MethylR-manual

Contents

Keywords. ***DNA methylation***, ***ChAMP***, ***minfi***, ***Volcano plot***, ***multi-dimensional analysis***, ***gene features analysis***, ***pairwise/heatamp***, ***Volcano plot***, ***Chromosome plot***, ***gene ontology***, ***pathway enrichment***, ***set analysis***

Welcome to methylR: DNA Methylation Data Analysis Pipeline

DNA Methylation is one of the most studied epigenetic modifications in humans, playing a critical role in cellular response, development and differentiation [DVGL19]. 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 [MLF13]. 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 [DIS+21]. 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)[HPM11]. 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 Mac and Linux based system.

How to Use

Web server

methylR has a dedicated web server to run all modules. methylr.research.liu.se

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

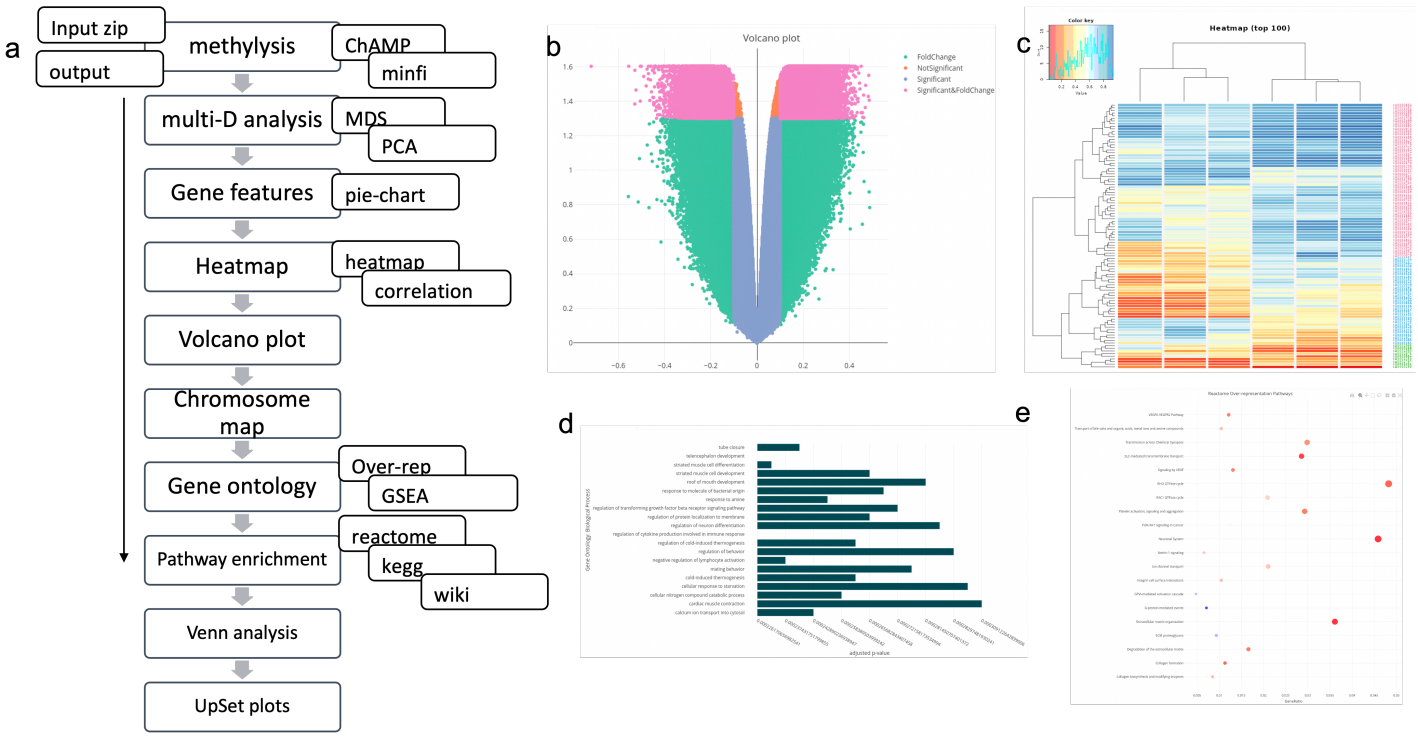
Local use

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.

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 [KM17], we modified as required for *methylr*.

Methylysis

Methylysis is a tool to analyze the DNA methylation data for two different Illumina arrays, 450K and 850K array. The current tool uses two most used well-defined pipelines, named The Chip Analysis Methylation Pipeline ChAMP and minfi pipeline. We introduced different input options for the user to run both pipelines in a more customized way. Based on the above-mentioned pipelines, we selected the user's options. Users can run either pipeline of their choice with different filtration criteria, and of course they can use two different runs and compare the results by themselves using downstream processing, like Venn analysis.



Please note

1. After uploading the zip file [See Chapter 13, how to create the zip file], the pipeline will start immediately by displaying the notification "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 Chapter 14: 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.
2. If the methylysis page goes **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. 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.

How to use

Details are provided below -

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.

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

Note If the user uses Microsoft Excel to build the Sample_sheet.csv, please check that

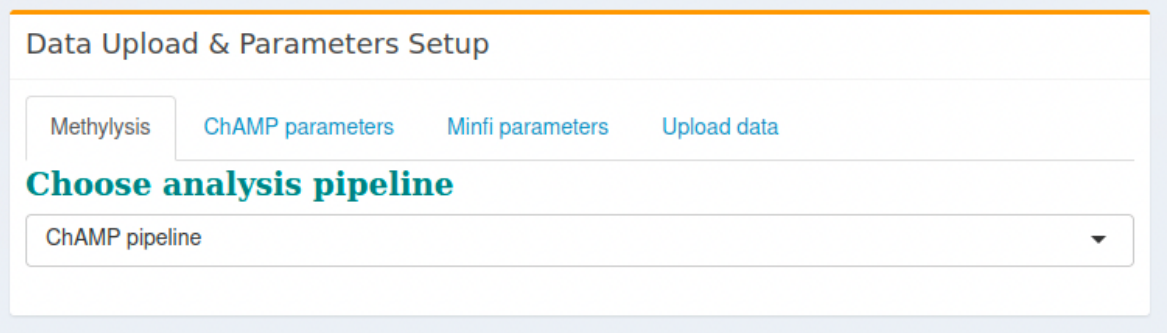
1. Sentrix_ID : are in text format (not in number format, which Excel will change to scientific numbers and will not properly displayed).
2. Optional check: Copy and paste the Excel table of Sample_sheet in some text editor like notepad or VS code and check the format.

To download a template of Sample_sheet.csv, please check here.

Parameters setup

Choose analysis algorithm

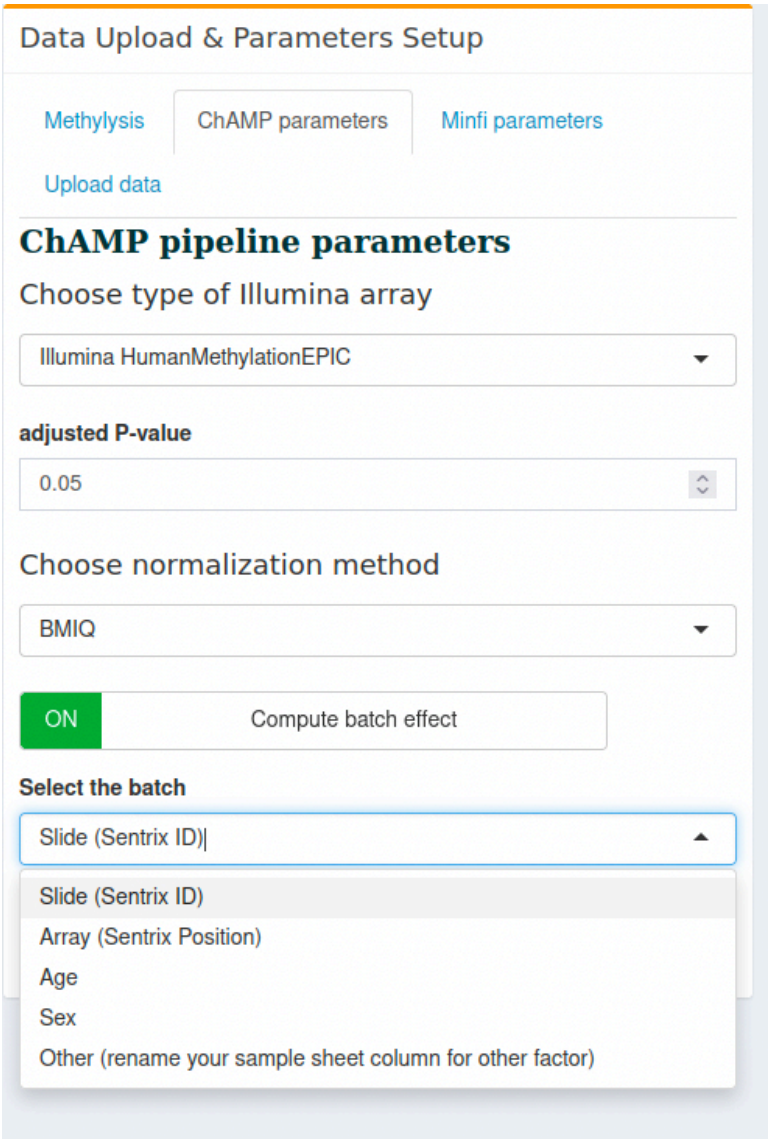
Currently, we have included two most usable algorithms to analyze the data - ChAMP and minfi.



ChAMP pipeline parameters

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) [TML+13],
    - SWAN (Sub-Quantile Within Area) [TT12],[MGO12],
    - PBC (Peak-Based Correction) [DDC+11],
    - FunctionalNormalization [FLL+14]. The default setup will run with the BMIQ normalization method. 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 (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. 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 (OFF), the pipeline will continue without analyzing the batch effect. Please check the reference for the batch effect correction using combat method [JLR07].

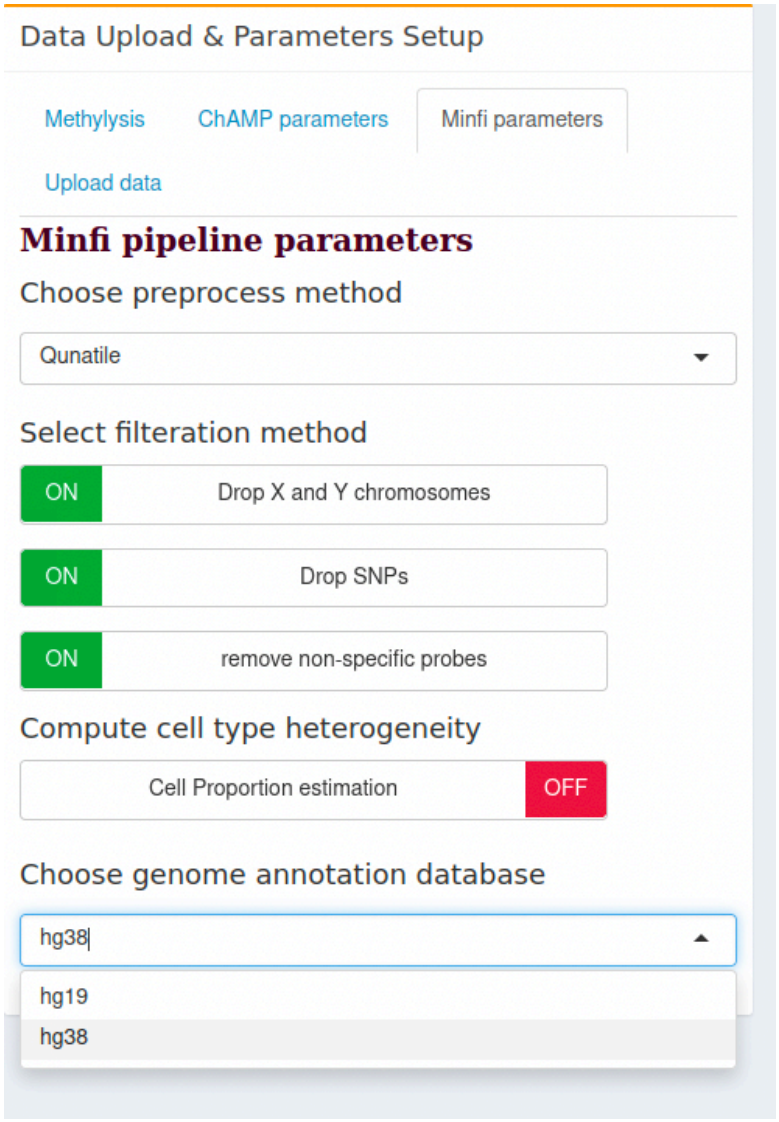
5. *Cell Type Heterogeneity*: Houseman et al (2013)[HAK+12] algorithm is applied to calculate the cell-type heterogeneity from PBMC dataset using the *refbase* function. This is deactivated as default. Press the button to activate and run during the analysis.



