Version: “21_07Interactive_heatmap_manuscript.doc”

21 July 2023, SJB edits added.

Title Page

HIVE browser: a user-friendly tool for interactive viewing and analysis of gene expression data

Stephen Bridgett, Sara El jadid, Guillermo Lopez-Campos, David A Simpson

s.bridgett@qub.ac.uk

s.eljadid@qub.ac.uk

g.lopezcampos@qub.ac.uk

david.simpson@qub.ac.uk

Wellcome-Wolfson Institute for Experimental Medicine

School of Medicine, Dentistry and Biomedical Sciences

Queen’s University Belfast

97 Lisburn Road

Belfast

BT9 7BL

ORCID IDs

David Simpson: 0000-0003-0157-1211

Stephen Bridgett: 0000-0001-9330-5820

Sara El jadid: 0000-0001-9793-5061

Guillermo Lopez-Campos: 0000-0003-3011-0940

Abstract

Summary: The potential of omics technologies is hindered by the need to navigate datasets of increasing size, number, and complexity. Interpreting multiple lists of genes differentially expressed by the many cell-types identified in single cell RNA-sequencing experiments is one common challenge. We therefore developed the '*Heatmap for Interactive Viewing of Expression data (HIVE)*' browser. It generates an instant heatmap of expression changes and allows users to submit selected genes to external tools for network and gene set enrichment analyses. Any tabular text file containing differential expression data (eg scRNA-Seq, proteomics etc.) can be loaded directly into the browser which operates as a standalone single-page-application (SPA) html file, enabling local use with restricted private data and facilitating the sharing of datasets in an accessible format. HIVE provides a user-friendly interactive overview of differential expression data within a familiar web-browser environment. The single webpage tabular format is familiar for those accustomed to spreadsheets and enables genes to be selected manually, or by filtering on fold change, probability, and comparison of interest. Selected genes can be visualised as a heatmap and analysed interactively with the most relevant interpretative tools.

Keywords: omics technologies, differential gene expression, single-cell RNA-sequencing, interactive visualization, data analysis, interpretative tools, gene set enrichment analysis, network analysis.

Availability and implementation: *HIVE* is implemented in html/css/javascript as a single page web application, with the String DB API javascript file as its only dependency.

HIVE is available free on GitHub-Pages: https://qub-simpson-lab.github.io/HIVE_browser/ or can be downloaded from Github respository at: https://github.com/QUB-Simpson-lab/HIVE_browser along with sample expression data in standard csv (comma-separated value) format. **Contact:** David Simpson, David.Simpson@qub.ac.uk

Supplementary information: A vignette is available at Bioinformatics online.

1 Introduction

Improved technology and reduced costs have made high throughput genome-wide expression analysis accessible for most life scientists, in addition many existing datasets are also available online. A single-cell RNA-sequencing experiment, for example, can reveal expression data for thousands of individual cells and even when aggregated by cell type, comparison of several conditions generates multiple differential expression files. While many tools have been developed for visualization of genomic data (Rue-Albrecht, et al., 2018; Tintori, et al., 2020) other refs?, they generally employ scripts which require installation and execution. For a majority of wet-lab users the use of these tools can be a barrier and therefore there is still an unmet need of tools that can make the exploration of such complex datasets easy in a an easily accessible way explore such data interactively is crucial to uncover potential functional effects of treatments and relationships between altered genes. To facilitate this process, we have developed the Heatmap for Interactive Viewing of Expression data (HIVE) web-browser. Although driven by needs of scRNA-Seq, HIVE can also be used to analyze bulk RNA-Seq or other differential expression datasets.

While many tools have been developed for visualization of genomic data (Rue-Albrecht, et al., 2018; Tintori, et al., 2020) other refs?, they generally employ scripts which require installation and execution. In contrast, HIVE is encoded within a single html file and therefore, although useful for those with computational experience, presents no barrier to new users who may be deterred by the need to install and run an R or python scripts, which often have dependencies on other scripts or C libraries. This also has the significant advantage that it runs locally within the user's web-browser and does not transfer data to a remote third-party server, retaining the security of sensitive data. While primarily designed with multi-condition gene expression datasets in mind, HIVE is broadly applicable to comparisons of any quantitative data from, for example, proteomic or epigenetic studies.

2 Results

2.1 Implementation

The HIVE webpage is written in HTML5, CSS 2, and Javascript ES5, which are widely supported across web-browsers. The only external file used is the "combined_embedded_network_v2.0.4.js" javascript file that is hosted on the string-db.org website and is used to send selected gene-ids to the string API and retrieve the network image or redirect to the network on the STRING website. Internally the web-page uses the "*FileReader.readAsText()*" API to read the input data files directly into the web-browser memory so does not need script on a webserver; and it also downloads the single file format file without needing a server script.

