multiSLIDE: a query-driven heatmap visualization tool for multi-omics data

Soumita Ghosh, Abhik Datta, Hyungwon Choi

1. Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
2. Saw Swee Hock School of Public Health, National University of Singapore, Singapore
3. Institute of Molecular and Cell Biology, Agency for Science, Technology, Research, Singapore
4. Centre for BioImaging Sciences, National University of Singapore, Singapore
5. Department of Biological Sciences, National University of Singapore, Singapore

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

Running title: multiSLIDE: a web-based multi-omics visualization tool
Abbreviations

ANOVA: analysis of variance
API: application programming interface
CPDB: Consensus Path DB
CPTAC: Clinical Proteomic Tumor Analysis Consortium
CSS: Cascading Style Sheet
DTT: dithiothreitol
EGFR: epidermal growth factor receptor
ER: endoplasmic reticulum
ERBB2: erb-b2 receptor tyrosine kinase 2
ESR1: estrogen receptor alpha
GO: Gene Ontology
HGNC: HUGO Gene Nomenclature Committee
JSON: JavaScript Object Notation
PDI: protein disulfide isomerase
PPI: protein-protein interaction
RefSeq: Reference Sequence
SVG: scalable vector graphics
TCGA: The Cancer Genome Atlas
TF: transcription factor
TS: Type Script
UPR: unfolded protein response
Abstract

We present multiSLIDE, an open-source tool for query-driven visualization of quantitative single- or multi-omics data. Using pathways and networks as the basis for data linkage, multiSLIDE provides an interactive platform for querying the multi-omics data by genes, pathways, and intermolecular relationships. Representing individual -omics levels as separate heatmaps, multiSLIDE visualizes quantitative data for selected genes at all omics levels in a single snapshot. The tool also provides functionalities to arrange data both ways, by their phenotypic characteristics and by their molecular interactions or co-membership to common pathways. Built-in statistical tests and clustering methods provide display subsets of interesting genes or rearrange the genes based on expression patterns. All visualization panels are fully customizable, and both the graphics and the analysis workspace can be saved and shared between collaborating parties.

We demonstrate the utility of multiSLIDE through two example studies. First, with a time-course data of HeLa cells, subjected to dithiothreitol induced endoplasmic reticulum stress, we visualized different stages of unfolded protein response and identified temporal patterns of gene expression response at the mRNA and protein levels. Second, through joint visualization of mRNA and protein expression of TCGA/CPTAC Invasive Breast Carcinoma data, we explored the estrogen receptor $\alpha$ regulon and prioritized clusters of genes associated with PAM50 basal-like subtype.

Keywords: Visualization, multi-omics
Introduction

Visualization is an important tool for understanding any data. Large-scale data sets from high-throughput -omics platforms, such as massively parallel sequencing or mass spectrometry, often require significant reduction of data using statistical filters or abstraction via a projection of data into a lower dimensional space to facilitate interpretation. Although data reduction is unavoidable for effective presentation of data, our dependence on these filters blinds us from other intrinsic features that fail to pass the filter. Therefore, there is need for visualization tools that enable detailed exploration of full data set prior to data filtering. Easy-to-use, versatile visualization tools are all the more in increasing demand, especially for emerging multi-omics data sets. With the leading -omics technologies maturing, we have learned that different omics-level measurements often lead to concordant and discordant observations from the same biological samples (1,2), although ultimately the most relevant data for biological phenomenon is the expression of the final gene products, i.e. proteins (3).

Indeed, there already exists a wealth of bioinformatics tools for multi-omics data visualization. Supplementary Table S1 gives a summary of existing tools in four categories: open-source data portals, networks-based tools, pathway-based tools, and heatmap-based omics integration tools. Open-source, data-rich web resources such as cBioPortal (4), UCSC Xena (5), and LinkedOmics (6) have made it possible for systematic exploration and visualization of public cancer-omics datasets. However, most of these tools, apart from UCSC Xena, do not visualize user’s own data. Moreover, visualization in these data portals are predominantly single gene-based or key cancer driver gene-based exploration modules.

In pathway-based visualizations (7–11), pathway diagrams are augmented with quantitative data to enable users to gain more meaningful insights for phenotypic variations at the level of gene groups. However, it is challenging to visualize interconnections between omics layers within the pathway-based visualization modules. Network-based visualization approaches (12–21) are popular alternatives for visualization of complex interconnectivities between biomolecules, yet visualized networks often run into the “hairball” problem even with a few hundred nodes (22,23).
Integrated heatmap-based visualizations (24,25) can help disentangle the crowding of network diagrams without overwhelming the graphical interface. While heatmaps have not been the primary mode of representation in multi-omics data visualization tools, many tools often include heatmaps as an additional visualization. For instance, PaintOmics3 supplements their pathway diagrams with additional heatmaps displaying omics expression profiles, where the interactivity and data handling capacity of these heatmaps are usually restricted even for moderately sized data sets, limiting its utility.

