Phantasus, a web application for visual and interactive gene expression analysis

Maksim Kleverov[1,2†], Daria Zenkova[1†], Vladislav Kamenev[1†], Margarita Sablina[1], Maxim N Artyomov[2], Alexey A Sergushichev[1,2*]

[1]ITMO University, Computer Technologies Laboratory, Saint Petersburg, Russian Federation; [2]Washington University in St. Louis School of Medicine, Department of Pathology and Immunology, St Louis, United States

Abstract Transcriptomic profiling became a standard approach to quantify a cell state, which led to the accumulation of huge amount of public gene expression datasets. However, both reuse of these datasets or analysis of newly generated ones requires significant technical expertise. Here, we present Phantasus: a user-friendly web application for interactive gene expression analysis which provides a streamlined access to more than 96,000 public gene expression datasets, as well as allows analysis of user-uploaded datasets. Phantasus integrates an intuitive and highly interactive JavaScript-based heatmap interface with an ability to run sophisticated R-based analysis methods. Overall Phantasus allows users to go all the way from loading, normalizing, and filtering data to doing differential gene expression and downstream analysis. Phantasus can be accessed online at https://alserglab.wustl.edu/phantasus or can be installed locally from Bioconductor (https://bioconductor.org/packages/phantasus). Phantasus source code is available at https://github.com/ctlab/phantasus under an MIT license.

Editor's evaluation

This study introduces Phantasus, a useful tool accessible through both web and local applications, designed to analyze transcriptome data derived from microarray or RNA-seq technologies. The compelling tool facilitates normalization, data visualization, and differential expression analysis. Phantasus represents a valuable contribution to the biomedical community, enabling individuals without extensive bioinformatics expertise to analyze new transcriptomic data or reproduce studies effectively.

Introduction

Transcriptomic profiling is a ubiquitous method for whole-genome-level profiling of biological samples (***Stark et al., 2019***). Moreover, the deposition of these data into one of the public repositories has become a standard in the field, leading to the accumulation of a huge amount of publicly available data. The most significant example is the NCBI Gene Expression Omnibus (GEO) project (***Barrett et al., 2013***), which stores information from more than 225,000 studies.

Sharing of transcriptomic data opens up possibilities for reusing them: instead of carrying out a costly experiment, a publicly available dataset can be used, thus decreasing the cost and accelerating the research (***Byrd et al., 2020***). However, the standard approach for gene expression analysis requires significant technical expertise. In particular, many analysis methods are implemented in R as a part of Bioconductor project ecosystem (***Gentleman et al., 2004***), and thus one has to have programming skills in R to use them. On the other hand, domain knowledge is beneficial to improve quality

***For correspondence:** alsergbox@gmail.com

†These authors contributed equally to this work

Competing interest: The authors declare that no competing interests exist.

Funding: See page 7

Preprinted: 12 December 2022
Received: 21 December 2022
Accepted: 13 May 2024
Published: 03 June 2024

Reviewing Editor: Marisa Nicolás, Laboratório Nacional de Computação Científica, Brazil

control of the data, which is especially important when working with the publicly available data, as well as generation of biological hypotheses (***Wang et al., 2016***).

A number of applications have been developed with the aim to simplify analysis of transcriptomic datasets (see Appendix 1 for details). In particular, web-based applications remove the burden of setup and configuration from the end users, thus lowering the entry threshold. Shiny framework (***Chang, 2022***) revolutionized the field as it became easy to create a web interface for R-based pipelines, which led to a significant growth of web applications for gene expression analysis (***Ge et al., 2018***; ***Hunt et al., 2022***; ***Mahi et al., 2019***; ***Nelson et al., 2017***). However, such applications generally have limited interactivity due to mainly server-side computations. Shiny-independent applications can be more interactive, but they suffer from lack of native R support and require reimplementation of existing methods from scratch (***Gould, 2016***; ***Alonso et al., 2015***).

Here, we present Phantasus: a web application for gene expression analysis that integrates highly interactive client-side JavaScript heatmap interface with an R-based back-end. Phantasus allows us to carry out all major steps of gene expression analysis pipeline: data loading, annotation, normalization, clustering, differential gene expression, and pathway analysis. Notably, Phantasus provides streamlined access to more than 96,000 microarray and RNA-seq datasets from the GEO database, simplifying their reanalysis. Phantasus can be accessed online at https://alserglab.wustl.edu/phantasus or can be installed locally from Bioconductor. Phantasus is open source, and its code is available at https://github.com/ctlab/phantasus under MIT license (***Sergushichev and Kamenev, 2024***).

Results

Phantasus web application

We developed a web application called Phantasus for interactive gene expression analysis. Phantasus integrates a JavaScript-rich heatmap-based user interface originated from Morpheus (***Gould, 2016***) with an R back-end via the OpenCPU framework (***Ooms, 2014***). The heatmap graphical interface provides an intuitive way to manipulate the data and metadata: directly in a web browser the user can create or modify annotations, edit color schemes, filter rows and columns, and so on. On the other hand, the R back-end provides a way to easily run a multitude of computational analysis methods available as R packages. Altogether, this architecture (***Figure 1***) ensures a smooth experience for performing all common analysis steps: loading datasets, normalization, exploration, visualization, differential expression and gene set enrichment analyses.

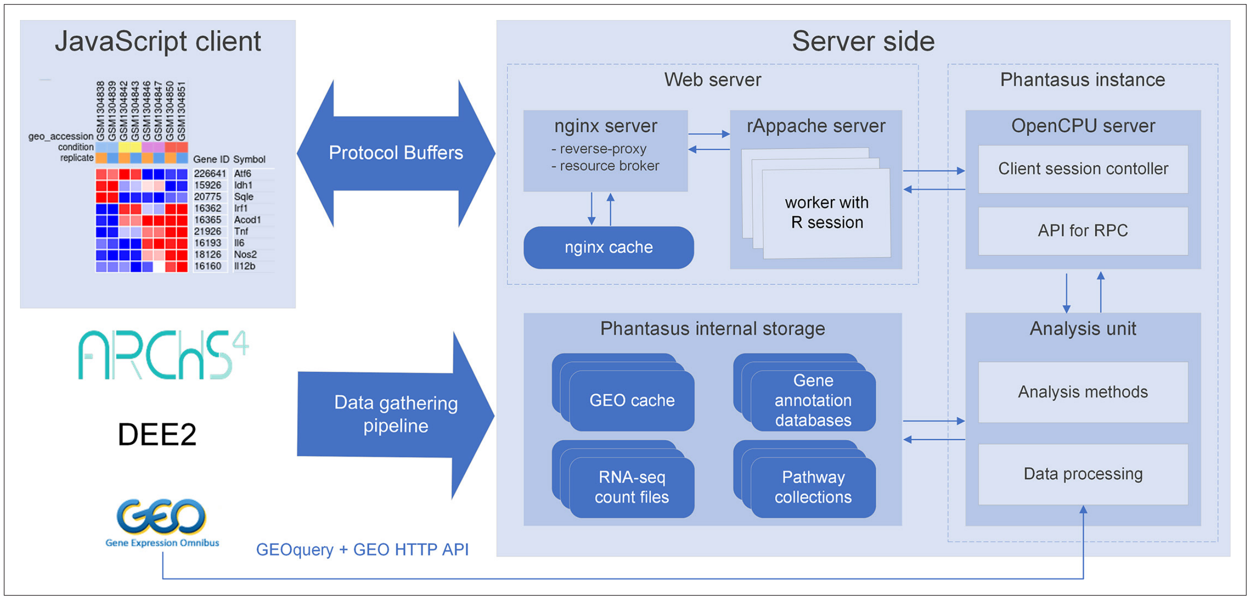


Figure 1. Overview of Phantasus architecture. The front-end interface is a JavaScript application that requests the web server to load the data and perform resource-consuming tasks. The core element of the back-end is the OpenCPU-based server, which triggers the execution of R-based analysis methods. Protocol Buffers are used for efficient client-server dataset synchronization.

Several options for loading the gene expression data into Phantasus are available. First, datasets from GEO (***Barrett et al., 2013***) can be loaded by their identifier. Phantasus supports microarray datasets, which are loaded directly from GEO, as well as RNA-sequencing datasets, for which counts data from third-party databases are used (see section 'Available datasets' for details). Second, datasets can be loaded from a gene expression table file in GCT, TSV, CSV, and XLSX formats (with a support for gzip-archived files). Finally, a set of curated datasets is available directly from the home page.

A number of methods can be used to prepare, normalize, and explore the gene expression table. In particular, it is possible to aggregate microarray probe-level data to gene levels, transform and filter the data, conduct a principal component analysis (PCA), perform k-means or hierarchical clustering, etc. These tools allow you to perform a thorough quality control of the dataset and remove outliers if present.

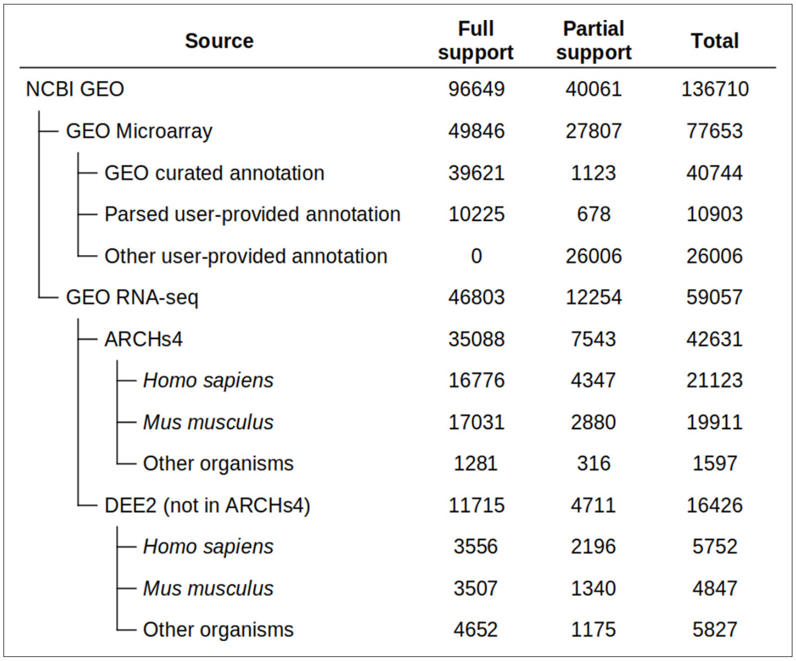
When the dataset is properly filtered and normalized, differential expression analysis using limma (***Ritchie et al., 2015***) or DESeq2 pipelines (***Love et al., 2014***) can be carried out. These results can then be used with other web services for downstream analysis, with shortcuts for pathway analysis with Enrichr (***Kuleshov et al., 2016***) and metabolic network analysis with Shiny GAM (***Sergushichev et al., 2016***). Additionally, gene set enrichment analysis (GSEA) can be done directly in Phantasus as implemented in the fgsea package (***Korotkevich et al., 2021***).

All the plots produced by Phantasus during the data exploration and analysis can be exported as vector images in SVG format. This includes heatmaps, PCA plots, gene profiles, enrichment plots, etc. The obtained images can be used for publications as is or adjusted in a vector graphics editor.

Another option for presenting final or intermediate results is session link sharing. When a link is generated, a snapshot with the current dataset and its representation – annotations, color scheme, sample dendrograms, etc. – are saved on the server. The link can be shared with other users, and, when opened, restores the session.

Stand-alone Phantasus distribution

Aside from using the official mirror (https://alserglab.wustl.edu/phantasus), there is a possibility to set up Phantasus locally. Phantasus can be installed as an R package from Bioconductor (https://



Source	Full support	Partial support	Total
NCBI GEO	96649	40061	136710
├ GEO Microarray	49846	27807	77653
├ GEO curated annotation	39621	1123	40744
├ Parsed user-provided annotation	10225	678	10903
└ Other user-provided annotation	0	26006	26006
└ GEO RNA-seq	46803	12254	59057
├ ARCHs4	35088	7543	42631
├ *Homo sapiens*	16776	4347	21123
├ *Mus musculus*	17031	2880	19911
└ Other organisms	1281	316	1597
└ DEE2 (not in ARCHs4)	11715	4711	16426
├ *Homo sapiens*	3556	2196	5752
├ *Mus musculus*	3507	1340	4847
└ Other organisms	4652	1175	5827

Figure 2. Dataset availability in Phantasus. For fully supported datasets, gene expression data is accompanied by gene annotations in a standardized format. Partial support datasets have either incomplete gene expression matrix or gene annotations.

The online version of this article includes the following source data for figure 2:

Source data 1. Source file for table shown in ***Figure 2***.

bioconductor.org/packages/phantasus) or loaded as a Docker image (https://hub.docker.com/r/alserglab/phantasus). In both cases, almost all of the Phantasus functions will be available from the start.

Some of the Phantasus features require additional server-side setup. Extended support of GEO datasets requires preprocessed expression matrices and platform annotations. Identifier mapping requires organism annotation databases. Pathways enrichment requires pathway databases. The R package and Docker image versions of Phantasus can automatically handle this setup by downloading the required files from https://alserglab.wustl.edu/files/phantasus/minimal-cache. The detailed instructions for installing Phantasus locally are available at https://ctlab.github.io/phantasus-doc/installation.

Importantly, we have introduced a highly scalable data service (HSDS) server (https://alserglab.wustl.edu/hsds/?domain=/counts,), facilitating access to HDF5 files containing precomputed gene count matrices from the official Phantasus mirror. Through this service and a helper R package phantasusLite (https://bioconductor.org/packages/phantasusLite), a stand-alone Phantasus instance selectively loads the count matrix exclusive to the specified dataset, avoiding the need to burden local installations with unnecessarily large files. However, for users prioritizing increased reactivity, the option remains to load these files from the aforementioned cache mirror.

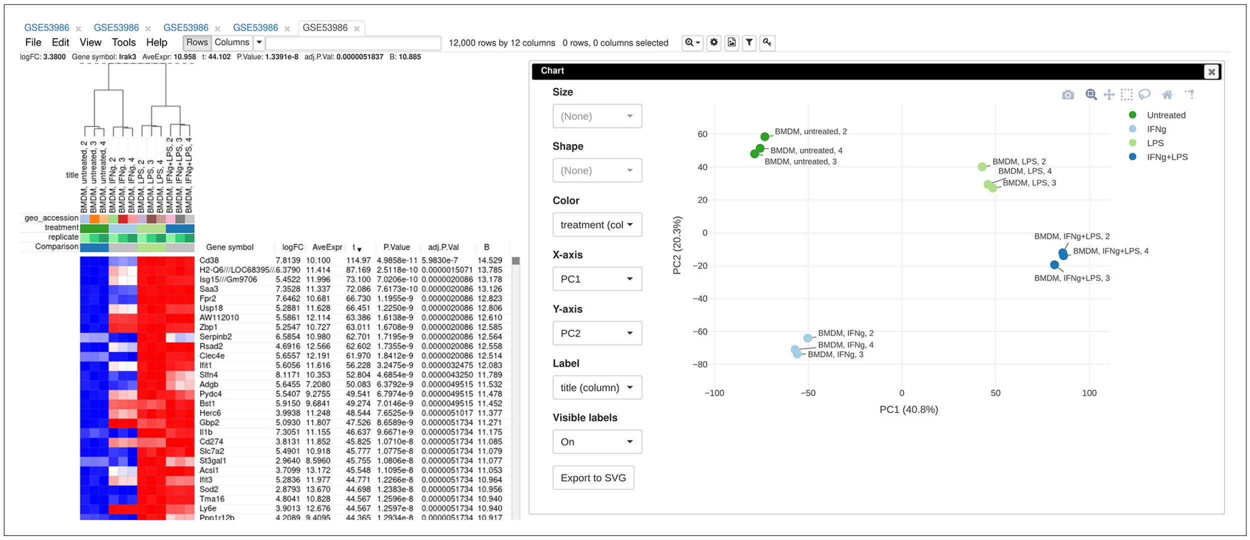
An important feature of the stand-alone version of Phantasus is the ability to share manually curated datasets. Similar to Phantasus session link sharing, one can generate a named session consisting of a dataset and its visual representation. Link to this named session (e.g. https://alserglab.wustl.edu/phantasus/?preloaded=GSE53986.Ctrl_vs_LPS) can then be shared for the other users to view. Such a predictable display of the data can be particularly useful in a publication context.

Available datasets

Phantasus provides streamlined access to more than 96,649 GEO datasets. For these datasets, the expression values and gene identifiers (Entrez, ENSEMBL, or Gene Symbol) are readily available (***Figure 2***). Moreover, these datasets are used to populate the initial Phantasus cache, and thus they have low loading times.

Of these 96,649 datasets, 49,846 are microarrays based on 2767 platforms. For 1347 platforms, GEO databases have machine-readable annotations in the *annot.gz* format with Entrez gene and Gene symbol columns, corresponding to 39,621 datasets. The remaining 10,225 datasets are obtained from platforms that do not have a GEO-provided annotation. For these 1420 platforms, we have automatically marked up user-provided annotations in *SOFT* format to extract gene identifiers and convert the annotations into *annot.gz* format.

The RNA-seq subset of the datasets with streamlined access consists of 46,803 datasets. As GEO does not store expression values for RNA-seq datasets, we rely on other databases for the expression data. The first-priority database for RNA-seq gene counts is ARCHS4 (Human, Mouse, and Zoo



12,000 rows by 12 columns 0 rows, 0 columns selected

logFC: **3.3800** Gene symbol: **Irak3** AveExpr: **10.958** t: **44.102** P.Value: **1.3391e-8** adj.P.Val: **0.0000051837** B: **10.885**

Gene symbol	logFC	AveExpr	t ▾	P.Value	adj.P.Val	B
Cd38	7.8139	10.100	114.97	4.9858e-11	5.9830e-7	14.529
H2-Q6///LOC68395//.	6.3790	11.414	87.169	2.5118e-10	0.0000015071	13.785
Isg15///Gm9706	5.4522	11.996	73.100	7.0206e-10	0.0000020086	13.178
Saa3	7.3528	11.337	72.086	7.6173e-10	0.0000020086	13.126
Fpr2	7.6462	10.681	66.730	1.1955e-9	0.0000020086	12.823
Usp18	5.2881	11.628	66.451	1.2250e-9	0.0000020086	12.806
AW112010	5.5861	12.114	63.386	1.6138e-9	0.0000020086	12.610
Zbp1	5.2547	10.727	63.011	1.6708e-9	0.0000020086	12.585
Serpinb2	6.5854	10.980	62.701	1.7195e-9	0.0000020086	12.564
Rsad2	4.6916	12.566	62.602	1.7355e-9	0.0000020086	12.558
Clec4e	5.6557	12.191	61.970	1.8412e-9	0.0000020086	12.514
Ifit1	5.6056	11.616	56.228	3.2475e-9	0.0000032475	12.083
Slfn4	8.1171	10.353	52.804	4.6854e-9	0.0000043250	11.789
Adgb	5.6455	7.2080	50.083	6.3792e-9	0.0000049515	11.532
Pydc4	5.5407	9.2755	49.541	6.7974e-9	0.0000049515	11.478
Bst1	5.9150	9.6841	49.274	7.0146e-9	0.0000049515	11.452
Herc6	3.9938	11.248	48.544	7.6525e-9	0.0000051017	11.377
Gbp2	5.0930	11.807	47.526	8.6589e-9	0.0000051734	11.271
Il1b	7.3051	11.155	46.637	9.6671e-9	0.0000051734	11.175
Cd274	3.8131	11.852	45.825	1.0710e-8	0.0000051734	11.085
Slc7a2	5.4901	10.918	45.777	1.0775e-8	0.0000051734	11.079
St3gal1	2.9640	8.5960	45.755	1.0806e-8	0.0000051734	11.077
Acsl1	3.7099	13.172	45.548	1.1095e-8	0.0000051734	11.053
Ifit3	5.2836	11.977	44.771	1.2266e-8	0.0000051734	10.964
Sod2	2.8793	13.670	44.698	1.2383e-8	0.0000051734	10.956
Tma16	4.8041	10.828	44.567	1.2596e-8	0.0000051734	10.940
Ly6e	3.9013	12.676	44.567	1.2597e-8	0.0000051734	10.940
Ppp1r12b	4.2089	9.4095	44.365	1.2934e-8	0.0000051734	10.917

Size: (None)

Shape: (None)

Color: treatment (col

X-axis: PC1

Y-axis: PC2

Label: title (column)

Visible labels: On

Figure 3. Example of analyzed dataset GSE53986 with normalized gene expression values, filtered outliers, hierarchically clustered columns, and rows annotated with differential expression analysis between untreated and LPS-treated macrophages.

versions), which covers 35,088 datasets. The other source is the DEE2 database (human, mouse, and other available organisms), which covers an additional 11,715 datasets. The DEE2 database contains run-level quantification, so it has been preprocessed to sum read counts into sample-level tables.

