frontiers | Frontiers in Bioinformatics

TYPE Technology and Code
PUBLISHED 18 July 2022
DOI 10.3389/fbinf.2022.831025

OPEN ACCESS

EDITED BY
Sean O'Donoghue,
Garvan Institute of Medical Research, Australia

REVIEWED BY
Daofeng Li,
Washington University in St. Louis, United States
Sergio Martinez Cuesta,
AstraZeneca, United Kingdom

*CORRESPONDENCE
Douglas J. Chapski,
dchapski@ucla.edu

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Data Visualization, a section of the journal Frontiers in Bioinformatics

RECEIVED 07 December 2021
ACCEPTED 28 June 2022
PUBLISHED 18 July 2022

CITATION
Chen J, Zhu AJ, Packard RRS, Vondriska TM and Chapski DJ (2022), genomeSidekick: A user-friendly epigenomics data analysis tool. *Front. Bioinform.* 2:831025. doi: 10.3389/fbinf.2022.831025

genomeSidekick: A user-friendly epigenomics data analysis tool

Junjie Chen[1†], Ashley J. Zhu[2†], René R. S. Packard[1,3,4,5], Thomas M. Vondriska[1,2,3] and Douglas J. Chapski[2*]

[1]Division of Cardiology, Department of Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, United States, [2]Department of Anesthesiology and Perioperative Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, United States, [3]Department of Physiology, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, United States, [4]Ronald Reagan UCLA Medical Center, Los Angeles, CA, United States, [5]Veterans Affairs West Los Angeles Medical Center, Los Angeles, CA, United States

Recent advances in epigenomics measurements have resulted in a preponderance of genomic sequencing datasets that require focused analyses to discover mechanisms governing biological processes. In addition, multiple epigenomics experiments are typically performed within the same study, thereby increasing the complexity and difficulty of making meaningful inferences from large datasets. One gap in the sequencing data analysis pipeline is the availability of tools to efficiently browse genomic data for scientists that do not have bioinformatics training. To bridge this gap, we developed genomeSidekick, a graphical user interface written in R that allows researchers to perform bespoke analyses on their transcriptomic and chromatin accessibility or chromatin immunoprecipitation data without the need for command line tools. Importantly, genomeSidekick outputs lists of up- and downregulated genes or chromatin features with differential accessibility or occupancy; visualizes omics data using interactive volcano plots; performs Gene Ontology analyses locally; and queries PubMed for selected gene candidates for further evaluation. Outputs can be saved using the user interface and the code underlying genomeSidekick can be edited for custom analyses. In summary, genomeSidekick brings wet lab scientists and bioinformaticians into a shared fluency with the end goal of driving mechanistic discovery.

KEYWORDS

epigenomics, chromatin, data visualization, Shiny app, bioinformatics

Introduction

Computational biology tools written in different languages and applied across diverse fields allow for creative interrogation of genomics data to make biological conclusions. Understandably, the breadth of online genomic data analysis resources may appear overwhelming to a novice programmer. Fortunately, global efforts to bring bioinformatics training to general researchers are well underway (Mulder et al., 2018). Nevertheless, learning how to code may be a barrier to entry for non-bioinformaticians into the field of epigenomics, yet it is important to incorporate these researchers into the data analysis

process. A logical solution to this training issue is an inclusive approach that brings non-bioinformaticians into computational workflows after completing most of the command line processes, thereby fostering scientific creativity and leveraging shared knowledge about how the data are processed, analyzed, and visualized.

While lab skillsets ideally include formal bioinformatics knowledge, genomic researchers who do not understand how to code can readily make meaningful conclusions using processed data. An unmet need within this realm is a software for visualizing genomics data and filtering epigenomic and transcriptomic results for downstream analyses, especially considering the combination of orthogonal genomic datasets required to reveal more comprehensive mechanisms of cell biology. In addition, while Excel is a common tool for management and visualization of data, gene lists can be imported incorrectly into Excel and cause permanent edits to gene names (Ziemann et al., 2016). To prevent this issue and to promote independence from the bioinformatician, the next logical step is to furnish tools to perform data operations that a novice researcher might otherwise try in Excel.

