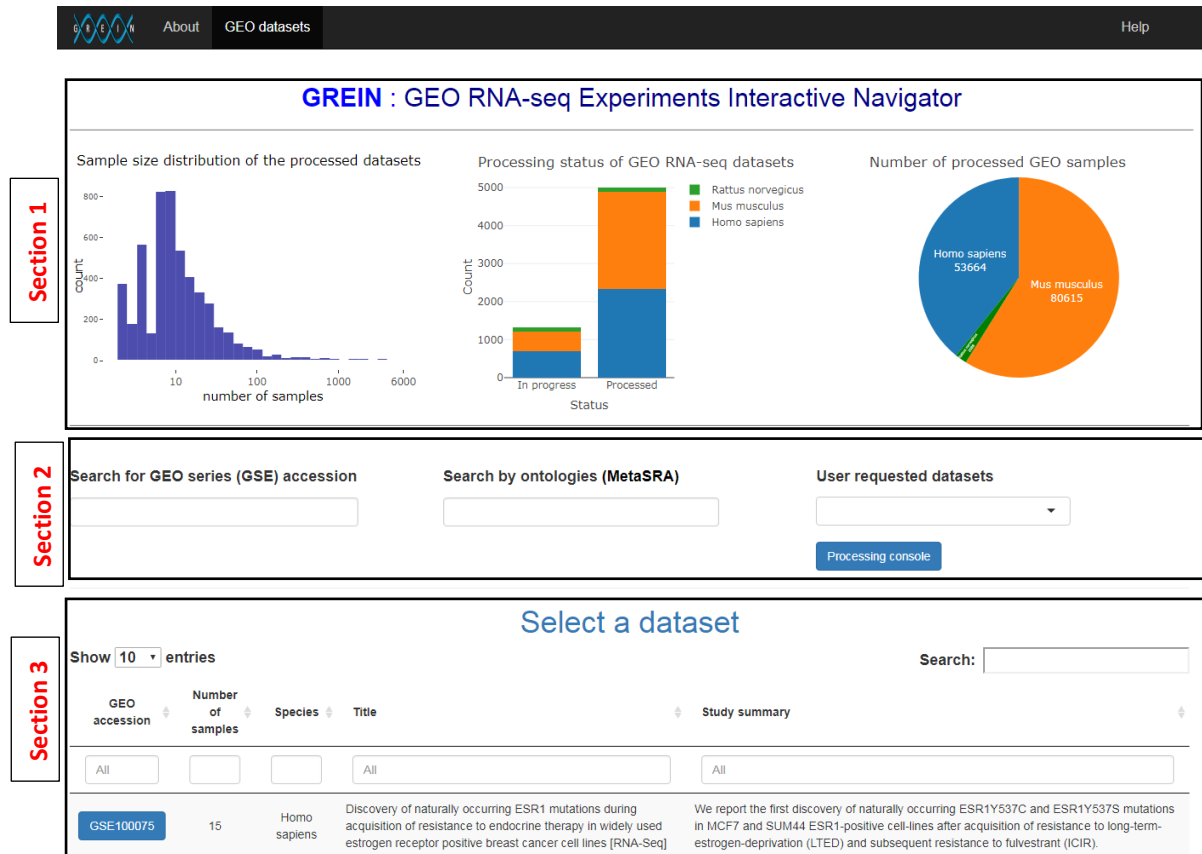


S1. Step-by-step guide of GREIN with an example dataset

S1.1. Landing page (GEO datasets)

To illustrate the usability and efficacy of GREIN, we will walk through the available features for exploring and analyzing data sets with an example. GREIN is a platform independent web application. All you need is to open a web browser and type this url: <https://shiny.ilincs.org/grein> in the address bar which will load the following landing page with three sections as shown below:



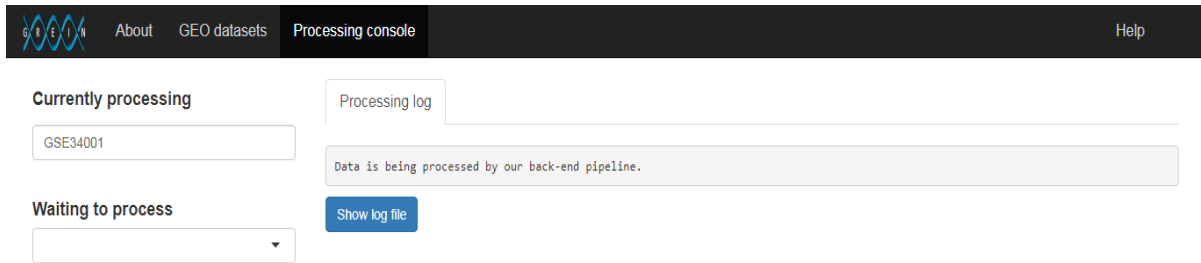
Supplementary Figure 1. GREIN landing page.

S1.1.1. Section 1

The first section provides information regarding the sample size distribution of the processed datasets, total number of data sets already processed or waiting to be processed, and the number of processed human, mouse, or rat samples by our GREP2 pipeline.

S1.1.2. Section 2

The first panel in section 2 provides the option to process your GEO dataset of interest if it is not already processed. You can search for a GEO series accession to see if it exists in the dataset table (Section 4). If not, then 'Start processing' button will appear right below this box and you can initialize the process by clicking this button which will take you to the 'Processing console' window (See supplementary figure 2). You will see your data set id in the 'Currently processing' or at the bottom of the 'Waiting to process' menu. This window also shows the logs of the currently processing dataset requested by a user. A single server processing pipeline is continuously running and processing datasets whenever requested. This pipeline is dedicated to process the user requested datasets only. Depending on the size of the data and queue, the requested data sets are automatically uploaded to the portal as soon as they are processed.



The screenshot shows the 'Processing console' tab in the GREIN application. At the top, there is a navigation bar with 'About', 'GEO datasets', 'Processing console', and 'Help'. The main area is divided into two sections: 'Currently processing' and 'Waiting to process'. In the 'Currently processing' section, a text box contains 'GSE34001'. Below it, a 'Waiting to process' section has a dropdown menu. To the right, a 'Processing log' box shows the message 'Data is being processed by our back-end pipeline.' and a 'Show log file' button.

Supplementary Figure 2. Processing console window.

You can also search by biomedical ontologies (for example, cancer, basal cell, kidney, etc.) in the second panel of this section. We use ontology terms mapped to GEO samples by MetaSRA project (<http://metasra.biostat.wisc.edu/>) (Bernstein *et al.*, 2017). Your search term associated ontologies can be found in the ‘**Metadata**’ under ‘**Explore dataset**’ tab.