Motivated by the difficulties presented in the existing tools, we developed multiSLIDE, an interactive data visualization tool for easy exploration of multi-omics data. In multiSLIDE, quantitative multi-omics data are visualized for genes in specific pathways or gene ontology (GO) categories preselected by the user, simultaneously at different molecular levels. The tool displays the abundance measurements of DNA copy number, mRNA transcripts, and proteins in functional groups (e.g. pathways and GO terms) selected by the user in aligned panels of heatmaps, synchronizing them by gene identifiers and sample names.

multiSLIDE also integrates biological networks in queries, including transcription factor (TF) regulatory networks and protein-protein interaction networks, which capture the dependencies across and within molecular levels, respectively. These networks can be quickly queried for genes of interest, which allows the user to add other genes interacting with them to the current heatmap visualization in real time. At the same time, sample specific phenotype labels are visualized as side bars in the heatmaps. This integrated visualization enables exploring molecular profiles linked with phenotypes at various omics levels concurrently. The visualizations in multiSLIDE are fully customizable, with multiple ordering options for samples and genes, such as hierarchical clustering, multi-level phenotypic sorting, and statistical filtering.

We demonstrate multiSLIDE using two publicly available multi-omics datasets. In the first case study, we visualize the time-course mRNA and protein expression data in HeLa cells subjected to endoplasmic reticulum (ER) stress (26). The joint analysis reveals time-dependent patterns of unfolded protein response (UPR), distinctly regulated at the mRNA and protein levels. In the second case study, we visualized mRNA and protein data for 73 tumors from TCGA/CPTAC invasive ductal breast carcinoma
(TCGA-BRCA) cohort, aiming to visualize key hormone receptor ESR1 and growth factor EGFR proteins in the four intrinsic molecular subtypes (27).
Experimental Procedures

The Visualization Workflow

Figure 1A shows the web-based visualization interface of multiSLIDE. In multiSLIDE, data analysis begins with the user’s selection of pathways, GO terms, or individual genes. multiSLIDE provides an intuitive keyword-based search syntax for searching multiple pathways, GO terms, and genes. The user selects relevant genes and gene groups from the search results and those genes are visualized with clicks on the group names, as shown in Figure 1B. In addition, the user can choose to add network neighbors of a target gene on protein-protein interaction (PPI) and TF regulatory networks via a network neighborhood search, all enabled by a simple right-click on the gene of interest. Selected network neighbors are added to the visualization in real time (Figure 1C). Alternatively, the user can visualize genes in GO terms or pathways, with the side tracks indicating the membership of genes to the functional groups (side bars on the left side of heatmaps, Figure 1A).

The scales and dynamic ranges of detection and quantification techniques can be inconsistent for different -omics data. Therefore, we made the graphical parameters customizable in each omics data separately. The user can set a suitable binning range, select different color schemes differently in individual omics panels (see heatmap settings panel, Figure 1D). Settings applicable to all heatmaps, such as zoom (or resolution) and orientation of heatmaps, are applied to all heatmaps simultaneously using the global settings panels (Figures 1E and 1F). The heatmaps in different panels (different omics data) are anchored by the genes and samples so as to enable simultaneous scrolling at any instant, using the scrolling panel (Figure 1G).

multiSLIDE has no restrictions on the amount of data that can be viewed in a single snapshot, although the screen size of the user’s computer has natural limits to which the visualization can be effective. As different systems and browsers have variable computing capabilities, this choice is left to the user. Using the layout panel (Figure 1E), 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 multiple sorting, clustering, and filtering methods built in to help users discover patterns in the data. Interesting genes can be difficult 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, however, previously undetectable structures in the data can become apparent. In multiSLIDE, features can be sorted by gene groups, based on significance level in differential expression analysis, or based on hierarchical clustering. Samples can be ordered by a combination of phenotypes or based on hierarchical clustering. The hierarchical clustering can be further modified by selecting different linkage functions, distance metrics, and leaf ordering schemes, although the dendrograms are not visualized.

The user can also remove statistically non-significant genes from the panels using differential expression analysis. In large pathways and networks, a substantial number of genes are not differentially expressed between sample groups, and removal of these genes may improve the visualization clarity. The statistical test for differential expression analysis depends on the type of the selected phenotype. For binary, categorical, and continuous phenotypes, two-sample t-test, analysis of variance (ANOVA), and linear least squares regression are performed, respectively. multiSLIDE automatically classifies phenotypes into one of the three variable types.

The changes in the customizations are reflected in the visualizations in real time. The user can create curated feature lists by adding individual genes or gene groups from various heatmap panels. The “User List” panel in the far-right side of Figure 1A shows an example gene list curated by the user. These gene subsets can be easily re-visualized within the analysis with a single click. Feature lists can also be created by uploading a list of gene identifiers as a delimited text file. The user can also maintain multiple save points in multiSLIDE, to which they can go back to and restart the analysis. The visualizations can be saved as high-resolution PDF files, and the analysis workspaces with customized heatmaps can be saved as “.mslide” files. The feature lists can be also exported as text files. The saved analysis workspaces can later be loaded back into multiSLIDE for continued analysis, or simply to revisit the analysis steps leading to the current visualizations. This functionality makes multiSLIDE a useful tool to share analysis results in collaborative projects.
Software Design

Databases: Pathways, GO terms and Molecular Networks. Comprehensive genome-wide annotations and gene ontology databases for mouse and human genes were extracted using libraries in R Bioconductor (28, 29). Data is highly structured in these R packages and are routinely used by bioinformaticians to analyze their data. multiSLIDE recognizes Entrez, HUGO Gene Nomenclature Committee (HGNC) Gene Symbols, ENSEMBL identifiers, NCBI Reference Sequence (RefSeq) identifiers and UniProt identifiers. Internally within multiSLIDE, the gene identifiers are standardized by conversion to Entrez.

