PAWER: Protein Array Web ExploreR

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

1 Institute of Computer Science, University of Tartu, J. Liivi 2, 50409, Tartu, Estonia
2 Quretec Ltd, Ülikooli 6a, 51003, Tartu, Estonia

*** To whom correspondence should be addressed: hedi.peterson@ut.ee**

Abstract

Protein microarray is a well-established approach for characterising activity levels of thousands of proteins in a parallel and high-throughput manner. Analysis of protein microarray data is complex and time consuming, and existing solutions are either outdated, or challenging to use without programming skills. Typical pipeline consists of a normalisation step, followed by differential reactivity analysis. Then the differentially reactive proteins are visualised and put into context via functional enrichment analysis. Usually, biologists would need to assemble their own workflow by combining an independent set of tools to perform all these tasks. We present PAWER, a freely available web tool, that enables biologists to carry out all the analysis steps in one go. PAWER uses state-of-the-art computational methods, and generates publication-ready illustrations. We also provide an R package for more advanced use cases, such as bespoke analysis workflows. PAWER is freely available at `https://biit.cs.ut.ee/pawer`.

1 Introduction

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. Intense fluorescent signal 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, three major tools for protein microarray analysis are Prospector, Protein Array Analyser (PAA) [10] and Protein Microarray Analyser (PMA) [5]. 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 [11]. 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. Moreover, neither of these protein microarray analysis tools allow the user to complete the full analysis workflow and thus demand external tools or packages, to be used downstream for visualisation and putting the results into context of previous knowledge.

Here, we present Protein Array Web Explorer (PAWER), the first web tool dedicated to analysing protein microarray data. PAWER incorporates the strengths of Prospector and PAA, while eliminating their major limitations. PAWER is suitable for experimental biologists who want to analyse their own data without the need to write code.

2 MATERIALS AND METHODS

2.1 Key features

PAWER implements the following key features:

1. Public web service that can be used by anyone with protein microarray data
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 [12])
5. Downloadable intermediate results after each analysis step
6. Connection with g:Profiler [13] tool through its R package (gprofiler2) for fast enrichment analysis of differential protein features
7. An open source R package that the PAWER web server is built upon

