Dprofiler Documentation

Release 1.0.0

Artur Manukyan, Alper Kucukural, Manuel Garber, Onur Yukselen

Nov 01, 2021

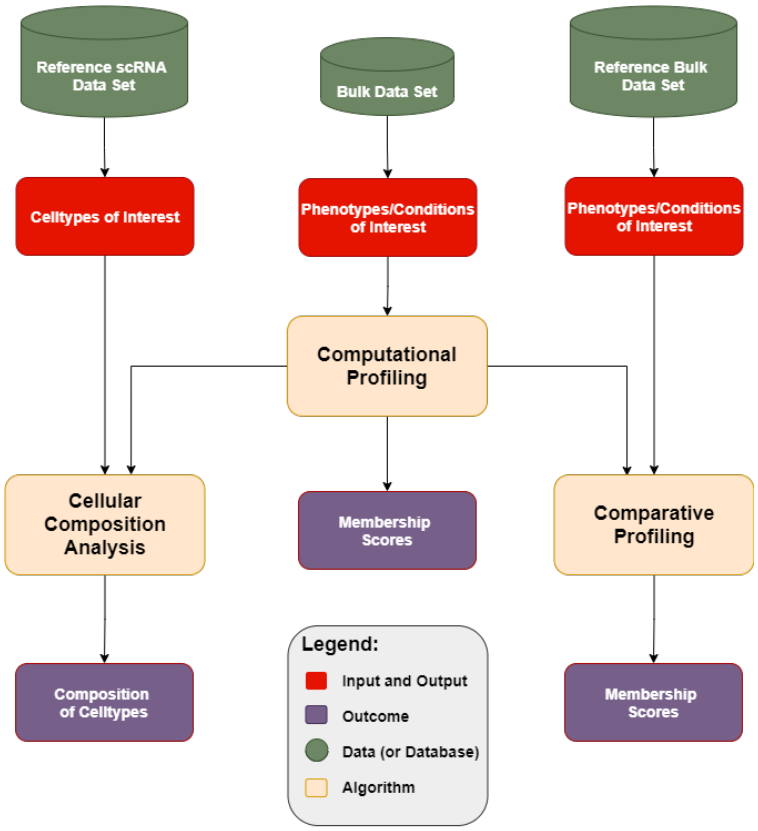
CONTENTS

What is Dprofiler ?

Biospecimen collected from multiple sources of experimental and technical conditions often exhibit **diverse molecular profiles and patterns**. Each individual sample presents an additional source of information towards elucidating biological mechanisms, and each sample may even be more informative than the differential expression between phenotypic conditions

Understanding complex patterns of expression profiles are essential to solving diseases, and such knowledge can only be generated by using advanced statistical measures and methods that are capable of computationally modeling and **separately investigating molecular profiles of each sample**.

To deciphering these complex molecular profiles within gene expression datasets, we have developed **Dprofiler**.



This application computationally profiles a set of targeted samples by connecting them to reference expression datasets with phenotypic profiles of interest. Building on these reference profiles of phenotypic groups, Dprofiler evaluates bulk RNA samples, detect possibly existing anomalies and heterogeneities, and further explores causes and sources of distinct molecular patterns with the aid of single cell maps and other reference gene expression datasets.

Users are allowed to use a variety of algorithms to calculate a **Membership Score** of samples associated to some phenotypic profiles of interest. Dprofiler derives reference phenotypic profiles from submitted datasets, cell-types of single cell maps and conditions from external gene expression data sets. Membership scores are universally interpretable, and indicate the similarity of a sample to these reference profiles.

Dprofiler offers multiple capabilities using these three components:

- **Computational Profiling**: Scoring and Profiling submitted samples using homogeneouos subpopulations of phenotypic reference profiles within the same dataset.
- **Cellular Composition Analysis**: Infering the cellular composition of each scored sample with a reference scRNA data and estimate fractions of cellular compositions.

- **Comparative Profiling**: Scoring and Profiling samples using targeted phenotypic profiles of a reference bulk expression data set.

Dprofiler is based on a recently developed Shiny (R) application, DEBrowser, an interactive tool for DE analysis and visualization. DEBrowser incorporates DESeq2, EdgeR, and Limma coupled with shiny to produce real-time changes within your plot queries and allows for interactive browsing of your DE results. DEBrowser also manipulates your results in a way that allows for interactive plotting by which changing padj or fold change limits also changes the displayed graph(s).

Contents:

CHAPTER

ONE

QUICK-START GUIDE

This guide is walkthrough for the Dprofiler from start to finish.

1.1 Getting Started

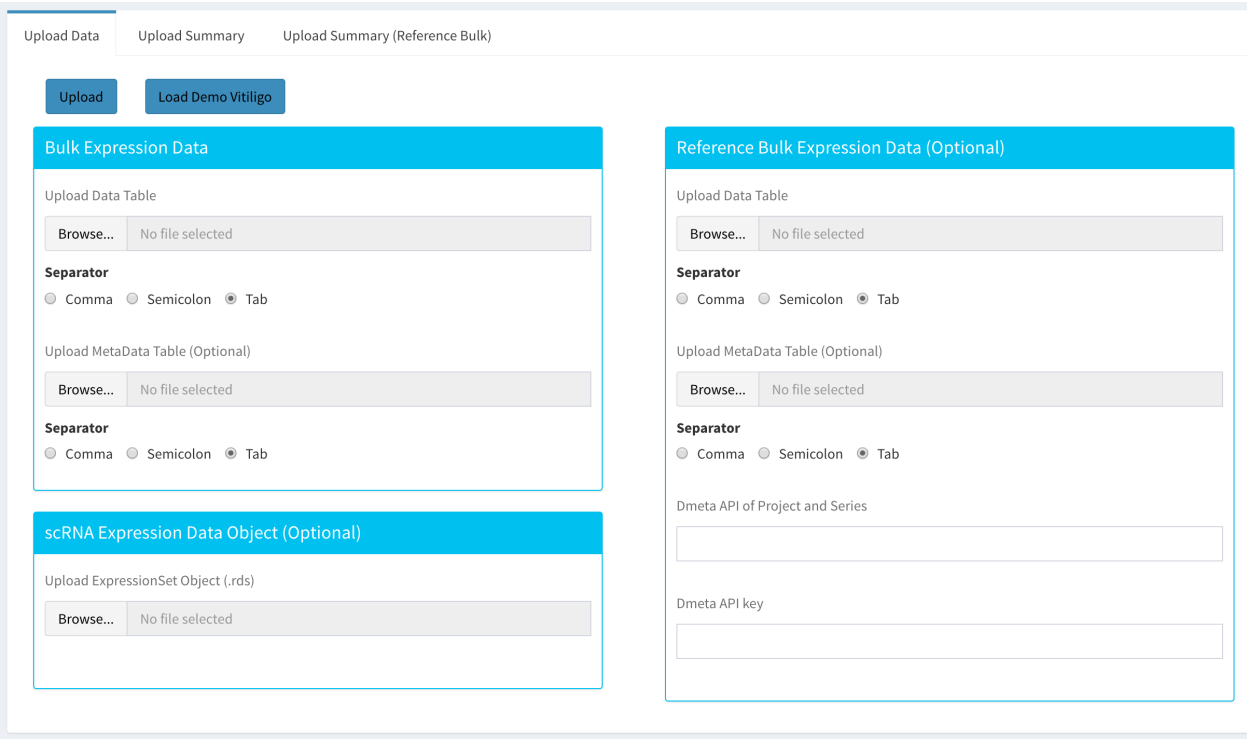
First off, we need to install R package of Dprofiler from GitHub:

```
if (!requireNamespace("remotes", quietly=TRUE))
  install.packages("remotes")
remotes::install_github("UMMS-Biocore/dprofiler")
```

One you have installed the R package, you can call these R commands:

```
library(dprofiler)
startDprofiler()
```

Once you've made your way to the website, or you have a local instance of Dprofiler running, you will be greeted with data loading section:



To begin the analysis, you need to upload your data files (comma or semicolon separated, i.e. ".csv", or tab separated, i.e. ".tsv", format) to be analyzed and choose appropriate separator for the file (comma, semicolon or tab).

There are three types input data in Dprofiler. These are:

- **Bulk Expression Data**: used for profiling samples within and establish homogeneous reference profiles.
- **scRNA Expression Data Object**: used for deconvoluting Bulk RNA data and infering cellular compositions of these bulk samples.
- **Reference Bulk Expression Data**: used for comparative profiling of the Bulk data set using reference phenotypic profiles of reference bulk data set(s).

If you do not have a dataset to upload, you can use the built in demo data file by clicking on the 'Load Demo Vitiligo' button that loads a case study. To view the entire demo data file, you can download.

- PRJNA554241: a bulk RNA-Seq count data of lesional and non-lesional vitiligo skin samples processed by the RNA-Seq pipeline of DolphinNext .
- scVitiligo: a reference scRNA-Seq count data of lesional and non-lesional vitiligo skin samples.
- GSE65127: an external bulk microarray data of vitiligo skin samples where the gene set is union of differentially expressed genes across four conditions: healthy, lesional, non-lesional and peri-lesional.

Otherwise, you can start uploading your own data given instructions below.

1.2 Input for Bulk Expression Data

For both submitted Bulk and Reference Bulk expression data, you need to upload your data files (comma or semicolon separated, i.e. ".csv", or tab separated, i.e. ".tsv", format) to be analyzed and choose appropriate separator for the file (comma, semicolon or tab). However, for scRNA data set, the user should provide an ExpressionSet object. In addition, users may connect to DolphinMeta using their credentials to import reference bulk expression data from any Dmeta project.