Comprehensive biological pathways were obtained from ConsensusPathDB (CPDB) (http://cpdb.molgen.mpg.de/) (30, 31). Validated miRNA - target interactions on pathways and GO were extracted from miRWalk2.0 (http://zmf.umm.uni-heidelberg.de/apps/zmf/mirwalk2/) (32).

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, have been integrated in multiSLIDE. The networks come from diverse sources and the interactions are either experimentally validated or computationally predicted. multiSLIDE also integrates Human Transcription Factor (TF) - targets network information from additional databases: TRED (33), ITFP (34), ENCODE (35), Neph2012 (36), TRRUST (37), Marbach2016 (38). Mouse Transcription Factor (TF) - targets network information was obtained directly from TRRUST. Physical interactions between proteins was sourced from iRefIndex (http://irefindex.org/wiki/index.php?title=iRefIndex) (39), which indexes protein-protein interaction networks from a number of databases. multiSLIDE also includes miRNA-mediated gene regulation information from TargetScan (http://www.targetscan.org) (40), a database housing mostly predicted targets of miRNAs.

Software Implementation. multiSLIDE is built on a three-tier, client-server architecture separating the core computational logic, user interface and data storage tiers. A multi-tier architecture has the flexibility of
developing and extending each tier independently without affecting the other, provided that the tiers communicate with each other using predefined application programming interfaces (APIs). Moreover, such architectures facilitate highly parallel communication. Similar to SLIDE, a tool we had developed for online visualization of single -omics data (41), multiSLIDE is available online and can also be used as a standalone software. Due to its modular design, multiSLIDE can also easily scale to distributed multi-node environments.

Supplementary Figure S1 shows a schematic view of multiSLIDE’s software architecture. The server side of multiSLIDE consists of an HTTP server and a database server. The client can be any modern web browser. The HTTP server hosts the analytics module, which is the main computation engine, as well as the graphics server. The analytics module carries out the bulk of the computation. The HTTP server and the browser communicate via highly optimized JavaScript Object Notation (JSON) objects. The data tier, implemented using MongoDB, manages the physical storage of all curated gene annotation, regulatory networks, biological pathways and Gene Ontology (GO) tables. In multiSLIDE, individual components within each tier are also highly compartmentalized. At the client, the data and presentation layers are also decoupled. Layouts and graphics can therefore be altered without the need for fetching the data again from the server. User interactions and styles were implemented using TypeScript (TS) and Cascading Style Sheets (CSS). The visualizations are rendered using resolution independent Scalable Vector Graphics (SVG).

A key design philosophy in multiSLIDE is to visualize only user queried genes (molecules). As a result of the extensive user interactions available in multiLSIDE, there is frequent communication between client and server. At the same time, owing to multiple -omics datasets, the server has to manage a much larger amount of data. In multiSLIDE, lazy execution combined with memoization is extensively used to balance the server-side memory footprint and response times. The computation intensive modules are developed by leveraging the advantages of Java and Python. Aggressive caching in the browser and using delta loads for data transfer gives multiSLIDE a desktop-like user experience.
Input Data Format. In multiSLIDE, the user creates an analysis by uploading a set of required input files. Inputs to multiSLIDE should be delimited ASCII text files (one for each -omics dataset), containing quantitative measurements across samples. These files can be created and edited using any text editor and there is no restriction on the number of input files that can be uploaded for visualization. Measurements can be counts, categorical data or continuous data that have already undergone standard pre-processing and transformations. The databases integrated into multiSLIDE contain functionally annotated genes assigned to pathways and GO terms. To fully utilize multiSLIDE’s querying capabilities, individual -omics input files should have at least one column with standard gene identifiers available in multiSLIDE. Also, high throughput -omics data, such as DNA methylation, are molecular features at specific genomic coordinates. Since multiSLIDE’s search capabilities are gene-based, for omics data with no designated gene identifiers (e.g. sequence variants, CpG islands in DNA methylation), genomic coordinates of the molecular features have to be mapped to the nearest gene and labeled as such in the current version of the software.

In addition, a separate file containing sample attributes (e.g. clinical data) is also required, formatted as a separate delimited ASCII text file. The information in this file should map the samples in -omics data files to their corresponding phenotype information. The attributes file may include optional sample information such as descriptive sample names, replicate names, and time points.

