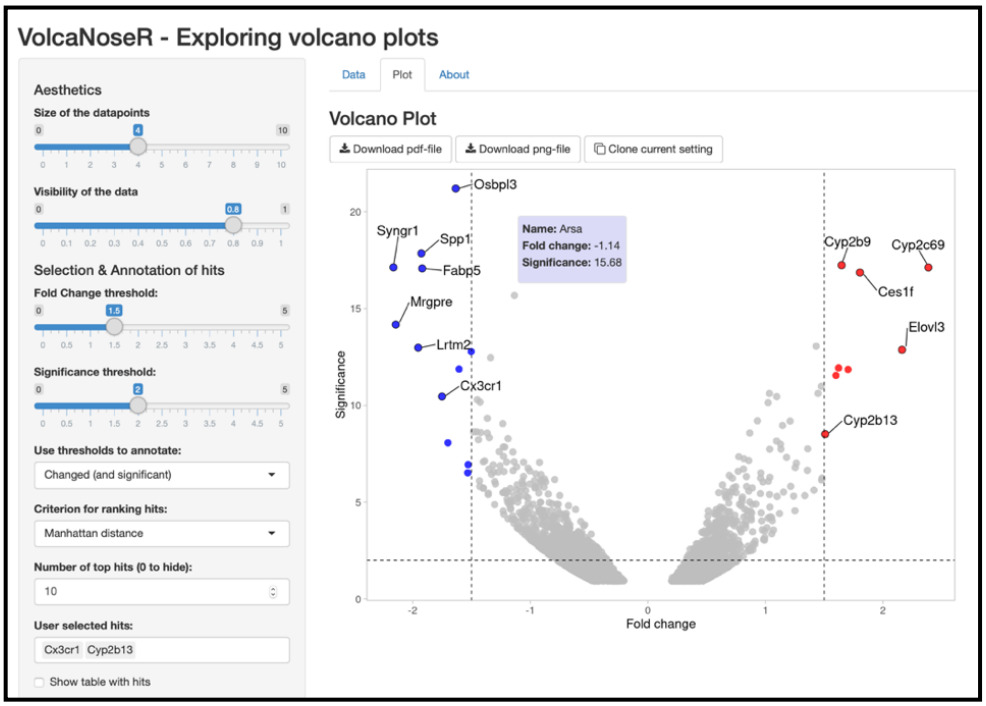
VolcaNoseR – a web app for creating, exploring and sharing volcano plots

Joachim Goedhart[1, ‡] and Martijn S. Luijsterburg[2]

[1]Swammerdam Institute for Life Sciences, Section of Molecular Cytology, van Leeuwenhoek Centre for Advanced Microscopy, University of Amsterdam, P.O. Box 94215, NL-1090 GE Amsterdam, The Netherlands.
Email: j.goedhart@uva.nl | Twitter: @joachimgoedhart | ORCID: 0000-0002-0630-3825
[2]Department of Human Genetics, Leiden University Medical Center, Einthovenweg 20, 2333 ZC, Leiden, The Netherlands.
Email: M.Luijsterburg@lumc.nl | Twitter: @luijsterburglab | ORCID: 0000-0001-5796-6541



Abstract

Comparative genome- and proteome-wide screens yield large amounts of data. To efficiently present such datasets and to simplify the identification of hits, the results are often presented in a type of scatterplot known as a volcano plot, which shows a measure of effect size versus a measure of significance. The data points with the largest effect size and a statistical significance beyond a user defined threshold are considered as hits. Such hits are usually annotated in the plot by a label with their name. Volcano plots can represent ten thousands of data points, of which typically only a handful is annotated. The information of data that is not annotated is hardly or not accessible. To simplify access to the data and enable its re-use, we have developed an open source and online web tool with R/shiny. The web app is named VolcaNoseR and it can be used to create, explore, annotate and share volcano plots (https://huygens.science.uva.nl/VolcaNoseR). When the data is stored in an online data repository, the web app can retrieve that data together with user-defined settings to generate a customized, interactive volcano plot. Users can interact with the data, adjust the plot and share their modified plot together with the underlying data. Therefore, VolcaNoseR increases the transparency and re-use of large comparative genome- and proteome-wide datasets.

