BMC Bioinformatics

PAWER: protein array web exploreR

Dmytro Fishman[1,2], Ivan Kuzmin[1], Priit Adler[1,2], Jaak Vilo[1,2] and Hedi Peterson[1,2*]

*Correspondence:
hedi.peterson@ut.ee
[1]Institute of Computer Science,
University of Tartu, Narva mnt 18,
51009 Tartu, Estonia
[2]Quretec Ltd, Ülikooli 6a, 51003
Tartu, Estonia

Abstract

Background: Protein microarray is a well-established approach for characterizing activity levels of thousands of proteins in a parallel manner. Analysis of protein microarray data is complex and time-consuming, while existing solutions are either outdated or challenging to use without programming skills. The typical data analysis pipeline consists of a data preprocessing step, followed by differential expression analysis, which is then put into context via functional enrichment. Normally, biologists would need to assemble their own workflow by combining a set of unrelated tools to analyze experimental data. Provided that most of these tools are developed independently by various bioinformatics groups, making them work together could be a real challenge.

Results: Here we present PAWER, the online web tool dedicated solely to protein microarray analysis. PAWER enables biologists to carry out all the necessary analysis steps in one go. PAWER provides access to state-of-the-art computational methods through the user-friendly interface, resulting in publication-ready illustrations. We also provide an R package for more advanced use cases, such as bespoke analysis workflows.

Conclusions: PAWER is freely available at https://biit.cs.ut.ee/pawer.

Keywords: Protein microarray, Data analysis, Web tool, Normalisation, Visualisation

Background

Protein microarray is the leading high-throughput method to study protein interactions [1], antibody specificity or autoimmunity [2]. In functional protein microarrays, full-length functional protein targets or protein domains are attached to the surface of the slide and then incubated with a biological sample that contains interacting molecules (e.g. autoantibodies) [3]. After molecules bind to their targets, labelling is done via secondary antibody with a fluorescent marker attached. Resulting fluorescent signal of high intensity indicates the reaction, which can be registered by the specialised scanner. The most popular microarray platforms (e.g. Human Proteome Microarray (HuProt), ProtoArray, NAPPA arrays, Human Protein Fragment arrays and Immunome arrays) allow to measure autoantibody reaction to thousands of unique human protein abundances simultaneously [4, 5].

Hundreds of studies that use different types of protein microarrays are conducted every year [6]. All these studies largely depend on well executed data analysis. Usual analysis workflow starts with pre-processing of raw data obtained from GenePix Pro - one of *de facto* standard softwares used to read the microarrays [7]. The pre-processing step involves quality control and normalisation. It is followed by the differential protein analysis, in which protein reactivity levels that are significantly different between studied conditions are identified. These reactive protein levels are visualised, e.g. with boxplots. Finally, the results are interpreted using the body of prior knowledge via applying functional enrichment analysis tools. Setting up and executing these steps requires a lot of time and care from the researchers as each analysis step needs to be documented to ensure reproducibility.

Protein microarrays are similar to DNA microarrays as both technologies measure abundance of thousands of probes immobilised on the surface of the slide [8]. In the early days of protein microarrays research, this technological resemblance allowed practitioners to adapt methods and computational tools, originally developed for DNA microarrays [9]. However, a number of studies have shown that the same set of assumptions is not necessarily applicable to both types of microarrays, especially in terms of normalisation [5, 8, 9]. For example, in DNA microarrays the overall amount of signal is considered to be roughly the same between samples, while in protein microarrays only a small number of proteins are expected to show reactivity to probed serum. Applying quantile normalisation, that is usually utilised in DNA microarray analysis, may eliminate the relevant biological signal [8]. Thus, analytical pipelines tailored to protein microarrays are required in order to enable correct data analysis and consequently, biologically relevant results.