An example structure of the count data files are shown below:

gene	P39_NL	P39_L	P33_NL	P33_L	P22_NL	P22_L	P19_NL	P19_L	P65_NL	P65_L
A1BG	46	29	104	27	42	17	65	101	27	32
A1BG-AS1	27	18	48	13	10	3	39	54	23	24
A1CF	5	0	38	15	2	0	4	7	0	3

In addition to the count data file; you might need to upload metadata files to correct for batch effects or any other normalizing conditions you might want to address that might be within your results. To handle for these conditions, simply create a metadata file by using the example table at below. Metadata file also simplifies condition selection for complex data. The columns you define in this file can be selected in condition selection page. Make sure you have defined two conditions per column. If there are more than two conditions in a column, those can be defined empty. Please note that, if your data is not complex, metadata file is optional, you don't need to upload.

In the example below, the "patient" column may serve as a batch or a normalizing condition.

sample	patient	treatment
P39_NL	P39	NL
P39_L	P39	L
P33_NL	P33	NL
P33_L	P33	L
P22_NL	P22	NL
P22_L	P22	L
P19_NL	P19	NL
P19_L	P19	L
P65_NL	P65	NL
P65_L	P65	L

Metadata file can be formatted with comma, semicolon or tab separators similar to count data files. These files used to establish different batch effects for multiple conditions. You can have as many conditions as you may require, as long as all of the samples are present.

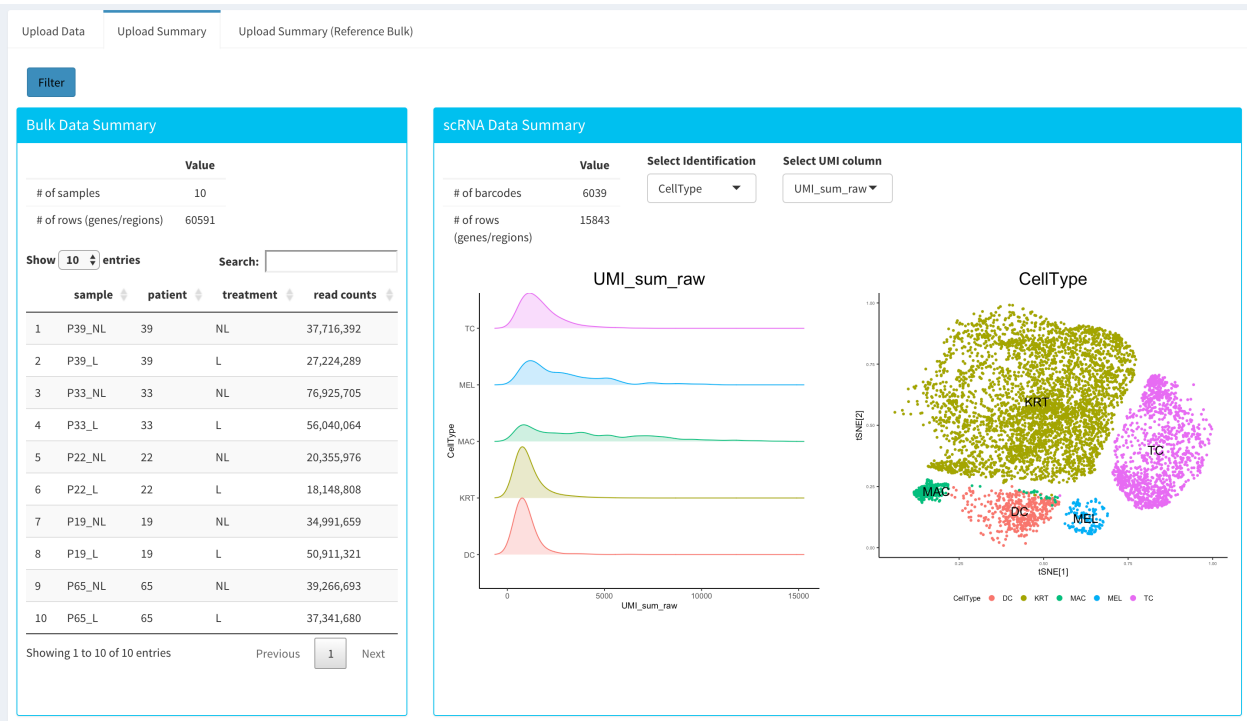
Alternatively, you can connect to your DolphinMeta (Dmeta) account using your access token and import external expression profiles and associated metadata.

1.3 Input for scRNA Expression Data

However, for scRNA data set, the user should provide an .rds file containing an ExpressionSet object whose metadata (pData(*You Expression Set Object*)) should include the following variables or columns:",

- **(i)** sample associated to each barcode
- **(ii)** total UMI counts of each barcode
- **(iii)** cell annotation or label of each barcodes
- **(iv)** other categorical and numerical variables relavant to barcodes

Once the count data and metadata files have been loaded in Dprofiler, you can click upload button to visualize your data as shown below:



You have the option to search samples or other terms within submitted or reference bulk Expression data sets, and you also have the option to visualize the t-SNE and other numeric measures of your barcodes within the uploaded scRNA expression data object.

After reviewing your uploaded data in "Upload Summary" panels, and if specified the metadata file containing your batch correction fields, you then have the option to filter low counts and conduct batch effect correction prior to your analysis. Alternatively, you may skip these steps and directly continue with Computational Profiling analysis.

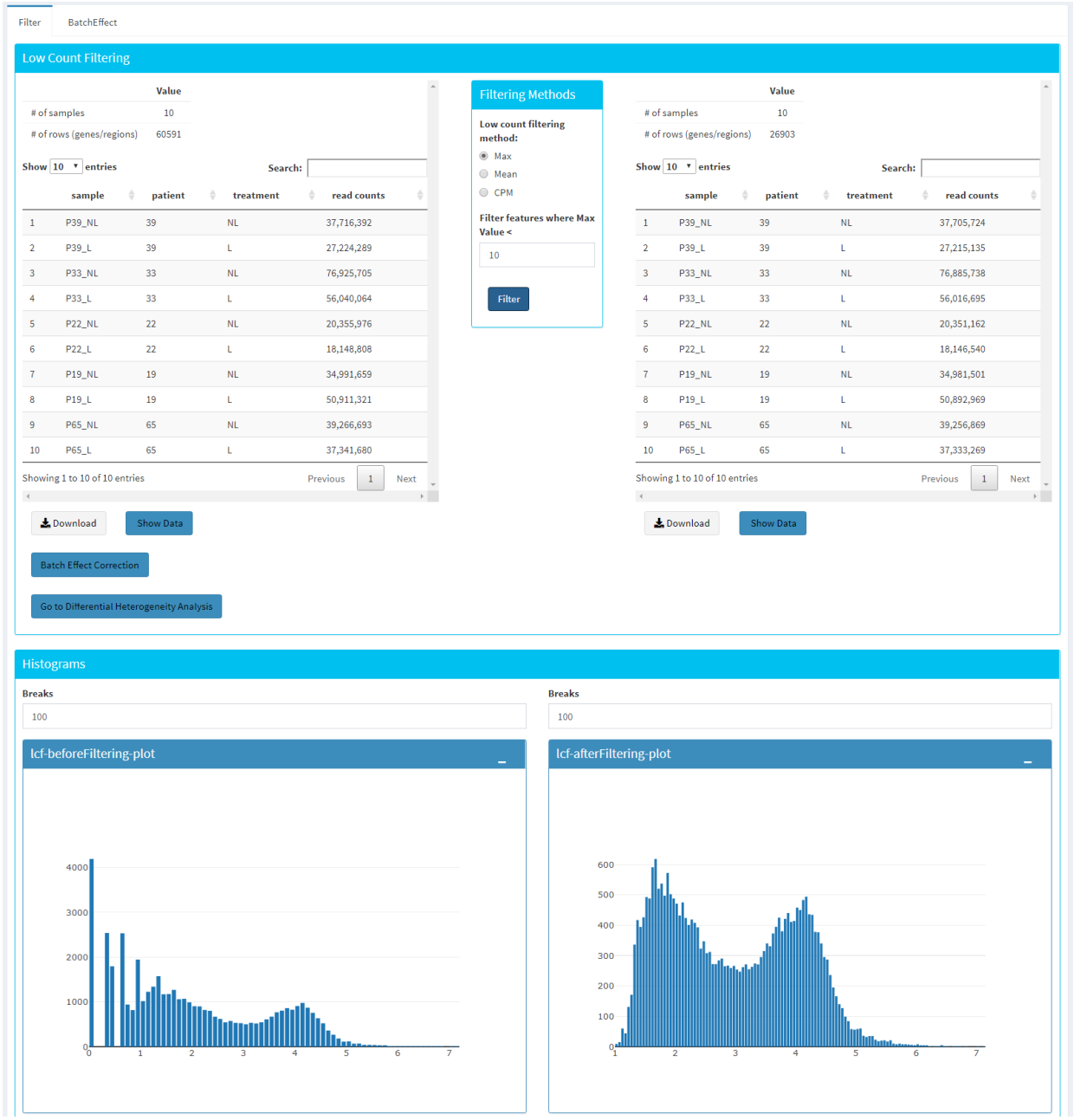
Data analysis steps such as "Low Count Filtering", "Batch Effect Correction", "Computational Profiling" are only applicable to the submitted bulk expression data, and other submitted reference scRNA and bulk RNA datasets are used for referential purposes and assumed to be already filtered and analyzed before submission.

1.4 Low Count Filtering

In this section, you can simultaneously visualize the changes of your submitted bulk RNA expression data while filtering out the low count genes. Choose your filtration criteria from **Filtering Methods** box which is located just center of the screen. Three methods are available to be used:

- **Max:** Filters out genes where maximum count for each gene across all samples are less than defined threshold.
- **Mean:** Filters out genes where mean count for each gene are less than defined threshold.
- **CPM:** First, counts per million (CPM) is calculated as the raw counts divided by the library sizes and multiplied by one million. Then it filters out genes where at least defined number of samples is less than defined CPM threshold.