Introduction

The volcano plot visualizes complex datasets generated by genomic screening or proteomic approaches. It is essentially a scatter plot, in which the coordinates of datapoints are defined by effect size and statistical significance (Li, 2012; Cui and Churchill, 2003).

Volcano plots typically show the data of hundreds to ten thousands of genes or proteins. However, only a handful of data points are usually highlighted by a label to indicate specific genes or proteins of interest that are considered hits. This enables the visual identification of such obvious hits and simplifies the interpretation of the complex dataset. Nevertheless, the data points that are not annotated may be of equal interest. Therefore, it is highly desirable to have easy access to the information of all data points from such large datasets.

Volcano plots are usually generated using commercial software or with software that requires coding skills. A viable alternative is provided by dedicated free web apps that allow users to generate plots through a graphical user interface (GUI). Several web apps are available (Singh et al., 2016; Naumov et al., 2017), but these do not generate interactive plots and only have limited options for customization and annotation. Therefore, we decided to generate a web-based online tool for generating volcano plots, similar to previously plotting apps (Goedhart, 2020; Postma and Goedhart, 2019). Here, we report an open source web app for generating, exploring, labeling and sharing Volcano plots. The web app is created with R/Shiny and is dubbed VolcaNoseR. Below we discuss the features of the app.

Availability, code & issue reporting

The VolcaNoseR webtool is available at: https://huygens.science.uva.nl/VolcaNoseR or at (as long as the bandwidth limit is not reached): http://goedhart.shinyapps.io/VolcaNoseR/

The code was written in R using R (https://www.r-project.org) and Rstudio (https://www.rstudio.com). To run the app several freely available packages are required: shiny, ggplot2, magrittr, dplyr, ggrepel, shinycssloaders, DT and RCurl. The code of version 1.0.1 reported in this manuscript is archived at Zenodo.org: https://doi.org/10.5281/zenodo.3758543

Up-to-date code and new release will be made available on Github, together with information on running the app locally: https://github.com/JoachimGoedhart/VolcaNoseR

The Github page of VolcaNoseR is the preferred way to communicate issues and request features (https://github.com/JoachimGoedhart/VolcaNoseR/issues). Alternatively, the users can contact the developers by email or Twitter. Contact information is found on the “About” page of the app.

Data input and format

The data can be supplied via upload of a text file with different delimiters, including the Comma Separate Values (CSV) format. Alternatively, a CSV file from an online data repository can be used through a URL. To demonstrate the features of the app, example data is included of which the details can be found elsewhere (Becares et al., 2019; Gillingham et al., 2019).

After data upload, the user selects the columns that hold the information on the fold change (for the x-coordinate) and the significance (for the y-coordinate). Selecting a column with gene or protein names is optional.

