ComplexBrowser: A Tool for Identification and Quantification of Protein Complexes in Large-scale Proteomics Datasets

Authors
Wojciech Michalak, Vasileios Tsiamis, Veit Schwämmle, and Adelina Rogowska-Wrzesińska

Correspondence
adelinar@bmb.sdu.dk

In Brief
ComplexBrowser can identify protein complexes registered in manually curated databases that are present in datasets obtained from large scale quantitative proteomic experiments (label free or isobaric mass tag based). It provides, in the form of the CFC factor, a quantitative measure of the coordinated changes in complex components. This facilitates assessing the trends in the processes governed by the identified complexes providing a new and complementary way of interpreting proteomic experiments.

Highlights
- Automated analysis of protein complexes in proteomic experiments.
- Quantitative measurement of the coordinated changes in protein complex components.
- Interactive visualizations for exploratory analysis of proteomic results.
ComplexBrowser: A Tool for Identification and Quantification of Protein Complexes in Large-scale Proteomics Datasets*

Wojciech Michalak§, Vasileios Tsiamis, Veit Schwämmle¶, and Adelina Rogowska-Wrześnińska‡¶

We have developed ComplexBrowser, an open source, online platform for supervised analysis of quantitative proteomic data (label free and isobaric mass tag based) that focuses on protein complexes. The software uses manually curated information from CORUM and Complex Portal databases to identify protein complex components. For the first time, we provide a Complex Fold Change (CFC) factor that identifies up- and downregulated complexes based on the level of complex subunits coregulation. The software provides interactive visualization of protein complexes’ composition and expression for exploratory analysis and incorporates a quality control step that includes normalization and statistical analysis based on the limma package. ComplexBrowser was tested on two published studies identifying changes in protein expression within either human adenocarcinoma tissue or activated mouse T-cells. The analysis revealed 1519 and 332 protein complexes, of which 233 and 41 were found coordinately regulated in the respective studies. The adopted approach provided evidence for a shift to glucose-based metabolism and high proliferation in adenocarcinoma tissues, and the identification of chromatin remodeling complexes involved in mouse T-cell activation. The results correlate with the original interpretation of the experiments and provide novel biological details about the protein complexes affected. ComplexBrowser is, to our knowledge, the first tool to automate quantitative protein complex analysis for high-throughput studies, providing insights into protein complex regulation within minutes of analysis. Molecular & Cellular Proteomics 18:2324–2334, 2019. DOI: 10.1074/mcp.TIR119.001434.

Proteomics has become one of the methods of choice for large scale analysis of biological systems. Recent advances in multidimensional separation methods, improved instrument speed, sensitivity, and resolving power allow for the generation of near complete proteomes in the scope of 35 h of analysis, providing quantitative information for more than 12,000 gene products and covering 4–6 orders of magnitude (1, 2). Proteomics provides countless research opportunities, but also challenges, especially in the domain of biological interpretation of the results. Currently, this process remains time-consuming and requires considerable expertise.

Commonly used analysis pipelines involve annotating proteins according to their molecular function, cellular component, and biological processes, based on information gathered in Gene Ontology (GO)1 databases (3, 4). Further GO-term enrichment methods define over-represented annotations in users’ data, providing a general understanding of the biological processes affected (5).

Pathway analysis is a different approach concentrated on protein specific biochemical activity. Tools such as IPA®, KEGG or Reactome map proteins to molecular pathways (6–8) and visualize processes those gene products are known to be involved in. A clear advantage of this approach is that pathway databases are mostly based on experimental, manually curated data, whereas most GO-annotations come from in silico predictions and text mining (9).

An alternative practice, often employed in studies of less researched organisms, is protein domain and motif analysis (10). This strategy uses sequence alignment and secondary structure prediction tools to find similarities between a protein of interest and better-annotated analogs in other species. Identification of a specific sequence motif enables the assignment of function to the previously undescribed protein.

Analysis of protein-protein interactions (PPI) is a complementary approach used very often in parallel with the above methods. Platforms such as STRING (11) use information from co-expression studies, cross-species predictions, experimental evidence, and literature mining to build protein interaction graphs, in which nodes represent gene products and edges correspond to interactions. These maps facilitate the identification of genes involved in similar processes or influenced by common regulators. Platforms for PPI investigation provide comprehensive information on proteins’ function and involvement in various biological processes. How-
ever, the enormous knowledgebase of these platforms and the amount of interactions drawn from large gene/protein lists is often very difficult to handle and interpret.

Protein complexes are molecular machines that perform many of the key biochemical activities essential to the cell e.g. replication, transcription, translation, cell signaling, cell-cycle regulation, and oxidative phosphorylation. Their role in maintaining cell homeostasis and involvement in disease development (12) prove that the detailed characterization of protein complex expression would be very helpful in understanding the often highly intertwined processes in the cell.

