EpigenCentral User Guide

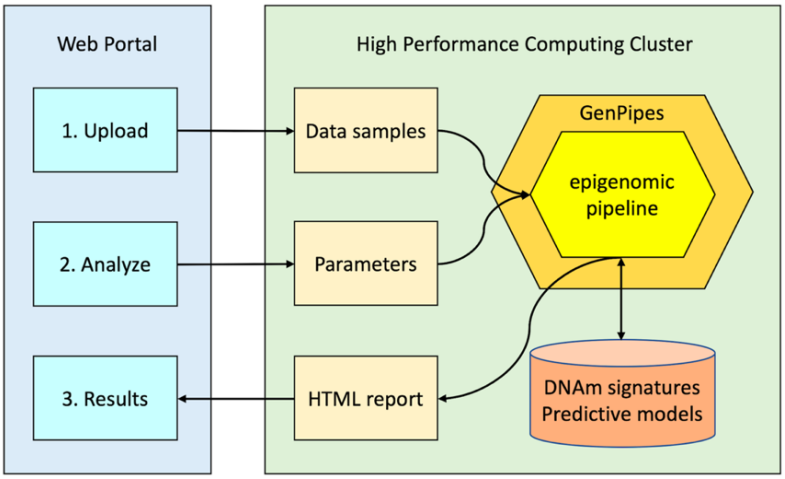
June 08, 2020

1 Introduction

EpigenCentral is a web resource for the interactive analysis of epigenomic datasets. It enables the classification of DNA methylation samples related to rare diseases and neurodevelopmental disorders (NDDs) and the discovery of new epigenetic patterns of disease. EpigenCentral consists of three interrelated components: (i) a web portal through which users can upload their own data for analysis and visualization; (ii) a set of computational pipelines that enable pre-processing, analysis and classification of the user's data samples; and (iii) a collection of known DNA methylation patterns and predictive models associated with various NDDs, which are used by the pipelines. The epigenetic patterns identified by our team (Butcher et al. 2017; Chater-Diehl et al. 2019; Choufani et al. 2015; Choufani et al. 2020; Siu et al. 2019) as well as those from additional datasets and studies (Bacalini et al. 2015; Strong et al. 2015) have been used to build the growing collection of predictive models currently available in EpigenCentral for classification tasks.

By submitting a dataset of DNA methylation (DNAm) samples to the EpigenCentral portal, the user should be able to assess the likelihood of each sample belonging to one of the NDD types, based on the presence of known molecular patterns and biomarkers in their DNAm profile. The generated disease scores help quantify the pathogenicity of genetic sequence variants. Exploratory data analysis is also available to help the user find new patterns in the data: e.g. if the submitted dataset contains multiple sample groups, such as disease cases vs. controls, EpigenCentral enables the detection of methylation differences between the groups.

From a user's perspective the workflow includes three main stages: upload the DNAm dataset and sample sheet, select analysis tasks and parameters, and review the results. These stages are reflected in the three menu items at the top of the EpigenCentral page: *Upload*, *Analyze*, *Results*.



2 Quick Start

A sample dataset has already been pre-uploaded into EpigenCentral and can be accessed through a *guest* account. A new user logged in as *guest* may proceed to the Analysis page to customize and submit a new disease classification or exploratory analysis run, then monitor the Results page where the analysis report should become available.

For a first quick exploration of the portal please follow these steps:

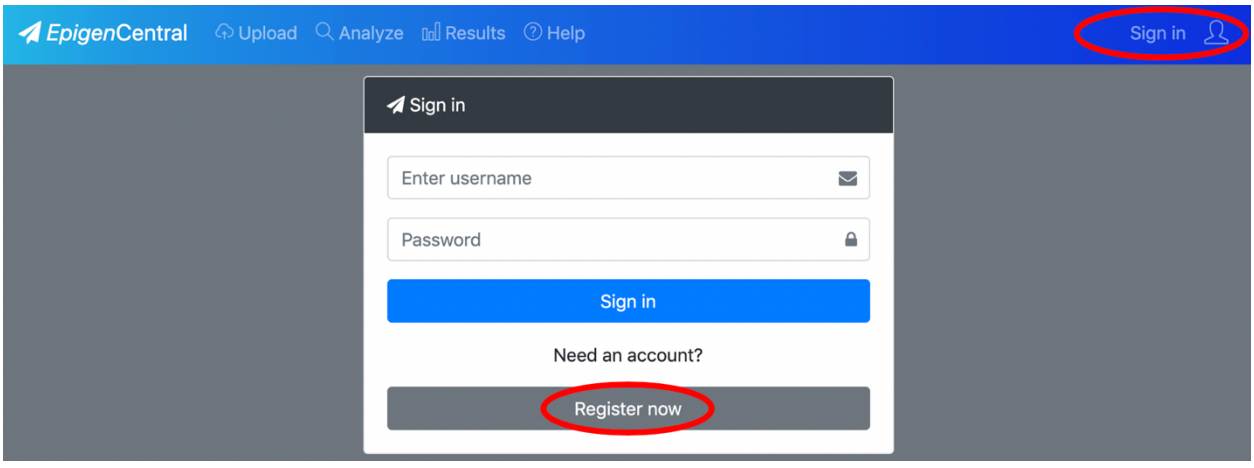
1. Login: use ***guest*** as both the username and the password.
2. Go to the Analyze page. In the dropdown list Dataset at the top of the page, check that the selected dataset is "Kabuki".
3. In the *Disease classification* tab, choose the option "Kabuki syndrome: KMT2D gene".
4. Click on the button Create Run. This should submit the analysis and automatically take you to the Results page.
5. Monitor the progress status until the analysis report becomes available.