The availability of distinct measurements to understand genomic mechanisms governing complex cellular and organ phenotypes has increased over time, resulting in a need to combine datasets (Chapski and Vondriska, 2021). Our recent study using RNA-seq, ATAC-seq, reduced representation bisulfite sequencing (RRBS), and chromatin structural data is an example of such integration of orthogonal data to make meaningful conclusions about chromatin architectural dynamics during heart failure (Chapski et al., 2021). Another investigation established an Atlas of murine ATAC-seq and RNA-seq data across 86 immune cell types and integrated the two datasets to identify a subset of cell types containing open regulatory elements bound by retinoic acid receptor-related orphan receptor gamma (RORγ) or paired-box protein PAX5 (as measured by ChIP-seq), thereby linking chromatin accessibility, transcription, and transcription factor binding in specific cell types (Yoshida et al., 2019). A common feature of all 'omics investigations is the need to ask questions of the massive datasets once acquired—to prioritize for further mechanistic evaluation. We also appreciate that even professional bioinformaticians may not have the time to perform bespoke analyses for collaborators: thus, a tool for transforming lists of genes into functional targets for a focused, mechanistic experiment is an opportunity to bring non-computational scientists and clinicians into the genomic analysis process.

To bridge the gap between processed data and biological inference, we built user-friendly genomic data visualization and manipulation tools for investigators without computational training. This GUI-based software called genomeSidekick allows for investigation of transcriptomic (RNA-seq) data in addition to chromatin accessibility (ATAC-seq) and chromatin immunoprecipitation-sequencing (ChIP-seq) data in a web browser. Based on a Shiny (Chang et al., 2021) dashboard written in R (Team, R. C, 2020), our tool—which we have named genomeSidekick—generates commonly used, intuitive graphs with interactive information retrieval. Moreover, we wrapped data visualization features for each individual experiment (RNA-seq, ATAC-seq, and ChIP-seq) into individual tabs to make switching between experiments easier. We also provide a tab to integrate RNA-seq, ATAC-seq, and ChIP-seq datasets, so modulation of the transcriptome and epigenome can be examined based on multiple criteria from independent experiments. Lastly, we provide links to external tools (and offer to perform small analyses locally) to facilitate Gene Ontology analysis and PubMed searches.

Freely available on GitHub (https://www.github.com/dchapski/genomeSidekick), genomeSidekick also contains extensive user-friendly documentation in a README markdown file with informational links so that most novice bioinformaticians can achieve results quickly. Lastly, genomeSidekick is a customizable tool that allows for code editing to support a shared collaboration between bioinformaticians and non-computational personnel in the biological research setting, thereby promoting increased computational engagement by non-bioinformaticians.

Methods

To run genomeSidekick, users should download the software from the repository on GitHub (https://www.github.com/dchapski/genomeSidekick) and then open the app.R file using Rstudio and click the "Run" button in the upper right corner of the script. Alternatively, users can download the code and run the app directly from the terminal using "R -e shiny:runApp ("/path/to/app.R")." Comma-separated or tab-delimited input RNA-seq data should include gene names (either identifiers or common names) with an adjusted *p*-value and log2FoldChange (preferably from a tool such as DESeq2 (Love et al., 2014) or edgeR (Robinson et al., 2010), which corrected *p*-values for multiple testing and provide fold change information). Comma-separated or tab-delimited input ATAC-seq data should include adjusted *p*-values and log2FoldChange information about accessibility peaks [the output from DiffBind (Ross-Innes et al., 2012) works well], in addition to either the closest gene or an overlapping gene for each feature. Gene names should be included for the ATAC-seq data as they are required for merging the RNA-seq and ATAC-seq dataset; however, independent analysis of ATAC-seq data alone does not require gene names. Importantly, other epigenomic experiments such as ChIP-seq outputs containing log2FoldChanges and adjusted *p*-values can be used on the genomeSidekick platform, either alone or in combination with RNA-seq data as described above for the ATAC-seq tab. We provide test data on GitHub and a hyperlink to the data directly within the app.

Extensive documentation regarding installation of R, RStudio, and dependencies for genomeSidekick is provided on