minfi pipeline parameters

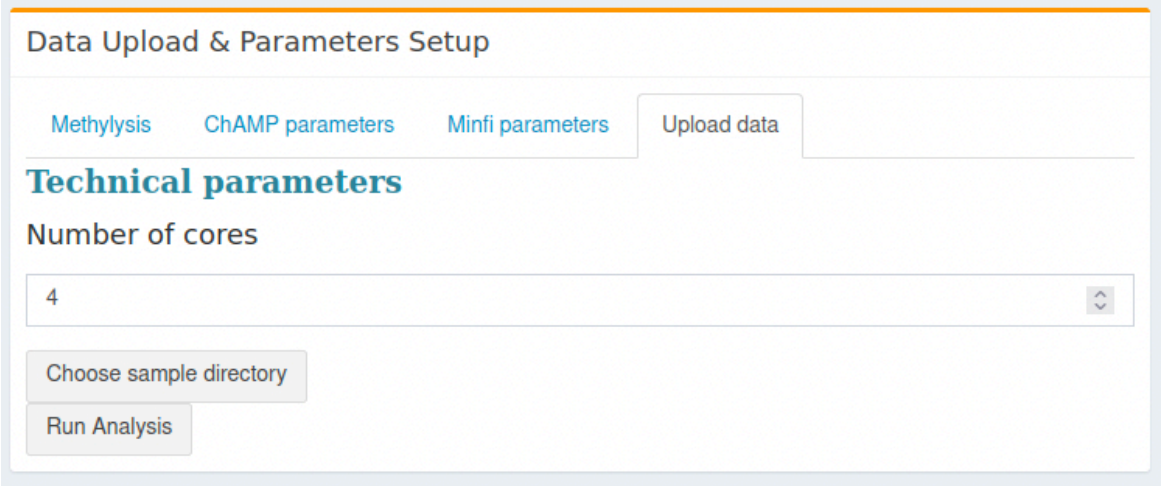
1. *Choose preprocess method* : There are several methods available for preprocessing or normalizing the raw data using the RGset. Here we listed them as the user's input options to select the preprocess/normalization method as per their choice:
    - Raw: No processing of the raw data,
    - SWAN: Subset Quantile Within array Normalization [MGO12],[TT12],
    - Quantile: Quantile normalization method [MGO12],
    - Noob: Noob preprocessing [TJWVDB+13],
    - Illumina: Illumina preprocessing, as performed by Genome Studio (reverse engineered by minfi authors) [AJCB+14]
    - Funnorm: Functional normalization [FLL+14]
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. *Please select the compare groups*: In this option, the pipeline will ask the user to type the group names which will be compared for the differential methylation analysis. To compute the ContrastsMatrix, these group names will be used. Please use the same groups as provide in the Sample_sheet.csv.

5. *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.



Technical setup:

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 2 cores and maximum is 4 cores.
2. *Data upload* The user should set the parameters first and choose all parameters as described above and then upload the data directory. As soon as the pipeline finishes the upload of the data directory, it will start running the analysis.



Requirement for data upload

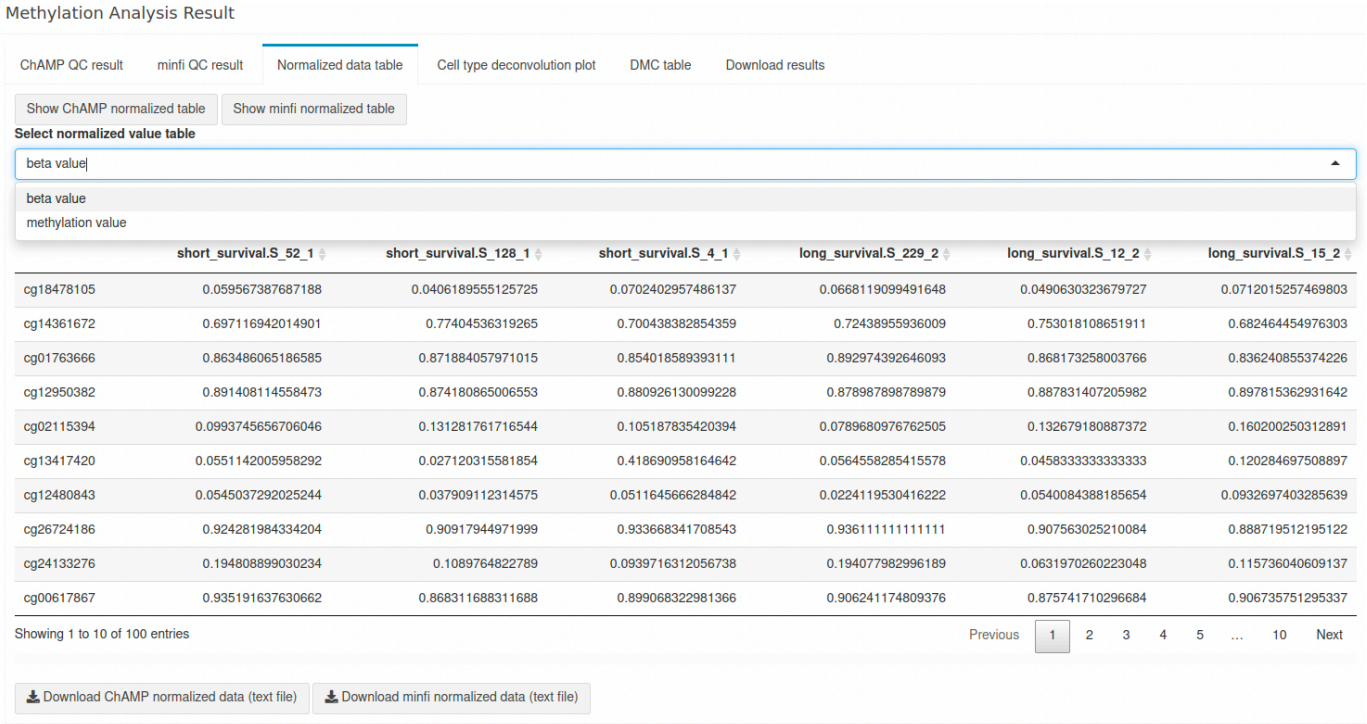
1. idat files: all idat files, green and red as received from Illumina sequencing array should be provided for the analysis. All files should be in one directory/folder.
2. Sample_sheet.csv: the “Sample_sheet.csv” should also be provided in the same directory with idat files.

Analysis Result

We implemented two different approaches to save/view the analysis result. A. First, all results from the analysis pipeline will be saved in the same folder from where the data loaded, and B. Second, using the on-screen visualization tool.

In the visualization tabs of methylysis tool, we introduced 4 tabs for differential methylation analysis result,

1. **QC result**: QC result from each pipeline mentioned above will be generated and presented here in PDF format. User can download the PDF, however the result will also be saved in the data directory.
2. **Normalized data table**: Only first 100 rows from the normalized \beta values will appear in the visualization tab. The full result will be saved on the data directory. For *minfi* analysis, two different options are available, beta value and M-value (methylation value).


Methylation Analysis Result

ChAMP QC result | minfi QC result | Normalized data table | Cell type deconvolution plot | DMC table | Download results

Show ChAMP normalized table | Show minfi normalized table

Select normalized value table

beta value

beta value
methylation value

	short_survival.S_52_1	short_survival.S_128_1	short_survival.S_4_1	long_survival.S_229_2	long_survival.S_12_2	long_survival.S_15_2
cg18478105	0.059567387687188	0.0406189555125725	0.0702402957486137	0.0668119099491648	0.0490630323679727	0.071201525746980
cg14361672	0.697116942014901	0.77404536319265	0.700438382854359	0.72438955936009	0.753018108651911	0.682464454976303
cg01763666	0.863486065186585	0.871884057971015	0.854018589393111	0.892974392646093	0.868173258003766	0.836240855374226
cg12950382	0.891408114558473	0.874180865006553	0.880926130099228	0.878987898789879	0.887831407205982	0.897815362931642
cg02115394	0.0993745656706046	0.131281761716544	0.105187835420394	0.0789680976762505	0.132679180887372	0.160200250312891
cg13417420	0.0551142005958292	0.027120315581854	0.418690958164642	0.0564558285415578	0.0458333333333333	0.120284697508897
cg12480843	0.0545037292025244	0.037909112314575	0.0511645666284842	0.0224119530416222	0.0540084388185654	0.093269740328563
cg26724186	0.924281984334204	0.90917944971999	0.933668341708543	0.936111111111111	0.907563025210084	0.888719512195122
cg24133276	0.194808899030234	0.1089764822789	0.0939716312056738	0.194077982996189	0.0631970260223048	0.115736040609137
cg00617867	0.935191637630662	0.868311688311688	0.899068322981366	0.906241174809376	0.875741710296684	0.906735751295337

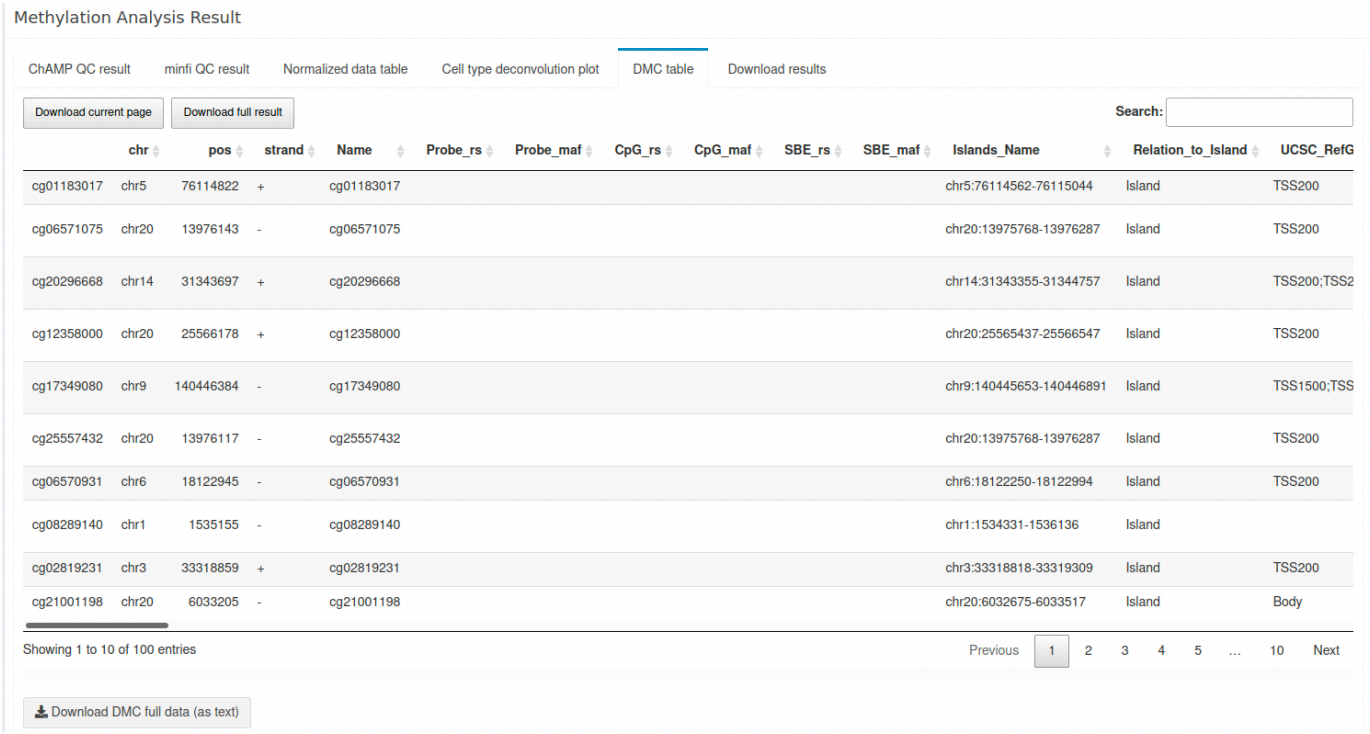
Showing 1 to 10 of 100 entries

Previous | 1 | 2 | 3 | 4 | 5 | ... | 10 | Next

Download ChAMP normalized data (text file) | Download minfi normalized data (text file)

3. **Cell type deconvolution plot**: If user choose to analyze the cell type deconvolution, the analysis figure will appear here. Please note, current version only supports the PBMC deconvolution.