Case study 1: Basic usage

To illustrate the basic usage of Phantasus, we will consider the dataset GSE53986 (***Noubade et al., 2014***) from the GEO database (***Figure 3***). This dataset consists of 16 samples of bone marrow-derived macrophages, untreated and treated with three stimuli: LPS, IFNg, and combined LPS + IFNg. The gene expression was measured with Affymetrix Mouse Genome 430 2.0 Array. Here, we give an overview of the steps; the full walk-through for the analysis is available in Appendix 2.

As the first step of the analysis, the dataset can be loaded and normalized. The dataset is loaded straightforwardly by the *GSE53986* identifier. Because this is a microarray dataset, internally the gene expression values are obtained from GEO. In this particular case, the expression values have not been normalized, but it can be done in Phantasus. From the available normalization options, we select log2 scaling and quantile normalization. Furthermore, we can aggregate microarray probe-level expression values into gene-level expression. We chose the *Maximum Median Probe* method, which retains only a single probe per gene, the one that has the highest median expression value. Finally, we can filter out lowly expressed genes, for example, by keeping only the top 12000 expressed genes.

After the normalization step, we can apply a number of exploratory techniques. In particular, we can do a PCA, k-means gene clustering, and hierarchical clustering of the samples. From these analyses, we can discover that there is an overall good concordance between the replicates of the same treatment, with an exception of the first replicate in each group. We can conclude that these samples are outliers and remove them before the downstream analysis.

Finally, we can do a comparison between the sample groups, for example, by comparing untreated and LPS-treated samples. As the data has been normalized, we can apply *limma* for differential gene expression analysis. The result appears as additional gene annotation columns: p-values, log-fold-changes, and other statistics. Next, we can use differential expression results for a pathway enrichment analysis: for example, we can use R-based GSEA via the *fgsea* package or we can use external tools, such as Enrichr.

Case study 2: Data reanalysis

To highlight Phantasus's ability to reanalyze publicly available data in a context of a biological study, let us consider a study by ***Mowel et al., 2017***. The study considers a genomic locus *Rroid* linked by the authors to homeostasis and function of group 1 innate lymphoid cells (ILC1). The authors hypothesized that *Rroid* locus controls ILC1s by promoting the expression of Id2 gene, a known regulator of ILCs. To confirm this hypothesis, the authors generated an Id2-dependent gene signature based on an existing transcriptomic data (***Shih et al., 2016***) and showed its deregulation in *Rroid*-deficient cells.

The computational analysis described above linking Rroid and Id2 can be replicated in Phantasus in a straightforward way (see Appendix 3 for the detailed walk-through).

First, we can open GEO dataset GSE76466 (***Shih et al., 2016***), containing gene expression data for Id2-deficient NK cells. Notably, GSE76466 is an RNA-sequencing dataset, without gene expression values stored directly in the GEO database; however, Phantasus loads the dataset leveraging precomputed expression values from the ARCHS4 project (***Lachmann et al., 2018***). Then we can perform differential gene expression analysis with the limma tool, comparing Id2-deficient and wild-type (WT) natural killer (NK) cells. Id2-dependent gene signature can be obtained by sorting the genes by *t* column.

Second, RNA-sequencing dataset GSE101458 (***Mowel et al., 2017***), generated by Mowel and colleagues for Rroid-deficient NK cells, can also be opened in Phantasus. There we can do differential gene expression analysis with DESeq2 and remove lowly expressed genes. Finally, we can input the generated Id2-dependent gene signature into Phantasus gene search field and use GSEA plot tool to obtain an enrichment plot, similar to the one presented by Mowel and colleagues, confirming a potential regulation via Id2.

Discussion

Here, we present Phantasus: a web tool designed to simplify the analysis of gene expression data and provide streamlined access to tens of thousands of publicly available datasets. This is an important area of development, as evidenced by the numerous tools addressing this and similar objectives. In Appendix 1, we compare some of these tools, and we anticipate the development of more of them in the future. However, this diversity is highly beneficial to the broad research community as it enhances the overall accessibility of gene expression analysis by allowing researchers to find the most suitable tool for their needs.

Phantasus's uniqueness lies in its highly interactive heatmap-based user interface, integrated with an R and Bioconductor-based back-end. In our experience, this combination effectively supports users from both computational and biological backgrounds. It also provides avenues for extending support to analysis methods available within the R and Bioconductor ecosystem.

A major focus of Phantasus's development was to provide easy access to publicly available datasets from the GEO database. To our knowledge, Phantasus is the only web application for gene expression analysis that offers streamlined access to both microarray and RNA-seq datasets. Currently, it combines approximately 96,000 datasets, with additional datasets available following some manipulations by the user. About half of these datasets are RNA-sequencing datasets, where quantified gene expression values are not stored in GEO. For this purpose, we integrated the ARCHS4 and DEE2 databases and also made them available remotely from the R environment via the phantasusLite package in a GEOquery-compatible manner.

Phantasus is readily available online at https://alserglab.wustl.edu/phantasus, but it can also be installed as a Bioconductor R package at https://bioconductor.org/packages/phantasus. For further convenience, we provide a Docker image at https://hub.docker.com/r/alserglab/phantasus. Phantasus documentation, including comprehensive installation instructions, is available at https://ctlab.github.io/phantasus-doc/. The Phantasus source code is available at https://github.com/ctlab/phantasus under an MIT license (***Sergushichev and Kamenev, 2024***).

Materials and methods

Web application architecture

Phantasus is a web application that combines an interactive graphical user interface with access to a variety of R-based analysis methods (***Figure 1***). The front-end, which is JavaScript-based, derives from the Morpheus web application designed for matrix visualization and analysis (***Gould, 2016***). The back-end is written in R, with an OpenCPU server (***Ooms, 2014***) translating HTTP-queries from the client into R procedure calls.

The JavaScript client is responsible for the matrix visualization, as well as certain analysis methods. In particular, steps like subsetting the dataset, working with annotations, and basic matrix modification (e.g., log-transformation, scaling, etc.) have client-side implementation. Furthermore, the client supports additional visualization methods such as row profile plots, volcano plot, and others.

The analysis methods that require external data or algorithms are implemented in the form of the `phantasus` R package to be carried out on the server side. The operations include differential gene expression analysis, PCA, pathway analysis, and others. Commonly, these methods rely on functions that are already available in the existing R packages; for such methods, only wrapper R functions are implemented.

The OpenCPU server is a core component of the Phantasus back-end. The server provides an HTTP API for calling computational methods implemented in R. For each call, OpenCPU creates a new R environment with the required data, in which the method is then executed. OpenCPU can manage these R environments both in a standard single-user R session and, with the help of rApache, in a multi-user manner inside an Apache web server.

The transfer of large objects between the server and the client exploits a binary Protobuf protocol. The Phantasus back-end uses the protolite R package (***Ooms, 2021***) for object serialization and deserialization. The front-end relies on the protobuf.js module (***Coe, 2020***).

For further performance improvement, Nginx server is used to wrap the OpenCPU server. Nginx server caches the results of the OpenCPU method calls. If the same method with the same data is called again, the cached result can be returned without any additional computations. Furthermore, Nginx is used to serve static content and manage permissions.

Data sources and data gathering

The main data source for Phantasus is the NCBI GEO database (***Barrett et al., 2013***). All of the GEO datasets are identified by a GSEnnnnn accession number (with a subset of the datasets having an additional GDSnnnnn identifier). However, depending on the type of dataset, the processing procedure is different.

The majority of gene expression datasets in the GEO database can be divided into two groups: microarray data and RNA-seq data. While the experiment metadata is available for all of the datasets, the expression matrices are provided only for the microarray datasets. Phantasus relies on the GEOquery package (***Davis and Meltzer, 2007***) to load the experiment metadata (for all datasets) and expression matrices (for microarray datasets) from GEO.

When a GEO RNA-seq dataset is requested by the user, Phantasus refers to precomputed gene counts databases available in the internal storage. In particular, data from ARCHS4 (***Lachmann et al., 2018***) and DEE2 (***Ziemann et al., 2019***) projects are used. Both of these projects contain gene counts and metadata for RNA-seq samples related to different model organisms including but not limited to mouse and human. For any requested RNA-seq dataset, the gene counts are loaded from a single database, whichever covers the highest number of samples.

Next, Phantasus stores gene annotation databases that are used to map genes between different identifier types. These databases are stored in SQLite format compatible with the *AnnotationDbi* R package (***Pagès, 2022***). Currently, only human and mouse databases are available, which are based on `org.Hs.eg.db` and `org.Mm.eg.db` R packages, respectively.

Pathway databases are stored to be used for GSEA. Currently, gene set collections include the GO biological processes database (***Ashburner et al., 2000***), the Reactome database (***Gillespie et al., 2022***), and the MSigDB Hallmark database (***Liberzon et al., 2011***) for human and mouse.

Finally, for faster access, Phantasus dataset cache is automatically populated by a large compendium of datasets. The automatic pipeline is Snakemake-based and consists of four steps.

First, the pipeline converts DEE2 files into ARCHs4-like HDF5 files. During this procedure, the expression values of runs provided by DEE2 are summed up to the sample level. The second step checks for which microarray platforms GEO contains a curated machine-readable annotation in *annot.gz* format. The third step tries to generate the machine-readable annotation for the rest of the microarray platforms from the annotations available in the *SOFT* format. Currently, this step has produced an additional 1300 *annot.gz* files. The last step goes over all of the microarray datasets with a machine-readable annotation and over all of the RNA-seq datasets with the counts available in ARCHS4 or DEE2. For each such dataset, the cached entry with all of the data and metadata is created and stored.

A snapshot of Phantasus internal storage is available at https://alserglab.wustl.edu/files/phantasus/minimal-cache. It contains preprocessed count files, automatically marked-up annotations, and gene and pathways databases. This snapshot can be used for a local Phantasus setup.

Acknowledgements

The project was supported by the Ministry of Science and Higher Education of the Russian Federation (Priority 2030 Federal Academic Leadership Program).

Additional information

Funding

Funder	Grant reference number	Author
Ministry of Science and Higher Education of the Russian Federation	Priority 2030 Federal Academic Leadership Program	Maksim Kleverov Alexey A Sergushichev

Funder	Grant reference number	Author

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Maksim Kleverov, Software, Writing – original draft, Data curation, Writing – review and editing; Daria Zenkova, Vladislav Kamenev, Software, Writing – original draft; Margarita Sablina, Software; Maxim N Artyomov, Conceptualization, Methodology; Alexey A Sergushichev, Conceptualization, Software, Supervision, Funding acquisition, Methodology, Writing – original draft

Author ORCIDs

Maksim Kleverov http://orcid.org/0009-0008-0353-8945
Alexey A Sergushichev http://orcid.org/0000-0003-1159-7220

Decision letter and Author response

Decision letter https://doi.org/10.7554/eLife.85722.sa1
Author response https://doi.org/10.7554/eLife.85722.sa2

Additional files

Supplementary files

- MDAR checklist

Data availability

The current article is a computational study, so no data have been generated. The application source code is available at https://github.com/ctlab/phantasus under an MIT licence. Previously published datasets were taken from the Gene Expression Omnibus; GEO series GSE53986 (***Noubade et al., 2014***); GEO series GSE76466 (***Delconte et al., 2016***); GEO series GSE101458 (***Mowel et al., 2017***).

The following previously published datasets were used:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Noubade R, Wong K, Ota N, Rutz S, Eidenschenk C, Ding J, Valdez PA, Peng I, Sebrell A, Caplazi P, DeVoss J, Soriano RH, Modrusan Z, Hackney JA, Sai T, Ouyang W	2014	NRROS negatively regulates ROS in phagocytes during host defense and autoimmunity	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE53986	NCBI Gene Expression Omnibus, GSE53986
Delconte RB, Shi W, Belz GT, Carotta S, Huntington ND	2016	The helix-loop-helix protein ID2 governs NK cell fate by tuning their sensitivity to interleukin-15	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE76466	NCBI Gene Expression Omnibus, GSE76466
Mowel WK, McCright SJ, Kotzin JJ, Collet M, Uyar A, Chen X, DeLaney A, Spencer SP, Virtue AT, Yang E, Villarino A, Kurachi M, Dunagin MC, Harms Pritchard G, Stein J, Hughes C, Fonseca-Pereira D, Veiga-Fernandes H, Raj A, Kambayashi T, Brodsky IE, O'Shea JJ, Wherry EJ, Goff LA, Rinn JL, Williams A, Flavell RA, Henao-Mejia J	2018	ILC1 lineage identity is determined by a cis-regulatory element marked by a novel lncRNA	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE101459	NCBI Gene Expression Omnibus, GSE101459

References

Alonso R, Salavert F, Garcia-Garcia F, Carbonell-Caballero J, Bleda M, Garcia-Alonso L, Sanchis-Juan A, Perez-Gil D, Marin-Garcia P, Sanchez R, Cubuk C, Hidalgo MR, Amadoz A, Hernansaiz-Ballesteros RD, Alemán A, Tarraga J, Montaner D, Medina I, Dopazo J. 2015. Babelomics 5.0: functional interpretation for new generations of genomic data. *Nucleic Acids Research* **43**:W117–W121. DOI: https://doi.org/10.1093/nar/gkv384, PMID: 25897133

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**:25–29. DOI: https://doi.org/10.1038/75556

Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A. 2013. NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Research* **41**:D991–D995. DOI: https://doi.org/10.1093/nar/gks1193, PMID: 23193258

Byrd JB, Greene AC, Prasad DV, Jiang X, Greene CS. 2020. Responsible, practical genomic data sharing that accelerates research. *Nature Reviews Genetics* **21**:615–629. DOI: https://doi.org/10.1038/s41576-020-0257-5

Chang W. 2022. Shiny: web application framework for R. version 1.7.3. R Package. https://cran.r-project.org/web/packages/shiny/index.html

Coe B. 2020. Protobuf.Js is a pure Javascript implementation with typescript support for node.Js and the Browser. version 6.11.3. JS Npm Package. https://protobufjs.github.io/protobuf.js/

Davis S, Meltzer PS. 2007. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* **23**:1846–1847. DOI: https://doi.org/10.1093/bioinformatics/btm254, PMID: 17496320

Delconte RB, Shi W, Sathe P, Ushiki T, Seillet C, Minnich M, Kolesnik TB, Rankin LC, Mielke LA, Zhang JG, Busslinger M, Smyth MJ, Hutchinson DS, Nutt SL, Nicholson SE, Alexander WS, Corcoran LM, Vivier E, Belz GT, Carotta S, et al. 2016. The helix-loop-helix protein ID2 governs NK cell fate by tuning their sensitivity to Interleukin-15. *Immunity* **44**:103–115. DOI: https://doi.org/10.1016/j.immuni.2015.12.007, PMID: 26795246

Ge SX, Son EW, Yao R. 2018. iDEP: an integrated web application for differential expression and pathway analysis of RNA-Seq data. *BMC Bioinformatics* **19**:534. DOI: https://doi.org/10.1186/s12859-018-2486-6, PMID: 30567491

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, et al. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* **5**:R80. DOI: https://doi.org/10.1186/gb-2004-5-10-r80, PMID: 15461798

Gillespie M, Jassal B, Stephan R, Milacic M, Rothfels K, Senff-Ribeiro A, Griss J, Sevilla C, Matthews L, Gong C, Deng C, Varusai T, Ragueneau E, Haider Y, May B, Shamovsky V, Weiser J, Brunson T, Sanati N, Beckman L, et al. 2022. The reactome pathway knowledgebase 2022. *Nucleic Acids Research* **50**:D687–D692. DOI: https://doi.org/10.1093/nar/gkab1028, PMID: 34788843

Gould J. 2016. Morpheus: Versatile matrix visualization and analysis software. 0.1. Broad Institute. https://software.broadinstitute.org/morpheus/index.html

Hait TA, Maron-Katz A, Sagir D, Amar D, Ulitsky I, Linhart C, Tanay A, Sharan R, Shiloh Y, Elkon R, Shamir R. 2019. The EXPANDER integrated platform for transcriptome analysis. *Journal of Molecular Biology* **431**:2398–2406. DOI: https://doi.org/10.1016/j.jmb.2019.05.013, PMID: 31100387

Hunt GP, Grassi L, Henkin R, Smeraldi F, Spargo TP, Kabiljo R, Koks S, Ibrahim Z, Dobson RJB, Al-Chalabi A, Barnes MR, Iacoangeli A. 2022. GEOexplorer: A webserver for gene expression analysis and visualisation. *Nucleic Acids Research* **50**:W367–W374. DOI: https://doi.org/10.1093/nar/gkac364, PMID: 35609980

Korotkevich G, Sukhov V, Budin N, Shpak B, Artyomov MN, Sergushichev A. 2021. Fast Gene Set Enrichment Analysis. *bioRxiv*. DOI: https://doi.org/10.1101/060012

Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, McDermott MG, Monteiro CD, Gundersen GW, Ma'ayan A. 2016. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research* **44**:W90–W97. DOI: https://doi.org/10.1093/nar/gkw377

Lachmann A, Torre D, Keenan AB, Jagodnik KM, Lee HJ, Wang L, Silverstein MC, Ma'ayan A. 2018. Massive mining of publicly available RNA-seq data from human and mouse. *Nature Communications* **9**:1366. DOI: https://doi.org/10.1038/s41467-018-03751-6, PMID: 29636450

La Ferlita A, Alaimo S, Di Bella S, Martorana E, Laliotis GI, Bertoni F, Cascione L, Tsichlis PN, Ferro A, Bosotti R, Pulvirenti A. 2021. RNAdetector: a free user-friendly stand-alone and cloud-based system for RNA-Seq data analysis. *BMC Bioinformatics* **22**:298. DOI: https://doi.org/10.1186/s12859-021-04211-7, PMID: 34082707

Law CW, Chen Y, Shi W, Smyth GK. 2014. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology* **15**:R29. DOI: https://doi.org/10.1186/gb-2014-15-2-r29, PMID: 24485249

Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. 2011. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**:1739–1740. DOI: https://doi.org/10.1093/bioinformatics/btr260, PMID: 21546393

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**:550. DOI: https://doi.org/10.1186/s13059-014-0550-8, PMID: 25516281

Mahi NA, Najafabadi MF, Pilarczyk M, Kouril M, Medvedovic M. 2019. GREIN: An interactive web platform for re-analyzing GEO RNA-seq data. *Scientific Reports* **9**:7580. DOI: https://doi.org/10.1038/s41598-019-43935-8, PMID: 31110304