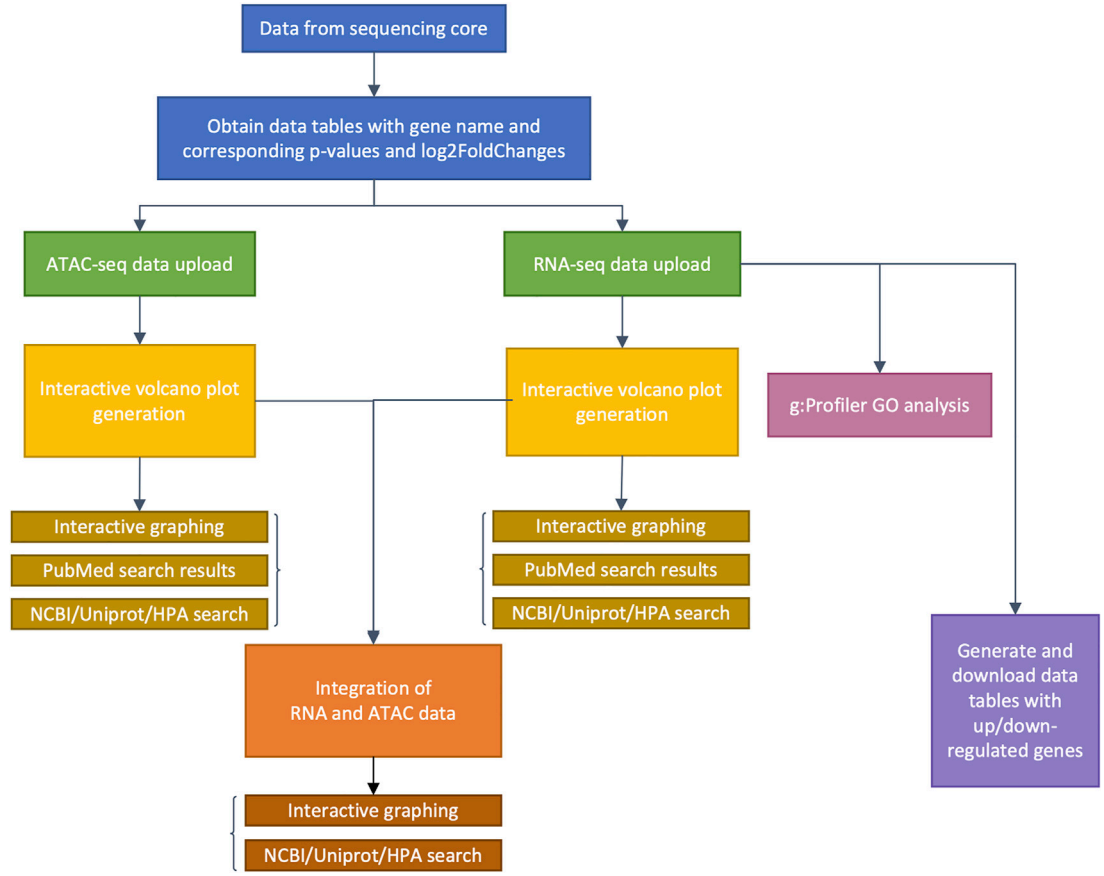

FIGURE 1
Flow chart detailing genomeSidekick functionalities. After raw sequencing data is obtained from the sequencing core, alignment, and statistical analyses on genes or chromatin features should be performed by a bioinformatician, with at least three columns as a result: feature name, adjusted *p*-values, and log2FoldChange (blue boxes). Then, genomeSidekick users can upload these data tables (.csv or tab-delimited files) as inputs to their corresponding tabs within the Shiny app (green boxes), which will contain functions for generation of independent interactive volcano plots (yellow boxes) as well as integrated ATAC-seq and RNA-seq graphs (orange boxes), for example. Note that in addition to ATAC-seq, users may analyze ChIP-seq and other epigenomics data inputs containing *p*-values and log2FoldChanges. Uploading an RNA-seq data set will also allow the application to perform g:Profiler GO analysis (magenta box) and generate downloadable lists of up- and downregulated genes (purple box). HPA: Human Protein Atlas.

the GitHub page. This documentation is also provided within the software so users can directly find information on how to run the software within the app. We also provide an install.R script on the GitHub page to facilitate installation of dependencies. To run the app in a password protected location online, a Shiny subscription can be purchased from the RStudio website (pricing starts at $9 USD/month in 2021). For exploration, we also provide online access at https://genomesidekick.shinyapps.io/genomesidekick/.

Results

The genomeSidekick software, written in R, can be run on a laptop and requires few dependencies to analyze RNA-seq, ATAC-seq, ChIP-seq, and any other epigenomics datasets that contain *p*-values and fold changes. The utility of genomeSidekick comes from its interface built on the Shiny framework in R (Team, 2020). This genomics dashboard allows separation of experimental strategies *via* individual tabs in the GUI (Figure 1). Inputs include processed data tables that can be loaded directly into the app. For example, an output from DESeq2 (Love et al., 2014) that contains the log2FoldChange and adjusted *p*-value information required for the visualizations. Other inputs include the output from DiffBind (Ross-Innes et al., 2012), a tool that statistically evaluates differentially bound or accessible genomic regions in the case of chromatin immunoprecipitation followed by sequencing (ChIP-seq) or ATAC-seq data, respectively.