Interrogation of the expression of components of known protein complexes in large scale proteomic data has been performed in several studies (13–15). It is now clear that many known protein complexes are translationally and post-translationally regulated, and therefore exhibit co-expression when compared across cell types and tissues. However, so far, no automated and user-friendly approach for the analysis of complex behavior in new datasets has been developed.

In this manuscript we present ComplexBrowser, which enables automated and quantitative analysis of protein complexes in proteomic experiments. The software interrogates CORUM (16) and EBI Complex Portal (17) databases to find known protein complexes present in a given protein list, and uses quantitative proteomic data (label free or isobaric mass tag based) and factor analysis to summarize the overall expression trends for each complex across the studied biological conditions. The re-analysis of two, previously published, large scale proteomic datasets show the large potential of the approach to gain in-depth knowledge about the regulation of protein complexes in different biological contexts.

**EXPERIMENTAL PROCEDURES**

**Protein Complex Databases**—The presented software relies on information from two established, manually curated protein complex databases: CORUM (16) and EBI Complex Portal (17) with 2693 and 2454 entries respectively, together covering 22 species (state for 24.05.2018). It is also possible to upload a user defined protein complex database; an example is provided in software manual (supplementary File S1).

**Software Design**

**Data Input**—ComplexBrowser accepts data Tables in .csv or .txt format as input. The file must contain single, unique UniProt (18) accessions in the first column and quantitative information in subsequent columns. Label free quantitative data in form of e.g. LFQ intensities (19) or isobaric tag based summarized reporter ion intensities for each analyzed sample from TMT or ITRAQ (20) experiments can be used. Optionally, columns with confidence scores from statistical test results can be appended. These values must be calculated in relation to the first condition appearing in the input file. If they are not included by the user, differential expression analysis using the limma package (21) is conducted, along with FDR estimation using the qvalue R package (22). Examples of data input files are provided in supplemental Tables S1 and S2. Additionally, the software provides an option for uploading the T-cell dataset as a test dataset (please see the detailed description of the source of the data in the section Test Datasets).

**Quality Control**—Prior to the analysis of protein complexes, the software creates data visualizations for quality control (QC) evaluation purposes. The QC visualizations include:
- Boxplot of log-transformed intensities to control for inconsistency in, for example, injection amounts.
- Missing value bar plots generated by summing up the number of missing values in each quantitative column to compare protein coverage between samples.
- Pairwise Scatter plots of log-transformed intensities of all proteins quantified in two selected conditions, displaying sample to sample correlations (Pearson, Kendall, Spearman), to test similarity between the samples.
- Histograms of coefficient of variation (CV) of protein absolute intensity measurements within each experimental condition to assess replicate variation.
- Q-value charts counting the number of differentially expressed features in relation to a set threshold.
- Volcano plots to depict the relationship between fold-changes and confidence for differentially regulated features.
- PCA results for visual comparison of all samples.

The software also implements four common, previously described normalization methods (23).

**Identification and Visualization of Protein Components**—To perform complex analysis additional parameters have to be set: (1) q value threshold, which will set the minimum q value considered for visualization purposes, (2) fold change threshold and (3) noise threshold used for the visualization of fold changes in complex components and for the generation of a summary report; (4) selection of database (CORUM or EBI Complex Portal); and (5) selection of species the samples were generated from. Upon pressing “Run analysis,” ComplexBrowser searches for relevant Uniprot accessions in users’ data and displays all protein complexes with at least one subunit found in the input. For complexes with at least 3 quantified subunits a complex fold change (CFC) is calculated (please see below). The composition of complexes and changes in the expression of their subunits are visualized using star graphs. Nodes (circles) denote subunits of the complex, the size of the nodes represents the degree of fold change, the colors indicate the type of regulation (red - downregulated, green - upregulated, blue - not changing, gray - not identified), thick edges (lines) indicate differentially regulated proteins. Selected components of the identified complex or the entire complex can be submitted for analysis in CoExpresso software (15), which assesses the co-regulatory behavior of arbitrary protein groups in more than 100 human cell types. CoExpresso is only compatible with human proteins, therefore the redirection to CoExpresso from ComplexBrowser is enabled only for human complexes.