The final panel in this section shows the user requested data sets. If you have already requested for a data set to process, then you will be able to see your data set id (GEO series accession) once you refresh the web page. You can also see the status of the processing queue by pressing the ‘**Processing console**’ button.

S1.1.3. Section 3

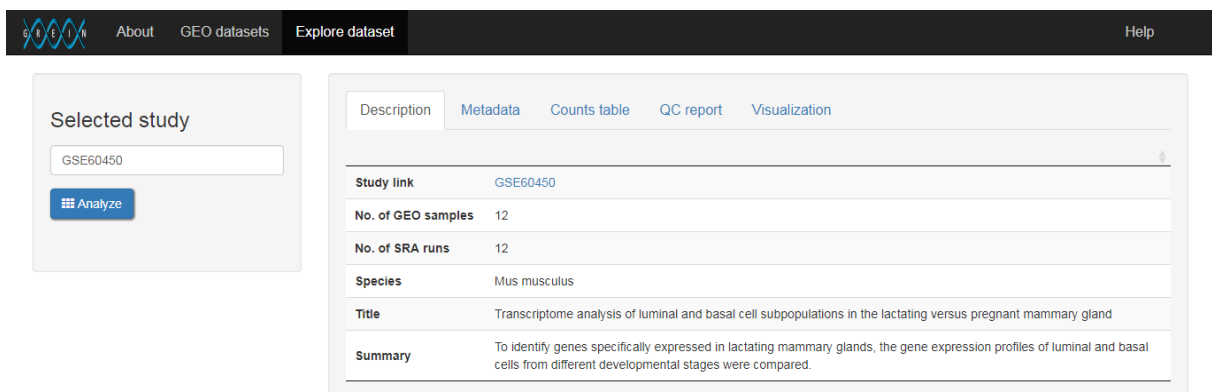
You can see the list of processed data sets with additional information in the data table shown in section 4 (See supplementary figure 1). Two types of search options are available in this table. Search box at the top-right of the table lets a user to search anything in this table. Other search boxes at the top of each column enables column-wise searching. You can click the GEO accession in the first column to start exploring a dataset.

S1.2. Explore dataset

Let us demonstrate the features of GREIN for exploring and analyzing an RNA-seq data by searching ‘**GSE60450**’ either at the top-right or first column's search box in the dataset table. If you click ‘**GSE60450**’, it will take you to the ‘**Explore dataset**’ tab. This experiment was conducted to examine the change in expression profiles between luminal and basal cells in mouse mammary glands of virgin, pregnant, and lactating mice (Fu *et al.*, 2015). The data set is available in GEO as GSE60450.

S1.3.1. Description

This tab panel provides descriptive information including study link, number of GEO samples, number of SRA runs, title, and study summary of the corresponding dataset.



The screenshot shows the 'Explore dataset' tab in the GREIN application. The navigation bar includes 'About', 'GEO datasets', 'Explore dataset', and 'Help'. On the left, a 'Selected study' section shows 'GSE60450' and an 'Analyze' button. The main area displays a table with tabs for 'Description', 'Metadata', 'Counts table', 'QC report', and 'Visualization'. The 'Description' tab is active, showing the following information:

Study link	GSE60450
No. of GEO samples	12
No. of SRA runs	12
Species	Mus musculus
Title	Transcriptome analysis of luminal and basal cell subpopulations in the lactating versus pregnant mammary gland
Summary	To identify genes specifically expressed in lactating mammary glands, the gene expression profiles of luminal and basal cells from different developmental stages were compared.

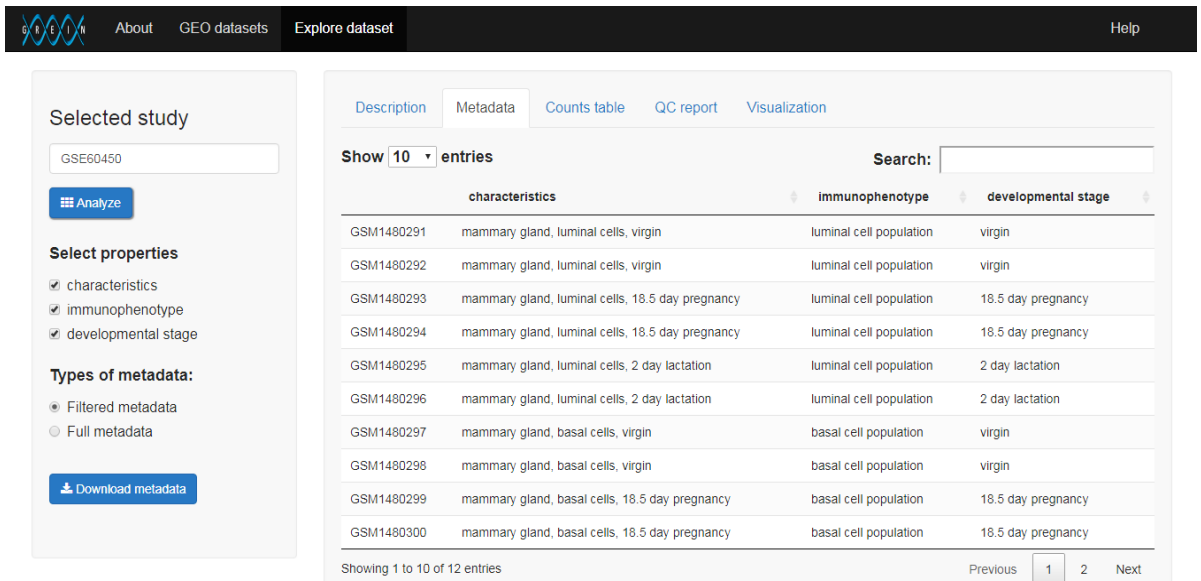
Supplementary Figure 3. Description tab panel.

S1.3.2. Metadata

GEO metadata contains a lot of information, although not all of these are useful for analysis or visualization purpose. So, we provide a filtered version of the metadata besides the full metadata. We filter metadata based on the following criteria:

1. Columns that contain a single value.
2. Columns with incoherent information regarding analysis and visualization such as dates, time, download path and so on.

