MethylR manual

methylR: from sequencer to publication

Massimiliano Volpe

Jyotirmoy Das

Contents

List of Figures

Acknowledgements

e-mail: methylr@googlegroups.com

Bioinformatics,
Core Facility,
Faculty of Medicine and Health Sciences,
Linköping University,
Linköping,
Sweden

We would like to acknowledge the Core Facility, Faculty of Medicine and Health Sciences, Linköping University, Linköping, Sweden and Clinical Genomics Linköping, Science for Life Laboratory, Sweden for their support.

DISCLAIMER: ***All packages used in*** **methylR* are publicly available and open-source license. We have modified the source as required for *methylR*. Venn and UpSet plots inspired by the intervene package (Khan and Mathelier 2017), we modified as required for *methylR*.***

Part I.

Welcome to methylR: DNA Methylation Data Analysis Pipeline

> **!** Important
>
> TO ALL OUR USERS, IF YOU ARE EXPERIENCING ANY TROUBLE WITH THE APP, BEFORE SENDING THE BUG REPORT, PLEASE RESTART THE DOCKER CONTAINER AND TRY AGAIN.

DNA Methylation is one of the most studied epigenetic modifications in humans, playing a critical role in cellular response, development and differentiation (Das, Verma, et al. 2019). The transfer of a methyl group onto the C5 position of the cytosine to form 5-methyl-cytosine is considered as the DNA methylation or epigenetic mechanism in human or mammalian genome (Moore, Le, and Fan 2013). Epigenetic research has its roots in plant science, which emerged in the early 20th century. In human medicine, cancer biology has driven the field forwards during the last two decades and in combination with modern, array-based techniques and nex-generation sequencing now provides the scientific community with an easily accessible tool to study epigenetics at a whole-genome level (Das, Idh, et al. 2021). To find the methylated site or the CpG site, Illumina$^{(R)}$ uses DNA methylation array-based technology. Till date three different array platforms are available from Illumina for human genome to identify the CpG site or specific DNA methylation location, namely 27K, 450K or 850K. A more detail history and timeline can be found here in this article by Harrison and Pari-McDermott (2011) (Harrison and Parle-McDermott 2011). After performing the array, the major part is the analysis of the raw data generated from the machine. Numerous tools are available to analyze the data using different operating system, various computational languages. And all of these tools require extensive handling of computational resources. For the Biologist or those who have limited computational knowledge, it is extremely difficult to handle all these tools.

Here, in *methylR*, we presented a shiny-based web server approach to minimize the above-mentioned difficulties. MethylR has graphical user interface to support and understand the various options used in the DNA methylation analysis with an extensive manual/tutorial how to use it. The background computational power depends on the user's computer which can also be optimized. We successfully tested the pipeline on Linux based system.

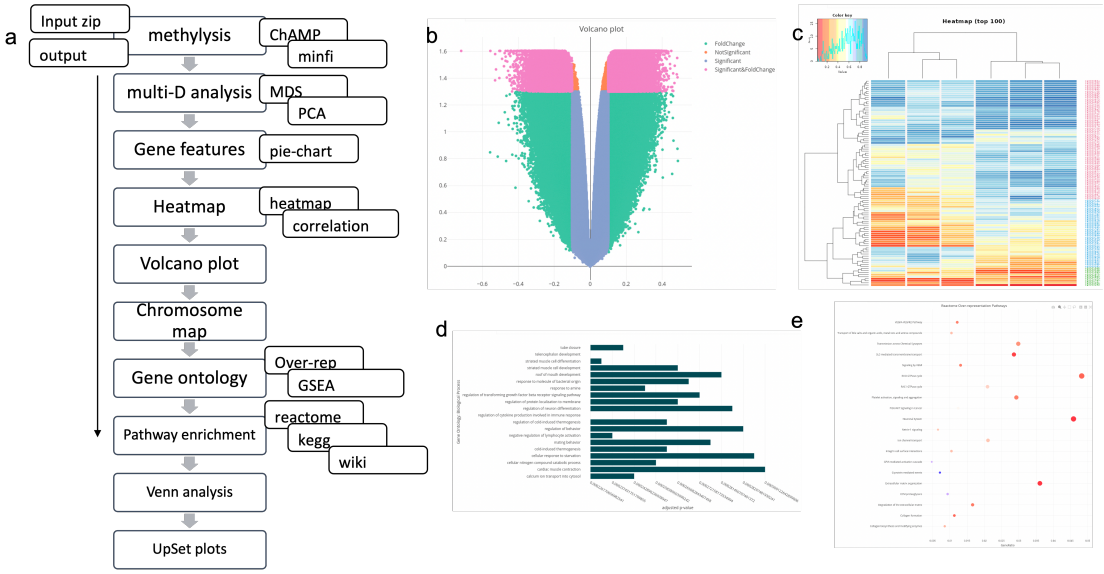

Figure 0.1.: Pipeline description and Figures

Requirements

- **LinuxOS** - (AMD64)
  - *Ubuntu 20.04LTS*
  - *Docker* (version 20.10.18)
  - *web-browser*: *Firefox* (version 105)
- **MacOS** - (AMD64)
  - *Monterey (version 12.5.1)*
  - *Docker* (version 20.10.17)
  - *Docker Desktop* (version 4.12.0)
  - *web-browsers*:
    - *Google Chrome* (version 106),
    - *Firefox* (version 106),
    - *Apple Safari* (version 15.6.1)
- **WindowsOS** - (AMD64)
  - *Windows 10* (version 21H2)
  - *Docker* (version 20.10.20)
  - *Docker Desktop* (version 4.13.0)
  - *WSL2* - (**Ubuntu 20.04LTS**)
  - web-browsers:
    - Firefox (version 106),

* Google Chrome (version 107),
* Microsoft Edge (version 106).

> **Note**
>
> MethylR cannot run on ARM64 chipset architecture.

How to Use

Local use

methylR is packed into docker container that is available online. Singularity can also be used to run the docker container directly from terminal. Please check the github link to run the container from your local computer.

> **Important**
>
> For convenient analysis, after one complete analysis (either with *ChAMP* or *minfi*), close the browser (to clear the temporary memory) and start again.

> **Note**
>
> Note: If you want to run with the test data, Please download the testdata from https://sourceforge.net/projects/methylr/files/testData.zip

1. Methylysis

Methylysis is the most important section in *methylR* since it contains the tools to analyze the DNA methylation data for two different Illumina arrays, 450K and 850K array. Two of the most used well-defined pipelines are currently available in *methylR*: The Chip Analysis Methylation Pipeline ChAMP and minfi. Those are alternative pipelines and, although the user has the freedom to decide to test both, only the results from one of them is required to run to visualize and explore the data in the downstream sections. For new users, we suggest to run **ChAMP** as it is the latest, well packed and well maintained pipeline for *Illumina HumanMethylation Array* data analysis. Both pipelines offer different input options and parameters that can be modified based on the user's actual data. If users wish to compare the results from both the pipelines, *methylR* offers this possibility through the downstream processing (Venn Analysis see Chapter 9 and UpSet Plot see Chapter 10 sections).

Tip

1. After uploading the zip file See Appendix A, the pipeline will start automatically by displaying the notification "Computing methylysis, please wait…" "Computing methylysis, please wait...". When the notification turns off, the user can go to different tabs to display the result. Please wait 5-10 seconds (see Appendix D, *Calculation Time for each process*) to display the result on the tab. Depending on the sample size, it may requires more time to display properly.

Tip

2. If the *methylysis* page goes "dim/disconnected"dim/disconnected, the analysis may encounter some errors during the run and the program stops working. Please refresh the page and run the same analysis again with different filters/parameters. Example, for ChAMP pipeline, if the user selects the "adjusted P-value" = 0.05 (as default), may be for the sample data, there is no differentially methylated CpGs at that value. Please change the adjusted P-value (recommend to set it at 1, and check the table after the run that what is appropriate cut-off) and run the pipeline again.

1.1. How to use

Details are provided below -

1.1.1. Data upload & Parameters setup

The current version can handle the upload of the data directory. Please put all RAW IDAT (intensity data) files (as generated by the Illumina sequencer) and the "Sample_sheet.csv" together in a directory.

1.1.2. Structure of sample_sheet.csv

The Sample_sheet.csv must have the following components -
1. Sample_Name
2. Sample_Group
3. Sentrix_ID
4. Sentrix_position

> ⚠ Warning
>
> *Note* If the user uses Microsoft Excel to build the Sample_sheet.csv, please check that
> 1. **Sentrix_ID** : are in text format (not in number format, which Excel will change to scientific numbers and will not properly displayed).
> 2. *Optional check*: Copy and paste the Excel table of Sample_sheet in some text editor like notepad or VS code and check the format.

Every section in *methylR* comes with the proper testing data linked at the bottom of the page, just search for the ***"example data"*** button.

1.2. Parameters setup

1.2.1. Choose analysis algorithm

Currently, we have included two most usable algorithms to analyze the data - ChAMP and minfi.

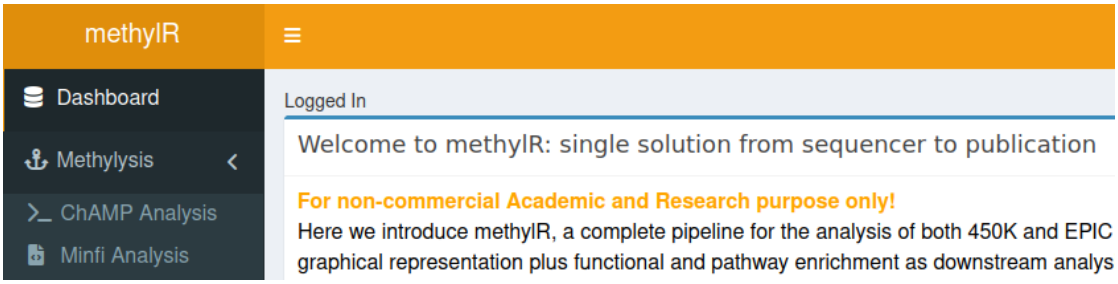

Figure 1.1.: Algorithm choice

1.2.1.1. ChAMP pipeline parameters

> **Note**
>
> **PLEASE NOTE**: ChAMP process will do all filtration automatically, independent of User's input. The following filtration will be done -
> i. *filtering probes with detection p-value > 0.01*,
> ii. *filter out probes with <3 beads in at least 5% of samples per probe*,
> iii. *filter all non-CpG probes*
> iv. *filter all SNP-related probes*
> v. *filter all multi-hit probes*
> vi. *filter all probes located in X and Y chromosomes*

1. *Choose type of Illumina array*: Two options are provided to choose, namely EPIC/850K array and 450K array from Illumina array analysis.
2. *Adjusted P-value*: User can define there own adjusted P-value to run the analysis. The default is 0.05.
3. *Normalization*: User can choose different normalization methods from the drop-down list,
   - BMIQ (Beta-Mixture Quantile Normalization) (Teschendorff et al. 2013),
   - PBC (Peak-Based Correction) (Dedeurwaerder et al. 2011),

The default setup will run with the BMIQ normalization method.

> **Tip**
>
> Please check references for different type of normalization method.

4. *Batch Effect Correction*: *ComBat* function is used to correct the batch effect. User can choose whether to compute the batch effect or not by clicking the button. When the button is "green" green (**ON**), it will prompt to select the factors for the batch effect correction, *Slide*, *Array*, *Age*, *Sex*, or *Other*. Select as necessary and should have the column in the Sample_sheet (check the Sample_sheet in the testdata set). If you have other option than *Slide*, *Array*, *Age*, or *Sex*, rename the column as **'Other'** and run batch effect correction.
   If the button is "red" red (**OFF**), the pipeline will continue without analyzing the batch effect.

> Tip
>
> Please check the reference for the batch effect correction using combat method (Johnson, Li, and Rabinovic 2007).

5. *Cell Type Heterogeneity*: Houseman et al (2013) (Houseman et al. 2012) algorithm is applied to calculate the cell-type heterogeneity from PBMC (*Peripheral Bllod Mononuclear Cells*) dataset using the *refbase* function. This is deactivated as default. Press the button to activate and run during the analysis.

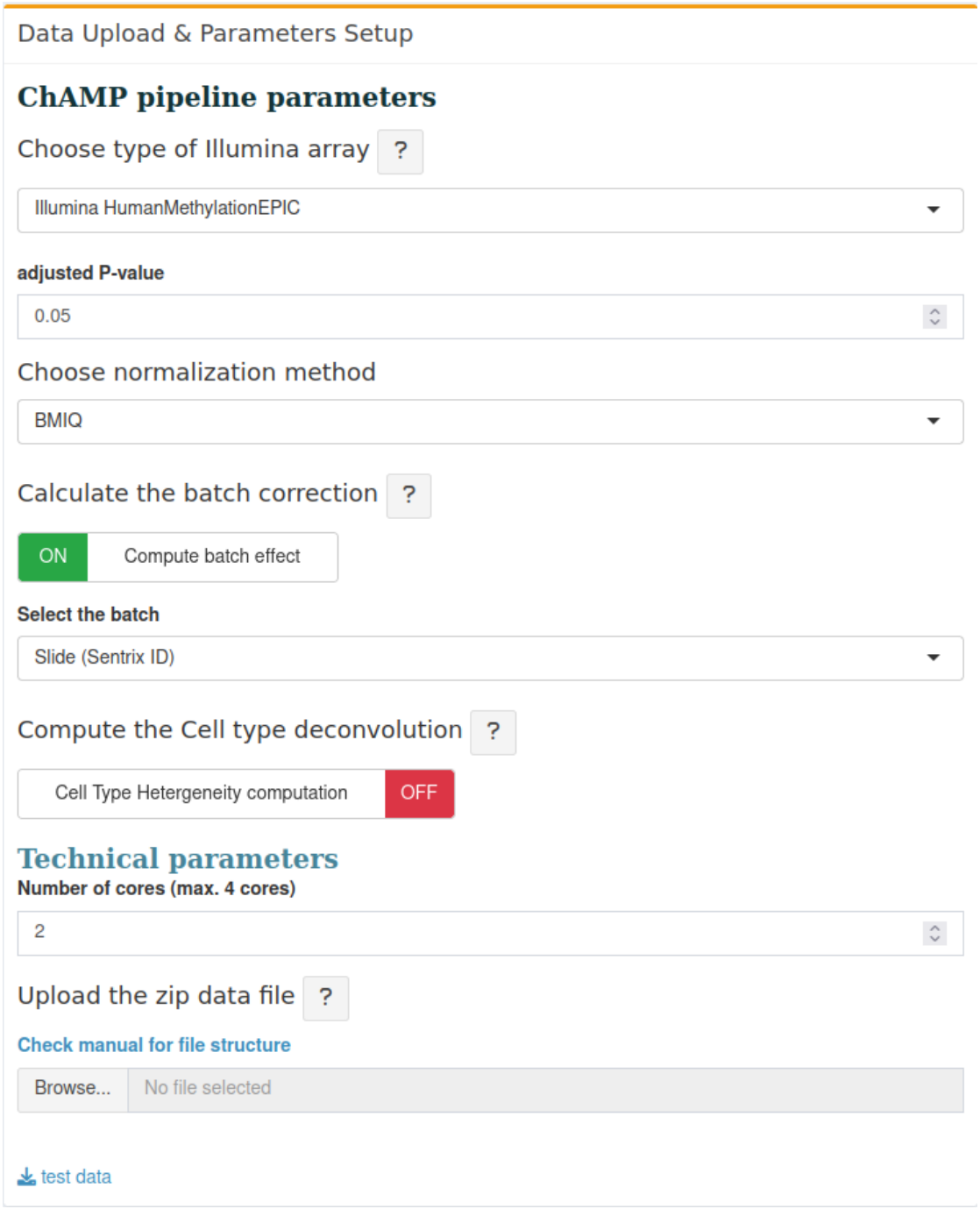

Figure 1.2.: ChAMP parameters setup

1.2.1.2. Minfi pipeline parameters

1. *Choose preprocess method* : There are several methods available for preprocessing or normalizing the raw data using the RGset. Here we listed them as the user's input options to select the preprocess/normalization method as per their choice:

    - Raw: No processing of the raw data,
    - SWAN: Subset Quantile Within array Normalization (Maksimovic, Gordon, and Oshlack 2012; Touleimat and Tost 2012),
    - Noob: Noob preprocessing (Triche Jr et al. 2013),
    - Illumina: Illumina preprocessing, as performed by Genome Studio (reverse engineered by minfi authors) (Aryee et al. 2014)
    - Funnorm: Functional normalization (Fortin et al. 2014)

2. *Select filtration method*: In this section, we assigned options for the user to perform the different filters, like removal of XY chromosomes from the analysis, removal of SNPs or removal of non-specific probes from the dataset. By default, the pipeline will use p-value detection 0.01.

