DOSCHEDA: a web application for interactive chemoproteomics data analysis

Bruno Contrino[1], Eric Miele[2], Ronald Tomlinson[2], M. Paola Castaldi[2] and Piero Ricchiuto[1]

[1] Department of Quantitative Biology, Discovery Sciences, AstraZeneca, Cambridge, United Kingdom
[2] Department of Chemical Biology, Discovery Biology, Discovery Sciences, AstraZeneca, Waltham, MA, United States of America

ABSTRACT

Background. Mass Spectrometry (MS) based chemoproteomics has recently become a main tool to identify and quantify cellular target protein interactions with ligands/drugs in drug discovery. The complexity associated with these new types of data requires scientists with a limited computational background to perform systematic data quality controls as well as to visualize the results derived from the analysis to enable rapid decision making. To date, there are no readily accessible platforms specifically designed for chemoproteomics data analysis.

Results. We developed a Shiny-based web application named DOSCHEDA (Down Stream Chemoproteomics Data Analysis) to assess the quality of chemoproteomics experiments, to filter peptide intensities based on linear correlations between replicates, and to perform statistical analysis based on the experimental design. In order to increase its accessibility, DOSCHEDA is designed to be used with minimal user input and it does not require programming knowledge. Typical inputs can be protein fold changes or peptide intensities obtained from Proteome Discover, MaxQuant or other similar software. DOSCHEDA aggregates results from bioinformatics analyses performed on the input dataset into a dynamic interface, it encompasses interactive graphics and enables customized output reports.

Conclusions. DOSCHEDA is implemented entirely in R language. It can be launched by any system with R installed, including Windows, Mac OS and Linux distributions. DOSCHEDA is hosted on a shiny-server at https://doscheda.shinyapps.io/doscheda and is also available as a Bioconductor package (http://www.bioconductor.org/).

Subjects Bioinformatics

Keywords Quantitative Chemoproteomics, Statistical Models, Web interface, Shiny, TMT, iTRAQ, Proteomics, Quantitative Chemical biology, Drug dose-response, Protein drug profiling

BACKGROUND

Many drugs fail in the clinic for lack of efficacy or for toxicity and as such, some of the most important steps in drug discovery are evaluation of target engagement and off-targets liabilities. Next Generation Sequencing (NGS) and mass spectrometry proteomics-based drug discovery (*Jones & Neubert, 2017*) approaches offer unique

opportunities to deeply characterize drug-targets biology and pharmaceutical agents (*Bantscheff et al., 2007*; *Bantscheff et al., 2008*).

Quantitative chemical proteomics (*Bantscheff & Drewes, 2012*) and MS-Cellular Thermal Shift Assay (MS-CETSA) (*Martinez & Nordlund, 2016*) have been recently deployed in drug discovery to aid target deconvolution in the context of phenotypic screens, to elucidate drugs' mechanisms of action and evaluation of drug repurposing (*Schirle, Bantscheff & Kuster, 2012*).

A generous number of computational tools has been developed for MS-spectral analysis and protein quantification, such as Proteome Discoverer (https://www.thermofisher.com/), MaxQuant (*Cox & Mann, 2008*) or PEAKS (http://www1.bioinfor.com) for de novo peptide sequencing. However, the increasing variety of MS based approaches for drug target deconvolution can produce data that need dedicated downstream analysis platforms for facilitating the biological interpretation of results.

Here, we focus on quantitative chemoproteomics used to determine protein interaction profiles of small molecules from whole cell or tissue lysates (*Manning, 2002*). Chemoproteomics includes reverse competition-based experiments that, in combination with quantitative MS (e.g., tandem mass tags (TMT) or isobaric tag for relative quantification (iTRAQ)), are used for rank-ordering of small molecule-protein interactions by binding affinity (*Bantscheff & Drewes, 2012*).

Although several comprehensive analysis pipelines, such as OCAP (*Wang, Yang & Yang, 2012*), ProSightPC (http://proteinaceous.net/software/), TopPIC (*Kou, Xun & Liu, 2016*), MSstats (*Choi et al., 2014*), Skyline (*MacLean et al., 2010*), MaxQuant & Perseus (*Tyanova et al., 2016*) and DAPAR & ProStaR (*Wieczorek et al., 2017*), have been developed for the downstream data analysis, to the best of our knowledge there are no tools specifically designed to facilitate chemoproteomics data analysis for scientists with a limited computational background and available as a public server application.