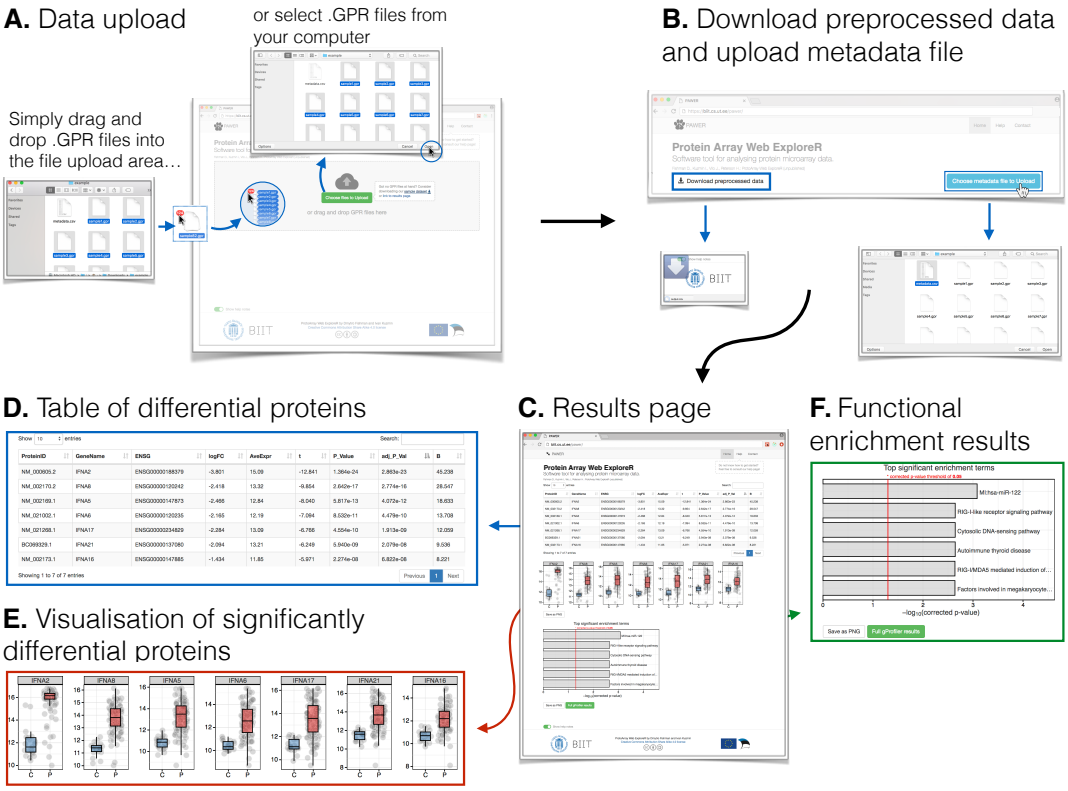

Figure 1. PAWER pipeline. Raw data upload page (A), intermediate results page (B) from which normalised data can be downloaded as a data matrix and metadata uploaded for further analysis. The results page (C). It includes the table with significantly differential proteins (D), boxplots with their expression values (E) and the top six significantly enriched annotation terms (F).

2.2 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 from the computer (Figure 1A). Once samples have been uploaded, a single data matrix is assembled using the limma R package [12] by first subtracting background intensities and then averaging the duplicate spots. Resulting signal values are log-transformed prior to applying robust linear model [8]. RLM is used to normalise the obtained matrix, removing differences between sample that are likely due to technical noise. Data upload and normalisation steps take a few minutes on typical datasets of a few tens of samples. After the normalisation is completed, user can either download the normalised data or proceed to differential reactivity analysis. For the latter user needs to provide a metadata file that specifies phenotypic groups (Figure 1B). The file must contain two columns, filenames in the first column and phenotypic groups in the second, separated by a comma.

2.3 PAWER output

Differential protein features are identified using a moderated t-test, implemented in the limma R package. 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 (Figure 1D). 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. Expression values of the proteins shown in the table are visualised with boxplots (Figure 1E). Each protein is visualised in its own facet, with log2 transformed expression values denoted by the y-axis and studied conditions by the x-axis. Circles that overlay boxes represent individual samples. The corresponding sample name appears on hover. Coloured box covers 50% of the values, providing an insight into the underlying distribution of the protein binding values. Boxplots can be downloaded in a form of a publication-ready figure. Additionally, enrichment of differential intensities is enabled by the gprofiler2 R package that provides interface for g:Profiler service [13]. g:Profiler gives functional enrichment results from a number of different categories, such as Gene Ontology [14], pathways and other structured data sources for instance KEGG [15], Reactome [16], Human Phenotype Ontology [17] and Human Protein Atlas [18]. The six most significant terms are visualised as a downloadable bar plot figure (Figure 1F). The full results can be accessed through a generated link.

2.4 Implementation

We developed the PAWER web service as a tool that covers all the necessary steps in protein microarray analysis. PAWER has been primarily designed and tested on data from ProtoArray and HuProt platforms. Its core has been implemented using R version 3.4.2, using limma [19] (v. 3.34.4) for reading in the GPR format files and performing differential analysis, MASS [20] (v. 7.3.47), reshape2 [21] (v. 1.4.2) for normalisation and preprocessing of protoarray data and gprofiler2 [13] (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 [22] and DataTables libraries. Both the R package and the web server code are freely available under the MIT and GNU GPL v2. licenses, respectively.

2.5 Comparison to existing tools

To the best of our knowledge there are four available tools, dedicated to protein microarray analysis – Prospector, PAA, PMA 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

Tool name	License	Last updated	Platform	GUI	Protein array specific normalisation	Potential Biomarker Identification	Detailed Functional Annotation	Downloadable visuals
Prospector	No license specified	2015	Windows 7 Desktop application	+	+	+	—	—
PAA	BSD 3	2018	R package	—	+	+	—	+
PMA	No license specified	2018	Java desktop application	+	+	—	—	—
PAWER	GNU GPL v2.	2019	Web server, R package	+	+	+	+	+

Table 1. Comparison between currently available protein microarray analysis tools: Prospector, PAA, PMA and PAWER.