2.2 Requirements

The HIVE browser has been tested in the Chrome, Edge, Firefox and Safari web-browsers, on Windows, MacOS, Linux and Android. From the Chrome Task-Manager, running on MacOS, our internal dataset of 2,878 genes and 34 fold-change columns used 625 MBytes of memory. No server infrastructure is needed because the full webpage and sample test data will be available long-term on Github and on Zenodo (Links above in: '**Availability and implementation'**).

2.3 Usage Use case?

The HIVE webpage includes expandable 'Help/Info' links; and detailed step-by-step instructions illustrated with test datasets are provided in the vignette available as Supplementary material. Here we outline the process of using the HIVE browser to investigate an exemplar single cell RNA-Seq dataset (Fig. 1) which is...

Possible datasets from GEO: GSE203499, GSE205123, GSE183206 Retinal degeneration, Our retina first dataset??

The first step is simply to enter the URL for the online version at github pages https://qub-simpson-lab.github.io/HIVE_browser/ , or download and open the html file (from: https://github.com/QUB-Simpson-lab/HIVE_browser). The HIVE browser webpage takes the user through the following steps and offers optional additional information boxes.

2.3.1 Data input

The ability to load data smoothly into the browser is critical to empower our wide target user base and we have therefore focused on simplifying this step. The default input data are differential gene (or protein) expression text files such as those from popular analytical packages such as Seurat, DeSeq2 or EdgeR. An example of the steps required to generate input files is provided by the Seurat vignette (https://satijalab.org/seurat/articles/de_vignette.html). Double quotes around values are optional, only being needed if the value contains the column separator (eg. a comma). For example:

	"p_val"	"avg_log2FC"	"pct.1"	"pct.2"	"p_val_adj"
"Rpgrip1"	2.4930235e-06	-0.70848114	0.025	0.186	0.04304205
"Ccdc126"	4.7435722e-06	-0.33199724	0.006	0.137	0.0818977
"Epha6"	3.5653145e-05	0.77590795	0.46	0.288	0.61555154

...etc...

The "avg_log2FC" column can also be named "avg_logFC". If the "p_val_adj" is missing, then the unadjusted "p_val" will be used instead. The "pct.1" "pct.2" are optional, being optionally displayed if available. More details are provided in the HIVE webpage and vignette.

After clicking the "*Choose files*" button, the user can select their input data files from the standard open files dialog (using Ctrl/cmd or Shift key to select multiple files). The browser will try to automatically identify the gene/protein identifier type (eg. Ensembl or Gene-symbol etc) or it can be set manually, followed by selection of the species.

Any data in appropriate tab or comma delimited file formats can be analysed and the browser will try to determine from the titles which columns to use for fold-change, p-value, etc. If necessary, the user can select the columns containing fold-change and p-values/FDDR/q-values, and edit the headings as appropriate.

When the data has been loaded you can download it as a single file in a custom HIVE csv format which can be loaded more quickly in future or shared with collaborators. The HIVE browser can also read that HIVE single-file directly from a Dropbox.com link, or other webserver, by entering the link in the URL input box in the HIVE webpage, or specifying it as a "?url=...." (url-encoded) after the link the HIVE-browser, for example: https://github.pages.com/HIVE_browser1.html?url=www.dropbox.com%2Flink..... (finish this link). We envisage core genomics/bioinformatics facilities producing and supplying these files to the end-user, perhaps using a semi-automated pipeline. The end-user could then use the HIVE browser to

investigate their data, without needing computational expertise, (and the link with "?url=..." could even be shared in a journal article so the reader could simply click on that link to explore the data).

2.3.2 Displaying the data

Gene expression is displayed as an interactive HTML table with the names of genes/proteins in the left and right-most columns and the different cell-types/cluster and timepoint/condition along the top, and fold-change plus p-values in each corresponding table cell with appropriate heatmap background colour (ranging from dark blue, through green to dark red). Display of p-values for the fold-changes and percentage of cells where the feature is detected in the first & second groups can be toggled on or off. At any time an image of the heatmap can be downloaded in png, jpeg, gif or webp format, for use in publications or presentations. Fig 1x

2.3.3 Selection of genes of interest

The genes can be filtered by various criteria, including the minimum number of columns that have known fold-changes, the minimum required fold-change value (eg. >0.5 or: >0.5 & <-0.5) and the minimum P-value for the required fold-change. Criteria can be applied to specific conditions or cell-types. Alternatively, specific genes can be selected using checkboxes in the left column of the heatmap, or the 'All', 'None' buttons.