Mowel WK, McCright SJ, Kotzin JJ, Collet MA, Uyar A, Chen X, DeLaney A, Spencer SP, Virtue AT, Yang E, Villarino A, Kurachi M, Dunagin MC, Pritchard GH, Stein J, Hughes C, Fonseca-Pereira D, Veiga-Fernandes H, Raj A, Kambayashi T, et al. 2017. Group 1 innate lymphoid cell lineage identity is determined by a cis-regulatory element marked by a long non-coding RNA. *Immunity* **47**:435–449.. DOI: https://doi.org/10.1016/j.immuni.2017.08.012, PMID: 28930659

Nelson JW, Sklenar J, Barnes AP, Minnier J. 2017. The START App: a web-based RNAseq analysis and visualization resource. *Bioinformatics* **33**:447–449. DOI: https://doi.org/10.1093/bioinformatics/btw624, PMID: 28171615

Noubade R, Wong K, Ota N, Rutz S, Eidenschenk C, Valdez PA, Ding J, Peng I, Sebrell A, Caplazi P, DeVoss J, Soriano RH, Sai T, Lu R, Modrusan Z, Hackney J, Ouyang W. 2014. NRROS negatively regulates reactive oxygen species during host defence and autoimmunity. *Nature* **509**:235–239. DOI: https://doi.org/10.1038/nature13152, PMID: 24739962

Ooms J. 2014. The Opencpu System: Towards a Universal Interface for Scientific Computing through Separation of Concerns. *arXiv*. https://arxiv.org/abs/1406.4806

Ooms J. 2021. Protolite: highly Optimized protocol buffer Serializers. version 2.1.1. R Package. https://CRAN.R-project.org/package=protolite

Pagès H. 2022. Annotationdbi: manipulation of Sqlite-based annotations in Bioconductor. Version 1.58.0. R Package. https://bioconductor.org/packages/AnnotationDbi

Powell DR. 2019. Degust: interactive RNA-Seq analysis. Zenodo. https://doi.org/10.5281/zenodo.3258932

Prieto C, Barrios D. 2019. RaNA-Seq: Interactive RNA-Seq analysis from FASTQ files to functional analysis. *Bioinformatics* **1**:btz854. DOI: https://doi.org/10.1093/bioinformatics/btz854, PMID: 31730197

Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. 2006. GenePattern 2.0. *Nature Genetics* **38**:500–501. DOI: https://doi.org/10.1038/ng0506-500

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43**:e47. DOI: https://doi.org/10.1093/nar/gkv007, PMID: 25605792

Santamaría R, Therón R, Quintales L. 2014. BicOverlapper 2.0: visual analysis for gene expression. *Bioinformatics* **30**:1785–1786. DOI: https://doi.org/10.1093/bioinformatics/btu120, PMID: 24590442

Sergushichev AA, Loboda AA, Jha AK, Vincent EE, Driggers EM, Jones RG, Pearce EJ, Artyomov MN. 2016. GAM: a web-service for integrated transcriptional and metabolic network analysis. *Nucleic Acids Research* **44**:W194–W200. DOI: https://doi.org/10.1093/nar/gkw266, PMID: 27098040

Sergushichev A, Kamenev V. 2024. Phantasus. GitHub. https://github.com/ctlab/phantasus

Shih H-Y, Sciumè G, Mikami Y, Guo L, Sun H-W, Brooks SR, Urban JF Jr, Davis FP, Kanno Y, O'Shea JJ. 2016. Developmental acquisition of regulomes underlies innate lymphoid cell functionality. *Cell* **165**:1120–1133. DOI: https://doi.org/10.1016/j.cell.2016.04.029, PMID: 27156451

Stark R, Grzelak M, Hadfield J. 2019. RNA sequencing: the teenage years. *Nature Reviews Genetics* **20**:631–656. DOI: https://doi.org/10.1038/s41576-019-0150-2

Torre D, Lachmann A, Ma'ayan A. 2018. BioJupies: automated generation of interactive notebooks for RNA-Seq Data Analysis in the Cloud. *Cell Systems* **7**:556–561.. DOI: https://doi.org/10.1016/j.cels.2018.10.007, PMID: 30447998

Wang Z, Monteiro CD, Jagodnik KM, Fernandez NF, Gundersen GW, Rouillard AD, Jenkins SL, Feldmann AS, Hu KS, McDermott MG, Duan Q, Clark NR, Jones MR, Kou Y, Goff T, Woodland H, Amaral FMR, Szeto GL, Fuchs O, Schüssler-Fiorenza Rose SM, et al. 2016. Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd. *Nature Communications* **7**:12846. DOI: https://doi.org/10.1038/ncomms12846, PMID: 27667448

Ziemann M, Kaspi A, El-Osta A. 2019. Digital expression explorer 2: a repository of uniformly processed RNA sequencing data. *GigaScience* **8**:giz022. DOI: https://doi.org/10.1093/gigascience/giz022, PMID: 30942868

Appendix 1

In this section, we consider existing software platforms for gene expression analysis. While they share common principles, the details of implementation vary, which affects their usability. For the comparison, we considered three aspects: (1) support for gene expression analysis steps, (2) data availability, and (3) user experience.

The general gene expression analysis workflow can be roughly divided into three stages: data preprocessing, exploratory analysis, and differential gene expression with pathway analysis. ***Appendix 1—table 1*** contains an overview of support of the key analysis steps implemented in the considered tools. At the data preprocessing stage, many tools allow us to load unnormalized gene expression matrices and provide a variety of matrix normalization approaches such as log-transformation, quantile normalization, scaling, and others. After the dataset is normalized, some of these tools provide instruments for dataset exploration like PCA and row/column clustering. Gene expression analysis usually ends by determining differentially expressed genes and pathway analysis to explain the observed difference. While the differential expression is available almost everywhere, the pathway analysis can be carried out only by about a half of the tools. In this half, the most notable tools are Gene Pattern, Babelomics, iDep, EXPANDER, and GEOexplorer, which support all the considered steps.

Appendix 1—table 1. Gene expression analysis key steps.

Application	Normalization	Principal component analysis	Clustering	Differential expression	Pathway analysis
Gene pattern ***Reich et al., 2006***	+	+	+	+	+
GEO2R ***Barrett et al., 2013***	-	-	-	+	-
BicOverlapper2 ***Santamaría et al., 2014***	-	-	-	+	+
Babelomics ***Alonso et al., 2015***	+	+	+	+	+
Morpheus ***Gould, 2016***	+	-	+	-	-
START ***Nelson et al., 2017***	+	+	-	+	-
BioJupies ***Torre et al., 2018***	+	+	+	+	+
iDEP ***Ge et al., 2018***	+	+	+	+	+
Degust ***Powell, 2019***	-	+	-	+	-
GREIN ***Mahi et al., 2019***	+	+	-	+	-
EXPANDER ***Hait et al., 2019***	+	+	+	+	+
RaNa-seq ***Prieto and Barrios, 2019***	-	+	-	+	+
RNAdetector ***La Ferlita et al., 2021***	-	-	-	+	-
GEOexplorer ***Hunt et al., 2022***	+	+	+	+	+
Phantasus	+	+	+	+	+

The huge difference between tools appears in the supported data sources and the availability of datasets (***Appendix 1—table 2***). In addition to a simple loading of gene expression matrices in tabular formats, which is a common practice, few tools allow us to load publicly available datasets from external sources, in particular from GEO database. However, depending on the type of the GEO dataset, microarray or RNA-seq, their support varies among the considered tools. The GEO database does not store gene expression matrices for RNA-seq datasets, thus tools like GREIN and RaNa-seq incorporate alignment and quantification of the raw data into their pipeline. Alternatively, precomputed count matrices from other projects, such as ARHCs4 (***Lachmann et al., 2018***) and DEE2 (***Ziemann et al., 2019***), can be used, as implemented in BioJupies and iDep.

Appendix 1—table 2 Continued on next page

Appendix 1—table 2. Gene expression sources.

Application	User-provided data	GEO microarray	GEO RNA-seq
Gene pattern ***Reich et al., 2006***	+	+	-
GEO2R ***Barrett et al., 2013***	-	+	-
BicOverlapper2 ***Santamaría et al., 2014***	+	+	-
Babelomics ***Alonso et al., 2015***	+	-	-
Morpheus ***Gould, 2016***	+	-	-
START ***Nelson et al., 2017***	+	-	-
BioJupies ***Torre et al., 2018***	+	-	+/- *
iDEP ***Ge et al., 2018***	+	-	+/-*
Degust ***Powell, 2019***	+	-	-
EXPANDER ***Hait et al., 2019***	+	-	-
GREIN ***Mahi et al., 2019***	-	-	+
RaNa-seq ***Prieto and Barrios, 2019***	+	-	+
RNAdetector ***La Ferlita et al., 2021***	+	-	-
GEOexplorer ***Hunt et al., 2022***	-	+	-
Phantasus	+	+	+/-*

*As processed in ARCHs4 and/or Dee2 projects.

Finally, we consider features that influence user experience (***Appendix 1—table 3***). There are three major types of architectures used in gene expression analysis applications, each with different trade-offs in reactivity, user setup, and development ease: Shiny-based web applications, non-Shiny web applications, and stand-alone applications.

Some non-Shiny web applications, such as GenePattern, BioJupies, and RaNa-seq, allow users to share sessions, enhancing reproducibility and facilitating collaborative work or publication. While many tools offer interactive gene expression heatmaps and other plots for data visualization and exploration, the support for working with metadata is often limited. Only four other tools (Babelomics, Morpheus, EXPANDER, RNAdetector) allow the editing of genes and sample annotations, and among them, only Babelomics and iDEP can map gene annotations using external gene annotation databases.

Appendix 1—table 3. User experience features.

Application	Architecture	Saved sessions	Interactive heatmap	Interactive plots	Editing sample annotations	Editing gene annotations
Gene pattern ***Reich et al., 2006***	Web non-Shiny	+	-	-	-	-
GEO2R ***Barrett et al., 2013***	Web non-Shiny	-	-	-	-	-
BicOverlapper2 ***Santamaría et al., 2014***	Local installation	-	+/-*	-	-	-
Babelomics ***Alonso et al., 2015***	Web non-Shiny	-	-	+	+	+
Morpheus ***Gould, 2016***	Web non-Shiny	+/-	+	+	+	-
START ***Nelson et al., 2017***	Web Shiny	-	+/-*	+	-	-
BioJupies ***Torre et al., 2018***	Web non-Shiny	+	+/-*	+	-	-
iDEP ***Ge et al., 2018***	Web Shiny	-	+/-*	+	-	+
Degust ***Powell, 2019***	Web non-Shiny	-	+	+	-	-
GREIN ***Mahi et al., 2019***	Web Shiny	-	+	+	-	-

Appendix 1—table 3 Continued on next page

Appendix 1—table 3 Continued

Application	Architecture	Saved sessions	Interactive heatmap	Interactive plots	Editing sample annotations	Editing gene annotations
EXPANDER ***Hait et al., 2019***	Local installation	-	+	+	+	-
RaNa-seq ***Prieto and Barrios, 2019***	Web non-Shiny	+	+/-*	+	-	-
RNAdetector ***La Ferlita et al., 2021***	Local installation	-	-	-	+	-
GEOexplorer ***Hunt et al., 2022***	Web Shiny	-	+	+	-	-
Phantasus	Web non-Shiny	+	+	+	+	+

*Limited number of genes.

Phantasus is a Shiny-independent web application capable of executing all key steps in gene expression analysis. It grants access to an extensive collection of publicly available datasets, encompassing both microarray and RNA-seq data. Through an interactive heatmap-based interface, users can explore data, modify annotations, create plots, etc. Furthermore, Phantasus delivers results in a publication-ready format, allowing users to export images to vector-based graphics formats and share session links. These links retain all user-defined settings, including colors, annotations, and filter options.

Appendix 2

Introduction

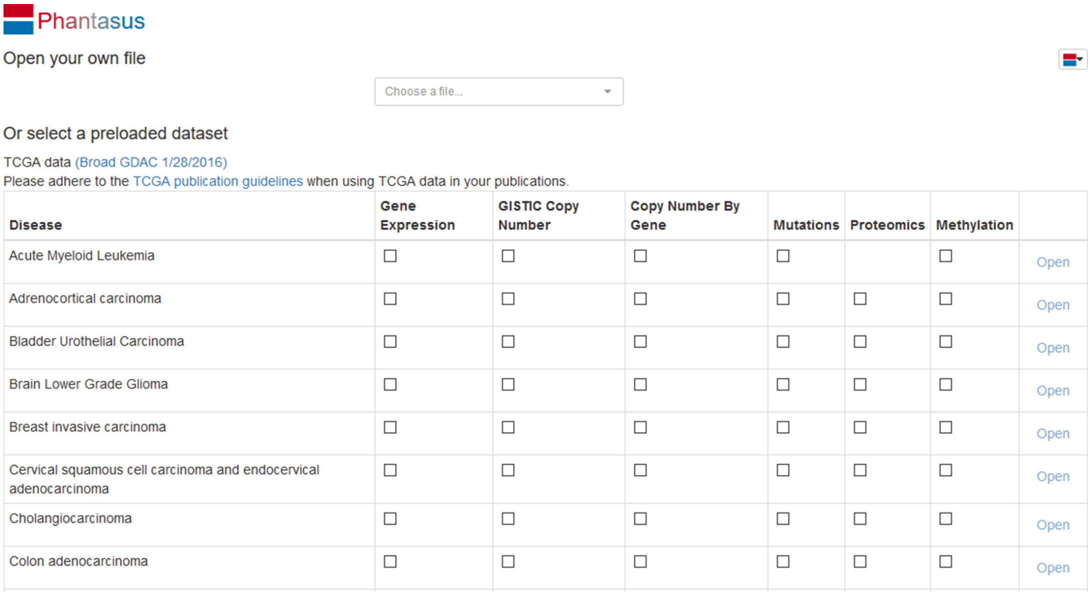
In this section, we show an example usage of Phantasus for the analysis of public gene expression data from the GEO database. It starts from loading data, normalization and filtering outliers, to doing differential gene expression analysis and downstream analysis.

To illustrate the usage of Phantasus, let us consider the public dataset from the GEO database GSE53986. This dataset contains data from experiments, where bone marrow-derived macrophages were treated with three stimuli: LPS, IFNg, and combined LPS + IFNg.

Opening Phantasus

The simplest way to try Phantasus application is to go to the website https://alserglab.wustl.edu/phantasus where the latest versions are deployed. Alternatively, Phantasus can be installed locally (see section 'Stand-alone Phantasus distribution').

When Phantasus opens, the starting screen (***Appendix 2—figure 1***) should appear.



Open your own file

Or select a preloaded dataset

TCGA data (Broad GDAC 1/28/2016)
Please adhere to the TCGA publication guidelines when using TCGA data in your publications.

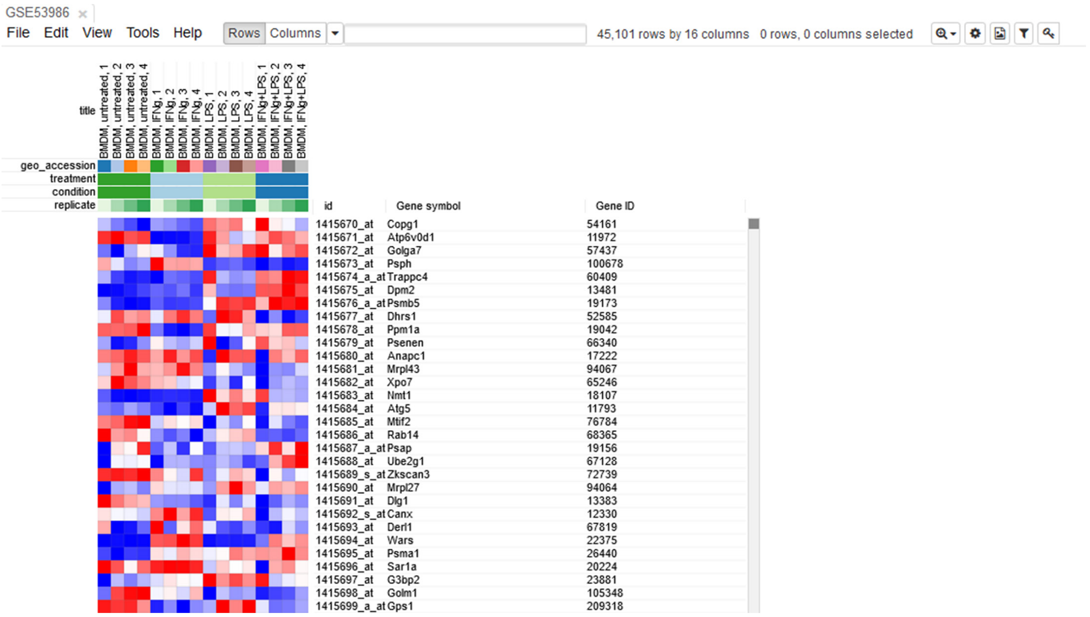
Disease	Gene Expression	GISTIC Copy Number	Copy Number By Gene	Mutations	Proteomics	Methylation	
Acute Myeloid Leukemia	□	□	□	□		□	Open
Adrenocortical carcinoma	□	□	□	□	□	□	Open
Bladder Urothelial Carcinoma	□	□	□	□	□	□	Open
Brain Lower Grade Glioma	□	□	□	□	□	□	Open
Breast invasive carcinoma	□	□	□	□	□	□	Open
Cervical squamous cell carcinoma and endocervical adenocarcinoma	□	□	□	□	□	□	Open
Cholangiocarcinoma	□	□	□	□	□	□	Open
Colon adenocarcinoma	□	□	□	□	□	□	Open

Appendix 2—figure 1. Phantasus starting screen.

Preparing the dataset for analysis

Opening the dataset

Let us open the dataset. To do this, select the *GEO Datasets* option in *Choose a file …* dropdown menu. There, a text field will appear where GSE53986 should be entered. Clicking the *Load* button (or pressing *Enter* on the keyboard) will start the loading. After a few seconds, the corresponding heatmap (***Appendix 2—figure 2***) should appear.



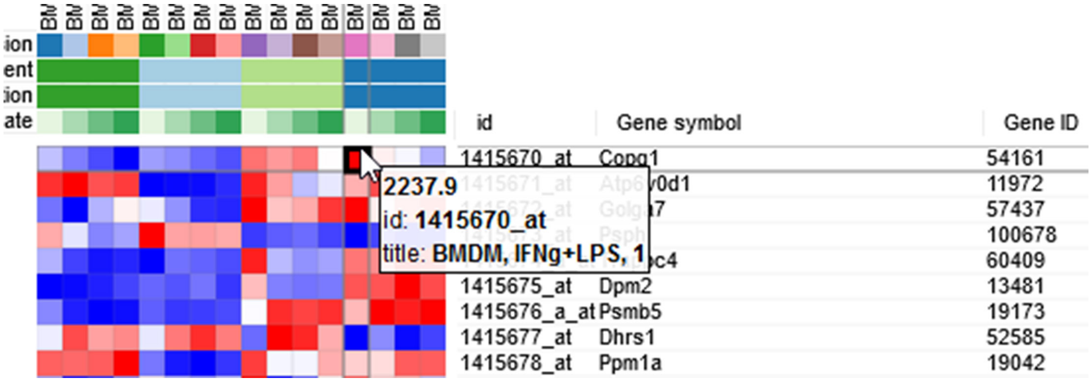
id	Gene symbol	Gene ID
1415670_at	Copg1	54161
1415671_at	Atp6v0d1	11972
1415672_at	Golga7	57437
1415673_at	Psph	100678
1415674_a_at	Trappc4	60409
1415675_at	Dpm2	13481
1415676_a_at	Psmb5	19173
1415677_at	Dhrs1	52585
1415678_at	Ppm1a	19042
1415679_at	Psenen	66340
1415680_at	Anapc1	17222
1415681_at	Mrpl43	94067
1415682_at	Xpo7	65246
1415683_at	Nmt1	18107
1415684_at	Atg5	11793
1415685_at	Mtif2	76784
1415686_at	Rab14	68365
1415687_a_at	Psap	19156
1415688_at	Ube2g1	67128
1415689_s_at	Zkscan3	72739
1415690_at	Mrpl27	94064
1415691_at	Dlg1	13383
1415692_s_at	Canx	12330
1415693_at	Derl1	67819
1415694_at	Wars	22375
1415695_at	Psma1	26440
1415696_at	Sar1a	20224
1415697_at	G3bp2	23881
1415698_at	Golm1	105348
1415699_a_at	Gps1	209318

Appendix 2—figure 2. Heatmap of the loaded dataset.