Based on this, we have developed a Shiny-based web application named DOSCHEDA (Down Stream Chemoproteomics Data Analysis), which includes: (i) an open-source code available on Bioconductor for R-users; (ii) a user-friendly Graphical User Interface (GUI) with no programming knowledge required (iii) a traffic-light system to enable the user to rapidly identify and address data incongruences; (iv) an OS-independent implementation which generates a comprehensive final report in addition to analysis results; (v) a flexible data-input routine which enables the user to import different file types (.txt, .xlsx, .csv), typically exported from MS-software such as Proteome Discover or MaxQuant; (vi) the CRAPome flagged proteins based on the contaminant repository database (*Mellacheruvu et al., 2013*)

DOSCHEDA addresses the need to perform fit for purpose statistical analysis of chemoproteomics experiments, including linear and non-linear models, to provide a ranking of the protein(s) most competed by the investigational compound (the competitor) as well as to generate standardized results that can be further used for downstream analysis or for different experiment comparisons.

Table 1 Required inputs for DOSCHEDA.

	Input data from PD2.1	Input data not from PD2.1
Peptide intensities	Peptide qvality score, protein accessions, peptide names, intensities	Peptide qvality score, protein accessions, peptide names, intensities
Fold changes	Protein fold changes	Protein accessions, protein fold changes, unique peptides
Log-fold changes	Protein log-fold changes	Protein accessions, protein log-fold changes, unique peptides

IMPLEMENTATION

DOSCHEDA is implemented using R (*R Core Team, 2014*), an open-source software for statistical computing and coded in Shiny (*Chang et al., 2015*).

DOSCHEDA processes the data based on a series of pipelines developed and integrated into the application. The user selects what pipeline to be executed based on the experimental design and data input.

Data input

The application is designed to take three different types of data:

1. **Peptide Intensities**. These are obtained from Proteome Discoverer, MaxQuant or similar software. The same procedure applied in Proteome Discoverer 2.1 has been implemented in DOSCHEDA for summing the reporter ions to protein relative quantification. The protein fold changes [Ctrl]/[treated] are then converted to log2 scale and then passed into the pipeline.
2. **Fold Changes.** These are the protein fold changes, [Ctrl]/[treated].
3. **Log2 Fold Changes**. These are the log2 protein fold changes.

DOSCHEDA has been optimized for Proteome Discoverer 2.1 (PD 2.1), but it can also use data from other software given that the input file contains specific columns as described in Table 1.

Statistical analyses

The experimental design as shown in Table 2 will determine the type of analysis that can be performed in DOSCHEDA. The standard pipeline utilizes a linear model with a quadratic form

$$f(x) = a_0 + a_1 x + a_2 x^2$$

where a_0 is the intercept, a_1 the slope and a_2 the quadratic coefficient. This pipeline is suitable for experimental designs with less than 5 channels (4-plex) and more than 1 replicate (Table 2). Increasing free drug concentrations and generation of a dose-response relationship with the protein target(s) provides information about specific drug-target interactions. The linear model analysis implemented in DOSCHEDA by using the limma R package (*Smyth, 2004*) will identify proteins with a significant p value (slope, intercept, quadratic) as the protein drug-target(s).

Alternatively, in cases where biological input material is not a limiting factor, chemoproteomics experiments consist of a full-scale dose-response, in which 5 or more

Table 2 Type of analysis performed in DOSCHEDA.

