built-in parametric and non-parametric statistical tests. Additionally, in case the search retrieves too many data features, Benjamini-Hochberg multiple testing correction for controlling false discovery rate can also be applied (24). We remark that the adjustment p-value calculation uses the list of molecules selected by the user for

visualization (prior to filtering) as the background list, and not the full set of molecules available in the dataset. multiSLIDE also has two distinct modes of data clustering, namely synchronized and independent modes. The synchronized mode rearranges the queried molecules in individual heatmaps based on the clustering of one omics data (anchor data), while the independent mode clusters each omics data on its own.

For some omics modalities, it is possible to summarize the data at the gene level. For instance, when visualizing transcriptome and proteome, both datasets have measurements at the gene level. In such cases, a linker can synchronize the relationship between the heatmaps, resulting in a one-to-one mapping between molecules. In this instance of multiSLIDE, a linker is simply a molecular identifier that is common between datasets. In the absence of shared molecular identifiers between two omics data, any pair of molecules in different omics data are considered independent. This independent mode is necessary, for instance, when visualizing microRNA and protein data, where molecular identifiers do not overlap.

Moreover, it is often necessary visualize quantitative data at a resolution higher than genes. For instance, data are summarized by genomic location, or by mRNA transcripts rather than the whole gene, or by peptides rather than whole protein, to name a few. These nested identifiers, unless specifically linked, are distinct across omics modalities and may have many-to-many relationships. multiSLIDE can automatically recognize nested identifiers and group them for visualization. For instance, in the second case study, the phosphorylation site-level data has nested identifiers: proteins and their specific sequence positions of modifiable residues (p-sites). For this dataset, when features are ordered by proteins, multiSLIDE performs two-level hierarchical clustering. The order of proteins is determined by clustering the average peptide intensities (default choice) of all p-sites in the protein, and the order of p-sites within each protein is determined by independently clustering the data of the p-sites. Users can choose the summary statistic used (average, maximum, minimum, or sum) to compute the protein-level summaries.

multiSLIDE can also infer the relationship between omics modalities when datasets have shared identifiers. For instance, when visualizing transcriptomic and proteomic data, if the transcriptomics dataset contains gene symbols and the proteomic dataset contains Uniprot identifiers, multiSLIDE will internally map both identifiers to Entrez, and connect genes and proteins that have the same Entrez. We refer to these standard identifiers, which can be mapped to Entrez, as linkers.

When visualizing omics modalities where the resident molecules are mutually exclusive, e.g. proteome and metabolome, multiSLIDE allows users to input externally curated relationship data for linked visualization. Further, externally curated biological networks, such as transcription

factor (TF) regulatory networks (25–30) and kinase-substrate networks (31–33), can be integrated and visualized through linkers in multiSLIDE.

The structure of delimited input text files, various querying mechanisms, and output figures and analysis file options are discussed in the **Visualization Workflow** section below. The workflow is also described in the context of multiSLIDE's interface, in **Supplementary Figure S1**.

Visualization Workflow

Figure 1 illustrates a typical workflow in multiSLIDE. The web-based visualization interface of multiSLIDE is shown in **Supplementary Figure S1a**. In multiSLIDE, data analysis begins with the user specifying or selecting pathways, GO terms, or individual genes to visualize using one of three options: keyword search, enrichment analysis, or 'Pathway Upload' (see Input Data section for details). For the first option multiSLIDE provides an intuitive keyword-based search syntax for quickly searching multiple pathways, GO terms, and genes. The relevant genes and gene groups are selected from the search results and they are visualized, as shown in **Supplementary Figure S1b**, by clicking the group names.

Once the visualization is ready, network neighbors of a target gene on protein-protein interaction (PPI) and TF regulatory networks can be added via a quick network neighborhood search, all enabled by a simple right-click on a gene of interest. A selection of neighbors can be added to the visualization in real time.

Pathways and GO terms represent biological processes, and concerted regulation of gene expression for these processes are essential for the mechanisms underlying a phenotype. Instead of visualizing functional groups one at a time, visualizing an aggregated set of genes across a multitude of biological processes can produce the global landscape of gene expression regulation all at once. In multiSLIDE, the gene-function associations are visualized by vertical tracks next to each heatmap, highlighting the intersections (common genes) between functional groups (side bars on the left side of heatmaps, **Supplementary Figure S1a**).

The scales and dynamic range of detection and quantification are different for omics platforms. As a result, direct comparison of absolute expression values is not meaningful. Customizing the graphical parameters in each omics data are therefore essential for successful visualization. Individual heatmaps in multiSLIDE are independently customizable (see heatmap settings panel, **Supplementary Figure S1c**). Settings common to all heatmaps, such as zoom (or resolution) and orientation of heatmaps, are applied to all heatmaps simultaneously using the global settings panels (**Supplementary Figure S1f**). multiSLIDE has no restrictions on the amount of data that can be loaded in a single snapshot. As different systems and browsers have

different computing capabilities, this choice is left to the user. Using the layout options (**Supplementary Figure S1f**), the size of a single snapshot can be optimized, depending on the data transfer rate between the multiSLIDE server and the browser, and the browser's latency in rendering the data.

multiSLIDE has a variety of sorting, clustering, and filtering methods to help users discover patterns in their data. Interesting genes can be hard to discern when they are incoherently mixed with other genes, particularly when visualizing large pathways or networks. With the appropriate ordering of genes and samples, previously unforeseen structures in the data can emerge. In multiSLIDE, molecules can be sorted by gene groups, based on the statistical significance level in a differential expression analysis, or based on hierarchical clustering. Samples can be ordered by a combination of phenotypes or based on hierarchical clustering for interrogating the strength of phenotype-genotype associations. The hierarchical clustering can be customized by selecting different linkage functions, distance metrics, and leaf ordering schemes.

In multiSLIDE, statistically non-significant genes can be removed from the visualizations by internal differential expression analysis. In large pathways and networks, a substantial number of genes do not show differential expression, and thus removing these stably expressed or non-expressed genes improves the visualization clarity. multiSLIDE automatically classifies phenotypes into one of three categories based on the data: binary, categorical, or continuous. For binary and categorical data, users can choose to perform either parametric or non-parametric tests. The parametric tests used for binary and categorical phenotypes are two-sample t-test and analysis of variance (ANOVA), respectively. The corresponding non-parametric tests are Mann-Whitney U test and Kruskal-Wallis test, respectively. For continuous data, linear least squares regression is used. Additionally, users can also choose to perform multiple testing correction using the Benjamini-Hochberg procedure to control the false discovery rate.

As mentioned earlier, the user can choose to apply hierarchical clustering and filtering of features either in a synchronized or in an independent mode. In the synchronized mode, hierarchical clustering and filtering are performed on the features of one of the datasets selected by the user, and the ordering of that dataset is used to order the features of other datasets. This mode is only meaningful when the datasets share a linker column. In contrast, in the independent mode hierarchical clustering and filtering are applied independently for each omics data. A combination of these modes can also be used, where a subset of omics is synchronized and a subset is kept independent, by customizing the omics relationships, as shown in **Supplementary Figure S1f**.

Software Implementation

Software Architecture. multiSLIDE is built on a distributed architecture, shown in the schematic in **Supplementary Figure S2**. The server side of multiSLIDE consists of a state server, an analytics server and a knowledge server. The client can be any modern web browser. These four components are separate applications that communicate with each other through well-defined application programming interfaces (APIs). Due to this modular design, multiSLIDE can scale to distributed multi-node environments, with many possible deployment configurations.

The state server is an HTTP server, implemented in Java, that maintains client state information, user uploaded data, and user selections. The analytics server, also an HTTP server, is implemented in python and is the main computation engine. The knowledge server, implemented using MongoDB, manages the physical storage of curated gene annotation, regulatory networks, biological pathways and GO terms. The analytics and knowledge servers are stateless. The client interacts only with the state server by sending HTTP requests and receiving data in response in optimized JavaScript Object Notation (JSON). The client is implemented using Angular, with the data and presentation layers completely decoupled. Layouts can therefore be altered without the need to re-fetch data from the server. The visualizations are rendered using resolution independent Scalable Vector Graphics (SVG).

Databases: Pathways, GO terms and Molecular Networks. multiSLIDE includes comprehensive genome-scale annotations and gene ontology databases for mouse and human, extracted using R Bioconductor (34). The data in these R packages are well-structured and routinely used by bioinformaticians in their analyses. multiSLIDE can recognize five standard gene identifiers (linkers): Entrez, HUGO Gene Nomenclature Committee (HGNC) Gene Symbols, ENSEMBL identifiers, NCBI Reference Sequence (RefSeq) identifiers and UniProt identifiers, as well as miRNA identifiers from miRBase (35). To facilitate pathway and GO keyword-based search, multiSLIDE includes comprehensive biological pathways obtained from ConsensusPathDB (CPDB) (http://cpdb.molgen.mpg.de/) (36, 37). Validated miRNA - target interactions on pathways and GO from miRWalk2.0 (http://zmf.umm.uni-heidelberg.de/apps/zmf/mirwalk2/) (38) are also included in multiSLIDE.