3. *Compute cell type heterogeneity*: similar as ChAMP, we used Houseman method to correct the cell type heterogeneity. In minfi pipeline we used default minfi function *estimateCellCounts*. As before, user can choose to avoid this if the samples are not from PBMC cell types.

4. *Choose genome annotation database*: To annotate the DMC list from the analysis, use the human genome reference annotation data file. For 850K array, make sure to use the hg38 array to compare the result with the output of ChAMP pipeline.

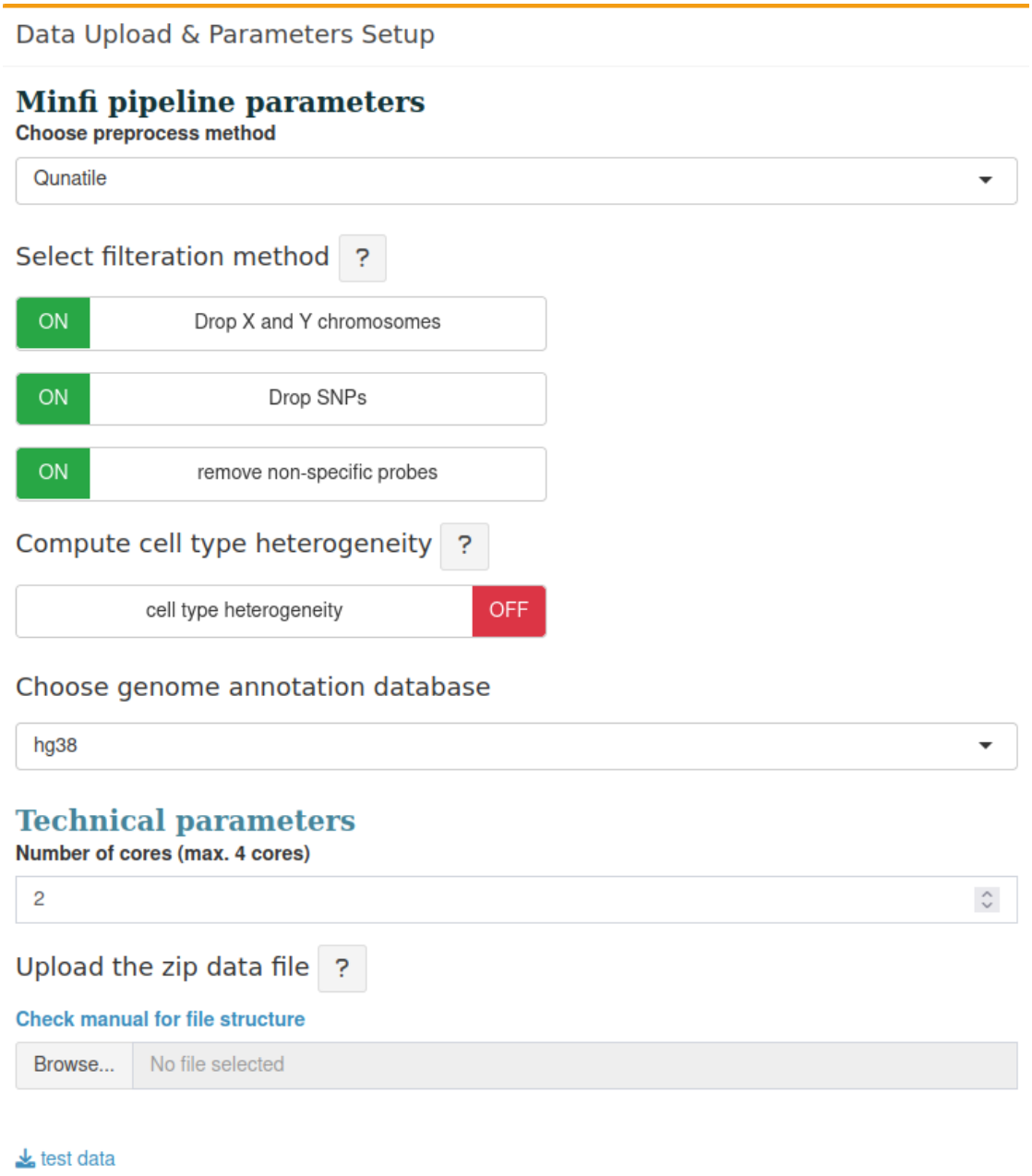

Figure 1.3.: Minfi parameters setup

1.3. Technical setup:

Both pipelines require the user to set the following technical parameters:

1. *Number of cores*: Both pipeline can run on 1 core which will take more time to compute the entire process. User can choose to setup the number of cores depending the availability. The default is set to 1 core for Minfi, and 2 cores for ChAMP and maximum is 4 cores.

> **Note**
>
> Multi-threading for DNA methylation analysis is mainly used in the **Normalization** process. If you have data with more than 50 samples, using 4 cores may reduce the calculation time. If the dataset has less number of samples, increasing the number of cores will not do effectively any better time reduction (See Appendix D).

2. *Data upload* The user should set the parameters first and choose all parameters as described above and then upload the data directory. To do that, just click on the button "Browse..." and locate the zip archive containing raw files. As soon as the pipeline finishes the upload of the data directory, it will start running the analysis.

1.4. Requirement for data upload

1. ***idat files***: all idat files, green and red as received from Illumina sequencing array should be provided for the analysis. All files should be in one directory/folder.
2. ***Sample_sheet.csv***: the "Sample_sheet.csv" should also be provided in the same directory with idat files.

> **Tip**
>
> Check the github repository for sample data file. You may download the Sample_sheet.csv file and use as a template for your sample_sheet.

Part II.

Feature Analysis

2. Multi-D Analysis

[1]

In *methylR*, multiple dimensional analysis includes two type of analysis -
1. **MDS**: Multidimensional Scaling
2. **PCA**: Prinicipal Component Analysis

Multidimensional scaling is a visual representation of distances or dissimilarities between set of objects. **MDS** finds set of vectors in p-dimensional space such that the matrix of Euclidean distance among them corresponds as closely as possible to some function of the input matrix. The input to multidimensional scaling is a distance matrix. To get some more details on how to use **MDS** in biological data, read (Mugavin 2008; D. Lacher 1987; David A Lacher and O'Donnell 1988)
Principal Component Analysis (PCA) is the original vectors in n-dimensional space and the data are projected onto the directions in the data with the most variance.

2.1. How to use

For both analysis, user need to provide a TEXT (tab-delimited) file with numeric values, *e.g.* the output normalized table from methylysis, *i.e.* the normalized value table. However the user can use similar tables for the analysis.

2.1.1. Data Upload

1. Select the text file (tab-delimited file) and upload it.
2. User can choose the option to use number of variables from the uploaded data file.
3. On the right tab, under "MDS plot", user will find the button to "Run MDS Analysis"
4. Next tab is designed for the PCA plot and here user can have an input of text file to highlight the group.

[1] TO ALL OUR USERS, IF YOU ARE EXPERIENCING ANY TROUBLE WITH THE APP, BEFORE SENDING THE BUG REPORT, PLEASE RESTART THE DOCKER CONTAINER AND TRY AGAIN.

2.2. Analysis result

2.2.1. MDS plot

1. After the click on "Run MDS Analysis", the program will take some time to generate the plot and will appear as soon the run finishes. The zoom bar can be used to zoom in/out the plot.
2. The generated figure can be downloaded in different format, vector graphics support - PDF and SVG or PNG and TIFF. User can download all different options for the same figure.

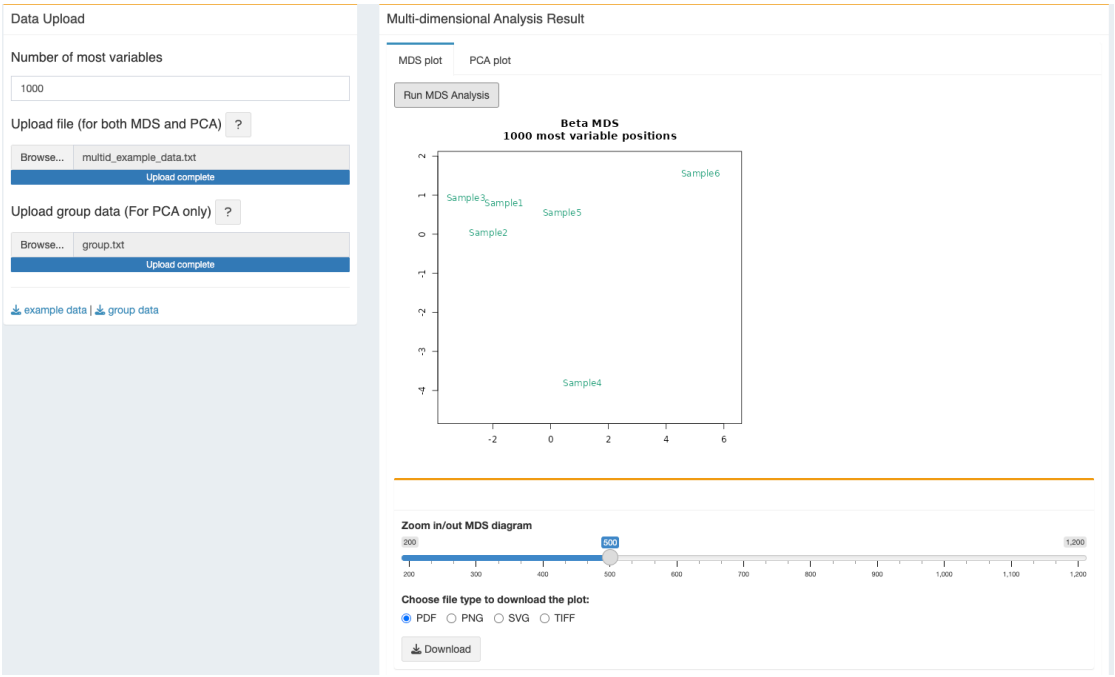

Figure 2.1.: MDS plot

2.2.2. PCA plot

1. The plot will be generated after computing the PCA, "Run PCA" with the group colors and legend.
2. The generated plot is dynamic and positional details with the group name can be seen with mouse hovering.
3. The plot is generated using the plotly application, it can zoom in/out, save figure as PNG format, and do all other available functionality for plotly figures.
4. User can also download the dynamic figure as html file.

5. To generate the PCA plot one additional TEXT file is needed. This single-column file must have "group" as header and should store the sample group in the same order as you have in the Sample_Group column of the Sample_shee.csv file. Please find here an example of this file that you can use with the test data provided with this distribution.

> **Note**
>
> Please note that, the single column with header **"group"** should be supplied in this file. Match the column names of the variable data file with the group.
> ***Example:*** If in the variable data file, the column names are - sampleA1 sampleA2 sampleA3 sampleB1 sampleB2 sampleB3 the group text file should be like this - group
> A
> A
> A
> B
> B
> B
> For more, please see the test data files.

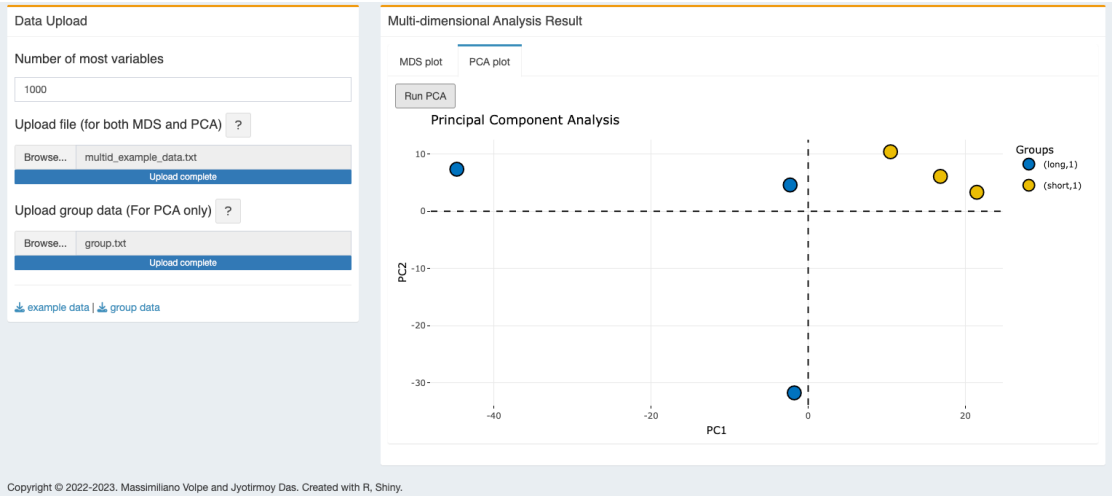

Figure 2.2.: Principal Component Analysis plot

2.3. R packages used

1. FactoMineR

2. factoExtra
3. explor

3. Gene Features analysis

[1]

Gene Features, here in *methylR*, we presented only the structural part of the gene that can be classified as promoter, exon, intron, untranslated regions (Skolnick, Fetrow, and Kolinski 2000). These features are essential for DNA methylation study as example, methylation on the promoter can alter the gene expression. Here we used a simple tool to find out how many differentially methylated CpGs are distributed over the different regions of the gene. However, we separated this from the methylysis because user can use the same tool for different datasets, such as differentially expressed genes data.