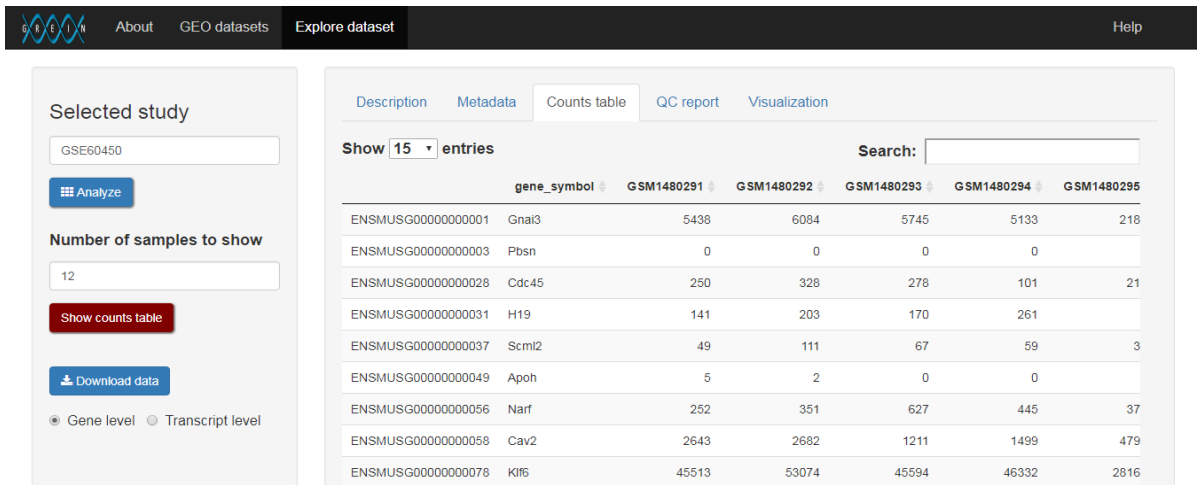
This dataset (GSE60450) has two cell types and three developmental stages and each combination has two biological replicates. You can also download both the filtered and full metadata.



Supplementary Figure 4. Metadata tab panel.

S1.3.3. Counts table

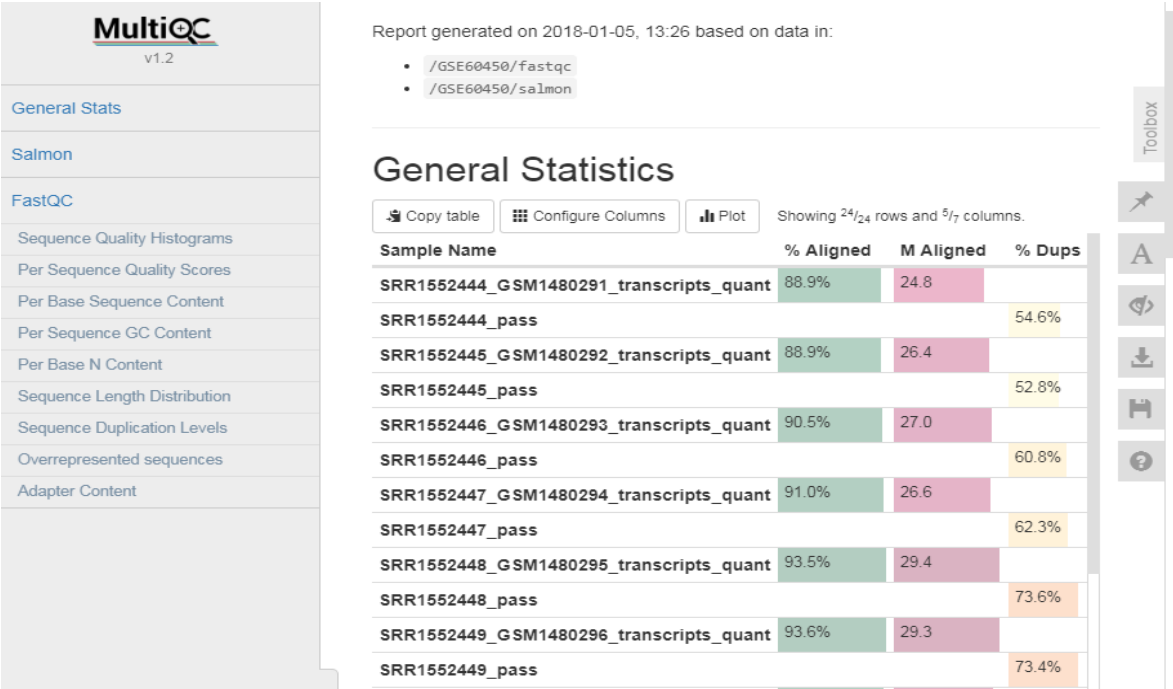
This table shows gene wise estimated read abundance (rounded to the nearest integer) for each sample. We use *Salmon* (Patro *et al.*, 2017) to quantify transcript abundances for each sample. These transcript level estimates are then summarized to gene level using Bioconductor package *tximport* (Soneson *et al.*, 2015) which gives estimated counts scaled up to library size while taking into account for transcript length. We obtained gene annotation for Homo sapiens (GRCh38), Mus musculus (GRCm38), and Rattus norvegicus (Rnor_6.0) from Ensemble (release-91). Both gene and transcript level expression data are downloadable.



Supplementary Figure 5. Counts table tab panel.

S1.3.4. QC report

After running *FastQC* (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) and *Salmon*, we generate a combined quality control report of all the samples using *MultiQC* (Ewels *et al.*, 2016). This downloadable report contains information regarding read mapping and quality scores of the FastQ files. In the general statistics of the FastQ files. In the general statistics table, each sample corresponds to two rows, the first one for the Salmon read mapping and the second one for *FastQC* (See supplementary figure 6).



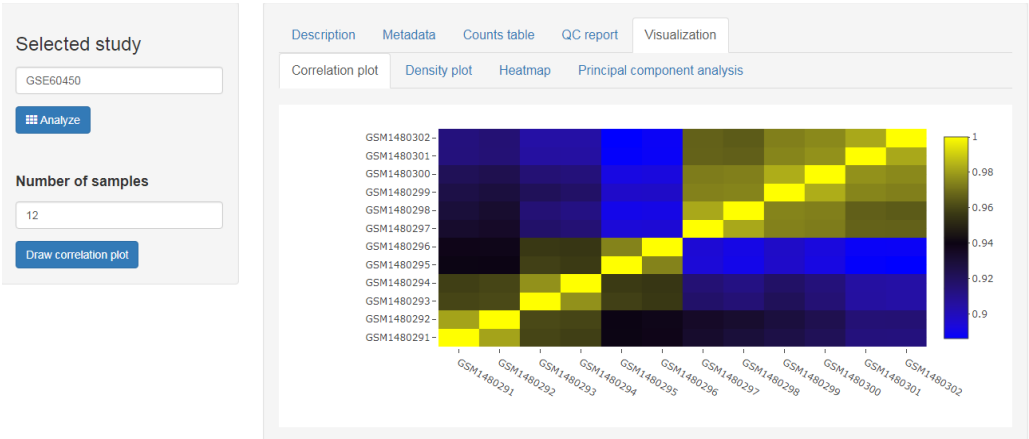
Supplementary Figure 6. MultiQC report.