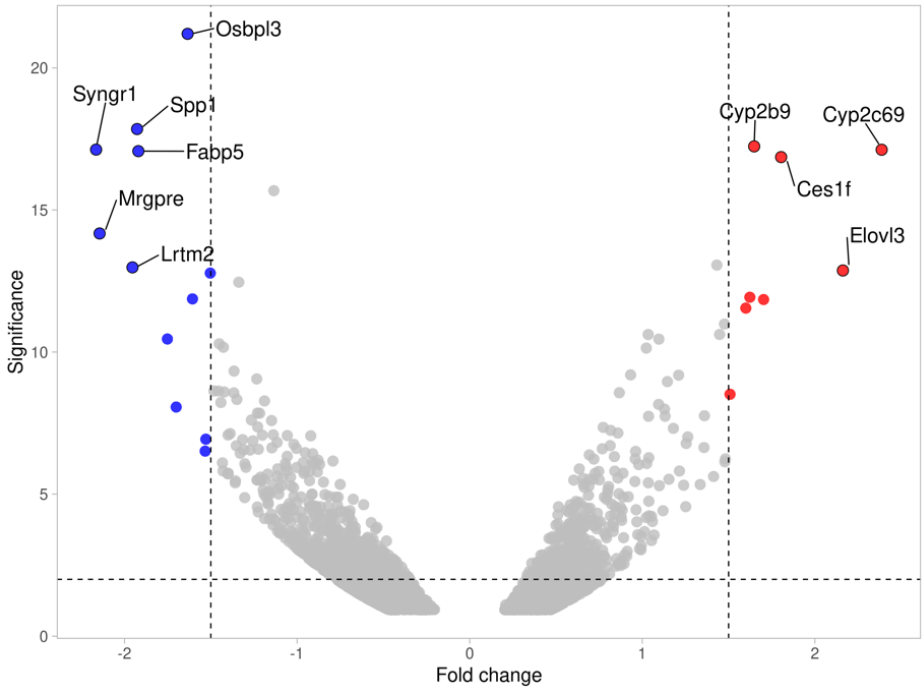


Figure 1: The standard output of the VolcaNoseR app for example dataset 1. The annotated dots are the ten datapoints that have the largest (Manhattan) distance from the origin and are above the thresholds indicated by the dashed line. Direct access to an interactive plot and all the data is provided by [this hyperlink](#).

Data visualization

A typical volcano plot shows the $\log_2$ of the fold change on the x-axis and minus $\log_{10}$ of the p-value on the y-axis. The data is shown as dots and their size and transparency can be adjusted. The position of the individual points is defined by these coordinates. By hovering over the data points the information about the data can be accessed immediately and dynamically. When the pointer (mouse) is near a data point, the x- and y-coordinate and the name is retrieved, providing the user with information on the spot.

In some cases, data from screens is shown in a 90 degrees rotated plot. This option is available and this will depict the fold change on the y-axis and the significance on the x-axis.

Thresholds and hits

The user can set threshold values for the fold change and the significance. The threshold values are indicated by dashed lines in the plot and used to classify the data as 'unchanged', 'decreased' or 'increased'. The data are colored according to this classification and this can be shown in a legend. The 'top hits' can be automatically detected and ranked based on a number of criteria. The default criterion is the Manhattan distance ($|\Delta X|+|\Delta Y|$) of the data from the origin (0,0). The other criteria are Euclidean distance ($SQRT(\Delta X^2+\Delta Y^2)$), absolute fold change or significance. The data are sorted based on the selected criterion and the 10 top-ranking data points are selected. The number of top ranking data points can be adjusted by the user.

Also, it is possible to annotate only 'increased' or 'decreased' or all significantly changed ('increased' and 'decreased') data points. The top-ranking hits are shown in the plot and there is an option to list them in a table. Finally, the user can manually search and select the names of genes or proteins of interest, which will be annotated in the plot and added to the table.
The standard colors to indicate 'unchanged', 'increased' and 'decreased' are respectively grey, red and blue. Another color combination that is available is grey, blue and green. Users can also define their own color scheme.

Output

Users can customize the titles for the axes and their size. The plot that is generated by the app can be directly retrieved by drag-and-drop from the web browser. In addition, the plot can be downloaded as a PNG or PDF file. The PNG is a lossless bitmap format. The PDF allows for downstream processing/editing with software that can handle vector-based graphics.

Sharing data and plot settings

All settings that are defined in the GUI can be stored as a URL, as was previously implemented for PlotsOfData and PlotTwist (Postma and Goedhart, 2019; Goedhart, 2020). When the data is retrieved from an external online resource, this hyperlink is included in the URL. The URL with settings is sufficient to (i) launch the app, (ii) retrieve the data, and (iii) plot the data according to user-defined settings. Once the plot is available, it can be adjusted and a new URL reflecting the new settings can be obtained. This feature enables transparent reporting of all the data and simplifies re-use of the data.