**Complex Expression Analysis**—ComplexBrowser employs the fast-FARMS algorithm to summarize the abundance of protein complexes to determine complex fold changes (CFC). Fast-FARMS is a modified version of FARMS (24), a Bayesian factor analysis method with an assumption of Gaussian measurement noise originally implemented for Affymetrix probe level data summarization. It has previously been used to estimate protein abundance based on peptide concentrations.
in the protein summarization process. It has proven useful in detecting outliers in peptide expression profiles and limiting their influence on protein quantitation (25). In ComplexBrowser, we created our own implementation of the fast-FARMS algorithm for performing weighted average summarization of log-transformed expression changes of protein complex subunits. Fast-FARMS in ComplexBrowser assumes that the complex subunit abundances are proportional to the protein complex concentration. Therefore, the subunit abundances can be modeled linearly with an additional Gaussian noise as:

\[ x = \lambda z + \epsilon, \]

where \( x, \lambda \in R_1, \) and \( z \) are the subunit abundances and the true complex concentration in log scale respectively. \( \lambda \) expresses the contribution of each subunit to \( z \) and \( \epsilon \) describes the noise. Fast-FARMS solves the above model by a maximum a posteriori estimation of \( \lambda \) that describes the covariation of \( x \) the best under minimal noise \( \epsilon \). The complex expression is calculated in two steps. First, log scaled intensities of all subunits of the same complex are processed in fast-FARMS, which assigns individual weights based on \( \lambda \) to each subunit and estimates the noise \( \epsilon \). Complex expression is then calculated by a weighted average summarization per condition process applied to the subunit intensities. The relative change in complex expression between two given conditions is defined as the complex fold change (CFC). ComplexBrowser also provides a summary measure, describing the amount of variability in the expression profile of a given complex. This is presented as a signal-to-noise ratio (S/N), or in short - Noise, by comparing \( \lambda \) with noise \( \epsilon \). Noise of 0 indicates perfect co-expression, whereas Noise = 1 indicates very poor correlation. The default noise threshold set in the software is 0.5.

**Linearity of Complex Components Co-expression**—Building on the idea of using the linearity of subunit co-expression as a measure of data quality (26), ComplexBrowser draws supplementary visualizations to investigate co-regulation between different conditions. For a selected protein complex, it takes log-transformed abundances of all its subunits in two conditions and displays them on a scatter plot, where each point corresponds to one protein. Orthogonal distance regression (ODR) is employed to determine the quality of co-expression similarity, because unlike ordinary least squares regression, ODR considers variability in both \( x \) and \( y \) values, therefore it fits a model that minimizes errors of both measurements (27). The procedure returns a single \( R^2 \) value per complex for each pair of conditions as a measure of co-expression.

**Heatmaps and Hierarchical Clustering**—Combined with dendrograms produced by hierarchical clustering algorithms, heatmaps allow the detection of proteins not following the common trend of a protein complex. These could be subunits that participate in several different complexes or interact with a given complex transiently. ComplexBrowser displays two different heatmaps—expression and correlation (between protein expressions). The expression heatmap displays log-transformed, mean-normalized expression values for all subunits within a selected complex across all experimental conditions. The same data input is used to calculate pairwise correlations between expression profiles of all complex subunits. Based on this information a correlation matrix is computed and displayed as the correlation heatmap. Both graphs use aggregative hierarchical clustering implemented in R’s hclust function to provide dendrograms. The function uses distance measures and linkage functions selected by the user in the graphical interface.

**Test Datasets**—ComplexBrowser’s performance was tested using protein quantification data from two previously published studies (28, 29). Further on in the manuscript we refer to the studies as adenocarcinoma dataset (28) and T-cell dataset (28).

The adenocarcinoma dataset investigates protein expression differences between formalin-fixed, paraffin embedded tissue samples from patients with colon cancer compared with healthy colon mucosa and nodal metastatic tumors using label-free quantitation based on LFO intensities. The MS proteomic data of the adenocarcinoma dataset were obtained from supplementary tables of the original publication available from the publisher’s site (29). For this analysis we have discarded samples denoted as “CA2” and “NO2” to ensure an equal number of replicates in each condition. We have filtered the protein intensities table to retain proteins with at least 4 valid quantitative values within each condition. We have removed isoform identifiers from the original accession numbers and rows with non-unique identifiers were removed. This resulted in a dataset containing LFO values for 6824 proteins from 3 conditions with 7 biological replicates each. The input file used is this study can be found in supplemental Table S1.

The T-cell dataset studies activation of quiescent mouse T cells over four time points (0, 2, 8, and 16 h) in two biological replicates. Proteins were quantified using tandem mass tag (TMT) labeling and were analyzed on an Orbitrap Elite MS instrument. The data were obtained from the original publication from the PRIDE database (accession numbers PXD004367 and PXD005492) (30). The dataset contained normalized intensities for 8,431 proteins. The input file used is this study can be found in supplemental Table S2.

**Software Implementation**—ComplexBrowser was implemented in R (31). The user interface was developed using Shiny, Plotly, netWorkX, heatmapper, DT and data.table libraries, allowing interactive and adjustable data visualization. PreprocessCore, stringr, pracma, dplyr, limma, and qvalue packages were used for data manipulation and statistical analysis.

**Software Accessibility**—The tool can be accessed via web service at http://computproteomics.bmb.sdu.dk/Apps/ComplexBrowser or can be run locally after installation of Rstudio and the required libraries.