S1.3.5. Visualization

This section provides access to four different types of interactive exploratory plots. These plots are important in order to uncover underlying relationship of the samples and gain deeper insight of the data structure. We leverage several state-of-the-art R and Bioconductor packages for this purpose.

S1.3.1.1. Correlation plot

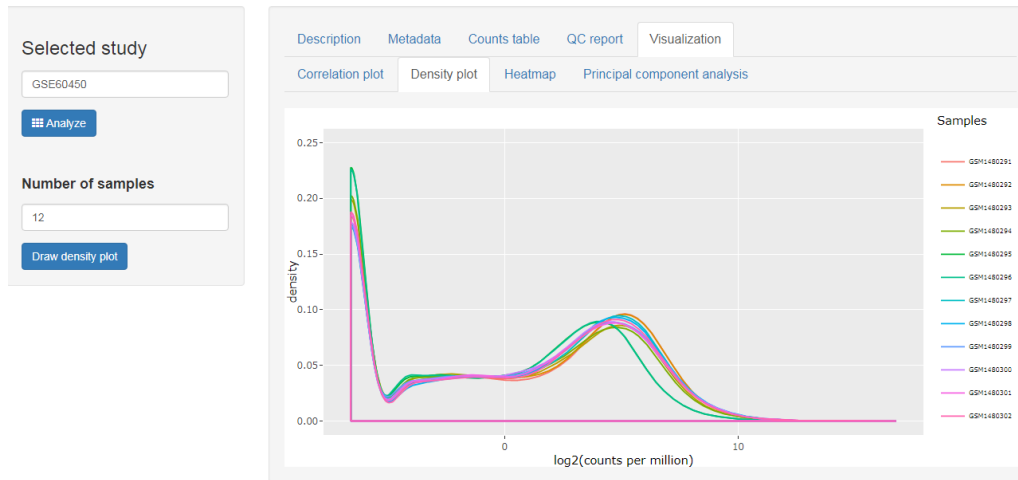
Sample-wise correlation heatmap is generated using *Plotly* (Sievert *et al.*, 2016). User can hover over the heatmap to see the correlation coefficient values or zoom in to any specific area and double click to zoom out. Here, overall sample-to-sample correlation is quite high, although basal or luminal cells have higher within group correlation compared to between group correlation.



Supplementary Figure 7. Correlation plot tab panel.

S1.3.1.2. Density plot

Distribution of the data on the $\log_2(\text{Counts per million})$ scale is shown in the density plot using *Plotly*. You can deselect any sample from the legend in the right side by just clicking on the sample names or generate the density plot for a certain number of samples. The latter option is particularly useful when the sample size is large.



Supplementary Figure 8. Density plot tab panel.

S1.3.1.3. Heatmap

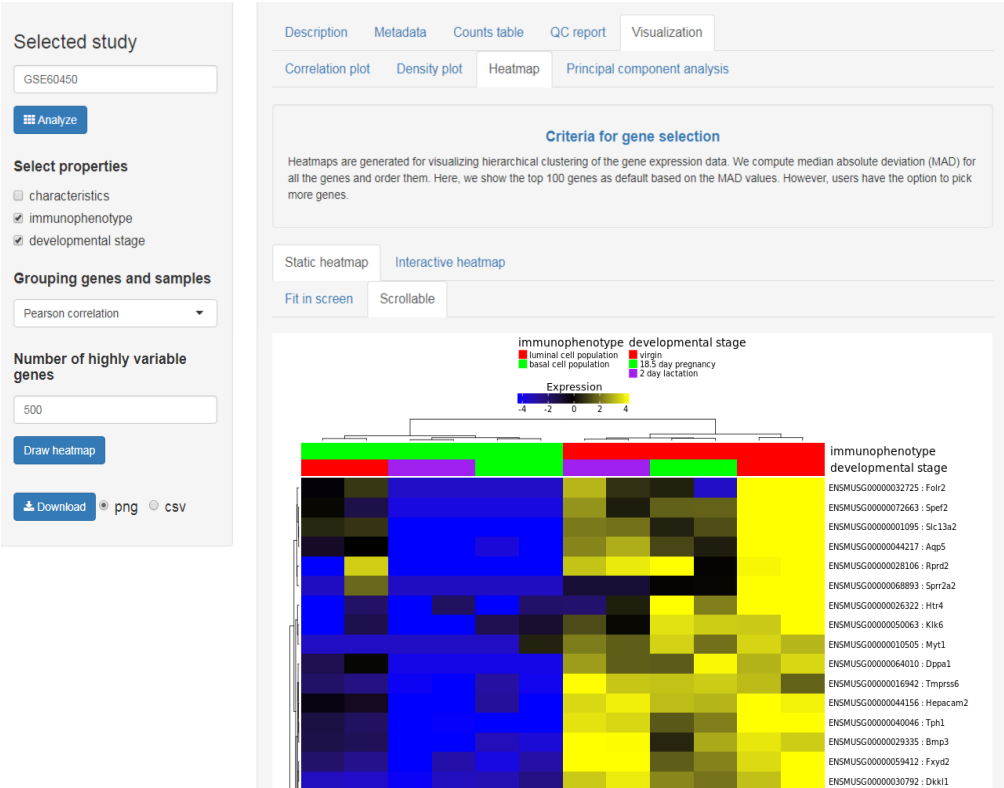
Heatmap is one of the most popular ways to visualize hierarchical clustering of gene expression data. We display heatmap of the top most highly variable genes (sorted by median absolute deviation values of $\log_2(\text{Counts per million})$ and data is centered to the mean) in this section. You can pick any number of genes for clustering based on any of the three methods: Pearson correlation, Euclidean distance, or group by properties. These heatmaps are available in both single and multiple annotations. You can download both the heatmap and the associated data. Two types of heatmaps are available: Static and interactive. We use Bioconductor packages *ComplexHeatmap* (Gu *et al.*, 2016) and R package *iheatmapr* (Schep and Kummerfeld, 2017) for static and interactive heatmaps respectively.