Please note that others may also log in as guest and examine or delete the data or analysis results within the guest account. Therefore, we recommend using the guest account only for demonstration purposes with datasets that are not sensitive.

A more extensive tutorial is available on the portal's Help page, which is accessible through the top menu. The tutorial contains videos that demonstrate how to upload different types of input data, how to select analysis options and submit a new run.

3 Login

To analyze your own datasets, please create an account with your username (or email) and password: first click on the *Sign In* link in the top-right corner of the EpigenCentral web page, then click on *Register now* link in the popup window.



4 Data preparation

EpigenCentral currently supports data generated using the Illumina Human methylation microarray platforms such as HumanMethylation450 and HumanMethylationEPIC (also known as 450k and EPIC arrays, respectively). The data may be represented as either the original IDAT files of color intensities along with their metadata, or as a pre-processed plain-text table of DNAm β-values. The data bundle prepared for analysis may include the following components.

- A sample sheet file
- A design file of contrasts for the detection of DNAm differences
- Folders containing IDAT file pairs, one folder per array chip
- Tab-delimited table of DNAm β-values

4.1 Sample sheet files

A sample sheet file is required for a dataset consisting of Illumina IDAT files; it is optional (but recommended) for a dataset submitted as a pre-processed table of DNAm β-values. The sample sheet file is a comma-separated plain text file in which rows represent data samples and columns represent various sample attributes. It can be prepared using a text editor, Microsoft Excel, or a custom-made script. The sample sheet follows the Illumina's Infinium HD Methylation Sample Sheet format, which is also supported by the *minfi* Bioconductor package.

The following columns or their equivalents should be present in the sample sheet:

- *Sample_Name* : user-specified names or IDs of all samples in the dataset, such as "Case12" or "ctrl34". **Important**: Each sample name must contain only letters A-Z, a-z, numbers 0-9, hyphens (-) or underscores (_). I.e. characters like "/", "\", "|", "#", parentheses or spaces should not appear in the sample names. If the column *Sample_Name* is missing, the sample sheet must contain another equivalent column that holds sample names.
- *Sample_Group* : user-specified name of the sample group, such as "Kabuki syndrome" or "Control". The groups are used to identify patterns of differential methylation. If the column *Sample_Group* is missing, the sample sheet must contain another equivalent column that holds sample-group name(s).
- *Sentrix_ID* : the unique identifier of an Illumina BeadChip array, such as *8655685138*. This column is mandatory.
- *Sentrix_Position* : the unique position of the sample on the BeadChip indicating the row and column, such as R05C02. This column is mandatory.

The sample sheet may contain other columns describing various sample characteristics and confounding factors, such as sex, age, tissue of origin, mutation status or batch information. The following example shows a fragment of a sample sheet for a dataset on Kabuki syndrome (KS) with information on three controls and 3 KS patient samples (this information is derived from the GEO dataset GSE116300 generated by (Sobreira et al. 2017)).

```
Sample_Name,GEO_accession,Sample_Group,Sex,Sentrix_ID,Sentrix_Position
Con1,GSM3227755,control,M,8655685138,R05C02
Con2,GSM3227756,control,M,8655685138,R06C02
Con3,GSM3227757,control,F,8655685063,R05C02
Pat1,GSM3227764,KS,F,8655685138,R01C01
Pat2,GSM3227775,KS,F,8655685138,R02C01
Pat3,GSM3227786,KS,M,8655685138,R03C01
```

Please note: There should be no more than one CSV file within any dataset prepared for submission. Any comma-separated text file with a file extension ***.csv*** is assumed to be the unique sample sheet for the dataset.

4.2 Design file

To enable the comparison between groups of DNAm samples, EpigenCentral requires a so-called design file, which follows the GenPipes approach to pipeline development. For details see http://www.computationalgenomics.ca/tutorials/.

The design file is a tab-separated plain text file with a file extension ***.design*** and the following two columns:

- *Sample* : the first column, which should match the corresponding column *Sample_Name* (or its equivalent) in the sample sheet. Each sample name must contain only letters A-Z, a-z, numbers 0-9, hyphens (-) or underscores (_). This column is mandatory.

- Column of contrast : the second column defines an experimental design contrast. The column name defines the contrast name, e.g. “KS” to indicate the Kabuki syndrome. The following values represent the sample group membership for this contrast:

  “**2**”: the sample is in the disease, mutation or treatment group of the discovery cohort
  “**1**”: the sample belongs to the control group.
  “**0**” or “” (empty): the sample does not belong to any group and will not be used in comparisons. This option should be used e.g. for genetic-variant samples the validation cohort, which would be examined for the presence of the disease pattern, once the latter is found using the discovery-cohort cases (“2”) and matched controls (“1”).