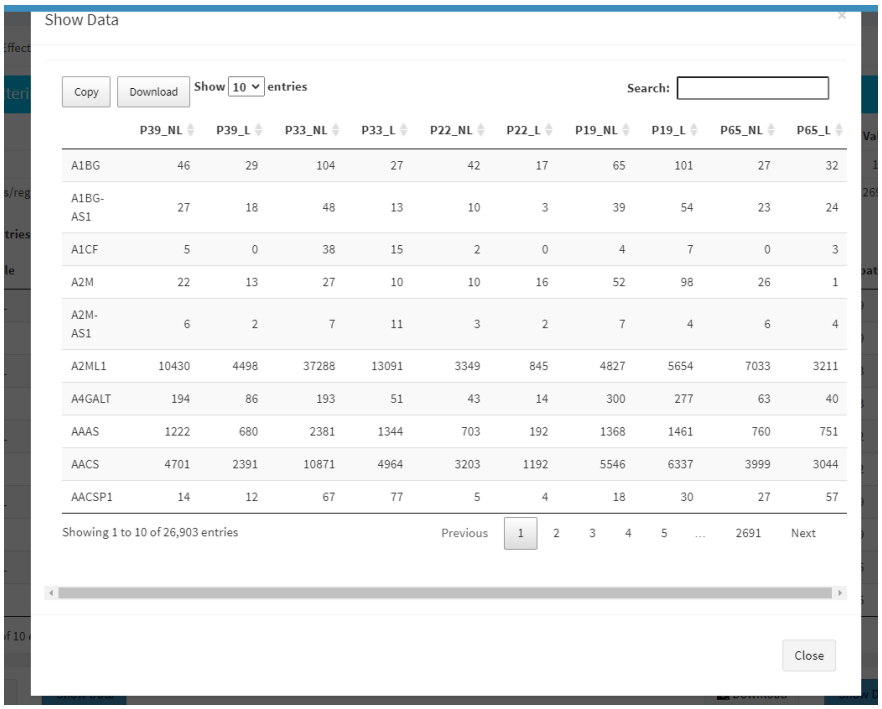
After selection of filtering methods and entering threshold value, you can proceed by clicking **Filter** button which is located just bottom part of the **Filtering Methods** box. On the right part of the screen, your filtered dataset will be visualized for comparison as shown at figure below.



You can easily compare following features, before and after filtering:

- Number of genes/regions.
- Read counts for each sample.
- Overall histogram of the dataset.
- gene/region vs samples data

Important: To investigate the gene/region vs samples data in detail as shown at below, you may click the **Show Data** button, located bottom part of the data tables. Alternatively, you may download all filtered data by clicking **Download** button which located next to **Show Data** button.



	P39_NL	P39_L	P33_NL	P33_L	P22_NL	P22_L	P19_NL	P19_L	P65_NL	P65_L
A1BG	46	29	104	27	42	17	65	101	27	32
A1BG-AS1	27	18	48	13	10	3	39	54	23	24
A1CF	5	0	38	15	2	0	4	7	0	3
A2M	22	13	27	10	10	16	52	98	26	1
A2M-AS1	6	2	7	11	3	2	7	4	6	4
A2ML1	10430	4498	37288	13091	3349	845	4827	5654	7033	3211
A4GALT	194	86	193	51	43	14	300	277	63	40
AAAS	1222	680	2381	1344	703	192	1368	1461	760	751
AACS	4701	2391	10871	4964	3203	1192	5546	6337	3999	3044
AACSP1	14	12	67	77	5	4	18	30	27	57

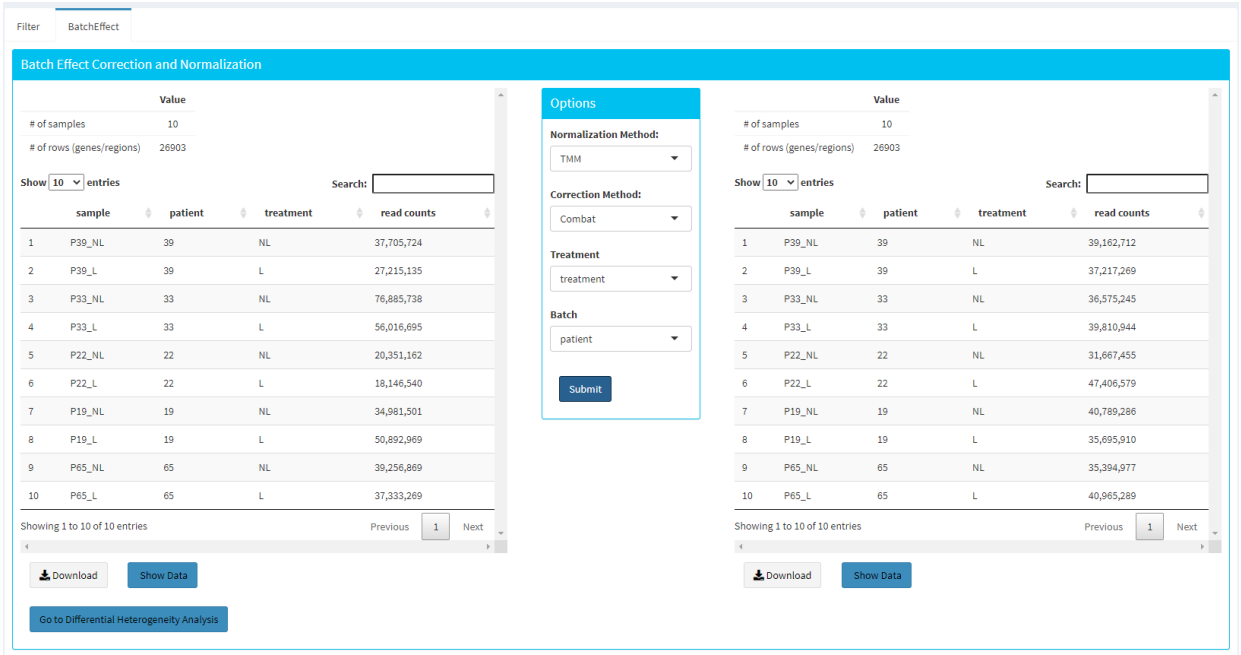
Afterwards, you may continue your analysis with **Batch Effect Correction** or directly jump to **Computational Profiling** of your dataset.

1.5 Batch Effect Correction and Normalization

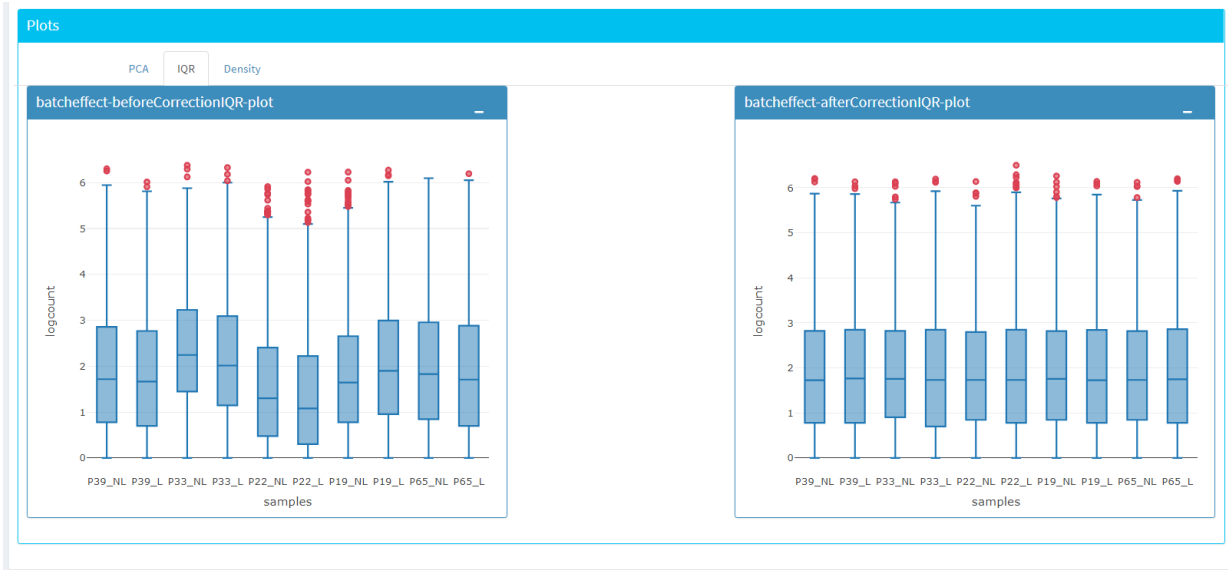
If specified metadata file containing your batch correction fields, then you have the option to conduct batch effect correction prior to your analysis. By adjusting parameters of **Options** box, you can investigate your character of your dataset. These parameters of the options box are explained as following:

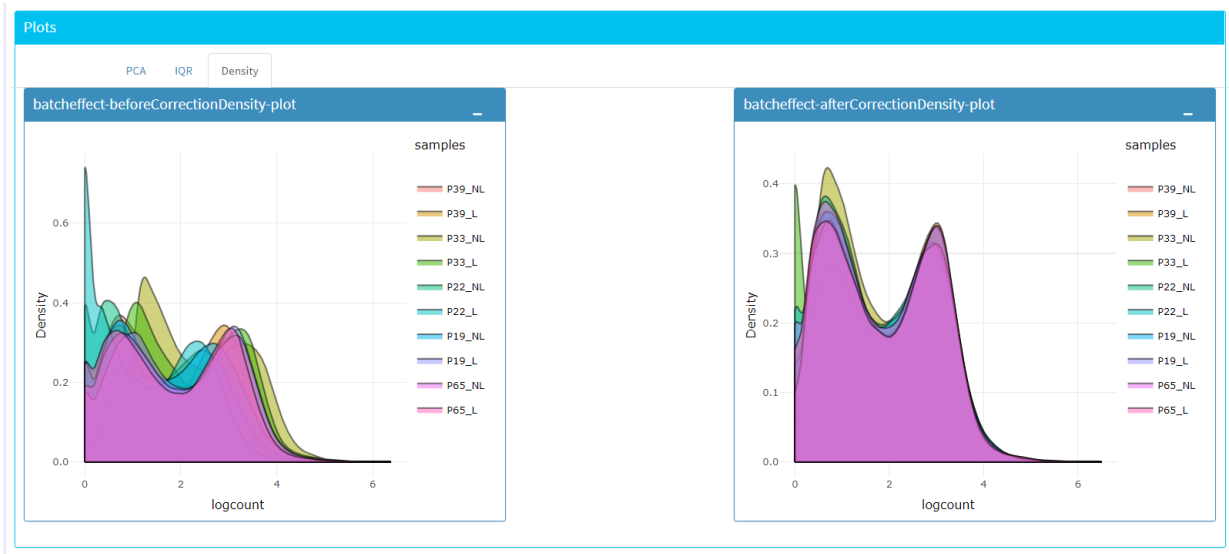
- **Normalization Method:** Dprofiler allows performing normalization prior the batch effect correction. You may choose your normalization method (among MRN (Median Ratio Normalization), TMM (Trimmed Mean of M-values), RLE (Relative Log Expression) and upperquartile), or skip this step by choosing **none** for this item.
- **Correction Method:** Dprofiler uses ComBat (part of the SVA bioconductor package) or Harman to adjust for possible batch effect or conditional biases.
- **Treatment:** Please select the column that is specified in metadata file for phenotypic comparisons, such as cancer vs control.
- **Batch:** Please select the column name in metadata file which differentiate the batches.