To date, four major tools for protein microarray analysis are Prospector, Protein Array Analyser (PAA) [10], Protein Microarray Analyser (PMA) [5] and online tool available as part of Protein Microarray Database (PMD) [11]. Prospector, provided by ThermoFisher Scientific, allows easy point and click analysis. However, it has not been updated since 2015, is a closed source software and runs only on the Windows 7 operating system [12]. PAA [10] builds on top of Prospector's core functionality, and provides workflow customisation and tools for biomarker discovery in R. Although PAA is flexible and robust, it requires substantial programming skills from the user. PMA is a multi-platform desktop application, built in Java and published in 2018. It can be used via simple graphical user interface as well as executed from the command line. Although, PMA implements state-of-the-art normalisation and pre-processing strategies, working with it can be challenging, as to this date no relevant documentation is available. The only web-based tool for analysing protein microarray experiments, developed prior to current work can be found on Protein Microarray Database website [11]. Unlike previously mentioned software packages, PMD tool offers an all-encompassing analysis according to the original publication. The PMD website openly prioritises depositing and archiving of protein microarray datasets, but its accompanying analysis tool lacks user-interaction and clear guidance.

Here, we present Protein Array Web Explorer (PAWER), the only interactive web tool solely dedicated to analysing protein microarray data. PAWER builds upon the strengths of previously-described tools, while eliminating their major limitations. PAWER is suitable for experimental biologists who want to analyse their own data without the need to

write code. PAWER has already been used for multiple projects, with the underlying R codebase central for analysis in two recent studies of APECED syndrome [2, 13].

Implementation

Key features

PAWER implements the following key features:

1. Public web service that can be used by anyone with protein microarray data in standard format
2. Interactive results table for convenient exploration of the results
3. Clear interactive visuals that can be downloaded in publication-ready formats
4. Parameterised algorithms at key steps (robust linear model (RLM), moderated T-test [14])
5. Downloadable intermediate results after each analysis step
6. Connection with g:Profiler [15] tool through its R package (gprofiler2) for fast enrichment analysis of differential protein features
7. An open source R package that the PAWER web service is built upon

Data upload and preprocessing

To start using PAWER, the user first needs to upload the fluorescent signal array readings - GenePix Results (GPR) files by either dragging and dropping files into the upload area or selecting them directly from the file system (via file upload window). Upon upload, PAWER automatically checks if submitted files come from the same platform and have the same extension. Detailed error message is shown in case any of these assumptions are not met. Once files have been successfully uploaded, the user is asked to select features that represent foreground and background intensities. In the case of ProtoArray and HuProt platforms, these values are chosen automatically, for other platforms user may have to manually search through the possible options from the drop-down menu. As soon as this is done, a global data matrix for the entire experiment is assembled from uploaded files using the limma R package [14]. Next, the background intensities are subtracted from the foreground values and signal from technical replicates is averaged. Resulting values are then log-transformed.

To reduce the technical noise in the data, we used a robust linear model trained on the set of protein features that are assumed to exhibit constant level of signal regardless of biological differences between samples. Such proteins are called positive controls and used in most of the platforms. Usually they are uniquely denoted in GPR files so that computer algorithms could identify them automatically. Hence, after files are uploaded, PAWER searches for such proteins and creates a list of potential positive controls. The list is then shown to the user for validation. User can alter it, by either removing or adding individual proteins. Robust linear model [8] is then used to predict the signal of control proteins based on their location (array and block) and type. In an ideal noise-free scenario, the resulting model will rely solely on protein type when predicting its signal, as any non-negative coefficient associated with array index or block id would indicate technical bias. In practice, unfortunately, noise is hard to avoid. Therefore, non-zero coefficients associated with individual protein arrays and blocks are subtracted from corresponding protein signals to remove technical bias. Data upload and normalisation steps normally take a few

minutes, for example it took about 2 minutes to preprocess a dataset of size 770 Mb, with 100 samples.

After the normalisation step is complete, user can download the normalised data as a separate file. The file can be used as an input to other tools for additional analysis. Namely, in order to enable more elaborate cluster analysis, PAWER is linked to ClustVis [16]. ClustVis is a stand-alone online tool for cluster analysis and visualisation. ClustVis implements heatmaps and principal component analysis.

The final step in the PAWER data analysis pipeline is differential expression analysis which aims to identify proteins, which signal levels significantly differ between the sample groups. To execute this step, metadata (e.g. information about patients and controls) for each sample is required. User can either upload a separate metadata file or manually annotate every GPR file using the set of radio-buttons. The metadata file should contain only two columns: the list of filenames and corresponding sample groups.