Data preprocessing for TCGA/CPTAC breast cancer data

Genes that had fewer than 80% samples expressed in Luminal A, Luminal B, and basal-like subtypes in both -omics level were removed. For the HER2-enriched subtype, genes with fewer than ten samples expressed at either -omics levels were removed from the corresponding -omics level. For the remainder of genes, weighted $k$-nearest neighbor method, where $k = 10$, (42) was used to impute the missing data prior to importing into multiSLIDE. The mRNA expression values were log-transformed (base 2) and each gene was centered by the mean of their expression levels in multiSLIDE visualization. For the protein level data, the iTRAQ ratios were also median centered for visualization.
Results

We present two case studies using previously published datasets to illustrate multiSLIDE's features. The first case study explores multiple pathways and GO terms simultaneously in a time-course multi-omics dataset. The second case study illustrates exploration of the TF regulatory networks and statistical filtering-based gene prioritization functionalities on a paired transcriptomic and proteomic TCGA/CPTAC breast cancer dataset. In addition, we also demonstrate the exploration of the PPI network around a receptor tyrosine kinase HER2/ERBB2 on the same dataset in Supplementary Figure S2.

UPR reveals independent gene expression regulation at the mRNA and protein level

Using multiSLIDE, we first visualize the time-course mRNA and protein expression data of HeLa cells in response to dithiothreitol (DTT) treatment, which induces ER stress (26). In the study, cells were sampled at eight time points (0, 0.5, 1, 2, 8, 16, 24, and 30h) and their transcriptome and proteome were measured. Both -omics data were normalized by dividing the measurements at post-treatment time points by their respective measurement at 0h. During the ER stress, we expect to observe complex signaling cascades involved in UPR, translation attenuation, ER-associated protein degradation, and cellular apoptosis (43–46).

Before we explore this multi-omics data, we first visualize the entire data sets separately using the SLIDE tool, another related tool that we had previously developed for full-scale single -omics data visualization (41). We applied hierarchical clustering with Euclidean distance and complete linkage to the whole transcriptome data comprising of 16,704 genes in SLIDE (Supplementary Figure S3A). The global view of the clustering profiles in Supplementary Figure S3A shows three phases of the ER stress response characterized by Cheng et al: early phase (< 2 h), intermediate phase (2 - 8 h), and late phase (> 8 h). The whole mRNA expression regulation suggested a spike-like pattern in the transition from the early phase to the intermediate phase, peaking in the intermediate phase before returning to original levels in the late phase.
One of the direct consequences of ER stress is the aggregation of misfolded, unassembled proteins in the organelle. As a survival mechanism to avert the loss of homeostasis, the ER responds by increasing the protein folding capacity. This activated pro-survival cellular mechanism, otherwise known as the UPR, is involved in extensive reprogramming of the transcriptional and translational regulation (43, 47, 48). First, to identify action patterns of UPR related genes, the global view in Supplementary Figure S3A was tagged to highlight genes belonging to the GO term ‘endoplasmic reticulum unfolded protein response’ (green bars to the right of the heatmap). As an adaptive response, a hallmark of UPR is to reduce ER stress and restore homeostasis by the coordinated transcriptional upregulation of ER chaperones and protein folding enzymes. The initial visual inspection of the whole mRNA data confirmed the upward expression trends of the ER chaperones, heat shock protein family A (Hsp70) member 5 (HSP5A/GRP78/BiP) and folding chaperones such as protein disulfide isomerase (PDI) family genes (Supplementary Figure S3B).

Cheng et al. also analyzed the dynamics of 1,237 mRNA/protein pairs that passed the filtering for missing data and noise. Supplementary Figure S3C shows the protein abundances of these 1,237 proteins in SLIDE. Most UPR related proteins expressed upward trends in the proteomics data, similar to the mRNA expression regulation, but with a delayed temporal response. The global heatmap clearly shows that this up-regulation persisted even at the late phase of the stress response.

To investigate these patterns in the transcriptomic and proteomic data simultaneously, we loaded the data sets into multiSLIDE. Hierarchical clustering of UPR-related genes at the mRNA level in multiSLIDE reveals a cluster of upregulated genes. The master sensor of misfolded proteins in the ER, HSPA5 (also known as GRP78 or BiP), is a member of the heat shock protein 70 family, which is upregulated at both mRNA and protein level. In Figure 2, the mRNA level data shows early up-regulation of HSPA5, at 0.5 h of stress induction. HSPA5 regulates the activation of ER stress transducers, including PERK (protein kinase R-like ER kinase, also known as EIF2AK3/PEK), inositol-requiring 1 (IRE1) and activating transcription factor 6 (ATF6/ACHM7). Under basal conditions, HSPA5 keeps the three ER stress transducers inactive (47). The downstream effectors of these three key UPR regulators converge on promoting ER chaperone synthesis, ER-associated protein degradation and ER membrane biogenesis (49).
Observing the dynamics of transcriptome regulation in multiSLIDE in Figure 2, we also notice the early phase activation of ER chaperones, heat shock protein 90 beta family member 1 (HSP90B1), DNAJC3/P58IPK (member of the HSP40 chaperone family) and protein disulfide isomerases (PDIs), PDIA4, PDIA6 and PDIA3, that remains activated in the intermediate and late phases. PDIs are known to be responsible for the oxidation (formation), reduction (break down) and isomerization (rearrangement) of protein disulfide bonds via disulfide interchange activity. The other major role of PDIs is in general chaperone activity and recent studies have also identified PDI for its role as PERK activator (50–52). In Figure 2, it is interesting to observe that mRNA-level upregulation is countered by protein-level down-regulation of DnaJ heat shock protein family (Hsp40) member C3 (DNAJC3), which is a known inhibitor of PERK (53). This suggests that an active feedback control defense mechanism is in place to mitigate ER stress, allowing the PERK pathway to remain uninhibited (54,55).