Various networks indicating relationships between molecules within the same molecular level such as PPI network (within proteins), as well as networks indicating relationships between molecules at different levels such as TF regulatory networks, are also integrated in multiSLIDE, to facilitate network search. These networks come from diverse sources and are either experimentally validated relationships or putative interactions that rely on computational

prediction. For instance, TargetScan (http://www.targetscan.org) (39), a database widely used to represent miRNA-mediated gene regulation, houses predicted targets of miRNAs and is included in multiSLIDE. multiSLIDE also integrates Human Transcription Factor (TF) - targets network information from additional databases: TRED (25), ITFP (26), ENCODE (27), Neph2012 (28), TRRUST (29), and Marbach2016 (30). Mouse Transcription Factor (TF) - target network information was obtained directly from TRRUST. Physical interactions between proteins was sourced from iRefIndex (http://irefindex.org/wiki/index.php?title=iRefIndex) (40), which indexes protein-protein interaction networks from a number of databases.

Input Data. multiSLIDE provides a simple and intuitive interface for users to upload their own data and create an analysis. Each omics data should be uploaded into multiSLIDE in the form of a separate delimited ASCII text file, containing quantitative measurements across samples. These files can be created and edited using any text editor. Rows in the data file correspond to features or molecules, whereas a column can either be a measurement column or an identifier column. Data columns correspond to samples or experimental conditions and contain measurements in the form of counts, numerically encoded categorical data, or continuous data. Metadata columns contain feature identifiers and can be numeric or non-numeric. In case feature identifiers are standard identifiers such as Entrez or gene symbol, they can be tagged as such during analysis creation. multiSLIDE assumes that measurements have already undergone necessary pre-processing and transformations. Each column in the data file is required to have a unique column header.

In addition to data files, a separate “sample information” file containing sample attributes (e.g. clinical data or phenotypes) is also required. The sample information file should also be formatted as a delimited ASCII text file, with rows corresponding to samples and all columns, except the first column, corresponding to sample attributes or phenotypes. The first column in this file must contain sample names that are identical to the measurement column headers in data files. There are no restrictions on the number of sample attributes or phenotypes the sample information file can contain. The visualization interface allows users to select a subset of (at most five) phenotypes for visualization. The sample information file can also be used to include additional sample information such as descriptive sample names, replicate names, or time points. The associations between any two omics can be customized by uploading a "Network" file. These files have two columns, one for each omics, with rows containing identifiers that are connected. Users can upload multiple network files into multiSLIDE during visualization and dynamically

choose to apply any one. Even when linker columns are available and multiSLIDE automatically infers the associations, network files can be used to override them.

Querying Mechanisms. multiSLIDE provides three querying mechanisms for users to select a subset of molecules to visualize from the uploaded data files. The first option, keyword-based search, was discussed in the Visualization Workflow section above. The second querying mechanism is enrichment analysis, which requires data files to contain standard gene identifiers. Unlike keyword-based search, this option is useful when the user does not have clearly identified pathways, GO terms, or genes of interest. To perform enrichment analysis, the user specifies a set of parameters, such as the phenotype to use for identifying differentially expressed genes, the significance level, to name a few. The results of enrichment analysis are displayed in a tabular format where the user can click on enriched pathways or GO terms and add them to the visualization. Several options such as size of pathways, the number of differentially expressed genes in the pathway, are also available to filter enrichment analysis results. Finally, the third querying mechanism is to upload a pathway file containing user-specified functionally relevant molecules. This option is for data files that do not contain any standard identifiers. For example, when visualizing metabolomics data, since the map between genes and metabolites has not been fully charted by experimental means, standard shared identifiers cannot be provided. A pathway file contains four columns: *functional group name, data filename, identifier name, and identifier value*. Each row of the pathway file selects molecules with identifiers matching the criteria *identifier name* = *identifier value* from the specified *data filename* and adds them to the *functional group* specified. A single pathway file can therefore be used to describe multiple groups of functionally relevant molecules.

Output Data. multiSLIDE treats both the data analysis and the generated visualizations as resources to be shared and disseminated for collaborative research, a feature that is often missing in existing visualization tools. In multiSLIDE, visualizations in the current view can be rendered in a ready to print form, using the “save visualization” option. This re-renders the current view in a more compact form in a separate window in resolution independent SVG formats. The “print” or “save as” options of the browser can be used to save the contents of this window as high-resolution PDF files (**Supplementary Figure S1d**).

The analysis workspace in multiSLIDE can be saved as “.mslide” files and shared between collaborating parties (**Supplementary Figure S1d**). These files can be loaded back into a different instance of multiSLIDE running on any web browser for continued analysis. The saved

analysis retains all user customizations and data selections, as well as the raw and processed data in JSON format.

Availability

An online version of multiSLIDE is available with multiple demo analyses that users can open with a single click and explore. Users can also upload their own data here. For continued use, users or facility managers are encouraged to install multiSLIDE on their own computers or servers using a pre-built docker image. Links to the online version and the docker image can be found at https://github.com/soumitag/multiSLIDE.

Data sets

Dynamic Transcriptome and Proteome in HeLa cells during ER stress

The first case study explores the whole transcriptome data from Cheng *et al*. (21) which consists of 16704 genes. In their dual-omics time-course experiment, HeLa cells were treated with a sublethal dose of dithiothreitol (DTT) inducing endoplasmic reticulum (ER) stress, and then sampled at eight time points (0, 0.5. 1, 2, 8, 16, 24, and 30 h after treatment) for transcriptome and proteome profiling. In their study, a series of quality filters resulted in the final set of 1237 genes with matched mRNA and protein data.

Prior to visualization in multiSLIDE, we inspected the entire transcriptome data using SLIDE (41), a related tool that we have previously developed for full-scale single-omics data visualization. The log-transformed mRNA and protein measurements of each gene were normalized by subtracting the pre-treatment measurement at 0h. This normalization was performed independently for each replicate, turning the abundance values into (log) ratios to the baseline time point. The normalized data is visualized after applying hierarchical clustering in **Supplementary Figure S3**. In **Supplementary Figure S3a**, search tags highlight genes belonging to the GO term ‘endoplasmic reticulum unfolded protein response’ (green bars to the right of the heatmap).

Proteome and phosphoproteome in CPTAC ovarian cancer cohort

In the second case study, we visualize high-grade serous ovarian carcinoma data from mass spectrometry-based untargeted proteomics and phosphoproteomics experiments conducted by the Clinical Proteomics Tumor Analysis Consortium (CPTAC). Our visualization follows that of Zhang *et al.* (22). In their analysis, the authors retained the 3586 proteins, whose total variance was greater than the technical variance, out of the 9600 that were quantified in all 169 tumors.

For the phosphoproteome data, Zhang *et al*. quantified the relative abundance for a total of 69 tumor samples. Among these 69 tumor samples, 67 of them were quantified at both omics level and were used in the final visualization. Phosphosites with more than 50% of missing data were filtered out and the remaining missing values were imputed using KNNImpute (42). Since ischemia of the TCGA tumor samples was found to be a confounding variable that altered phosphopeptide abundance, sites that were shown to be regulated in ovarian carcinoma (43) were also removed. The abundances were converted to z-scores before visualizing in multiSLIDE. A total of 16718 kinase-substrate interactions were curated from PhosphoSitePlus (31), PhosphoNetworks (32) and a predictive network inference approach (33) to build the kinase-substrate map.

Human plasma proteome and microRNAome associated with insulin resistance.

In the third case study, we visualize plasma proteins and miRNAs between eight insulin resistant (IR) and nine insulin sensitive (IS) subjects (23). The authors analyzed plasma protein and miRNA data for a total of 17 subjects, 8 were insulin resistant and 9 were insulin sensitive. The abundance values of 368 miRNAs and 1499 proteins were log transformed (base 2) and mean centered prior to visualization in multiSLIDE. The miRNAs were mapped to the corresponding miR Family obtained from TargetScan (http://www.targetscan.org) (39). Predicted miR Family to target genes network information from TargetScan were curated externally and uploaded to multiSLIDE to explore miRNA-mediated gene regulation.

Results: Case Studies

Case I: Dynamic transcriptome and proteome in HeLa cells during ER stress

In this case study, we visualize the preprocessed and filtered data from Cheng *et al*. (see Methods section) in multiSLIDE using search keywords chosen from the original paper. During the ER stress, authors characterized acute transcription regulation and delayed translation control for genes involved in the following functions: unfolded protein response, translation attenuation, ER-associated protein degradation, and cellular apoptosis. One of the direct consequences of ER stress is the aggregation of misfolded and unassembled proteins in the organelle. As a survival mechanism to avert the loss of homeostasis, the ER responds by increasing protein folding capacity. Otherwise known as UPR, this mechanism is involved in extensive reprogramming of the transcriptional and translational regulation (44–46). An activated UPR initiates adaptive stress response to regulate downstream effectors, and further, through feedback control, switches on/off transcriptional regulation and protein synthesis to restore ER homeostasis (47, 48). To visualize

this without filtering out any genes, we searched in multiSLIDE using keywords ‘unfolded protein response’ and ‘endoplasmic reticulum’ to retrieve all related pathways and GO terms, and their genes. Genes for pathways and GO terms selected from the search results were retrieved and visualized for both transcriptome and proteome data.