	1 Replicate	More than 1 replicate
Less than 5 channels	Not enough data	Linear model
5 or more channels	Sigmoidal	Linear model

concentrations are tested (e.g., 6-plex or 10-plex). In this case, the user can run the sigmoidal pipeline where the four-parameter log- logistic (*Ritz et al., 2015*) model is applied (Table 2). Here, the *p* values of the model parameters (min, max, slope, RB_{50}) will be computed to rank proteins based on their selectivity profile (*Smyth & Collins, 2009*). The half maximal residual binding (RB_{50}) is a measure of the effectiveness of a drug in binding to a protein. Thus, this quantitative measure indicates how much of a drug or small molecule is needed to saturate binding to a protein by half, and can be used to compare drug-protein profiles. The RB_{50} values are typically expressed as molar concentration and are computed in the sigmoidal pipeline for each protein within DOSCHEDA. Furthermore, the corrected RB_{50}, according to *Daub (2015)* corresponds to the ratio (r) of the proteins enriched in the second incubation (supernatant) versus those retained in the first incubation (DMSO or blank) with the affinity matrix. This pulldown of pulldown or depletion factor (r) enables the calculation of a K_d for each protein and it will be part of DOSCHEDA's outputs.

Peptide removal process using Pearson correlation

When peptide intensities and two replicates are available, DOSCHEDA implements a prior step to the peptide aggregation process to consider of potential technical experimental errors.

The peptide removal function is tailored having in mind the typical chemoproteomics experimental design and aims to leverage the empirical peptides information of experimental replicates avoiding data imputation or using spectra features of peptides (*Fischer & Renard, 2016*) as they might not be necessarily available information to the final user.

The peptide removal algorithm is described in Fig. 1, in this procedure peptides that are inconsistent (e.g., anti-correlate) between replicates are excluded. This implementation is not intended to address the ratio compression (*Savitski et al., 2013*) but to leverage the information available on the same peptide between the two replicates.

In fact, the peptide removal process is based on the calculation of the Pearson correlation coefficient between the same peptide abundances of two replicates as a pre-filtering step before summing the remaining peptide intensities to finally infer back the protein abundance.

Initially, the peptides are filtered by their Peptide Quality score ($Q_s < 0.05$) column, a mandatory input column in this case. Next, data are filtered to have a minimum of 2 peptides (shared or unique) per protein and per replicate such that the Pearson correlation coefficient (R^2) can be calculated between matched peptides. Peptides with $R^2 < 0.4$ are removed and summed intensities are computed from the remaining peptides; although this cut-off can be modified by the user within DOSCHEDA, we observed that 0.4 is a reasonable default threshold that removes extreme peptide quantifications (e.g., low correlated and
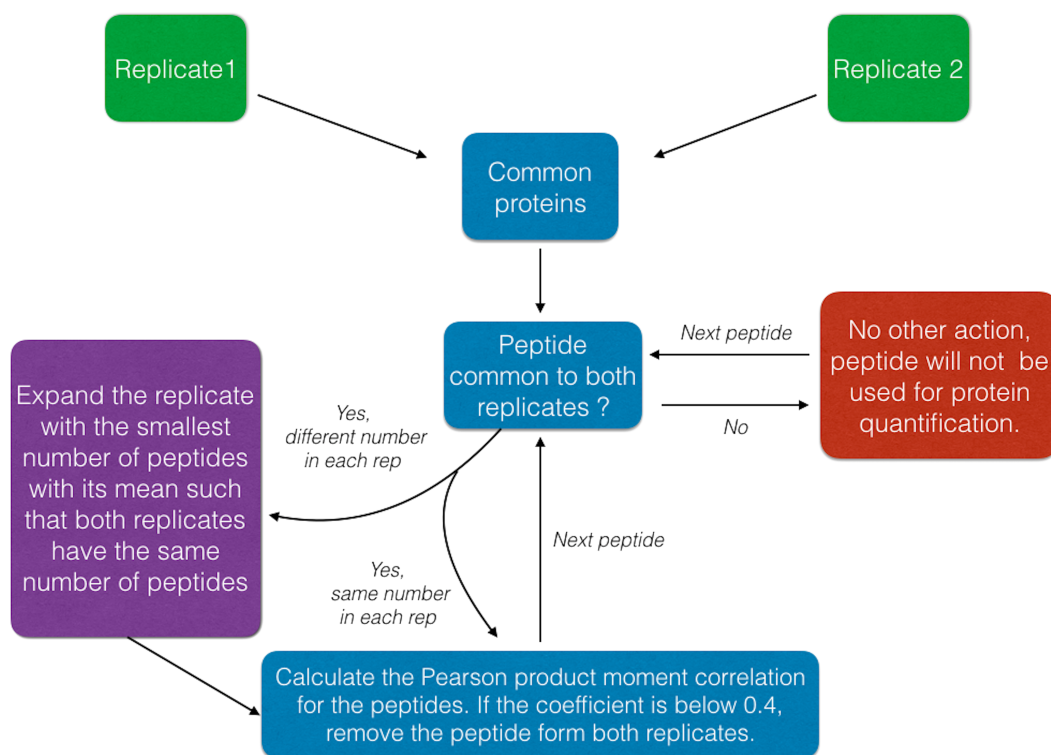
Figure 1 Schematic of the Peptide Removal process algorithm.