Upon clicking submit button, comparison tables and plots will be created on the right part of the screen as shown below.


Filter
BatchEffect
Batch Effect Correction and Normalization
Value
# of samples 10
# of rows (genes/regions) 26903
Show 10 entries
Search:
sample patient treatment read counts
1 P39_NL 39 NL 37,705,724
2 P39_L 39 L 27,215,135
3 P33_NL 33 NL 76,885,738
4 P33_L 33 L 56,016,695
5 P22_NL 22 NL 20,351,162
6 P22_L 22 L 18,146,540
7 P19_NL 19 NL 34,981,501
8 P19_L 19 L 50,892,969
9 P65_NL 65 NL 39,256,869
10 P65_L 65 L 37,333,269
Showing 1 to 10 of 10 entries
Previous 1 Next
Download
Show Data
Go to Differential Heterogeneity Analysis
Options
Normalization Method:
TMM
Correction Method:
Combat
Treatment
treatment
Batch
patient
Submit
Value
# of samples 10
# of rows (genes/regions) 26903
Show 10 entries
Search:
sample patient treatment read counts
1 P39_NL 39 NL 39,162,712
2 P39_L 39 L 37,217,269
3 P33_NL 33 NL 36,575,245
4 P33_L 33 L 39,810,944
5 P22_NL 22 NL 31,667,455
6 P22_L 22 L 47,406,579
7 P19_NL 19 NL 40,789,286
8 P19_L 19 L 35,695,910
9 P65_NL 65 NL 35,394,977
10 P65_L 65 L 40,965,289
Showing 1 to 10 of 10 entries
Previous 1 Next
Download
Show Data


Plots
PCA
IQR
Density
PCA Plot
PC1 (64.91%)
PC2 (12.92%)
Loadings
explained
PCs
PCA Controls
PC selection[x]
1
PC selection[y]
2
Legend On/Off
On
Text On/Off
On
Select legend
color
Color field
None
Shape field
None
PCA Plot
PC1 (53.63%)
PC2 (14.74%)
Loadings
explained
PCs





You can investigate the changes on the data by comparing following features:

- Read counts for each sample.
- PCA, IQR and Density plot of the dataset.
- Gene/region vs samples data

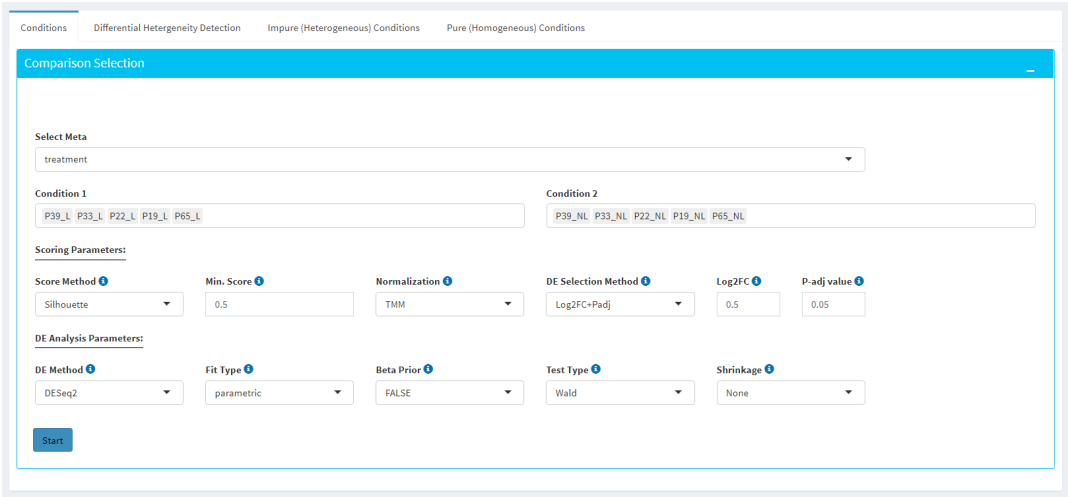
Tip: You can investigate the gene/region vs samples data in detail by clicking the **Show Data** button, or download all corrected data by clicking **Download** button.

Since we have completed **batch effect correction and normalization** step, we can continue with 'Go to Computational Profiling'. This takes you to page where computational profiling is conducted with popular DE analysis methods like DESeq2, EdgeR or Limma.

1.6 Computational Profiling

The first option, 'Go to Computational Profiling', takes you to the next step where an iterative differential expression analysis and scoring of samples takes place.

- **Sample Selection:** In order to run the analysis, you first need to select the initial set of samples which will be compared or may be removed throughout the analysis. To do so, choose **Select Meta** box as **treatment** to simplify fill `Condition 1` and `Condition 2` based on the **treatment** column of the metadata as shown below.



If you need to remove samples from a condition, simply select the sample you wish to remove and hit the delete/backspace key. In case, you need to add a sample to a condition you can click on one of the condition text boxes to bring up a list of samples and then click on the sample you wish to add from the list and it will be added to the textbox for that comparison.

Scoring Parameters: Two scoring methods are available for Dprofiler: Silhouette and NNLS-based.

- Silhouette method incorporates Spearman correlation measures between samples of the same phenotypic condition to estimate the magnitude of similarity between a particular sample and all other samples in the same group.
- NNLS-based method fits a non-negative regression model with a sample being the response and condition-specific (mean) expression profiles of conditions are input variables.

Both methods produce scores between (0,1) where lower values are associated with low membership score indicating that the sample is dissimilar to other samples in the same group/phenotype/condition. You can determine a threshold for low membership scores from the **Min. Score** option which is between (0,1). You can also determine additional criteria for selecting differentially expressed genes by **DE Selection Method** where additional options are provided to choose thresholds for parameters such as **log2FC** and **P-adj value**.

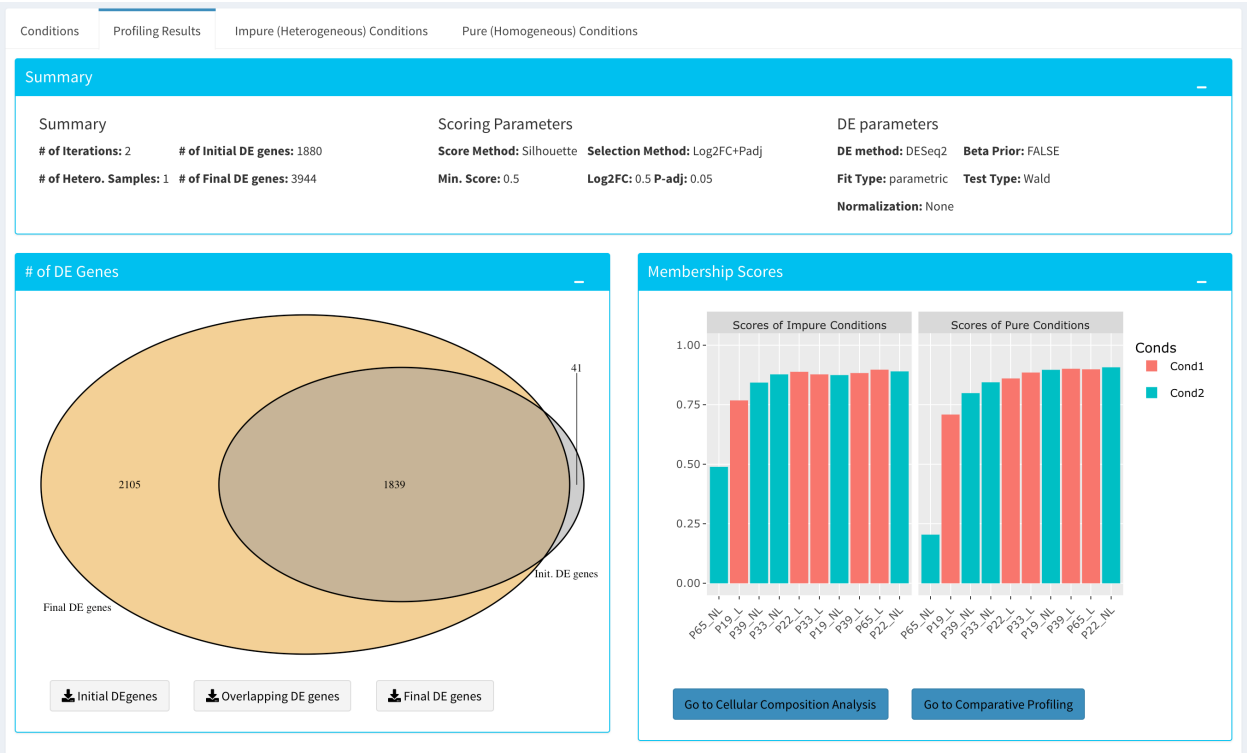
DE Parameters: Thera are three DE methods that are available for Dprofiler: DESeq2, EdgeR, and Limma. DESeq2 and EdgeR are designed to normalize count data from high-throughput sequencing assays such as RNA-Seq. On the other hand, Limma is a package to analyse of normalized or transformed data from microarray or RNA-Seq assays. Upon selecting any of three DE analysis methods, additional options will appear for parameters specific to the selected DE method.