In addition, prior to visualization using multiSLIDE, the 1237 mRNA were also visualized in the context of the whole transcriptome data (N=16704) using a related visualization tool SLIDE (40) (see Methods section and **Supplementary Figure S3**). This global view shows the three phases of ER stress response characterized by Cheng *et al.*: early phase (< 2 h), intermediate phase (2 - 8 h), and late phase (> 8 h). The transcriptome regulation suggests a spike-like pattern in the transition from the early phase to the intermediate phase of the response, peaking in the intermediate phase before returning to original levels in the late phase.

In **Figure 2a**, we show heatmap visualizations of the selected UPR and ER stress-related genes from the dual-omics data in multiSLIDE. We clustered the genes by hierarchical clustering of mRNA data with Euclidean distance and average linkage, which automatically synchronizes the order of display in the protein data. The lines between the two omics data connect mRNA and protein molecules from the same genes. In the web interface, the user is able to click on specific genes to highlight their corresponding molecules in the other omics data. The user can rearrange this visualization by performing further clustering on the protein data (**Figure 2b**), generating an “independent” clustering outcome. As the connecting lines follow the original map throughout these operations, the user can easily track the concordance and discordance of mRNA and protein expression patterns across the samples (time points here). These dynamically changing visualization instances clearly show two key observations: (i) the time course patterns in these selected genes are highly consistent between replicates, and (ii) the time course patterns are discrepant for many genes, and they highlight a 2-hour time gap in response time between the transcripts and the proteins, suggesting these HeLa cells underwent considerable cellular reorganization at 2 hours post treatment and subsequent control of protein translation.

For the genes reported by the keyword searches, the independent clustering of both data show that a multitude of genes are transcriptionally regulated as early as 0.5h. Not surprisingly, the most pronounced gene expression response was observed for the heat shock protein family A (Hsp70) member 5 (HSPA5/GRP78/BiP), the master sensor of misfolded proteins in ER (46), as well as heat shock protein 90 beta family member 1 (HSP90B1), and protein disulfide isomerases (PDIs). In contrast to the transcription regulation of these genes, the corresponding protein abundances do not increase until after 2 hours (top portion of **Figure 2a**), which

corresponds to the time point until which the cells were under cell cycle arrest, as verified by the authors.

Jointly visualizing the mRNA and protein expression data in a time-course dependent manner helps the user to interpret sophisticated response stages during UPR. Clustering genes at the mRNA level and applying the same ordering at the protein level helped visualize whether the clusters propagate between the two omics levels. In this regard, our visualization recapitulated the well-characterized ER stress response in mammalian cells.

Case II: Proteome and phosphoproteome visualization in an ovarian cancer cohort

We next visualize high-grade serous ovarian carcinoma data (22) from mass spectrometry-based untargeted proteomics and phosphoproteomics experiments conducted by the Clinical Proteomics Tumor Analysis Consortium (CPTAC). A total of 67 tumor samples with both proteomic data (3329 unique proteins) and phosphoproteomic data (5746 phosphosites) were visualized in multiSLIDE (see **Methods** section for details).

In this example, we demonstrate the custom network feature of multiSLIDE. By connecting the proteome data (protein abundance) with the phosphoproteome data (protein modification level at individual residues), we pursue two visualization objectives. First, we visualize the relationships between abundance of proteins along with the site-specific phosphorylation level in the same proteins. As the protein identifiers are present in both datasets, multiSLIDE automatically derives the relationship. Second, we use multiSLIDE to simultaneously visualize the kinase proteins from the proteomics data and the substrate sites from the phosphoproteome data, with lines connecting known kinase-substrate pairs. In this instance, we uploaded the ‘custom’ network from externally curated (31–33) kinase-substrate map to the tool as a user (see **Methods** section for details).

We remark that this data example is different from the first example in terms of distinct granularity of identifiers between the two omics data. In the previous example of mRNA and protein data, each gene appeared at both molecular levels, and thus there was a one-to-one mapping between the omics data sets. By contrast, the proteome and phosphoproteome data are at different granularities - multiple phosphorylation sites reside in a protein sequence. This creates one-to-many mapping between the proteome and phosphoproteome.

In the original analysis of the data by Zhang *et al.*, the authors identified five proteome-based molecular subtypes: differentiated, immunoreactive, proliferative, mesenchymal and stromal, with enrichment of distinct pathways in the discriminating protein signatures. In multiSLIDE, we initiate visualization by searching keywords such as ‘DNA replication’, ‘cell-cell communications’, and ‘complement cascade’, corresponding to the authors’ enriched pathways

(22). Applying one-way ANOVA (p-value <= 0.05) and a cutoff of 5% for multiple testing correction (Benjamini-Hochberg procedure, built-in feature), there were a total of 610 proteins and 490 phosphorylation sites (p-sites) that are statistically significant in comparisons of the subtypes (**Supplementary Figure S4**).

Next, we looked more closely at the GO term 'extracellular matrix' (ECM) alone. Using one-way ANOVA (p-value <= 0.05, FDR cutoff of 10%), we found 155 proteins and 116 phosphosites that are differentially expressed (**Supplementary Figure S5**). Unsynchronized, independent hierarchical clustering of the proteins and the phosphosites, using Euclidean distance and average linkage, shows that a number of ECM proteins are elevated in the mesenchymal and stromal subtypes. Further, a subset of ECM proteins is dominantly upregulated in the stromal subtype and multiple ECM proteins, including plectin (PLEC), lamin A/C (LMNA), filamin A (FLNA) and vimentin (VIM), and they have multiple sites that are phosphorylated. Vimentin is an important marker for the epithelial-mesenchymal transition in tissues (EMT), a phenomenon where cells undergo transition from epithelial to mesenchymal phenotype, ultimately leading to cancer metastasis (49). The visualization immediately shows that the phosphosites for the PLEC, LMNA, FLNA, VIM also show consistent subtype specificity in the mesenchymal and stromal subtypes, suggesting possible involvement of protein modification in this process. Studies have shown that vimentin, a type III intermediate filament (IF) protein, is hyperphosphorylated during mitosis by serine/threonine protein kinases involved in cell-cycle which promotes the disassembly of its filamentous structure (50, 51). Filamin A, the actin binding protein, is also phosphorylated at multiple sites by different protein kinases. To understand kinase-dependent phosphorylation pathways, we next visualize the kinases present in the proteomic data jointly with the substrates in the phosphoproteomic data.

The human kinome consists of 518 kinases (52), among which 83 were present in the proteome dataset. In our curated kinase-substrate map, we found these 83 kinases phosphorylates 8269 substrates. Among these 8269 substrates 454 were present in the phosphoproteome data. In multiSLIDE, users can select molecules to visualize by using the search functionality as described before, or through enrichment analysis (see Methods section), or by uploading customized subsets of molecules. Here, using the third option we jointly visualized 83 protein kinases and 454 substrates (**Supplementary Figure S6**). At the protein level, the calcium/calmodulin-dependent protein kinases (CAMK2B, CAMK2G, CAMK2D and CAMK2A) are upregulated in mesenchymal and stromal subtypes, whereas the CMGC kinases, consisting of cyclin-dependent kinases (CDK1, CDK2 and CDK11A) and glycogen synthase kinases (GSK3A and GSK3B) are upregulated in the proliferative subtype. These results affirm the widely

known role of CDK1 and CDK2 in orchestrating mitotic progression, a phase in the cell cycle process during which protein phosphorylation is also known to peak (53).

To further investigate subtype-specific protein kinase activity, we calculated the Pearson correlation coefficients between the subtype-specific levels of kinase abundance and substrate site-level phosphorylation, outside of multiSLIDE. **Supplementary Figure S7** shows the histograms of calculated correlation coefficients for each subtype. The subtypes 'immunoreactive', 'proliferative' and 'stromal' have relatively greater numbers of highly positively correlated (>=0.8) kinase-phosphosite pairs. In the 'proliferative' subtype, among these kinase - substrate phosphosite pairs, those pairs upregulated in both omics were visualized in multiSLIDE (**Figure 3**), mimicking the kinase-substrate enrichment analysis (54). In addition, the Uniform Manifold Approximation and Projection (UMAP) visualization of the entire proteomics and phosphoproteomics data, performed outside multiSLIDE and shown in **Figure 3b** and **3c** respectively, also revealed stratification of the proliferative subtype patients, likely driven by the portion of the data visualized above (55).