the key features is highlighted in Table 1. Prospector was the first protein microarray analysis tool, 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, could design and apply custom analysis pipelines for their protein microarrays. But PAA requires users to be familiar with R programming language. On the other hand, PMA - a Java desktop application, provides a graphical user interface and implements cutting edge preprocessing techniques. However it lacks documentation and does not allow for the integrated downstream analysis [5]. We developed PAWER as a freely accessible web server as an alternative way to analyse protein microarrays. PAWER has a user friendly interactive graphical interface that helps researchers to go through the standard analysis pipeline (Figure 1 A-F). While still having the same core strengths as PAA (protein specific normalisation and biomarker identification capabilities), PAWER uniquely helps to interpret 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 are well documented in a separate help page.

2.6 Use case

In order to illustrate the functionality of PAWER, we used publicly available HuProt data from the study by Gupta and colleagues [23]. In their original study they analysed meningioma and used individual arrays from 15 patients and 15 healthy individuals. Their results highlight IGHG4, CRYM, EFCAB2, STAT6, HDAC7A and CCNB1 as significantly dysregulated proteins. We downloaded the GPR files by accession number GSE83507 from GEO repository [24]. As PAWER also requires metadata file and no such file was present in the repository, we generated it based on the metadata information provided in supplementary materials. We uploaded the GPR files along with the metadata file to PAWER and used the default settings. While PAWER identified many more abundant proteins (2387 versus 203 identified by the original paper) the overlap of 101 proteins between lists is highly significant (p-value < 0.00001) evaluated by the permutation test. Despite the large size difference of the result lists, 5 out of 6 proteins highlighted by the authors also appear to be significantly different between the conditions when using PAWER analysis. Moreover, the top two most significant proteins, IGHG4 and CRYM, are also the top two abundant proteins in our analysis. One potential reason behind the large difference in sizes is that original study has applied different pre-processing strategy. Namely, they mentioned quantile normalisation, which is more common in DNA microarray experiments. PAWER on the other hand utilises an RLM method that has been shown to be better

suited for protein array analysis [8]. Despite the large size difference in obtained protein lists, this use case shows PAWER's capacity to generate meaningful and reproducible results by applying the standard pipeline.

3 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 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 images. In order to encourage further development of protein microarray analysis methods, both the R and the web application code are made openly available. PAWER is already used for multiple projects, with the underlying R codebase central for analysis in two recent studies of APECED syndrome [2, 25].

3.1 Future directions

The mission of PAWER is to provide researchers with a clear and usable tool for the analysis of protein microarrays. Despite the overall well-chosen default parameters offered by PAWER, we feel that there is a room for customisation at every step of the analysis pipeline. For instance, a possibility to specify proteins that will be used as a reference for normalisation. Also adding more tools for exploratory analysis (e.g. PCA plots and heatmaps) may help users to gain important insights about their data. 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 tool home page. It allows users to post feature requests and provide feedback.

4 DATA AVAILABILITY

PAWER web server and accompanied R package are both freely available online at `https://biit.cs.ut.ee/pawer/`. R package repository at `https://gl.cs.ut.ee/biit/paweR`. Web interface repository at `https://gl.cs.ut.ee/biit/pawer_web_client`.

5 FUNDING

This work was supported by the Estonian Research Council grants [PSG59, IUT34-4]; European Regional Development Fund for CoE of Estonian ICT research EXCITE projects; European Union through the Structural Fund [Project No 2014-2020.4.01.16-0271, ELIXIR].

5.1 Conflict of interest statement.

None declared.

6 ACKNOWLEDGEMENTS

The authors would like to acknowledge Pärt Peterson and Kai Kisand for introducing us to protein microarrays and their expert insight into the autoimmunity field. Also, we would like to thank Leopold Parts and Liis Kolberg for critical reading and comments on the manuscript. Special thanks go to miss Boson for providing ideas that laid down the basis for the name and logo of PAWER.

References

1. Fan, Q., Huang, L. Z., Zhu, X. J., Zhang, K. K., Ye, H. F., Luo, Y., Sun, X. H., Zhou, P., and Lu, Y. (2014) Identification of proteins that interact with alpha A-crystallin using a human proteome microarray. *Mol. Vis.,* **20**, 117–124 [PubMed:24453475] [PubMedCentral: PMC3893783].