A fully functional demo version of ComplexBrowser is available online via http://computproteomics.bmb.sdu.dk/Apps/ComplexBrowser/.

The source code can be downloaded from: https://bitbucket.org/michalakw/complexbrowser.

**RESULTS**

We developed ComplexBrowser to enable the identification and quantification of protein complexes in large scale proteomic experiments. The unique feature of the software is its ability to quantify changes in the abundance of protein complexes and the co-expression of their components in different conditions of the experiment. The general analysis pipeline implemented in the program is presented in Fig. 1. In brief, a table containing the quantitative information of the identified proteins is uploaded using the web browser interface. After defining the parameters of the analysis (e.g. number of conditions and replicates), the analysis of the quality of the data is carried out and visualized. In a following window the analysis of the presence and changes in abundance of protein complexes are carried out. Interactive tables and graphics allow the user to conveniently evaluate the results of the analysis. Tables containing results and a summary report are available for download. All figures generated by ComplexBrowser, as well as the QC report can be exported as vectorised graphics and edited in any PDF editor or graphics design software. An extensive description of ComplexBrowser’s procedures and example results can be found in software manual in supplementary File S1.
To test the performance of the developed platform, we have used two published proteomics studies: adenocarcinoma dataset (29) and T-cell dataset (28). The results of both the quality control and protein complex analysis steps are presented below.

**Quality Control Of Proteomic Data in ComplexBrowser—**

The Adenocarcinoma dataset containing quantitative proteomic values from 3 biological conditions with 7 replicates each was uploaded to ComplexBrowser in the following order: C1 - control, C2 - metastasis, C3 - cancer. A QC report file generated by ComplexBrowser summarizing the quality analysis is provided in supplementary File S2.

Analysis of Boxplot graphs of log-transformed intensities, Fig. 2A, indicated that normalization was necessary to reduce the variability between intensity distributions and ensure sample to sample comparability; therefore, a quantile normalization was performed. Despite the normalization, the mean CV values were 65, 78, and 77% for normal, metastasis and cancer samples respectively, Fig. 2B, indicating relatively large variability within measurements, most likely because of the clinical character of the samples.

Label free experiments may contain proteins that have been quantified in only a few samples and in other samples their quantitative value is missing (referred to as “missing values”). Many missing values in a few samples could indicate lack of technical reproducibility and/or quality of the obtained data. In the adenocarcinoma dataset the number of missing values per sample varied from 113 to 510 and consisted, in total, of only 3.5% of all valid measurements, Fig. 2C. It did not show any persistent bias of the data.

Differentially expressed features were determined using a paired test and FDR estimation included in the limma package, supplemental Table S3. Considering the large deviation in measurements and clinical sample character, features with 0.01 FDR value were considered, resulting in the identification of 802 proteins differentially regulated for metastasis and 813 for cancer samples. PCA analysis has shown a good separation of control samples but an overlap between the carcinoma and metastatic tissues, Fig. 2D.

Subsequently ComplexBrowser data quality analysis was applied to the T-cell dataset, which consisted of four sets (0, 2, 8, and 16 h) with two replicates and was generated using TMT-based quantitation. The complete report can be found in supplementary File S3. The Boxplot graph distributions of the T-cell dataset results show only very small variation in the median values for the different samples and did not require any further normalization. No missing values were observed. The variability of measurements was noticeably lower compared with the Adenocarcinoma experiment with an average CV from all conditions of 4.56% (versus 73.51% for the previous dataset), supplementary File S2 and S3. Fig. 3A illustrates a high level of sample to sample correlation of TMT intensities for each protein between selected samples. An increasing number of differentially expressed proteins (39, 1869, and 5600) were detected after 2, 4, and 16 h of T-cell activation, at FDR of 0.05. Fig. 3B presents volcano plots generated by the ComplexBrowser software. Results of the statistical analysis downloaded from ComplexBrowser are in supplemental Table S4.

**Protein Complex Analysis—** ComplexBrowser facilitates the analysis of protein complex expression in large-scale studies. The software queries the input data for proteins reported to be complex members, investigates their co-expression patterns and visualizes results according to specified statistical, noise and expression change thresholds. It quantifies changes of complex abundance by calculating the complex fold change factor (CFC). Resulting graphs and tables allow for data exploration using interactive figures.

We recommend that setting parameters for complex analysis should be performed based on the results obtained from
the quality analysis module. The adenocarcinoma dataset showed a relatively high variation between samples therefore it was analyzed using the CFC $\geq 1.5$, complex Noise $\geq 0.5$ and protein expression q value $\leq 0.05$. The T-cell dataset showed lower measurement variability and good correlation between the biological replicates therefore the CFC was set to $\geq 1.2$, complex noise threshold was kept at $0.5$ and protein expression q value $\leq 0.05$.