3.1. How to use

3.1.1. Data upload & Parameters setup

3.1.1.1. Data upload

1. User can upload the differentially methylated CpG (DMC) file that are generated from methylysis run or
2. they can use separate file which has similar annotation. The basic requirement to run the tool is to have the following gene feature in the supplied text file -
    - 1st Exon
    - 3´UTR
    - 5´UTR
    - Body
    - ExonBnd
    - TSS1500
    - TSS200
3. Remember to upload the file as TEXT (tab-delimited) format file.
4. After uploading the file, when the 'blue' bar finished uploading, click on 'Run Analysis' will generate the pie chart.

[1] TO ALL OUR USERS, IF YOU ARE EXPERIENCING ANY TROUBLE WITH THE APP, BEFORE SENDING THE BUG REPORT, PLEASE RESTART THE DOCKER CONTAINER AND TRY AGAIN.

3.2. Gene Features Analysis Result

3.2.1. Gene Feature Plot

An interactive pie chart will be generated with different regions and number of DMCs.

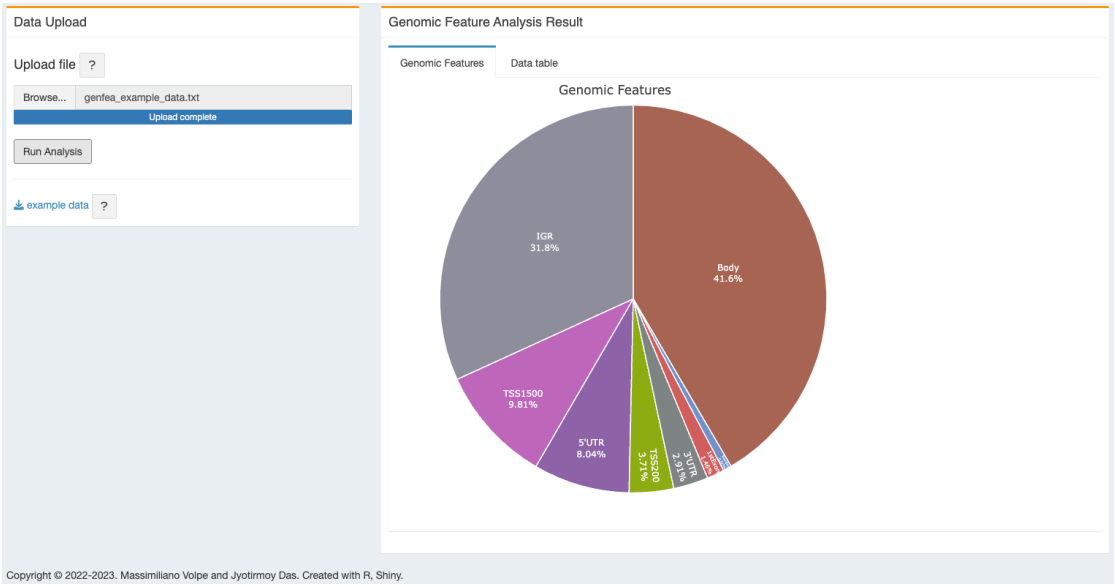

Figure 3.1.: Genomic Feature plot

3.2.2. Gene Feature Table

The result will also be displayed as table, available for download.

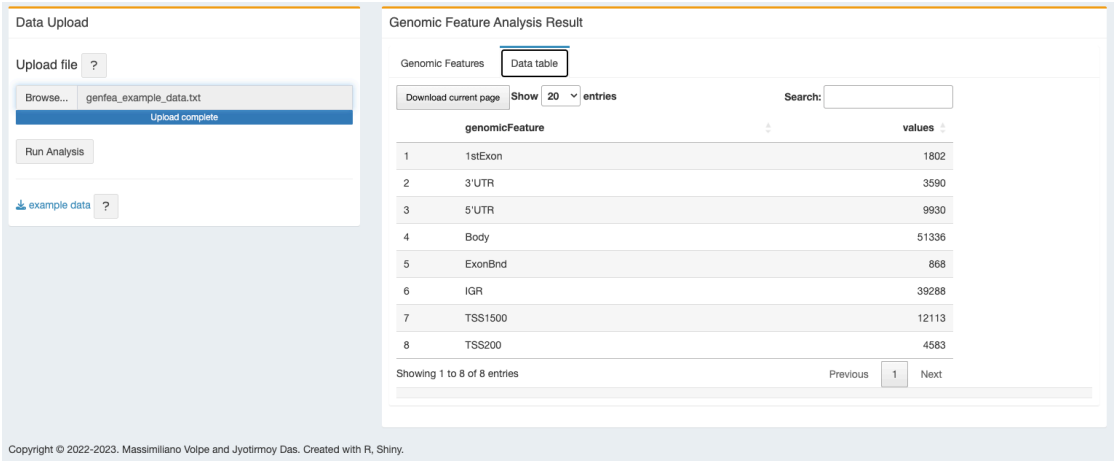

Figure 3.2.: Genomic Feature Table

3.3. R packages used

1. plotly

4. Pairwise analysis (Heatmap)

[1]

A heatmap module (Pairwise Plot) is added in *methylR* to show the β value distribution of the differentially methylated CpGs. A pairwise correlation analysis can also be performed in the module.

4.1. How to use

4.1.1. Upload

User can upload the input data matrix in **Tab** (.txt) or **Comma** (.csv) or **Semicolon** (.csv but with ;) separated format. If you don't have the matrix, this modules provides the functionality to make it starting from the main results coming from *methylysis* (see Chapter 1). Either ChAMP and minfi will provide the input data or you can get test data by clicking on "Matrix example data" and "List example data" buttons. For a better view of the result, we added the functionality to change the number of variables on the heatmap.

4.1.2. Settings

1. **Plot type** - User can choose the heatmap or correlation plot function for the analysis.
2. **Correlation Coefficient** - four different types of correlation coefficient added in the module, 1) Pearson, 2) Spearman, 3) Kendall and 4) no-correlation (better for heatmap). Default chosen 'non' (no-correlation).

> ! Important
>
> ***PLEASE NOTE: IF YOU WANT TO PERFORM THE CORRELATION PLOT, THEN PLEASE SELECT PEARSON, SPEARMAN OR KENDALL.***

[1] TO ALL OUR USERS, IF YOU ARE EXPERIENCING ANY TROUBLE WITH THE APP, BEFORE SENDING THE BUG REPORT, PLEASE RESTART THE DOCKER CONTAINER AND TRY AGAIN.

3. **Agglomeration method for hclust** - different agglomeration method for hierarchical cluster analysis provided in the module, 1) ward.D, 2) ward.D2, 3) single, 4) Complete, 5) Average, 6) Mcquitty, 7) Median, and 8) Centroid. Default chosen 'Complete'.
4. **No. of clusters for hclust** - user can set the number of hierarchical cluster for their data. Default is 3.
5. **Distance Matrix computation** - different types of distance matrix calculation can be applied to generate the heatmap, 1) Euclidean, 2) Manhattan, 3) Canberra, 4) Minkowski or 5) none. Default is 'Euclidean'.
6. **Dendrogram** - user can choose to show the dendrogram on the row and/or column list.
7. **Color key** - selecting color key will give option to change the size of the color key. However, user can choose not to show the color-key. Also color key title is user-defined.
8. **axis label** - both x and y-axis label is user-defined. User can change the label of the x and y axis.
9. **Title** - It will change the title of the heatmap/ correlation plot.
10. **Zoom in & out Heatmap** - for the static plot, user can set the zoom in/out option.

4.1.3. Font & Color

1. **Select theme** - with the pre-defined theme colors, custom-defined color for the heatmap is also enabled.
2. **label** - user can separately define the size, rotation and color of the label text.
3. **Color** - rectangle border, grid color and label color is also user-defined.

4.1.4. Matrix preparation

We added a tab for the user to build the heatmap matrix by starting from the results of the main analysis, regardless the user choice to perform it with ChAMP or minfi. The matrix can be uploaded directly on the *Upload* tab to run the heatmap analysis.
1. **Upload normalized data table** - user can upload the normalized table from the main analysis directly without any modification.
2. **Upload DMC data table** - user can upload the differentially methylated CpG data table from the main analysis directly without any modification.
3. **Select adjusted P-value** - for more filtration on the dataset, we set a adjusted p-value (BH-corrected as defined in the main analysis section, both ChAMP or minfi) parameter. Default is 0.05.
4. **Select logFC value** - for more filtration on the dataset, we set a logFC (as defined in the main analysis section, both ChAMP or minfi) parameter. Default is 0.1.

4.2. Analysis result

1. **Heatmap** - the figure will be shown in the adjacent panel. It can be downloaded in the following formats, PDF, PNG, SVG and TIFF.

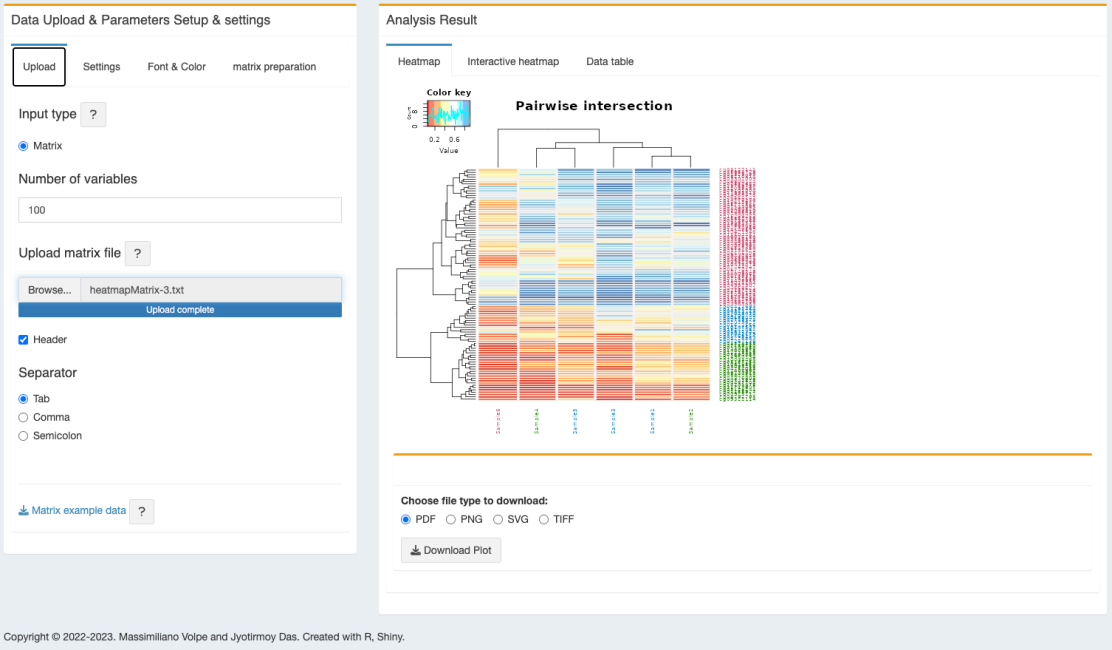

Figure 4.1.: Static Heatmap

> **Tip**
>
> **Correlation plot** - for correlation plot, please adjust the 'Settings', 'Plot type' to ***Corrplot*** and 'Correlation Coefficient' to ***Pearson/Spearman/Kendall***.

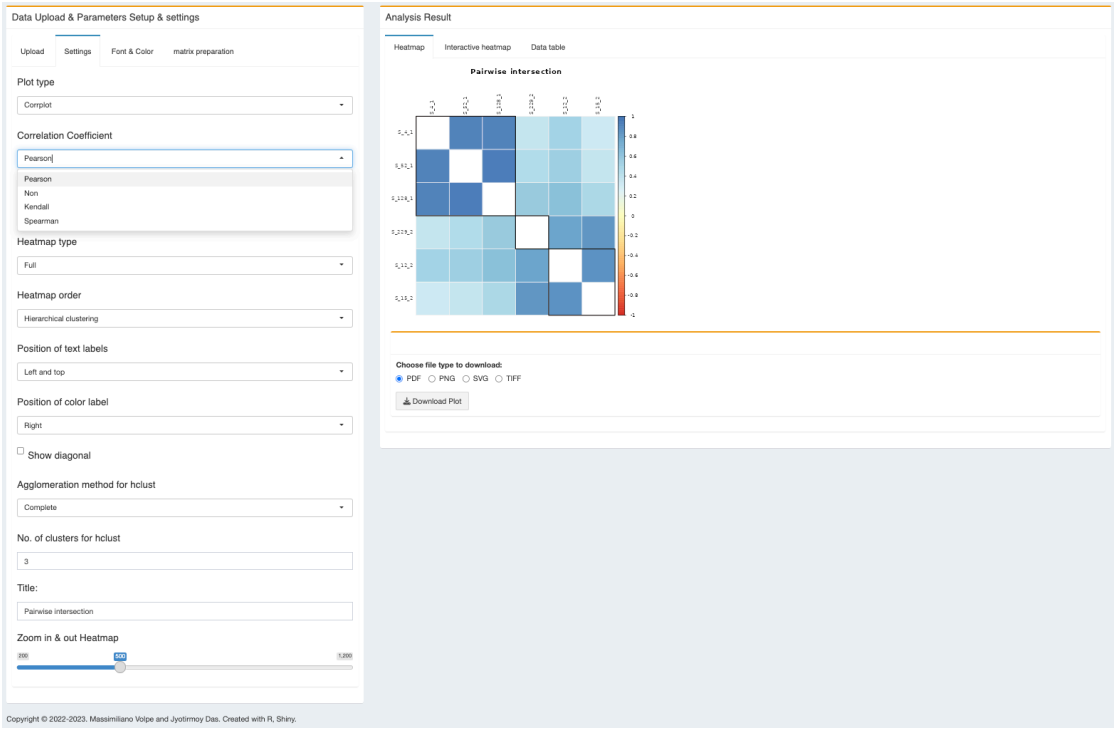

Figure 4.2.: Correlation plot

2. **Interactive heatmap** - an interactive heatmap will also be generated and can be downloaded as HTML file.

Figure 4.3.: Interactive Heatmap

3. **Data table** - a data table will be generated from the heatmap figure data.