Table 3 Summary scenarios in peptide removal process.

| Prot. name	Peptides Rep1	Peptides Rep2	Pearson correlation coefficient	Action
Protein 1	A, B	A, B	$R_A > 0.4$ $R_B > 0.4$	No removal, computed the summed intensities as in PD2.1
Protein 2	A, B	A	NULL	Lack of peptide measurements; Protein 2 will be removed
Protein 3	A, B, C	A, B, C	$R_A > 0.4$ $R_B > 0.4$ $R_C < 0.4$	A, B considered and summed. C is removed, reason of noise in the data

Notes.
A,B,C in the real case scenario are peptide sequences.

anti-correlated peptides) and keeps almost unchanged the proteome coverage, as shown in Fig. 2 $R^2 < 0.4$. Intuitively, larger R^2 values will reflect into a larger peptide removal and consequently proteins loss as shown by Figs. 2A–2D.

The R^2 can be adjusted by the user within DOSCHEDA allowing quick results comparison in a similar fashion of Fig. 2 and, empowering the analyst to fine tuning this threshold based on project aims.

Finally, only proteins with more than one unique peptide are retained for the downstream analysis. Table 3 summarizes all the possible scenarios handled by the peptide removal function.
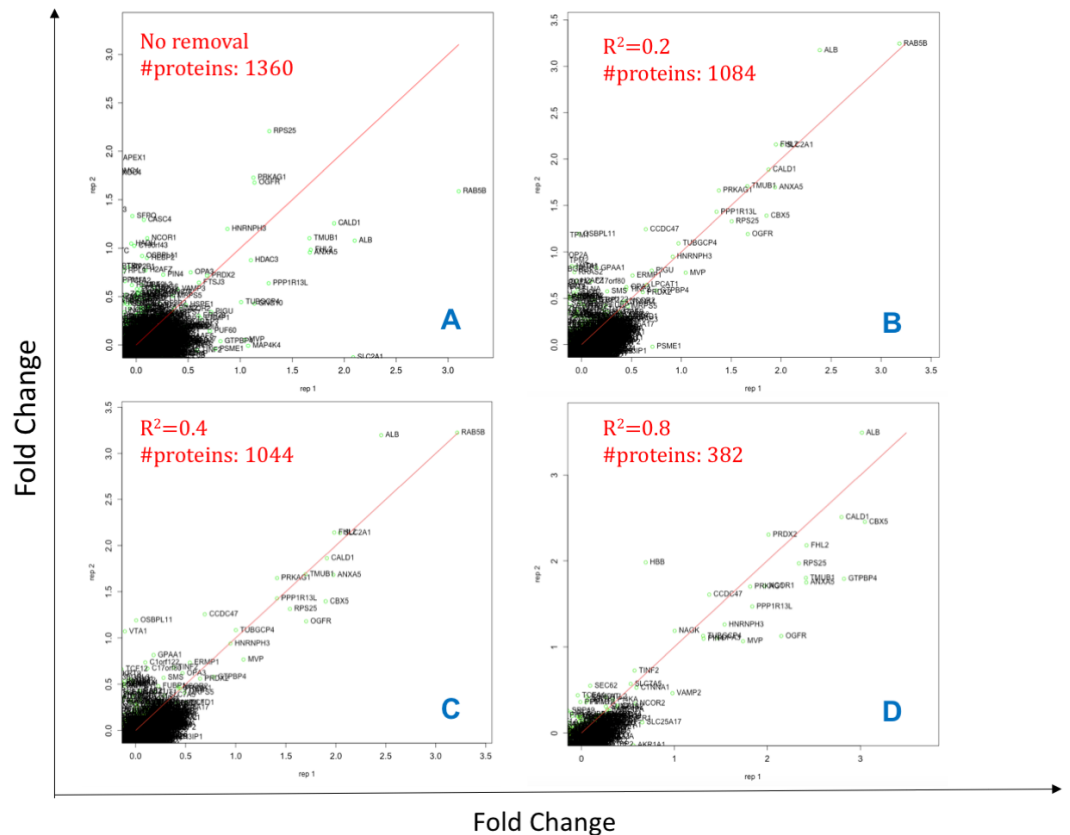
Figure 2 **Effect of different Pearson coefficient cut offs.** X- and Y-axis are Log2 fold changes. The total number of proteins 1,360 (A, 100%) reduces to 1,084 with Pearson coefficient $R^2 = 0.2$ (B, $\sim$80%), 1,044 with $R^2 = 0.4$ (C, $\sim$77%), and 382 with $R^2 = 0.8$ (D, 28%).