Visualizations for volcano plots are coded using ggplot2 (Wickham, 2016) based code and visualized using ggplotly (Sievert, 2020), an open-source R package allows for

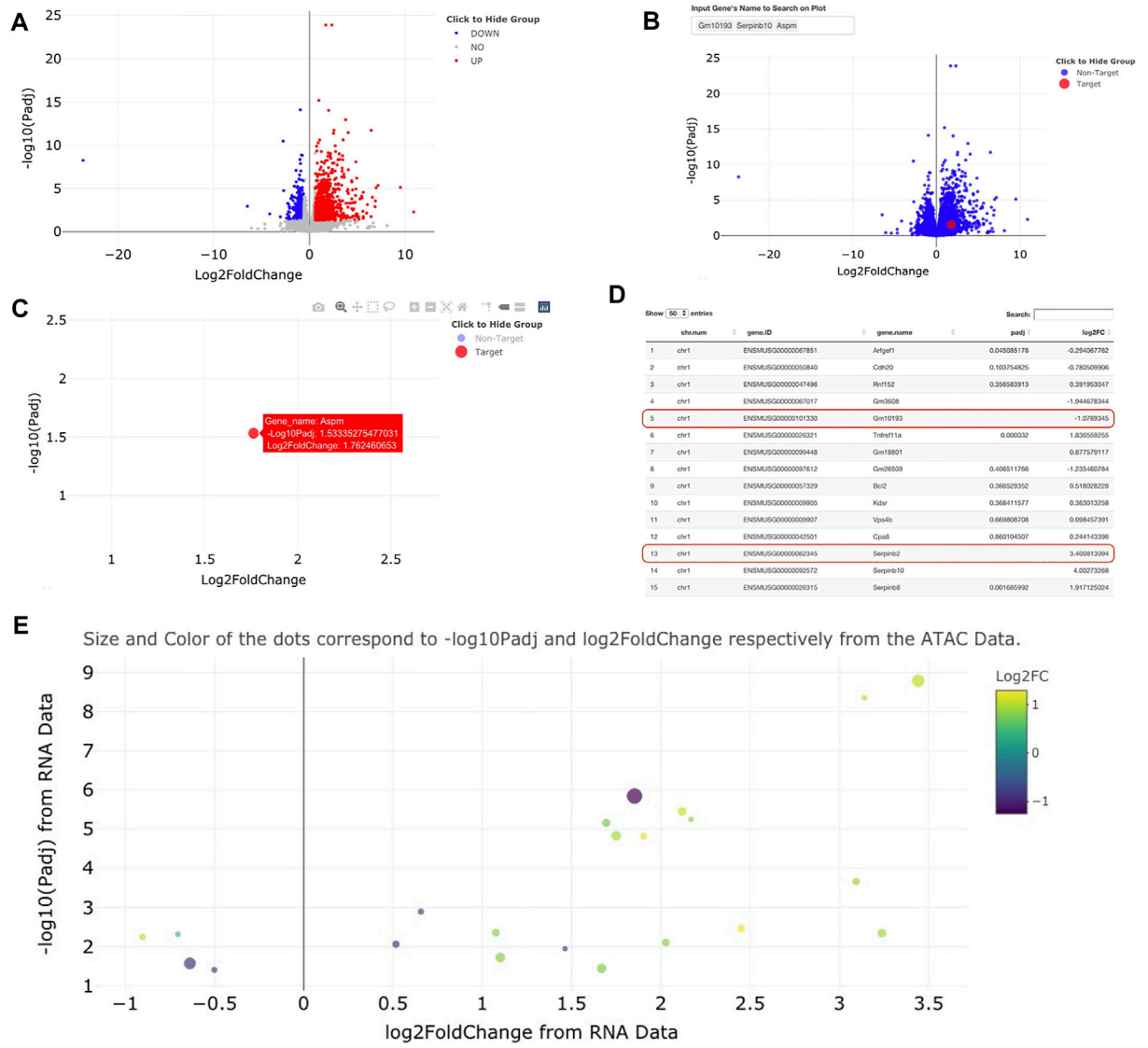

FIGURE 2
Example visualizations of sample RNA-seq and ATAC-seq data using genomeSidekick. **(A)** Volcano plot of RNA-seq data shows down-regulated genes in blue, upregulated genes in red, and non-significant genes in gray. Thresholds for *p*-value and log2FoldChange can be adjusted within genomeSidekick. **(B)** Searching for a particular gene(s) will highlight the target gene in red and mark the non-target genes in blue. In the example query for three genes, only one is found and visualized on the graph. **(C)** The non-target genes can be hidden by clicking on the Non-Target label in the legend (resulting in a grey label) and un-hidden by clicking the label a second time. Details about any gene shown on the interactive volcano plot can be shown by hovering over the dot on the plot. **(D)** Looking at the RNA-seq data upload page, genes without a *p*-value are not plotted on the volcano plot and therefore will not return any results when queried on the graph (see red lines). Common RNA-seq analysis packages do not evaluate all genes in the genome due to low detection, and this varies by experiment. Panels **(A)** through **(D)** use RNA-seq data as an example, but the functionalities are the exact same for ATAC-seq data. **(E)** Once both RNA-seq and ATAC-seq data sets are uploaded, they can be integrated into one graph using either the RNA-seq or ATAC-seq data as the base. Example data shown have padj <0.05 and FDR <0.1 in the RNA and ATAC-seq datasets, respectively.

interactive inspection of graphs (Figures 2A–C). The ggplotly visualizations allow for truly interactive point-by-point investigation to reveal individual metrics about each data point (gene name, adjusted *p*-value, log2FoldChange, and other custom information within the table). Superimposed on these visualizations are gene names highlighted by small lines [visualized using ggrepel (Slowikowski, 2021)] to indicate the *n* most significant points in the dataset. Notably, when a gene point is clicked within a volcano plot, genomeSidekick links the user to either the NCBI database, the UniProt (UniProt, 2021) website, or the Human Protein Atlas (Uhlen et al., 2015) for further

FIGURE 3