After clicking on the 'start' button, Dprofiler will analyze your selected comparison and conditions, and store the results into separate data tables. Upon finishing the Computational Profiling, three separate results panels will be produced:

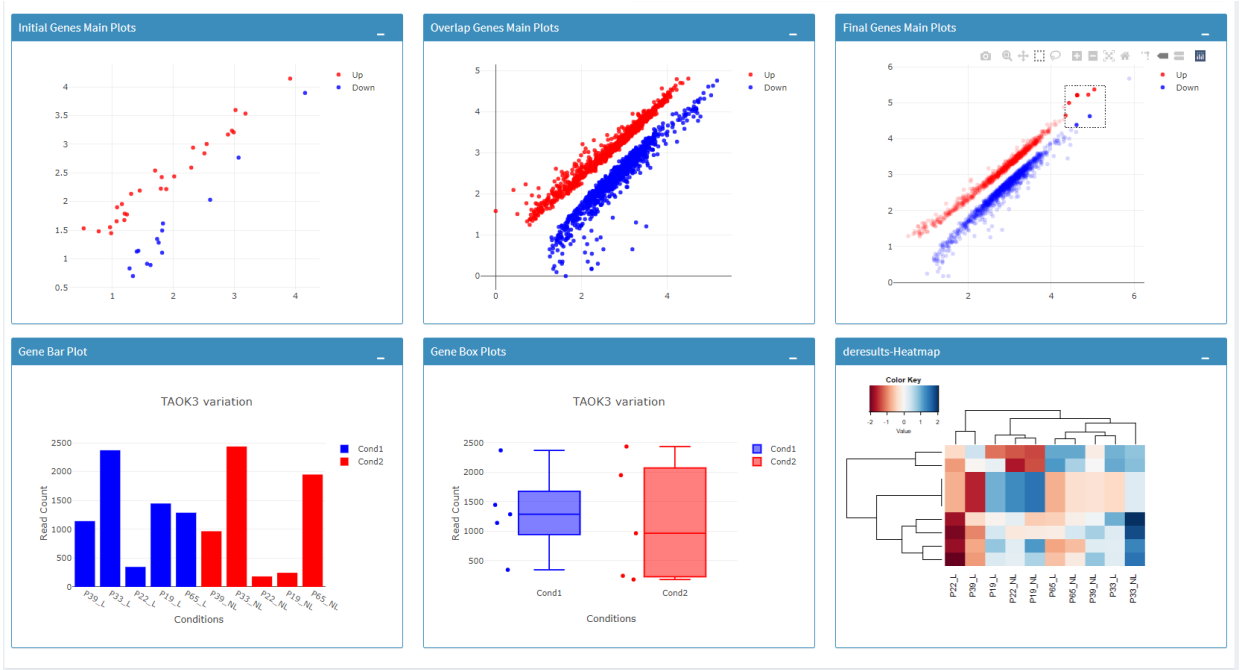
- Profiling Results

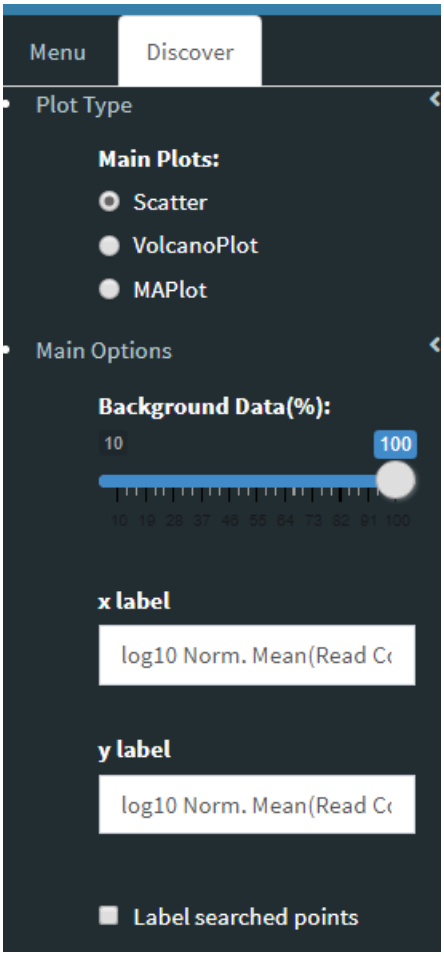
- Impure (Heterogeneous) Conditions
- Pure (Homogeneous) Conditions

Upon finishing the Computational Profiling, the application will switch to "Profiling Results" panel showing results of the analysis. Differentially expressed genes of initial DE analysis and Final DE analysis are compared: that is the number of DE genes at the analysis at the first and last iteration are compared. The app also informs you about the parameters of the Scoring and DE analysis.



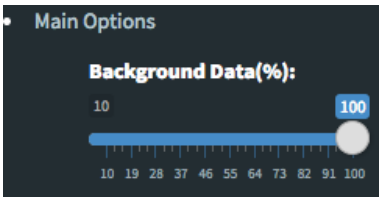
Additional information of initial and final DE genes can be found on plots below. Three **Scatter Plots** of initial and final genes, as well as the common genes in both list of DE genes will be plotted. You can switch to **Volcano Plot** and **MA Plot** by using **Plot Type** section at the left side of the *Discover** menu. Since these plots are interactive, you can click to **zoom** button on the top of the graph and select the area you would like to zoom in by drawing a rectangle. Please see the plots at below:





You can hover over the scatterplot points to display more information about the point selected. A few bargraphs will be generated for the user to view as soon as a scatterplot point is hovered over.

Tip: Please keep in mind that to increase the performance of the generating graph, by default 10% of non-significant(NS) genes are used to generate plots. You might show all NS genes by please click **Main Options** button and change Background Data(%) to 100% on the left sidebar.

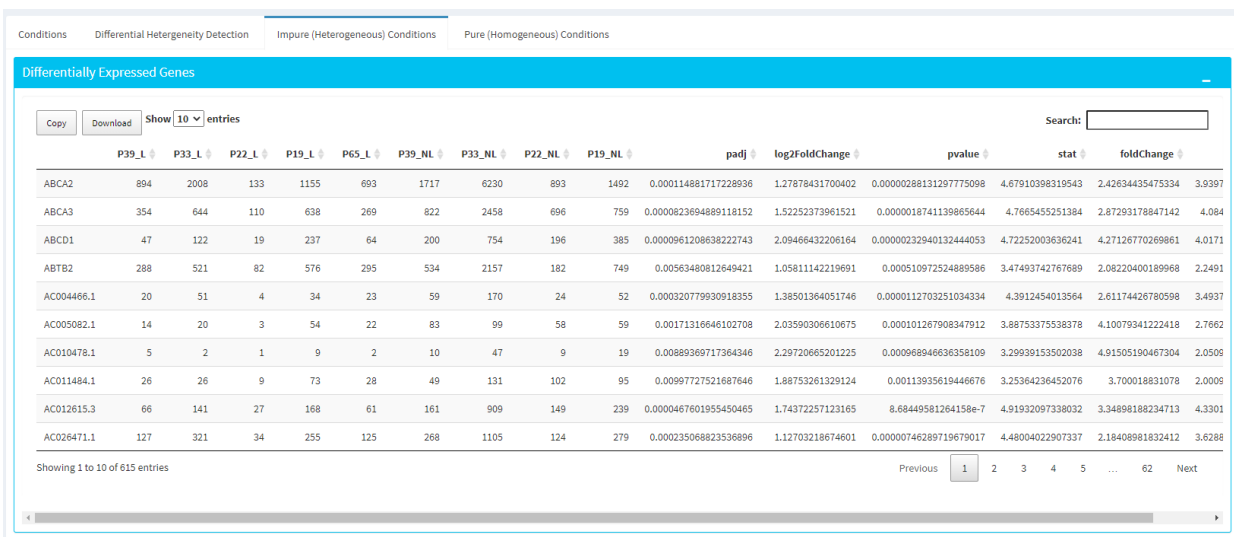


Next, you can initiate a Cellular composition analysis using either the Homogeneouos, Heterogeneous conditions or marker genes, and deconvolute the Reference bulk expression data using the reference scRNA expression data by clicking "Go to Cellular Composition Analysis". Or, you can click to "Go to Comparative Profiling" for the comparative analysis between the submitted bulk RNA expression and reference bulk RNA data.

But before that, you can take a look and investigate DE genes of either initial or Final DE analysis from remaining panels.

1.7 Impure and Pure Conditions

There are two more panels on the right of Profiling Results panel which take a closer look at initial and final DE genes of the conditions.