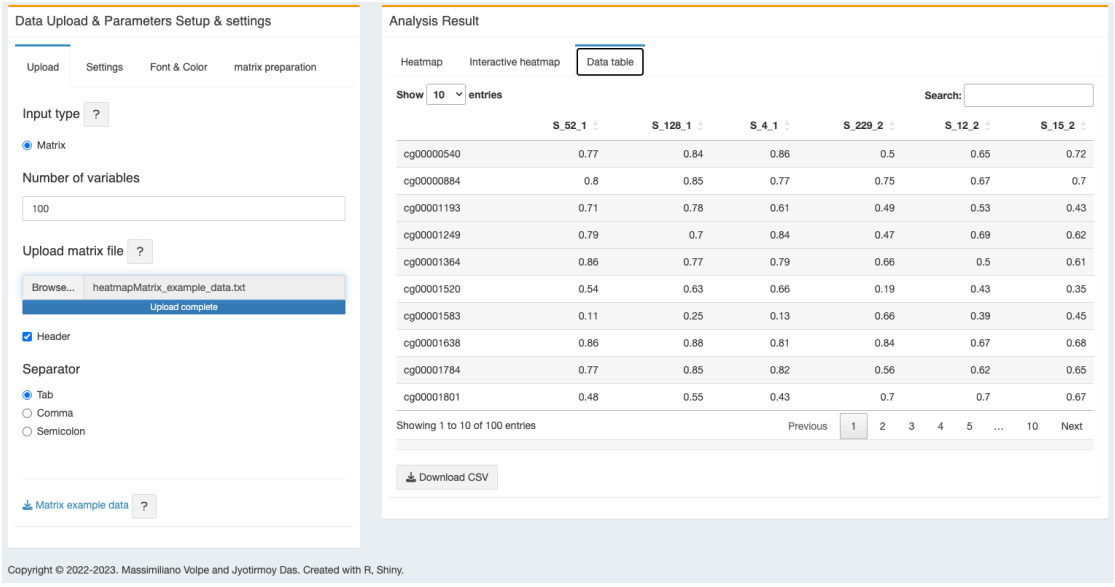

Figure 4.4.: Heatmap/Correlation data table

4.3. R packages used

1. heatmap2

2. D3heatmap
3. intervene

5. Volcano plot

[1]

Volcano plot is a nice tool to visualize in a two-dimensional way for differentially methylated CpG site or differentially expressed genes using the statistical p-values as well as the fold change value. Like a volcano, the plot can show the significant or insignificant data in a scatter plot manner. Here with *methylR*, we used plotly output to visualize the volcano plot to see the CpG or gene name (if the data from the differential analysis) with their respective p-values (adjusted p-values) and the logFC (or mean methylation difference).

5.1. How to use

5.1.1. Data upload & Parameters setup

5.1.1.1. Data upload

1. User needs to upload a **Tab** (.txt) file with adjusted p-values and logFC values. At present, user can use the DMCs data file directly generated from the main analysis (see Chapter 1).
2. To setup the adjusted p-value, user can change the cut-off. Default is setup to 0.01.
3. LogFC cut-off can also be changed as per user requirement. Default is setup to 0.3.
4. After the file upload and setting up the cut-off for adjusted p-value and logFC, click the "Run Analysis" button.

[1] TO ALL OUR USERS, IF YOU ARE EXPERIENCING ANY TROUBLE WITH THE APP, BEFORE SENDING THE BUG REPORT, PLEASE RESTART THE DOCKER CONTAINER AND TRY AGAIN.

5.2. Analysis result

5.2.1. Volcano plot

1. The figure will generated as soon as the computation finishes. However, it might takes some more time depending on the size of the file. If user upload a file with 750K rows, it will take 3-5 minutes to generate the figure (See Appendix D). It is noteworthy that this big data in volcano plot, may be unstable in the browser.
2. User can download the plot as figure (same as before) and the dynamic figure as a html file.

> **Note**
>
> Please note displaying of the volcano plot will take some time, even after the warning "generating plot, please wait..." "generating plot, please wait..." disappears. Please wait for 1-2 minutes to get the visualization. The same may happen to the volcano plot data table.

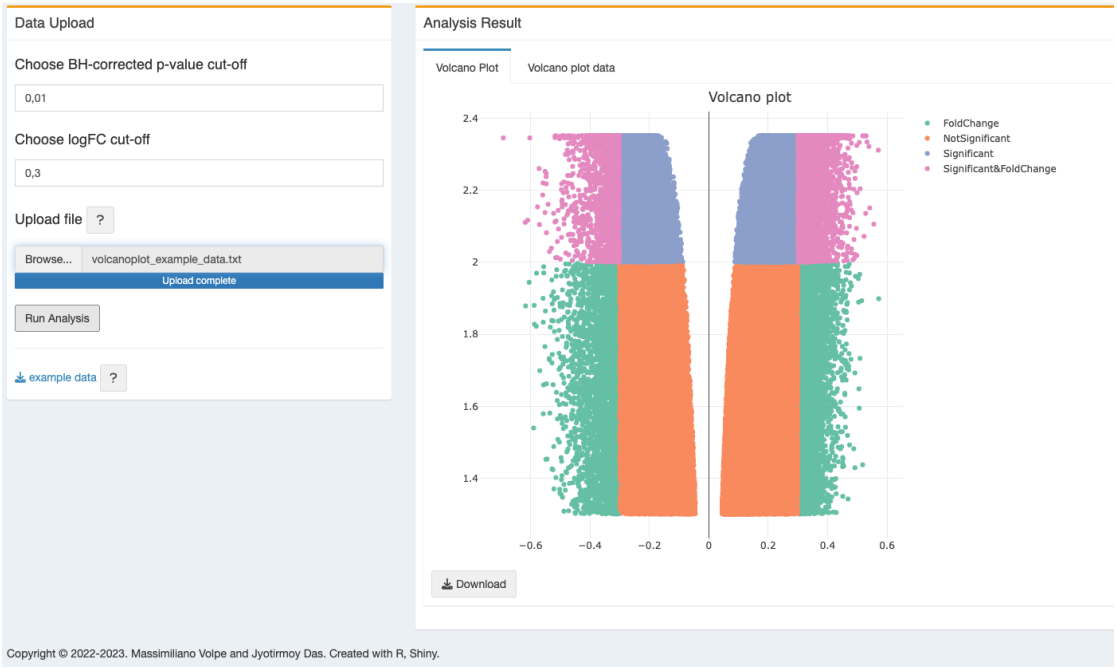

Figure 5.1.: Volcano plot

> **Tip**
>
> Volcano plot figure colors annotation- Significant & Fold Change"Significant & Fold Change", Significant "Significant", FoldChange FoldChange or NotSignificant "Notsignificant"

3. On the right tab, user can also see the volcano data table which is useful when they are using the full dataset from the main analysis result (See Chapter 1).

5.2.2. Volcano data

One data table will be generated using the input data and will have a column marked with Significant & Fold Change"Significant & Fold Change", Significant "Significant", FoldChange FoldChange or NotSignificant "Notsignificant" depends on the adjusted p-value and logFC cut-off.

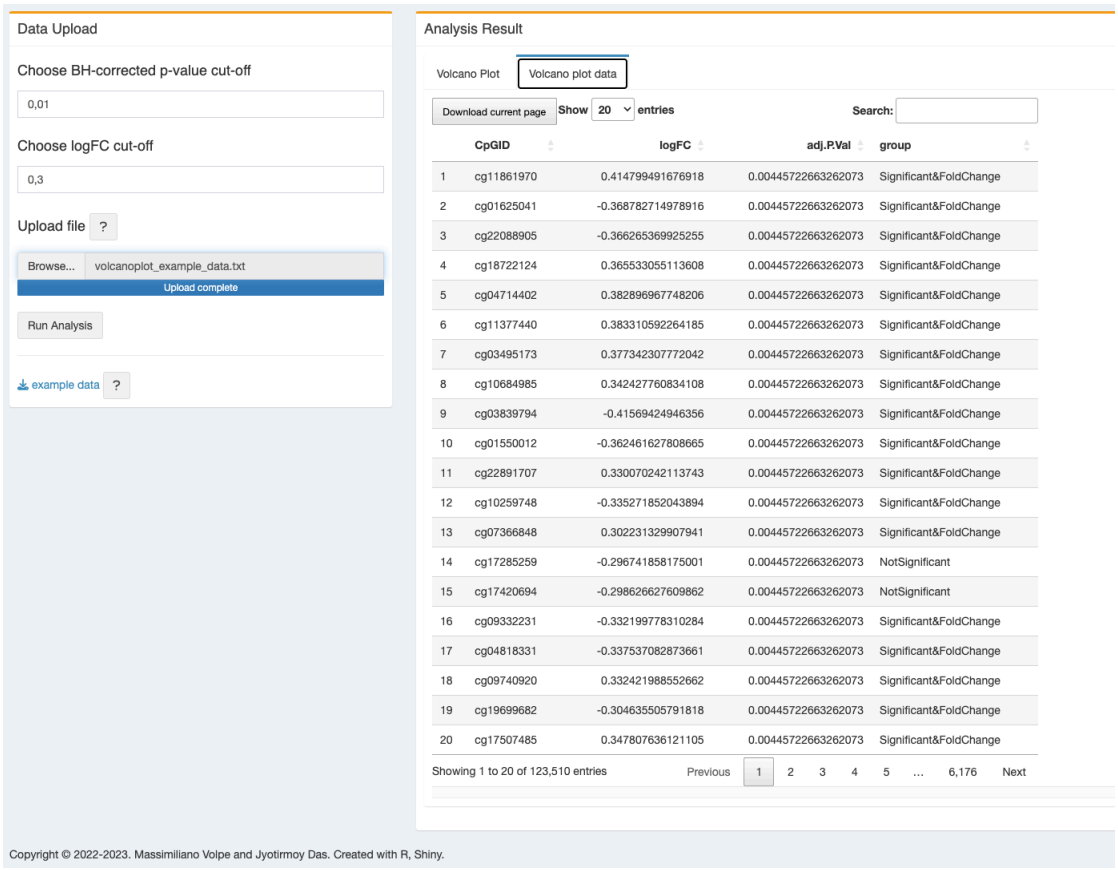

Figure 5.2.: Volcano data table

5.3. R packages used

1. plotly

6. Chromosome plot

[1]

Chromosome plot is a way to visualize coordinates at DMC positions over the chromosome structure. In *methylR*, Users can change the cut-off for adjusted *p*-value as well as the fold change value. It is possible to visualize one chromosome at a time or all the chromosomes on the same figure.

6.1. How to use

6.1.1. Data upload & Parameters setup

6.1.1.1. Data upload

1. User needs to upload a text (tab-delimited) file with adjusted *p*-values and logFC values. At present, user can use the DMCs data file directly generated from the main analysis (See Chapter 1).
2. To setup the adjusted p-value, user can change the cut-off. Default is setup to 0.05.
3. LogFC cut-off can also be changed as per user requirement. Default is 0.3.
4. After the file upload and setting up the cut-off for adjusted *p*-value and logFC, click the "Create plot" button.

> Tip
>
> 1. After creating the plot, if the user needs to add more chromosomes to the plot, please add the chromosome number from the "Select Chromosome" drop-down list and the plot will be updated automatically (***do not need to click "Create plot" again.***) and the same will hapeen if the user wants to remove one chromosome from the figure, just ***DELETE** it from the "Select Chromosome" drop-down menu.
> 2. "Change font size" will also updated automatically after the figure generation.

[1] TO ALL OUR USERS, IF YOU ARE EXPERIENCING ANY TROUBLE WITH THE APP, BEFORE SENDING THE BUG REPORT, PLEASE RESTART THE DOCKER CONTAINER AND TRY AGAIN.

Just change the font size as desired (max. 2).

6.2. Analysis result

6.2.1. Chromosome plot

1. The figure will be generated as soon as the computation finishes and it will allow you to vary the font size on the flow.
2. User can download the plot as a static figure (PDF, PNG, SVG, TIFF).

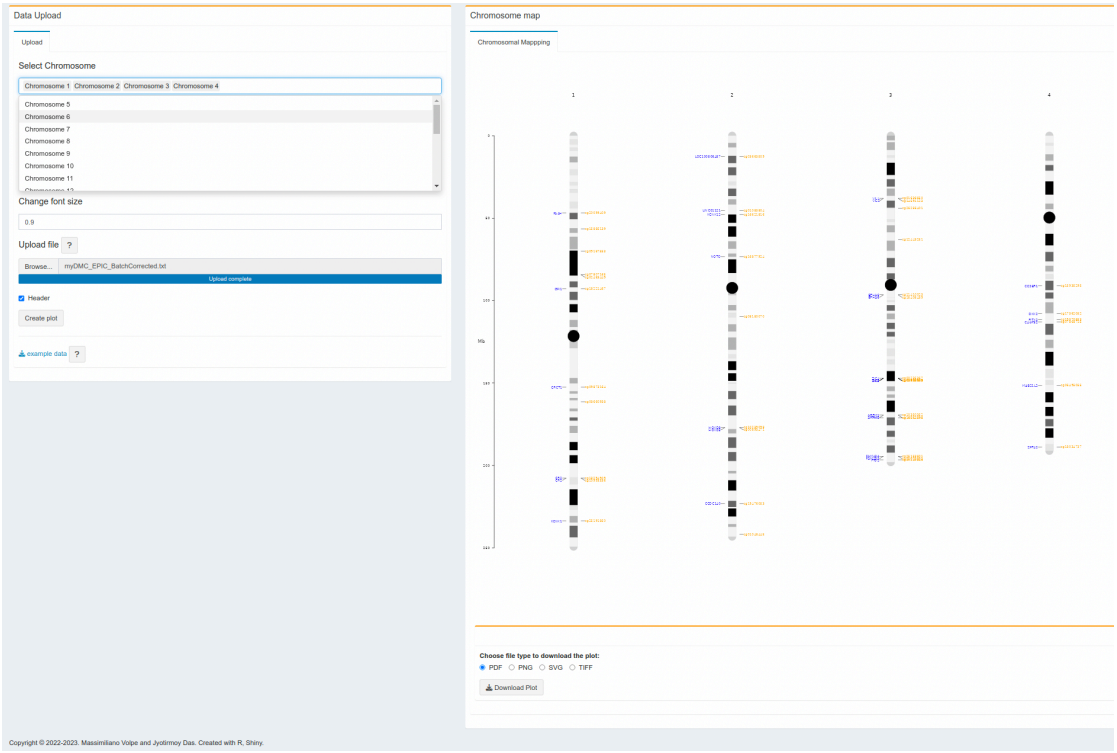

Figure 6.1.: Chromosome plot

Note

In the Chromosome plot, the blue blue text (e.g. FAAH FAAH) in the left side is the gene symbol and the text orange in the right side is the CpG ID from Illumina annotation (e.g. cg20099409 cg20099409). Please note, CpGs without the gene

> symbol only shows the CpG ID on the right side.

3. On the right tab, user can also see the volcano data table which is useful when they are using the full dataset from the main analysis (See Chapter 1).