There may be one or more contrast columns specified in the design file. The following example shows a fragment of a design file for Kabuki syndrome (KS) dataset in which two KS samples are compared to two controls, whereas one KS case and one control sample are excluded:

```
Sample KS
Con1   1
Con2   1
Con3   0
Pat1   2
Pat2   2
Pat3   0
```

4.3 Illumina IDAT files

A typical Illumina microarray dataset consists of pairs of IDAT files, one pair per data sample, each pair representing the red and green channel intensities. The file names follow the Illumina naming format, e.g. a sample on an array chip 8655685138 in the position R05C02 will correspond to the two files *8655685138_R05C02_Red.idat* and *8655685138_R05C02_Grn.idat*. Samples should be organized into folders (directories) whose name match the BeadChip IDs of the corresponding microarrays. For example, the following two file folders and 12 files (or 6 file pairs) are required for a sample sheet of three patient and three KS samples as shown above.

```
8655685063/
       8655685063_R05C02_Grn.idat
       8655685063_R05C02_Red.idat
8655685138/
       8655685138_R01C01_Grn.idat
       8655685138_R01C01_Red.idat
       8655685138_R02C01_Grn.idat
       8655685138_R02C01_Red.idat
       8655685138_R03C01_Grn.idat
       8655685138_R03C01_Red.idat
       8655685138_R05C02_Grn.idat
       8655685138_R05C02_Red.idat
       8655685138_R06C02_Grn.idat
       8655685138_R06C02_Red.idat
```

Please note: every sample listed in the sample sheet should have its pair of IDAT files present in the dataset. The file names should match the sample-sheet columns *Sentrix_ID* and *Sentrix_Position* thus uniquely identifying each sample. Additional IDAT files not described in the sample sheet may also be present in the uploaded dataset but will be ignored.

Please note: individual IDAT files may be submitted in their gzipped form (file extension .gz). However, the directory structure should still match the chip IDs as shown above, e.g.:

```
8655685063/
        8655685063_R05C02_Grn.idat.gz
        8655685063_R05C02_Red.idat.gz
```

4.4 Tab-delimited table of DNAm values

EpigenCentral allows the upload of a pre-processed table of DNAm values in the form of a tab-delimited plain text file, instead of the original collection of IDAT files. In this case the data file should satisfy the following requirements:

- The file should have the extension ***.tsv*** (i.e. tab separated values)
- Rows starting with the exclamation mark '!' are ignored. This facilitates the uploading of NCBI GEO series-matrix files in which metadata lines start with '!'
- Table rows correspond to Illumina array probes.
- The first column should contain the Illumina array probe IDs corresponding to CpG sites. The column name does not matter.
- All other table columns correspond to data samples.
- The first row of the table should be a header row containing sample names.
- The values in the data table are DNAm β-values, i.e. values between 0 and 1 representing the percentage of methylated cytosines for the corresponding CpG site and data sample.

Please note: some of the analysis options are unavailable for data submitted as TSV files. E.g. a single TSV table without any sample-sheet or design files cannot be analyzed for differentially methylated patterns. However, a TSV data file may be accompanied by a sample sheet file and/or a design file, which extends the range of available analysis options.

The following example shows a fragment of a tab-separated file representing a Down's syndrome dataset GSE52588 from the GEO repository, extracted directly from the corresponding GEO series matrix. (Quotation marks are optional.)

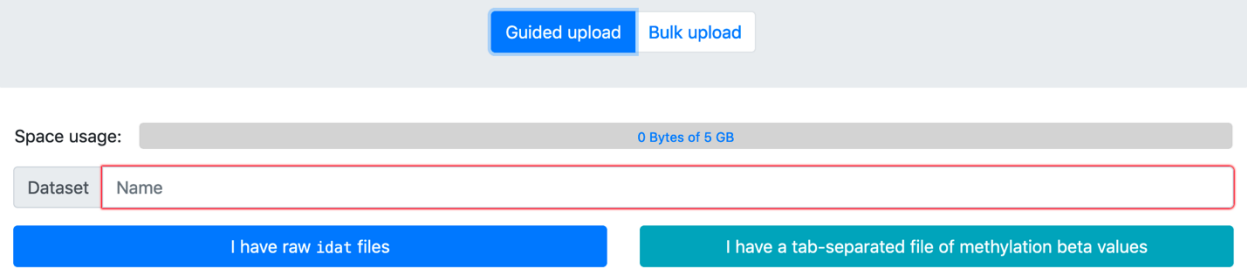
```
"ID_REF"      "GSM1272122"  "GSM1272123"  "GSM1272124"
"cg00000029"  0.5744937     0.6312186     0.6389823
"cg00000108"  0.9205178     0.9374648     0.9379248
"cg00000109"  0.889981      0.8370166     0.8247315
"cg00000165"  0.144547      0.1385563     0.1685321
"cg00000236"  0.7515112     0.694276      0.6910655
"cg00000289"  0.659462      0.7139714     0.6510754
```

5 Upload page

The Upload page requires the user to first assign a new dataset name. The dataset name must contain only letters A-Z, a-z, numbers 0-9, hyphens (-) or underscores (_). Afterwards the user may proceed to drag & drop the components of the prepared data bundle as appropriate. There are two main types of upload: Guided and Bulk. We recommend Guided upload for new users.