4. **DMC table**: First 100 rows of the differentially methylated CpGs will appear here and the full result will be saved as mentioned above.



5. **Download results**: Users can download all analysis result as a zip, including QC, normalized data, cell type deconvolution data, differentially methylated annotated data. Please note, all these files will be deleted once the user close the browser.

Multi-D Analysis

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 (“Objects” can be anything). 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 [Mug08], [Lac87], [LODonnell88] 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.

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 \beta value table. However the user can use similar tables for the analysis.

Data Upload

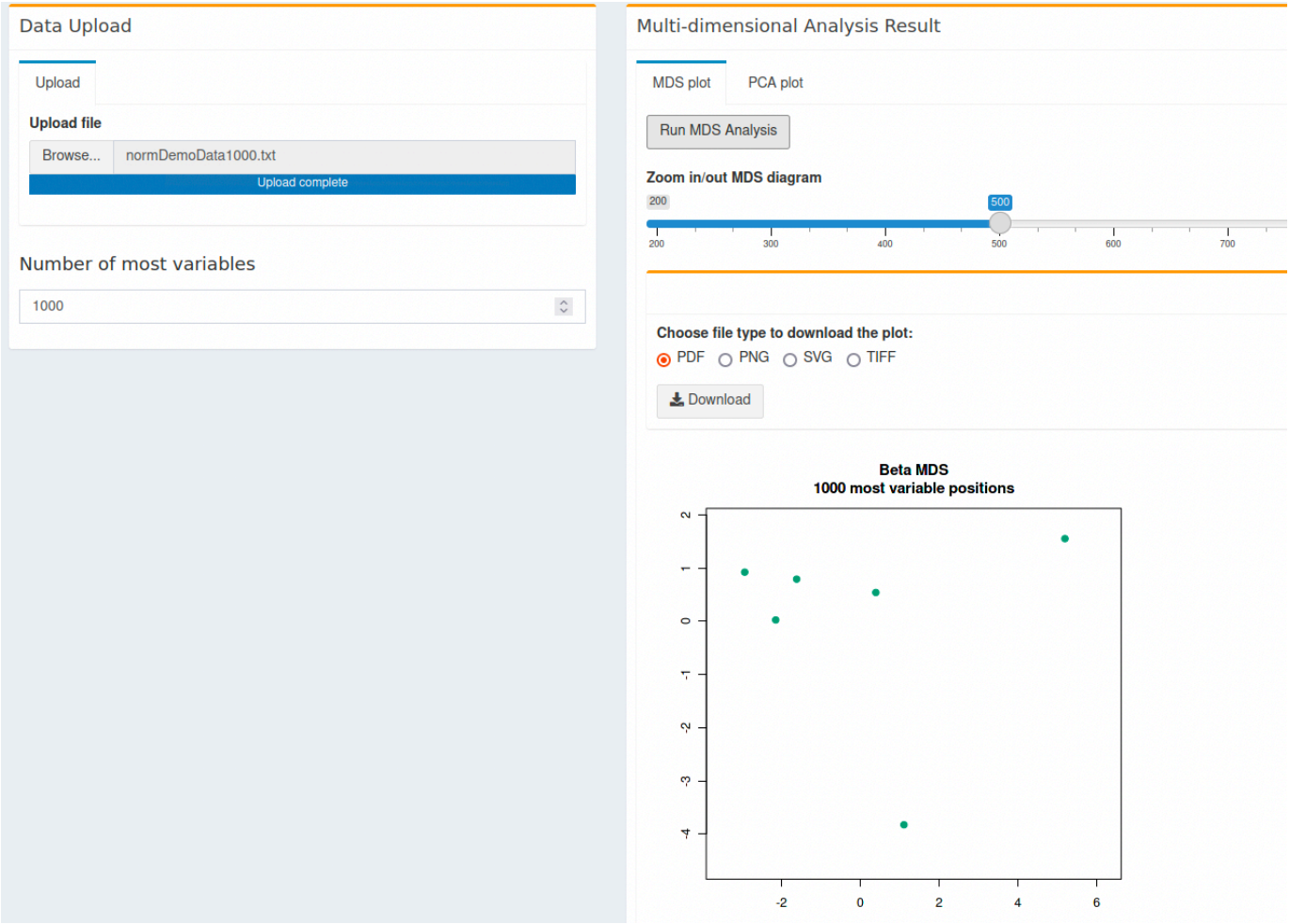
1. Select the text file (tab-delimited file) and upload it.
2. User can choose the option to use number of variables from the uploaded data file.
3. On the right tab, under “MDS plot”, user will find the button to “Run MDS Analysis”
4. Next tab is designed for the PCA plot and here user can have an input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

Analysis result

MDS plot

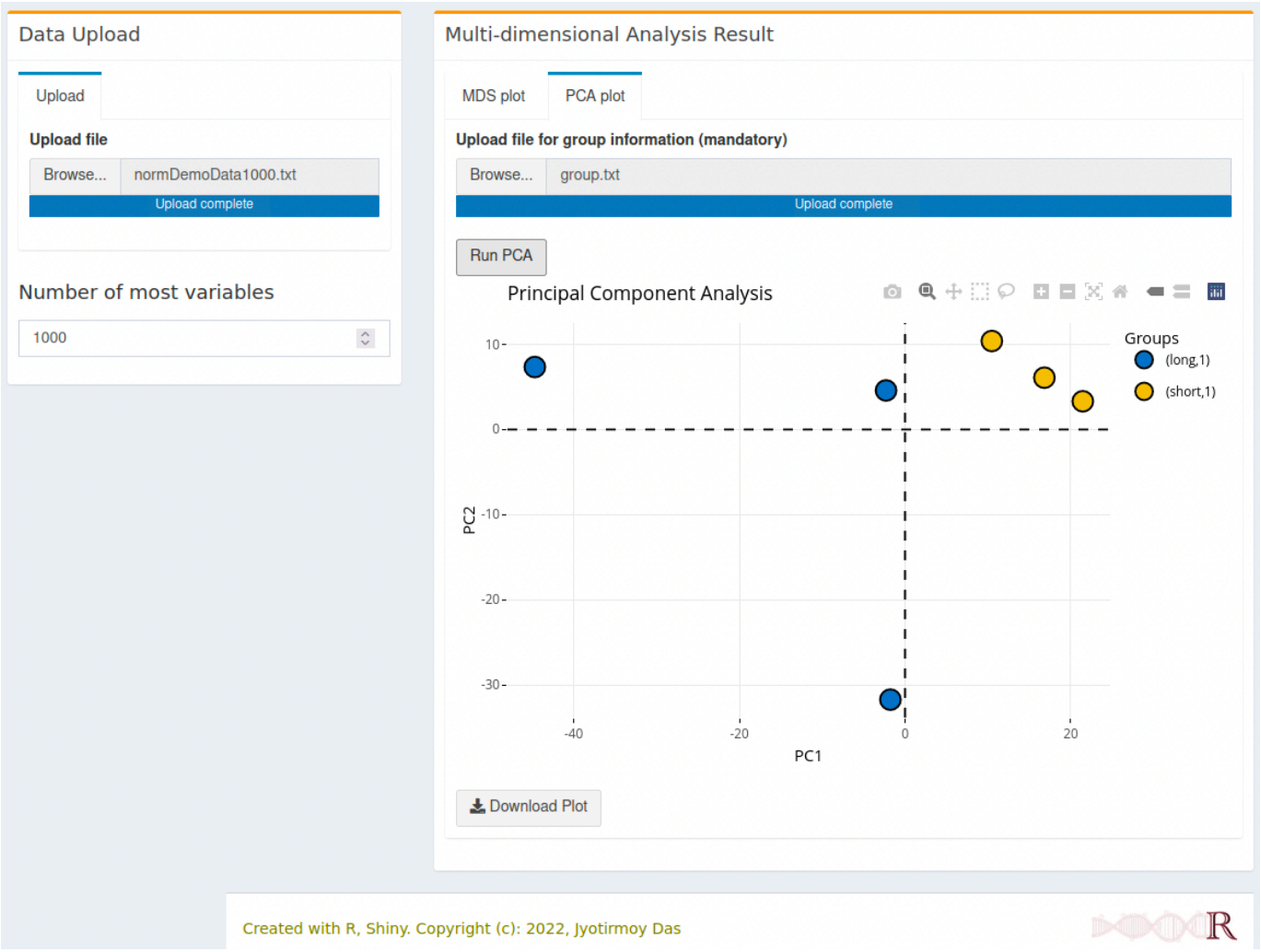
1. After the click on “Run MDS Analysis”, the program will take some time to generate the plot and will appear as soon the run finishes. The zoom bar can be used to zoom in/out the plot.
2. The generated figure can be downloaded in different format, vector graphics support - PDF and SVG or PNG and TIFF.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



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.



R packages used

1. minfi
2. FactoMineR
3. factoExtra
4. explor

Gene Features analysis

Gene Features, here we presented only the structural part of the gene that can be classified as promoter, exon, intron, untranslated regions [SFK00]. 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.