2.3.4 Functional analyses

Selected genes (and species) can be submitted directly for third party analysis. The *STRING* gene Network(Szklarczyk, et al., 2021) for the selected genes can be displayed within the browser or opened in a separate STRING tab to facilitate more in-depth interrogation. Functional enrichment within the selected gene list can be displayed by submitting it to *g:Profiler* (Raudvere, et al., 2019) with a single click. Further functional annotation can be achieved by submitting the selected genes to the *Database for Annotation, Visualization and Integrated Discovery (DAVID)* (Sherman, et al., 2022).

3. Conclusion

The HIVE browser provides an accessible way of exploring gene expression data. It is particularly useful for managing the multiple differential expression comparisons generated by single-cell RNA-Seq analyses. Our aim is to support a wide range of researchers and additional analytical features and direct import of more input data types will be added in response to feedback.

Acknowledgements

Funding information

This work was supported by a grant to DS from Fight for Sight [grant ref: 5091/5092].

References

Raudvere, U., *et al.* g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res* 2019;47(W1):W191-W198.

Rue-Albrecht, K., *et al.* iSEE: Interactive SummarizedExperiment Explorer. *F1000Res* 2018;7:741.

Sherman, B.T., *et al.* DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res* 2022;50(W1):W216-W221.

Szklarczyk, D., *et al.* The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res* 2021;49(D1):D605-D612.

Tintori, S.C., Golden, P. and Goldstein, B. Differential Expression Gene Explorer (DrEdGE): a tool for generating interactive online visualizations of gene expression datasets. *Bioinformatics* 2020;36(8):2581-2583.