The main results of the complex analysis module are presented in the form of a tabular output that can be downloaded directly from ComplexBrowser. Such output for the adenocarcinoma dataset for CORUM and EBI databases can be found in supplemental Table S5 and S6 respectively. In ComplexBrowser the table can be sorted, filtered and searched to provide easy access to the relevant complexes. It contains complex ID, complex names, number of proteins (subunits) of the complex identified and quantified in the analyzed dataset, number of all subunits and a % complex coverage that allows the user to identify complexes that are highly represented in the analyzed dataset. CFC - complex fold change, together with Noise factor, is given for the analyzed conditions providing a means to evaluate the coordinated changes in expression of complex components. Accession numbers of all identified complex subunits and gene ontology annotations for the complex are also listed in the table.

Protein Complexes Identified in the Adenocarcinoma Dataset Reflect Biological Features of Cancer and Metastatic Tissues—Protein complex analysis of the adenocarcinoma data
set identified 1519 protein complexes from CORUM and 366 from Complex Portal, Table I. Typical visualization of a protein complex and its components generated by ComplexBrowser is presented in Fig. 4. The top 5 most up-regulated and top 5 most downregulated complexes in cancer tissue selected based on the CFC are shown in Table II. Mitochondrial respiratory chain I (−4.368 CFC) and F1F0- cytochrome C oxidase (−4.188 CFC) were identified as the key complexes down-regulated in both metastatic and cancer samples, Table II. This finding combined with a significant down-regulation of ATP synthase (−1.263 CFC) indicates a shift in cancer cells' metabolism from oxidative phosphorylation to glycolytic path-

Fig. 3. Example of ComplexBrowser generated visualization of quality control analysis of the T-cell dataset. A, Pearson correlation of TMT intensities between 4 selected samples. Left panel - C1_1 versus C2_1; Middle panel - C1_1 versus C3_1; Right panel - C1_1 versus C4_1; B, Volcano plots at FDR of 0.05; C1, C2, C3 and C4 depict nonstimulated T-cells (0 h) and T-cells stimulated for 2 h, 8 h, and 16 h respectively; C1_1 and C1_2 etc. depict different replicates within each condition.

| Adenocarcinoma | T-cell |
|----------------|--------|
| **Species** | Homo sapiens | Mus musculus |
| **Quantitation method** | Label-free | TMT |
| **Mean CV** | 73.51% | 4.56% |
| **Total number of missing values (%)** | 3.55% | 0% |
| **Number of conditions** | 3 | 4 |
| **Number of replicates** | 7 | 2 |
| **Number of proteins in the dataset** | 6824 | 8431 |
| **Number of complexes (No. of proteins participating in complexes)** | CORUM 1519 (1687) | Complexity Portal 366 (437) |
| **Regulated in CORUM** | | |
| **Condition** | Cancer/Normal | Metastasis/Normal | 2/0 h | 8/0 h | 16/0 h |
| Up | 228 | 219 | 0 | 10 | 33 |
| Down | 8 | 9 | 1 | 3 | 9 |
| **Regulated in EBI Complex Portal** | | |
| Up | 59 | 52 | 1 | 6 | 29 |
| Dow | 11 | 5 | 0 | 1 | 5 |
ways, which is consistent with the results of the original publication (29).

Among the top 5 regulated complexes we have identified over-expression of 3 complexes related to mitotic cell cycle progression and activation: RalBP1-CDC2-CCNB1 complex (7.420 CFC), CDC2-CCNA2-CDK2 complex (6.031 CFC) and Cell cycle kinase complex CDC2 (4.820 CFC), for more details see Table II and supplemental Table S4. This points to increased activation of cell division and replication, as well as reduction in respiration, which are known characteristics of cancer (32, 33).

In addition to previously described results, ComplexBrowser detected a 3.58-fold up-regulation of MTA1 complex.
involved in metastatic tumor formation in nodal tissue samples (34). The same complex was changed 2.33-fold in the main tumor tissue, supplemental Table S4.

Activation of Mouse T-cells is Reflected in Coordinated Changes of Protein Complex Components—ComplexBrowser was used to analyze protein complexes during activation of murine T-cells. The dataset consists of 4 consecutive time points (0, 2, 8, and 16 h) collected after activation. It is expected that trends of coordinated changes of protein complexes would follow the timeline of the activation events and would be reflected in the CFC changes.

Despite the large number (8431) of proteins present in the T-cell dataset, complex analysis had identified only 332 and 374 protein complexes in the CORUM and EBI Complex Portal respectively, Table II. This is most likely because there are fewer mouse specific protein complexes present in those databases. Additionally, the complexes identified with the two databases are very different and share only 16% of the involved proteins.