How to use

Data upload & Parameters setup

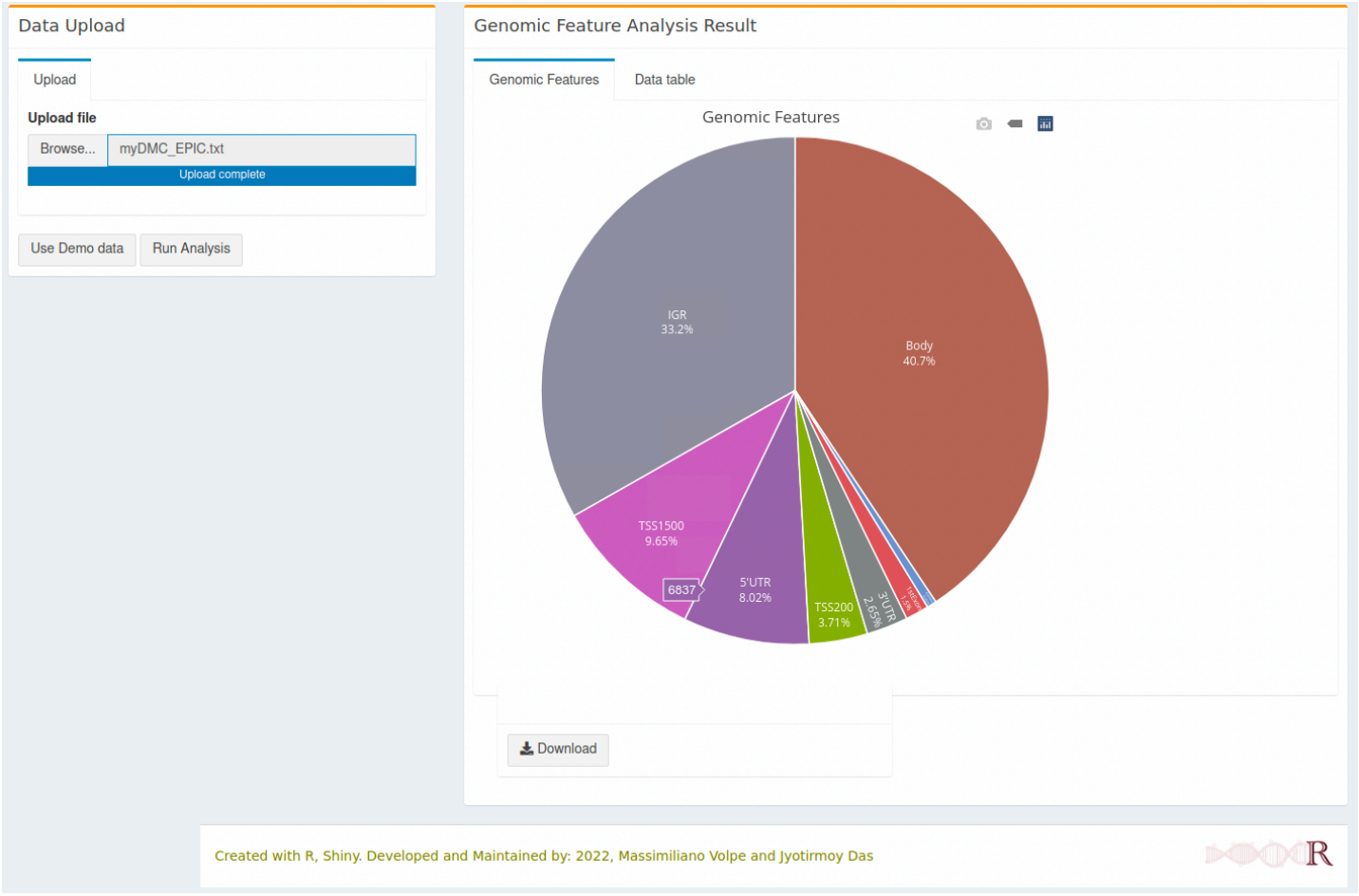
Data upload

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.

Gene Features Analysis Result

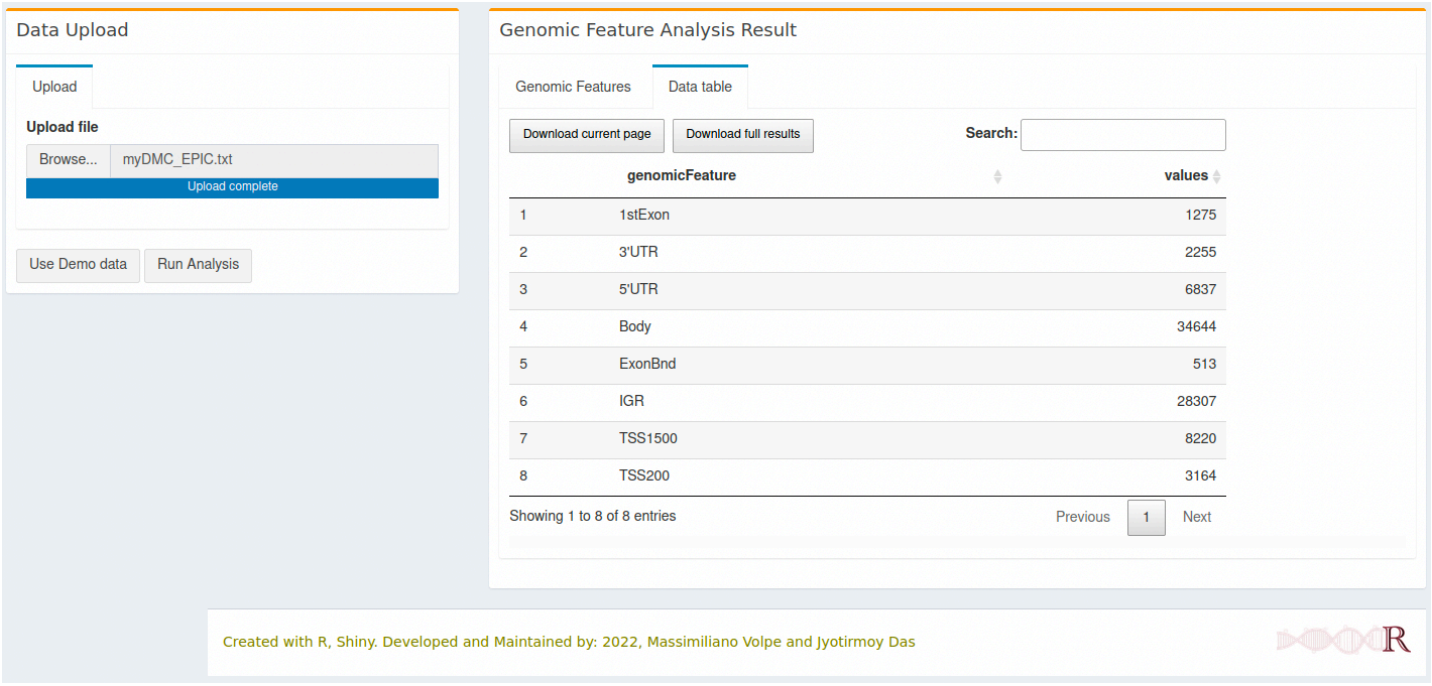
Gene Feature Plot

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



Gene Feature Table

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



R packages used

1. plotly

Heatmap analysis

A heatmap module 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.

How to use

Upload

User can upload the input data matrix in CSV, text or semicolon-separated format. For a better view of the result, we added the functionality to change the number of variables on the heatmap.

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).
3. **Agglomeration method for hclust** - different agglomeration method for hierarchical cluster analysis provided in the module, 1) war.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.

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.

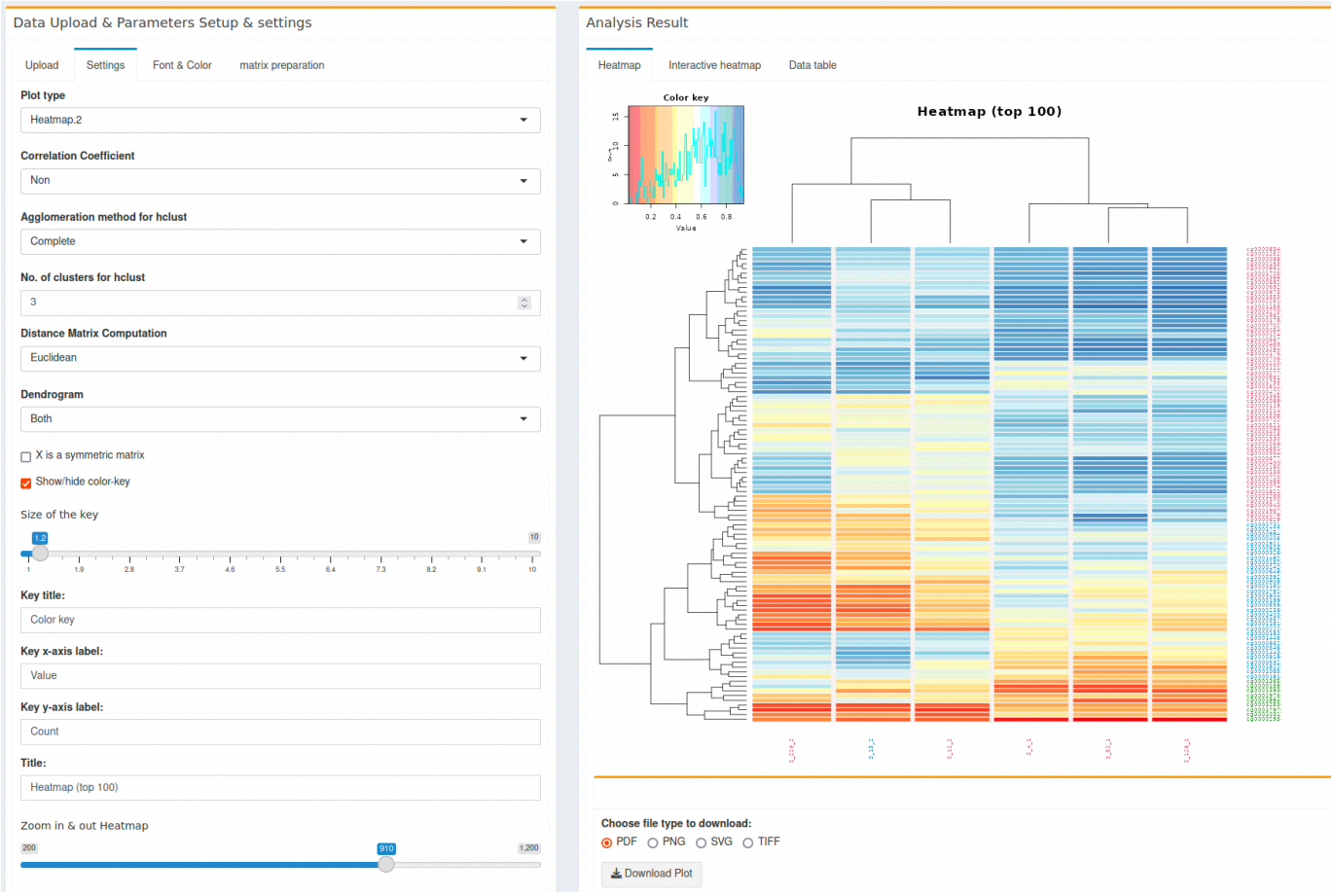
Matrix preparation

We added a tab for the user to build the heatmap matrix from the *methylysis* analysis. The matrix can be uploaded directly on the *Upload* tab to run the heatmap analysis.

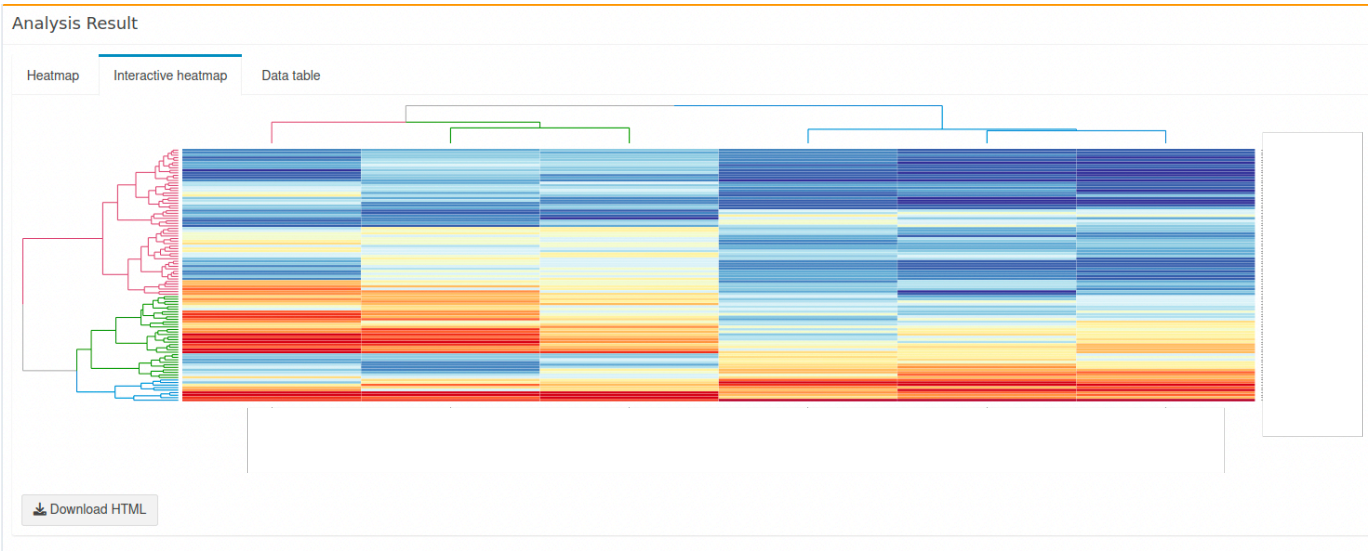
1. **Upload normalized data table** - user can upload the normalized table from the *methylysis* analysis directly without any modification.
2. **Upload DMC data table** - user can upload the differentially methylated CpG data table from the *methylysis* analysis directly without any modification.
3. **Select adjusted P-value** - for more filtration on the dataset, we set a adjusted p-value (BH-corrected as defined in the *methylysis*) parameter. Default is 0.05.
4. **Select logFC value** - for more filtration on the dataset, we set a logFC (as defined in the *methylysis*) parameter. Default is 0.1.