PAWER output

Differential protein features are identified using a moderated t-test, implemented using limma R package [14]. In order to perform a moderated t-test, the number of samples must be larger than the number of conditions (at least by one). Therefore, PAWER requires at least three samples (in total) to perform the differential analysis. To account for multiple testing, obtained p-values are adjusted by the Benjamini-Hochberg method. Proteins with adjusted p-values of less than 0.05 are considered significant and shown to the user in a table. User can filter the table by any value (e.g. protein name) and sort each field. By default it is sorted by the adjusted p-value. Results can be downloaded as a CSV text file for further analysis, as an Excel file to supplement a publication or as a PDF file to include into a presentation. To explore the underlying data distribution, individual protein expression values are visualised using interactive boxplots, which can be downloaded in a form of a publication-ready figure.

Additionally, enrichment of differential proteins is enabled by the gprofiler2 R package that provides interface for g:Profiler service [15]. g:Profiler gives functional enrichment results from a number of different categories, such as Gene Ontology [17], pathways and other structured data sources for instance KEGG [18], Reactome [19], Human Phenotype Ontology [20] and Human Protein Atlas [21]. The six most significant terms are visualised as a downloadable bar plot figure. The complete list of significantly enriched terms is accessible at the g:Profiler website.

Tool development

We developed the PAWER web service as a tool that covers all the necessary steps in protein microarray analysis. Its core has been implemented using R version 3.4.2, limma [22] (v. 3.34.4) for reading in the GPR format files and performing differential analysis, MASS [23] (v. 7.3.47), reshape2 [24] (v. 1.4.2) for normalisation and preprocessing of protoarray data and gprofiler2 [15] (v. 0.1.4) to enable protein identifier conversion and enrichment analysis. Web interface was implemented as a single page application using React.js and Redux architecture with node.js on the server side. Figures have been created and rendered with a help of D3.js [25] and DataTables libraries. Both the R package and the web server code are freely available under the GNU GPL v2. license.

Supported platforms

PAWER has been initially designed to support data produced mainly by ProtoArray and HuProt platforms. Later, support for the ArrayCam imaging system was added on request. Eventually, the decision was made to support as many platforms as possible by enabling customization at every step of the pipeline. Therefore, PAWER is in principle compatible with any protein microarray system or technology as long as the latter outputs text files with identical headers for each sample, and user knows several key properties of the system (background, foreground intensities and control proteins).

Comparison to existing tools

To the best of our knowledge there are five available tools, dedicated to protein microarray analysis — Prospector, PAA, PMA, PMD and PAWER. All the alternatives perform protein array specific normalisation and all but one (PMA) have capacity to identify potential biomarkers. The detailed comparison of the key features is highlighted in Table 1. Prospector was the first protein microarray analysis tool on the market, introduced by the Invitrogen company. It was originally developed for the Windows XP and later in 2015 updated to be compatible with Windows 7. Strict operating system dependency makes the number of potential Prospector users limited. In 2013 an R package, called PAA emerged [10]. Now users, independent from the platform, had an opportunity to design and apply custom analysis pipelines for their protein microarrays. At the same time, PAA requires users to be familiar with R programming language. Another tool, PMA - a Java desktop application, provided a graphical user interface and implemented cutting edge preprocessing techniques. However, it lacks documentation and does not allow for the integrated downstream analysis [5]. Finally, Protein Microarray Database website offers a possibility to analyse protein microarray experiments using their online tool. According to the original publication, PMD offers functionality for enrichment analysis, detection of differentially expressed proteins and generation of user reports based on the results [11]. However, upon closer examination, we were not able to execute the analysis using available GPR files and thus failed to confirm these claims. Neither documentation page nor original publication provide exhaustive details as to which specific methods were implemented in PMD. Also a pdf file with guidelines for interpreting the output of the tool linked from the help page was not accessible. In response to all the