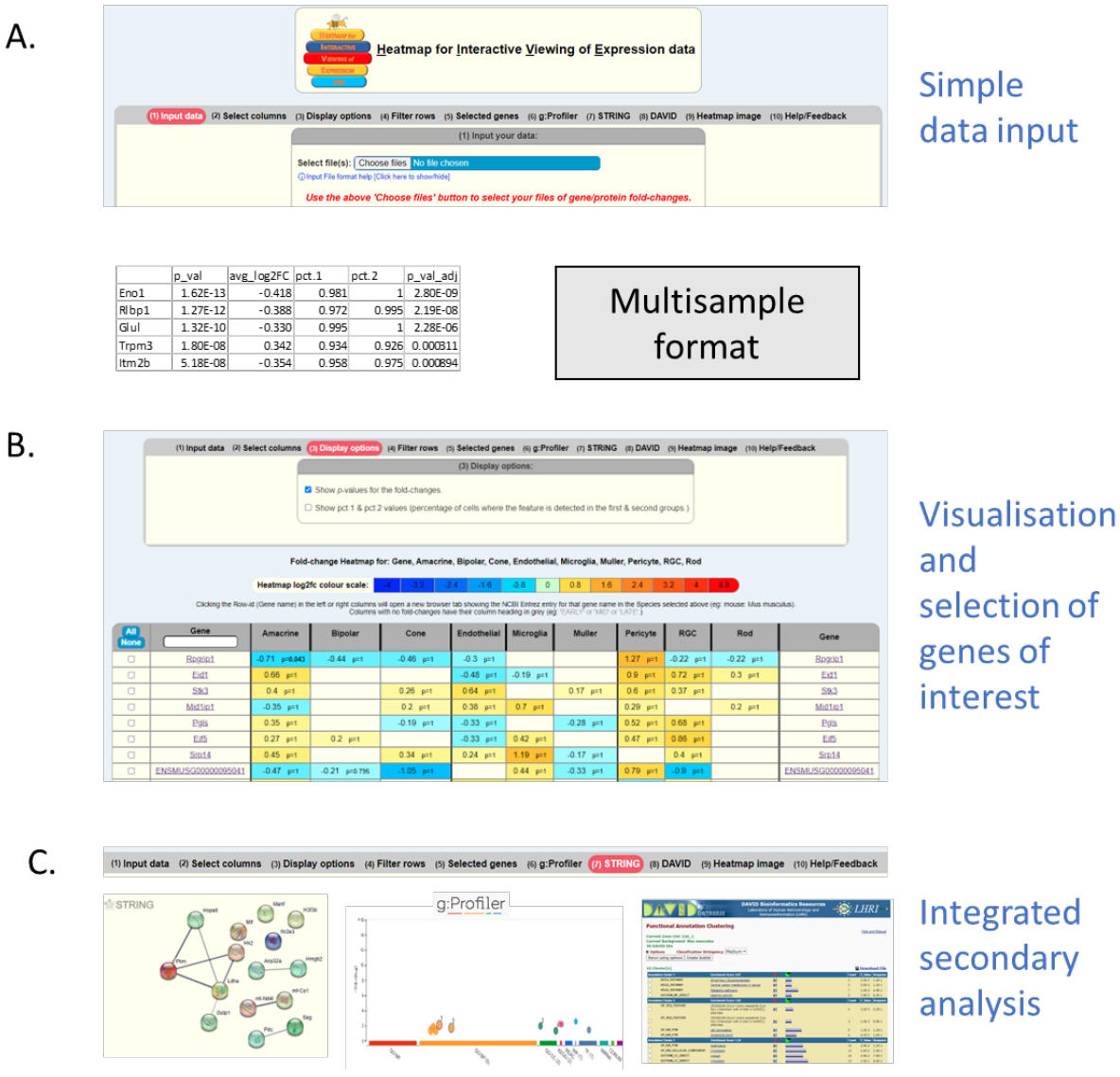
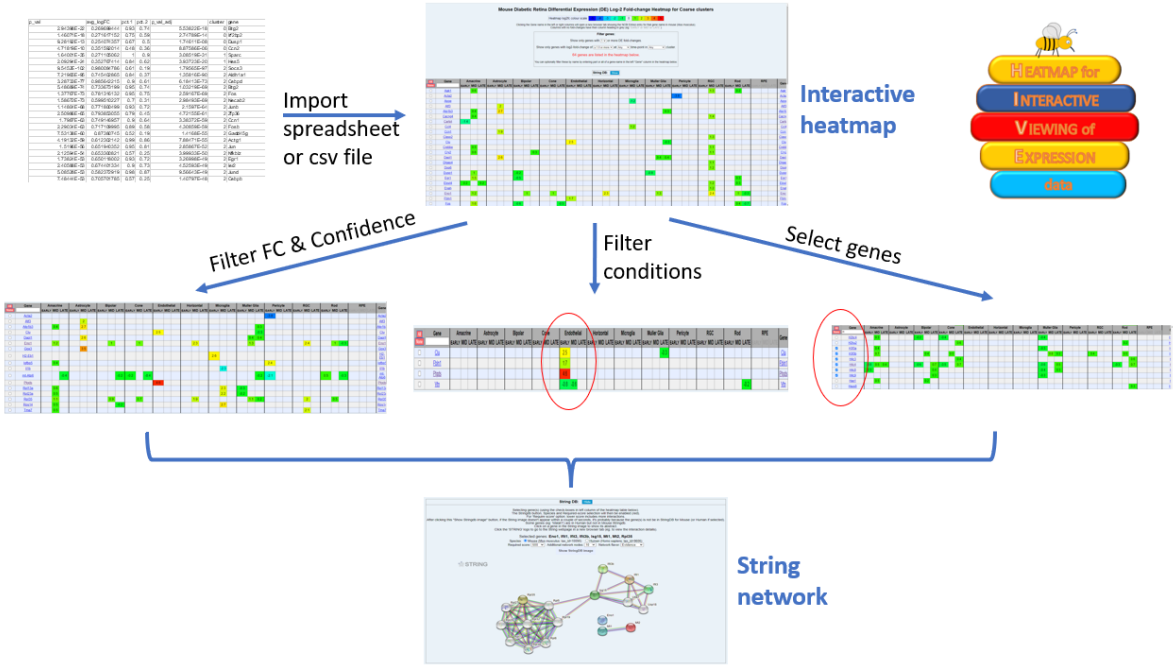


Figure 1. Features of the HIVE browser. Users navigate across a numbered series of interfaces within the browser window. First data input is by selecting text files in simple format… (A). Once loaded, data is displayed in a heatmap format and can be filtered to select genes of interest (B). The selected genes can be submitted directly to external tools for secondary analysis (C).

Graphical abstract?



Check this applies to Applications Notes: Authors are encouraged to submit a graphical abstract as part of the article, in addition to the text abstract. The graphical abstract should clearly summarize the focus and findings of the article, and will be published as part of the article online and in PDF. The graphical abstract should be submitted for peer review as a separate file, selecting the appropriate file-type designation in the journal's online submission system. The file should be clearly named, e.g. graphical_abstract.tiff. See Preparing and submitting your manuscript page for guidance on appropriate file format and resolution for graphics. Please ensure graphical abstracts are in landscape format.