The reversal of expression levels in the ER resident proteins SELK and MANF to their pre-treatment levels at the very late phase can also be attributed to the ER stress attenuation mechanism of UPR. Up-regulation of selenoprotein K (SELK) and mesencephalic astrocyte-derived neurotrophic factor (MANF) are a protective mechanism to avoid ER stress mediated cell death.

In summary, jointly visualizing the mRNA- and protein-level expression data in multiSLIDE helps the user to uncouple the distinct expression regulation patterns at different response phases of UPR. Clustering genes at the mRNA level and applying the same ordering at the protein level helped visualize whether the clusters propagate across -omics levels. An activated UPR initiates an adaptive stress response to regulate downstream effectors and further through a feedback control switches on/off transcriptional regulation and protein synthesis to restore ER homeostasis. A failure to attain homeostasis leads to programmed cell death (56,57).

Exploring ESR1 regulon in basal-like breast cancer tumors

In recent years, genome-scale characterization of breast cancer subtypes has helped us understand its tissue heterogeneity and identify additional therapeutic targets such as human epidermal growth factor receptor 2
(HER2/ERBB2). The four major intrinsic subtypes of breast cancer are defined as luminal A, luminal B, HER2–enriched and basal-like (27).

However, immunohistochemistry (IHC)-based surrogate subtyping, which are still routinely used for classification in pathological settings, is based on three key hormone receptors: estrogen receptor (ESR1), progesterone receptor (PGR), and HER2 (58). Basal-like tumors tend to have the worst prognosis, with all three receptors negative, making them unresponsive to endocrine therapy. Specifically, the expression of ESR1 is widely linked to better survival and considered a major regulator of the phenotypic properties of these breast cancers. ESR1 is a hormone-regulated transcription factor, a member of a superfamily of nuclear receptors, and plays a key role in cell proliferation. Typically, cell growth in breast cancer is stimulated either by the hormone estrogen (17β-estradiol) or growth factors such as EGF. Interestingly, transcription of EGF receptor is regulated by ESR1. Therefore, exploring the behavior of TF targets of ESR1, particularly in the basal-like subtype, can reveal genes that bypass ESR1 negativity.

We visualized proteomics data from CPTAC (59) and transcriptomics data (60) from TCGA invasive ductal breast carcinoma (TCGA-BRCA) in multiSLIDE. To understand the role of ESR1 as a master regulator of pathways, we first query and visualize all its transcription target genes at the mRNA and protein levels in multiSLIDE. A query for the TF targets of ESR1 in multiSLIDE's databases resulted in a total of 3,787 targets, of which the 2,312 unique targets present in the datasets were included in the analysis.

Next, using the PAM50 subtype classification as phenotype, we performed differential expression analysis at the protein level in multiSLIDE. With (unadjusted) p-value < 0.001 from analysis of variance as the statistical filter, we were able to select 172 differentially expressed genes. We applied hierarchical clustering to these genes with Euclidean distance and complete linkage, resulting in four distinct clusters consistent at mRNA and protein level. For visual clarity, we visualized the four clusters in two parts, as shown in Figures 3A and 3B (the side track on top of each heatmap indicating the subtypes). The genes in Figure 3A show two distinct clusters of genes (top-half and bottom-half of the figure) where the ESR-tumors (basal-like and HER2-enriched subtypes) are predominantly down-regulated (blue) while the other
subtypes are mostly up-regulated (red), across both -omics levels. These genes had signatures that were strongly coherent with that of ESR1. The clusters shown in Figure 3B had expression patterns that were opposite to that of ESR1 across subtypes. The latter of the two clusters (bottom-half of Figure 3B) had pronounced up-regulation only in the basal-like subtype but no distinct signature in the other subtypes.

In Figure 3A, the two clusters included key tumorigenic genes down-regulated in basal-like tumor samples and some HER2-enriched samples, such as progesterone receptor (PGR), carbonic anhydrase 12 (CA12), androgen receptor (AR), growth regulating estrogen receptor binding 1 (GREB1), and GATA binding protein 3 (GATA3) that control cell proliferation. The under-expression of ESR1 and these targets in the basal-like subtype suggests the role of alternative pathways for growth and proliferation of basal-like tumors, bypassing ESR1-dependent stimulation. One such pathway is epidermal growth factor (EGF)-induced tumor growth. In Figure 3B, EGFR is over-expressed in the basal-like subtype at both mRNA and protein levels. An overexpression of EGF receptor (EGFR) drives an aggressive form of cell proliferation, which is often associated with poor survival (61–63). In addition, Figure 3B also shows that the FAT atypical cadherin 1 (FAT1) gene, a member of the cadherin superfamily, widely considered as a tumor suppressor, is overexpressed in the basal-like subtype at both mRNA and protein levels. This is indeed consistent with recent literature indicating a more multifaceted role for FAT1, suggesting it as either a tumor suppressor or tumor promoter in a context dependent manner (64,65). Most of the other genes in Figure 3B, that are upregulated in the basal-like subtype are known for their role in cell motility, growth and positive regulation of GTPase activity and has to do with tumor survival, migration and metastasis.