On the heatmap, the rows correspond to genes (or microarray probes). The rows are annotated with *Gene symbol* and *Gene ID* annotations (as loaded from the GEO database). Columns correspond to samples. They are annotated with titles, GEO sample accession identifiers, and treatment field. The annotations, such as treatment, are loaded from user-submitted GEO annotations (they can be seen, for example, in *Characteristics* section at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1304836). We note that not for all of the datasets in GEO such proper annotations are supplied.

Adjusting expression values

By hovering at a heatmap cell, gene expression values can be viewed. The large values there (***Appendix 2—figure 3***) indicate that the data is not log-scaled, which is important for most types of gene expression analysis.

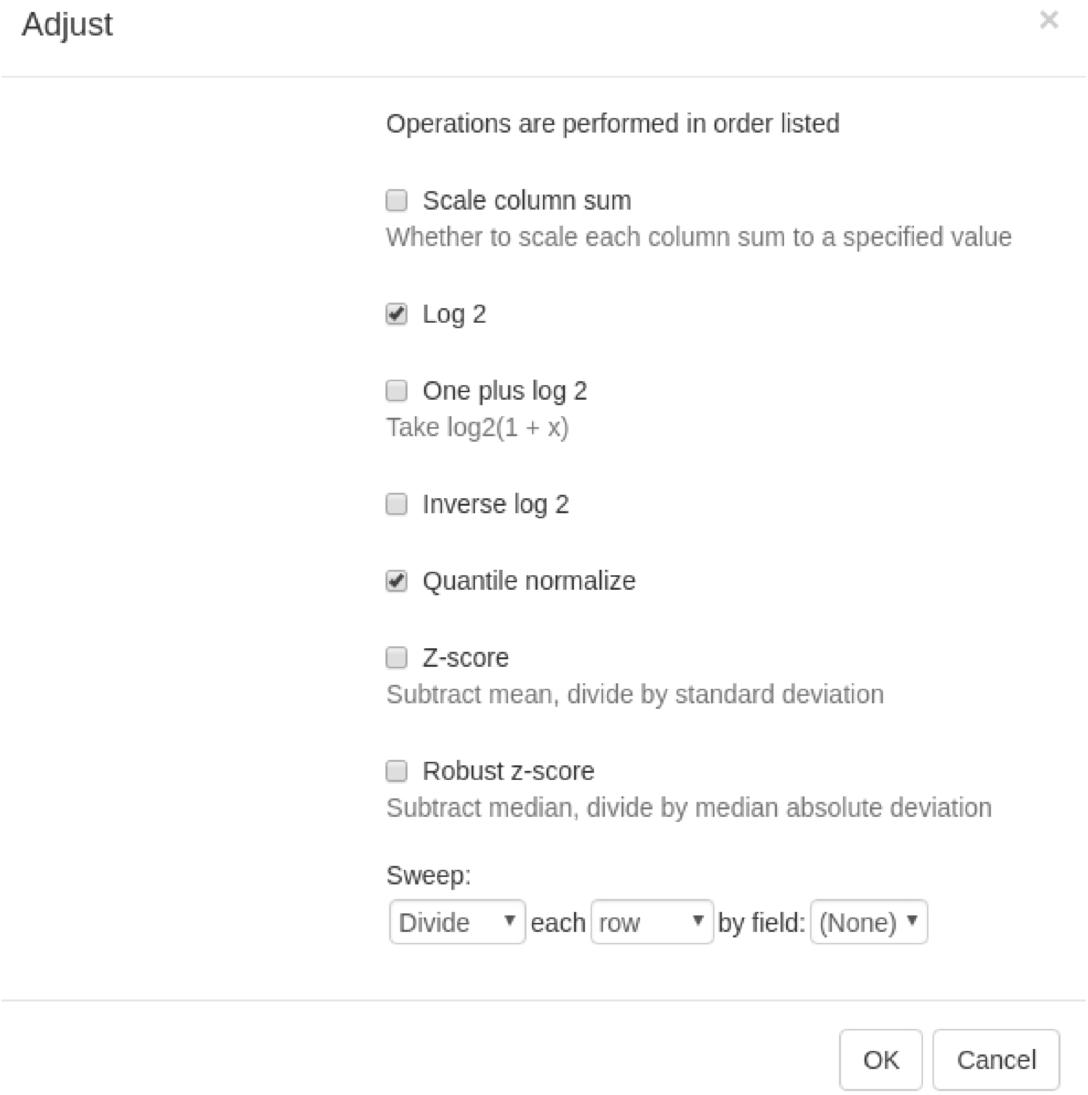


2237.9
id: 1415670_at
title: BMDM, IFNg+LPS, 1

id	Gene symbol	Gene ID
1415670_at	Copg1	54161
1415671_at	Atp6v0d1	11972
1415672_at	Golga7	57437
1415673_at	Psph	100678
	c4	60409
1415675_at	Dpm2	13481
1415676_a_at	Psmb5	19173
1415677_at	Dhrs1	52585
1415678_at	Ppm1a	19042

Appendix 2—figure 3. Heatmap cell tooltip.

For proper further analysis, it is recommended to normalize the matrix. To normalize values, go to *Tools/Adjust* menu and check *Log 2* and *Quantile normalize* adjustments (***Appendix 2—figure 4***).



Adjust

Operations are performed in order listed

□ Scale column sum
Whether to scale each column sum to a specified value

☑ Log 2

□ One plus log 2
Take log2(1 + x)

□ Inverse log 2

☑ Quantile normalize

□ Z-score
Subtract mean, divide by standard deviation

□ Robust z-score
Subtract median, divide by median absolute deviation

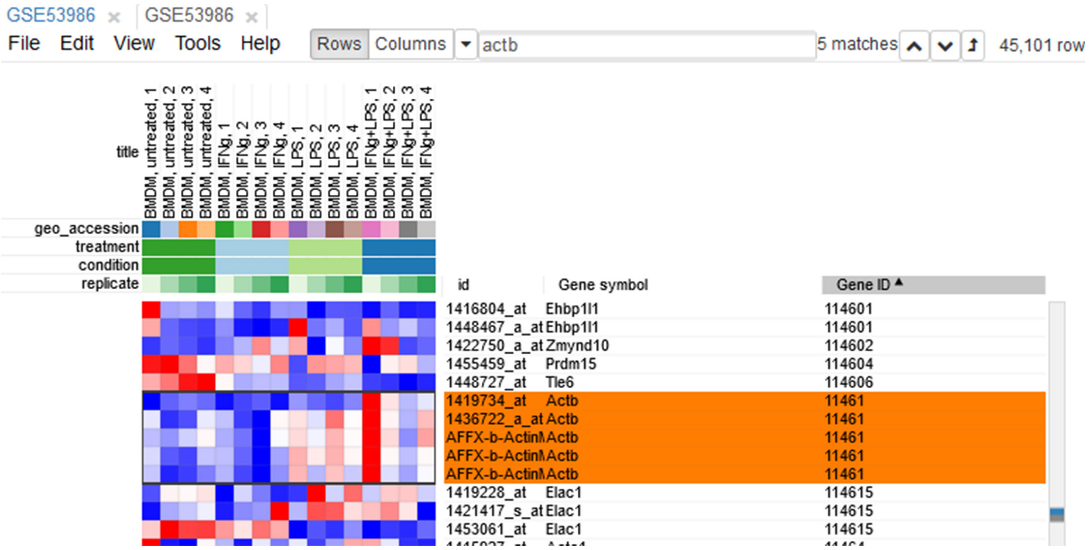
Sweep:
Divide ▾ each row ▾ by field: (None) ▾

Appendix 2—figure 4. Adjust tool.

A new tab with adjusted values will appear. All operations that modify gene expression matrix (such as adjustment, subsetting, and several others) create a new tab. This allows the user to revert the operation by going back to one of the previous tabs.

Removing duplicate genes

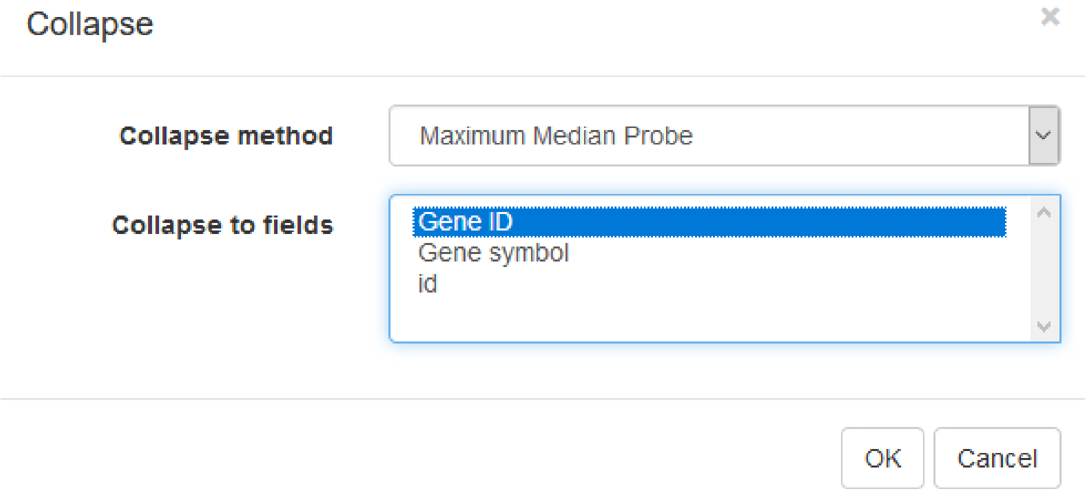
Since the dataset is obtained with a microarray, a single gene can be represented by several probes. This can be seen, for example, by sorting rows by *Gene symbol* column (one click on column header), entering Actb in the search field and going to the first match by clicking the down-arrow next to the field. There are five probes corresponding to Actb gene in the considered microarray (***Appendix 2—figure 5***).



id	Gene symbol	Gene ID ▲
1416804_at	Ehbp1l1	114601
1448467_a_at	Ehbp1l1	114601
1422750_a_at	Zmynd10	114602
1455459_at	Prdm15	114604
1448727_at	Tle6	114606
1419734_at	Actb	11461
1436722_a_at	Actb	11461
AFFX-b-ActinM	Actb	11461
AFFX-b-ActinM	Actb	11461
AFFX-b-ActinM	Actb	11461
1419228_at	Elac1	114615
1421417_s_at	Elac1	114615
1453061_at	Elac1	114615

Appendix 2—figure 5. Actb duplicates.

To simplify the analysis, it is better to have one row per gene in the gene expression matrix. One of the easiest ways is to choose only one row that has the maximal median level of expression across all samples. Such a method removes the noise of lowly expressed probes. Go to *Tools/Collapse* and choose *Maximum Median Probe* as the method and *Gene ID* as the collapse field (***Appendix 2—figure 6***).



Collapse

Collapse method: Maximum Median Probe

Collapse to fields: Gene ID (selected), Gene symbol, id

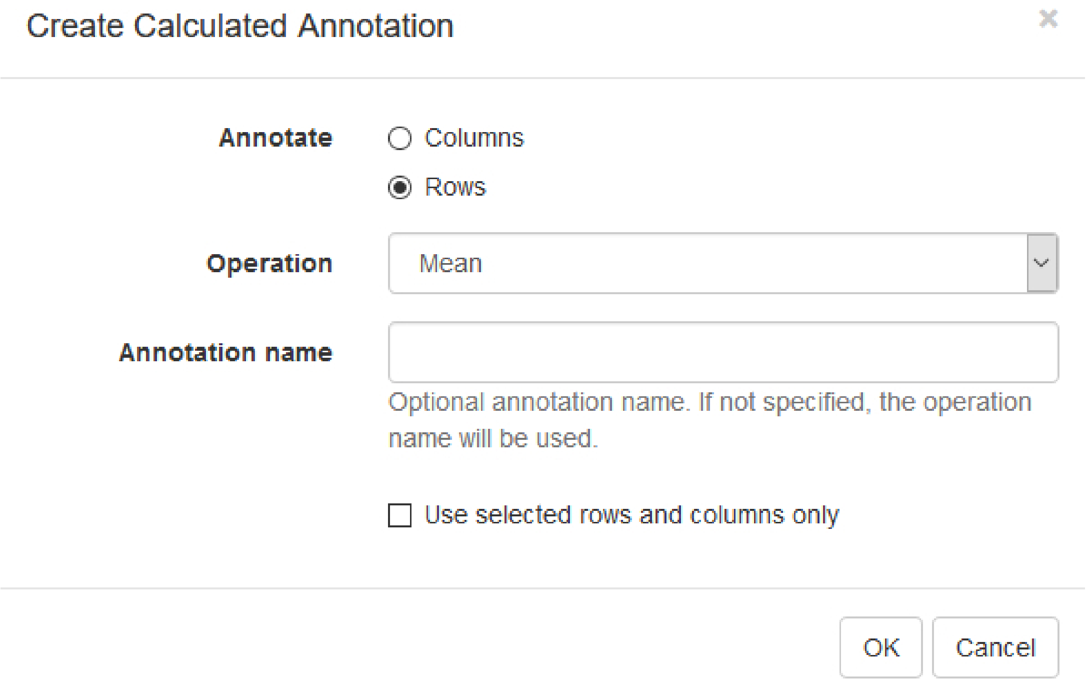
Appendix 2—figure 6. Collapse tool settings.

The result will be shown in a new tab.

Filtering lowly expressed genes

Additionally, genes expressed at low levels can be explicitly filtered. It helps reduce noise and increase the power of downstream analysis methods.

First, we calculate mean expression of each gene using the *Tools/Create Calculated Annotation* menu. Select Mean operation, optionally enter a name for the resulting column (***Appendix 2—figure 7***), and click *OK*. The result will appear as an additional column in row annotations (***Appendix 2—figure 8***).



Create Calculated Annotation

Annotate: ○ Columns ● Rows

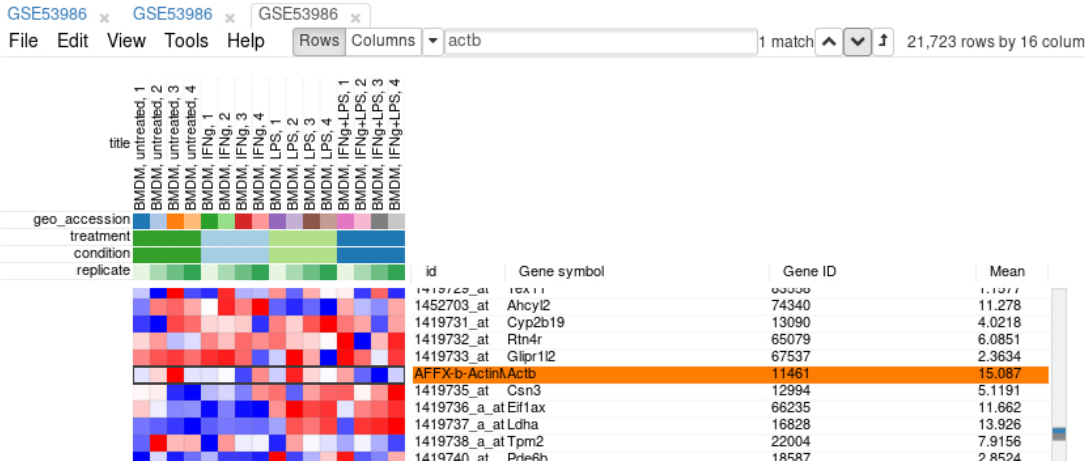
Operation: Mean

Annotation name:

Optional annotation name. If not specified, the operation name will be used.

□ Use selected rows and columns only

Appendix 2—figure 7. Row mean calculated annotation settings.

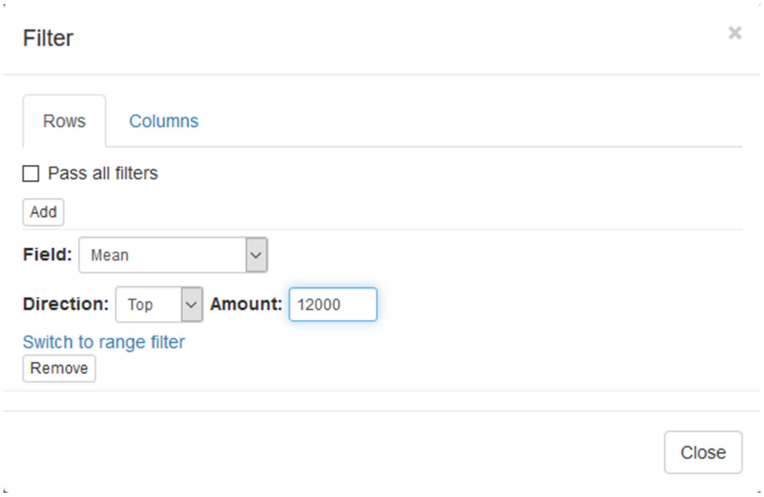


Search: actb

id	Gene symbol	Gene ID	Mean
1452703_at	Ahcyl2	74340	11.278
1419731_at	Cyp2b19	13090	4.0218
1419732_at	Rtn4r	65079	6.0851
1419733_at	Glipr1l2	67537	2.3634
AFFX-b-ActinM	Actb	11461	15.087
1419735_at	Csn3	12994	5.1191
1419736_a_at	Eif1ax	66235	11.662
1419737_a_at	Ldha	16828	13.926
1419738_a_at	Tpm2	22004	7.9156

Appendix 2—figure 8. Heatmap with loaded row mean annotation.

Now this annotation can be used to filter genes. Open *Tools/Filter* menu. Click *Add* to add a new filter. Choose mean_expression as a *Field* for filtering. Then press *Switch to top filter* button and input the number of genes to keep. A good choice for a typical mammalian dataset is to keep around 10–12, 000 most expressed genes (***Appendix 2—figure 9***). Filter is applied automatically, so after closing the dialog with *Close* button only the genes passing the filter will be displayed (***Appendix 2—figure 10***).