6.3. R packages

1. Chromplot

Part III.

Association Study

7. Gene Ontology (GO) enrichment analysis

[1]

Gene Ontology is a very well-known method for accessing the functions of the gene identified through methylation analysis or expression analysis. Here in *methylR*, we used ontology analysis using the clusterProfiler package (Yu et al. 2012; Wu et al. 2021).

7.1. How to use

7.1.1. Data upload & Parameters setup

7.1.1.1. Parameters setup

1. *Choose GO analysis type*: user can choose to do the analysis whether **over-representation analysis** or the **gene-set enrichment analysis** (GSEA) from the drop-down menu.
2. *Select the adjusted p-value*: user can also choose the adjusted p-value for the analysis. Default is set to 0.05.
3. *Select the adjusted q-value*: q-value or the FDR can also be adjusted as per user's requirement. Default is 0.05.
4. *Select number of ontology classes*: to see the number of ontologies on the graph, user can setup different number. Default is 20.
5. *Select P-value adjustment method* As per clusterProfiler, we set different p-value adjustment methods, Benjamini-Hochberg, Benjamini-Yeketuli, Bonferroni, Holm, Hommel, Hochberg, FDR or none. Default is Benjamini-Hochberg.
6. *Select ontology class*: As defined in GO classification, we included all three ontology classes which user can select to show the plot.

7.1.1.2. Data upload

At present, user can upload the DMC data produced by the main analysis (see Chapter 1) directly. The input file should be in a **text (tab-delimited)** format.

[1] TO ALL OUR USERS, IF YOU ARE EXPERIENCING ANY TROUBLE WITH THE APP, BEFORE SENDING THE BUG REPORT, PLEASE RESTART THE DOCKER CONTAINER AND TRY AGAIN.

7.2. Analysis result

1. On the right tab, the analysis result the plot will be generated as soon as computation finished. The plot is generated with plotly and it will be dynamic in nature as before. User can download the plot as PNG format, zoom in/out or do other stuffs as per plotly figures. The dynamic figure can also be downloaded as a html file.

> **Note**
>
> 1. All horizontal bars (Gene Ontology terms) are clickable and will open a new tab with the respective gene onlogy detail from the AmiGO database.
> 2. Each interactive figure can be downloaded as HTML file and PNG file. The HTML file is clickable and each gene ontology term can open the respective detail from AmiGO database.

7.2.1. Biological processes

Figure 7.1.: Gene Ontology - Biological Processes

7.2.2. Cellular component

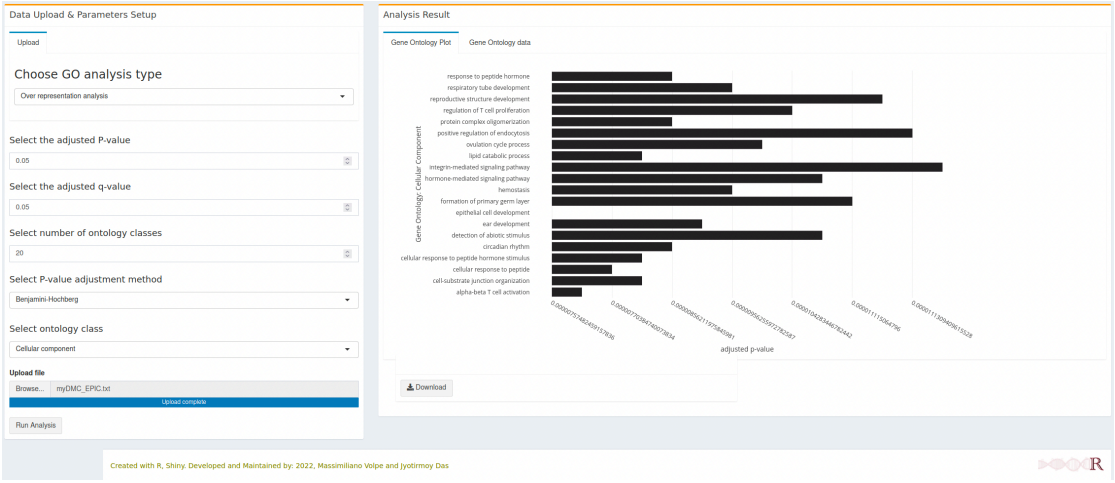

Figure 7.2.: Gene Ontology - Cellular Component

7.2.3. Molecular function

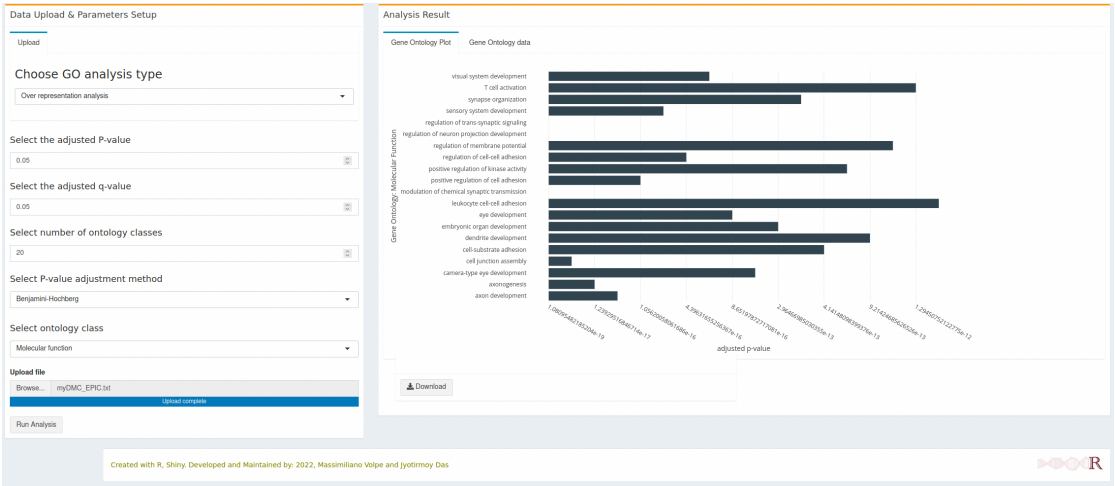

Figure 7.3.: Gene Ontology - Molecular Function

2. On the second right tab, user will get the result as a table format. It might takes some time to compute result and generates the table. User can download the result as an Excel file from the current page or the entire result.

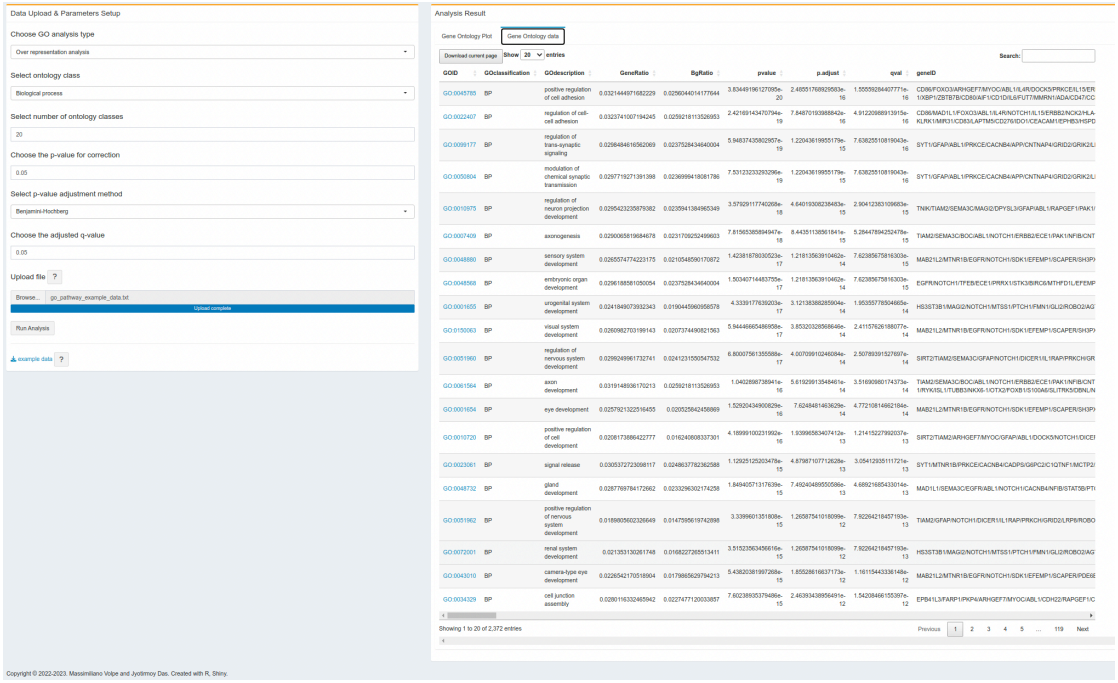

Figure 7.4.: Gene Ontology data table

> **Warning**
>
> 1. In the gene onlogy enrichment table, the GO ID is clickable and will open the respective GO class from the AmiGO database. However, this feature is only avaible on the browser, if the user download the table, there is no such link to check the GO source.
> 2. The GO enrichment table will download the full result (i.e., all GO classes, Cellular Component, Molecular Function and Biological Processes), user donot need to run the table again for different GO classes.

7.3. R packages used

1. clusterProfiler

8. Pathway enrichment analysis

[1]

Pathway enrichment analysis helps the user to get the mechanistic insights of the important genes from genome-wide data analysis. In *methylR*, we introduced the pathway analysis module that can compute the enriched pathways from three different databases, KEGG (Kanehisa and Goto 2000), Reactome (Gillespie et al. 2021) and Wikipathways (Pico et al. 2008; Martens et al. 2020).

8.1. How to use

8.1.1. Data upload & Parameters setup

8.1.1.1. Data upload

User can upload the direct output result from the main analysis. At present, user can upload the DMC data produced by the main analysis (See Chapter 1). The input file should be in a **text (tab-delimited)** format.

8.1.1.2. Parameters setup

1. *Choose pathway analysis type*: Please select pathway analysis type from the drop-down list
    - Over representation analysis (ORA);
    - Gene set enrichment analysis (GSEA); By default, the tool will use the over representation analysis.
2. *Choose pathway database*: user can choose to use different pathway database, namely
    - Reactome,
    - KEGG or

[1] TO ALL OUR USERS, IF YOU ARE EXPERIENCING ANY TROUBLE WITH THE APP, BEFORE SENDING THE BUG REPORT, PLEASE RESTART THE DOCKER CONTAINER AND TRY AGAIN.

- Wikipathways.

3. *Choose number of pathways*: Please select number of pathways for graphical display. The default is Top 20 pathways. The Top 20 enriched pathways is selected based on the adjusted P-values.

> Tip
>
> 1. If the analysis result does not get 20 pathways (as default setup) or the number selected by the user, then the plot will only shows the result with less number of pathways. User can change the parameters to see if they can get more number of enriched pathways.
> 2. If there is no enriched pathways with selected parameters, the figure tab may show warning like "check the logs or contact the author" check the logs or contact the author, please change the parameters and run again the analysis. If you are experiencing trouble, donot hesitate to contact us.

4. *Select P-value cut-off for correction*: The default value for *p*-value correction is set to 0.05. User can set their own cut-off values.

5. *Select P-value correction method*: The default method for adjustment of P-value is the Benjamini-Hochberg (BH) correction method. User can choose different method using the drop-down list:

   - Benjamini-Hochberg (BH)
   - Benjamini-Yeketuli (BY)
   - Bonferroni
   - Holm
   - Hommel
   - Hochberg
   - FDR
   - none

6. *Upload data file*: The input file should be in a **text (tab-delimited)** format. The user can upload the *ChAMP* result file (DMC file) directly for the analysis.

8.2. Analysis result

1. *Pathway enrichment plot*: after "Run Analysis", the plot will be generated as soon as computation has been done. Depends on the size of data, it might take few minutes (See Appendix D). At present the plot will be generated as a dot plot which is also a product of plotly, hence dynamic and have similar functionalities with mouse pointing. At present, with the mouse hover over, each dot will show

the pathway name, count of genes from the input list for that particular pathway, the corrected p-value and gene ratio. The color scale bar shows in the legend. User can download the figure as PNG as described above and the dynamic figure as a html file.

> **Note**
>
> 1. All dots (pathway enrichment terms) are clickable and will open a new tab with the respective pathway detail from the selected database (Reactome/KEGG/Wiki).
> 2. The interactive figure can be downloaded as HTML file and PNG file. The HTML file is clickable and each pathway enrichment term can open the respective database for pathway details.

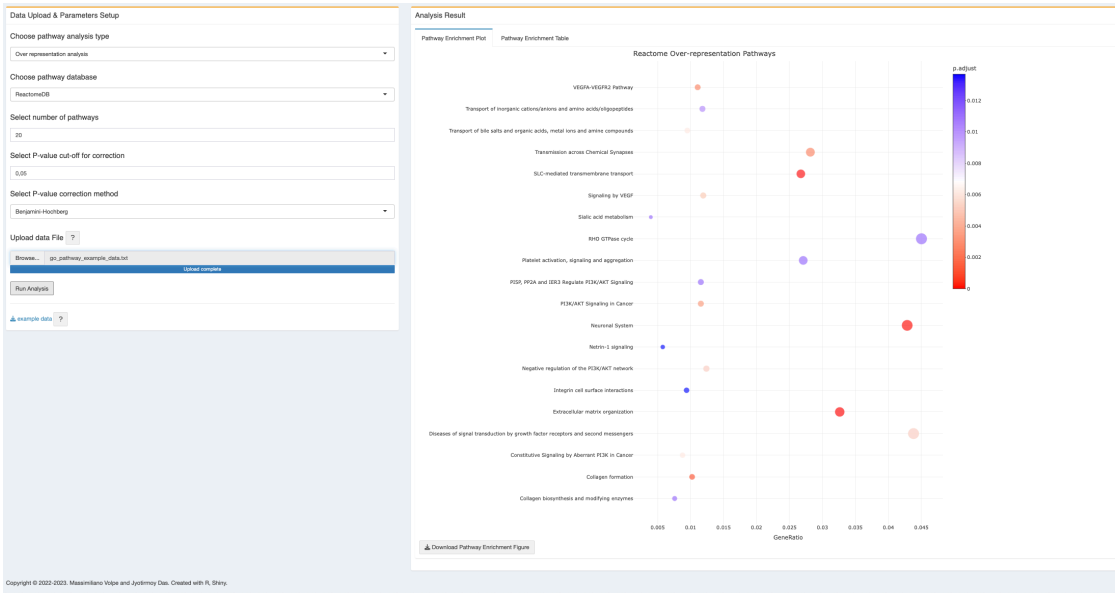

Figure 8.1.: Pathway Enrichment Plot

2. *Pathway enrichment table*: with the same input file and parameter setup, user can also get the result as an excel file (current page as well as full table).