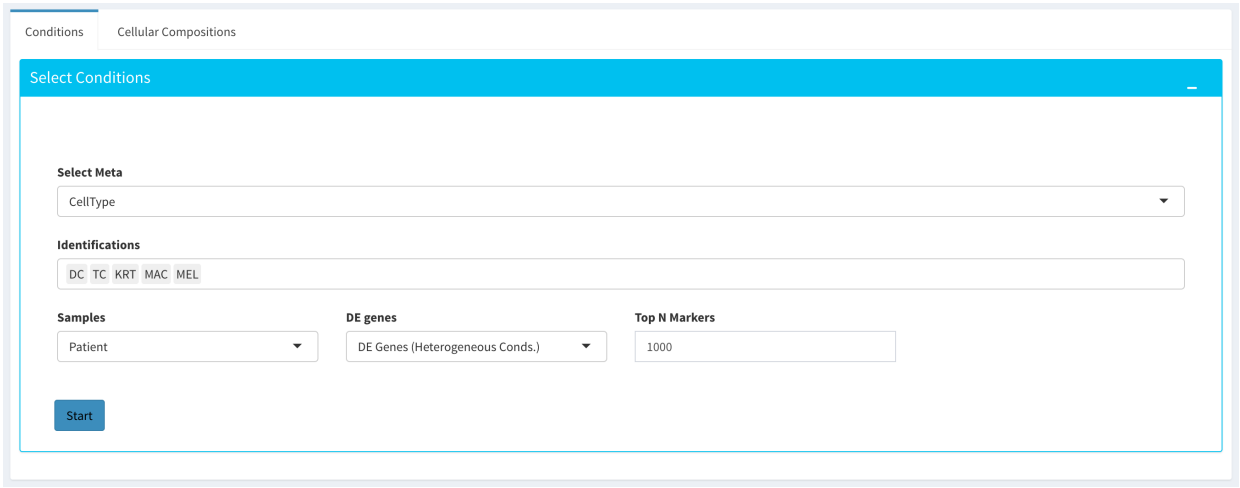


	P39_L	P33_L	P22_L	P19_L	P65_L	P39_NL	P33_NL	P22_NL	P19_NL	padj	log2FoldChange	pvalue	stat	foldChange
ABCA2	894	2008	133	1155	693	1717	6230	893	1492	0.000114881717228936	1.27878431700402	0.00000288131297775098	4.67910398319543	2.42634435475334	3.9397
ABCA3	354	644	110	638	269	822	2458	696	759	0.000082369488911815 2	1.52252373961521	0.0000018741139865644	4.7665455251384	2.87293178847142	4.084
ABCD1	47	122	19	237	64	200	754	196	385	0.0000961208638222743	2.09466432206164	0.00000232940132444053	4.72252003636241	4.27126770269861	4.0171
ABTB2	288	521	82	576	295	534	2157	182	749	0.00563480812649421	1.05811142219691	0.000510972524889586	3.47493742767689	2.08220400189968	2.2491
AC004466.1	20	51	4	34	23	59	170	24	52	0.000320779930918355	1.38501364051746	0.0000112703251034334	4.3912454013564	2.61174426780598	3.4937
AC005082.1	14	20	3	54	22	83	99	58	59	0.00171316646102708	2.03590306610675	0.000101267908347912	3.88753375538378	4.10079341222418	2.7662
AC010478.1	5	2	1	9	2	10	47	9	19	0.00889369717364346	2.29720665201225	0.000968946636358109	3.29939153502038	4.91505190467304	2.0509
AC011484.1	26	26	9	73	28	49	131	102	95	0.00997727521687646	1.88753261329124	0.00113935619446676	3.25364236452076	3.700018831078	2.0009
AC012615.3	66	141	27	168	61	161	909	149	239	0.0000467601955450465	1.74372257123165	8.68449581264158e-7	4.91932097338032	3.34898188234713	4.3301
AC026471.1	127	321	34	255	125	268	1105	124	279	0.000235068823536896	1.12703218674601	0.000007462897196790 17	4.48004022907337	2.18408981832412	3.6288

You can always download these results in CSV format by clicking the **Download** button. You can also download the plot or graphs by clicking on the **download** button at top of each plot or graph.

1.8 Cellular Composition Analysis

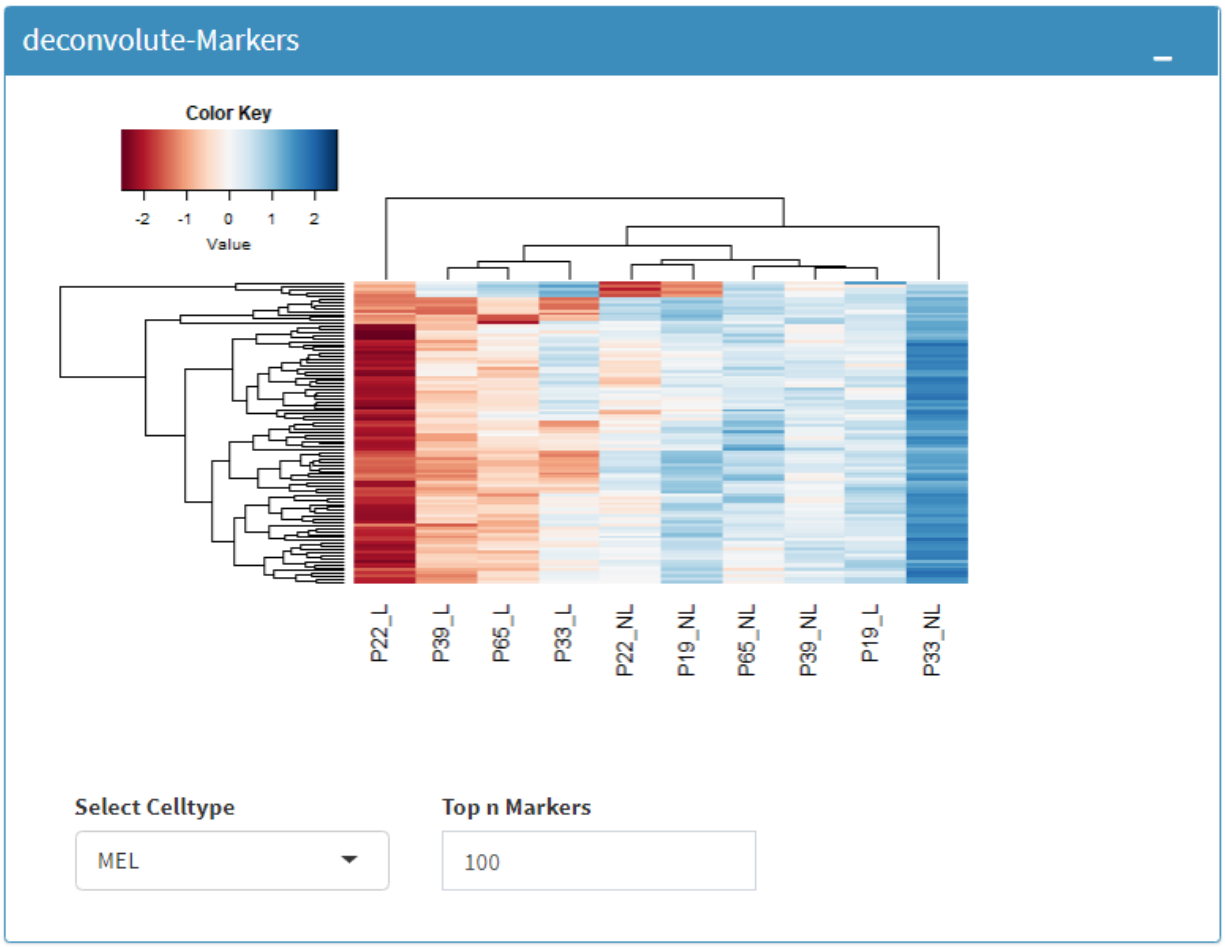
By using the "Cellular Comp. Analysis" tab, you can determine which idents (or identifications, categories) are to be used to deconvolute the submitted bulk expression data. You can also choose which of those cell types within each ident are to be used for the deconvolution as well. Then you can also decide whether DE genes of initial or final DE analysis are used to deconvolute the data. You should decide which column in the scRNA metadata that the samples are introduced, this is required by the MUSIC algorithm to give weight to genes that are less variant across different samples. You can also determine the set of genes to deconvolute bulk samples where you can either use DE genes of impure or pure conditions associated to the initial and final DE analysis, or you can use the marker genes of all selected cell types stored in fData(*You scRNA Expression object*).



After clicking the "Start" button, the results will be given in the "Cellular Compositions" panel. Membership Scores and estimated cell type fractions are given for each sample where each box of the table are highlighted with respect to cell type.

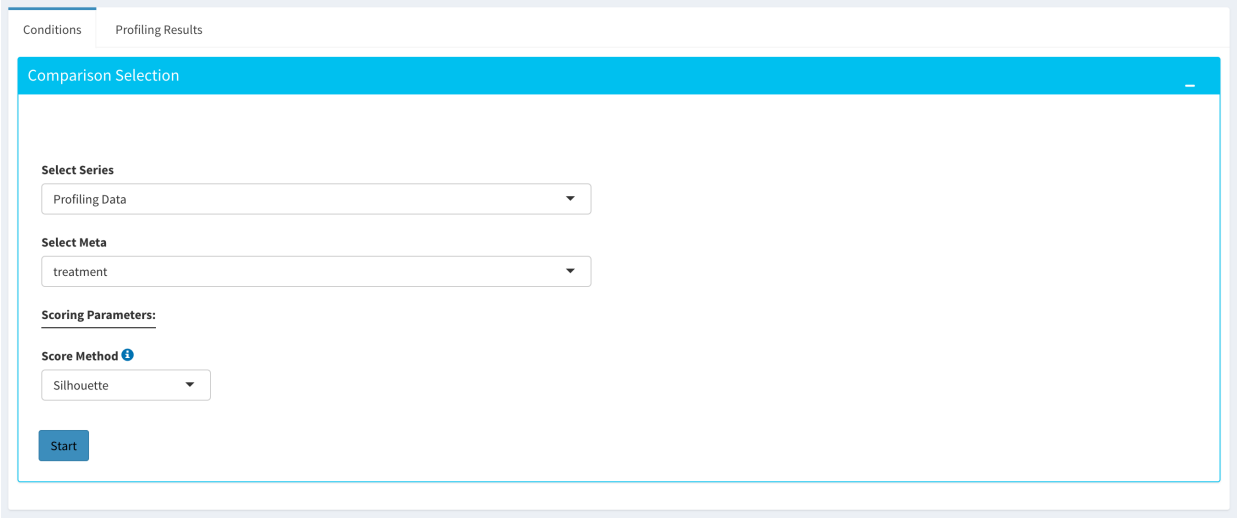
Samples	Conds	Scores	DC	TC	KRT	MAC	MEL
P39_L	Cond1	0.901	0.159	0.183	0.436	0.063	0.159
P33_L	Cond1	0.885	0.034	0.188	0.462	0.099	0.216
P22_L	Cond1	0.860	0.178	0.241	0.503	0.032	0.046
P19_L	Cond1	0.709	0.005	0.521	0.198	0.065	0.211
P65_L	Cond1	0.898	0.150	0.155	0.545	0.032	0.119
P39_NL	Cond2	0.798	0.000	0.044	0.110	0.139	0.707
P33_NL	Cond2	0.844	0.000	0.000	0.005	0.127	0.873
P22_NL	Cond2	0.907	0.000	0.209	0.222	0.036	0.534
P19_NL	Cond2	0.896	0.000	0.200	0.282	0.013	0.505
P65_NL	Cond2	0.204	0.090	0.231	0.274	0.089	0.316

You can also visualize count data of Reference bulk expression data set with respect to cellular markers using interactive heatmaps.