Furthermore, using a network-based integration method for multi-omics data, Koh et al. identified overexpressed TFs such as CCAAT enhancer binding protein beta (CEBPB), nuclear factor I X (NFIX), WW domain containing transcription regulator 1 (WWTR1), and WD repeat domain 74 (WDR74), as unique drivers of basal-like subtypes (66). Among the targets of ESR1 in Figure 3B, we also observed CEBPB and NFIX were overexpressed at both -omics levels. Given that these are TFs that in turn affect expression of their target genes, their overexpression in the basal-like subtype is suggestive that they play a role as key regulators of cell proliferation and aggressive tumor development in the subtype.
Supplementary Figures S4A and S4B visualize the targets of these four transcription factors: CEBPB, NFIX, WWTR1, WDR74. Statistical filtering and hierarchical clustering of this data revealed a cluster downregulated in the ESR- basal-like subtype (Supplementary Figures S4A) as well as a cluster upregulated in the ESR- basal-like subtypes (Supplementary Figures S4B). The TF search keys are highlighted by green arrows in Supplementary Figures S4B.

In summary, concurrent visualization of the curated ESR1 regulon both at the protein and mRNA level reveals target genes that are under-expressed in basal-like subtype due to ESR1 negativity, consistent at both molecular levels. At the same time, we were able to zoom into a set of specific target genes that bypass ESR1 negativity and are activated in the basal-like subtype. The same type of query-driven inspection of multi-omics data can be generically extended to other applications.
Discussion

In this paper, we described multiSLIDE, a new web-based tool for query-driven interactive heatmap-based visualization of multi-omics data. With steady growth in multi-omics experiments across many domains of biomedical research, it has become increasingly important to develop open-source analysis and visualization tools. This was the primary motivation behind the development of multiSLIDE. We demonstrated how multiSLIDE enables targeted exploration of large multi-omics datasets within a biological context. From a practical point of view, 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 current tools.

Methods in multi-omics integration are often devised making many assumptions. For instance, in correlation-based integration methods, genes with correlated expression profiles across -omics levels are considered closely associated. However, there could be numerous other factors such as time delay in response between -omics levels or proteo-static regulation of protein expression that violates this assumption. At the same time, anti-correlated patterns may be of interest as well. In the ER stress time-course multi-omics data, we identified the example of DNAJC3 gene, which was up-regulated at the transcriptome level but down-regulated at the protein level. Such patterns, if not known beforehand, may not be accounted for in the assumptions and thus remain undiscovered. Visualizing pathways in their entirety is a useful alternative for exploring the complex response patterns.

In multiSLIDE, the ordering of genes across all -omics datasets can be determined by clustering the expression profiles of any one of them. This sorting functionality, imposed on one of the datasets and propagated to the others simultaneously, can reveal clusters with concordant or discordant gene expression regulation across different -omics layers. In the breast cancer data, for instance, the PAM50 classifier used for identifying breast cancer subtypes was created using mRNA data (27). In Figures 3A and 3B, upon hierarchical clustering in multiSLIDE, distinct clusters corresponding to PAM50 classifications were found.
at the transcriptome level, as expected, but the same clusters were also visible at the proteome level just by enforcing the same ordering as that of the transcriptome level.

Currently, multiSLIDE automatically links different datasets using standard gene identifiers. Therefore, visualizing molecular features that are not gene products is not possible in the current implementation, and we are currently exploring a new version of the tool to address this limitation. For visualizing proteomics data and metabolomics data concurrently, for example, the map between genes and metabolites has not been completely charted by experimental means. Hence an ideal architecture for a future version of multiSLIDE may need to infer networks between genes and other types of molecules from the data directly, in conjunction with known connections.

Other data types that may clash with gene-based summaries are found in existing -omics data such as DNA methylation data, ChIP/ATAC-seq, and histone modification profiles to name a few, which cover more vaguely defined units over the genome. For example, DNA Methylation at a gene promoter or enhancer region acts to repress gene transcription. When multiple regions close to the gene are methylated or in regions with dense population of genes, care has to be taken to denote them as gene promoter or gene body. Such one to many mappings will require further modification to the software architecture and will be included in future iterations of multiSLIDE.

Acknowledgments

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.

Data and Software Availability
The source code, installation instructions and a demo online version of multiSLIDE are available at https://github.com/soumitag/multiSLIDE. The tool is accessible by any modern web-browser with the server deployed either in a distributed environment or installed locally.
References

1. McManus J, Cheng Z, Vogel C. Next-generation analysis of gene expression regulation–comparing the roles of synthesis and degradation. Mol Biosyst. 2015;11(10):2680–9.

2. Liu Y, Aebersold R. The interdependence of transcript and protein abundance: new data–new complexities. Mol Syst Biol. 2016 Jan 1;12(1):856.

3. Ebhardt HA, Root A, Sander C, Aebersold R. Applications of targeted proteomics in systems biology and translational medicine. Proteomics. 2015 Sep 1;15(18):3193–208.

4. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the eBioPortal. Sci Signal. 2013 Apr 2;6(269):pl1–pl1.

5. Goldman M, Craft B, Kamath A, Brooks A, Zhu J, Haussler D. The UCSC Xena Platform for cancer genomics data visualization and interpretation. bioRxiv. 2018 Jan 1;326470.

6. Vasaikar S V, Straub P, Wang J, Zhang B. LinkedOmics: analyzing multi-omics data within and across 32 cancer types. Nucleic Acids Res. 2017/11/09. 2018 Jan 4;46(D1):D956–63.

7. King ZA, Dräger A, Ebrahimia N, Sönnerschein N, Lewis NE, Pålsson BO. Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways. PLoS Comput Biol. 2015 Aug 1;11(8).

8. Tokimatsu T, Sakurai N, Suzuki H, Ohta H, Nishitani K, Kyoya T, et al. KaPPA-view: a web-based analysis tool for integration of transcript and metabolite data on plant metabolic pathway maps. Plant Physiol. 2005 Jul;138(3):1289–300.

9. Thimm O, Bläsing O, Gibon Y, Nagel A, Meyer S, Krüger P, et al. mapman: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J. 2004 Mar 1;37(6):914–39.

10. Hernández-de-Diego R, Tarazona S, Martínez-Mira C, Balzano-Nogueira L, Furió-Tari P, Pappas Jr GJ, et al. PaintOmics 3: a web resource for the pathway analysis and visualization of multi-omics data. Nucleic Acids Res. 2018/05/25. 2018 Jul 2;46(W1):W503–9.

11. Kutmon M, van Iersel MP, Bohler A, Kelder T, Nunes N, Pico AR, et al. PathVisio 3: an extendable pathway analysis toolbox. PLoS Comput Biol. 2015;11(2):e1004085.

12. Wang Q, Tang B, Song L, Ren B, Liang Q, Xie F, et al. 3DScapeCS: application of three dimensional, parallel, dynamic network visualization in Cytoscape. BMC Bioinformatics. 2013;14(1):322.

13. Ang Y, Yu N, Seo J, Kim S, Lee S. MONGKIE: an integrated tool for network analysis and visualization for multi-omics data. Biol Direct. 2016 Mar 18;11(1):10.

14. Brown KR, Otasek D, Ali M, McGuflin MJ, Xie W, Devani B, et al. NAViGaTOR: Network Analysis, Visualization and Graphing Toronto. Bioinformatics. 2009/10/16. 2009 Dec 15;25(24):3327–9.

15. Junker BH, Klukas C, Schreiber F. VANTED: A system for advanced data analysis and visualization in the context of biological networks. BMC Bioinformatics. 2006;7(1):109.

16. Hu Z, Hung J-H, Wang Y, Chang Y-C, Huang C-L, Huyck M, et al. VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology. Nucleic Acids Res. 2009 May 21;37(suppl_2):W115–21.

17. Xia J, Gill EE, Hancock REW. NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data. Nat Protoc. 2015;10(6):823.

18. Theocharidis A, Van Dongen S, Enright AJ, Freeman TC. Network visualization and analysis of gene expression data using BioLayout Express 3D. Nat Protoc. 2009;4(10):1535.

19. Zhou G, Xia J. OmicsNet: a web-based tool for creation and visual analysis of biological networks in 3D space. Nucleic Acids Res. 2018/06/07. 2018 Jul 2;46(W1):W514–22.

20. Pavlopoulos GA, O’Donoghue SI, Satagopam VP, Soldatos TG, Pafilis E, Schneider R. Arena3D: visualization of biological networks in 3D. BMC Syst Biol. 2008;2(1):104.

21. Shi Z, Wang J, Zhang B. NetGestalt: integrating multidimensional omics data over biological networks. Nat Methods. 2013 Jul;10(7):597–8.

22. Schulz H-J, Hurter C. Grooming the hairball–how to tidy up network visualizations? In 2013.

23. Nocaj A, Ortmann M, Brandes U. Untangling hairballs. In: International Symposium on Graph Drawing. Springer; 2014. p. 101–12.

24. Lex A, Streit M, Schulz H, Partl C, Schmalstieg D, Park PJ, et al. StratomeX: visual analysis of large-scale heterogeneous genomics data for cancer subtype characterization. In: Computer graphics forum. Wiley Online Library; 2012. p. 1175–84.
Kranz P, Neumann F, Wolf A, Classen F, Pompisch M, Ocklenburg T, et al. PDI is an essential redox-sensitive activator of PERK during the unfolded protein response (UPR). Cell Death & Dis. 2017 Aug 10;8:e2986.

Yan W, Frank CL, Korth MJ, Sopher BL, Novoa I, Ron D, et al. Control of PERK eIF2α kinase activity by the endoplasmic reticulum stress-induced molecular chaperone P58IPK. Proc Natl Acad Sci. 2002;99(25):15920–5.

Rainbolt TK, Saunders JM, Wiseman RL. Stress-responsive regulation of mitochondria through the ER unfolded protein response. Trends Endocrinol Metab. 2014;25(10):528–37.