The user should initially run the exploratory data analysis in DOSCHEDA up to the Principal Component Analysis (PCA) and the correlogram plot to evaluate the correlation between replicates. Only at this point the user can make an informed decision whether to apply the peptide removal function or continue to analyze the data without applying the removal process. Yet, being an open source tool the peptide removal process could be replaced by any other similar function that might be available in literature (see the DOSCHEDA Bioconductor vignette).

Additional Features. The server version of DOSCHEDA comprises additional uploads that are useful when researchers are comparing datasets, including: (i) intersections of the enriched proteins with a user uploaded GeneID list (e.g., protein kinesis); (ii) default mapping from Uniprot accession to GeneIDs using InterMine or in case the organism of interest is not present in the drop-down list, DOSCHEDA allows the user to specify the mapping file.

Furthermore, DOSCHEDA allows: (i) two normalization/scaling options (see DOSCHEDA manual); (ii) interactive 2D plots with user-defined x- and y- axes; (iii) quality control (QC) traffic lights flags; (iv) downloadable results and detailed report of the analyses.
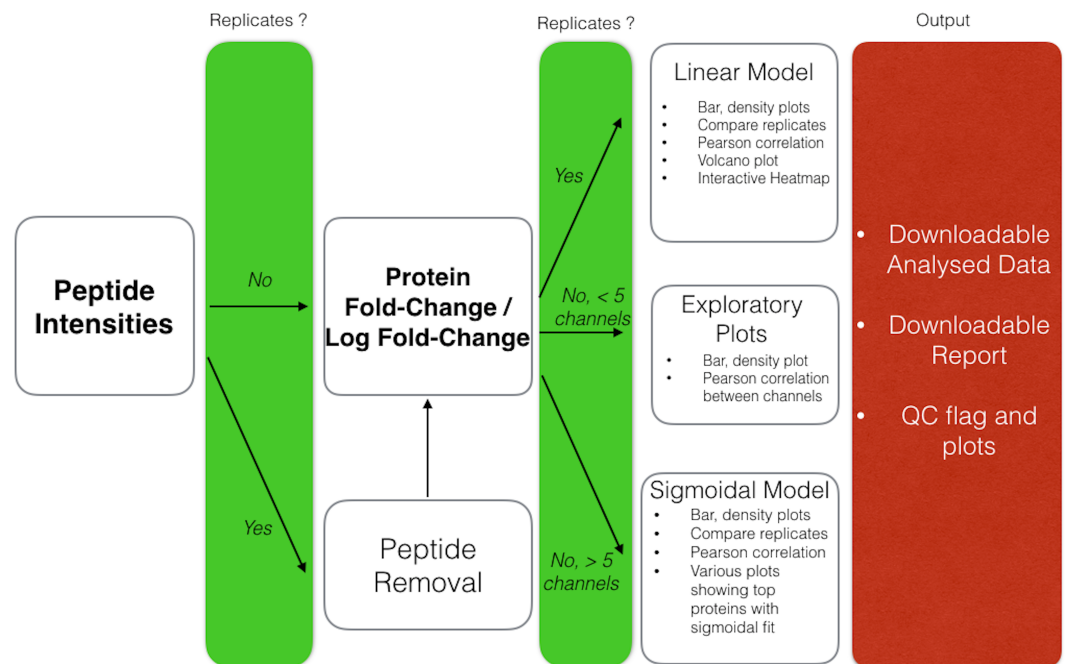
Possible inputs in **bold**

Figure 3 DOSCHEDA workflow.