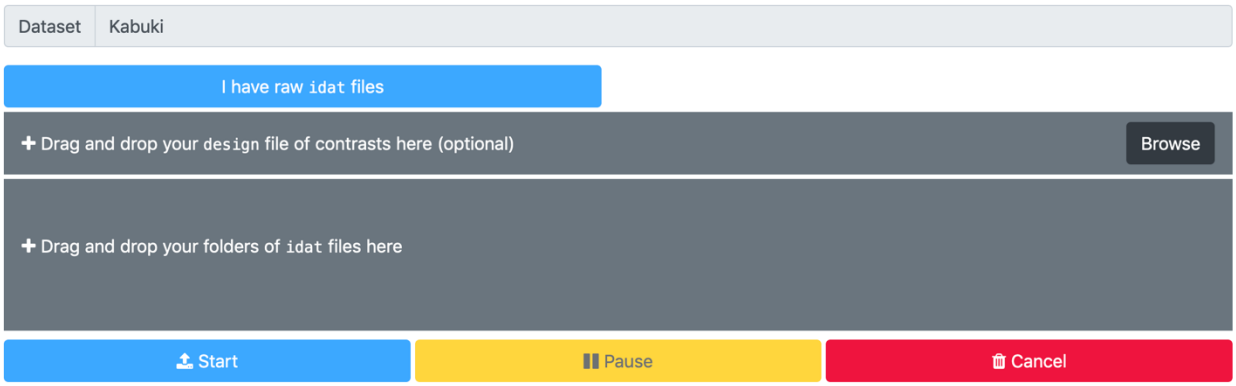
5.1 Guided upload: Illumina IDAT files

Click on *Guided upload*, type in the dataset name, and click the button "*I have raw idat files*".

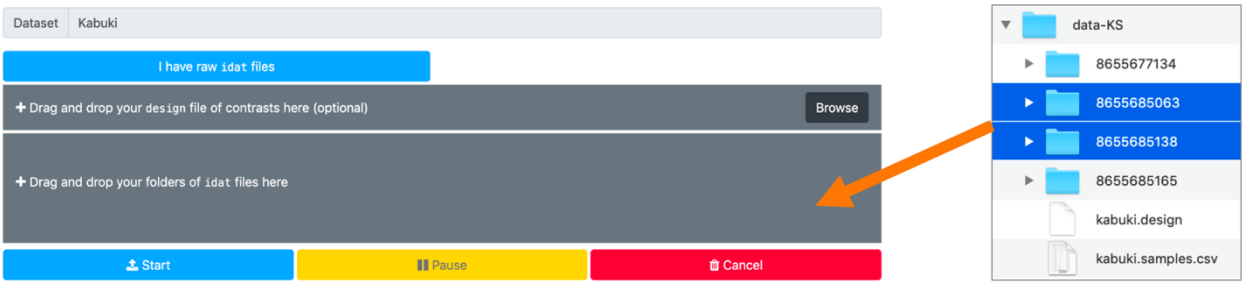


A grey rectangular area for selecting the sample-sheet file appears, where the file can be dragged & dropped or selected using the *Browse* button.

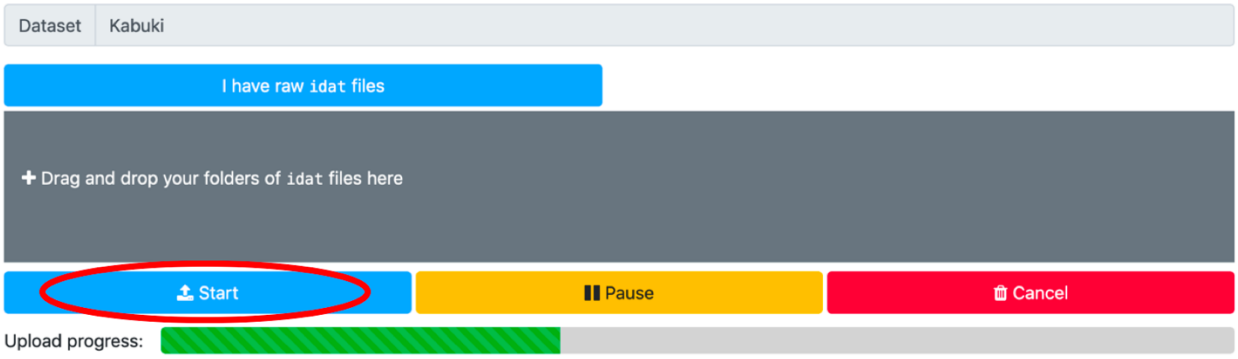
After the sample sheet is selected, two more grey rectangular areas appear. One is for the (optional) design file, which can be either dragged & dropped or selected using the *Browse* button.



The other grey area is for dragging and dropping the whole folders containing IDAT files, where each folder corresponds to the Illumina array chip and should match that chip's *Sentrix_ID* as specified in the sample sheet. Select the entire chip folder (not just the IDAT files therein) using your file-system graphical interface, and drag & drop it into the grey rectangle; then proceed to drag & drop the next folder. Or select and drag & drop several chip folders at once.