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.



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



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

R packages used

1. heatmap2
2. D3heatmap

Volcano plot

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 this tool, 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).

How to use

Data upload & Parameters setup

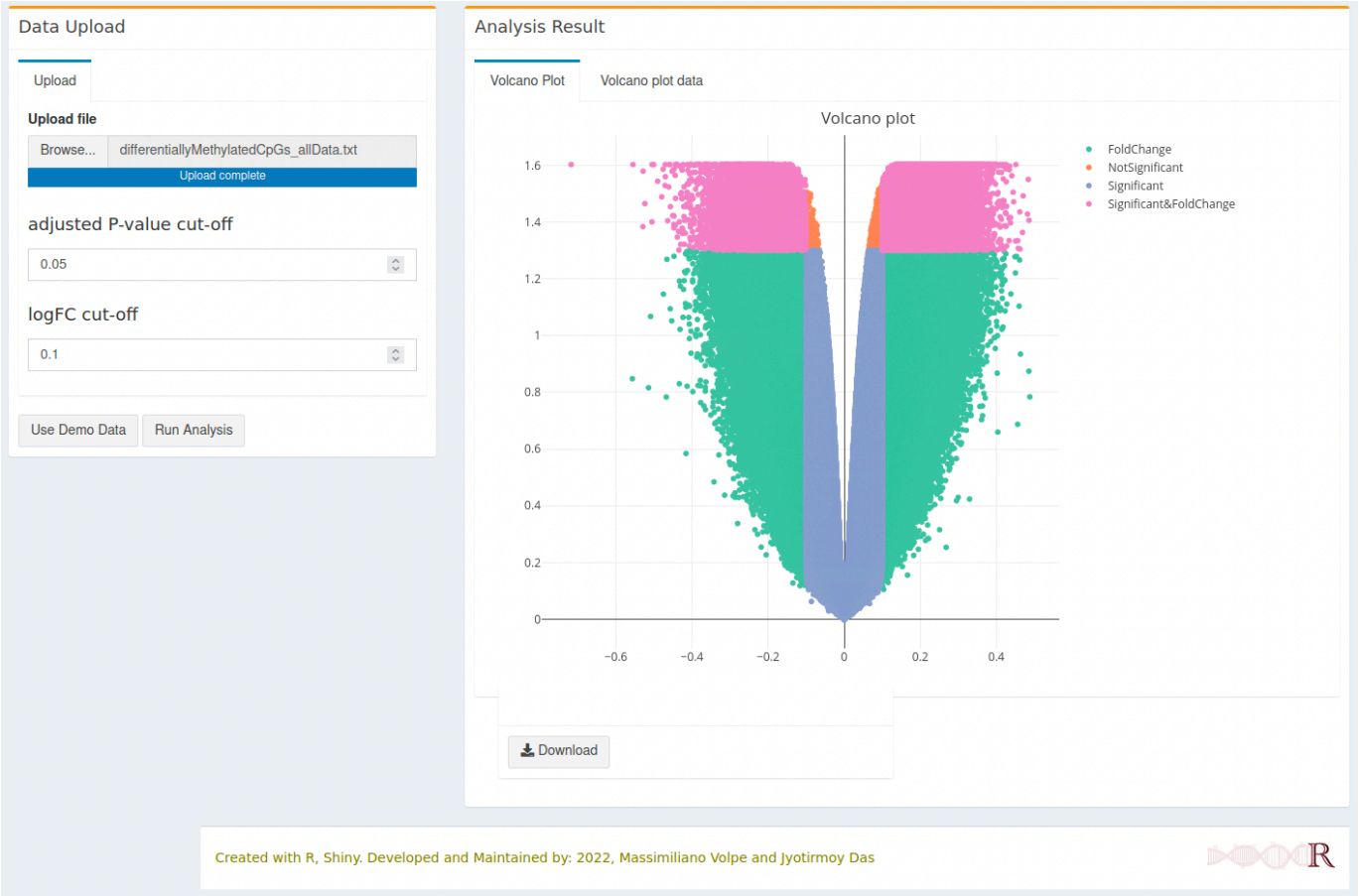
Data upload

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 methylysis run.
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.
4. After the file upload and setting up the cut-off for adjusted P-value and logFC, click the “Run Analysis” button.

Analysis result

Volcano plot

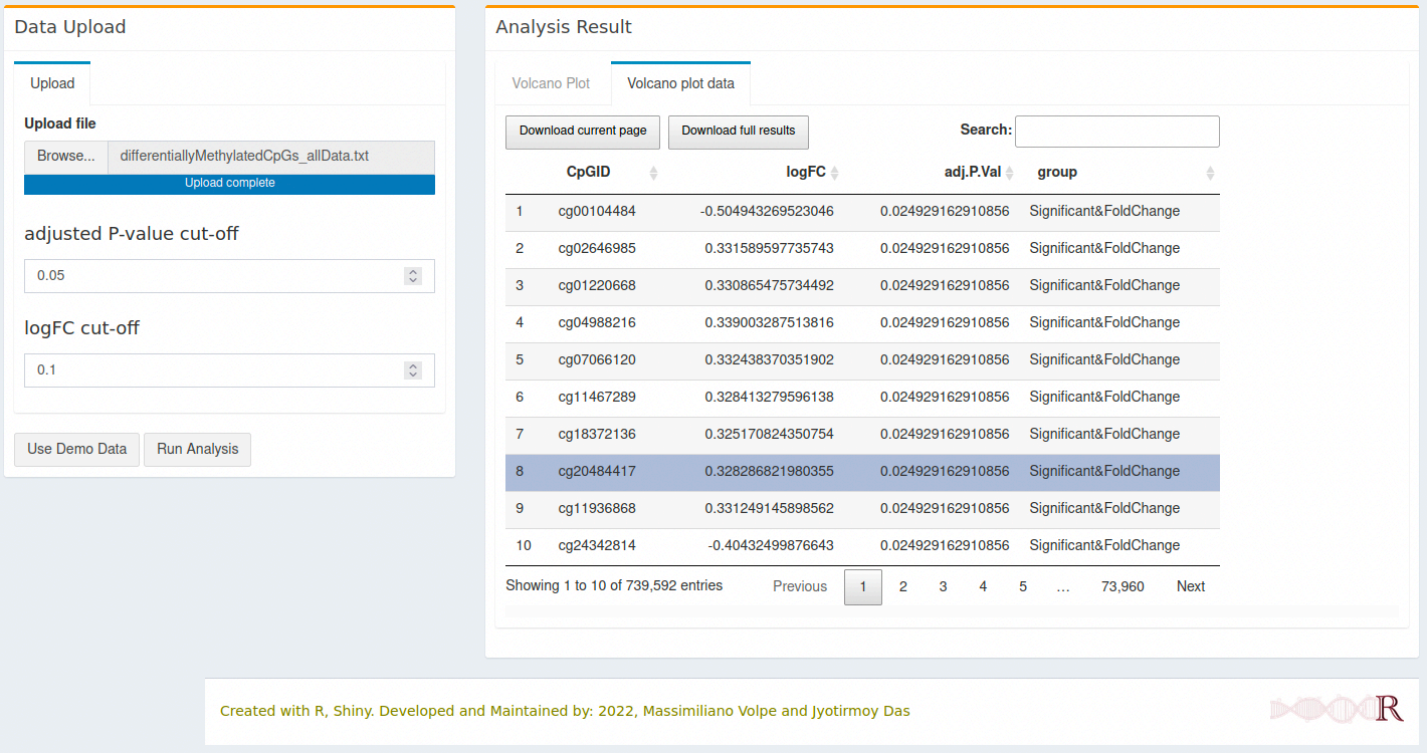
1. The figure will generated as soon as the computation finishes. However, it might takes some more time depending on the size of the file. If user upload a file with 750K rows, it will take 2-3 minutes to generate the figure [See Chapter 14]. It is noteworthy that this big data in volcano plot, may be unstable in the browser.
2. User can download the plot as figure (same as before) and the dynamic figure as a html file.



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 methylysis result.

Volcano data

One data table will be generated using the input data and will have a column marked with “significant & Fold Change”, “Significant” or “insignificant” depends on the adjusted *p*-value and logFC cut-off.



R packages used

1. plotly

Chromosome plot

Chromosome plot is a way to visualize coordinates at DMC positions over the chromosome structure. Users can vary 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.

How to use

Data upload & Parameters setup

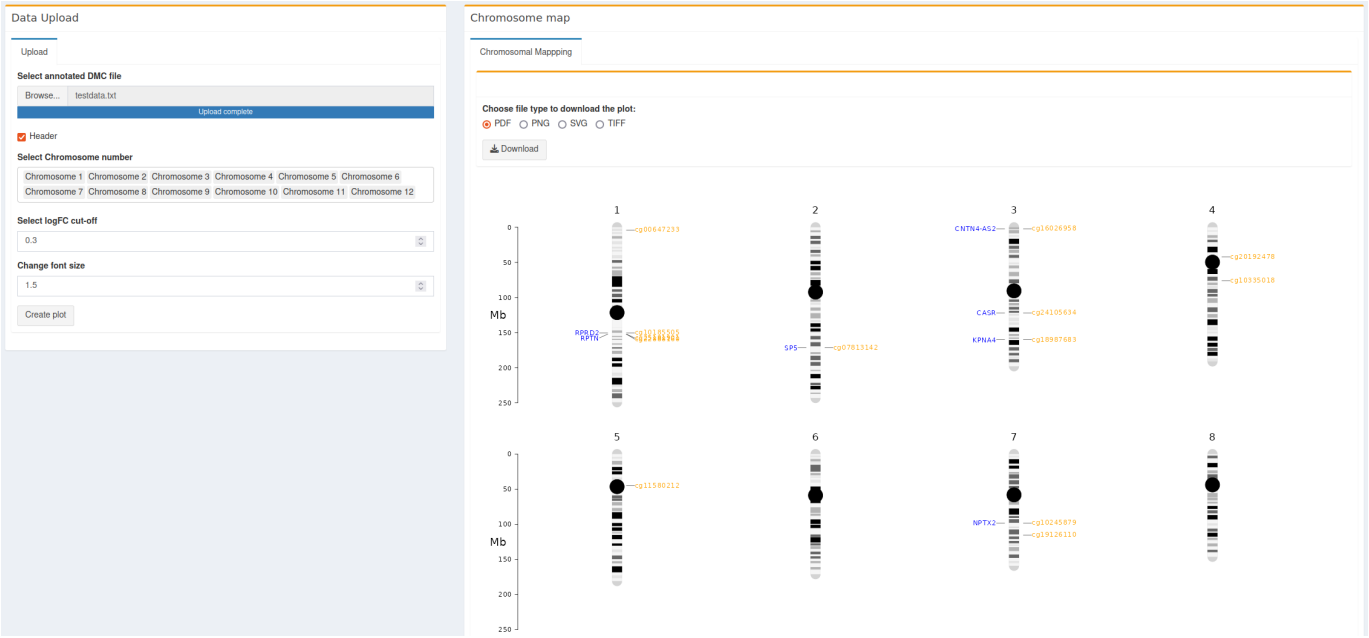
Data upload

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 methylysis run.
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.