Example integrated RNA-seq/ATAC-seq visualization from a collaborator's first time using genomeSidekick. A collaborator recently visualized change in transcription and promoter accessibility at the murine *Xirp2* locus after 3 days cardiac pressure overload. This gene, also upregulated at the transcriptional level after angiotensin II treatment, also becomes more accessible with pressure overload, suggesting that the observed increase in transcription is a consequence of increased accessibility. Dot size and color scale indicate ATAC-seq-log10(adjusted *p*-value) and log2FoldChange, respectively. Example data shown have padj and FDR less than 0.05 in the RNA and ATAC-seq datasets, respectively.

TABLE 1 Example graphical user interfaces for genomics tasks (genomeSidekick in bold).

Software	Advantages and considerations	Reference
DEApp	Differential expression and data visualization in one tool, many options to calculate statistics	Li and Andrade, (2017)
DEBrowser	End-to-end analysis (filtering, heatmaps, dimensionality reduction), may require more than basic knowledge of statistics	Kucukural et al. (2019)
VisRseq	End-to-end analysis (filtering, heatmaps, dimensionality reduction), requires knowledge of JavaScript Object Notation (JSON)	Younesy et al. (2015)
genomeSidekick	**Volcano plots, experiment integration, Gene Ontology analysis, PubMed search, suitable for early beginners**	**This paper**

investigation of candidate gene functions. Some genes are not available for data visualization since many tools that calculate differential expression/accessibility only statistically evaluate loci containing experimental data, thereby resulting in unmeasured regions without a *p*-value (Figure 2D shows an example of this phenomenon). When RNA-seq and ATAC-seq (or ChIP-seq) inputs include common genomic feature information (for example, gene names), genomeSidekick can merge and filter these tables to produce a list of the *n* most upregulated and downregulated genes with accessibility information. In addition, the merge computation is performed in a way that gene names do not become corrupted from loading data in Excel [for more, see Introduction above and (Ziemann et al., 2016)]. This merged dataset can then be visualized as a volcano plot: one dataset (RNA-seq, for example) is plotted along the axes and the other dataset visualized using different point size and coloring to show additional information (Figure 2E).