1. a) **Static heatmap (Fit in screen):** A complete picture that fits to the window without the gene symbols:



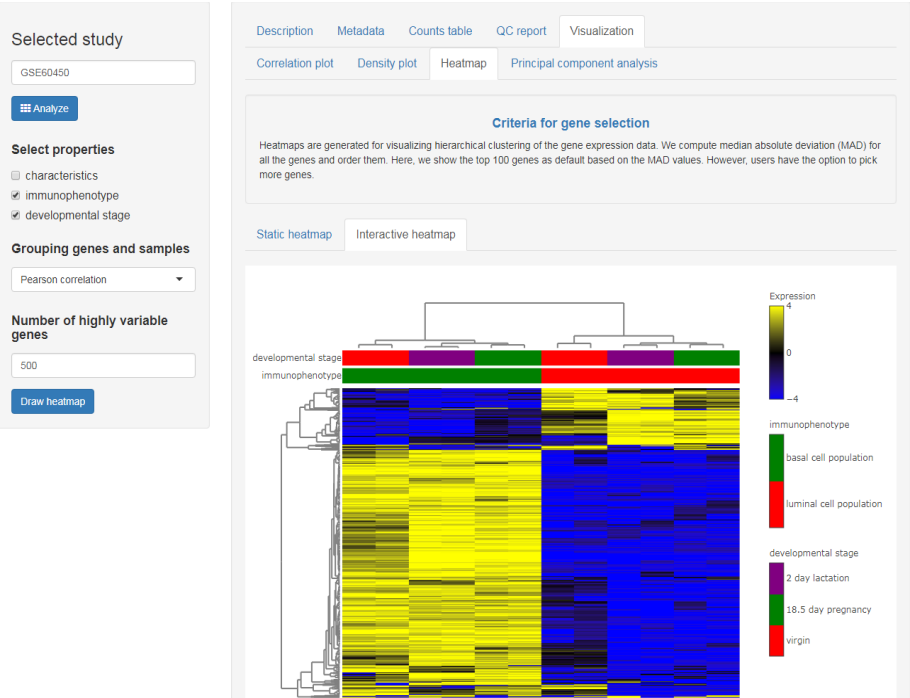
Supplementary Figure 9.a. Static heatmap (Fit in screen) tab panel.

1. **b) Static heatmap (Scrollable):** Shows the gene symbols.



Supplementary Figure 9.b. Static heatmap (Scrollable) tab panel.

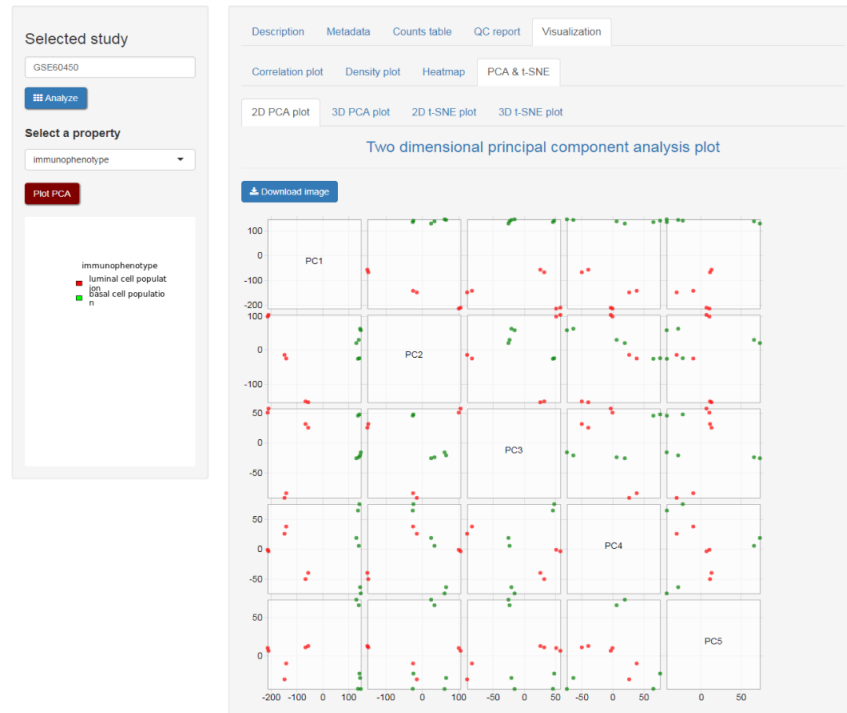
2. **Interactive heatmap:** This heatmap provides the option to see the values and gene symbols while hovering over the heatmap as well as zooming in and out. You can select any area to zoom in and double-click to zoom out.



Supplementary Figure 9.c. Interactive heatmap tab panel.

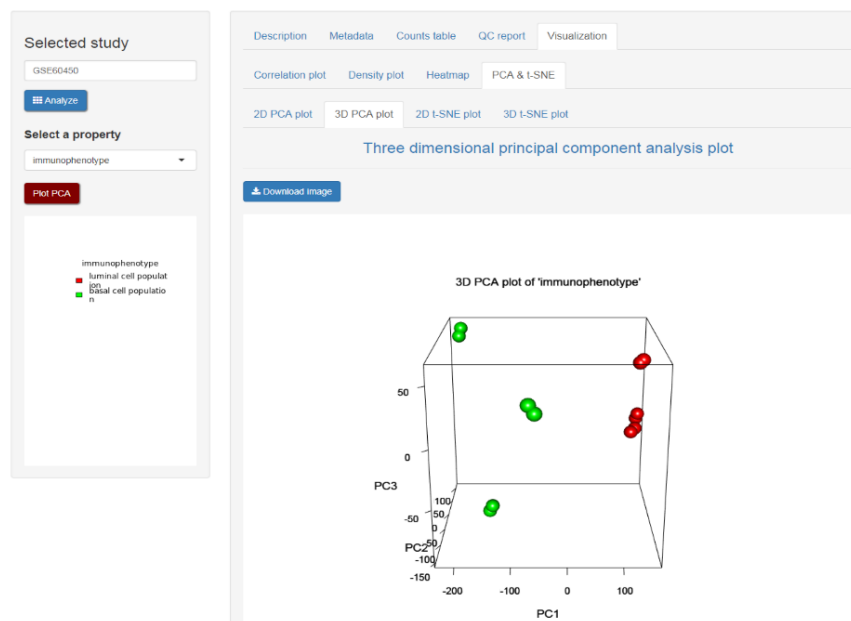
S1.3.1.4. PCA & t-SNE

1. **2-D PCA plot:** Scatter plot matrix of the first five principal components in $\log_2(\text{Counts per million})$ scale is generated using R package *pairsD3* (Tarr, 2015). You can mouse hover to see the sample names or make a square box on single or multiple points to see the location of these points in the graph.