2. Meyer, S., Woodward, M., Hertel, C., Vlaicu, P., Haque, Y., Karner, J., Macagno, A., Onuoha, S. C., Fishman, D., Peterson, H., Metskula, K., Uibo, R., Jantti, K., Hokynar, K., Wolff, A. S. B., Krohn, K., Ranki, A., Peterson, P., Kisand, K., Hayday, A., Meloni, A., Kluger, N., Husebye, E. S., Podkrajsek, K. T., Battelino, T., Bratanic, N., and Peet, A. (Jul, 2016) AIRE-Deficient Patients Harbor Unique High-Affinity Disease-Ameliorating Autoantibodies. *Cell,* **166**(3), 582–595 [PubMed:27426947] [doi:] [PubMedCentral: PMC4967814].

3. Sharon, D. and Snyder, M. (2014) Serum Profiling Using Protein Microarrays to Identify Disease Related Antigens. In *Methods in Molecular Biology* pp. 169–178 Springer New York.

4. Jeong, J. S., Jiang, L., Albino, E., Marrero, J., Rho, H. S., Hu, J., Hu, S., Vera, C., Bayron-Poueymiroy, D., Rivera-Pacheco, Z. A., Ramos, L., Torres-Castro, C., Qian, J., Bonaventura, J., Boeke, J. D., Yap, W. Y., Pino, I., Eichinger, D. J., Zhu, H., and Blackshaw, S. (Jun, 2012) Rapid identification of monospecific monoclonal antibodies using a human proteome microarray. *Mol. Cell Proteomics,* **11**(6), O111.016253 [PubMed:22307071] [doi:] [PubMedCentral: PMC3433917].

5. Duarte, J. D. G., Goosen, R. W., Lawry, P. J., and Blackburn, J. M. (February, 2018) PMA: Protein Microarray Analyser, a user-friendly tool for data processing and normalization. *BMC Research Notes,* **11**(1).

6. Yu, X., Petritis, B., Duan, H., Xu, D., and LaBaer, J. (December, 2017) Advances in cell-free protein array methods. *Expert Review of Proteomics,* **15**(1), 1–11.

7. Abel, L., Kutschki, S., Turewicz, M., Eisenacher, M., Stoutjesdijk, J., Meyer, H. E., Woitalla, D., and May, C. (May, 2014) Autoimmune profiling with protein microarrays in clinical applications. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics,* **1844**(5), 977–987.

8. Sboner, A., Karpikov, A., Chen, G., Smith, M., Mattoon, D., Dawn, M., Freeman-Cook, L., Schweitzer, B., and Gerstein, M. B. (Dec, 2009) Robust-linear-model normalization to reduce technical variability in functional protein microarrays. *J. Proteome Res.,* **8**(12), 5451–5464 [PubMed:19817483] [doi:].

9. Duarte, J. G. and Blackburn, J. M. (07, 2017) Advances in the development of human protein microarrays. *Expert Rev Proteomics,* **14**(7), 627–641 [PubMed:28644690] [doi:].

10. Turewicz, M., Ahrens, M., May, C., Marcus, K., and Eisenacher, M. (05, 2016) PAA: an R/bioconductor package for biomarker discovery with protein microarrays. *Bioinformatics,* **32**(10), 1577–1579 [PubMed:26803161] [doi:] [PubMedCentral: PMC4866526].

11. Turewicz, M., May, C., Ahrens, M., Woitalla, D., Gold, R., Casjens, S., Pesch, B., Brüning, T., Meyer, H. E., Nordhoff, E., Böckmann, M., Stephan, C., and Eisenacher, M. (Jul, 2013) Improving the default data analysis workflow for large autoimmune biomarker discovery studies with ProtoArrays. *Proteomics,* **13**(14), 2083–2087 23616427[pmid].

12. Smyth, G. K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol,* **3**, Article3 [PubMed:16646809] [doi:].

13. Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., and Vilo, J. (May, 2019) g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research,* **47**(W1), W191–W198.