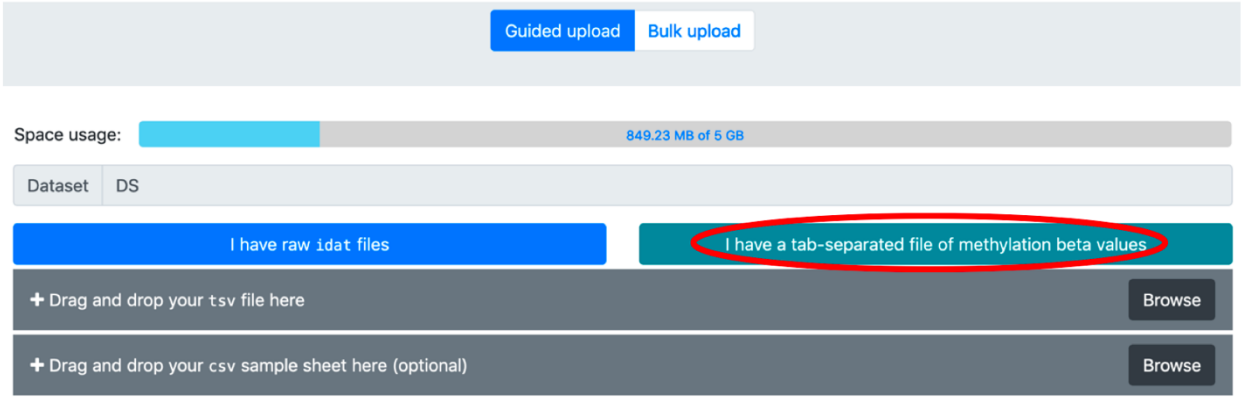


Once all the IDAT folders are selected, click the Start button, which initiates the data upload to the EpigenCentral server for processing, as indicated with the green progress bar.



5.2 Guided upload: Tab-delimited table of DNAm values

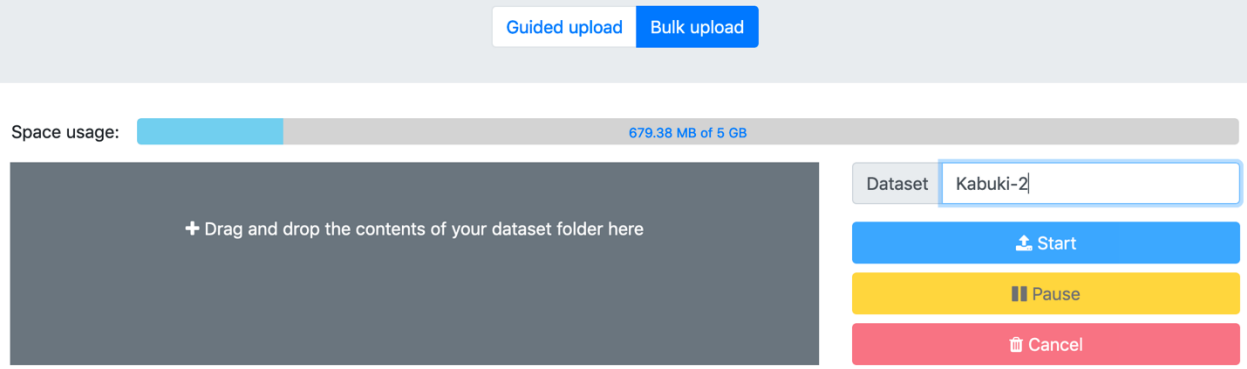
For TSV files containing DNAm value tables, click on *Guided upload*, then on the button "*I have a tab-separated file of methylation beta values*". Grey drag & drop areas will appear: first for the main TSV data file, then for the (optional) sample sheet and (optional) design file. After selecting all the files click on the Start button to initiate the upload.



5.3 Bulk upload: full data bundle

Bulk upload allows the drag & drop of all data files, folders and metadata at once into a single grey rectangular area, which could be a faster alternative for an experienced user.

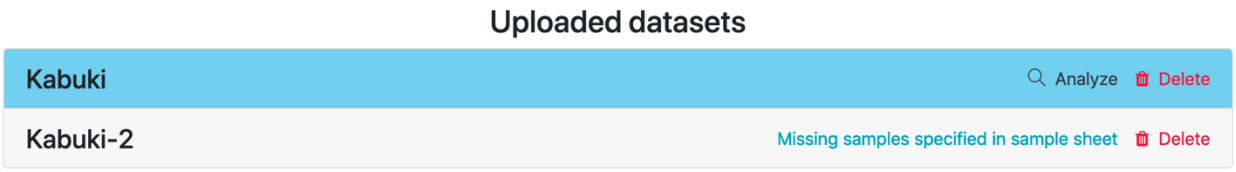
Click on the button *Bulk upload*, then type in the dataset name. Once the grey rectangle appears, select and drag & drop all your files and folders there, then click Start to initiate the upload. EpigenCentral will assign the role to each file based on the file extension: *.csv* for the sample sheet, *.design* for a design file and *.tsv* for a single tab-delimited data table.