RESULTS

DOSCHEDA generates a variety of different plots depending on the user's data input, as outlined in Fig. 3. Overall there are two sets of output plots, one designed for the quality control (QC), the other one to visualize the results of the analyzed data.

The DOSCHEDA application provides standard QC plots such as box, whisker plots, data distributions within each iTRAQ/TMT channel, correlogram plots to quantify the correlation between replicates, a 2D-plot to verify the variance independence from the data mean, and a plot of the first two principal components. Based on the QC outcomes, if data are consistent the user can proceed to the inspection of the statistical analysis results, otherwise, if peptide intensities are available the peptide removal process can be applied.

For the analysis, depending on the chosen statistical model, DOSCHEDA will generate different outputs.

In case of a linear model the p values distributions for the three coefficients (a_0, a_1 and a_2) and their corresponding interactive volcano plots are displayed.

In case of a sigmoidal model DOSCHEDA will output three plots: the first showing the sigmoidal curves with a difference higher than 30% between the top and the bottom value; the second and the third showing the proteins that have a significant RB_{50} (p value <0.05) and a significant slope (p value <0.05), respectively.

Independently of the applied statistical model, the R package *d3heatmap* is used to generate an interactive heatmap of the input data.
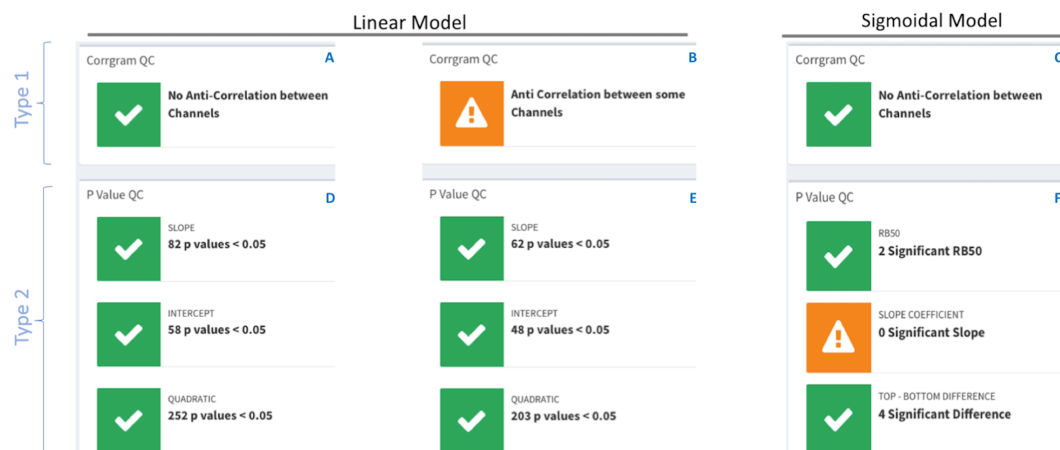
Figure 4 DOSCHEDA traffic lights flags system. Two types of QCs' are available, type 1 based on pairwise samples correlation obtained from the correlogram plot (A–C) and, type 2 based on model's p values coefficients (D–E) obtained by running the linear or sigmoidal statistical analysis. In both types, green flags will be displayed when there is no anti-correlation between any samples and the number of proteins with p value(s) <0.05 is larger than zero. On the contrary orange flags will be displayed when anti-correlation is observed (B) in at least one of the pairwise samples (i.e., Pearson Correlation Coefficient, $R^2 < 0$) and/or there are no proteins that passes the threshold of p value(s) < 0.05 (F).