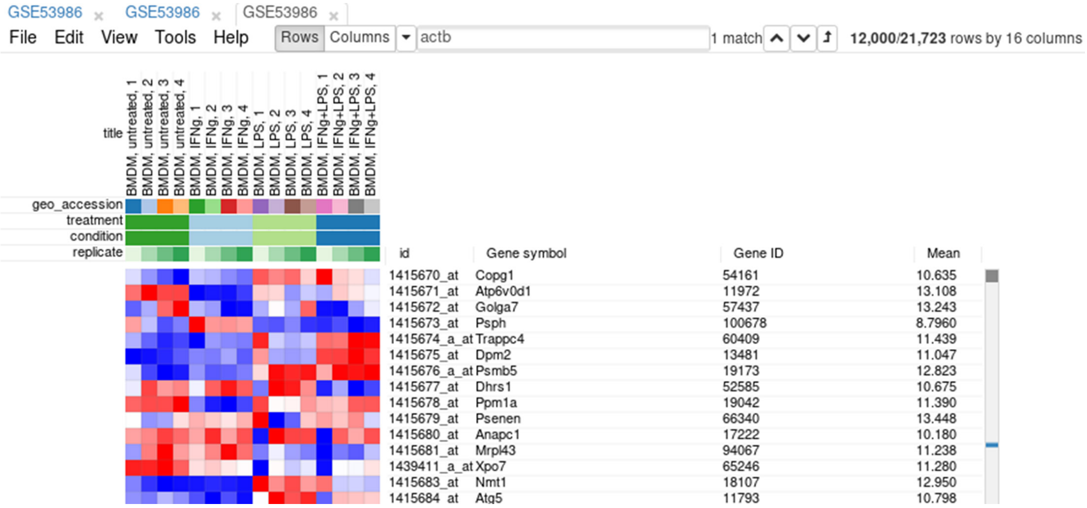


□ Pass all filters

Field: Mean

Direction: Top **Amount:** 12000

Appendix 2—figure 9. Filter tool settings.



actb 1 match **12,000/21,723** rows by 16 columns

id	Gene symbol	Gene ID	Mean
1415670_at	Copg1	54161	10.635
1415671_at	Atp6v0d1	11972	13.108
1415672_at	Golga7	57437	13.243
1415673_at	Psph	100678	8.7960
1415674_a_at	Trappc4	60409	11.439
1415675_at	Dpm2	13481	11.047
1415676_a_at	Psmb5	19173	12.823
1415677_at	Dhrs1	52585	10.675
1415678_at	Ppm1a	19042	11.390
1415679_at	Psenen	66340	13.448
1415680_at	Anapc1	17222	10.180
1415681_at	Mrpl43	94067	11.238
1439411_a_at	Xpo7	65246	11.280
1415683_at	Nmt1	18107	12.950
1415684_at	Atg5	11793	10.798

Appendix 2—figure 10. Heatmap with filtered genes.

Extracting these genes into a new tab is more convenient. For this, select all genes (click on any gene and press *Ctrl + A*) and use the *Tools/New Heat Map* menu (or press *Ctrl + X*).

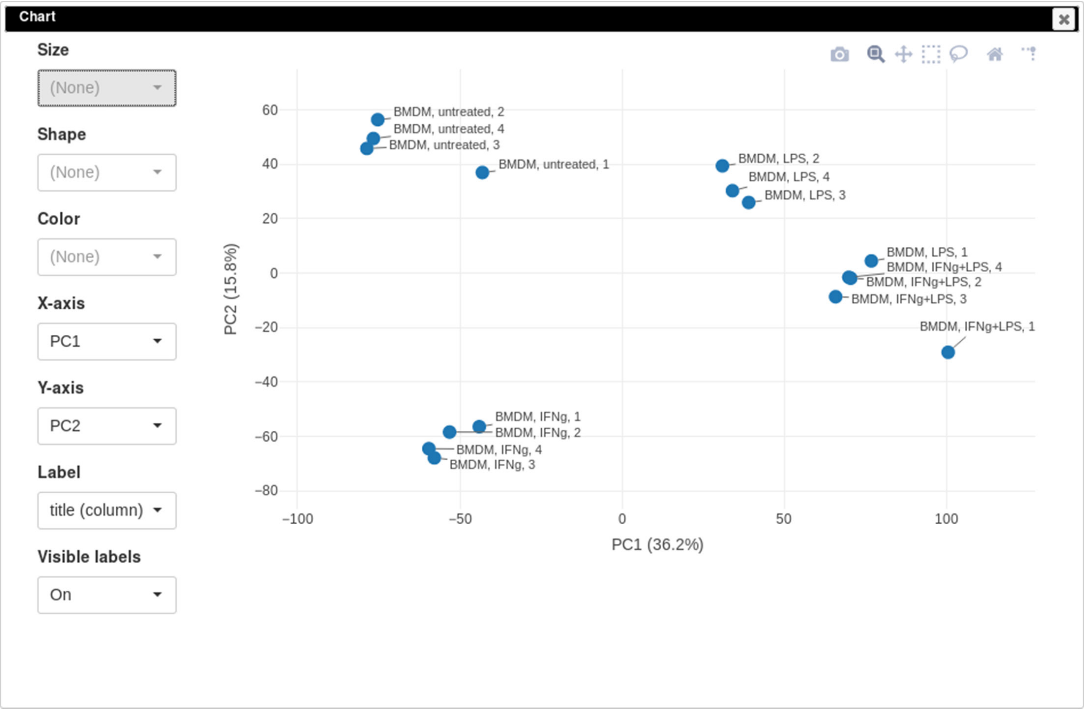
Now you have the tab with a fully prepared dataset for further analysis. To easily distinguish it from other tabs, you can rename it by right clicking on the tab and choosing the *Rename* option. Let us rename it to `GSE53986_norm`.

It is also useful to save the current result to be able to return to it later. In order to save it, use the *File/Save Dataset* menu. Enter an appropriate file name (e.g., `GSE53986_norm`) and press OK. A file in text GCT format will be downloaded.

Exploring the dataset

PCA plot

One of the ways to assess the quality of the dataset is to use PCA method. This can be done using the *Tools/Plots/PCA Plot* menu (***Appendix 2—figure 11***).



Chart

Size: (None)

Shape: (None)

Color: (None)

X-axis: PC1

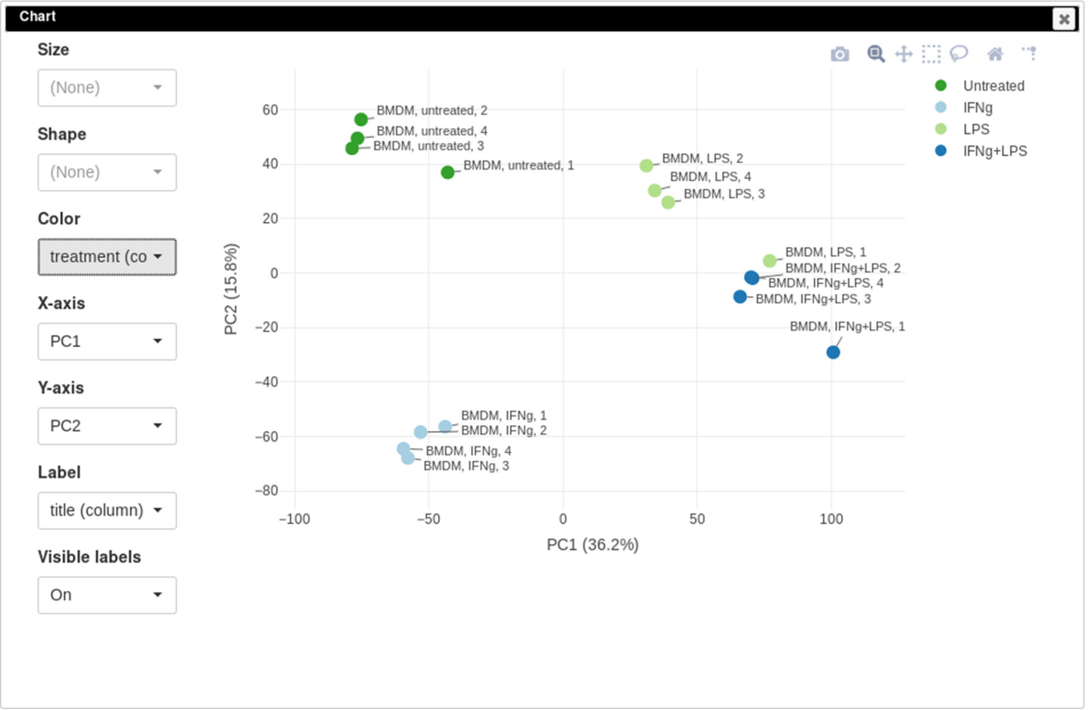
Y-axis: PC2

Label: title (column)

Visible labels: On

Appendix 2—figure 11. Principal component analysis (PCA) plot for dataset GSE53986.

You can customize color, size, and labels of points on the chart using values from annotation (***Appendix 2—figure 12***). Here we set color to come from *treatment* annotation.



Chart

Size: (None)

Shape: (None)

Color: treatment (co

X-axis: PC1

Y-axis: PC2

Label: title (column)

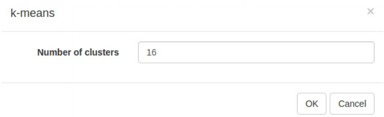
Visible labels: On

Appendix 2—figure 12. Customized principal component analysis (PCA) plot for the dataset GSE56986.

It can be seen that in this dataset the first replicates in each condition are outliers.

k-means clustering

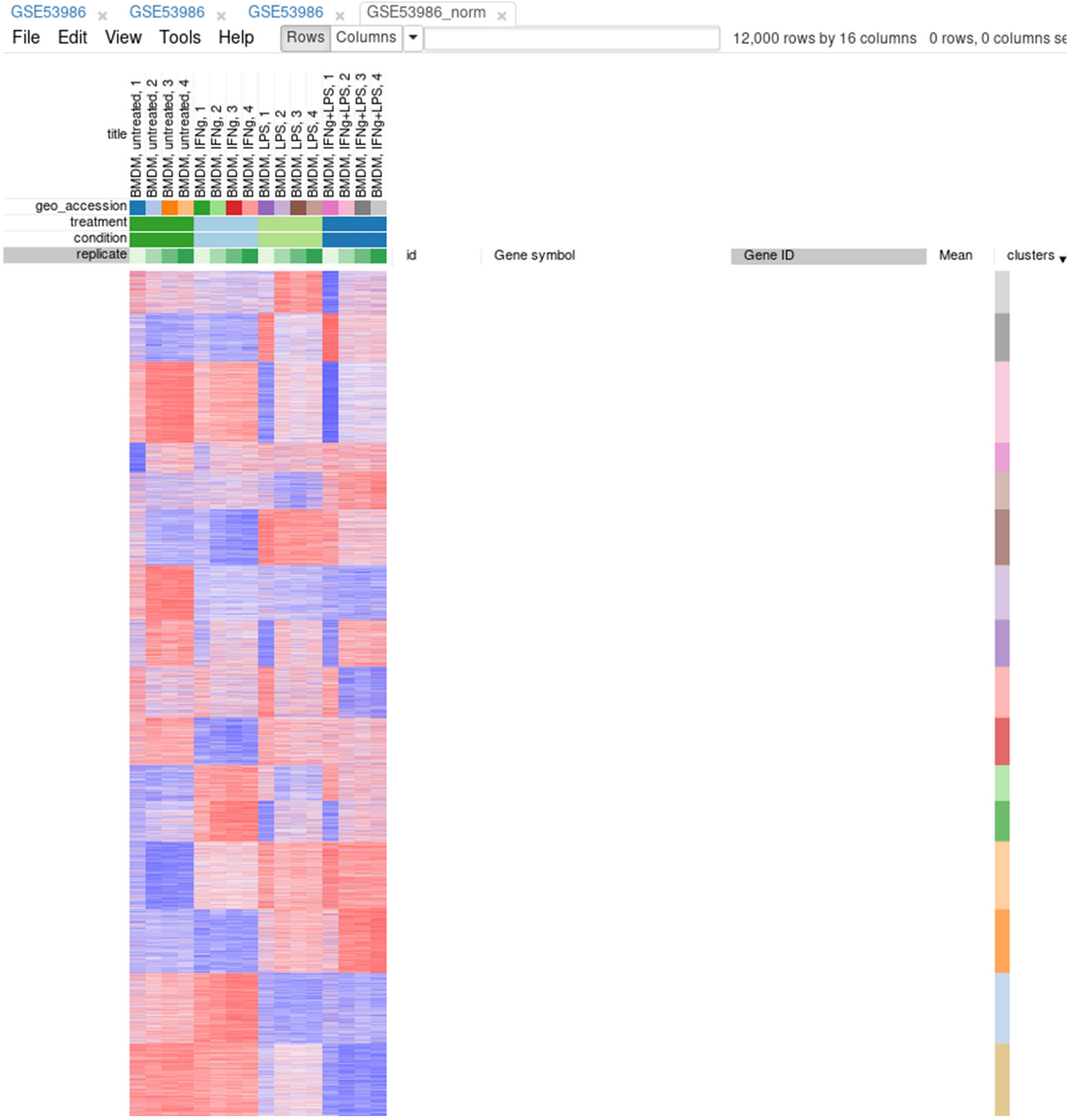
Another useful dataset exploration tool is k-means clustering. Use *Tools/Clustering/k-means* to cluster genes into 16 clusters (***Appendix 2—figure 13***).


k-means

Number of clusters: 16

Appendix 2—figure 13. k-means tool settings.

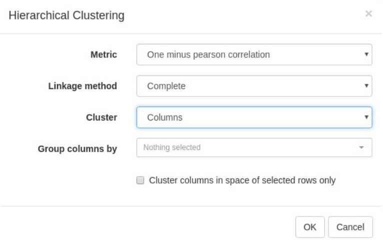
Afterward, rows can be sorted by the *clusters* column. By using the menu *View/Fit to window*, one can get a 'bird's-eye view' on the dataset (***Appendix 2—figure 14***). Here also one can clearly see the outlying samples.



Appendix 2—figure 14. Clusters obtained using the k-means algorithm.

Hierarchical clustering

Tool/Hierarchical clustering menu can be used to cluster samples and highlight outliers (and concordance of other samples) even further (***Appendix 2—figure 15***).



Hierarchical Clustering

Metric: One minus pearson correlation

Linkage method: Complete

Cluster: Columns

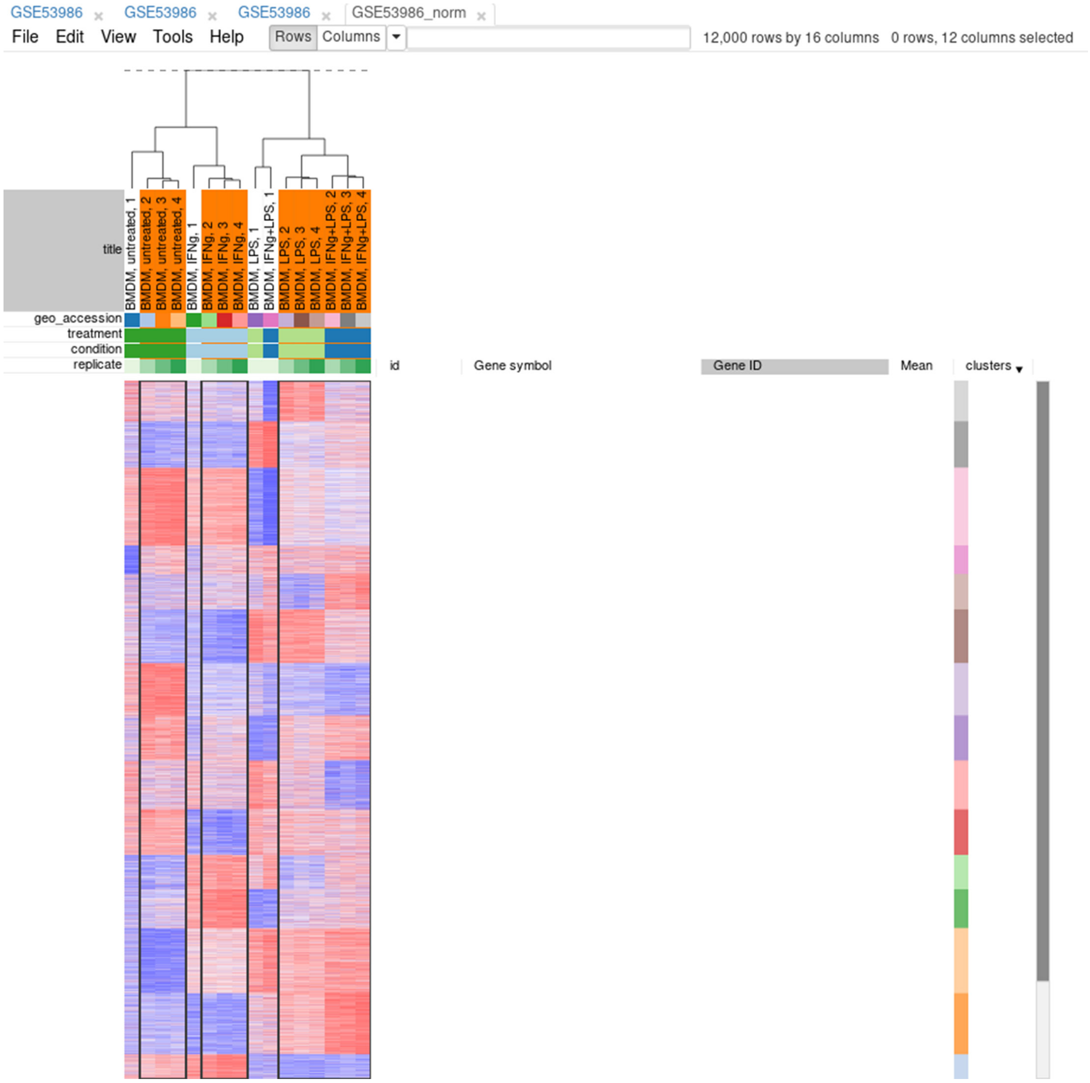
Group columns by: Nothing selected

□ Cluster columns in space of selected rows only

Appendix 2—figure 15. Hierarchical clustering settings.

Filtering outliers

Now, when outliers are confirmed and easily viewed with the dendrogram from the previous step, you can select the good samples (***Appendix 2—figure 16***) and extract them into another heatmap (by clicking *Tools/New Heat Map* or pressing *Ctrl + X*).