5.4 Additional uploads

The list of uploaded datasets appears at the bottom of the Upload page. Clicking on each row expands it to show the contents of the uploaded data. Datasets that are ready for analysis are shown with the "Analyze" link, which if clicked takes the user to the Analyze page.

EpigenCentral scans the uploaded dataset to ensure minimal compliance between the data and its metadata. E.g. if the dataset has fewer IDAT files than described in its sample sheet, a message will be shown and the dataset will not be available for analysis until the issue is resolved.

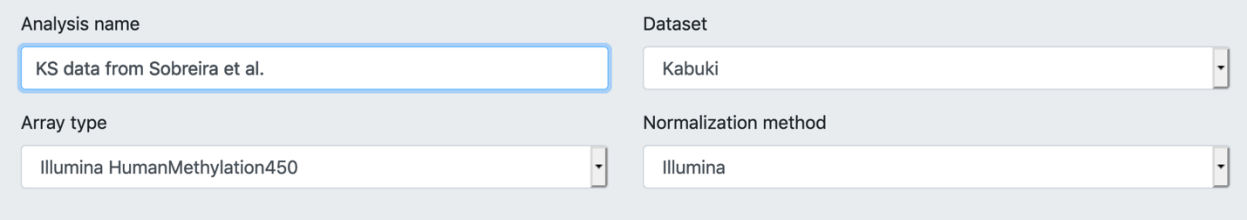


Additional files may be uploaded by using the same dataset name in subsequent uploads, e.g. to add missing files or to replace an old file with an updated version. If a file with the same name already exists in the dataset, it will be replaced with the new version. Existing files may be deleted using the dataset table at the bottom of the Upload page.

Please note that the user is responsible for maintaining data integrity e.g. by allowing no more than one CSV file per dataset, which is assumed to be the sample sheet file in the proper format.

6 Analysis page

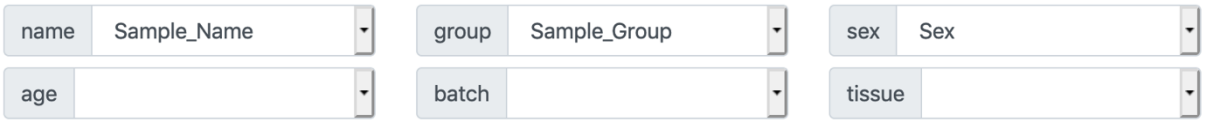
The Analyze page requires the user to first select the dataset for analysis from the dropdown list. The user has an option to provide a descriptive name for the analysis. Thereafter there are two main options for the analysis, represented by two tabs: the *Disease classification* tab allows users to compare their DNAm data to known disease profiles; and *Exploratory analysis* tab enables the search for new differential methylation patterns in the data (as long as a sample sheet and design files were provided). DNA methylation analysis may be applied to DNAm profiles generated on any human-derived tissues, disease settings or environmental exposures as long as they have proper tissue-matching controls.



Array type: Currently 450k and EPIC arrays are supported.

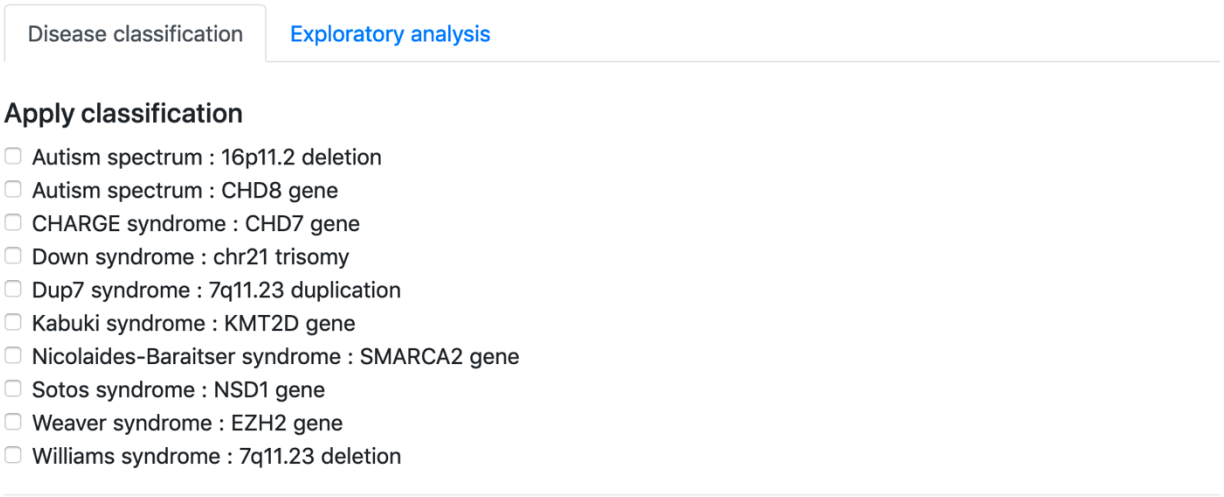
Normalization method: Options available in the *minfi* package are implemented, such as Raw (no processing), Illumina, SWAN, Quantile, Noob and Funnorm. See *minfi* vignette https://bioconductor.org/packages/release/bioc/vignettes/minfi/inst/doc/minfi.html for details.