Despite those drawbacks, use of both databases identified an increasing number of significantly changing complexes over the different time points of T-cell activation reflecting the gradual changes in the proteome of activated cells. For example, based on EBI Complex Portal, one complex was significantly regulated 2 h after T-cell activation and 7 and 34 at 8 and 16 h time points respectively, Table II. A summary of the top 5 most up-regulated and top 5 most downregulated complexes after 16 h of T-cell activation can be found in Table III.

The largest quantified complex was Parvulin-associated pre-rRNP complex (CORUM) with 49 out of 62 subunits quantified, Fig. 5A. Its abundance was highly correlated (Noise = 0.001) and increased over the time course of the experiment (~1.03, 1.44, 2.81 CFC after 2, 8, and 16 h of activation respectively). This protein complex is involved in ribosome biogenesis and contains several ribosomal subunits (35). BAT3 complex (EBI Complex Portal), responsible for the targeting of the transmembrane domain-containing proteins from the ribosome toward membranes, was also up-regulated in the last condition (CFC 1.42), Fig. 5A, suggesting an overall increase in processes related to protein synthesis.

Combined data from CORUM and EBI Complex Portal showed that 62 complexes were induced 16 h after activation, with a substantial increase in abundance between 8 and 16 h after activation, supplemental Table S7 and S8.

The p19-Cdk4-cyclinD2 (3.96 CFC after 16 h) and Cyclin D1-associated (2.47 CFC after 16 h) protein complexes were found to be among the most up-regulated 8 and 16 h after activation. These two complexes are involved in the regulation of cell cycle and transition through G1 phase, and their coordinated up-regulation reflects the transition of T-cells into the proliferative state. These changes were accompanied by an increase in the expression of various DNA polymerase complexes.

### Table III

Top 10 up and downregulated protein complexes in the T-cell experiment 16 h after activation based on CORUM and EBI Complex Portal databases, result for 2 h and 8 h after activation are shown for comparison; NQS = number of quantified subunits; NUS = number of unique subunits; CFC 2/0 h, CFC 8/0 h and CFC 16/0 h — complex fold change between 2, 8, and 16 h after activation and control samples (0 h) respectively.

| Complex name | NQS/NUS | CFC 2/0 h | CFC 8/0 h | CFC 16/0 h |
|--------------|---------|-----------|-----------|------------|
| p19-Cdk4-cyclinD2 complex | 3/3 | 1.074 | 3.189 | 3.674 |
| Parvulin-associated pre-rRNP complex | 49/62 | −1.032 | 1.440 | 2.810 |
| PCNA-DNA ligase complex | 4/4 | 1.130 | 1.436 | 2.197 |
| 9S-cytoplasmic 40S ribosomal subunit (40S) | 3/4 | 1.047 | 1.377 | 1.917 |
| BCL2L1-p53-PUMA complex, DNA damage induced | 3/3 | 1.052 | 1.431 | 1.862 |
| G protein complex (Btk, Gng2, Gnb1) | 3/3 | 1.009 | −1.097 | −1.284 |
| Stx7-Unc13d-Vamp8 complex | 3/3 | −1.086 | −1.144 | −1.33 |
| Itgav-Itgb3-Gsn complex | 3/3 | 1.028 | −1.08 | −1.361 |
| Cd3d-Cd3g-Cd3e-Cd247 complex | 4/4 | −1.065 | −1.328 | −1.385 |
| Cd3g-Cd3e-Cd247-Canx complex | 4/4 | −1.097 | −1.280 | −1.414 |
| Bcl-xL-p53-PUMA complex, DNA damage induced | 3/3 | 1.052 | 1.431 | 1.862 |
| G protein complex (Btk, Gng2, Gnb1) | 3/3 | 1.009 | −1.097 | −1.284 |
| Stx7-Unc13d-Vamp8 complex | 3/3 | −1.086 | −1.144 | −1.33 |
| Itgav-Itgb3-Gsn complex | 3/3 | 1.028 | −1.08 | −1.361 |
| Cd3d-Cd3g-Cd3e-Cd247-Canx complex | 4/4 | −1.065 | −1.328 | −1.385 |
| A2/p19-Cdk4-cyclinD2 complex | 3/3 | 1.074 | 3.189 | 3.674 |
| Parvulin-associated pre-rRNP complex | 49/62 | −1.032 | 1.440 | 2.810 |
| PCNA-DNA ligase complex | 4/4 | 1.130 | 1.436 | 2.197 |
| 9S-cytoplasmic 40S ribosomal subunit (40S) | 3/4 | 1.047 | 1.377 | 1.917 |
plexes, e.g., DNA synthesome (1.49 CFC after 16 h), DNA polymerase alpha, delta and epsilon complexes (1.45, 1.25, 2.09 CFC after 16 h), and Telomerase holoenzyme complex (1.40 CFC after 16 h) pointing to an increase in processes associated with DNA replication and cell proliferation.