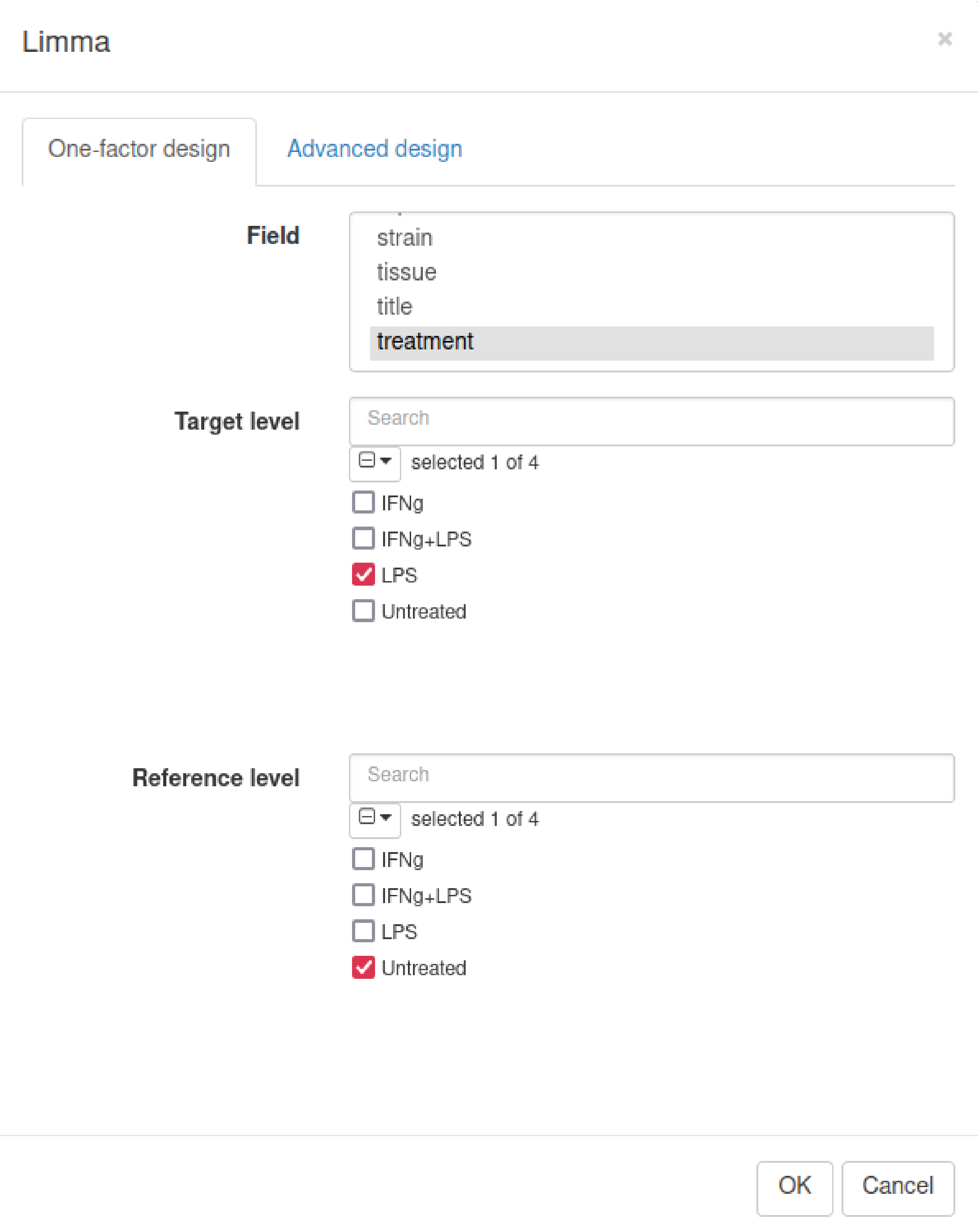
12,000 rows by 16 columns 0 rows, 12 columns selected

Appendix 2—figure 16. Good samples selection.

Differential gene expression

Applying *limma* tool

Differential gene expression analysis can be carried out with the *Tool/Differential Expression/limma* menu. Choose *treatment* as a *Field*, with *Untreated* and *LPS* as classes (***Appendix 2—figure 17***). Clicking *OK* will call the differential gene expression analysis method with the *limma* R package.



Limma

One-factor design | Advanced design

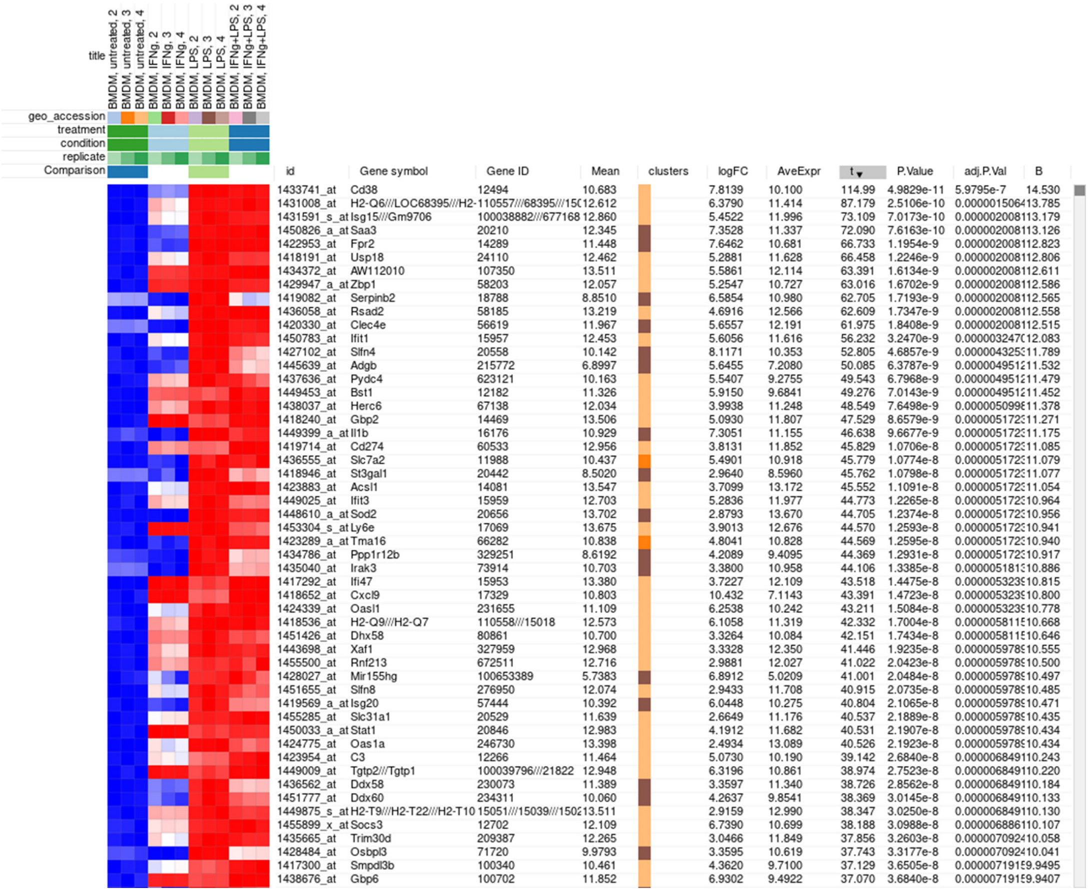
Field
- strain
- tissue
- title
- treatment (selected)

Target level
Search
selected 1 of 4
- ☐ IFNg
- ☐ IFNg+LPS
- ☑ LPS
- ☐ Untreated

Reference level
Search
selected 1 of 4
- ☐ IFNg
- ☐ IFNg+LPS
- ☐ LPS
- ☑ Untreated

Appendix 2—figure 17. Settings for differential expression analysis by limma package.

The rows can be ordered by decreasing the *t*-statistic column to see which genes are the most upregulated upon LPS treatment (***Appendix 2—figure 18***).



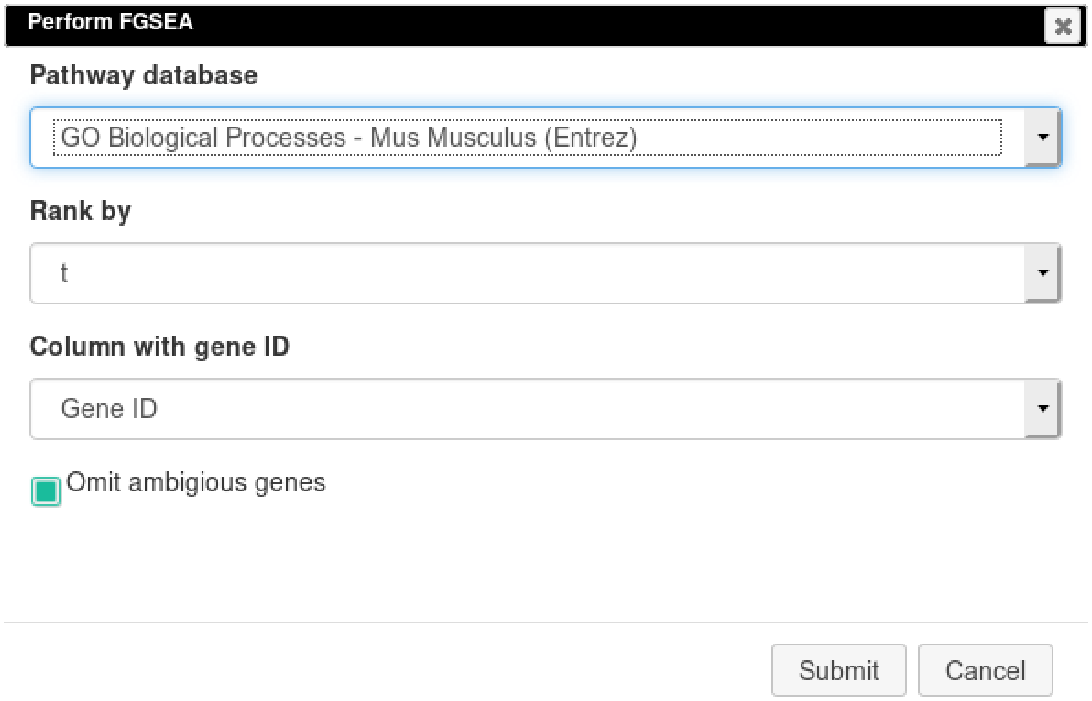
id	Gene symbol	Gene ID	Mean	clusters	logFC	AveExpr	t ▼	P.Value	adj.P.Val	B
1433741_at	Cd38	12494	10.683	🟨	7.8139	10.100	114.99	4.9829e-11	5.9795e-7	14.530
1431008_at	H2-Q6///LOC68395///H2-	110557///68395///15(	12.612	🟨	6.3790	11.414	87.179	2.5106e-10	0.0000015064	13.785
1431591_s_at	Isg15///Gm9706	100038882///677168	12.860	🟨	5.4522	11.996	73.109	7.0173e-10	0.0000020081	13.179
1450826_a_at	Saa3	20210	12.345	🟫	7.3528	11.337	72.090	7.6163e-10	0.0000020081	13.126
1422953_at	Fpr2	14289	11.448	🟫	7.6462	10.681	66.733	1.1954e-9	0.0000020081	12.823
1418191_at	Usp18	24110	12.462	🟨	5.2881	11.628	66.458	1.2246e-9	0.0000020081	12.806
1434372_at	AW112010	107350	13.511	🟨	5.5861	12.114	63.391	1.6134e-9	0.0000020081	12.611
1429947_a_at	Zbp1	58203	12.057	🟨	5.2547	10.727	63.016	1.6702e-9	0.0000020081	12.586
1419082_at	Serpinb2	18788	8.8510	🟫	6.5854	10.980	62.705	1.7193e-9	0.0000020081	12.565
1436058_at	Rsad2	58185	13.219	🟨	4.6916	12.566	62.609	1.7347e-9	0.0000020081	12.558
1420330_at	Clec4e	56619	11.967	🟫	5.6557	12.191	61.975	1.8408e-9	0.0000020081	12.515
1450783_at	Ifit1	15957	12.453	🟨	5.6056	11.616	56.232	3.2470e-9	0.0000032470	12.083
1427102_at	Slfn4	20558	10.142	🟫	8.1171	10.353	52.805	4.6857e-9	0.0000043253	11.789
1445639_at	Adgb	215772	6.8997	🟫	5.6455	7.2080	50.085	6.3787e-9	0.0000049512	11.532
1437636_at	Pydc4	623121	10.163	🟨	5.5407	9.2755	49.543	6.7968e-9	0.0000049512	11.479
1449453_at	Bst1	12182	11.326	🟨	5.9150	9.6841	49.276	7.0143e-9	0.0000049512	11.452
1438037_at	Herc6	67138	12.034	🟨	3.9938	11.248	48.549	7.6498e-9	0.0000050998	11.378
1418240_at	Gbp2	14469	13.506	🟨	5.0930	11.807	47.529	8.6579e-9	0.0000051723	11.271
1449399_a_at	Il1b	16176	10.929	🟫	7.3051	11.155	46.638	9.6677e-9	0.0000051723	11.175
1419714_at	Cd274	60533	12.956	🟨	3.8131	11.852	45.829	1.0706e-8	0.0000051723	11.085
1436555_at	Slc7a2	11988	10.437	🟧	5.4901	10.918	45.779	1.0774e-8	0.0000051723	11.079
1418946_at	St3gal1	20442	8.5020	🟫	2.9640	8.5960	45.762	1.0798e-8	0.0000051723	11.077
1423883_at	Acsl1	14081	13.547	🟨	3.7099	13.172	45.552	1.1091e-8	0.0000051723	11.054
1449025_at	Ifit3	15959	12.703	🟨	5.2836	11.977	44.773	1.2265e-8	0.0000051723	10.964
1448610_a_at	Sod2	20656	13.702	🟫	2.8793	13.670	44.705	1.2374e-8	0.0000051723	10.956
1453304_s_at	Ly6e	17069	13.675	🟨	3.9013	12.676	44.570	1.2593e-8	0.0000051723	10.941
1423289_a_at	Tma16	66282	10.838	🟧	4.8041	10.828	44.569	1.2595e-8	0.0000051723	10.940
1434786_at	Ppp1r12b	329251	8.6192	🟫	4.2089	9.4095	44.369	1.2931e-8	0.0000051723	10.917
1435040_at	Irak3	73914	10.703	🟫	3.3800	10.958	44.106	1.3385e-8	0.0000051813	10.886
1417292_at	Ifi47	15953	13.380	🟨	3.7227	12.109	43.518	1.4475e-8	0.0000053239	10.815
1418652_at	Cxcl9	17329	10.803	🟨	10.432	7.1143	43.391	1.4723e-8	0.0000053239	10.800
1424339_at	Oasl1	231655	11.109	🟨	6.2538	10.242	43.211	1.5084e-8	0.0000053239	10.778
1418536_at	H2-Q9///H2-Q7	110558///15018	12.573	🟨	6.1058	11.319	42.332	1.7004e-8	0.0000058115	10.668
1451426_at	Dhx58	80861	10.700	🟨	3.3264	10.084	42.151	1.7434e-8	0.0000058115	10.646
1443698_at	Xaf1	327959	12.968	🟨	3.3328	12.350	41.446	1.9235e-8	0.0000059789	10.555
1455500_at	Rnf213	672511	12.716	🟨	2.9881	12.027	41.022	2.0423e-8	0.0000059789	10.500
1428027_at	Mir155hg	100653389	5.7383	🟫	6.8912	5.0209	41.001	2.0484e-8	0.0000059789	10.497
1451655_at	Slfn8	276950	12.074	🟨	2.9433	11.708	40.915	2.0735e-8	0.0000059789	10.485
1419569_a_at	Isg20	57444	10.392	🟫	6.0448	10.275	40.804	2.1065e-8	0.0000059789	10.471
1455285_at	Slc31a1	20529	11.639	🟨	2.6649	11.176	40.537	2.1889e-8	0.0000059789	10.435
1450033_a_at	Stat1	20846	12.983	🟨	4.1912	11.682	40.531	2.1907e-8	0.0000059789	10.434
1424775_at	Oas1a	246730	13.398	🟨	2.4934	13.089	40.526	2.1923e-8	0.0000059789	10.434
1423954_at	C3	12266	11.464	🟨	5.0730	10.190	39.142	2.6840e-8	0.0000068491	10.243
1449009_at	Tgtp2///Tgtp1	100039796///21822	12.948	🟨	6.3196	10.861	38.974	2.7523e-8	0.0000068491	10.220
1436562_at	Ddx58	230073	11.389	🟫	3.3597	11.340	38.726	2.8562e-8	0.0000068491	10.184
1451777_at	Ddx60	234311	10.060	🟫	4.2637	9.8541	38.369	3.0145e-8	0.0000068491	10.133
1449875_s_at	H2-T9///H2-T22///H2-T10	15051///15039///150:	13.511	🟨	2.9159	12.990	38.347	3.0250e-8	0.0000068491	10.130
1455899_x_at	Socs3	12702	12.109	🟨	6.7390	10.699	38.188	3.0988e-8	0.0000068861	10.107
1435665_at	Trim30d	209387	12.265	🟨	3.0466	11.849	37.856	3.2603e-8	0.0000070924	10.058
1428484_at	Osbpl3	71720	9.9793	🟫	3.3595	10.619	37.743	3.3177e-8	0.0000070924	10.041
1417300_at	Smpdl3b	100340	10.461	🟨	4.3620	9.7100	37.129	3.6505e-8	0.0000071915	9.9495
1438676_at	Gbp6	100702	11.852	🟨	6.9302	9.4922	37.070	3.6840e-8	0.0000071915	9.9407

Appendix 2—figure 18. Differential expression analysis results for the dataset GSE53986.

Pathway analysis with FGSEA

The results of differential gene expression can be used for pathway enrichment analysis with the *FGSEA* tool.

Open *Tools/Pathway Analysis/Perform FGSEA*, then select Pathway database, which corresponds to the specimen used in the dataset (*Mus musculus* in this example), ranking column, and column with ENTREZID or Gene IDs (***Appendix 2—figure 19***).



Pathway database: GO Biological Processes - Mus Musculus (Entrez)

Rank by: t

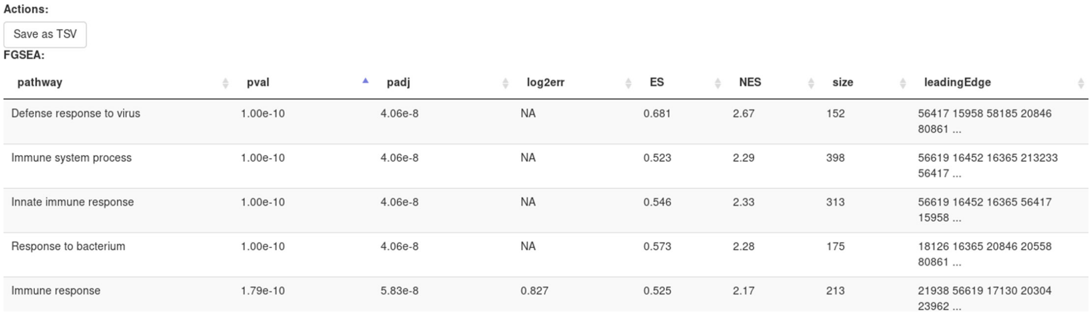
Column with gene ID: Gene ID

☑ Omit ambigious genes

Appendix 2—figure 19. FGSEA settings.

Clicking Submit will open a new tab with pathways table (***Appendix 2—figure 20***).

Clicking on the table row will provide additional information on pathway: pathway name, genes in pathway, leading edge. You can save the result of the analysis in TSV format.



FGSEA:

pathway	pval	padj	log2err	ES	NES	size	leadingEdge
Defense response to virus	1.00e-10	4.06e-8	NA	0.681	2.67	152	56417 15958 58185 20846 80861 ...
Immune system process	1.00e-10	4.06e-8	NA	0.523	2.29	398	56619 16452 16365 213233 56417 ...
Innate immune response	1.00e-10	4.06e-8	NA	0.546	2.33	313	56619 16452 16365 56417 15958 ...
Response to bacterium	1.00e-10	4.06e-8	NA	0.573	2.28	175	18126 16365 20846 20558 80861 ...
Immune response	1.79e-10	5.83e-8	0.827	0.525	2.17	213	21938 56619 17130 20304 23962 ...

Appendix 2—figure 20. Pathways enriched by the FGSE tool for the dataset GSE54986.