Supplementary Figure 10.a. Two-dimensional principal component analysis tab panel.

2. **3-D PCA plot:** Three-dimensional PCA plot is available to provide more visual flexibility of the principal components on a 3-D plane. We use R package *rgl* (Adler *et al.*, 2017) for 3-D PCA plot.



Supplementary Figure 10.b. Three-dimensional principal component analysis tab panel.

3. **2-D t-SNE plot:** Two-dimensional t-distributed stochastic neighboring embedding (t-SNE) plot is also available to visualize the data in a reduced dimension. We use R package *Rtsne* (Krijthe,J.H. *et al.*, 2015) for 2-D t-SNE plot.



Supplementary Figure 10.c. Two-dimensional t-SNE plot.

4. **3-D t-SNE plot:** Three-dimensional t-distributed stochastic neighboring embedding (t-SNE).



Supplementary Figure 10.d. Three-dimensional t-SNE plot.

S1.3. Analyze dataset

An action button ‘Analyze’ is attached in the left panel of each section of ‘Explore dataset’ tab. Once you click the button, it will take you to the ‘Analyze dataset’ tab.

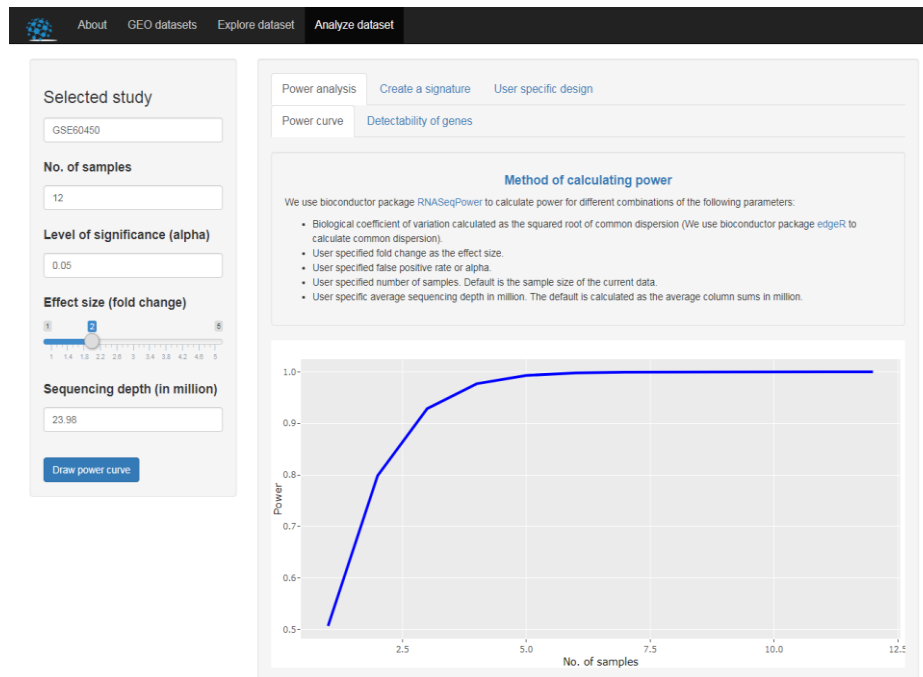
S1.3.1. Power analysis

This section is dedicated to assist users in power analysis which is an essential step in designing an RNA-seq experiment with a goal to achieve the desired power to detect differentially expressed genes. This section is comprised of two sub-sections: Power curve and detectability of genes.

S1.3.1.1. Power curve

We use Bioconductor package *RNASeqPower* (Hart *et al.*, 2013) to calculate power using the following parameters:

1. Biological coefficient of variation calculated as the squared root of the common dispersion (We use Bioconductor package *edgeR* (Robinson *et al.*, 2010) to calculate common dispersion).
2. Number of samples. Default is the sample size of the data in use.
3. Fold change as the effect size. The default value is 2.
4. Level of significance or alpha. The default value is 0.05.
5. Average sequencing depth in million. The default is calculated as the average column sums in million.



Supplementary Figure 11.a. Power curve tab panel.

S1.3.1.2. Detectability of genes

The plot of biological coefficient of variation (BCOV) vs. average $\log_2(\text{Counts per million})$ gives an idea of how the detectability of genes as differentially expressed (DE) may vary based on their BCOV. The line of detectability (LOD) is the estimated values of BCOV for different sequencing depths. Genes above this line have lower chances to be detected as differentially expressed compared to the genes below the line. Basically, LOD tells us that even if we increase the depth of sequencing to a very high number, genes above this line will not have enough power to be detected as DE. You can modify the parameters as per your interest. Also, you can search for a gene to see the location of the gene or hovering over the points which will display gene symbols.



Supplementary Figure 11.b. Detectability of genes tab panel.

S1.3.2. Create a signature

Generating differential expression signature is one of the most important segments of GREIN. This section begins with selecting a variable of interest to test for differential expression between the groups of this variable. We select the variable **‘immunophenotype’** as our group of interest. Depending on the number of available properties and levels, three different types of comparisons are available: two group without covariate, two group with covariate, and multi group. Here, we have two groups available for the selected variable **‘immunophenotype’** and we test for differential expression between **‘basal cell population’** and **‘luminal cell population’**. We choose **‘basal cell population’** as the experimental group and **‘luminal cell population’** as the control group. You can see the selected groups in the **‘Metadata’** table (See supplementary figure 12.a). The variable **‘Selected groups’** in this table is created on the fly based on your selected groups.

A signature table will be generated once you click the **‘Generate signature’** button (See supplementary figure 12.b). The analysis pipeline starts by filtering genes with very low counts. Genes that have counts per million (CPM) values of more than 1 in at least the minimum number of samples in any of the comparison groups are kept for the downstream analysis. We apply trimmed mean of M values (TMM) for normalizing libraries which is a built-in normalization method in *edgeR*. A design matrix is constructed based on the selected variable and groups. We use gene-wise negative binomial generalized log-linear models with quasi-likelihood tests and gene-wise exact tests from Bioconductor package *edgeR* to calculate differential expression between groups with and without covariates respectively. P-values are adjusted for multiple testing correction using Benjamini-Hochberg method. A gene is considered up-regulated in the **‘basal cell population’** group if $\log_2(\text{fold change})$ (Log_FoldChange) is positive and a gene is down-regulated if $\log_2(\text{fold change})$ is negative. You can sort the table by any of the columns or filter the table within a range of Log_FoldChange or Adjusted_pvalue by typing values in the search boxes under each column names.