To test the ease of dataset integration in a setting outside our institution, a collaborator provided a use case for custom analysis of RNA-seq and ATAC-seq data from (Chapski et al., 2021). Specifically, this collaborator sought to determine how the expression and chromatin accessibility at gene loci change with 3 days cardiac pressure overload (a pathological model that eventually leads to heart failure) in mice. Interestingly, the integrated RNA-seq and ATAC-seq output of genomeSidekick showed a significant increase in transcription and chromatin accessibility at the *Xirp2* gene locus (Figure 3), consistent with an earlier study showing that the cardiac stressor angiotensin II elicits an increase in *Xirp2* transcription mediated by the transcription factor MEF2A (McCalmon et al., 2010). This exercise, performed on the collaborator's first exploration of the software, suggests that genomeSidekick is useful for quick exploration of datasets and can provide meaningful scientific insights to first time users.

Importantly, the gene list outputs from each genomeSidekick tab are displayed for direct use as inputs for other software. For example, genomeSidekick includes a feature to perform local Gene Ontology analysis on smaller gene list outputs using the gprofiler2 (Kolberg et al., 2020) package in R in addition to a link to the g:Profiler website (Raudvere et al., 2019) for analyses of larger output gene lists from genomeSidekick that might take longer on a local machine. Lastly, we include a feature for quick PubMed searches of genes of interest that outputs query results directly in the app. This feature is based on the easyPubMed package in R (Fantini, 2019) and is designed to keep users' eyes on their data instead of opening a new tab to perform queries on data points of interest. Taken together, these features allow a non-bioinformatician to increase their computational fluency without having to learn how to code.

Discussion

We built a tool called genomeSidekick to facilitate inclusion of non-bioinformaticians into computational workflows for RNA-seq, ATAC-seq, ChIP-seq, or any other datasets that undergo statistical testing. This GUI-based software written in R allows individuals to focus their efforts on biological inference without having to frontload the bioinformatics training required to maneuver the command line. Specifically, genomeSidekick facilitates integration of gene expression and chromatin accessibility data, for example, to narrow down gene lists for further analyses. In addition, the software provides an opportunity for non-bioinformaticians to perform small edits to the code to customize their visualizations and filtering criteria without having to learn R. Overall, genomeSidekick will bring wet lab scientists onto a more level playing field for common data analysis questions, thereby reducing dependence on bioinformaticians.

Additional software exists for analysis of gene expression and epigenomics data and may be useful for more computationally versed individuals. For example, DEApp can be used to perform differential expression testing and data visualization (Li and Andrade, 2017), although a significant hurdle to using this tool is knowing which statistical approach to use for differential expression within the software. In addition, DEBrowser (Kucukural et al., 2019) and VisRseq (Younesy et al., 2015) are useful for performing end-to-end bioinformatics analyses of datasets, and both programs complete complicated tasks such as heatmap generation and principal component analysis. Importantly, these tools may require knowledge of data transformations at each step of a given analysis and/or training in statistics. In contrast, genomeSidekick provides a platform for users to explore and integrate processed transcriptomics and epigenomics data and create figures without the complexity seen in other tools (Table 1).

The simplicity of genomeSidekick allows researchers with no bioinformatics background to investigate their own datasets after initial mapping, quantification, and differential testing by a bioinformatician. Thresholding of *p*-values for individual experiments can be edited for custom stringency, which allows wet lab researchers to perform independent analyses without requesting individual gene lists from a bioinformatician. Moreover, extensive documentation providing explanations of individual functions and links to learning resources is condensed into an intuitive README file on GitHub with an intuitive interface and examples.

The genomeSidekick application allows bioinformaticians to send data to collaborators and then have them interact with multiple datasets independently. Importantly, the app can be hosted online for a small monthly fee using https://www.shinyapps.io, thereby facilitating longer distance collaborations. Accordingly, for simple data exploration, we provide genomeSidekick online at https://genomesidekick.shinyapps.io/genomesidekick/. Overall, genomeSidekick will help bring wet lab researchers into the computational realm by fostering creativity with data visualization and integrative analyses in a user-friendly format.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/dchapski/genomeSidekick.

Author contributions

JC and DC conceived of the study. JC, AZ, and DC wrote the software. RS and TV provided infrastructure. DC supervised the project and wrote the manuscript. All authors read and approved the final manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Acknowledgments

The authors would like to thank collaborator Dr. Christoph D. Rau for software testing and members of the Vondriska Lab for comments and suggestions. Research in the Vondriska Lab is supported by the NIH, UCLA Clinical and Translational Science Institute, the Department of Anesthesiology and Perioperative Medicine, and the David Geffen School of Medicine at UCLA. RP is supported by VA Merit BX004558 and the UCLA Cardiovascular Discovery Fund/Lauren B. Leichtman and Arthur E. Levine Investigator Award.