The subsection on *Sample sheet CSV column labels* allows the user to map different columns from the sample sheet to predefined roles. By default, EpigenCentral scans the sample sheet for common column names such as *Sample_Name, Sample_Group* or *Tissue*. However, the user may also map non-standard names e.g. “Gender” to indicate sex or “GEO Accession” for sample names.



6.1 Disease classification tab

Users can scan their DNAm datasets for the presence of disease-associated patterns supported by EpigenCentral. This requires only minimal pre-processing of the data, followed by the application of pre-built classification models to the data. Such scenario does not require a sample sheet or a design file. The DNAm patterns identified by our team (Butcher et al. 2017; Chater-Diehl et al. 2019; Choufani et al. 2015; Choufani et al. 2020; Siu et al. 2019) as well as those from additional datasets and studies (Bacalini et al. 2015; Strong et al. 2015) have been used to build the growing collection of machine-learning models currently available in EpigenCentral for classification tasks. All current predictive models are based on blood DNAm.



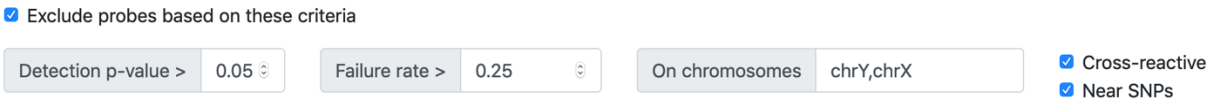
6.2 Exploratory analysis tab

In a more advanced analysis scenario users may explore patterns of differential methylation between sample groups in the data. This functionality is enabled by the R/Bioconductor packages *minfi, bumphunter, limma* and *LOLA*. Prior to the exploratory analysis the user is asked to apply filters to the array probes and also, for blood samples, to estimate their cell subtype compositions.

Probe filtering is based on several quality criteria:

- *Detection p-value*: DNAm values with their detection p-values above the threshold are treated as missing values
- *Failure rate*: Removing the CpG sites with the proportion of missing values (including values with poor detection p-value) above the selected rate

- *On chromosomes*: Removing the CpGs sites on the listed chromosomes. This option is most useful for removing sex chromosomes chrX and chrY from analysis. Chromosomes should have the prefix 'chr'. Multiple chromosomes should be separated by commas.
- *Cross-reactive*: Removing the CpGs sites whose Illumina array probes are known to hybridize to genomic regions other than the targeted CpG. The probes were identified in (Chen et al. 2013) for the 450k arrays and in (McCartney et al. 2016) for EPIC arrays.
- *Near SNPs*: Removing the CpG sites near known SNP mutation sites, as identified by the function *getSnpInfo* of the *minfi* package.



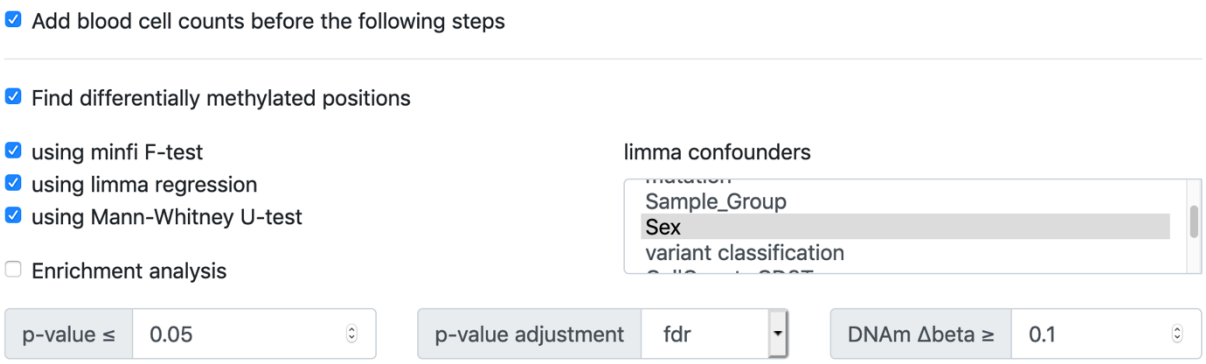
If the DNAm samples were collected from blood and the original IDAT files are available in the dataset, the EpigenCentral can estimate the proportions of six different blood cell subtypes for each sample. This feature uses the *minfi* function *estimateCellCounts* for the 450k arrays and *estimateCellCounts2* for the EPIC arrays. The six supported cell types are CD8+ T cells, CD4+ T cells, CD56+ NK cells, CD19+ B cells, CD14+ monocytes, and granulocytes (the latter subtype typically has the largest range).

Differentially methylated positions may be detected using three available methods: F-test for residuals implemented in the *minfi* function *dmpFinder*, regression analysis implemented in *limma* Bioconductor package, and the non-parametric Mann-Whitney U test. Additional confounders for limma regression analysis may be selected from the dropdown list of sample-sheet columns and (if available) the six estimated blood cell type proportions. We recommend using *Sex*, *Age* and (for blood-derived samples) *CellCounts.Gran* as the initial confounders.