Appendix 3

Overview

This section covers the replication of the study by ***Mowel et al., 2017***. In the study, the authors show that *Rroid* locus control innate lymphoid cells (ILCs) by promoting *Id2* gene expression. Here we consider two stages of the analysis that can be performed using Phantasus:

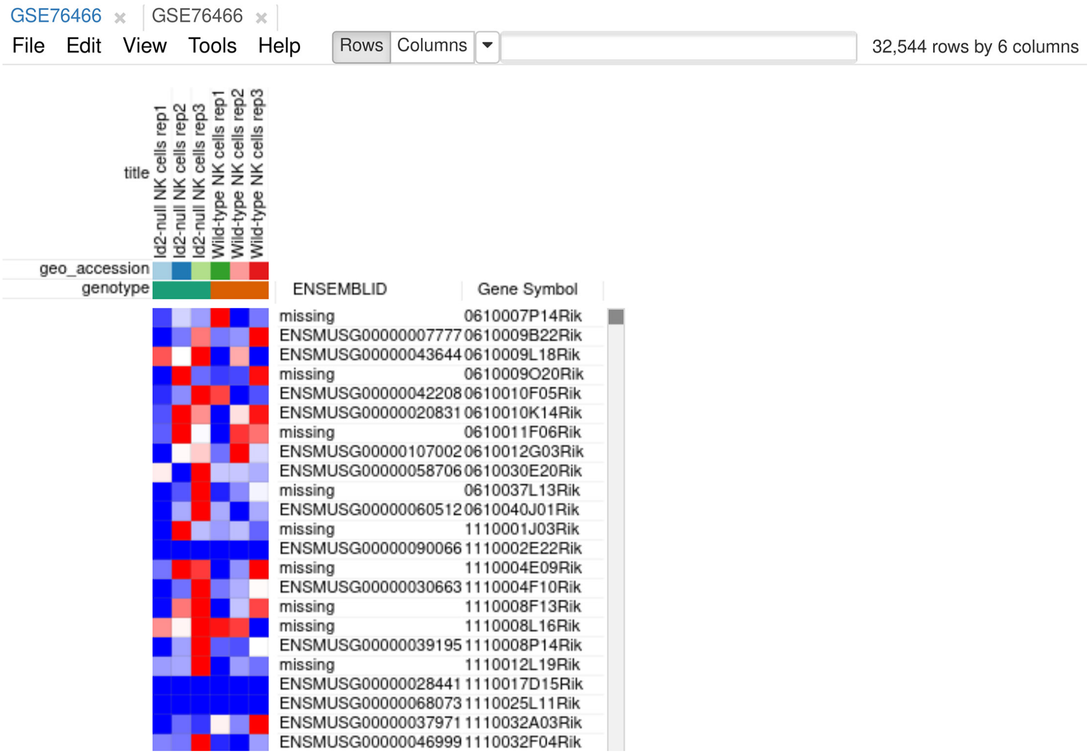
- Obtaining *Id2*-dependent gene signature based on the dataset GSE76466
- Establishing *Rroid*-mediated regulation of the Id2-signature in the dataset GSE101458

Throughout this section, we will provide screenshots to demonstrate control settings for replication.

Obtaining Id2-dependent gene signature

In the original paper, the authors use GSE76466 (***Shih et al., 2016***) dataset that contains WT and *Id2*-null NK cells samples to determine *Id2*-dependent signature. Using Phantasus, this dataset can be explored in a straightforward manner.

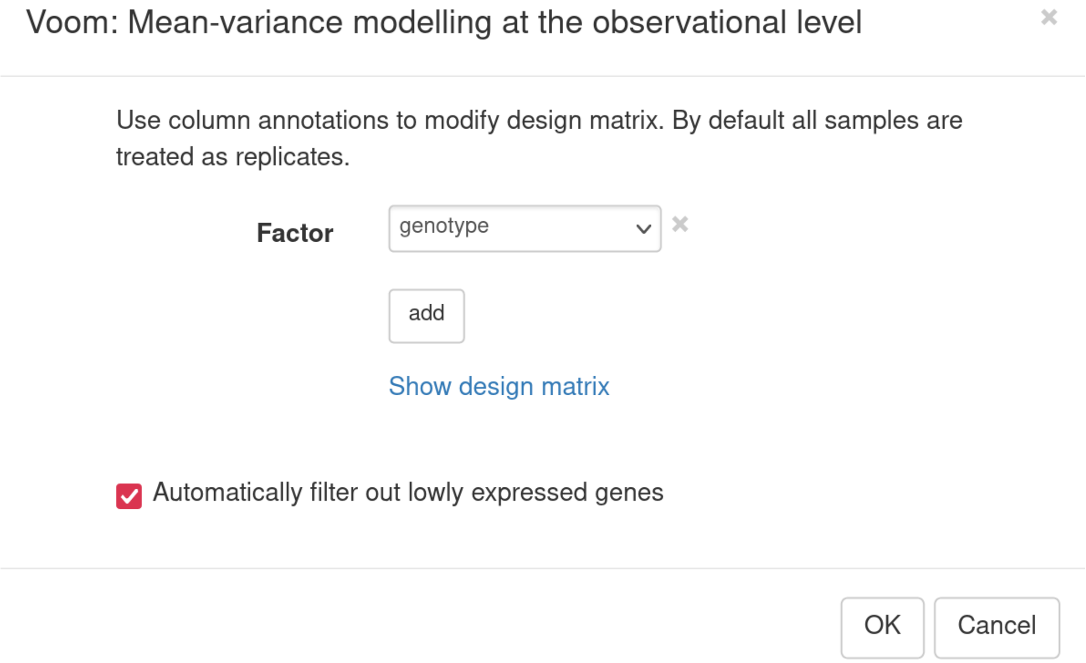
As the first step of the analysis, we will prepare the dataset. The dataset can be opened directly in Phantasus by its GEO identifier. For all the further steps, we will keep only WT and *Id2*-null samples. Select these samples and use the *Tools/New Heat Map* menu to create a new heatmap (***Appendix 3—figure 1***).



ENSEMBLID	Gene Symbol
missing	0610007P14Rik
ENSMUSG00000007777	0610009B22Rik
ENSMUSG00000043644	0610009L18Rik
missing	0610009O20Rik
ENSMUSG00000042208	0610010F05Rik
ENSMUSG00000020831	0610010K14Rik
missing	0610011F06Rik
ENSMUSG00000107002	0610012G03Rik
ENSMUSG00000058706	0610030E20Rik
missing	0610037L13Rik
ENSMUSG00000060512	0610040J01Rik
missing	1110001J03Rik
ENSMUSG00000090066	1110002E22Rik
missing	1110004E09Rik
ENSMUSG00000030663	1110004F10Rik
missing	1110008F13Rik
missing	1110008L16Rik
ENSMUSG00000039195	1110008P14Rik
missing	1110012L19Rik
ENSMUSG00000028441	1110017D15Rik
ENSMUSG00000068073	1110025L11Rik
ENSMUSG00000037971	1110032A03Rik
ENSMUSG00000046999	1110032F04Rik

Appendix 3—figure 1. Heatmap for the dataset GSE76466 after samples filtration.

In the appeared tab, we need to normalize gene counts and filter out genes with low expression. With the help of *Tools/Advanced normalization/Voom* menu, it can be done in one step. Choose genotype as the factor of interest to build the design matrix for Voom and check automatic filtration option (***Appendix 3—figure 2***). Voom (***Law et al., 2014***) normalization converts expression counts into log2-counts per million values and will keep about 10,000 genes for further visualization and analysis.



Voom: Mean-variance modelling at the observational level ✖

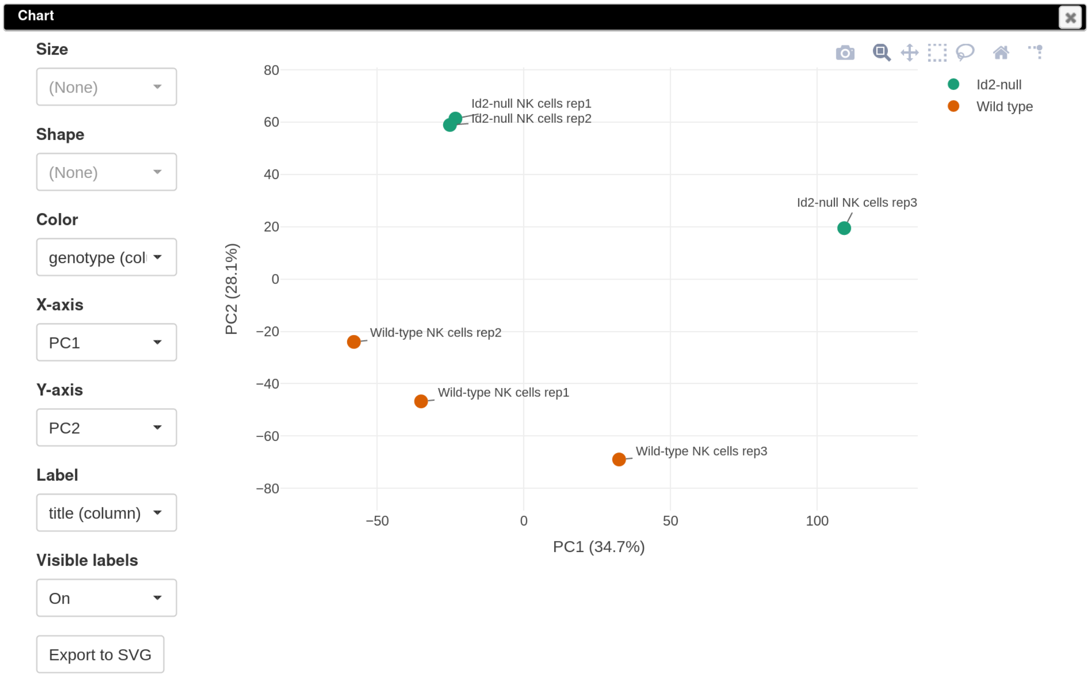
Use column annotations to modify design matrix. By default all samples are treated as replicates.

Factor genotype ✖

☑ Automatically filter out lowly expressed genes

Appendix 3—figure 2. Settings for the Voom tool.

To explore the normalized dataset, click on the *Tools/Plots/PCA plot* menu. The obtained PCA plot (***Appendix 3—figure 3***) shows clear separation between the WT and *Id2*-null samples. This encourages us to do next step and determine the genes that are differentially expressed between these two genotypes.



Chart

Size: (None)

Shape: (None)

Color: genotype (col

X-axis
PC1

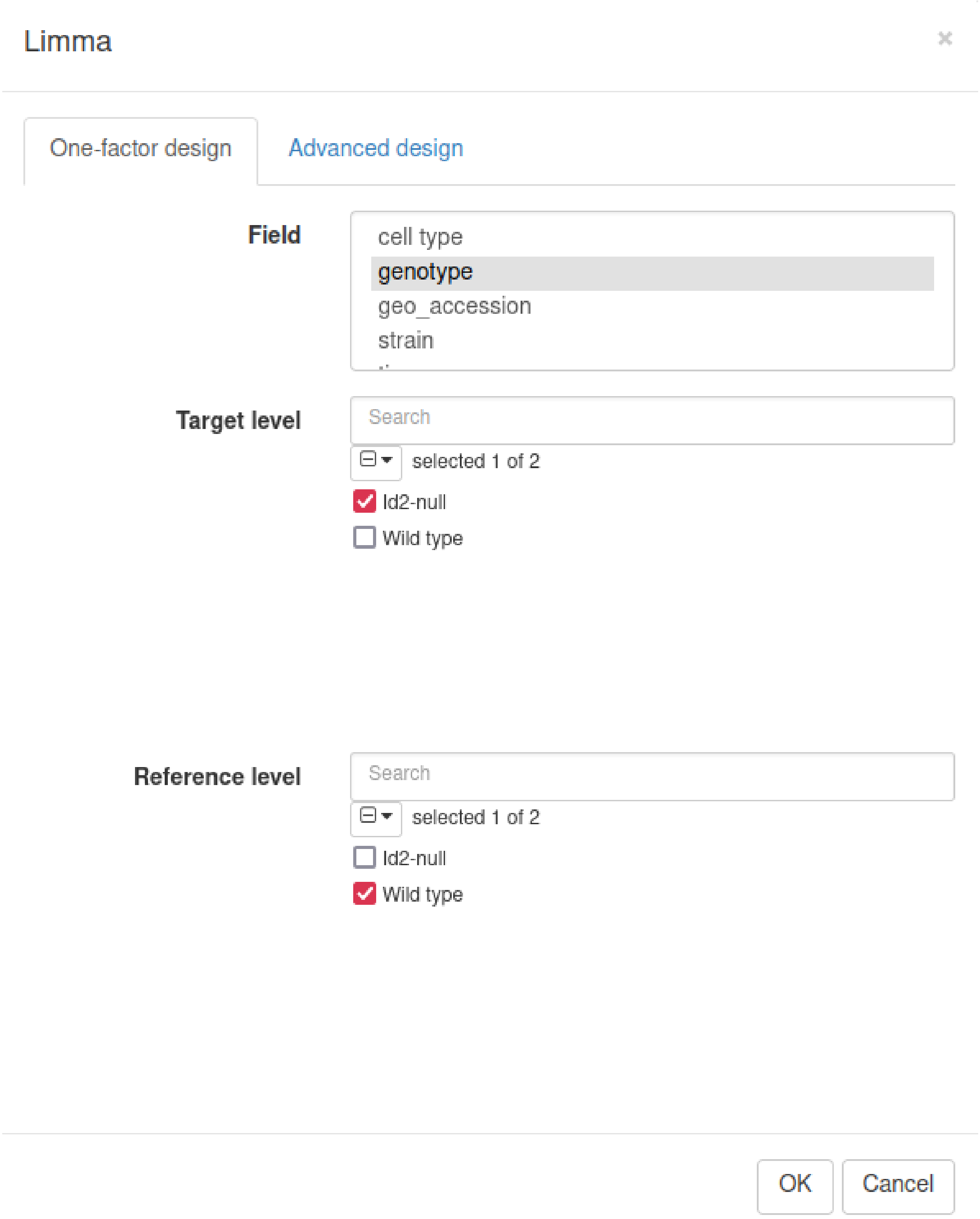
Y-axis
PC2

Label: title (column)

Visible labels: On

Appendix 3—figure 3. Principal component analysis (PCA) plot for the dataset GSE76466.

We will use limma tool (***Ritchie et al., 2015***) for the differential gene expression analysis. It will use the previously obtained weights from the Voom normalization to correct for mean–variance relation in RNA-seq data. In the *Tool/Differential Expression/Limma* menu, set *Id2*-null samples as the target level and WT as the reference (***Appendix 3—figure 4***).



Limma

One-factor design | Advanced design

Field

cell type
genotype
geo_accession
strain

Target level

Search

selected 1 of 2

☑ Id2-null
□ Wild type

Reference level

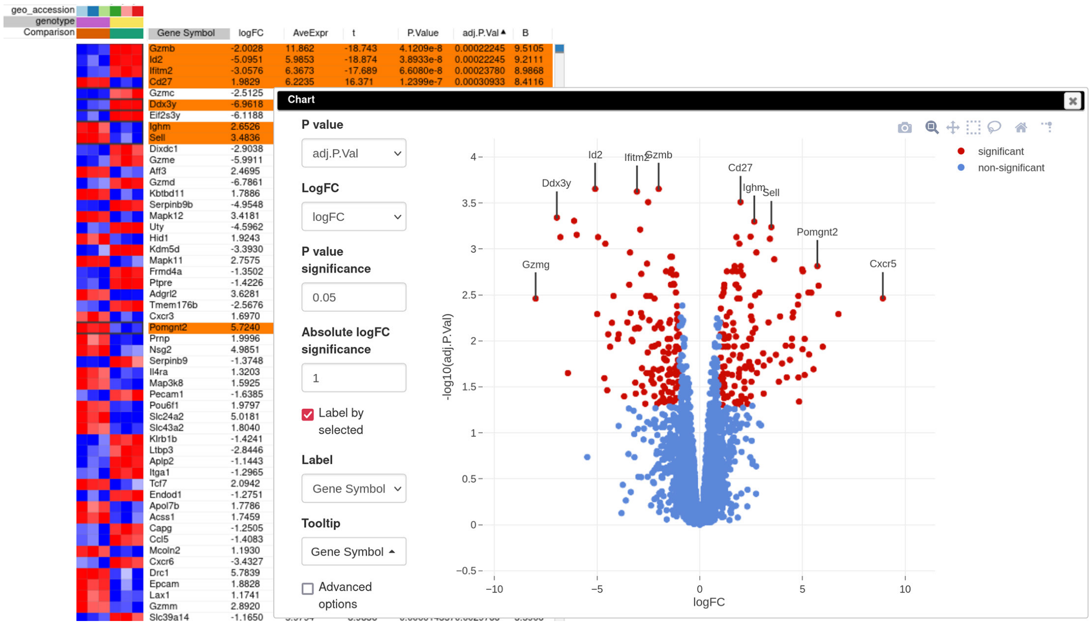
Search

selected 1 of 2

□ Id2-null
☑ Wild type

Appendix 3—figure 4. Limma tool settings for the dataset GSE76466.

A volcano plot generated using the *Tools/Plots/Volcano Plot* menu gives an overview of differential expression (***Appendix 3—figure 5***). In order to obtain the same picture, sort rows by *adj.P.Val* column, select the first few rows, and check the *Label by selected* option to add point annotations for the plot. Notably, *Id2* mRNA expression is downregulated in the knockout samples, which serves as a good positive control.



Gene Symbol	logFC	AveExpr	t	P.Value	adj.P.Val ▲	B
Gzmb	-2.0028	11.862	-18.743	4.1209e-8	0.00022245	9.5105
Id2	-5.0951	5.9853	-18.874	3.8933e-8	0.00022245	9.2111
Ifitm2	-3.0576	6.3673	-17.689	6.6080e-8	0.00023780	8.9868
Cd27	1.9829	6.2235	16.371	1.2399e-7	0.00030933	8.4116
Gzmc	-2.5125					
Ddx3y	-6.9618					
Eif2s3y	-6.1188					
Ighm	2.6526					
Sell	3.4836					
Dixdc1	-2.9038					
Gzme	-5.9911					
Aff3	2.4695					
Gzmd	-6.7861					
Kbtbd11	1.7886					
Serpinb9b	-4.9548					
Mapk12	3.4181					
Uty	-4.5962					
Hid1	1.9243					
Kdm5d	-3.3930					
Mapk11	2.7575					
Frmd4a	-1.3502					
Ptpre	-1.4226					
Adgrl2	3.6281					
Tmem176b	-2.5676					
Cxcr3	1.6970					
Pomgnt2	5.7240					
Prnp	1.9996					
Nsg2	4.9851					
Serpinb9	-1.3748					
Il4ra	1.3203					
Map3k8	1.5925					
Pecam1	-1.6385					
Pou6f1	1.9797					
Slc24a2	5.0181					
Slc43a2	1.8040					
Klrb1b	-1.4241					
Ltbp3	-2.8446					
Aplp2	-1.1443					
Itga1	-1.2965					
Tcf7	2.0942					
Endod1	-1.2751					
Apol7b	1.7786					
Acss1	1.7459					
Capg	-1.2505					
Ccl5	-1.4083					
Mcoln2	1.1930					
Cxcr6	-3.4327					
Drc1	5.7839					
Epcam	1.8828					
Lax1	1.1741					
Gzmm	2.8920					
Slc39a14	-1.1650					

P value: adj.P.Val

LogFC: logFC

P value significance: 0.05

Absolute logFC significance: 1

☑ Label by selected

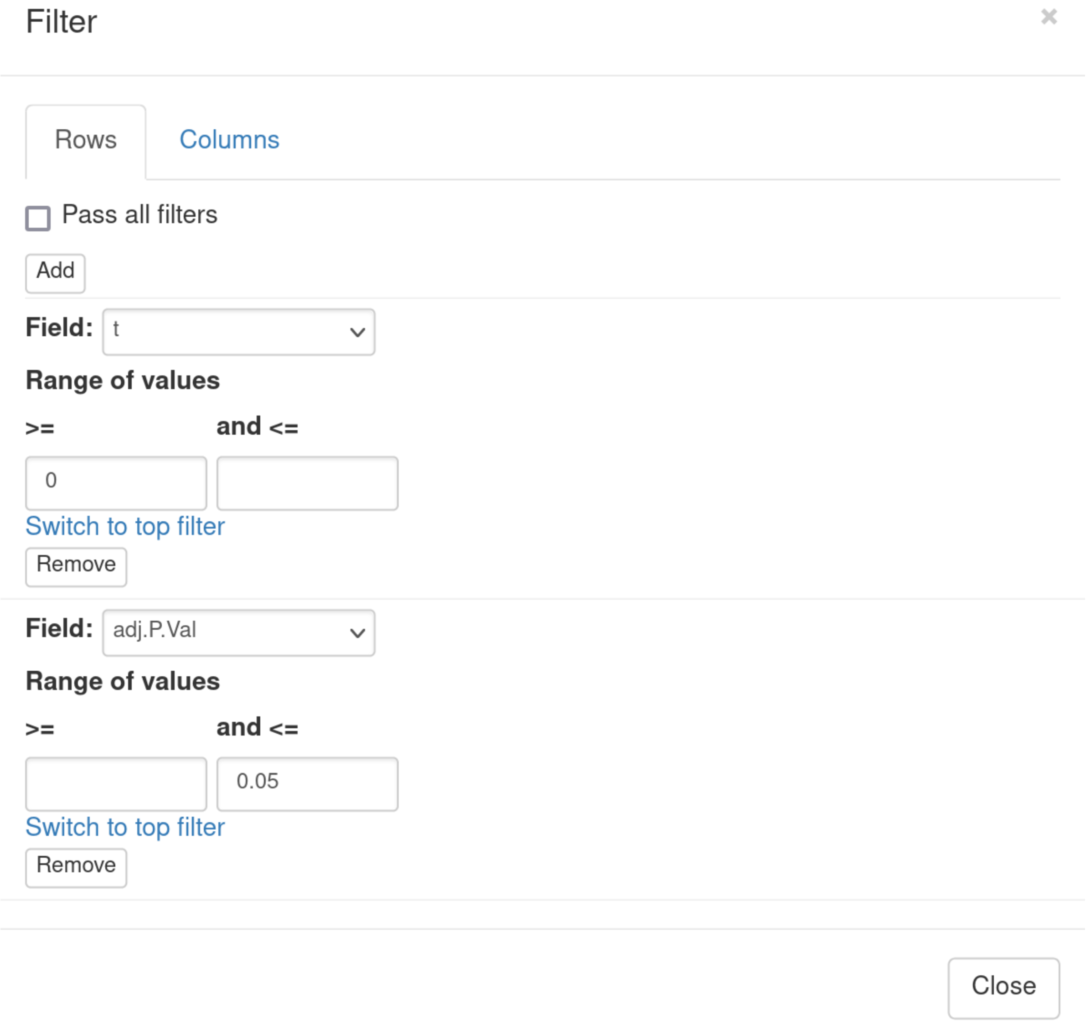
Label: Gene Symbol

Tooltip: Gene Symbol

□ Advanced options

Appendix 3—figure 5. Volcano plot with differential gene expression results between wild-type (WT) and *Id2*-deficient samples for dataset GSE76466.