Once the analysis has been completed, the summary section of the application allows the user to visualize the results in a table format, where functions like "search" and "sorting" are also enabled for quick queries.

The summary section contains quality control (QC) traffic light flags with green for data consistency and orange for data incongruences with an accompanying warning message. There are two types of traffic lights flags; see Fig. 4.

Finally, in the download section of the application the user can download (i) a .csv file containing the results of the analysis which also includes the user-input dataset and (ii) an html report with all the plots produced in the current run, the key tables as well as the session information to facilitate data reproducibility.

CONCLUSION

DOSCHEDA enables researchers with limited programming experience to perform, evaluate, and interactively visualize chemoproteomics data analysis. DOSCHEDA includes linear and non-linear statistical analysis whose results can be exported in excel spreadsheet format. Also, the user will be able to generate a full report of the executed data analysis, facilitating data reproducibility. Being open source, DOSCHEDA can be easily extended and modified to fit specific additional analyses.

Availability and requirements

DOSCHEDA server lives at https://doscheda.shinyapps.io/doscheda/ and it will be constantly maintained by the authors; the users by following the link above can upload and analyze their own datasets without additional requirements. Finally, to facilitate traceability and reproducibility DOSCHEDA is also available as an open source Bioconductor package.

Abbreviations

DOSCHEDA	Down Stream Chemoproteomics Data Analysis
QC	Quality Controls
iTRAQ	Isobaric tag for relative and absolute quantitation
TMT	Tandem Mass Tag
MS	Mass Spectrometry
GUI	Graphical User Interface

Competing Interests

Bruno Contrino, Eric Miele, Ronald Tomlinson, M. Paola Castaldi, and Piero Ricchiuto are employees of Discovery Science, AstraZeneca.

Author Contributions

- Bruno Contrino analyzed the data, contributed reagents/materials/analysis tools, prepared figures and/or tables, performed the computation work.
- Eric Miele performed the experiments, contributed reagents/materials/analysis tools.
- Ronald Tomlinson contributed reagents/materials/analysis tools, reviewed drafts of the paper.
- M. Paola Castaldi contributed reagents/materials/analysis tools, reviewed drafts of the paper.
- Piero Ricchiuto conceived and designed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, performed the computation work, reviewed drafts of the paper.

Data Availability

The following information was supplied regarding data availability:
The code, manual, walkthrough descriptions and dummy data are available here:
https://github.com/brunocontrino/DOSCHEDA_APP.
Bioconductor repository and package status:
https://github.com/Bioconductor/Contributions/issues/406.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.129#supplemental-information.

REFERENCES

Bantscheff M, Boesche M, Eberhard D, Matthieson T, Sweetman G, Kuster B. 2008. Robust and sensitive iTRAQ quantification on an LTQ Orbitrap mass spectrometer. *Mol Cell Proteomics* **7(9)**:1702–1713 DOI 10.1074/mcp.M800029-MCP200.

Bantscheff M, Drewes G. 2012. Chemoproteomic approaches to drug target identification and drug profiling. *Bioorganic & Medicinal Chemistry* **20(6)**:1973–1978 DOI 10.1016/j.bmc.2011.11.003.

Bantscheff M, Eberhard D, Abraham Y, Bastuck S, Boesche M, Hobson S, Mathieson T, Perrin J, Raida M, Rau C, Reader V, Sweetman G, Bauer A, Bouwmeester T, Hopf C, Kruse U, Neubauer G, Ramdsen N, Rick J, Kuster B, Drewes G. 2007. Quantitative chemical proteomics reveals mechanisms of action of clinical ABL kinase inhibitors. *Nature Biotech* **25**:1035–1044 DOI 10.1038/nbt1328.

Chang W, Cheng J, Allaire J, Xie Y, McPherson J. 2015. Shiny: web application framework for R. R package version 0.12. *Available at http://CRAN.R-project.org/package=shiny*.

Choi M, Chang CY, Clough T, Broudy D, Killeen T, MacLean B, Vitek O. 2014. MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics* **20(17)**:2524–2526.

Cox J, Mann M. 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology* **26**:1367–1372 DOI 10.1038/nbt.1511.