The AP-1 transcription factor complex FOS-JUN-NFATC2 (CORUM) (CFC 1.87 at 2 h, 2.13 at 8 h and 1.62 at 16 h) is an example of a complex with a highly coordinated early response to the stimulus, Fig. 5C. Similar patterns in the AP-1 complex were also reported by Tan et al. (28).

A trend of downregulation was found in the expression of the Cd3d-Cd3g-Cd3e-Cd247 complex (1.07, 1.2, 1.36 CFC), a membrane glycoprotein assembly which is a part of the T-cell co-receptor, Fig. 5D, that is shut down during T-cell activation (28).

Additionally, we have identified a variety of complexes involved in post-translational modification of histones and chromatin remodeling, the majority of which showed increased expression at 8 and 16 h of T-cell activation, Table IV. These complexes and regulation of epigenetic processes in the maturation of T-cells had been largely overlooked in the original study. Describing the regulation of transcription by post-translational modifications and chromatin remodeling requires their quantification which was not done in this case. An extensive follow-up analysis is necessary to interpret these results.

In the T-cell dataset ComplexBrowser analysis identified upregulated complexes involved in DNA replication and chromatin remodeling, protein synthesis initiation and cell cycle progression, which allowed us to conclude that the T-cells undergo significant cellular reprogramming and exit a quiescent state 8 to 16 h after stimulation. Moreover, there was a decrease in T-cell signaling receptor expression from the first time point of the experiment.

**DISCUSSION**

ComplexBrowser is, to our knowledge, the first automated tool that enables quantitative analysis of protein complexes in proteomic experiments. It is available through a web-browser and does not require any installations or programming experience. Thus, it has high potential for integration into data analysis workflows commonly used by the scientific community.

Based on the datasets analyzed in this manuscript we have shown that it is compatible with label free and TMT based quantitation experiments. And there are no technical limita-
Chromatin remodelling and histone modifications associated protein complexes identified in T-cells 16 h after activation; NQS - number of quantified subunits; NUS - number of unique subunits; CFC 2/0 h, CFC 8/0 h, and CFC 16/0 h—complex fold change between 2, 8, and 16 h after activation and control samples (0 h) respectively

| Complex name                                         | NQS/NUS | CFC 2/0 h | CFC 8/0 h | CFC 16/0 h |
|------------------------------------------------------|---------|-----------|-----------|------------|
| B-WICH chromatin remodelling complex                 | 9/7     | −1.104    | 1.221     | 1.837      |
| Methylosome                                          | 3/3     | 1.016     | 1.185     | 1.689      |
| Chromatin assembly factor 1 complex                  | 3/3     | 1.03      | −1.031    | 1.635      |
| CHRAC chromatin remodeling complex                   | 4/4     | −1.032    | 1.138     | 1.52       |
| eNoSc complex                                        | 3/3     | −1.075    | 1.046     | 1.371      |
| BAT3 complex                                         | 3/3     | 1.009     | 1.091     | 1.36       |
| NuA4 histone acetyltransferase complex               | 19/19   | −1.002    | 1.108     | 1.321      |
| PPP4C-PPP4R2-PPP4R3B protein phosphatase 4 complex   | 3/3     | 1.058     | 1.15      | 1.268      |
| MBD3/NuRD nucleosome remodeling and deacetylase complex | 12/12   | −1.135    | 1.04      | 1.261      |
| SRCAP histone exchanging complex                     | 9/8     | −1.001    | 1.073     | 1.252      |

ComplexBrowser relies on information stored in CORUM (16) and Complex Portal databases (17) and is therefore dependent on the efforts of their administrators. The composition of these resources introduces a bias in the analysis, because the largest proportion of complexes described in both databases are of human origin (66.36% of CORUM and 25.79% of Complex Portal) Thus, currently, ComplexBrowser is most suitable for analysis of human proteins. This is visible when comparing the number of proteins found to be involved in complexes from the adenocarcinoma (human) and T-cell (mouse) datasets, Table I. Additionally the databases contain entries that are not fully annotated. Further developments of the databases will improve the results provided by the software.

Acknowledgments—We thank Ole N. Jensen and Lauren Elizabeth Smith for their critical comments to the project and manuscript.

DATA AVAILABILITY

The source code can be downloaded from: https://bitbucket.org/michalakw/complexbrowser.

* A.R.-W. was supported by a grant from the Independent Research Fund Denmark - Natural Sciences and VILLUM Foundation for a grant to the VILLUM Center for Bioanalytical Sciences at SDU. V.S. was supported by ELIXIR DK. W.M. was supported by student grant from MC2 Therapeutics ApS.

† This article contains supplemental Figures and Tables.

‡ To whom correspondence should be addressed: Department of Biochemistry and Molecular Biology and VILLUM Center for Bioanalytical Sciences, University of Southern Denmark, Campusvej 55, DK-5230, Odense M, Denmark. E-mail: adelinar@bmb.sdu.dk.