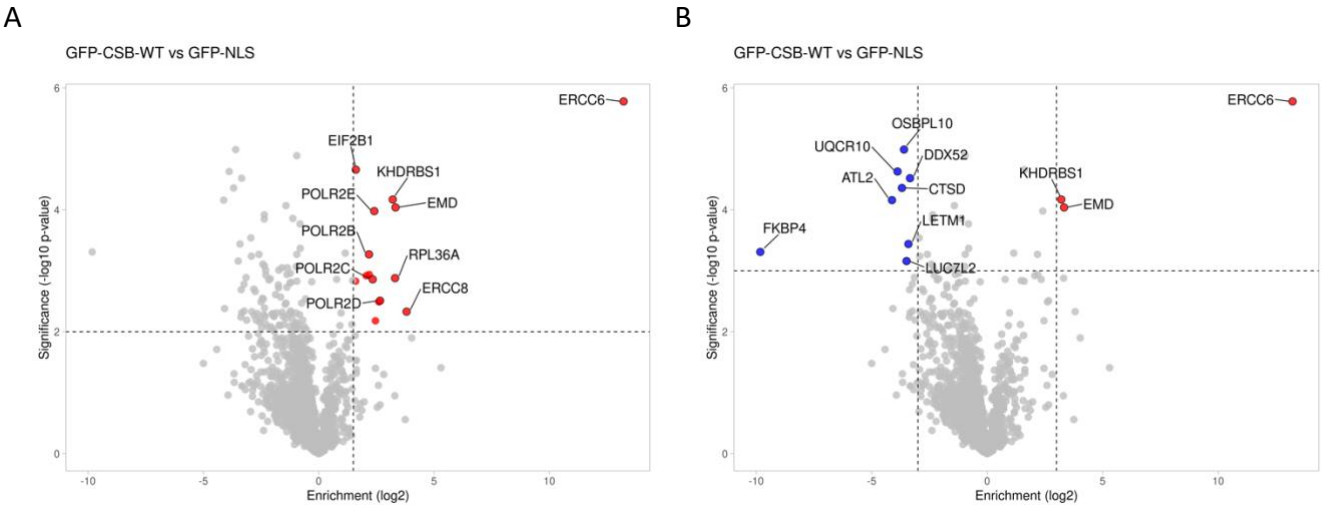


Figure 2: (A) Volcano plot generated from a published data (van der Weegen et al., 2020). The interactive plot and all of the data is accessible through this hyperlink. (B) Volcano plot from the same data as (A) but replotted to show both significantly increased and decreased hits with the Enrichment and Significance threshold set to 3 and with a different color scale. The plot is accessible through this hyperlink.

We illustrate this feature with data from proteomic screens that we have recently published (van der Weegen et al., 2020). These data are deposited and publicly available at the data repository zenodo.org, doi: 10.5281/zenodo.3713174. Volcano plots are generated with VolcaNoseR using the data from the CSV files in the repository. Next, the URL that encodes all necessary information was generated using the 'clone current setting' button. With this unique URL, the data is retrieved and a plot is generated by VolcaNoseR based on the parameters that are stored in the URL. For instance, this URL produces an interactive plot of which a static version is shown in figure 2A:

https://huygens.science.uva.nl:/VolcaNoseR/?data=5;;Difference_CSB_GFP;p_value_CSB_GFP;Gene_names&vis=4;0.8;1.5;2;increased&can=10;;;&layout=;;;;;log2;minus_log10;X;600;800&label=TRUE;GFP-CSB-WT vs GFP-NLS;TRUE;Enrichment (log2);Significance (-log10 p-value);;24;24;18;6;&url=https://zenodo.org/record/3713174/files/CSV_1-GFP-CSB-WT_vs_GFP-NLS.csv

Users can easily access the data and plot through this URL, inspect the data and replot it. Suppose that a user is interested in both showing and annotating increased and decreased proteins with more stringent threshold levels, the user can replot the data as shown in figure 2B. The URL can be copied and shared. This URL would be:

https://huygens.science.uva.nl/VolcaNoseR/?data=5;;Difference_CSB_GFP;p_value_CSB_GFP;Gene_names&vis=4;0.8;-3,3;3;significant;manh&can=20;;;&layout=;;;;;X;600;800&color=3;none&label=TRUE;GFP-CSB-WT%20vs%20GFP-NLS;TRUE;Enrichment%20(log2);Significance%20(-log10%20p-value);;24;24;18;6;&url=https://zenodo.org/record/3713174/files/CSV_1-GFP-CSB-WT_vs_GFP-NLS.csv

A list of settings that can be stored in the URL is available in a supplemental document (S1 text).