References

Chang, W., Cheng, J., Allaire, J. J., Sievert, C., Schloerke, B., Xie, Y., et al. (2021). *Shiny: web application framework for R*.

Chapski, D. J., Cabaj, M., Morselli, M., Mason, R. J., Soehalim, E., Ren, S., et al. (2021). Early adaptive chromatin remodeling events precede pathologic phenotypes and are reinforced in the failing heart. *J. Mol. Cell. Cardiol.* 160, 73–86. doi:10.1016/j.yjmcc.2021.07.002

Chapski, D. J., and Vondriska, T. M. (2021). Taking data science to heart: Next scale of gene regulation. *Curr. Cardiol. Rep.* 23 (5), 46. doi:10.1007/s11886-021-01467-6

Fantini, D. (2019). easyPubMed: *search and retrieve scientific publication records from PubMed*.

Kolberg, L., Raudvere, U., Kuzmin, I., Vilo, J., and Peterson, H. (2020709). gprofiler2– an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler. *F1000Res.* 9 (ELIXIR). doi:10.12688/f1000research.24956.1

Kucukural, A., Yukselen, O., Ozata, D. M., Moore, M. J., and Garber, M. (2019). DEBrowser: interactive differential expression analysis and visualization tool for count data. *BMC Genomics* 20 (1), 6. doi:10.1186/s12864-018-5362-x

Li, Y., and Andrade, J. (2017). DEApp: An interactive web interface for differential expression analysis of next generation sequence data. *Source Code Biol. Med.* 12, 2. doi:10.1186/s13029-017-0063-4

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15 (12), 550. doi:10.1186/s13059-014-0550-8

McCalmon, S. A., Desjardins, D. M., Ahmad, S., Davidoff, K. S., Snyder, C. M., Sato, K., et al. (2010). Modulation of angiotensin II-mediated cardiac remodeling by the MEF2A target gene Xirp2. *Circ. Res.* 106 (5), 952–960. doi:10.1161/CIRCRESAHA.109.209007

Mulder, N., Schwartz, R., Brazas, M. D., Brooksbank, C., Gaeta, B., Morgan, S. L., et al. (2018). The development and application of bioinformatics core competencies to improve bioinformatics training and education. *PLoS Comput. Biol.* 14 (2), e1005772. doi:10.1371/journal.pcbi.1005772

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., et al. (2019). g:profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* 47 (W1), W191–W198. doi:10.1093/nar/gkz369

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26 (1), 139–140. doi:10.1093/bioinformatics/btp616

Ross-Innes, C. S., Stark, R., Teschendorff, A. E., Holmes, K. A., Ali, H. R., Dunning, M. J., et al. (2012). Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* 481 (7381), 389–393. doi:10.1038/nature10730

Sievert, C. (2020). *Interactive web-based data visualization with R, plotly, and shiny*. Chapman and Hall/CRC.

Slowikowski, K. (2021). *ggrepel: Automatically position non-overlapping text labels with 'ggplot2*.

Team, R.C (2020). *R: A language and environment for statistical computing*.

Uhlen, M., Fagerberg, L., Hallstrom, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., et al. (2015). Proteomics. tissue-based map of the human proteome. *Science* 347 (6220), 1260419. doi:10.1126/science.1260419

UniProt, C., Martin, M. J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., et al. (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* 49 (D1), D480–D489. doi:10.1093/nar/gkaa1100

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag.

Yoshida, H., Lareau, C. A., Ramirez, R. N., Rose, S. A., Maier, B., Wroblewska, A., et al. (2019). The cis-regulatory atlas of the mouse immune system. *Cell.* 176 (4), 897–912. doi:10.1016/j.cell.2018.12.036

Younesy, H., Moller, T., Lorincz, M. C., Karimi, M. M., and Jones, S. J. (2015). VisRseq: R-based visual framework for analysis of sequencing data. *BMC Bioinforma.* 16 (Suppl. 11), S2. doi:10.1186/1471-2105-16-S11-S2

Ziemann, M., Eren, Y., and El-Osta, A. (2016). Gene name errors are widespread in the scientific literature. *Genome Biol.* 17 (1), 177. doi:10.1186/s13059-016-1044-7