There are three separate buttons in this tab panel: show heatmap, download signature, and upload signature to iLINCS. A pop-up window of heatmap of the top most differentially expressed genes will appear if you click the **‘Show heatmap’** button (See supplementary figure 12.c). This heatmap shows the change in relative expression of the genes. User can select to show the heatmap across all the samples or the comparison samples only. The **‘Download signature’** button lets you download the signature data table. Finally, pressing **‘upload signature to iLINCS’** (See supplementary figure 12.d) button will take you to iLINCS (<http://www.ilincs.org>), the mother domain of GREIN. Integrative LINCS or iLINCS is an integrative and user-friendly web platform with a number of tools for analysis of LINCS and non-LINCS data and signatures. User can upload or select a signature, find concordant signatures, and analyze them to identify meaningful biological pathways. It is a part of NIH LINCS (<http://www.lincsproject.org/>) common fund program.

Construct differential expression signature for the selected dataset

Selected study

GSE60450

Variable of interest

immunophenotype

Type of comparison

Two group without covariate

Experimental group

basal cell population

Control group

luminal cell population

Generate signature

Power analysis

Create a signature

User specific design

Method of constructing a signature

Genewise negative binomial generalized log-linear models with quasi-likelihood tests and genewise exact tests from bioconductor package edgeR are used to calculate differential expression between groups with and without covariates respectively. Control group is considered as reference in computing log of fold change and p-value is adjusted for multiple comparison.

Metadata

Signature

Show 10 entries

Search:

	Selected groups	immunophenotype
GSM1480297	experimental	basal cell population
GSM1480298	experimental	basal cell population
GSM1480299	experimental	basal cell population
GSM1480300	experimental	basal cell population
GSM1480301	experimental	basal cell population
GSM1480302	experimental	basal cell population
GSM1480291	control	luminal cell population
GSM1480292	control	luminal cell population
GSM1480293	control	luminal cell population
GSM1480294	control	luminal cell population

Showing 1 to 10 of 12 entries

Previous12Next

Supplementary Figure 12.a. Metadata table in the 'Create a signature' tab panel.

Construct differential expression signature for the selected dataset

Selected study

GSE60450

Show heatmap

Download signatures

Upload signatures to iLincs

Power analysis

Create a signature

User specific design

Method of constructing a signature

Genewise negative binomial generalized log-linear models with quasi-likelihood tests and genewise exact tests from bioconductor package edgeR are used to calculate differential expression between groups with and without covariates respectively. Control group is considered as reference in computing log of fold change and p-value is adjusted for multiple comparison.

Metadata

Signature

Show 10 entries

Search:

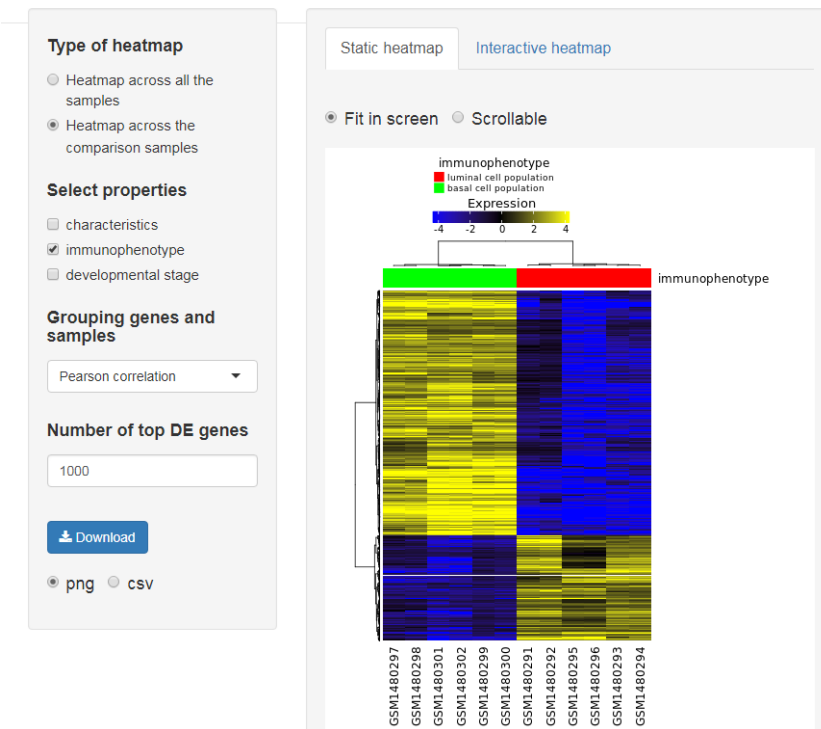
Ensembl_ID	Gene_symbol	Log_FoldChange	Adjusted_pvalue
All	All	All	All
ENSMUSG00000027750	Postn	-7.751	0
ENSMUSG00000021604	Irx4	-8.51	0
ENSMUSG00000022371	Col14a1	-8.226	0
ENSMUSG00000040118	Cacna2d1	-8.685	0
ENSMUSG00000049420	Tmem200a	-8.295	0
ENSMUSG00000061527	Krt5	-9.019	0
ENSMUSG00000002799	Jag2	-6.471	0
ENSMUSG00000024578	Il17b	-8.508	0
ENSMUSG00000058297	Spock2	-6.239	0
ENSMUSG00000034612	Chst11	-6.863	0

Showing 1 to 10 of 12,715 entries

Previous12345...1272Next

Supplementary Figure 12.b. Signature table in the 'Create a signature' tab panel.

Heatmap of top DE genes (ranked by adjusted p-values)



Supplementary Figure 12.c. Heatmap of top 1000 differentially expressed genes.

iLINC Genes Datasets Signatures Maps About Support

Search for signatures / Upload a signature / Uploaded Signature

Uploaded Signature

Session ID: Wed_Jan_3_18_05_50_2018_9384658
File name: GSE60450_signatureData.txt
Found 11375 out of 12722 submitted entries.