Lebeau J, Saunders JM, Moraes VWR, Madhavan A, Madrazo N, Anthony MC, et al. The PERK Arm of the Unfolded Protein Response Regulates Mitochondrial Morphology during Acute Endoplasmic Reticulum Stress. Cell Rep. 2018 Mar 13;22(11):2827–36.

Schröder M, Kaufman RJ. ER stress and the unfolded protein response. Mutat Res Mol Mech Mutagen. 2005;569(1–2):29–63.

Sano R, Reed JC. ER stress-induced cell death mechanisms. Biochim Biophys Acta (BBA)-Molecular Cell Res. 2013;1833(12):3460–70.

Nielsen TO, Hsu FD, Jensen K, Cheang M, Karaca G, Hu Z, et al. Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma. Clin cancer Res. 2004;10(16):5367–74.
Figure Legend

Figure 1: Visualization workflow of multiSLIDE. (A) The web-based interface of multiSLIDE, visualizing expression patterns of 25 significant genes in the ‘EGFR1’ pathway across mRNA and protein levels. In multiSLIDE, sample phenotypes, gene groups, and user-selected molecular interactions are visualized alongside the expression profiles. Here, phenotypes, molecular interactions, and gene groups are shown above, to the left of, and to the right of the heatmap, respectively. (B) Visualization in multiSLIDE starts with the user selecting datasets, gene groups (relevant pathways, GO terms, or genes), and phenotypes to visualize using the Selection Panel. Multiple gene groups can be simultaneously searched and selected using an intuitive search syntax. (C) Putative transcription factor targets or physically interacting proteins of a selected gene can be queried with a single right-click on the gene name. Interesting regulons or their interacting genes can be selected from the list and added to the heatmaps, with the molecular interactions visualized as tracks alongside the heatmaps. (D) For each -omics level heatmap, the data range, the number of color bins, and color scheme can be adjusted individually. Given the variability between -omics data, this customization is essential for a meaningful comparative analysis. (E) Settings common to all -omics levels (global settings), such as heatmap cell size and orientation, can be applied immediately to all heatmaps using the Layout Panel. (F) Using the Sorting and Filtering Panel, genes and samples can be ordered in multiple different ways, while non-significant genes can be filtered out. Appropriate ordering and filtering are essential for revealing structure in the data. (G) All the heatmaps are synchronized by samples and genes and can be scrolled simultaneously using the Scrolling Panel. H. Multiple curated lists of genes (feature lists) can be created and maintained within multiSLIDE. These lists can be re-visualized in multiSLIDE or downloaded as text files with a single click.

Figure 2: Visualization of unfolded protein response in mammalian cells responding to stress. mRNA and protein-level expression across eight time-points and two replicates, as visualized in multiSLIDE. Gene
expression profiles belonging to the GO terms and pathways shown in the “Legends” panel are visualized in the heatmaps. Association between genes and GO terms/pathways are indicated by colored tags in vertical tracks alongside the heatmap. Sample information, such as replicate number and timepoint, is visualized above the heatmaps. Genes highlighted with a green band show early (< 2 h) spikes in response to stress at the mRNA level. Among these, except DNAJC3, all others have a delayed upregulation at the protein level, beginning in the intermediate (2 - 8 h) phase and persisting through the late phase (> 8 h).

**Figure 3: Visualization of significant transcription regulation targets of ESR1.** The TF targets of ESR1 were queried and visualized simultaneously across the mRNA and protein level in multiSLIDE. Significance based filtering (one-way ANOVA, p-value <0.001) in multiSLIDE left 172 targets, which were clustered using Euclidean distance and complete linkage. Two of these clusters are shown in (A) and the other two clusters are shown in (B).
A. Datasets
- Genes
  - * ADD GENES
  - * indicates required field
- Phenotypes

B. Query Pathways, GO Terms and Genes to visualize
- Genes
  - * ADD GENES
  - * indicates required field
- Phenotypes

C. Select a gene to explore its regulatory and protein-protein interaction networks

D. Customize individual heatmaps
- Settings for mRNA Heatmap
  - Number of Colors: 21
  - Width Values: 9-256
  - Biological Range: Use Min and Max of Data (Min=1.21, Max=9.83)
  - Use Symmetric Bins (default)
  - Use Range
    - Start: a
    - End: b
  - Heatmap Color Scheme
    - red
    - blue
    - green
    - yellow
    - purple
    - orange
    - gray
    - black
    - white

E. Customize layouts of all heatmaps concurrently
- Display Genes Along
  - X-axis
  - Y-axis
- Cell Size
- Layout
  - Sample Label Width: 100
  - Gene Label Width: 100

F. Sort samples and genes, and filter genes
- Order Samples By
  - PhenoTypes
  - Hierarchical Clusters
  - Significance Level
- Order Genes By
  - Phenotypes
  - Gene Groups
  - Show Only Significant Genes

G. Scroll through all heatmaps simultaneously
- Scroll By
  - Samples 1 to 73 of 73
  - Samples 1 to 11 of 11
  - Samples 73
  - Samples 11 to 55
  - Genes 25
  - Genes 11 to 35

H. Create and visualize curated (feature) lists of genes