Table 1 Comparison between currently available protein microarray analysis tools: Prospector, PAA, PMA, PMD and PAWER
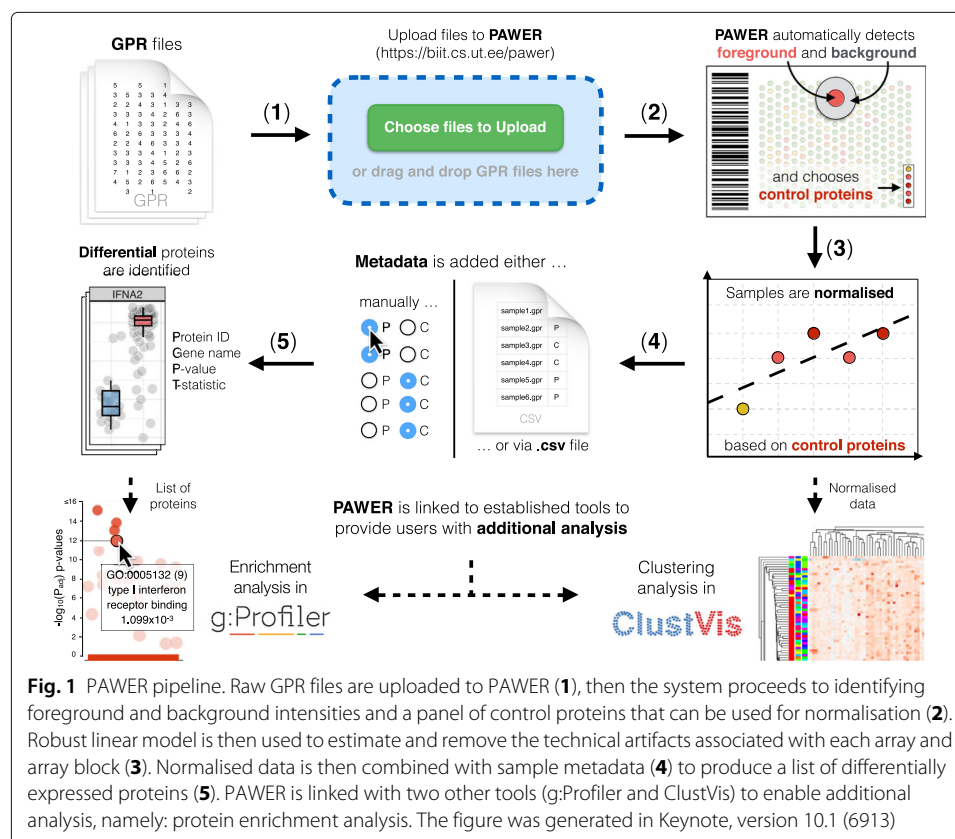
Tool name	License	Last updated	Platform	GUI	Normalisation	Biomarker identification	Functional annotation	Downloadable visuals
Prospector	No license specified	2015	Windows 7 desktop application	+	+	+	-	-
PAA	BSD 3	2019	R package	-	+	+	-	+
PMA	No license specified	2018	Java desktop application	+	+	-	-	-
PMD	No license specified	2020	Web server R code	+	+	+	+	+
PAWER	GNU GPL V2.	2020	Web server, R package	+	+	+	+	+

Presence or absence of relevant features (in columns) are shown as pluses highlighted in green (present features) or minuses in red (absent features). We were not able to obtain results using PMD tool, thus all the relevant entries are based on the claims made in the original publication [11] and highlighted in gray

challenges described above, we developed PAWER - a freely accessible web service as an alternative way to analyse protein microarrays. PAWER has a user-friendly interactive graphical interface that helps researchers to apply standard protein microarray analysis pipeline with ease (Fig. 1). Being comparable to PAA in its core strengths (protein specific normalisation and biomarker identification capabilities), PAWER also interprets the results of the analysis by providing detailed functional annotation of the identified differential proteins. Other key features are an interactive results table and accompanying attractive figures. Both figures and the table can be downloaded and used in the publications or scientific presentations. Notably, all the key steps of the analysis pipeline are well documented and presented on a separate help page.

Conclusions

PAWER is the state-of-the-art protein microarray analysis pipeline with clean and intuitive web interface. The result of the analysis is presented as a searchable and filterable table. Interactive figures related to the table allow to explore reactivities in a more detailed manner. Both the table and the figures can be downloaded in various file formats, including in publication-ready visuals. In order to encourage further development of protein microarray analysis methods, both the R and the web application code are made openly available. PAWER has already been used in multiple projects, with the underlying R codebase central for analysis in two recent studies of APECED syndrome [2, 13].

Fig. 1 PAWER pipeline. Raw GPR files are uploaded to PAWER (**1**), then the system proceeds to identifying foreground and background intensities and a panel of control proteins that can be used for normalisation (**2**). Robust linear model is then used to estimate and remove the technical artifacts associated with each array and array block (**3**). Normalised data is then combined with sample metadata (**4**) to produce a list of differentially expressed proteins (**5**). PAWER is linked with two other tools (g:Profiler and ClustVis) to enable additional analysis, namely: protein enrichment analysis. The figure was generated in Keynote, version 10.1 (6913)

Fishman *et al. BMC Bioinformatics*   (2020) 21:411

Page 7 of 8

To enable a closer interaction with our users and facilitate continuous improvement of PAWER, we have made available the PAWER feature roadmap, which can be accessed from the help page. It allows users to post feature requests and provide feedback.