§ Current affiliation: Biostatistics NADs & Complications, Novo Nordisk A/S, Vandtårnsvej 114, DK-2860 Søborg, Denmark.

¶ These authors contributed equally to this work.

Author contributions: W.M., V.T., V.S., and A.R.-W. designed research; W.M., V.T., and V.S. performed research; W.M., V.T., V.S., and A.R.-W. analyzed data; W.M., V.T., V.S., and A.R.-W. wrote the paper.

REFERENCES

1. Bekker-Jensen, D. B., Kelstrup, C. D., Batth, T. S., Larsen, S. C., Haldrup, C., Bramsen, J. B., Sørensen, K. D., Høyer, S., Ørntoft, T. F., Andersen,
C. L. Nielsen, M. L., and Olsen, J. V. (2017) An optimized shotgun strategy for the rapid generation of comprehensive human proteomes. *Cell Systems* 4, 587–599.e584

2. Kelstrup, C. D., Bekker-Jensen, D. B., Arrey, T. N., Hogrebe, A., Harder, A., and Olsen, J. V. (2018) Performance evaluation of the Q Exactive HF-X for shotgun proteomics. *J. Proteome Res.* 17, 727–738

3. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. *The Gene Ontology Consortium. Nat. Genet.* 25, 25–29

4. The Gene Ontology. (2017) Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.* 45, D331–D338

5. Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* 10, 48

6. Kramer, A., Green, J., Pollard, J., Jr, and Tugendreich, S. (2014) Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics* 30, 523–530

7. Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 27, 29–34

8. Fabregat, A., Sidirooulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., Jassal, B., Jupe, S., Körner, F., McKay, S., Matthews, L., May, B., Milacic, M., Rothfels, K., Shamsowski, V., Webber, M., Weiser, J., Williams, M., Wu, G., Stein, L., Hermjakob, H., and D'Eustachio, P. (2016) The Reactome pathway knowledgebase. *Nucleic Acids Res.* 44, D481–D487

9. Rhee, S. Y., Wood, V., Dolinski, K., and Draghici, S. (2008) Use and misuse of the gene ontology annotations. *Nat. Rev. Genet.* 9, 509–515

10. Schmidt, A., Forne, I., and Imhof, A. (2014) Bioinformatic analysis of proteomics data. *BMC Syst. Biol.* 8, S3

11. Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N. T., Roth, A., Bork, P., Jensen, L. J., and von Mering, C. (2017) The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* 45, D362–D368

12. David, A., Razali, R., Wass, M. N., and Sternberg, M. J. (2012) Protein-protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. *Hum. Mutat.* 33, 359–363

13. Ori, A., Iskar, M., Buczak, K., Kastritis, P., Andres-Pons, A., David, A., Razali, R., Wass, M. N., and Sternberg, M. J. (2012) Protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* 40, D331–D338

14. Chawade, A., Alexandersson, E., and Levander, F. (2014) Normalyzer: A tool for rapid evaluation of normalization methods for omics data sets. *J. Proteome Res.* 13, 3114–3120

15. Hochreiter, S., Clevert, D. A., and Obermayer, K. (2006) A new summarization method for affymetrix probe level data. *Bioinformatics* 22, 943–949

16. Zhang, B., Pirmoradian, M., Zubarev, R., and Kall, L. (2017) Covariation of peptide abundances accurately reflects protein concentration differences. *Mol. Cell Proteomics* 16, 936–948

17. Rogowska-Wrzesinska, A., Wrzesinski, K., and Fey, S. J. (2014) Heteromer score-using internal standards to assess the quality of proteomic data. *Proteomics* 14, 1042–1047

18. Boggs, P. T., Spiegelman, C. H., Donaldson, J. R., and Schnabel, R. B. (1988) A computational examination of orthogonal distance regression. *J. Econometrics* 38, 169–201

19. Tan, H. Y., Yang, K., Li, Y. X., Wang, T. I., Wang, Y. Y., Bianco, D. B., Wang, X. S., Cho, J. H., Wang, H., Rankin, S., Guy, C., Peng, J. M., and Chi, H. B. (2017) Integrative proteomics and phosphoproteomics profiling reveals dynamic signaling networks and bioenergetics pathways underlying T cell activation. *Immunity* 46, 488–503

20. Wierniewski, J. R., Ostaszewicz, P., Dus, K., Zielinska, D. F., Gnad, F., and Mann, M. (2012) Extensive quantitative remodeling of the proteome between normal colon tissue and adenocarcinoma. *Mol. Syst. Biol.* 8, 611

21. Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015) limma powers differential expression analyses for RNA-seq and microarray studies. *Nucleic Acids Res.* 43, e47

22. Cox, J., Hein, M. Y., Luber, C. A., Paron, I., Nagaraj, N., and Mann, M. (2012) Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell Proteomics* 13, 1753–1762