Here, we uploaded the kinase-substrate relationships into multiSLIDE using the network upload feature, which allows users to visualize externally created inter-omics connections. The lines from CDK1 and CDK2 point to all their known substrates. We see that most of the substrates of CDK1 and CDK2 show elevated phosphorylation levels in the 'proliferative' and the 'immunoreactive' subtypes. These proteins are involved in the G1/S phase transition which is known to be initiated by cyclin dependent kinases (56). These examples showcase the different querying mechanisms available in multiSLIDE, giving users the flexibility to visualize any subset of the data while retaining the relational information between omics data.

Case III: Human plasma proteome and microRNAome associated with insulin resistance

In this last example, we visualize non-matching molecular entities between omics modalities, i.e. plasma proteins and miRNAs between eight insulin resistant (IR) and nine insulin sensitive (IS) subjects (23). miRNAs are small non-coding RNAs that control the fate of target mRNAs through mRNA degradation or translation repression (57). Specifically, by binding to the sequence motifs in the 3' UTR region of the target gene, miRNA diverts the mRNA away from the ribosome and thereby inhibits protein translation, or primes the mRNAs for degradation via deadenylation and decapping (58). The integrative analysis in the original paper incorporates this negative relationship into the search to identify IR-associated plasma proteins and circulating miRNAs that are negatively correlated with the proteins. The authors made the assumptions that an elevated level of a circulating miRNA reflects increasing transcription of the miRNA in the originating donor

tissues (or organ systems) and it would have resulted in reduced secretion of the target protein into the blood as well.

What sets this visualization apart from the previous two examples is that there is no direct link between the mapping identifiers between the two omics modalities (**Figure 4**). The only genome-scale relationship data between proteins and miRNAs are the prediction of miRNA target sites, which can be achieved by a variety of tools (39, 59–61). Using multiSLIDE, we searched keywords 'metabolism', 'inflammatory response', 'glucose transport', and 'lipid homeostasis'. The original query resulted in a long list of 414 proteins and 69 miRNAs (**Supplementary Figure S8**), and thus we performed hypothesis testing with false discovery rate (FDR) control within multiSLIDE (Mann-Whitney U test and Benjamini-Hochberg method). This functionality gives users the flexibility to adjust the size of data for display. This internal filtering produced 63 proteins and 42 miRNAs significantly different between IR and IS subjects (see **Methods**).

In **Figure 4**, although there is variation within the IR and IS groups, hierarchical clustering using correlation distance (1 minus Pearson correlation) and complete linkage function in multiSLIDE reveals two clear clusters of the miRNA data. At the protein level, however, we find that most plasma proteins are higher in the IR subjects than in the IS subjects, forming a homogeneous cluster. This observation prompts a reasonable speculation that the top cluster of 21 miRNAs with reduced circulation in plasma, from let-7a-5p to mir-654-3p, may have to do with increased secretion of the selected 63 proteins at their respective donor tissues, likely from the endocrine organ systems such as liver and adipose tissues.

To further probe whether the proteins are potential targets of translation control by these miRNAs, we integrated the heatmaps in multiSLIDE with the TargetScan map (39). Any known miRNA - target protein is indicated by a connecting line in **Figure 4**. Interestingly, some of the lines showed that higher circulating levels of miR-130a-3p and miR-301b-3p in the IS subjects were associated with lower level of their predicted target gene, low density lipoprotein receptor (LDLR), a type-I transmembrane glycoprotein. It is widely known that LDLR plays a critical role in maintaining cholesterol homeostasis in the blood, and while insulin resistance is defined by abnormal glucose metabolism, its pathogenesis is increasingly being studied in the context of disordered lipid metabolism (62). The negative correlation between the two miRNAs (miR-130a-3p and miR-301b-3p) and the expression of LDLR in IR subjects, evident from **Figure 4**, is therefore interesting. This has also been previously shown, through GWAS meta-analysis (63), where miR-301b was identified as a key to controlling LDL-C uptake by regulating the expression of LDLR. Also, investigations into the role of miR-130a-3p suggest that its overexpression improves insulin sensitivity both *in vitro* and *in vivo* (64). All in all, this example shows that the

ability to integrate externally curated networks in multiSLIDE allows visualizing relationships between mutually exclusive omics data sets.

Discussion

We have described multiSLIDE, a new web-based tool for interactive heatmap-based visualization of multi-omics data. With steady growth in multi-omics experiments, it has become increasingly appealing to develop open-source analysis and visualization tools. Using multiSLIDE's flexible user interface, users can retrieve segments of data through keyword searches to generate visualizations, with the flexibility to control the size and readability of the display contents. Existing tools for multi-omics visualization tended to focus on displaying a small number of genes within publicly available datasets, such as TCGA, or visualize patterns at abstract levels without showing the actual quantitative data. multiSLIDE addresses this gap -- visualizing the relational data along with actual quantitative data (65). We demonstrated how multiSLIDE enables targeted exploration of large multi-omics datasets within the proper biological contexts through three independent examples. Because the architecture of multiSLIDE was built as a web-based tool, the saved contents can be opened in any other computer with a standard web browser. As such, multiSLIDE was designed to treat both the data analysis and visualization as resources to be shared and disseminated for collaborative research, a feature that is often missing in currently available tools.

multiSLIDE has its own limitations. multiSLIDE is primarily a search-driven visualization tool, which requires the users to have prior biological hypotheses. Therefore, it may not be particularly well suited for the unrestricted exploration of whole multi-omics datasets. Such global exploration can be done using other tools such as multiSLIDE's sister tool SLIDE (41). As an alternative querying mechanism, multiSLIDE provides the option to detect differentially expressed genes and perform enrichment analysis to identify key pathways and GO terms enriched in the user's data. From the result of enrichment analysis users can select pathways and GO terms for visualization instead of having to provide search keywords. A full-scale global visualization of the data in SLIDE can be a useful precursor to analyzing the data in multiSLIDE. In addition, multiSLIDE does not provide functionalities to pre-process and normalize data within the tool and expects the user to prepare display-ready data sets prior to using multiSLIDE. This was an unavoidable choice as these preprocessing operations are often better handled by validated domain-specific tools.

Data Availability

The quantitative expression data used as inputs to multiSLIDE are available at: https://github.com/soumitag/multiSLIDE

Funding

This work was supported in part by grants from Singapore Ministry of Education (MOE2016-T2-1-001 and MOE2018-T2-2-058 to H.C.), National Medical Research Council of Singapore (NMRC-CG-M009 to H.C.), and the support by the Institute of Molecular and Cellular Biology, Agency for Science, Technology and Research.

Conflict of Interest

The authors declare that they have no conflict of interests.

References

1. Hasin,Y., Seldin,M. and Lusis,A. (2017) Multi-omics approaches to disease. *Genome Biol.*, **18**, 83.

2. Gao,J., Aksoy,B.A., Dogrusoz,U., Dresdner,G., Gross,B., Onur Sumer,S., Sun,Y., Jacobsen,A., Sinha,R., Larsson,E., *et al.* (2013) Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. *Sci. Signal.*, **6**, l1-pl1.

3. Goldman,M., Craft,B., Hastie,M., Repečka,K., McDade,F., Kamath,A., Banerjee,A., Luo,Y., Rogers,D., Brooks,A.N., *et al.* (2019) The UCSC Xena platform for public and private cancer genomics data visualization and interpretation. 10.1101/326470.

4. Vasaikar,S.V., Straub,P., Wang,J. and Zhang,B. (2018) LinkedOmics: analyzing multi-omics data within and across 32 cancer types. *Nucleic Acids Research*, **46**, D956–D963.

5. Weinstein,J.N., Collisson,E.A., Mills,G.B., Shaw,K.R.M., Ozenberger,B.A., Ellrott,K., Shmulevich,I., Sander,C., Stuart,J.M., Network,C.G.A.R., *et al.* (2013) The cancer genome atlas pan-cancer analysis project. *Nat. Genet.*, **45**, 1113.

6. Curtis,C., Shah,S.P., Chin,S.-F., Turashvili,G., Rueda,O.M., Dunning,M.J., Speed,D., Lynch,A.G., Samarajiwa,S., Yuan,Y., *et al.* (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, **486**, 346–352.

7. Hernández-de-Diego,R., Tarazona,S., Martínez-Mira,C., Balzano-Nogueira,L., Furió-Tarí,P., Pappas,G.J.,Jr and Conesa,A. (2018) PaintOmics 3: a web resource for the pathway analysis and visualization of multi-omics data. *Nucleic Acids Res.*, **46**, W503–W509.

8. King,Z.A., Dräger,A., Ebrahim,A., Sonnenschein,N., Lewis,N.E. and Palsson,B.O. (2015) Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways. *PLoS Comput. Biol.*, **11**, e1004321.

9. Kutmon,M., van Iersel,M.P., Bohler,A., Kelder,T., Nunes,N., Pico,A.R. and Evelo,C.T. (2015) PathVisio 3: an extendable pathway analysis toolbox. *PLoS Comput. Biol.*, **11**, e1004085.

10. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

11. Kuo,T.-C., Tian,T.-F. and Tseng,Y.J. (2013) 3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. *BMC Syst. Biol.*, **7**, 64.

12. Jang,Y., Yu,N., Seo,J., Kim,S. and Lee,S. (2016) MONGKIE: an integrated tool for network analysis and visualization for multi-omics data. *Biol. Direct*, **11**, 10.

13. Ihaka,R. and Gentleman,R. (1996) R: A Language for Data Analysis and Graphics. *null*, **5**, 299–314.

14. Shen,R., Olshen,A.B. and Ladanyi,M. (2009) Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, **25**, 2906–2912.

15. Vaske,C.J., Benz,S.C., Sanborn,J.Z., Earl,D., Szeto,C., Zhu,J., Haussler,D. and Stuart,J.M. (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, **26**, i237-45.

16. Bonnet,E., Calzone,L. and Michoel,T. (2015) Integrative multi-omics module network inference with Lemon-Tree. *PLoS Comput. Biol.*, **11**, e1003983.

17. Argelaguet,R., Velten,B., Arnol,D., Dietrich,S., Zenz,T., Marioni,J.C., Buettner,F., Huber,W. and Stegle,O. (2018) Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.*, **14**, e8124.

18. Koh,H.W.L., Fermin,D., Vogel,C., Choi,K.P., Ewing,R.M. and Choi,H. (2019) iOmicsPASS: network-based integration of multiomics data for predictive subnetwork discovery. *NPJ Syst Biol Appl*, **5**, 22.

19. Gehlenborg,N. and Wong,B. (2012) Heat maps. *Nat. Methods*, **9**, 213.

20. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.*, **95**, 14863–14868.

21. Cheng,Z., Teo,G., Krueger,S., Rock,T.M., Koh,H.W.L., Choi,H. and Vogel,C. (2016) Differential dynamics of the mammalian mRNA and protein expression response to misfolding stress. *Mol. Syst. Biol.*, **12**, 855.

22. Zhang,H., Liu,T., Zhang,Z., Payne,S.H., Zhang,B., McDermott,J.E., Zhou,J.-Y., Petyuk,V.A., Chen,L., Ray,D., *et al.* (2016) Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell*, **166**, 755–765.

23. Choi,H., Koh,H.W.L., Zhou,L., Cheng,H., Loh,T.P., Parvaresh Rizi,E., Toh,S.A., Ronnett,G.V., Huang,B.E. and Khoo,C.M. (2019) Plasma Protein and MicroRNA Biomarkers of Insulin Resistance: A Network-Based Integrative -Omics Analysis. *Front. Physiol.*, **10**, 379.

24. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.*

25. Jiang,C., Xuan,Z., Zhao,F. and Zhang,M.Q. (2007) TRED: a transcriptional regulatory element database, new entries and other development. *Nucleic Acids Res.*, **35**, D137-40.

26. Zheng,G., Tu,K., Yang,Q., Xiong,Y., Wei,C., Xie,L., Zhu,Y. and Li,Y. (2008) ITFP: an integrated platform of mammalian transcription factors. *Bioinformatics*, **24**, 2416–2417.

27. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

28. Neph,S., Vierstra,J., Stergachis,A.B., Reynolds,A.P., Haugen,E., Vernot,B., Thurman,R.E., John,S., Sandstrom,R., Johnson,A.K., *et al.* (2012) An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, **489**, 83–90.

29. Han,H., Shim,H., Shin,D., Shim,J.E., Ko,Y., Shin,J., Kim,H., Cho,A., Kim,E., Lee,T., *et al.* (2015) TRRUST: a reference database of human transcriptional regulatory interactions. *Sci. Rep.*, **5**, 11432.

30. Marbach,D., Lamparter,D., Quon,G., Kellis,M., Kutalik,Z. and Bergmann,S. (2016) Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nat. Methods*, **13**, 366–370.

31. Hornbeck,P.V., Kornhauser,J.M., Tkachev,S., Zhang,B., Skrzypek,E., Murray,B., Latham,V. and Sullivan,M. (2012) PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.*, **40**, D261–D270.

32. Hu,J., Rho,H.-S., Newman,R.H., Zhang,J., Zhu,H. and Qian,J. (2014) PhosphoNetworks: a database for human phosphorylation networks. *Bioinformatics*, **30**, 141–142.

33. Corwin,T., Woodsmith,J., Apelt,F., Fontaine,J.-F., Meierhofer,D., Helmuth,J., Grossmann,A., Andrade-Navarro,M.A., Ballif,B.A. and Stelzl,U. (2017) Defining Human Tyrosine Kinase Phosphorylation Networks Using Yeast as an In Vivo Model Substrate. *Cell Syst*, **5**, 128-139.e4.

34. Carlson,M.R.J., Pagès,H., Arora,S., Obenchain,V. and Morgan,M. (2016) Genomic Annotation Resources in R/Bioconductor. *Methods in Molecular Biology*, 10.1007/978-1-4939-3578-9_4.

35. Griffiths-Jones,S. (2004) The microRNA Registry. *Nucleic Acids Research*, **32**, 109D – 111.

36. Kamburov,A., Pentchev,K., Galicka,H., Wierling,C., Lehrach,H. and Herwig,R. (2011) ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Res.*, **39**, D712-7.

37. Kamburov,A., Stelzl,U., Lehrach,H. and Herwig,R. (2013) The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res.*, **41**, D793-800.

38. Dweep,H. and Gretz,N. (2015) miRWalk2.0: a comprehensive atlas of microRNA-target interactions. *Nat. Methods*, **12**, 697.

39. Agarwal,V., Bell,G.W., Nam,J.-W. and Bartel,D.P. (2015) Predicting effective microRNA target sites in mammalian mRNAs. *Elife*, **4**.

40. Razick,S., Magklaras,G. and Donaldson,I.M. (2008) iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics*, **9**, 405.

41. Ghosh,S., Datta,A., Tan,K. and Choi,H. (2019) SLIDE--a web-based tool for interactive visualization of large-scale--omics data. *Bioinformatics*, **35**, 346–348.

42. Hastie,T., Tibshirani,R., Sherlock,G., Eisen,M., Brown,P. and Botstein,D. (1999) Imputing missing data for gene expression arrays.

43. Mertins,P., Yang,F., Liu,T., Mani,D.R., Petyuk,V.A., Gillette,M.A., Clauser,K.R., Qiao,J.W., Gritsenko,M.A., Moore,R.J., *et al.* (2014) Ischemia in tumors induces early and sustained phosphorylation changes in stress kinase pathways but does not affect global protein levels. *Mol. Cell. Proteomics*, **13**, 1690–1704.

44. Ron,D. and Walter,P. (2007) Signal integration in the endoplasmic reticulum unfolded protein response. *Nat. Rev. Mol. Cell Biol.*, **8**, 519–529.

45. Walter,P. and Ron,D. (2011) The unfolded protein response: from stress pathway to homeostatic regulation. *Science*, **334**, 1081–1086.

46. Rutkowski,D.T., Arnold,S.M., Miller,C.N., Wu,J., Li,J., Gunnison,K.M., Mori,K., Sadighi Akha,A.A., Raden,D. and Kaufman,R.J. (2006) Adaptation to ER stress is mediated by differential stabilities of pro-survival and pro-apoptotic mRNAs and proteins. *PLoS Biol.*, **4**, e374.

47. Schröder,M. and Kaufman,R.J. (2005) ER stress and the unfolded protein response. *Mutat. Res.*, **569**, 29–63.

48. Sano,R. and Reed,J.C. (2013) ER stress-induced cell death mechanisms. *Biochim. Biophys. Acta*, **1833**, 3460–3470.

49. Satelli,A. and Li,S. (2011) Vimentin in cancer and its potential as a molecular target for cancer therapy. *Cellular and Molecular Life Sciences*, **68**, 3033–3046.

50. Bollong,M.J., Pietilä,M., Pearson,A.D., Sarkar,T.R., Ahmad,I., Soundararajan,R., Lyssiotis,C.A., Mani,S.A., Schultz,P.G. and Lairson,L.L. (2017) A vimentin binding small molecule leads to mitotic disruption in mesenchymal cancers. *Proc. Natl. Acad. Sci. U. S. A.*, **114**, E9903–E9912.

51. Yasui,Y., Goto,H., Matsui,S., Manser,E., Lim,L., Nagata Ki and Inagaki,M. (2001) Protein kinases required for segregation of vimentin filaments in mitotic process. *Oncogene*, **20**, 2868–2876.

52. Manning,G. (2002) The Protein Kinase Complement of the Human Genome. *Science*, **298**, 1912–1934.

53. Malumbres,M. (2014) Cyclin-dependent kinases. *Genome Biology*, **15**, 122.

54. Casado,P., Rodriguez-Prados,J.-C., Cosulich,S.C., Guichard,S., Vanhaesebroeck,B., Joel,S. and Cutillas,P.R. (2013) Kinase-substrate enrichment analysis provides insights into the heterogeneity of signaling pathway activation in leukemia cells. *Sci. Signal.*, **6**, rs6.