Availability and requirements

Project name: PAWER

Project home page: https://biit.cs.ut.ee/pawer

Operating system(s): Platform independent

Programming language: JavaScript for web interface and R for package

Other requirements: R 3.4.2

License: Both R package and the web server code are available under GNU GPL v2 license

Any restrictions to use by non-academics: Not applicable

References
1. Fan Q, Huang LZ, Zhu XJ, Zhang KK, Ye HF, Luo Y, Sun XH, Zhou P, Lu Y. Identification of proteins that interact with alpha A-crystallin using a human proteome microarray. Mol Vis. 2014;20:117–24.
2. Meyer S, Woodward M, Hertel C, Vlaicu P, Haque Y, Macagno A, Onuoha SC, Fishman D, Peterson H, Metskula K, Uibo R, Jantti K, Hokynar K, Wolff ASB, Krohn K, Ranki A, Peterson P, Kisand K, Hayday A, Meloni A, Kluger N, Husebye ES, Podkrajsek KT, Battelino T, Bratanic N, Peet A. AIRE-Deficient Patients Harbor Unique High-Affinity Disease-Ameliorating Autoantibodies. Cell. 2016;166(3):582–95. https://doi.org/10.1016/j.cell.2016.06.024.
3. Sharon D, Snyder M. Serum profiling using protein microarrays to identify disease related antigens. In: Methods in Molecular Biology, vol 1176. New York: Springer; 2014. p. 169–78.
4. Jeong JS, Jiang L, Albino E, Marrero J, Rho HS, Hu J, Hu S, Vera C, Bayron-Poueymiroy D, Rivera-Pacheco ZA, Ramos L, Torres-Castro C, Qian J, Bonaventura J, Boeke JD, Yap WY, Pino I, Eichinger DJ, Zhu H, Blackshaw S. Rapid identification of monospecific monoclonal antibodies using a human proteome microarray. Mol Cell Proteomics. 2012;11(6):111–016253. https://doi.org/10.1074/mcp.O111.016253.