Signature Analysis

» Modify the list of selected genes
» Other analyses with selected genes

Show complete signature (11375) Show selected genes (100) Download processed signature Download selected genes

Signature analysis result

Enrichment Analysis L1000CDS2 GeneMANIA Reactome ToppFun

Connected Signatures

Complete signature Selected genes

- ☐ LINCS individual shRNA signatures
- ☐ LINCS consensus (CGS) gene knockdown signatures
- ☐ LINCS chemical perturbagen signatures
- ☐ LINCS RNA-Seq signatures
- ☐ Cancer Therapeutics Response signatures
- ☐ Disease Related signatures
- ☐ ENCODE Transcription Factor Binding signatures
- ☐ Connectivity Map signatures

900
900
900
11290
10827
11320
11303
8164

iLINC (Integrative LINCS) genomics data portal
BD2K-LINCS DATA COORDINATION AND INTEGRATION CENTER
a part of NIH LINCS Program

LINCS Tools
Data Portal Enrichr iLINC Life pLINC

Contact
Support

Supplementary Figure 12.d. Uploaded signature to iLINC portal.

S1.3.3. User specific design

In the ‘Create a signature’ section, user does not have the option to select specific samples from each group or reorganize the samples within groups. For example, in the ‘Create a signature’ section, it is not possible to compare lactating and pregnant samples from the basal population only. The ‘User specific design’ is an extension of the previous section. It provides the flexibility to reconstruct the experimental design by selecting any samples for both experimental and control groups.

After selecting the variable of interest and type of comparison, you can choose samples from the drop-down menus of both experimental and control groups. After forming the design table, you can generate signature by clicking the ‘Generate signature’ button.

Construct differential expression signature for user modified design

Selected study

GSE60450

Variable of Interest

characteristics

Type of comparison

Two group without covariate

Generate signature

Power analysisCreate a signatureUser specific design

Metadata

Construct your own experimental design

This section provides the flexibility of picking the groups or variables you want and modify the design table according to your choice for further analysis.

Select experimental samples

cells, virgin

☐ GSM1480299 : mammary gland, basal cells, 18.5 day pregnancy

☐ GSM1480300 : mammary gland, basal cells, 18.5 day pregnancy

☒ GSM1480301 : mammary gland, basal cells, 2 day lactation

☒ GSM1480302 : mammary gland, basal cells, 2 day lactation

Select control samples

Search:

Selected groupscharacteristics

experimentalmammary gland, basal cells, 2 day lactation

experimentalmammary gland, basal cells, 2 day lactation

controlmammary gland, basal cells, 18.5 day pregnancy

controlmammary gland, basal cells, 18.5 day pregnancy

Showing 1 to 4 of 4 entries

Previous1Next

Supplementary Figure 13.a. Metadata table in the user specific design tab.

Once you click the ‘Generate signature’ button, it will take you the ‘Signature’ tab. Similar to the ‘Create a signature’ tab, you can visualize the top DE genes in a heatmap, download the signature table, and upload the signature to iLINCS.

Construct differential expression signature for user modified design

Selected study

GSE60450

Show heatmap

Download signature

Upload signature to iLINCS

Power analysisCreate a signatureUser specific design

MetadataSignature

Construct your own experimental design

This section provides the flexibility of picking the groups or variables you want and modify the design table according to your choice for further analysis.

Show10entries

Search:

Ensembl_IDGene_symbolLog_FoldChangeAdjusted_pvalue

AllAllAllAll

ENSMUSG000000061388Csn1s2b6.2270

ENSMUSG000000047501Cldn4-5.3370

ENSMUSG000000046623Gjb4-4.5450

ENSMUSG000000031070Mrgprf5.1280

ENSMUSG000000042367Gjb3-3.5190

ENSMUSG000000031097Tnni25.6220

ENSMUSG000000014599Csfl-3.0160

ENSMUSG000000024164C3-4.0710

ENSMUSG0000000260511500015O10RIk-2.8210

ENSMUSG000000026147Col9a1-2.9240

Showing 1 to 10 of 12,871 entries

Previous12345...1288Next

Supplementary Figure 13.b. Signature table in the user specific design tab.

Table 1. List of R and Bioconductor packages included in GREIN with corresponding version.

R/Bioconductor packages	Version	References
ComplexHeatmap	1.17.1	Gu <i>et al.</i> , 2016
edgeR	3.20.8	Robinson <i>et al.</i> , 2010
GEOquery	2.46.14	Davis and Meltzer, 2007
iheatmapr	0.4.3	Schep and Kummerfeld, 2017
plotly	4.7.1	Sievert <i>et al.</i> , 2016
rgl	0.99.9	Adler <i>et al.</i> , 2017
RNASeqPower	1.18.0	Hart <i>et al.</i> , 2013
Rtsne	0.13	Krijthe <i>et al.</i> , 2015
shiny	1.0.5	Chang <i>et al.</i> 2015
tximport	1.6.0	Soneson <i>et al.</i> , 2015

References

- Adler,D. *et al.* (2017) rgl: 3D Visualization Using OpenGL. R package version 0.98.1. <https://CRAN.R-project.org/package=rgl>.
- Bernstein,M.N. *et al.* (2017) MetaSRA: normalized human sample-specific metadata for the Sequence Read Archive. *Bioinformatics*, **33**, 2914-2923.
- Bolger,A. *et al.* (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114-2120.
- Chang,W. *et al.* (2015) Shiny: web application framework for R. *R package version 0.11*, **1**, 106.
- Davis,S. and Meltzer,P.S. (2007) GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, **23**, 1846-1847.
- Ewels,P. *et al.* (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, **32**, 3047-3048.
- Fu,N.Y. *et al.* (2015) EGF-mediated induction of Mcl-1 at the switch to lactation is essential for alveolar cell survival. *Nature cell biology*, **17**, 365.
- Gu, Z. *et al.* (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, **32**, 2847-2849.
- Hart,S.N. *et al.* (2013) Calculating sample size estimates for RNA sequencing data. *Journal of computational biology*, **20**, 970-978.
- Krijthe,J.H. *et al.* (2015) Rtsne: T-distributed stochastic neighbor embedding using Barnes-Hut implementation. R package version 0.13. <https://github.com/jkrijthe/Rtsne>
- Patro,R. *et al.* (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*, **14**, 417.
- Robinson,M.D. *et al.* (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139-140.
- Schep,A.N. and Kummerfeld,S.K. (2017) iheatmapr: Interactive complex heatmaps in R. *The Journal of Open Source Software*, **2**, 359.
- Sievert, C. *et al.* (2017) plotly: Create Interactive Web Graphics via ‘plotly.js’. R package version 4.7.1. <https://CRAN.R-project.org/package=plotly>.
- Soneson,C. *et al.* (2015) Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, **4**.
- Tarr,G. (2015) pairsD3: D3 Scatterplot Matrices. R package version 0.1.0. <https://CRAN.R-project.org/package=pairsD3>.