1.9 Comparative Profiling

By using the "Comparative Profiling" tab, you can to choose which metadata variables to use as a reference to compare samples and conditions across submitted and reference bulk rna expression datasets. You can select a subset of the data with **Select Series** option, select a metadata variable with **Select Meta** option, and choose membership scoring method by **Score Method** similar to in Computational profiling.



Once you click **Start** button, Dprofiler calculates the membership scores given conditions/phenotypes in the reference bulk RNA expression data, and visualizes the scores as below.

CHAPTER

TWO

COMPUTATIONAL PROFILING

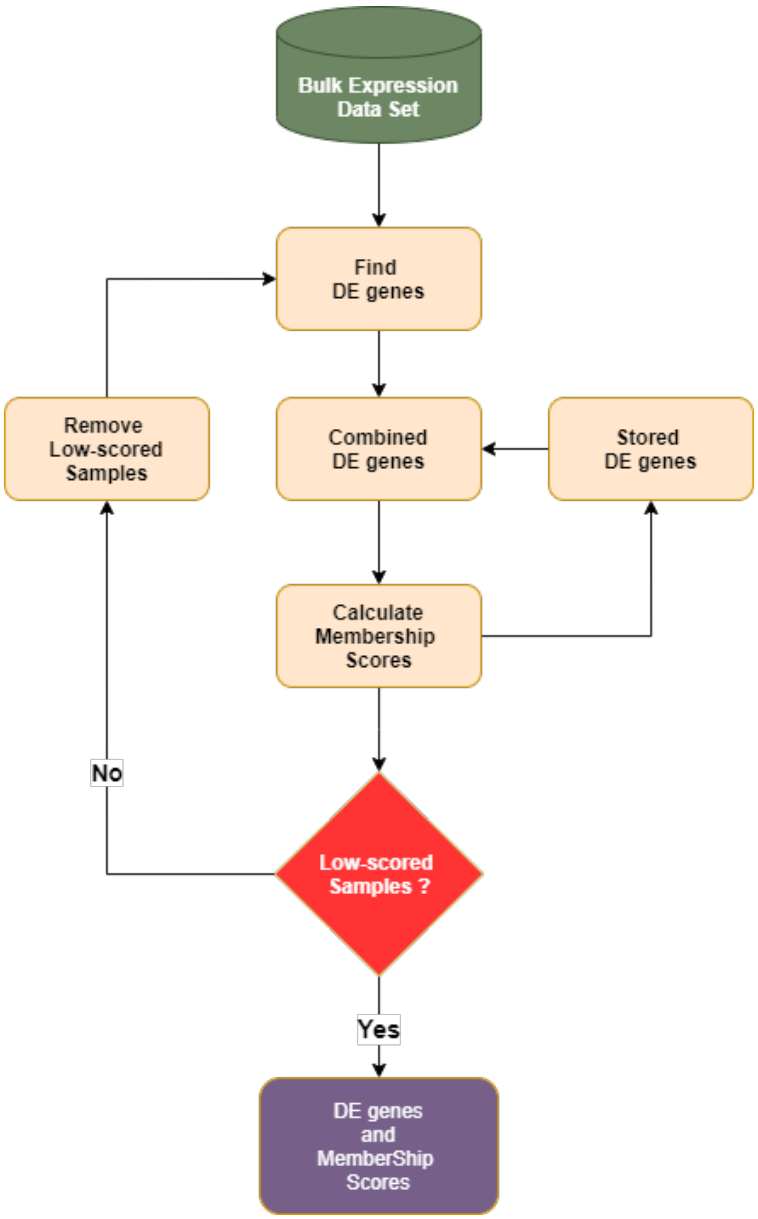
This guide contains a brief discription of the Computational Profiling Analysis used within Dprofiler.

2.1 Introduction

Dprofiler provides methods for calculating membership scores to be used to profiling samples associated to phenotypic profiles of interest with the submitted dataset. By iteratively removing samples with low scores and repeatatively testing for differentially expressed genes, computational profiling analysis of Dprofiler converges if there are no more low-scored samples left in the data set. The final list of samples with high membership scores establish homogeneous (pure) reference profiles of these phenotypic groups, and they are used to calculate the final score to establish the profile of each individual sample:

- Conducts a DE analysis (DESeq2, EdgeR or Limma) given remaining samples within the data.
- Estimates the membership scores (based on either Silhouette and NNLS) of all samples given expression profiles limited to differentially expressed genes
- Removes samples with low membership scores.
- Repeats until no more samples have to be removed from the data.

The Membership scores of samples are measured by two distinct similarity measures and methods. First of these measures is the Silhouette index that allows quality control of partitioning algorithms once datasets are clustered into meaningful subsets of samples. However, we utilize the silhouette index to detect those samples that do not well clustered or classified into their associated groups or conditions with the same label. The second method is based on a linear regression method whose coefficients are regularized to non-negative values as to calculate the percentage of input variables. Such a method allows us to model expression profile of each submitted sample given mean expression profiles of phenotypes/conditions where coefficient are associated to scores, representing the similarity between the condition and the submitted sample.



2.2 Silhouette Measure

Silhouette measure of a sample is calculated given a known partitioning (or classification) of data set and a measure of distance between all samples within the data set. We use Spearman correlation as a distance measure between expression profiles of each sample since it has been shown to be quite robust in many biological data analysis platforms and software tools (citation). The silhouette measure of each sample is calculated by separately measuring the average distance to all samples with the label and the minimum of all averages distances to other clusters with different labels, then these two measures are subtracted and normalized to calculate an index universally between -1 and 1. A silhouette measure of -1 would indicate that the sample is misclustered to its associated group and it is highly likely that its expression profile is more similar to samples of other conditions/groups. A silhouette measure of 1 would indicate a

perfect clustering of the sample, and silhouette measure 0 would indicate an ambiguous similarity of expression profiles between at least two conditions. We normalize silhouette measure of each submitted sample between (0,1) to establish the membership score.

2.3 Non-negative Least Squares

The second type of membership score available to Dprofiler users is non-negative least squares (NNLS) regression-based score where the non-negative beta coefficients are provided by the lawson-hanson implementation of NNLS regression. Such regression analysis has been applied to various problems where target profiles were confounded by a mixture of baseline profiles and hence target profiles are detected to exhibit heterogeneous properties. Applications include proteomics, genomics, imaging and economics. We use NNLS to detect the heterogeneous samples whose expression profiles are abundant in sets of biomarkers of multiple conditions within the disease study, hence deemed as heterogeneous. We use the mean expression profiles of all the conditions as an input to the non-negative regression problem where the response variable is the sample we would like to detect its degree of heterogeneity. We use the estimated coefficients are the membership scores.

CHAPTER

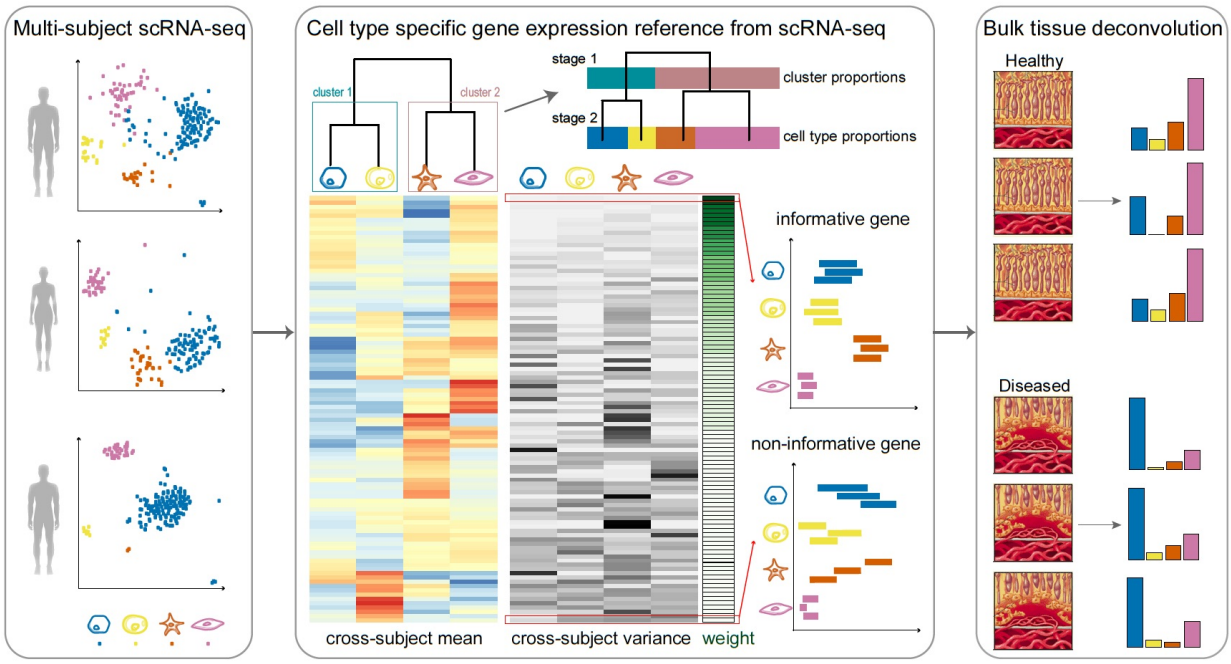
THREE

CELLULAR COMPOSITION ANALYSIS

This guide contains a brief discription of MuSiC algorithm used within Dprofiler for estimating cellular compositions of Reference bulk expression data set using the scRNA expression data.