Daub H. 2015. Quantitative proteomics of kinase inhibitor targets and mechanisms. *Chamical Biology (ACS)* **10**:201–212 DOI 10.1021/cb5008794.

Fischer M, Renard BY. 2016. iPQF: a new peptide-to-protein summarization method using peptide spectra characteristics to improve protein quantification. *Bioinformatics* **32(7)**:1040–1047 DOI 10.1093/bioinformatics/btv675.

Jones LH, Neubert H. 2017. Clinical chemoproteomics—opportunities and obstacles. *Science Translational Medicine* **9(386)**:eaaf7951 DOI 10.1126/scitranslmed.aaf7951.

Kou Q, Xun L, Liu X. 2016. TopPIC: a software tool for top-down mass spectrometry-based proteoform identification and characterization. *Bioinformatics* **15(32(22))** 3495–3497.

MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, Frewen B, Kern R, Tabb DL, Liebler DC, MacCoss MJ. 2010. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26(7)**:966–968 DOI 10.1093/bioinformatics/btq054.

Manning. 2002. The protein kinase complement of the human genome. *Science* **298**:1912–1934 DOI 10.1126/science.1075762.

Martinez MD, Nordlund P. 2016. The cellular thermal shift assay: a novel biophysical assay for *in situ* drug target engagement and mechanistic biomarker studies. *Annual Review of Pharmacology and Toxicology* **56**:141–161 DOI 10.1146/annurev-pharmtox-010715-103715.

Mellacheruvu D, Wright Z, Couzens AL, Lambert JP, St-Denis NA, Li T, Miteva YV, Hauri S, Sardiu ME, Low TY, Halim VA, Bagshaw RD, Hubner NC, Al-Hakim A, Bouchard A, Faubert D, Fermin D, Dunham WH, Goudreault M, Lin ZY, Badillo BG, Pawson T, Durocher D, Coulombe B, Aebersold R, Superti-Furga G, Colinge J, Heck AJ, Choi H, Gstaiger M, Mohammed S, Cristea IM, Bennett KL, Washburn MP, Raught B, Ewing RM, Gingras AC, Nesvizhskii AI. 2013. The CRAPome: a contaminant repository for affinity purification mass spectrometry data. *Nature Methods* **10**:730–736 DOI 10.1038/nmeth.2557.

R Core Team. 2014. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. *Available at http://www.r-project.org*.

Ritz C, Baty F, Streibig JC, Gerhard D. 2015. Dose-response analysis using R. *PLOS ONE* **10(12)**:e0146021 DOI 10.1371/journal.pone.0146021.

Savitski MM, Mathieson T, Zinn N, Sweetman G, Doce C, Becher I, Pachl F, Kuster B, Bantscheff M. 2013. Measuring and managing ratio compression for accurate iTRAQ/TMT quantification. *Journal of Proteome Research* **12(8)**:3586–3598 DOI 10.1021/pr400098r.

Schirle M, Bantscheff M, Kuster B. 2012. Mass spectrometry-based proteomics in preclinical drug discovery. *Chemistry and Biology* **27(19)**:72–84 DOI 10.1016/j.chembiol.2012.01.002.

Smyth GK. 2004. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **3**: Article 3.

Smyth LA, Collins I. 2009. Measuring and interpreting the selectivity of protein kinase inhibitors. *Journal of Chemical Biology* **2(3)**:131–151 DOI 10.1007/s12154-009-0023-9.

Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, Mann M, Cox J. 2016. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature Methods* **13**:731–740 DOI 10.1038/nmeth.3901.

Wang P, Yang P, Yang JY. 2012. OCAP: an open comprehensive analysis pipeline for iTRAQ. *Bioinformatics* **28(10)**:1404–1405 DOI 10.1093/bioinformatics/bts150.

Wieczorek S, Combes F, Lazar C, Gianetto QG, Gatto L, Dorffer A, Hesse AM, Couté Y, Ferro M, Bruley C, Burger T. 2017. DAPAR & ProStaR: software to perform statistical analyses in quantitative discovery proteomics. *Bioinformatics* **33(1)**:135–136 DOI 10.1093/bioinformatics/btw580.