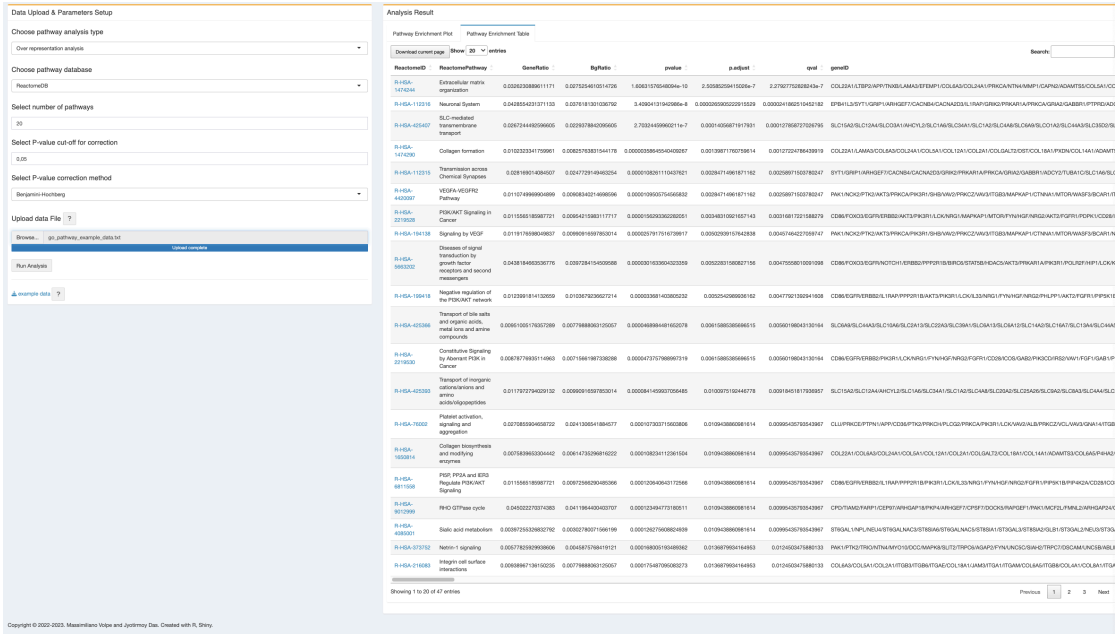

Figure 8.2.: Pathway Enrichment Table

> **Warning**
>
> In the pathway enrichment table, the pathway ID is clickable and will open the respective pathway from the database. However, this feature is only avaible on the browser, if the user download the table, there is no such link to check the pathway source.

8.3. R packages used

1. clusterProfiler

Part IV.

Set Analysis

9. Venn analysis

[1]

Venn analysis can be performed to show the logical relation between sets. In this module, user will need two or more analyses (max 6 datasets) to perform the Venn analysis. We adopt the part from intervene (Khan and Mathelier 2017) application and modified as required for *methylR* use.

9.1. How to use

Below given the details for the use of Venn analysis module.

9.1.1. Data upload & Parameters setup

9.1.1.1. Parameters setup

1. *Upload*: Data can be uploaded as **Tab** (.txt) or **Comma** (.csv) or **Semicolon** (.csv but with ;) separated format. A demo test dataset is running by default and it is available for download by clicking on the "example data" button.
2. *Settings*: Under settings, there are multiple options to display the plot -

    i. *Select sets*: will select sets from the uploaded data. User can remove the set as they need.
    ii. *Venn type*: different type of venn diagram can be selected from the drop-down menu

    - Chow-Ruskey
    - Classical
    - Edwards
    - Square
    - Battle

[1] TO ALL OUR USERS, IF YOU ARE EXPERIENCING ANY TROUBLE WITH THE APP, BEFORE SENDING THE BUG REPORT, PLEASE RESTART THE DOCKER CONTAINER AND TRY AGAIN.

The diagram can be *weighted* or *Eular*.

iii. *Border line width*: border line can be drawn with the slider option.

iv. *Border line type*: border line type can be selected from the drop-down menu.

v. *Zoom in/out Venn diagram*: select the zoom option on the slide bar.

3. *Font & Color*: multiple ooptions are included for font and colours -

i. *Select color theme*: Colour theme can be chosen from the drop-down menu.

ii. *Label font size*: Change the font size of the Label.

iii. *Number font size*: Change the font size of the number.

9.2. Results

User can download the figure in different format, PDF, PNG, SVG or TIFF.

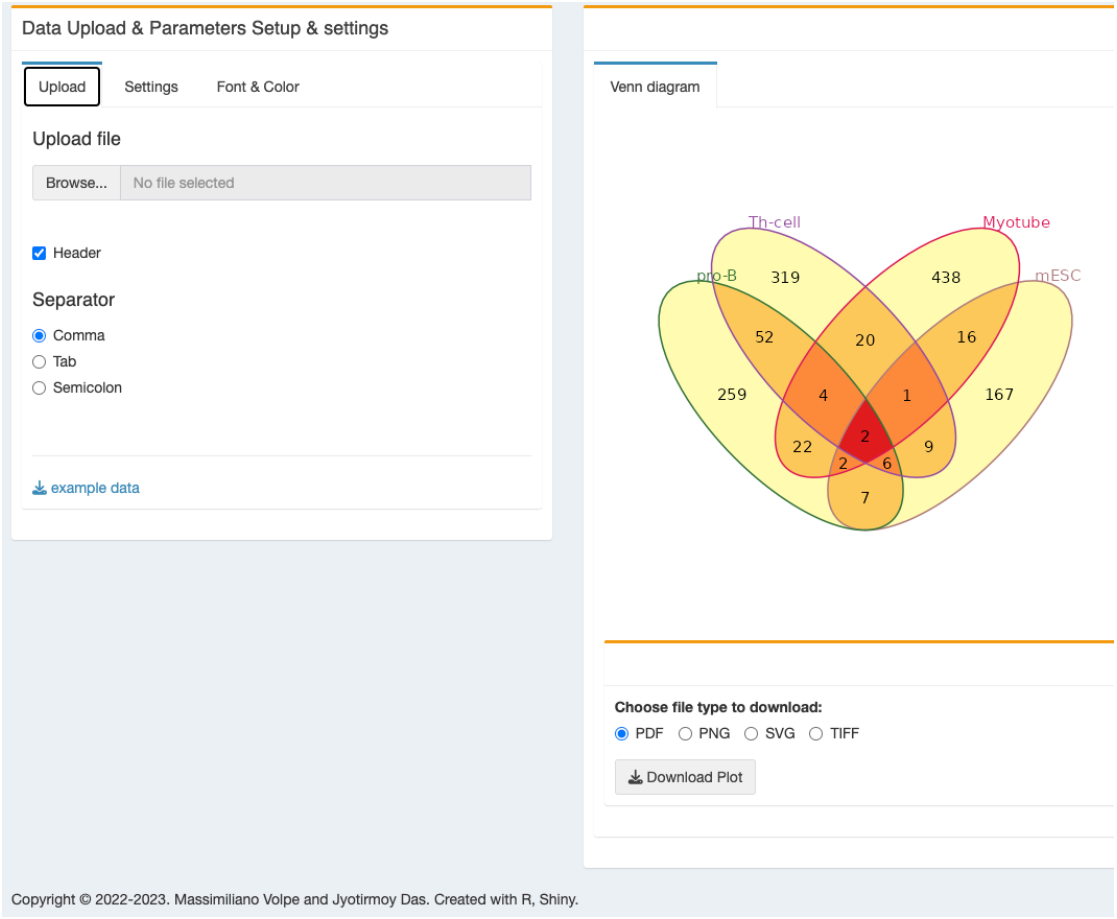

Figure 9.1.: Venn Result

9.3. R packages used

1. Vennerable
2. readr
3. intervene

10. UpSet Plots

[1]

UpSet plot will show the relation between different sets. We adopt the part from intervene (Khan and Mathelier 2017) application and modified as required for *methylR* use.

10.1. How to use

10.1.1. Data upload

1. *Upload*: Data can be uploaded as **Tab** (.txt) or **Comma** (.csv) or **Semicolon** (.csv but with ;) separated format. Please look into the example data file. A demo test dataset is running by default and it is available for download by clicking on the "List example data" button. UpSet module takes three types of inputs.

    i. *List type data*: List data is a correctly formatted csv/text file, with lists of names. Each column represents a set, and each row represents an element (names/gene/SNPs). Header names (first row) will be used as set names.
    ii. *Binary type data*: In the binary input file each column represents a set, and each row represents an element. If a names is in the set then it is represented as a 1, else it is represented as a 0.
    iii. *Combination/expression type data:* Combination/expression type data is the possible combinations of set intersections.

> **Note**
>
> PLEASE NOTE: ***"OR enter set combinations/expression"*** has the priority over **"Upload file"**. If you use the "set combinations", and then want to "upload file", please remove the "set combination" from the input box. To see how the "OR enter set combinations/expression" works, please use the example below the box (not the list/binary example file).

[1] TO ALL OUR USERS, IF YOU ARE EXPERIENCING ANY TROUBLE WITH THE APP, BEFORE SENDING THE BUG REPORT, PLEASE RESTART THE DOCKER CONTAINER AND TRY AGAIN.

10.1.2. Parameters setup

2. *Settings*: there are multiple options to display the plot -

   i. *select sets*: select the dataset from the input data.

   ii. *Number of intersections to show*: Please add the number to calculate the intersection.

   iii. *Order intersections by*: From the drop-down menu, please select the intersection order -

   - Frequency
   - degree

   iv. *Increasing/Decreasing*: Please select the order of the frequency/degree.

   v. *Scale intersections*: Please select the scale intersection from the drop-down menu -

   - Original,
   - log10,
   - log2

   vi. *scale sets*: Please select the scale intersection from the drop-down menu -

   - Original,
   - log10,
   - log2

   vii. *Plot width*: select the plot width from the slider.

   viii. *Plot height*: select the plot height from the slider.

   ix. *Bar matrix ratio*: select the bar matrix ratio from the slider.

   x. *Angle of number on the bar*: slider to change the angle of the numbers on the bar.

   xi. *Connecting point size*: change the connecting point size .

   xii. *Connecting line size*: change the connecting line size.

3. *Font & Color*: multiple options are included for font and colours -

   i. *Select main bar colour*: Change colour of the bars of intersection size.

   ii. *Select set bar colour*: Change the set bar colour on the side (set).

   iii. *Font size of intersection size label*: Change the font size of the intersection size.

iv. *Set size label font*: Change the font size of the set label.

v. *Set size ticks font*: Change the tick size (numerical value) on the set size bar.

vi. *Intersection size numbers font size*: Change the tick size (numerical value) on the intersection set bar.

vii. *Set names font size*: Change the font size for the set names.

10.2. Result

User can download the figure in different format, PDF, PNG, SVG or TIFF.

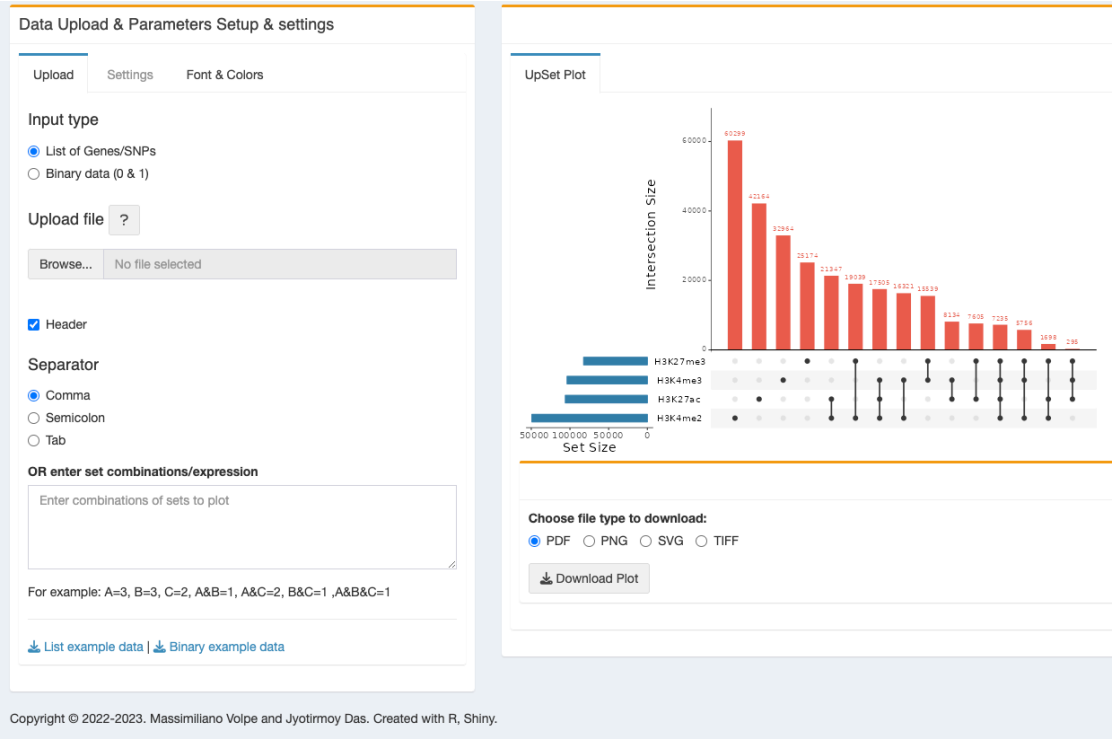

Figure 10.1.: UpSet plot

10.3. R packages used

1. UpSetR
2. intervene

A. Create the input zip file for *methylR*

This section describes how to create a zip archive containing the input files to start the methylysis analysis.