Fishman *et al. BMC Bioinformatics*   (2020) 21:411

Page 8 of 8

5.   Duarte JDG, Goosen RW, Lawry PJ, Blackburn JM. PMA: Protein microarray analyser, a user-friendly tool for data processing and normalization. BMC Res Notes. 2018;11(1):. https://doi.org/10.1186/s13104-018-3266-0.
6.   Yu X, Petritis B, Duan H, Xu D, LaBaer J. Advances in cell-free protein array methods. Expert Rev Proteomics. 2017;15(1):1–11. https://doi.org/10.1080/14789450.2018.1415146.
7.   Abel L, Kutschki S, Turewicz M, Eisenacher M, Stoutjesdijk J, Meyer HE, Woitalla D, May C. Autoimmune profiling with protein microarrays in clinical applications. Biochim Biophys Acta (BBA) Protein Proteomics. 2014;1844(5): 977–98. https://doi.org/10.1016/j.bbapap.2014.02.023.
8.   Sboner A, Karpikov A, Chen G, Smith M, Mattoon D, Dawn M, Freeman-Cook L, Schweitzer B, Gerstein MB. Robust-linear-model normalization to reduce technical variability in functional protein microarrays. J Proteome Res. 2009;8(12):5451–64. https://doi.org/10.1021/pr900412k.
9.   Duarte JG, Blackburn JM. Advances in the development of human protein microarrays. Expert Rev Proteomics. 2017;14(7):627–41. https://doi.org/10.1080/14789450.2017.1347042.
10.   Turewicz M, Ahrens M, May C, Marcus K, Eisenacher M. PAA: an R/bioconductor package for biomarker discovery with protein microarrays. Bioinformatics. 2016;32(10):1577–9. https://doi.org/10.1093/bioinformatics/btw037.
11.   Xu Z, Huang L, Zhang H, Li Y, Guo S, Wang N, Wang S-H, Chen Z, Wang J, Tao S-C. PMD: A resource for archiving and analyzing protein microarray data. Sci Rep. 2016;6(1):. https://doi.org/10.1038/srep19956.
12.   Turewicz M, May C, Ahrens M, Woitalla D, Gold R, Casjens S, Pesch B, Brüning T, Meyer HE, Nordhoff E, Böckmann M, Stephan C, Eisenacher M. Improving the default data analysis workflow for large autoimmune biomarker discovery studies with protoarrays. Proteomics. 2013;13(14):2083–7. https://doi.org/10.1002/pmic.201200518.
13.   Fishman D, Kisand K, Hertel C, Rothe M, Remm A, Pihlap M, Adler P, Vilo J, Peet A, Meloni A, Podkrajsek KT, Battelino T, Bruserud, Wolff ASB, Husebye ES, Kluger N, Krohn K, Ranki A, Peterson H, Hayday A, Peterson P. Autoantibody Repertoire in APECED Patients Targets Two Distinct Subgroups of Proteins. Front Immunol. 2017;8: 976. https://doi.org/10.3389/fimmu.2017.00976.
14.   Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004;3:3. https://doi.org/10.2202/1544-6115.1027.
15.   Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, Vilo J. g:profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. 2019;47(W1):191–8. https://doi.org/10.1093/nar/gkz369.
16.   Metsalu T, Vilo J. ClustVis: a web tool for visualizing clustering of multivariate data using principal component analysis and heatmap. Nucleic Acids Res. 2015;43(W1):566–70. https://doi.org/10.1093/nar/gkv468.
17.   Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. the gene ontology consortium. Nat Genet. 2000;25(1):25–9. https://doi.org/10.1038/75556.
18.   Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2016;45(D1):353–61. https://doi.org/10.1093/nar/gkw1092.
19.   Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, Haw R, Jassal B, Korninger F, May B, Milacic M, Roca CD, Rothfels K, Sevilla C, Shamovsky V, Shorser S, Varusai T, Viteri G, Weiser J, Wu G, Stein L, Hermjakob H, D'Eustachio P. The reactome pathway knowledgebase. Nucleic Acids Res. 2018;46(D1):649–55. https://doi.org/10.1093/nar/gkx1132.
20.   Köhler S, Carmody L, Vasilevsky N, Jacobsen JOB, Danis D, Gourdine J-P, Gargano M, Harris NL, Matentzoglu N, McMurry JA, Osumi-Sutherland D, Cipriani V, Balhoff JP, Conlin T, Blau H, Baynam G, Palmer R, Gratian D, Dawkins H, Segal M, Jansen AC, Muaz A, Chang WH, Bergerson J, Laulederkind SJF, Yüksel Z, Beltran S, Freeman AF, Sergouniotis PI, Durkin D, Storm AL, Hanauer M, Brudno M, Bello SM, Sincan M, Rageth K, Wheeler MT, Oegema R, Lourghi H, Rocca MGD, Thompson R, Castellanos F, Priest J, Cunningham-Rundles C, Hegde A, Lovering RC, Hajek C, Olry A, Notarangelo L, Similuk M, Zhang XA, Gómez-Andrés D, Lochmüller H, Dollfus H, Rosenzweig S, Marwaha S, Rath A, Sullivan K, Smith C, Milner JD, Leroux D, Boerkoel CF, Klion A, Carter MC, Groza T, Smedley D, Haendel MA, Mungall C, Robinson PN. Expansion of the human phenotype ontology (HPO) knowledge base and resources. Nucleic Acids Res. 2019;47(D1):1018–27. https://doi.org/10.1093/nar/gky1105.
21.   Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A, et al. Tissue-based map of the human proteome. Science. 2015;347(6220):1260419. https://doi.org/10.1126/science.1260419.
22.   Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):47. https://doi.org/10.1093/nar/gkv007.
23.   Venables WN, Ripley BD. Modern Applied Statistics with S-PLUS. Berlin, Germany: Springer; 2013.
24.   Wickham H, et al. Reshaping data with the reshape package. J Stat Softw. 2007;21(12):1–20. https://doi.org/10.18637/jss.v021.i12.
25.   Bostock M, Ogievetsky V, Heer J. D³: Data-Driven Documents. IEEE Trans Vis Comput Graph. 2011;17(12):2301–9. https://doi.org/10.1109/TVCG.2011.185.

Publisher's Note