Analysis result

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).



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 methylysis result.

R packages

1. Chromplot

Gene Ontology (GO) enrichment analysis

Gene Ontology is a very well-known method for accessing the functions of the gene identified through methylation analysis or expression analysis. Here in this pipeline, we used ontology analysis using the clusterProfiler package (latest version) [YWHH12b], [WHX+21].

How to use

Data upload & Parameters setup

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).
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.

Data upload

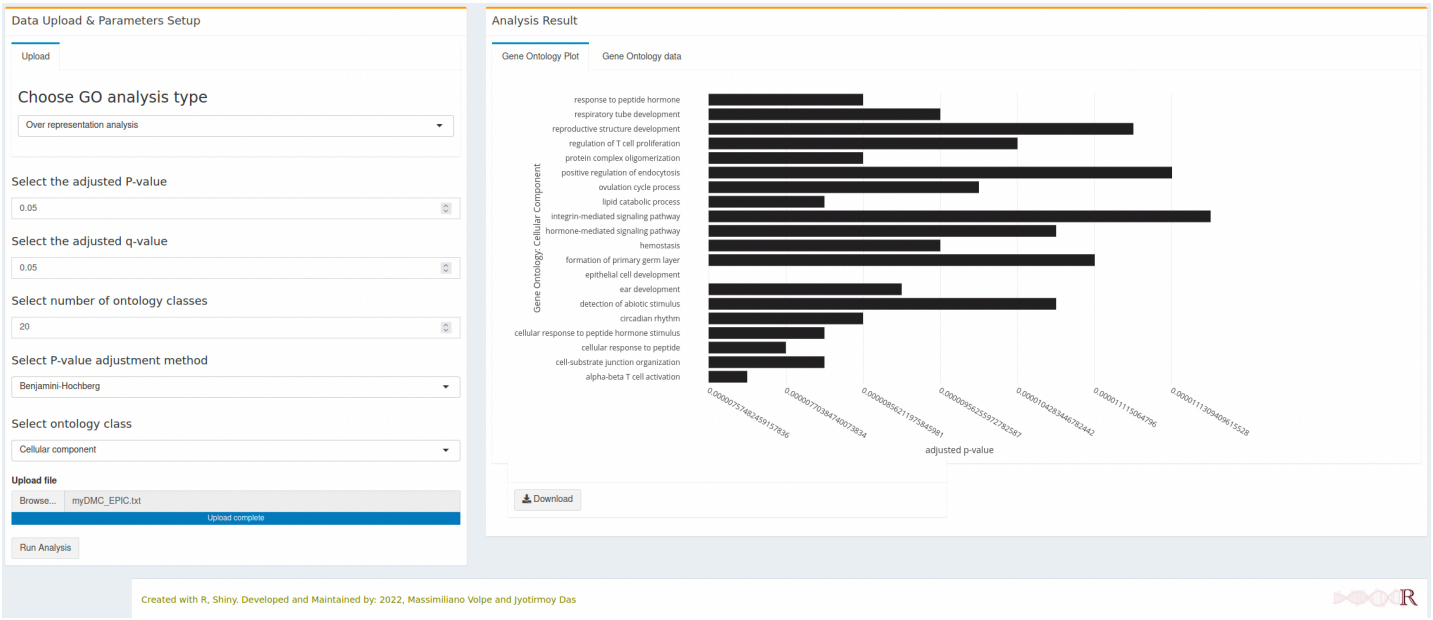
At present, user can upload the DMC data produced by the methylysis pipeline directly. The input file should be in a text (tab-delimited) format.

Analysis result

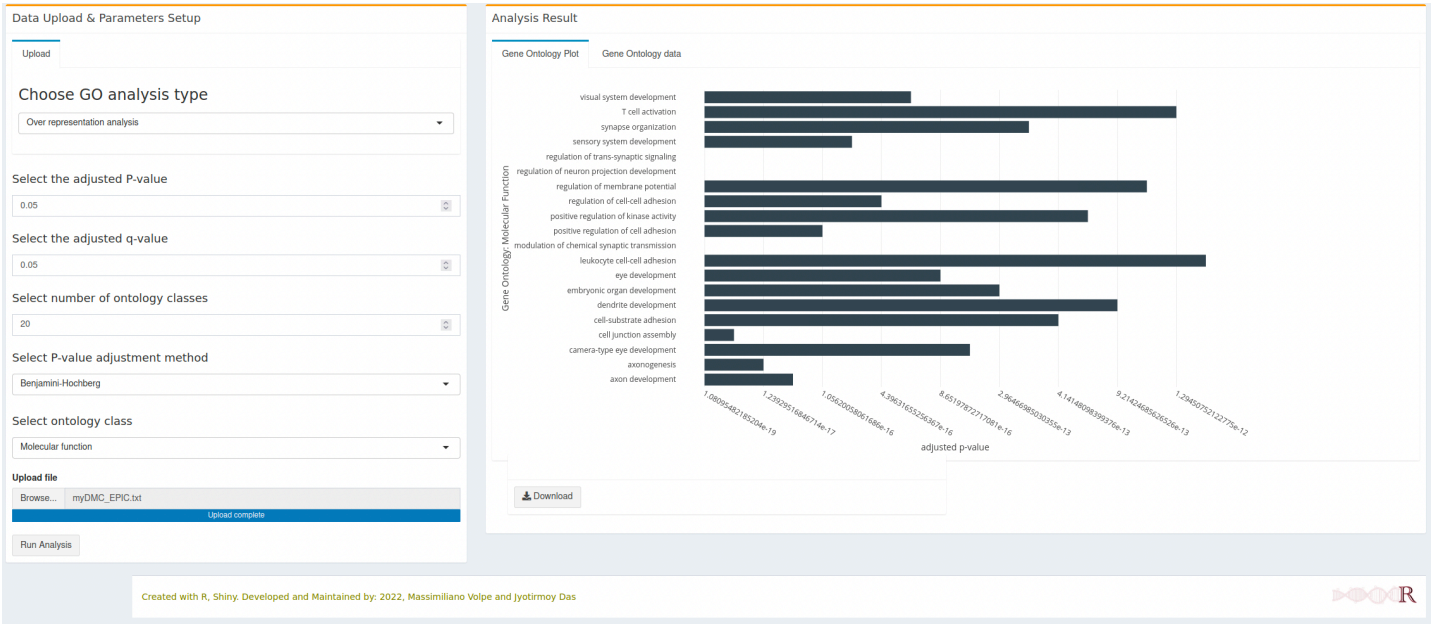
1. On the right tab, the analysis result the plot will be generated as soon as computation finished. The plot is generated with plotly and it will be dynamic in nature as before. User can download the plot as PNG format, zoom in/out or do other stuffs as per plotly figures. The dynamic figure can also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

Biological processes

Cellular component



Molecular function



1. 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.

R packages used

1. bitr
2. clusterProfiler

Pathway enrichment analysis

How to use

Data upload & Parameters setup

Data upload

User can upload the direct output result from the methylysis run.

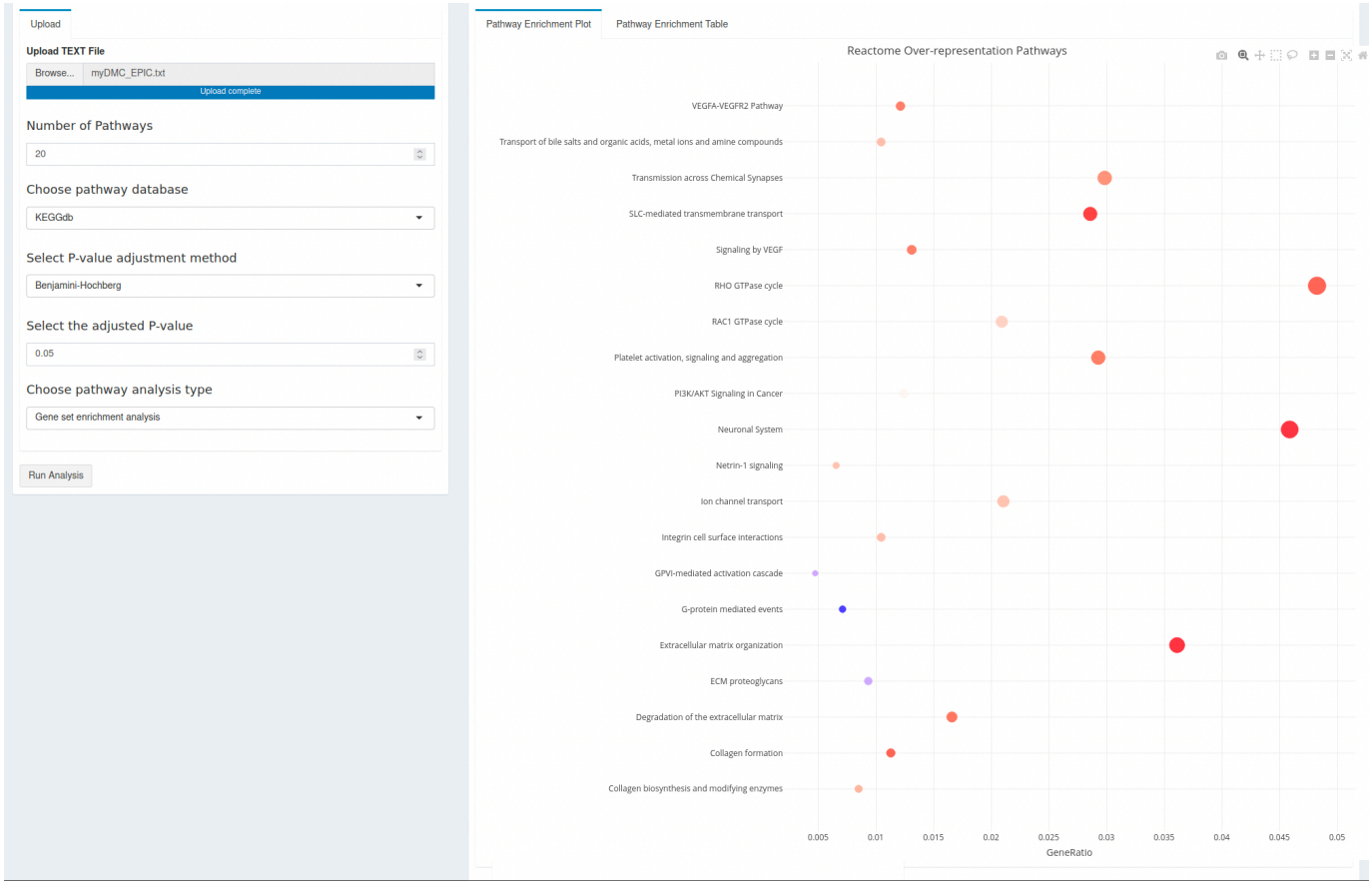
Parameters setup

1. *Choose pathway database*: Please select pathway database for the analysis from the drop-down list
    - ReactomeDB;
    - KEGGdb;
    - Wikipathways By default, the tool will use the ReactomeDB analysis.
2. *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.
3. *Choose pathway database*: user can choose to use different pathway database, namely ReactomeDB, KEGG or wikipathways.
4. *Select p-value adjustment 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)