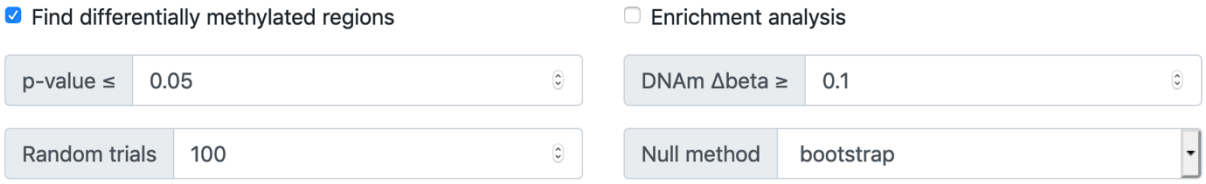
- *p-value*: significance level threshold for differentially methylated CpGs.
- *p-value adjustment*: the method of adjustment for multiple testing, with the options "fdr" (Benjamini-Hochberg method), "bonf" (Bonferroni method) or "none" (no adjustment).
- *DNAm Δbeta*: threshold for the difference in average DNAm levels between the two sample groups to be considered biologically significant.

By default, a CpG is considered to be differentially methylated if its $p<0.05$ after FDR adjustment and the DNAm change $\Delta\beta$ is at least 0.10 (i.e. 10 percentage points). These parameters may be customized as needed.

Enrichment analysis identifies significant overlaps between the differentially methylated sites on the one hand, and known histone marks and transcription factor binding sites on the other hand. The latter are extracted from external resources such as ENCODE, CEEHRC and DeepBlue. This feature is enabled by the *LOLA* Bioconductor package.



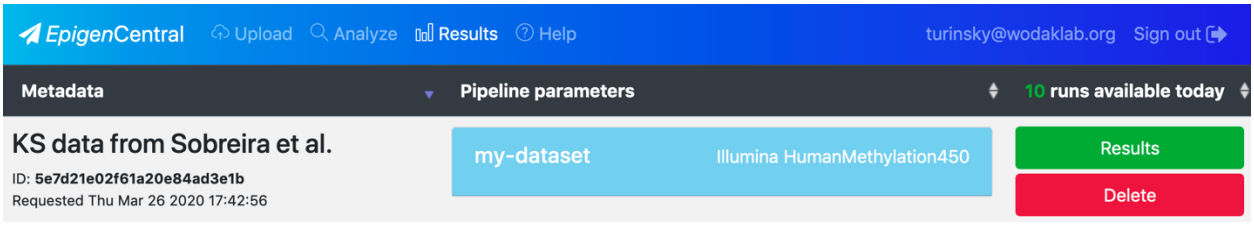
Differentially methylated regions are identified using the function *bumphunter* from the *bumphunter* Bioconductor package.



Once the analysis parameters are selected, the user can click the *Create run* button to submit the job for processing and move to the Results page; or click the *Submit another* button to submit the job and remain on the Analysis page.

7 Results page

The Results page presents a table of all analysis runs and their current status. A run that has just been submitted is shown as Pending. Once the data processing begins, a progress bar is shown. After the analysis is complete, the page displays a link to the analysis report and/or to any error messages encountered during the job processing. The automatically generated EpigenCentral analysis report is self-explanatory and presents summaries of various analysis steps as well as links to further files, tables and images.



8 References

Bacalini, M. G., et al. (2015), 'Identification of a DNA methylation signature in blood cells from persons with Down Syndrome', *Aging (Albany NY),* 7 (2), 82-96.

Butcher, D. T., et al. (2017), 'CHARGE and Kabuki Syndromes: Gene-Specific DNA Methylation Signatures Identify Epigenetic Mechanisms Linking These Clinically Overlapping Conditions', *Am J Hum Genet,* 100 (5), 773-88.

Chater-Diehl, E., et al. (2019), 'New insights into DNA methylation signatures: SMARCA2 variants in Nicolaides-Baraitser syndrome', *BMC Med Genomics,* 12 (1), 105.

Chen, Y. A., et al. (2013), 'Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray', *Epigenetics,* 8 (2), 203-9.

Choufani, S., et al. (2015), 'NSD1 mutations generate a genome-wide DNA methylation signature', *Nat Commun,* 6, 10207.

Choufani, S., et al. (2020), 'DNA Methylation Signature for EZH2 Functionally Classifies Sequence Variants in Three PRC2 Complex Genes', *Am J Hum Genet*.

McCartney, D. L., et al. (2016), 'Identification of polymorphic and off-target probe binding sites on the Illumina Infinium MethylationEPIC BeadChip', *Genom Data,* 9, 22-4.

Siu, M. T., et al. (2019), 'Functional DNA methylation signatures for autism spectrum disorder genomic risk loci: 16p11.2 deletions and CHD8 variants', *Clin Epigenetics,* 11 (1), 103.

Sobreira, N., et al. (2017), 'Patients with a Kabuki syndrome phenotype demonstrate DNA methylation abnormalities', *Eur J Hum Genet,* 25 (12), 1335-44.

Strong, E., et al. (2015), 'Symmetrical Dose-Dependent DNA-Methylation Profiles in Children with Deletion or Duplication of 7q11.23', *Am J Hum Genet,* 97 (2), 216-27.