Conclusion

Volcano plots are data visualizations that are based on large datasets. Unfortunately, only a fraction of the data is annotated in static figures and, therefore, the vast majority of the information is inaccessible. To provide access to all of the data represented in a Volcano plot, we developed an interactive online plotting tool. By hovering over the plot with a pointer, each data point can be inspected. In addition, user-defined genes or proteins can labeled in the plot and listed in a table. Together, these features enable access to all the information that the plot is based on. Finally, the web app can be used to share the data and the plot to allow other users to interact with the data and reuse it. Therefore, VolcaNoseR increases the transparency and re-use of large comparative genome- and proteome-wide datasets.

Acknowledgments

Some of the VolcaNoseR code is taken from PlotTwist and it is partially inspired by the VolcanoR app (https://github.com/vovalive/volcanoR). We are grateful to Graham Dellaire (Dalhousie University, Canada) and Inés Pineda-Torra (University College of London, UK) for their input and thank Auke Folkerts (UvA, The Netherlands) for help with the server that runs shiny. The feedback, suggestions, enthusiastic responses and example plots that are shared on Twitter are highly appreciated.

References

Becares, N., M.C. Gage, M. Voisin, E. Shrestha, L. Martin-Gutierrez, N. Liang, R. Louie, B. Pourcet, O.M. Pello, T.V. Luong, S. Goñi, C. Pichardo-Almarza, H. Røberg-Larsen, V. Diaz-Zuccarini, K.R. Steffensen, A. O'Brien, M.J. Garabedian, K. Rombouts, E. Treuter, and I. Pineda-Torra. 2019. Impaired LXRα Phosphorylation Attenuates Progression of Fatty Liver Disease. *Cell Rep.* 26:984-995.e6. doi:https://doi.org/10.1016/j.celrep.2018.12.094.

Cui, X., and G.A. Churchill. 2003. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol.* 4:210. doi:10.1186/gb-2003-4-4-210.

Gillingham, A.K., J. Bertram, F. Begum, and S. Munro. 2019. In vivo identification of GTPase interactors by mitochondrial relocalization and proximity biotinylation. *Elife*. 8:e45916. doi:10.7554/eLife.45916.

Goedhart, J. 2020. PlotTwist: A web app for plotting and annotating continuous data. *PLOS Biol.* 18:e3000581.

Li, W. 2012. Volcano plots in analyzing differential expressions with mRNA microarrays. *J. Bioinform. Comput. Biol.* 10:1231003. doi:10.1142/S0219720012310038.

Naumov, V., I. Balashov, V. Lagutin, P. Borovikov, and A. Alexeev. 2017. VolcanoR - web service to produce volcano plots and do basic enrichment analysis. *bioRxiv*. 165100. doi:10.1101/165100.

Postma, M., and J. Goedhart. 2019. PlotsOfData—A web app for visualizing data together with their summaries. *PLOS Biol.* 17:e3000202. doi:10.1371/journal.pbio.3000202.

Singh, S., M.Y. Hein, and A.F. Stewart. 2016. msVolcano: a flexible web application for visualizing quantitative proteomics data. *bioRxiv*. 38356. doi:10.1101/038356.

van der Weegen, Y., H. Golan-Berman, T.E.T. Mevissen, K. Apelt, R. González-Prieto, J. Goedhart, E.E. Heilbrun, A.C.O. Vertegaal, D. van den Heuvel, J.C. Walter, S. Adar, and M.S. Luijsterburg. 2020. The cooperative action of CSB, CSA, and UVSSA target TFIIH to DNA damage-stalled RNA polymerase II. *Nat. Commun.* 11:2104. doi:10.1038/s41467-020-15903-8.