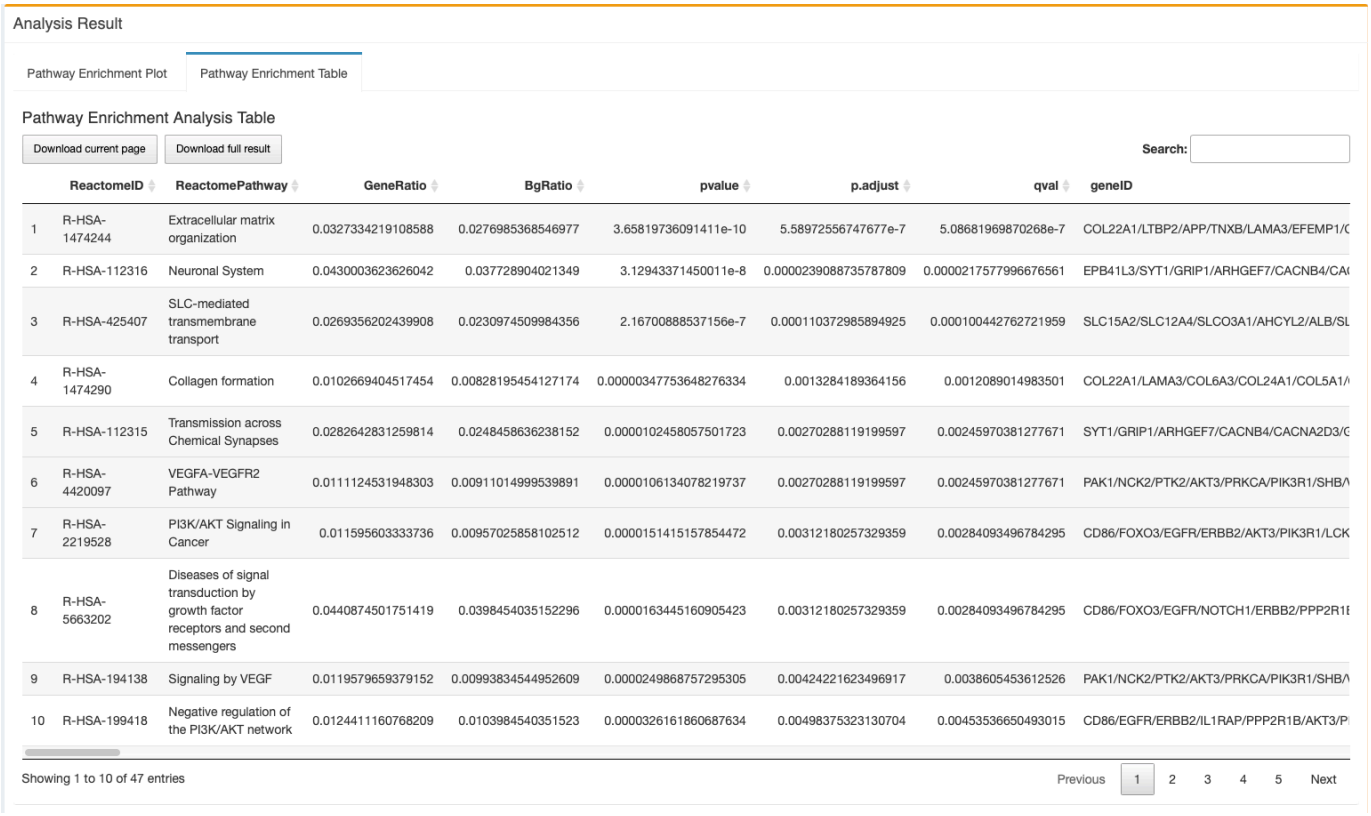
    - Bonferroni
    - Holm
    - Hommel
    - Hochberg
    - FDR
    - none
5. *Choose the adjusted P-value*: The default value for BH-correction is set to 0.05. User can set their own cut-off values.
6. *Choose pathway analysis type*: The pathway analysis type can be chosen from the drop-down list -
    - Over representation analysis (ORA)
    - Gene set enrichment analysis (GSEA).

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 some more time. 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.



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).


Analysis Result

Pathway Enrichment Plot | Pathway Enrichment Table

Pathway Enrichment Analysis Table

Download current page | Download full result | Search:

	ReactomeID	ReactomePathway	GeneRatio	BgRatio	pvalue	p.adjust	qval	geneID
1	R-HSA-1474244	Extracellular matrix organization	0.0327334219108588	0.0276985368546977	3.65819736091411e-10	5.58972556747677e-7	5.08681969870268e-7	COL22A1/LTBP2/APP/TNXB/LAMA3/EFEMP1/(
2	R-HSA-112316	Neuronal System	0.0430003623626042	0.037728904021349	3.12943371450011e-8	0.0000239088735787809	0.0000217577996676561	EPB41L3/SYT1/GRIP1/ARHGEF7/CACNB4/CA(
3	R-HSA-425407	SLC-mediated transmembrane transport	0.0269356202439908	0.0230974509984356	2.16700888537156e-7	0.00011037298589492 5	0.000100442762721959	SLC15A2/SLC12A4/SLCO3A1/AHCYL2/ALB/SL
4	R-HSA-1474290	Collagen formation	0.0102669404517454	0.00828195454127174	0.0000347753648276334	0.0013284189364156	0.0012089014983501	COL22A1/LAMA3/COL6A3/COL24A1/COL5A1/(
5	R-HSA-112315	Transmission across Chemical Synapses	0.0282642831259814	0.0248458636238152	0.000102458057501723	0.00270288119199597	0.00245970381277671	SYT1/GRIP1/ARHGEF7/CACNB4/CACNA2D3/(
6	R-HSA-4420097	VEGFA-VEGFR2 Pathway	0.0111124531948303	0.00911014999539891	0.000106134078219737	0.00270288119199597	0.00245970381277671	PAK1/NCK2/PTK2/AKT3/PRKCA/PIK3R1/SHB/(
7	R-HSA-2219528	PI3K/AKT Signaling in Cancer	0.011595603333736	0.00957025858102512	0.000151415157854472	0.00312180257329359	0.00284093496784295	CD86/FOXO3/EGFR/ERBB2/AKT3/PIK3R1/LCK
8	R-HSA-5663202	Diseases of signal transduction by growth factor receptors and second messengers	0.0440874501751419	0.0398454035152296	0.000163445160905423	0.00312180257329359	0.00284093496784295	CD86/FOXO3/EGFR/NOTCH1/ERBB2/PPP2R1I
9	R-HSA-194138	Signaling by VEGF	0.0119579659379152	0.00993834544952609	0.000249868757295305	0.00424221623496917	0.0038605453612526	PAK1/NCK2/PTK2/AKT3/PRKCA/PIK3R1/SHB/(
10	R-HSA-199418	Negative regulation of the PI3K/AKT network	0.0124411160768209	0.0103984540351523	0.000326161860687634	0.00498375323130704	0.00453536650493015	CD86/EGFR/ERBB2/IL1RAP/PPP2R1B/AKT3/P

Showing 1 to 10 of 47 entries | Previous | 1 | 2 | 3 | 4 | 5 | Next

R packages used

1. clusterProfiler
2. bitr

Venn analysis

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.

How to use

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

Data upload & Parameters setup

Parameters setup

1. *Upload*: Data can be uploaded as TEXT or CSV or semicolon separated format.
2. *Settings*: Under settings, there are multiple options to display the plot - i. *Venn type*: different type of venn diagram can be selected from the drop-down menu
    - Chow-Ruskey
    - Classical
    - Edwards
    - Square

- Battle

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

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

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

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

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

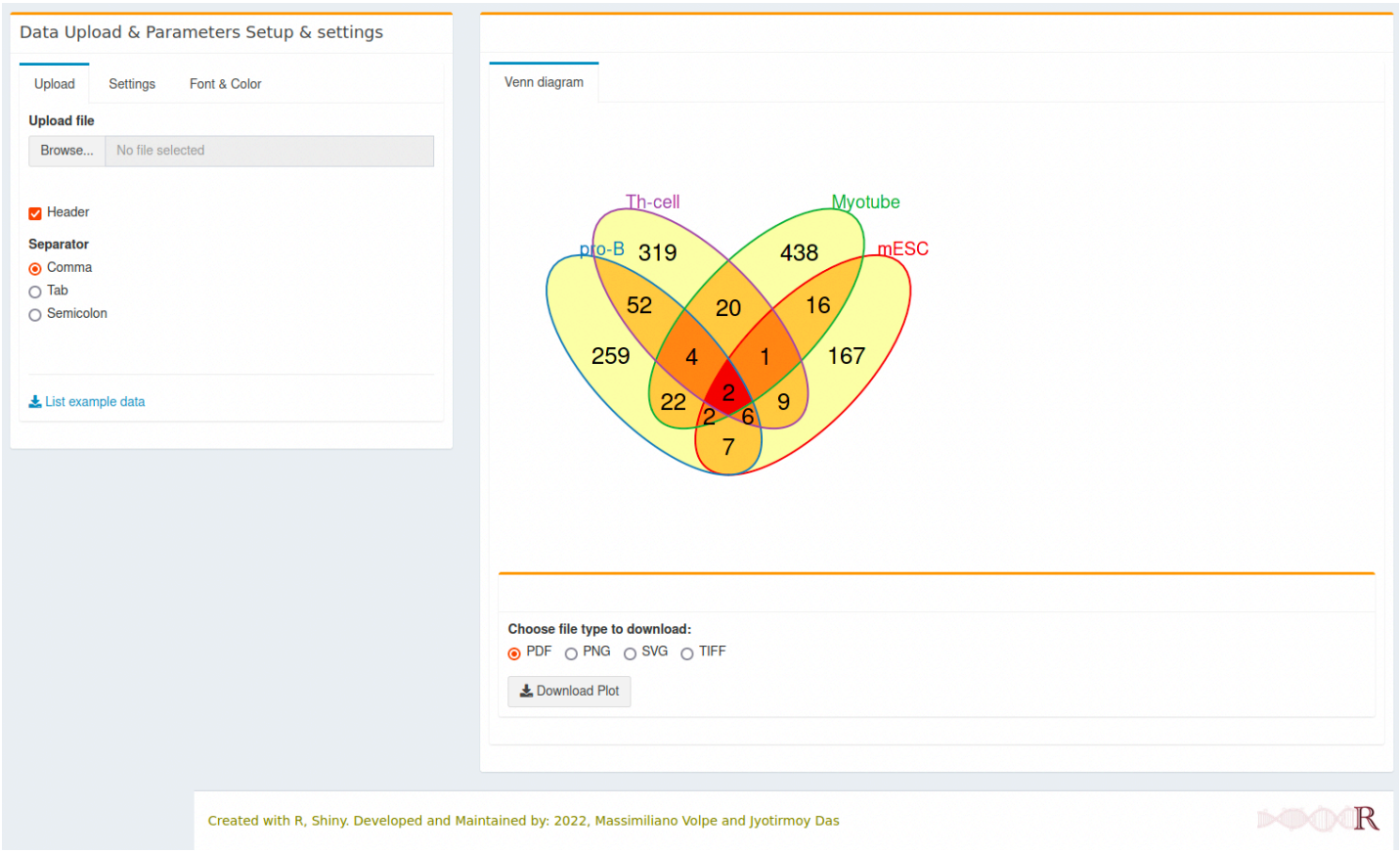
i. *Select color theme*: Colour theme can be chosen from the drop-down menu.

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

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

Results

User can download the figure in different format.



R packages used

1. Vennerable
2. readr

UpSet Plots

UpSet plot will show the relation between different sets.

How to use

Data upload

1. *Upload*: Data can be uploaded as TEXT or CSV or Semicolon separated format. Please look into the example data file. 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.

Parameters setup

1. *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.
2. *Font & Color*: multiple options 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.

Result

User can download the figure in different format.

R packages used

1. UpSetR

References

Ana19 Lakshay Anand. chromoMap: An R package for Interactive Visualization and Annotation of Chromosomes. *bioRxiv*, 2019. URL: http://dx.doi.org/10.1101/605600, doi:10.1101/605600.

[AJCB+14] Martin J Aryee, Andrew E Jaffe, Hector Corrada-Bravo, Christine Ladd-Acosta, Andrew P Feinberg, Kasper D Hansen, and Rafael A Irizarry. Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium dna methylation microarrays. *Bioinformatics*, 30(10):1363–1369, 2014.