A.1. Methods

We will describe three methods to create a zip file:

1. Windows zip utility
2. 7-zip (https://www.7-zip.org/)
3. Bash script (https://www.github.com/)
4. Command line (Ubuntu Linux)

A.1.1. Description

Users need to collect the Sample_sheet.csv file and all the idat files belonging to the analysis as they come from the sequencer. All the methods require to create a New folder (you can give any name, for example **testData**) and move the *Sample_sheet.csv* file inside. Enter the testData directory and then create a folder named **idat**, then move all the directories generated with the analysis and containing the *idat files* (green and red) into this idat folder. In the end you will get this kind of organisation:

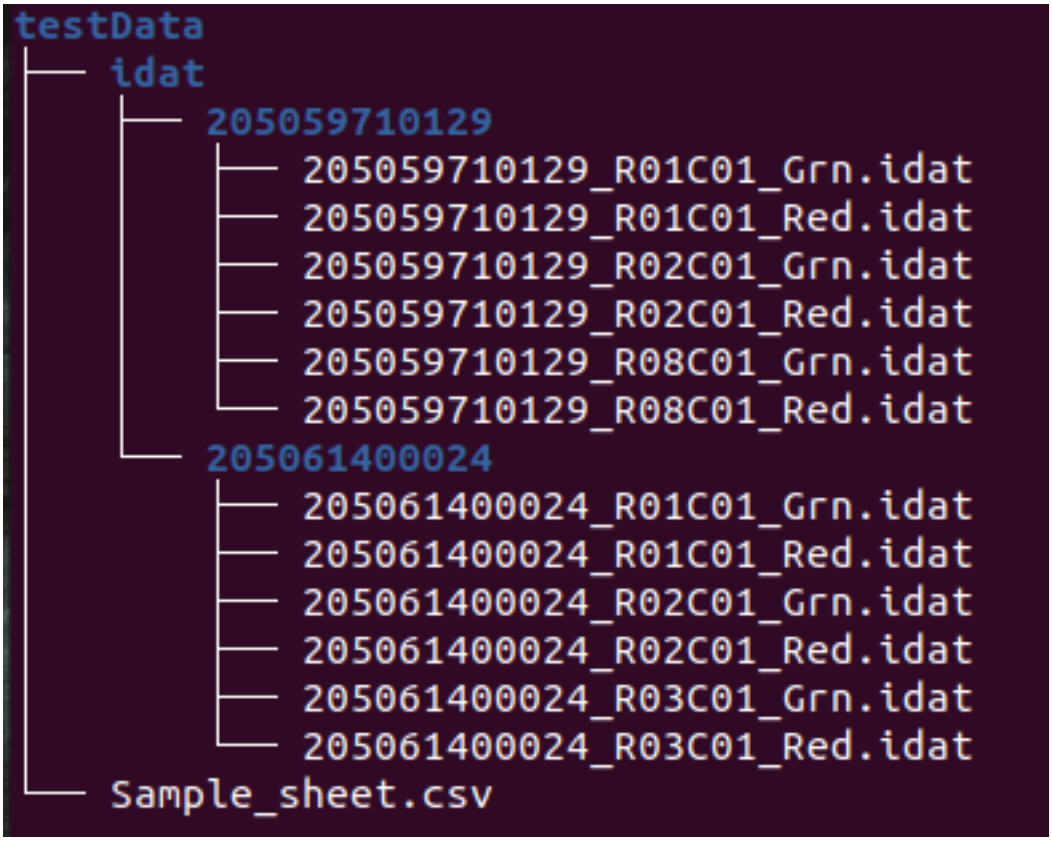

Figure A.1.: How to create zip: Figure 1

A.2. 1. Windows zip utility (Windows 7, 8, 10, 11)

1. Right-click on the New folder you created with the file structure discussed above.
2. Then click Send to > Compressed (zipped) folder

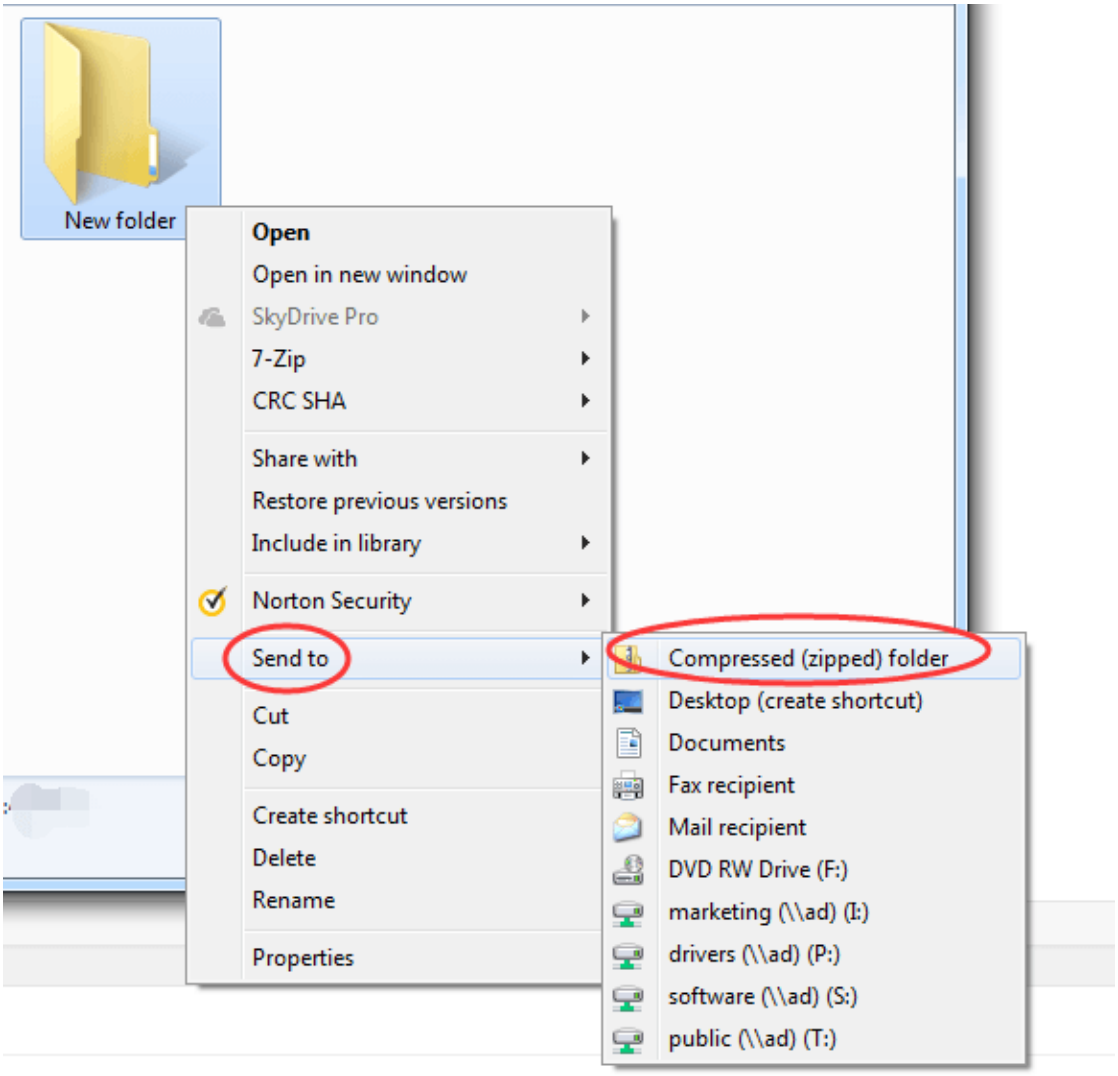

Figure A.2.: How to create zip: Figure 2

A.3. 2. 7-zip utility (Windows 7, 8, 10, 11)

7-Zip is a free open-source file archiver with a high compression ratio. You can use 7-Zip on any computer, including a computer in a commercial organization. You don't need to register or pay for 7-Zip. You can download 7-zip for Windows at (https://www.7-zip.org/). If you have installed 7-zip and want to create the input file for *methylR* you just:

1. Right-click on the New folder you created with the file structure discussed above.
2. Then click 7-Zip > Add to archive...

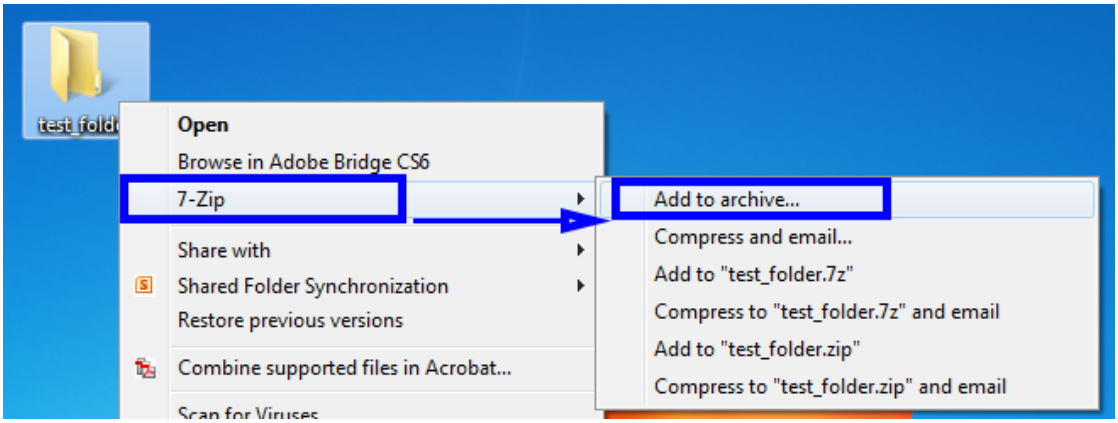

Figure A.3.: How to create zip: Figure 3

A.4. 3. Bash script (MacOS/Linux)

We provide an automathized bash script that is able to create the file structure discussed above for you.

A.4.1. Linux:

Depending on which interface you use (e.g., GNOME, KDE, Xfce), the terminal will be accessed differently. We recommend you check Ubuntu's Using the Terminal page for the several ways to access the terminal.

1. Click Start and search for "Terminal". Alternatively, press *Alt* + *Ctrl* + *t* and type "cmd" then click OK.
2. Then type the following command:

```
cd /path/to/data/
sh script.sh
```

A.4.2. MacOS:

1. You can access the terminal by pressing + space on your keyboard and searching for "terminal".
2. Then type the following command and press Enter:

```
cd /path/to/data/
sh script.sh
```

A.5. 4. Command line (Linux)

Depending on which interface you use (e.g. GNOME, KDE, Xfce), the terminal will be accessed differently. We recommend you check Ubuntu's Using the Terminal page for the several ways to access the terminal.

1. Click Start and search for "Terminal". Alternatively, press *Alt + Ctrl + t* and type "cmd" then click OK.
2. Then move to the New folder and create the zip archive by typing the following command and press Enter:

```
cd /path/to/data/
zip folder/
```

B. Use of Docker Container

B.1. On Windows

> **Danger**
>
> PLEASE NOTE: Only AMD64 OS

1. Please make sure you have installed latest version of Docker Desktop on your Windows machine.
2. Using 'command-prompt' or 'Powershell', run the command `docker pull jd21/methylr:latest`.
3. Open Docker Desktop, under the tab 'images', on the LOCAL images, the docker image will be available as shown in the following figure

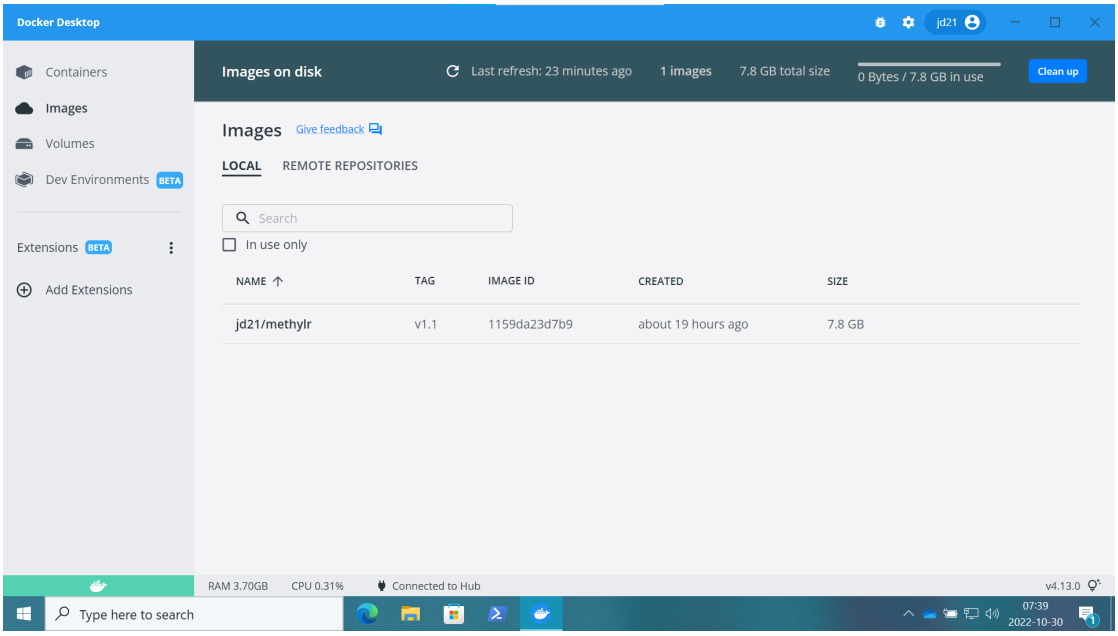

Figure B.1.: Docker container: Figure 1

4. Now, click on the RUN, it will open the 'Optional settings'. Under the 'Optional settings', select the 'Port (Host port)' and write **3838** and click 'RUN'.

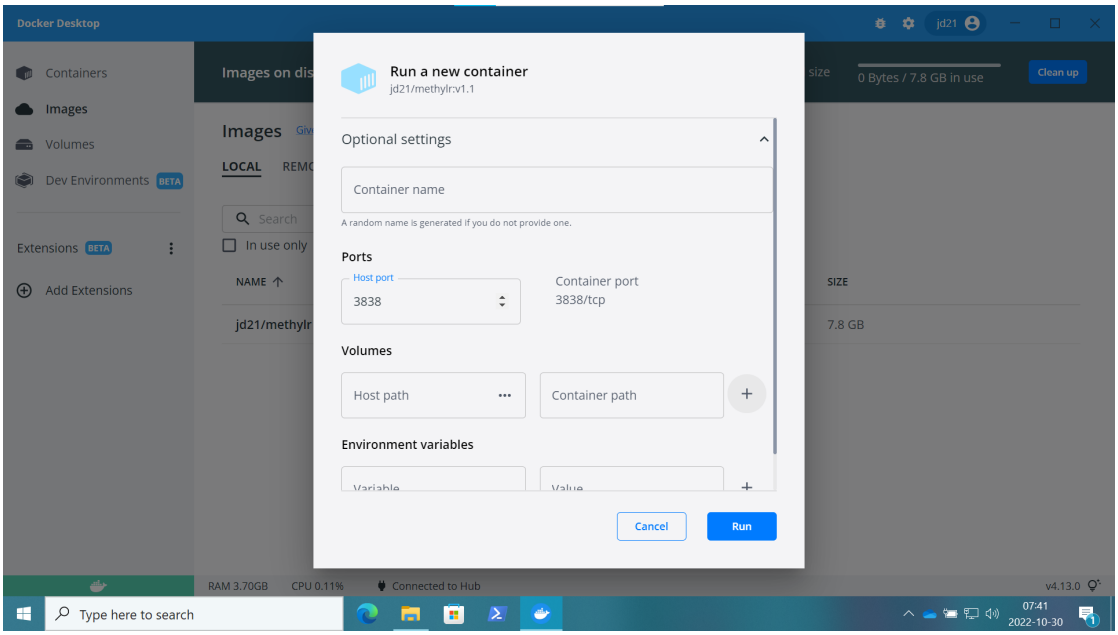

Figure B.2.: Docker image run

5. Click on the **Containers** on the side tab and then click **PORT(S)** '3838:3838'. The default web-browser will open and in a few minutes will start the app (It will take approximately 1-3 minutes to view the app).