55. McInnes,L., Healy,J. and Melville,J. (2018) UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv [stat.ML]*.

56. Bertoli,C., Skotheim,J.M. and de Bruin,R.A.M. (2013) Control of cell cycle transcription during G1 and S phases. *Nature Reviews Molecular Cell Biology*, **14**, 518–528.

57. Shyu,A.-B., Wilkinson,M.F. and van Hoof,A. (2008) Messenger RNA regulation: to translate or to degrade. *EMBO J.*, **27**, 471–481.

58. O'Brien,J., Hayder,H., Zayed,Y. and Peng,C. (2018) Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. *Front. Endocrinol.* , **9**, 402.

59. Betel,D., Koppal,A., Agius,P., Sander,C. and Leslie,C. (2010) Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biology*, **11**, R90.

60. Kertesz,M., Iovino,N., Unnerstall,U., Gaul,U. and Segal,E. (2007) The role of site accessibility in microRNA target recognition. *Nature Genetics*, **39**, 1278–1284.

61. Liu,H., Yue,D., Chen,Y., Gao,S.-J. and Huang,Y. (2010) Improving performance of mammalian microRNA target prediction. *BMC Bioinformatics*, **11**, 476.

62. Savage,D.B., Petersen,K.F. and Shulman,G.I. (2007) Disordered Lipid Metabolism and the Pathogenesis of Insulin Resistance. *Physiological Reviews*, **87**, 507–520.

63. Wagschal,A., Najafi-Shoushtari,S.H., Wang,L., Goedeke,L., Sinha,S., deLemos,A.S., Black,J.C., Ramírez,C.M., Li,Y., Tewhey,R., *et al.* (2015) Genome-wide identification of microRNAs regulating cholesterol and triglyceride homeostasis. *Nat. Med.*, **21**, 1290–1297.

64. Xiao,F., Yu,J., Liu,B., Guo,Y., Li,K., Deng,J., Zhang,J., Wang,C., Chen,S., Du,Y., *et al.* (2014) A novel function of microRNA 130a-3p in hepatic insulin sensitivity and liver steatosis. *Diabetes*, **63**, 2631–2642.

65. Vitrinel,B., Koh,H.W.L., Mujgan Kar,F., Maity,S., Rendleman,J., Choi,H. and Vogel,C. (2019) Exploiting Interdata Relationships in Next-generation Proteomics Analysis. *Mol. Cell. Proteomics*, **18**, S5–S14.

Figure Captions

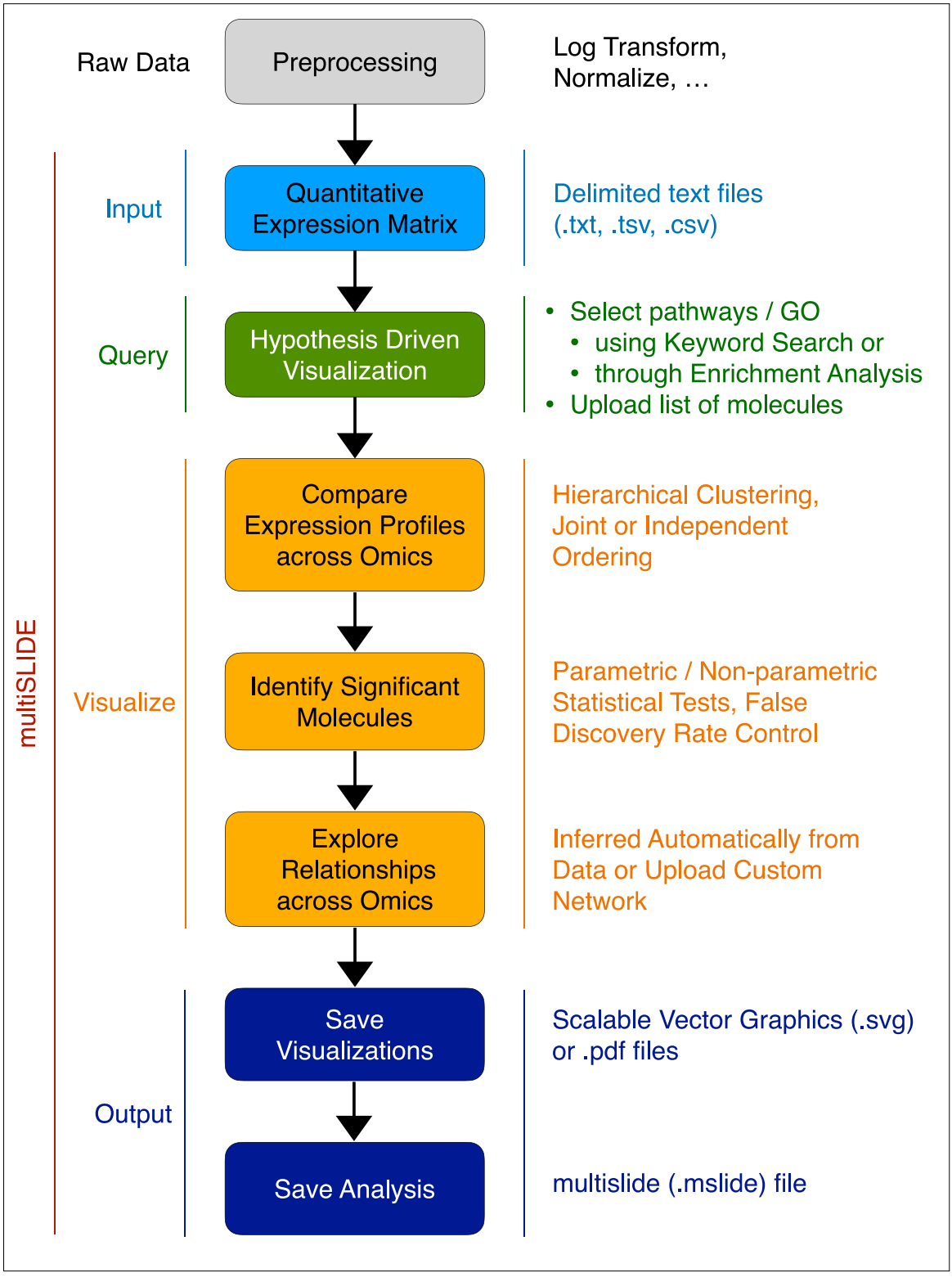
Figure 1: Visualization Workflow of multiSLIDE. Inputs to multiSLIDE are preprocessed quantitative expression profiles, formatted as delimited text files, with a separate file for each omics data. Users can select features to visualize, using keyword-based search, using the 'Upload Pathways' option, or through enrichment analysis. Users can interact with the selected data using the many options for ordering/clustering of molecules and samples, as well as the customizable filtering of molecules based on differential expression levels. Once the exploration of the data reveals interesting patterns, users can save the visualizations as Scalable Vector Graphics (SVG) or PDF files. The analysis workspace can also be saved as a ".mslide" file, retaining all user selections and interactions, for sharing among collaborators. Snapshots of the visualization interface corresponding to this workflow are presented in **Supplementary Figure 1**.