[DIS+21] Jyotirmoy Das, Nina Idh, Liv Ingunn Bjoner Sikkeland, Jakob Paues, and Maria Lerm. Dna methylome-based validation of induced sputum as an effective protocol to study lung immunity: construction of a classifier of pulmonary cell types. *Epigenetics*, pages 1–12, 2021.

[DVGL19] Jyotirmoy Das, Deepti Verma, Mika Gustafsson, and Maria Lerm. Identification of DNA methylation patterns predisposing for an efficient response to BCG vaccination in healthy BCG-naïve subjects. *Epigenetics*, pages 1–13, apr 2019. URL: https://www.tandfonline.com/doi/full/10.1080/15592294.2019.1603963, doi:10.1080/15592294.2019.1603963.

[DDC+11] Sarah Dedeurwaerder, Matthieu Defrance, Emilie Calonne, Hélene Denis, Christos Sotiriou, and François Fuks. Evaluation of the infinium methylation 450k technology. *Epigenomics*, 3(6):771–784, 2011.

[FLL+14](1,2) Jean-Philippe Fortin, Aurélie Labbe, Mathieu Lemire, Brent W Zanke, Thomas J Hudson, Elana J Fertig, Celia MT Greenwood, and Kasper D Hansen. Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome biology*, 15(11):1–17, 2014.

[HPM11] Alan Harrison and Anne Parle-McDermott. Dna methylation: a timeline of methods and applications. *Frontiers in genetics*, 2:74, 2011.

[HAK+12]

Eugene Andres Houseman, William P Accomando, Devin C Koestler, Brock C Christensen, Carmen J Marsit, Heather H Nelson, John K Wiencke, and Karl T Kelsey. Dna methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics*, 13(1):1–16, 2012.

[JLR07] W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

[KM17] Aziz Khan and Anthony Mathelier. Intervene: a tool for intersection and visualization of multiple gene or genomic region sets. *BMC bioinformatics*, 18(1):1–8, 2017.

[Lac87] DA Lacher. Interpretation of laboratory results using multidimensional scaling and principal component analysis. *Annals of Clinical & Laboratory Science*, 17(6):412–417, 1987.

[LODonnell88] David A Lacher and ED O'Donnell. Comparison of multidimensional scaling and principal component analysis of interspecific variation in bacteria. *Annals of Clinical & Laboratory Science*, 18(6):455–462, 1988.

MRS+18 Martin Maechler, Peter Rousseeuw, Anja Struyf, Mia Hubert, Kurt Hornik, Matthias Studer, Pierre Roudier, Juan Gonzalez, and Kamil Kozlowski. *Cluster Analysis Basics and Extensions. R package version 2.0.7-1*. online, 2018.

[MGO12] (1,2,3) Jovana Maksimovic, Lavinia Gordon, and Alicia Oshlack. Swan: subset-quantile within array normalization for illumina infinium humanmethylation450 beadchips. *Genome biology*, 13(6):1–12, 2012.

[MLF13] Lisa D Moore, Thuc Le, and Guoping Fan. Dna methylation and its basic function. *Neuropsychopharmacology*, 38(1):23–38, 2013.

[Mug08] Marie E Mugavin. Multidimensional scaling: a brief overview. *Nursing Research*, 57(1):64–68, 2008.

NBacklinW+13 Jessica Nordlund, Christofer L. Bäcklin, Per Wahlberg, Stephan Busche, Eva C. Berglund, Maija Leena Eloranta, Trond Flaegstad, Erik Forestier, Britt Marie Frost, Arja Harila-Saari, Mats Heyman, Ólafur G. Jónsson, Rolf Larsson, Josefine Palle, Lars Rönnblom, Kjeld Schmiegelow, Daniel Sinnett, Stefan Söderhäll, Tomi Pastinen, Mats G. Gustafsson, Gudmar Lönnerholm, and Ann Christine Syvänen. Genome-wide signatures of differential DNA methylation in pediatric acute lymphoblastic leukemia. *Genome Biology*, 2013. doi:10.1186/gb-2013-14-9-r105.

PS19 Emmanuel Paradis and Klaus Schliep. Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 2019. doi:10.1093/bioinformatics/bty633.

[SFK00] Jeffrey Skolnick, Jacquelyn S Fetrow, and Andrzej Kolinski. Structural genomics and its importance for gene function analysis. *Nature biotechnology*, 18(3):283–287, 2000.

[TML+13] Andrew E Teschendorff, Francesco Marabita, Matthias Lechner, Thomas Bartlett, Jesper Tegner, David Gomez-Cabrero, and Stephan Beck. A beta-mixture quantile normalization method for correcting probe design bias in illumina infinium 450 k dna methylation data. *Bioinformatics*, 29(2):189–196, 2013.

[TT12] (1,2) Nizar Touleimat and Jörg Tost. Complete pipeline for infinium® human methylation 450k beadchip data processing using subset quantile normalization for accurate dna methylation estimation. *Epigenomics*, 4(3):325–341, 2012.

[TJWVDB+13] Timothy J Triche Jr, Daniel J Weisenberger, David Van Den Berg, Peter W Laird, and Kimberly D Siegmund. Low-level processing of illumina infinium dna methylation beadarrays. *Nucleic acids research*, 41(7):e90–e90, 2013.

WSanchezCR15 Wencke Walter, Fátima Sánchez-Cabo, and Mercedes Ricote. GOplot: An R package for visually combining expression data with functional analysis. *Bioinformatics*, 2015. doi:10.1093/bioinformatics/btv300.

[WHX+21] Tianzhi Wu, Erqiang Hu, Shuangbin Xu, Meijun Chen, Pingfan Guo, Zehan Dai, Tingze Feng, Lang Zhou, Wenli Tang, Li Zhan, and others. Clusterprofiler 4.0: a universal enrichment tool for interpreting omics data. *The Innovation*, 2(3):100141, 2021.

YH16 Guangchuang Yu and Qing-Yu He. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Molecular bioSystems*, 12(2):477–9, feb 2016. URL: http://www.ncbi.nlm.nih.gov/pubmed/26661513, doi:10.1039/c5mb00663e.

YWHH12a Guangchuang Yu, Li Gen Wang, Yanyan Han, and Qing Yu He. ClusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS A Journal of Integrative Biology*, 2012. doi:10.1089/omi.2011.0118.

[YWHH12b] Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He. Clusterprofiler: an r package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology*, 16(5):284–287, 2012.

ZLS17 Wanding Zhou, Peter W. Laird, and Hui Shen. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic acids research*, 2017. doi:10.1093/nar/gkw967.

DTurner18 Stephen D. Turner. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *Journal of Open Source Software*, 2018. doi:10.21105/joss.00731.

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.

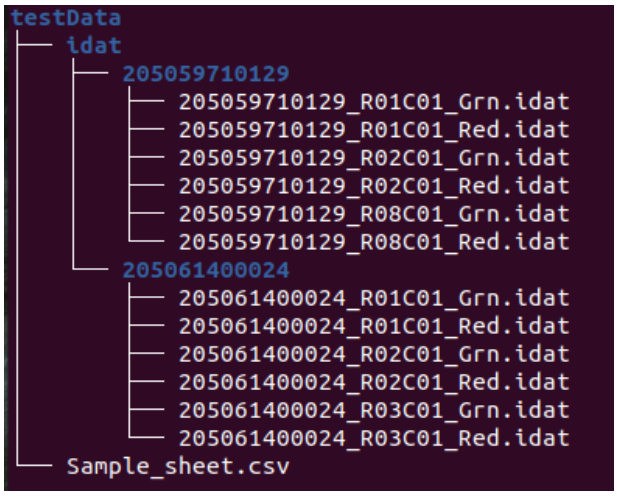
Methods

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)

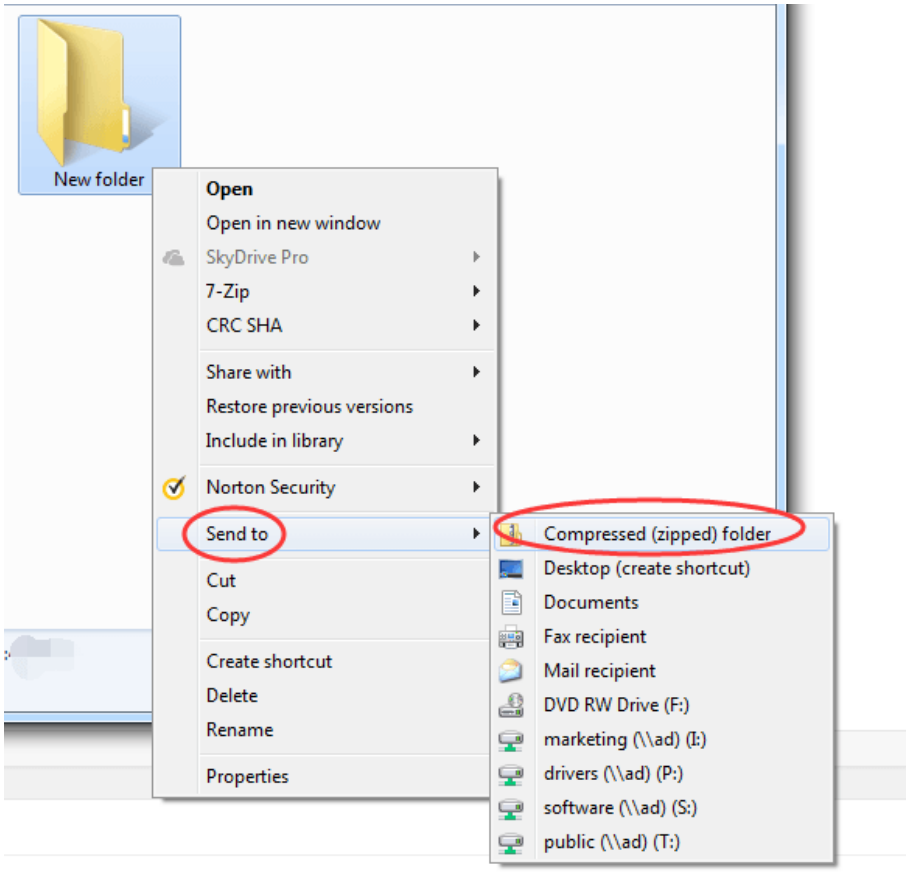
Description

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:



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

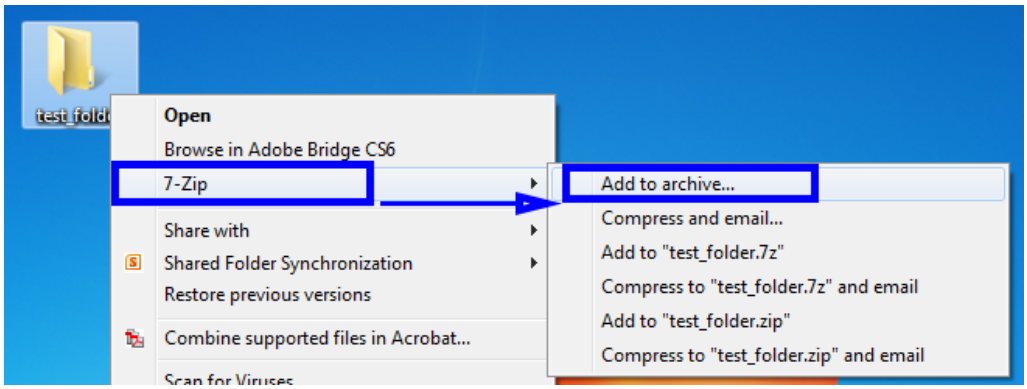
1. Right-click on the New folder you created with the file structure discussed above.
2. Then click Send to > Compressed (zipped) folder



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...



3. Bash script (MacOS/Linux)

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

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
```

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
```

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/
```

Calculation time

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: Quantile, filters, cores = 4)	04:01 s
	server run - minfi	
	(params: Quantile, 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

By Jyotirmoy Das