> **Note**
>
> **NOTE**: You can copy *http://localhost:3838* after running the container and open it on other web-browser to run the app.

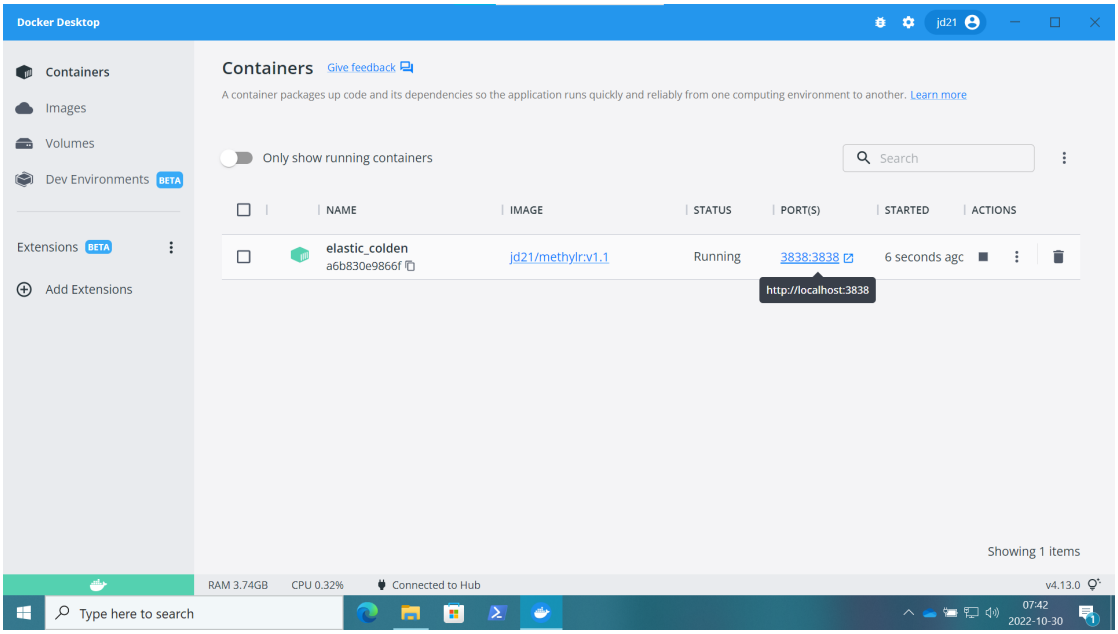

Figure B.3.: Docker container run

B.2. On MacOS

> **Danger**
>
> PLEASE NOTE: (Intel, Only AMD64 OS - not supported on Apple M1/M2 processors)

1. Please make sure you have installed latest version of Docker Desktop on your MacOS.
2. Run the command, `docker pull jd21/methylr:latest` on Mac *terminal*.
3. If you are using the **Docker Desktop** to use *methylR*, please follow the instructions from 3 to 5 as mentioned above for Windows.
4. Alternatively, if you want to use the MacOS *terminal* to run the app (**Only supported on Intel AMD64 OS architecture**), please use this command `docker run --rm -p 3838:3838 jd21/methylr:latest` directly and after pulling all the images by docker, *terminal* will display

```
[2022-10-30T07:57:41.311] [INFO] shiny-server - Shiny Server v1.5.18.979 (Node.js v12.22
[2022-10-30T07:57:41.312] [INFO] shiny-server - Using config file "/etc/shiny-server/shi
[2022-10-30T07:57:41.342] [WARN] shiny-server - Running as root unnecessarily is a secur
[2022-10-30T07:57:41.345] [INFO] shiny-server - Starting listener on http://[::]:3838
```

5. Now, open the web-browser and run `http://localhost:3838` will load the app within 1-3 minutes.

> **Note**
>
> PLEASE NOTE: It may possible that you run the docker container on MacOS Apple M1, but the application may not work as expected. We strongly recommend to use AMD64 OS architecture to run *methylR*

B.3. On Linux (Ubuntu 20.04LTS)

1. If you want to use the linux *terminal* to run *methylR*, use the following command on the *terminal*

```
docker run --rm -p 3838:3838 jd21/methylr:latest
```

2. If you want to use the Docker Desktop for Linux, first pull the docker container using `docker pull jd21/methylr:latest` from *terminal* and then follow Step 3-5 as mentioned above for Windows.

> **Note**
>
> 1. Please contact the IT support if Docker is running properly. You can also contact the developers using the *GitHub* or the *Google groups* or directly email the developer.
> 2. If after uploading the data for methylation analysis (see Chapter 1), the browser get disconnected, please check you have installed the docker or docker-desktop with administrative privilages. From terminal, user can run,
>
> ```
> $ sudo usermod -aG docker $USER
> or
> $ sudo chown $USER /var/run/docker.sock
> ```

C. Convert DMCs table to BED

This section describes how to convert the DMCs table to standard BED format using the ChAMP2bed.py script.

C.1. Method

Python3 must be installed on your system, no additional libraries are required. If your system lacks any python installation, please refer to this page: Python 3 Installation & Setup Guide.

C.1.1. Description

To use ChAMP2bed.py open a terminal and move to the directory storing your main *methylR* results. ChAMP2bed.py must be in the same directory storing your DMCs table:

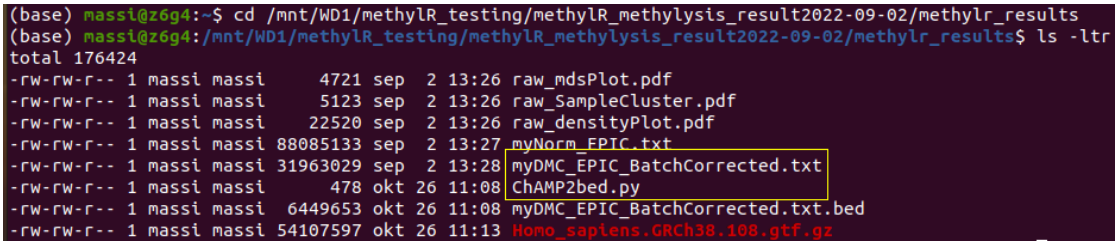

Figure C.1.: BED format: figure 1

Run the command:

```
python3 ChAMP2bed.py myDMC_EPIC_BatchCorrected.txt
```

Or adjust with the actual filename for your table. It will produce a new file with the same filename from your table but with the *.bed* extension:

```
(base) massi@z6g4:~$ cd /mnt/WD1/methylR_testing/methylR_methylysis_result2022-09-02/methylr_results
(base) massi@z6g4:/mnt/WD1/methylR_testing/methylR_methylysis_result2022-09-02/methylr_results$ ls -ltr
total 176424
-rw-rw-r-- 1 massi massi     4721 sep  2 13:26 raw_mdsPlot.pdf
-rw-rw-r-- 1 massi massi     5123 sep  2 13:26 raw_SampleCluster.pdf
-rw-rw-r-- 1 massi massi    22520 sep  2 13:26 raw_densityPlot.pdf
-rw-rw-r-- 1 massi massi 88085133 sep  2 13:27 myNorm_EPIC.txt
-rw-rw-r-- 1 massi massi 31963029 sep  2 13:28 myDMC_EPIC_BatchCorrected.txt
-rw-rw-r-- 1 massi massi      478 okt 26 11:08 ChAMP2bed.py
-rw-rw-r-- 1 massi massi  6449653 okt 26 11:08 myDMC_EPIC_BatchCorrected.txt.bed
-rw-rw-r-- 1 massi massi 54107597 okt 26 11:13 Homo_sapiens.GRCh38.108.gtf.gz
```

Figure C.2.: BED format: figure 2

You can use this file as input for Gviz or import it either in IGV or as a custom track in any other genome browser. Be sure to match the proper genome version used to perform the analysis and to download the correct GTF/GFF3 file if you want to display the CpG ("blue") together with additional features, such as genes ("green"):

Figure C.3.: BED format: figure 3

D. Time calculation

Here we showed the calculation time of each process in *methylR* for both full and lite versions.

module	processes	calculation time (mm:ss)
methylysis	local run - ChAMP	
	(params: BMIQ, batch correction, cores = 4)	03:11 s
	server run - ChAMP	
	(params: BMIQ, batch correction, cores = 2)	03:10 s*
	local run - minfi	
	(params: raw, filters, cores = 4)	04:01 s
	server run - minfi	
	(params: raw, filters, cores = 2)	03:40 s
multi-D		00:2 s
gene features		00:02 s
heatmap		00:01 s
volcano		00:18 s
chromosome		00:04 s
gene ontology		01:30 s
pathway analysis		00:18 s

E. FAQs/Troubleshooting

E.1. Troubleshooting

1. ***Issue with the server***: The University/IT needs to restart the server for maintenance, security updates and it may be down for few hours. Please use the docker container from your local computer or wait few hours before the server gets online again.
2. ***'reload, connection closed'***: Please reload/refresh the page or if the problem persists, close the browser, clear the browser cache and re-open the site.
3. ***calculation time***: We estimated the calculation time based on the provided test data. It varies with the amount of data and parameters chosen, please wait till the process finished.
4. ***error message on local run***: Check your docker permission and docker version.
5. ***check your log file or contact the app author*** Please restart the Docker container and launch the app again. Run it again. If the problem persists, contact us. Make sure your input file has the same format as described in the manual.

References

Bibliography

[1] Martin J Aryee et al. "Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays". In: *Bioinformatics* 30.10 (2014), pp. 1363–1369.

[2] Jyotirmoy Das, Nina Idh, et al. "DNA methylome-based validation of induced sputum as an effective protocol to study lung immunity: construction of a classifier of pulmonary cell types". In: *Epigenetics* (2021), pp. 1–12.

[3] Jyotirmoy Das, Deepti Verma, et al. "Identification of DNA methylation patterns predisposing for an efficient response to BCG vaccination in healthy BCG-naïve subjects". In: *Epigenetics* (Apr. 2019), pp. 1–13. ISSN: 1559-2294. DOI: 10.1080/15592294.2019.1603963. URL: https://www.tandfonline.com/doi/full/10.1080/15592294.2019.1603963.

[4] Sarah Dedeurwaerder et al. "Evaluation of the Infinium Methylation 450K technology". In: *Epigenomics* 3.6 (2011), pp. 771–784.

[5] Jean-Philippe Fortin et al. "Functional normalization of 450k methylation array data improves replication in large cancer studies". In: *Genome biology* 15.11 (2014), pp. 1–17.

[6] Marc Gillespie et al. "The reactome pathway knowledgebase 2022". In: *Nucleic Acids Research* 50.D1 (Nov. 2021), pp. D687–D692. ISSN: 0305-1048. DOI: 10.1093/nar/gkab1028. eprint: https://academic.oup.com/nar/article-pdf/50/D1/D687/42058295/gkab1028.pdf. URL: https://doi.org/10.1093/nar/gkab1028.

[7] Alan Harrison and Anne Parle-McDermott. "DNA methylation: a timeline of methods and applications". In: *Frontiers in genetics* 2 (2011), p. 74.

[8] Eugene Andres Houseman et al. "DNA methylation arrays as surrogate measures of cell mixture distribution". In: *BMC bioinformatics* 13.1 (2012), pp. 1–16.

[9] W Evan Johnson, Cheng Li, and Ariel Rabinovic. "Adjusting batch effects in microarray expression data using empirical Bayes methods". In: *Biostatistics* 8.1 (2007), pp. 118–127.

[10] Minoru Kanehisa and Susumu Goto. "KEGG: Kyoto Encyclopedia of Genes and Genomes". In: *Nucleic Acids Research* 28.1 (Jan. 2000), pp. 27–30. ISSN: 0305-1048. DOI: 10.1093/nar/28.1.27. eprint: https://academic.oup.com/nar/article-pdf/28/1/27/9895154/280027.pdf. URL: https://doi.org/10.1093/nar/28.1.27.

[11] Aziz Khan and Anthony Mathelier. "Intervene: a tool for intersection and visualization of multiple gene or genomic region sets". In: *BMC bioinformatics* 18.1 (2017), pp. 1–8.

[12] DA Lacher. "Interpretation of laboratory results using multidimensional scaling and principal component analysis". In: *Annals of Clinical & Laboratory Science* 17.6 (1987), pp. 412–417.

[13] David A Lacher and ED O'Donnell. "Comparison of multidimensional scaling and principal component analysis of interspecific variation in bacteria". In: *Annals of Clinical & Laboratory Science* 18.6 (1988), pp. 455–462.

[14] Jovana Maksimovic, Lavinia Gordon, and Alicia Oshlack. "SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips". In: *Genome biology* 13.6 (2012), pp. 1–12.

[15] Marvin Martens et al. "WikiPathways: connecting communities". In: *Nucleic Acids Research* 49.D1 (Nov. 2020), pp. D613–D621. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa1024. eprint: https://academic.oup.com/nar/article-pdf/49/D1/D613/35364599/gkaa1024.pdf. URL: https://doi.org/10.1093/nar/gkaa1024.

[16] Lisa D Moore, Thuc Le, and Guoping Fan. "DNA methylation and its basic function". In: *Neuropsychopharmacology* 38.1 (2013), pp. 23–38.

[17] Marie E Mugavin. "Multidimensional scaling: a brief overview". In: *Nursing Research* 57.1 (2008), pp. 64–68.

[18] Alexander R Pico et al. "WikiPathways: Pathway Editing for the People". In: *PLOS Biology* 6.7 (July 2008), pp. 1–4. DOI: 10.1371/journal.pbio.0060184. URL: https://doi.org/10.1371/journal.pbio.0060184.

[19] Jeffrey Skolnick, Jacquelyn S Fetrow, and Andrzej Kolinski. "Structural genomics and its importance for gene function analysis". In: *Nature biotechnology* 18.3 (2000), pp. 283–287.

[20] Andrew E Teschendorff et al. "A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data". In: *Bioinformatics* 29.2 (2013), pp. 189–196.

[21] Nizar Touleimat and Jörg Tost. "Complete pipeline for Infinium® Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation". In: *Epigenomics* 4.3 (2012), pp. 325–341.

[22] Timothy J Triche Jr et al. "Low-level processing of Illumina Infinium DNA methylation beadarrays". In: *Nucleic acids research* 41.7 (2013), e90–e90.

[23] Tianzhi Wu et al. "clusterProfiler 4.0: A universal enrichment tool for interpreting omics data". In: *The Innovation* 2.3 (2021), p. 100141.

[24] Guangchuang Yu et al. "clusterProfiler: an R package for comparing biological themes among gene clusters". In: *Omics: a journal of integrative biology* 16.5 (2012), pp. 284–287.