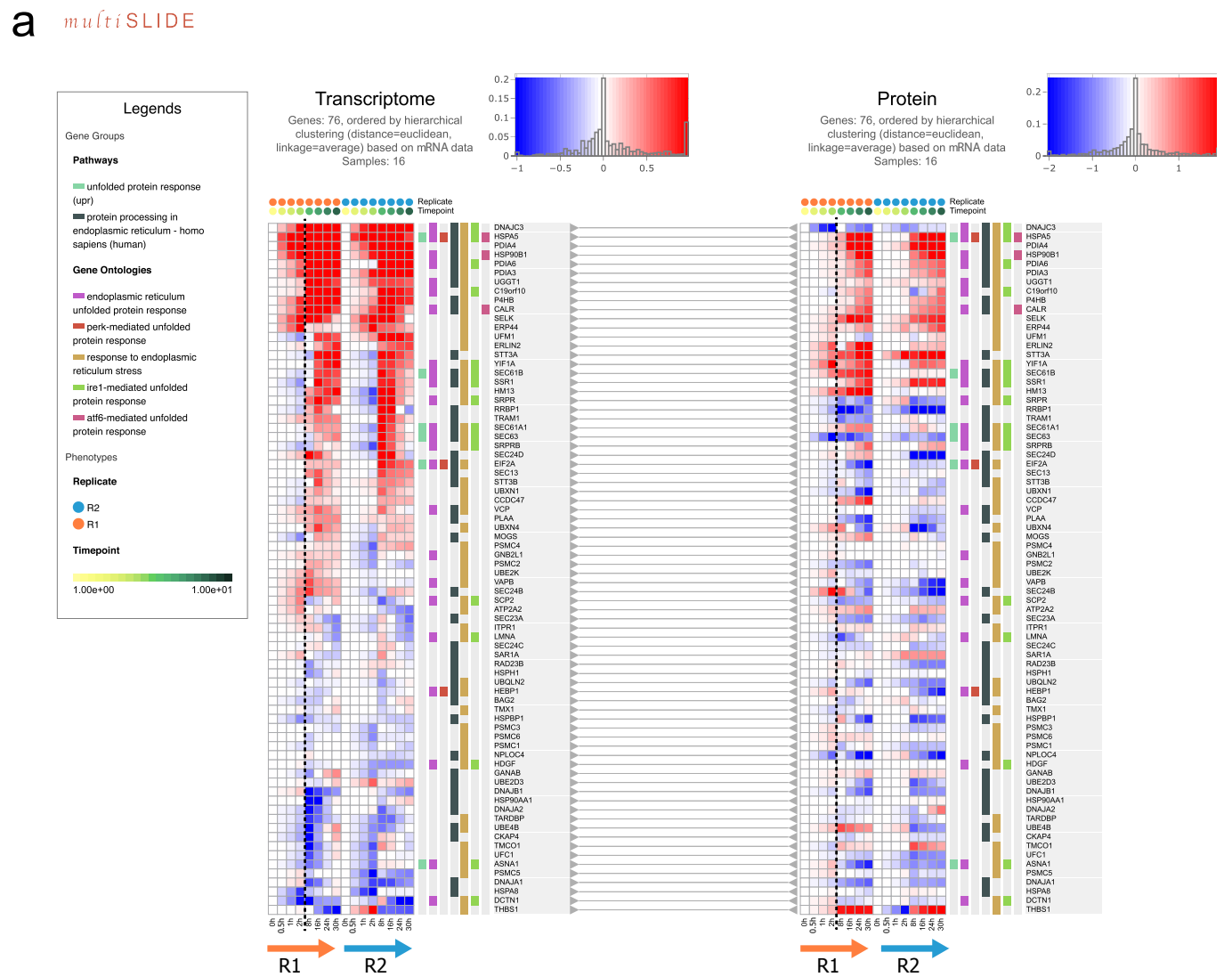
Figure 2: Visualization of unfolded protein response in mammalian cells responding to stress. mRNA and protein-level expressions, across eight time-points (0, 0.5. 1, 2, 8, 16, 24, and 30h after treatment) and two replicates, are jointly visualized in multiSLIDE to understand the dynamics of UPR under ER stress. The Legends panel, on the left, lists the selected GO terms and pathways. The colored tags in vertical tracks alongside the heatmap indicate associations between genes and GO terms/pathways. Panels a and b represent the two modes of visualization, synchronized and independent (unsynchronized), respectively. In the synchronized clustering mode, the same order of genes is applied to both the mRNA and protein levels. In the independent clustering mode, mRNA and protein data are hierarchically clustered independently, using Euclidean distance and complete linkage.

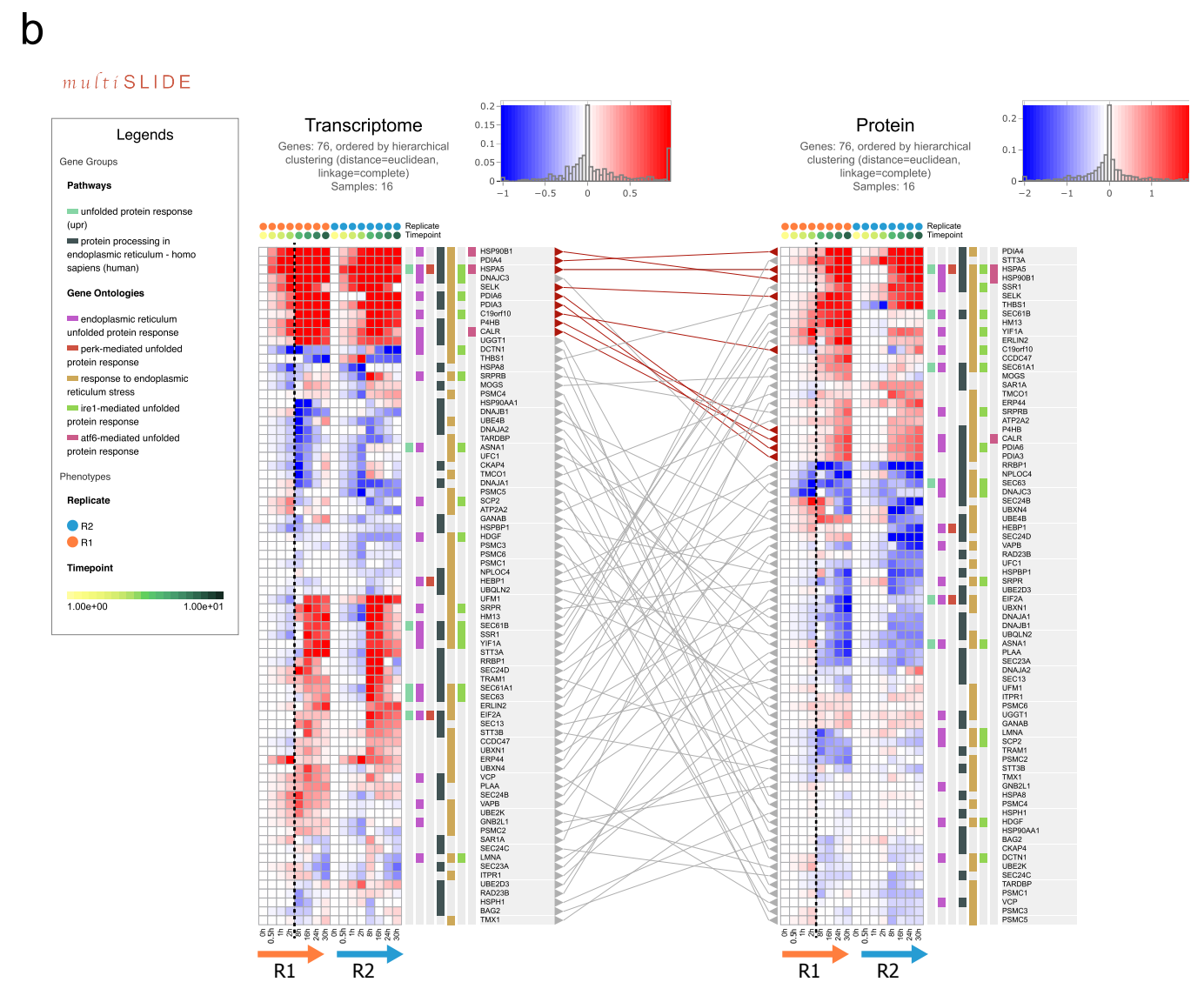
Figure 3: (**a**) Visualization of kinase-substrate relationship with CPTAC Ovarian Cancer Cohort. The figure visualizes the subset of kinases-substrate pairs, which are upregulated in the proliferative subtype. The custom 'Upload' option is used to select the molecules here. Kinase-substrate interactions were curated from PhosphoSitePlus (30), PhosphoNetworks (31), and a predictive network inference approach (32) to build a kinase-substrate map, which was uploaded into multiSLIDE using the upload network feature. The connecting lines show these curated relationships, with the highlighted (brown) lines connecting cyclin-dependent kinases CDK1 and CDK2 with known substrates. **Supplementary Figure S6** visualizes all the kinases-substrate pairs. **(b)** A UMAP visualization of the whole proteomics data for 3329 proteins. The ellipse highlights a cluster of proliferative subtype patients in the protein data. **(c)** A UMAP visualization