Desirable *Id2*-dependent gene signature, used by the authors, consists of the genes upregulated in *Id2*-null samples compared to WT. To obtain this signature, we can apply adj.P.Val < 0.05 and $t \geq 0$ filters using the *Tools/Filter* menu (***Appendix 3—figure 6***). Then we copy gene symbols from the filtered rows by doing a right click on any value in the *Gene symbol* column and selecting the option *Copy selected values from Gene symbol column* in the appeared pop-up menu. These genes then can be pasted into a text file to be used for the next stage.



□ Pass all filters

Field: t

Range of values

>= 0 and <=

Field: adj.P.Val

Range of values

>= and <= 0.05

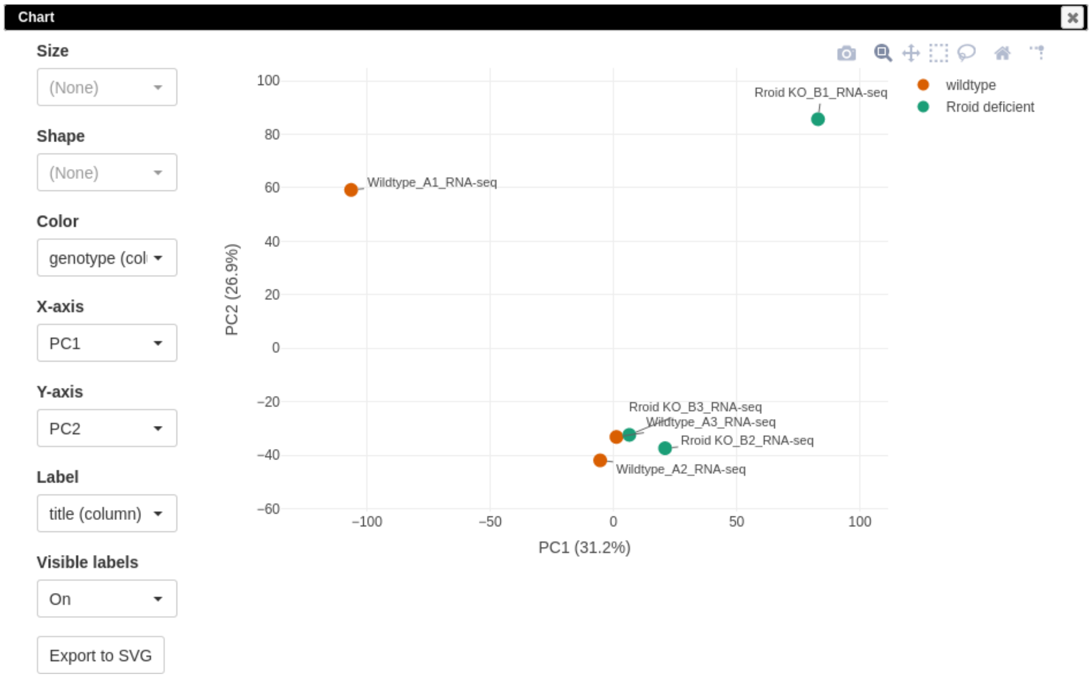
Appendix 3—figure 6. Filter tool settings for the *Id2*-dependent gene signature.

Rroid-mediated regulation in group 1 ILCs

Mowel and colleagues suggested that the discovered *Rroid* locus regulation of ILCs development happens via regulation of *Id2* gene. In order to computationally confirm this hypothesis, the authors examined the enrichment of the previously obtained *Id2* gene signature in the dataset with *Rroid*-deficient and WT samples. The *Rroid* dataset generated by the authors is available in the GEO database under accession number GSE101458.

Once the dataset GSE101458 is loaded into Phantasus, we can generate a PCA plot as we did for the dataset GSE76466. All samples will be used; however, preliminary filtration of lowly expressed genes and normalization is still required. As in the previous section, it can be done with the *Tools/Advanced normalization/Voom* menu. Alternatively, a reasonable result could be obtained by the following steps. Filter top 12,000 genes based on the row mean values: create the mean column using the *Tools/Create Calculated Annotation* menu and then use the top filter on it with *Tools/Filter*. In the *Tools/Adjust* menu, check *One plus log 2* and *Quantile normalize* options to complete all preparations. Finally, build the PCA plot using the adjusted heatmap.

The PCA plot (***Appendix 3—figure 7***) indicates that the first samples in each of the two groups differ significantly from the rest, which could indicate that they are outliers. We will proceed with all of the samples but the results should be taken cautiously.



Chart

Size
(None)

Shape
(None)

Color
genotype (col

X-axis
PC1

Y-axis
PC2

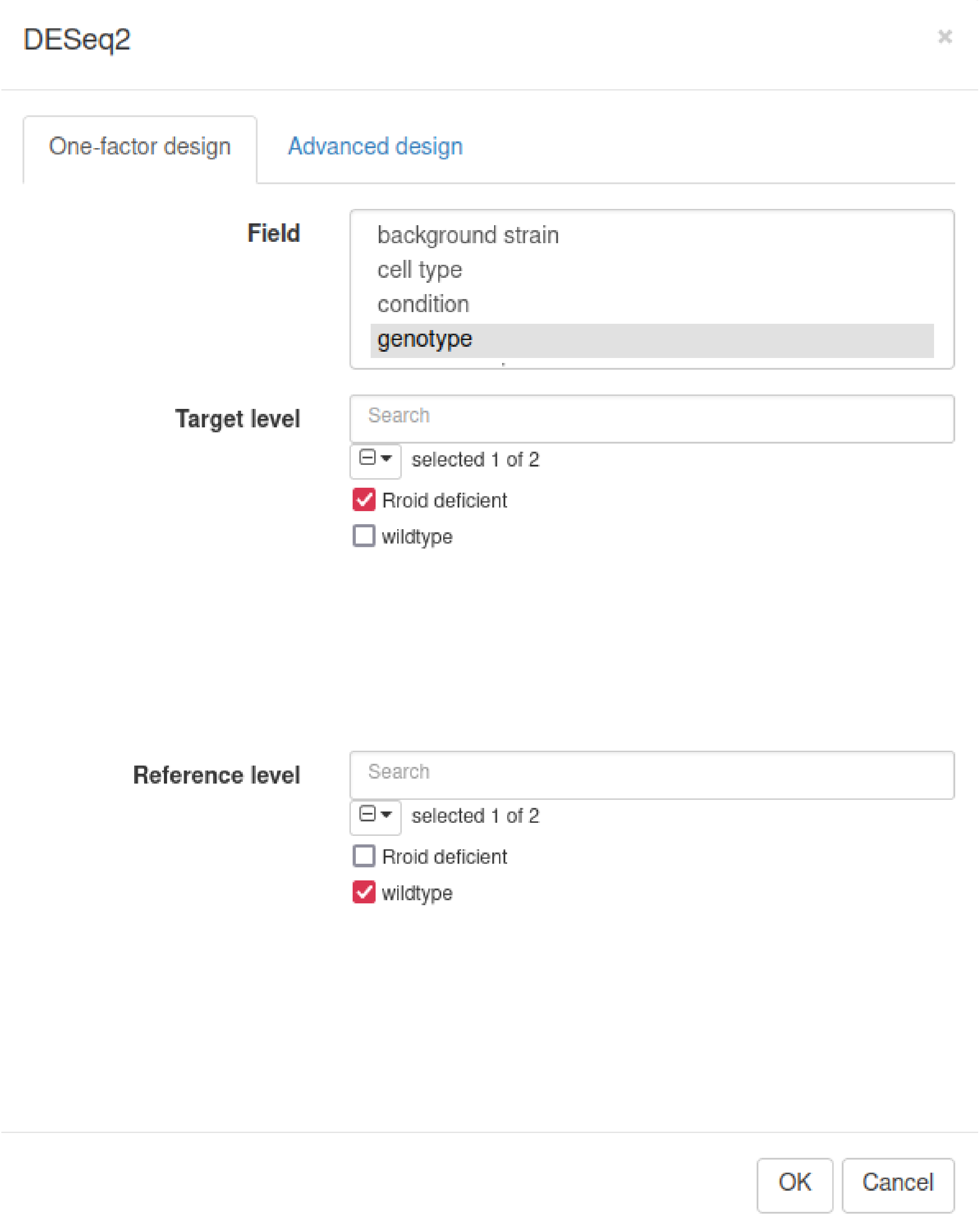
Label
title (column)

Visible labels
On

Export to SVG

Appendix 3—figure 7. Principal component analysis (PCA) plot for dataset GSE101458.

Following the authors, we will use the DESeq2 tool (***Love et al., 2014***) for differential gene expression analysis. Notice that DESeq2 requires raw count matrix with all of the genes as an input, so all of the further steps are performed with the original dataset. Switch the tab to the first one, remove all filters by clicking on the *Remove* button in the *Tools/Filter* menu, and then run the *Tool/ Differential Expression/DESeq2* tool. Use *genotype* field to set up the comparison of *Rroid* deficient versus WT samples (***Appendix 3—figure 8***).



DESeq2

One-factor design | Advanced design

Field

background strain
cell type
condition
genotype

Target level

Search

selected 1 of 2

☑ Rroid deficient
☐ wildtype

Reference level

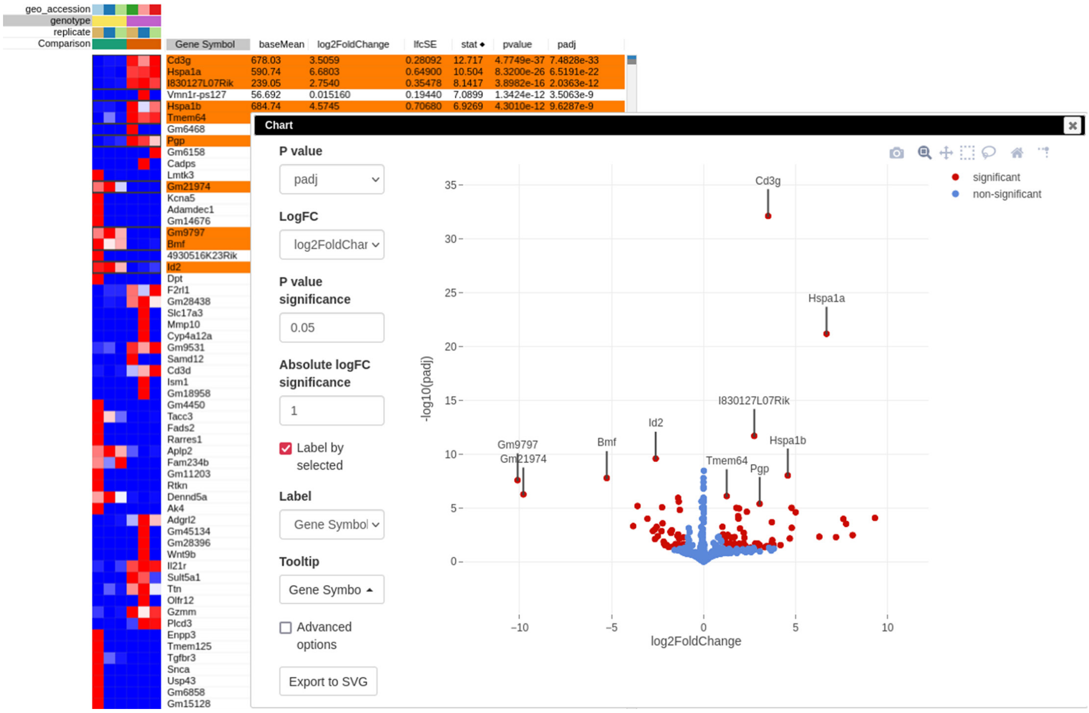
Search

selected 1 of 2

☐ Rroid deficient
☑ wildtype

Appendix 3—figure 8. DESeq2 tool settings for the dataset GSE101458.

After performing differential expression analysis, we can visualize it using the volcano plot. Click on the *stat* column three times to achieve the interleaved sorting of up- and downregulated genes, select several differentially expressed genes at the top of the table, and run *Tools/Plots/Volcano Plot* with the *Label by selected* option. The resulting volcano plot (***Appendix 3—figure 9***) reveals significant downregulation of *Id2* gene, which is consistent with the main hypothesis.



Gene Symbol	baseMean	log2FoldChange	lfcSE	stat ◆	pvalue	padj
Cd3g	678.03	3.5059	0.28092	12.717	4.7749e-37	7.4828e-33
Hspa1a	590.74	6.6803	0.64900	10.504	8.3200e-26	6.5191e-22
I830127L07Rik	239.05	2.7540	0.35478	8.1417	3.8982e-16	2.0363e-12
Vmn1r-ps127	56.692	0.015160	0.19440	7.0899	1.3424e-12	3.5063e-9
Hspa1b	684.74	4.5745	0.70680	6.9269	4.3010e-12	9.6287e-9
Tmem64						
Gm6468						
Pgp						
Gm6158						
Cadps						
Lmtk3						
Gm21974						
Kcna5						
Adamdec1						
Gm14676						
Gm9797						
Bmf						
4930516K23Rik						
Id2						
Dpt						
F2rl1						
Gm28438						
Slc17a3						
Mmp10						
Cyp4a12a						
Gm9531						
Samd12						
Cd3d						
Ism1						
Gm18958						
Gm4450						
Tacc3						
Fads2						
Rarres1						
Aplp2						
Fam234b						
Gm11203						
Rtkn						
Dennd5a						
Ak4						
Adgrl2						
Gm45134						
Gm28396						
Wnt9b						
Il21r						
Sult5a1						
Ttn						
Olfr12						
Gzmm						
Plcd3						
Enpp3						
Tmem125						
Tgfbr3						
Snca						
Usp43						
Gm6858						
Gm15128						

P value: padj

LogFC: log2FoldChar

P value significance: 0.05

Absolute logFC significance: 1

☑ Label by selected

Label: Gene Symbol

Tooltip: Gene Symbo

☐ Advanced options

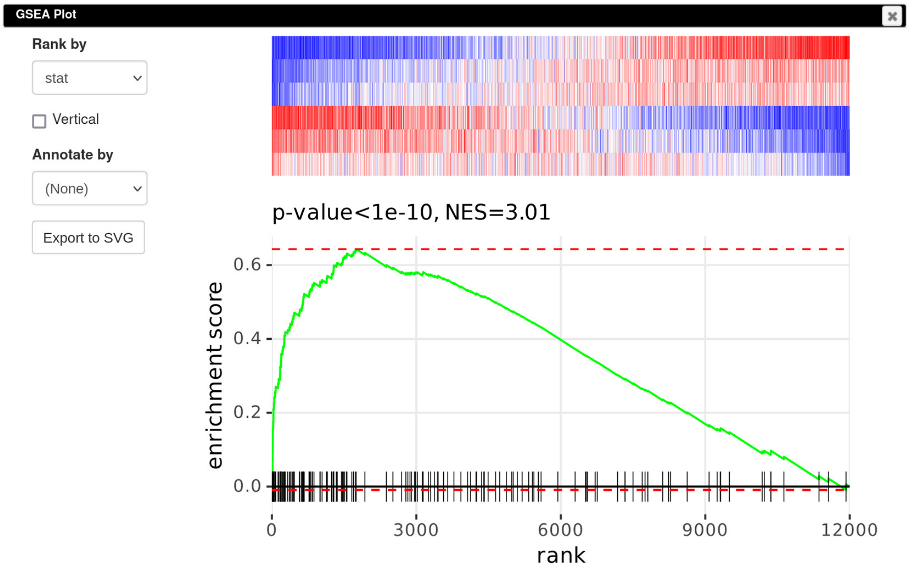
Export to SVG

Appendix 3—figure 9. Volcano plot with differential gene expression results between wild-type (WT) and *Rroid*-deficient samples for dataset GSE101458.

Following ***Mowel et al., 2017***, we aim to conduct the GSEA. GSEA requires to rank genes by the difference in the expression. Using the *stat* column values for that, we will build the enrichment plot for the *Id2*-dependent gene signature.

In order to get a cleaner heatmap representation of the gene expression matrix, we will use scaled values of the top 12,000 expressed genes. This can be reproduced in two steps. First, apply the top filter from the *Tools/Filter* tool on the *baseMean* column. Second, scale the filtered values using logarithmic (*One plus log2*) and quantile normalization options via *Tools/Adjust* menu.

Finally, to create a GSEA plot (***Appendix 3—figure 10***), we copy the previously saved gene signature (Section F), paste it into search field of the filtered DE results, and use the *Tools/Plots/GSEA Plot* menu. The resulting plot shows that the *Id2*-dependent genes are upregulated in the *Rroid*-deficient samples. This plot replicates figure 4E of the original paper by Mowel and colleagues.



Appendix 3—figure 10. Enrichment plot of the *Id2*-dependent gene signature in dataset GSE101458.