3.1 MuSiC Algorithm

The MuSIC algorithm employs single cell genomic expression profiles to acquire non-negative least squares estimates (Wang et al.). A specific feature of MUSIC allows the proportions of closely related cell types to be correctly estimated. To deal with collinearity, MuSiC employs a tree-guided procedure that recursively zooms in on closely related cell types. Rather than pre-selecting marker genes from scRNA-seq based only on mean expression, MuSIC gives weight to each gene allowing for the use of a larger set of genes in deconvolution. The weighting scheme prioritizes consistent genes across subjects: (i) up-weighing genes with low cross-subject variance (informative genes) and (ii) down-weighing genes with high cross-subject variance (non-informative genes). This requirement on cross-subject consistency is critical for transferring cell type-specific gene expression information from one dataset to another.



CHAPTER

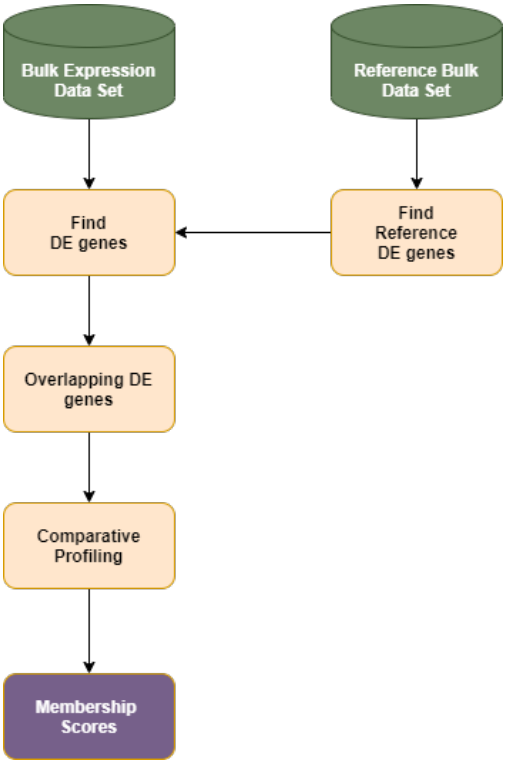
FOUR

COMPARATIVE PROFILING

Dprofiler allows users to incorporate silhouette measures and non-negative least squares (NNLS)-based membership scores to profile submitted bulk RNA samples given external reference bulk samples.

The gene expression profiles of the external reference bulk samples are often limited to genes of interest where Dprofiler uses an overlapping set of differentially expressed genes of submitted data set and gene profiles of reference bulk data sets to compute membership scores. The membership score of each submitted sample is calculated by:

- finding overlapping genes across submitted and reference bulk RNA expression datasets.
- choosing profiles and mean expression profiles of conditions within the reference expression dataset
- calculate the membership score using the similarity between submitted profiles and reference profiles .



Dprofiler also provides a connection to DolphinMeta (Dmeta) to import reference bulk expression profiles across nu-

mereous publically available data sets.

CHAPTER

FIVE

DE ANALYSIS

This guide contains a brief discription of DE analysis methods used within Dprofiler. These methods are primarily incorporated within each iteration of the Differential Heterogeneity Analysis.

5.1 Introduction

Differential gene expression analysis has become an increasingly popular tool in determining and viewing up and/or down experssed genes between two or more sets of samples. The goal of Differential expression analysis is to find genes, transcripts or regions whose difference in expression/count, when accounting for the variance within condition, is higher than expected by chance. DESeq2 is one of the highly used package in R available via Bioconductor and is designed to normalize count data from high-throughput sequencing assays such as RNA-Seq and test for differential expression (Love et al. 2014). For more information on the DESeq2 algorithm, you can read its documentation below;

DESeq2 userguide

With multiple parameters such as padjust values, log fold changes, and plot styles, altering plots, created with your DE data can be a hassle as well as time consuming. The Dprofiler uses DESeq2, EdgeR, and Limma coupled with shiny to produce real-time changes within your plot queries and allows for interactive browsing of your DE results. In addition to DE analysis, Dprofiler also offers a variety of other plots and analysis tools to help visualize your data even further.

5.2 DESeq2

For the details please check the user guide. DESeq2 userguide

DESeq2 performs multiple steps in order to analyze the data you've provided for it. The first step is to indicate the condition that each column (experiment) in the table represent. You can group multiple samples into one condition column. DESeq2 will compute the probability that a gene is differentially expressed (DE) for ALL genes in the table. It outputs both a nominal and a multiple hypothesis corrected p-value (padj) using a negative binomial distribution.

5.3 Un-normalized counts

DESeq2 requires count data as input obtained from RNA-Seq or another high-thorughput sequencing experiment in the form of matrix values. Here we convert un-integer values to integer to be able to run DESeq2. The matrix values should be un-normalized, since DESeq2 model internally corrects for library size. So, transformed or normalized values such as counts scaled by library size should not be used as input. Please use edgeR or limma for normalized counts.

5.4 Used parameters for DESeq2

- **fitType:** either "parametric", "local", or "mean" for the type of fitting of dispersions to the mean intensity. See estimateDispersions for description.
- **betaPrior:** whether or not to put a zero-mean normal prior on the non-intercept coefficients See nbinomWaldTest for description of the calculation of the beta prior. By default, the beta prior is used only for the Wald test, but can also be specified for the likelihood ratio test.
- **testType:** either "Wald" or "LRT", which will then use either Wald significance tests (defined by nbinomWaldTest), or the likelihood ratio test on the difference in deviance between a full and reduced model formula (defined by nbinomLRT)
- **rowsum.filter:** regions/genes/isoforms with total count (across all samples) below this value will be filtered out

5.5 EdgeR

For the details please check the user guide. EdgeR userguide.

5.6 Used parameters for EdgeR

- **Normalization:** Calculate normalization factors to scale the raw library sizes. Values can be "TMM","RLE","upperquartile","none".
- **Dispersion:** either a numeric vector of dispersions or a character string indicating that dispersions should be taken from the data object.
- **testType:** exactTest or glmLRT. exactTest: Computes p-values for differential abundance for each gene between two samples, conditioning on the total count for each gene. The counts in each group are assumed to follow a binomial distribution. glmLRT: Fits a negative binomial generalized log-linear model to the read counts for each gene and conducts genewise statistical tests.
- **rowsum.filter:** regions/genes/isoforms with total count (across all samples) below this value will be filtered out

5.7 Limma

For the details please check the user guide. Limma userguide.

Limma is a package to analyse of microarray or RNA-Seq data. If data is normalized with spike-in or any other scaling, tranforamtion or normalization method, Limma can be ideal. In that case, prefer limma rather than DESeq2 or EdgeR.

5.8 Used parameters for Limma

- **Normalization:** Calculate normalization factors to scale the raw library sizes. Values can be "TMM","RLE","upperquartile","none".
- **Fit Type:** fitting method; "ls" for least squares or "robust" for robust regression
- **Norm. Bet. Arrays:** Normalization Between Arrays; Normalizes expression intensities so that the intensities or log-ratios have similar distributions across a set of arrays.
- **rowsum.filter:** regions/genes/isoforms with total count (across all samples) below this value will be filtered out

5.9 ComBat

For more details on ComBat, please check the user guide. ComBat userguide.

ComBat is part of the SVA R Bioconductor package which specializes in corecting for known batch effects. No additional parameters are selected or altered when running SVA's ComBat.

CHAPTER

SIX

REFERENCES

1. Wang, X., Park, J., Susztak, K., Zhang, N. R., & Li, M. (2019). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nature communications, 10(1), 1-9.
2. Chang,W. et al. (2016) shiny: Web Application Framework for R.
3. Chang,W. and Wickham,H. (2015) ggvis: Interactive Grammar of Graphics.
4. Ritchie,M.E. et al. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res., 43, e47–e47.
5. Risso,D. et al. (2014) Normalization of RNA-seq data using factor analysis of control genes or samples. Nat. Biotechnol., 32, 896–902.
6. Anders,S. et al. (2014) HTSeq - A Python framework to work with high-throughput sequencing data.
7. Love,M.I. et al. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol., 15, 550.
8. Vernia,S. et al. (2014) The PPAR$alpha$-FGF21 hormone axis contributes to metabolic regulation by the hepatic JNK signaling pathway. Cell Metab., 20, 512–525.
9. Murtagh, Fionn and Legendre, Pierre (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? Journal of Classification 31 (forthcoming).
10. Reese,S.E. et al. (2013) A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal component analysis. Bioinformatics, 29, 2877–2883.
11. Trapnell,C. et al. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc., 7, 562–578.
12. Howe,E.A. et al. (2011) RNA-Seq analysis in MeV. Bioinformatics, 27, 3209–3210.
13. Kallio,M.A. et al. (2011) Chipster: user-friendly analysis software for microarray and other high-throughput data. BMC Genomics, 12, 507.
14. Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics, 12, 323.
15. Boutsidis, C., & Drineas, P. (2009). Random projections for the nonnegative least-squares problem. Linear algebra and its applications, 431(5-7), 760-771.
16. Johnson et al. (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics, 8, 118-127.
17. Reich,M. et al. (2006) GenePattern 2.0. Nat. Genet., 38, 500–501.
18. Giardine,B. et al. (2005) Galaxy: a platform for interactive large-scale genome analysis. Genome Res., 15, 1451–1455.

19. Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, 53-65.