14. 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., et al. (May, 2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics,* **25**(1), 25–29 [PubMed:10802651] [PubMed Central:PMC3037419] [doi:10.1038/75556].

15. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (November, 2016) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research,* **45**(D1), D353–D361.

16. Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., Milacic, M., Roca, C. D., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Viteri, G., Weiser, J., Wu, G., Stein, L., Hermjakob, H., and D'Eustachio, P. (November, 2017) The Reactome Pathway Knowledgebase. *Nucleic Acids Research,* **46**(D1), D649–D655.

17. Köhler, S., Carmody, L., Vasilevsky, N., Jacobsen, J. O. B., Danis, D., Gourdine, J.-P., Gargano, M., Harris, N. L., Matentzoglu, N., McMurry, J. A., Osumi-Sutherland, D., Cipriani, V., Balhoff, J. P., Conlin, T., Blau, H., Baynam, G., Palmer, R., Gratian, D., Dawkins, H., Segal, M., Jansen, A. C., Muaz, A., Chang, W. H., Bergerson, J., Lauledderkind, S. J. F., Yüksel, Z., Beltran, S., Freeman, A. F., Sergouniotis, P. I., Durkin, D., Storm, A. L., Hanauer, M., Brudno, M., Bello, S. M., Sincan, M., Rageth, K., Wheeler, M. T., Oegema, R., Lourghi, H., Rocca, M. G. D., Thompson, R., Castellanos, F., Priest, J., Cunningham-Rundles, C., Hegde, A., Lovering, R. C., Hajek, C., Olry, A., Notarangelo, L., Similuk, M., Zhang, X. A., Gómez-Andrés, D., Lochmüller, H., Dollfus, H., Rosenzweig, S., Marwaha, S., Rath, A., Sullivan, K., Smith, C., Milner, J. D., Leroux, D., Boerkoel, C. F., Klion, A., Carter, M. C., Groza, T., Smedley, D., Haendel, M. A., Mungall, C., and Robinson, P. N. (November, 2018) Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Research,* **47**(D1), D1018–D1027.

18. Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., et al. (2015) Tissue-based map of the human proteome. *Science,* **347**(6220), 1260419 [PubMed:25613900] [doi:10.1126/science.1260419].

19. Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (Apr, 2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.,* **43**(7), e47 [PubMed:25605792] [doi:] [PubMedCentral: PMC4402510].

20. Venables, W. N. and Ripley, B. D. (2013) Modern applied statistics with S-PLUS, Springer Science & Business Media, .

21. Wickham, H. et al. (2007) Reshaping data with the reshape package. *Journal of Statistical Software,* **21**(12), 1–20 [doi:].

22. Bostock, M., Ogievetsky, V., and Heer, J. (Dec, 2011) D: Data-Driven Documents. *IEEE Trans Vis Comput Graph,* **17**(12), 2301–2309 [PubMed:22034350] [doi:].

23. Gupta, S., Mukherjee, S., Syed, P., Pandala, N. G., Choudhary, S., Singh, V. A., Singh, N., Zhu, H., Epari, S., Noronha, S. B., Moiyadi, A., and Srivastava, S. (Aug, 2017) Evaluation of autoantibody signatures in meningioma patients using human proteome arrays. *Oncotarget,* **8**(35), 58443–58456 [PubMed:28938569] [doi:] [PubMedCentral: PMC5601665].

24. Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., and Soboleva, A. (Jan, 2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.,* **41**(Database issue), D991–995 [PubMed:23193258] [doi:] [PubMedCentral: PMC3531084].

25. Fishman, D., Kisand, K., Hertel, C., Rothe, M., Remm, A., Pihlap, M., Adler, P., Vilo, J., Peet, A., Meloni, A., Podkrajsek, K. T., Battelino, T., Bruserud, ., Wolff, A. S. B., Husebye, E. S., Kluger, N., Krohn, K., Ranki, A., Peterson, H., Hayday, A., and Peterson, P. (2017) Autoantibody Repertoire in APECED Patients Targets Two Distinct Subgroups of Proteins. *Front Immunol,* **8**, 976 [PubMed:28861084] [doi:] [PubMedCentral: PMC5561390].