Additional notes etc....

These should be included at the end of the text and not in footnotes. Please ensure you acknowledge all sources of funding, see funding section below.

Details of all funding sources for the work in question should be given in a separate section entitled 'Funding'. This should appear first in the 'Acknowledgements' section.

Funding information

This work was supported by:

The following rules should be followed:

The sentence should begin: ‘This work was supported by …’

The full official funding agency name should be given, i.e., ‘National Institutes of Health’, not ‘NIH’ (full RIN-approved list of UK funding agencies)

Grant numbers should be given in brackets as follows: ‘[grant number xxxx]’

Multiple grant numbers should be separated by a comma as follows: ‘[grant numbers xxxx, yyyy]’

Agencies should be separated by a semi-colon (plus ‘and’ before the last funding agency)

Where individuals need to be specified for certain sources of funding the following text should be added after the relevant agency or grant number 'to [author initials]'.

An example is given here: ‘This work was supported by the National Institutes of Health [AA123456 to C.S., BB765432 to M.H.]; and the Alcohol & Education Research Council [hfygr667789].’

Table of existing tools (from Dredge supplementary data)

Bioinformatics instructions

The title should be short, specific and informative, avoiding if possible version number and the words: tool, package, application, software (and similar). If novel software is being described, the name of the software should be included in the title. The surname and initials of each author should be followed by his/her department, Institution, city with postal code and country. Any changes of address may be added to the footnotes

Abstracts for Applications Notes are much shorter than those for an Original Paper. They are structured with four headings: Summary, Availability and Implementation, Contact and Supplementary Information.

- *Summary*: This section should summarize the purpose/novel features of the program in one or two sentences.
- *Availability and Implementation*: See above for advice and examples for this section.
- *Contact*: Full email address to be supplied, preferably an institutional address.
- *Supplementary information*: Links to additional figures/data available on a web site, or reference to online-only Supplementary data available at the journal's web site.

Application Notes (up to 2 pages; this is approx. 1,300 words or 1,000 words plus one figure): Applications Notes are short descriptions of novel software or new algorithm implementations, databases and network services (web servers, and interfaces). Software or data must be freely available to non-commercial users. Availability and Implementation must be clearly stated in the article. Authors must also ensure that the software is available for a full two years following publication. Web services must not require mandatory registration by the user. Additional supplementary data can be published online-only by the journal. This supplementary material should be referred to in the abstract of the Application Note. If describing software, the software should run under nearly all conditions on a wide range of machines. Web servers should not be browser specific. Application Notes must not describe trivial utilities, nor involve significant investment of time for the user to install. The name of the application should be included in the title.

Application notes: up to 2 pages; this is approx. 1300 words or 1000 words plus one figure.

Data Availability Statements

Data Availability Statements provide a standardised format to describe the availability of data underlying the research results of the article. The statement may refer to original data generated in the course of the study or to third-party data analysed in the article. The statement should describe and provide means of access, where applicable, by linking to the data or providing the required unique identifier.

The data underlying this article are available in [repository name, e.g. Zenodo], at https://dx.doi.org/[doi]. The datasets were derived from sources in the public domain: [list sources, including URLs].

ORCID

Bioinformatics requires submitting authors to provide an ORCID iD at submission to the journal.

Software

If the manuscript describes new software tools or the implementation of novel algorithms, the authors must:

- provide a self-contained and easy-to-use implementation of the software together with appropriate test data and detailed documentation on how to acquire, install, and run the software. This includes information on the purpose and values of all optional parameters.
- make the software and the source code freely available at stable URLs (e.g. github, bitbucket) without request to non-commercial users at the time of submission (availability must be clearly stated in the manuscript)
- ensure that the submitted software version and test data are archived at dedicated repositories such as Zenodo, Figshare, Software Heritage, CRAN, and Bioconductor and provide the corresponding URL in the availability section of the abstract.

We encourage authors to provide access through an open source license (see Open Source Initiative page for examples) and to implement the software using a non-proprietary programming language. We encourage authors to provide information on how to reproduce results/figures described in the manuscript. The name of the software should be included in the title of the manuscript wherever possible.

At least four suggested reviewers must be provided by the author at the submission stage

Cover letter

All submissions must be accompanied by a cover letter explaining the suitability of the paper for Bioinformatics. Submissions omitting a cover letter may be editorially rejected. If the submitted manuscript deviates from the requirements listed in the Instructions to Authors, the reasons for this must be clearly explained. For example, machine learning papers deviating from the listed requirements must explain why this is necessary and specifically suited to the problem addressed.