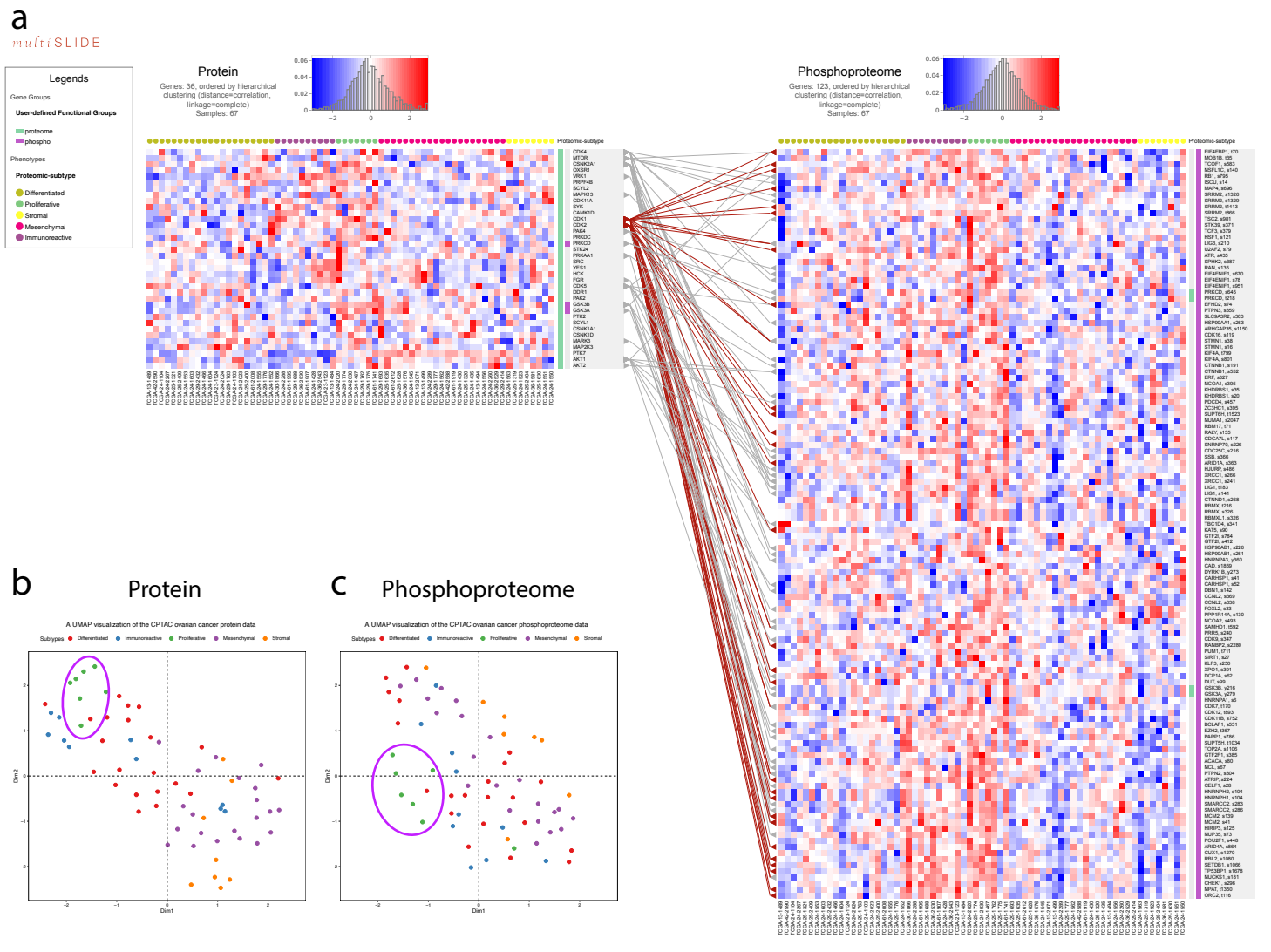
of the whole phosphoproteome data for 5746 phosphosites. The ellipsis highlights a cluster of proliferative subtype patients in the phosphoproteome data.

Figure 4: Visualization of human plasma proteome and microRNA-ome associated with insulin resistance related to the keywords: metabolism, inflammatory response, glucose transport, and lipid homeostasis, with filtering by Mann-Whitney U test (p-value $\leq$ 0.05, FDR 20%). Here, proteins and miRNA are independently clustered using correlation distance (1 minus Pearson correlation) and complete linkage function. The relationships between miRNA family names and their target proteins are extracted from TargetScanMap (38). The original list of 414 proteins and 69 miRNAs, before filtering, is shown in **Supplementary Figure S8**.


Raw Data
Preprocessing
Log Transform, Normalize, …
multiSLIDE
Input
Quantitative Expression Matrix
Delimited text files (.txt, .tsv, .csv)
Query
Hypothesis Driven Visualization
• Select pathways / GO
• using Keyword Search or
• through Enrichment Analysis
• Upload list of molecules
Visualize
Compare Expression Profiles across Omics
Hierarchical Clustering, Joint or Independent Ordering
Identify Significant Molecules
Parametric / Non-parametric Statistical Tests, False Discovery Rate Control
Explore Relationships across Omics
Inferred Automatically from Data or Upload Custom Network
Output
Save Visualizations
Scalable Vector Graphics (.svg) or .pdf files
Save Analysis
multislide (.mslide) file


a
multiSLIDE
Legends
Gene Groups
Pathways
unfolded protein response (upr)
protein processing in endoplasmic reticulum - homo sapiens (human)
Gene Ontologies
endoplasmic reticulum unfolded protein response
perk-mediated unfolded protein response
response to endoplasmic reticulum stress
ire1-mediated unfolded protein response
atf6-mediated unfolded protein response
Phenotypes
Replicate
R2
R1
Timepoint
1.00e+00
1.00e+01
Transcriptome
Genes: 76, ordered by hierarchical clustering (distance=euclidean, linkage=average) based on mRNA data
Samples: 16
Protein
Genes: 76, ordered by hierarchical clustering (distance=euclidean, linkage=average) based on mRNA data
Samples: 16
Replicate
Timepoint
R1
R2


b
multiSLIDE
Legends
Gene Groups
Pathways
unfolded protein response (upr)
protein processing in endoplasmic reticulum - homo sapiens (human)
Gene Ontologies
endoplasmic reticulum unfolded protein response
perk-mediated unfolded protein response
response to endoplasmic reticulum stress
ire1-mediated unfolded protein response
atf6-mediated unfolded protein response
Phenotypes
Replicate
R2
R1
Timepoint
1.00e+00
1.00e+01
Transcriptome
Genes: 76, ordered by hierarchical clustering (distance=euclidean, linkage=complete)
Samples: 16
Protein
Genes: 76, ordered by hierarchical clustering (distance=euclidean, linkage=complete)
Samples: 16
Replicate
Timepoint
R1
R2


a
multiSLIDE
Legends
Gene Groups
User-defined Functional Groups
proteome
phospho
Phenotypes
Proteomic-subtype
Differentiated
Proliferative
Stromal
Mesenchymal
Immunoreactive
Protein
Genes: 36, ordered by hierarchical clustering (distance=correlation, linkage=complete)
Samples: 67
Phosphoproteome
Genes: 123, ordered by hierarchical clustering (distance=correlation, linkage=complete)
Samples: 67
Proteomic-subtype
b
Protein
A UMAP visualization of the CPTAC ovarian cancer protein data
Subtypes
Differentiated
Immunoreactive
Proliferative
Mesenchymal
Stromal
Dim1
Dim2
c
Phosphoproteome
A UMAP visualization of the CPTAC ovarian cancer phosphoproteome data

multiSLIDE

Legends

Gene Groups

Pathways

- metabolism
- metabolism of amino acids and derivatives
- metabolism of lipids and lipoproteins
- srebf and mir33 in cholesterol and lipid homeostasis
- inflammatory response pathway
- glycolysis / gluconeogenesis - homo sapiens (human)
- glucose transport

Gene Ontologies

- cholesterol homeostasis
- inflammatory response

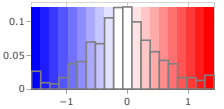
Phenotypes

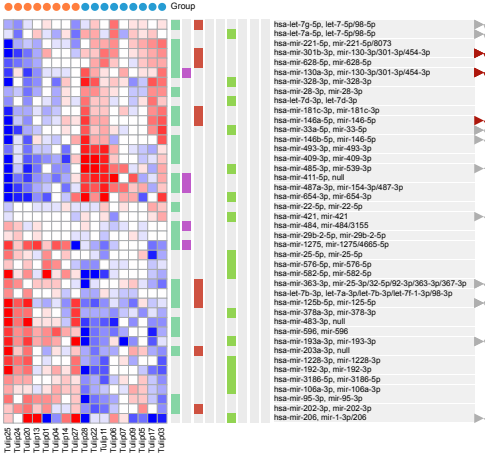
Group

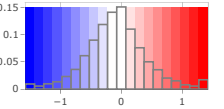
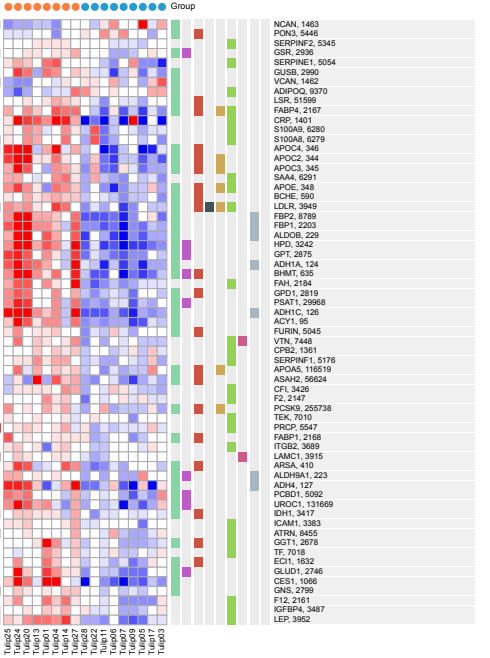
- IR
- IS

microRNA Expression

Genes: 42 (p-value=0.05, FDR=0.2), ordered by hierarchical clustering (distance=correlation, linkage=complete)
Samples: 17





Protein

Genes: 63 (p-value=0.05, FDR=0.2), ordered by hierarchical clustering (distance=correlation, linkage=complete)
Samples: 17