epimutacions: a Bioconductor package to identify outliers in rare diseases DNA methylation data

Supplementary material

Leire Abarrategui
Carlos Ruiz-Arenas
Carles Hernandez-Ferrer
Patricia Ryser-Welch
Juan R. Gonzalez

2021-09-27

Abstract

Epimutations are rare alterations in the methylation pattern at specific loci. Have been demonstrated that epimutations could be the causative factor of some genetic diseases. Nonetheless, no standard methods are available to detect and quantify these alterations. We have developed `epimutacions` package. The package implements 6 approaches to identify epimutations in genome-wide DNA methylation microarrays data. The document describes, package installation; data loading and preprocessing; epimutation identification, annotation and visualization; and GEO datasets simulations results.

Contents

1 Introduction **3**

1.1 Background . . . 3

1.2 Methodology . . . 3

2 Setup **4**

2.1 Installing the package . . . 4

2.2 Loading libraries . . . 4

2.3 Quick start . . . 4

3 Datasets **5**

3.1 IDAT files and reference panel . . . 6

3.2 Methy . . . 6

3.3 Candidate regions . . . 6

4 Preprocessing **7**

5 Epimutations **10**

5.1 Epimutations detection . . . 10

5.2 Unique parameters . . . 11

5.3 Results description . . . 12

5.4 Epimutations annotations . . . 14

5.5 Epimutation visualization . . . 15

6 Results **17**

6.1 Simulations . . . 17

6.2 Methods testing . . . 32

7 Acknowledgements **33**

8 Session Info **34**

References **36**

1 Introduction

1.1 Background

It is estimated that approximately 30 million people, in the EU, are affected by rare diseases. The European Commission has defined rare diseases as pathologies with a prevalence of less than 1 person in 2,000 people. It is estimated that there are between 6,000 and 8,000 different rare diseases, 80% of them with a genetic origin. These conditions can have an onset during childhood and affect 6-8% of the population during their lifetime (European-Commission 2020).

Despite the successful contributions of sequence-based approaches (e.g. exome and genome sequencing), about 60% of the patients remain undiagnosed (Lionel et al. 2018). It has been noted that even when the sequencing was appropriate some cases could not be diagnoses due to the origin may not be genetic. Previous studies have demonstrated that epimutations could be the causative factor of some genetic diseases (Aref-Eshghi et al. 2019; Garg et al. 2020; Barbosa et al. 2018). Epimutations can lead to cancers, such as Lynch syndrome, rare diseases such as Prader-Willi syndrome, and are associated with common disorders, such as autism. Consequently, we are assuming epimutations could help to solve the undiagnosis clinical cases using sequence based methods. To the best of our knowledge, no established method is available yet to detect and quantify suitably epimutations. However, some detection methods relying on methylation microarray data and two phases of computations have been reported. The first method identifies CpGs using outlier values and then clusterisation (Barbosa et al. 2018). The second one identifies candidate regions with Bumphunter (Aref-Eshghi et al. 2019) and then MANOVA to test statistical significance (Aref-Eshghi et al. 2019). However, the implementation of these methods has yet to be made publicly available. In addition to those two methods (i.e., `quantile` and `manova`), We implemented a beta distribution to detect CpG outliers (`beta`), or a different model to assess region significance (`mlm`, `mahdistmcd` and `isoforest`).

The package `epimutacions` provides tools to raw DNA methylation microarray intensities normalization and epimutations identification, visualization and annotation. The full `epimutacions` user´s guide is available in this document which describes: package installation; data loading and preprocessing; epimutation identification, annotation and visualization; and GEO datasets simulations results.

The name of the package is `epimutacions` (pronounced `pi mu ta 'sj ons`) which means epimutations in Catalan, a language from the northeast of Spain.

1.2 Methodology

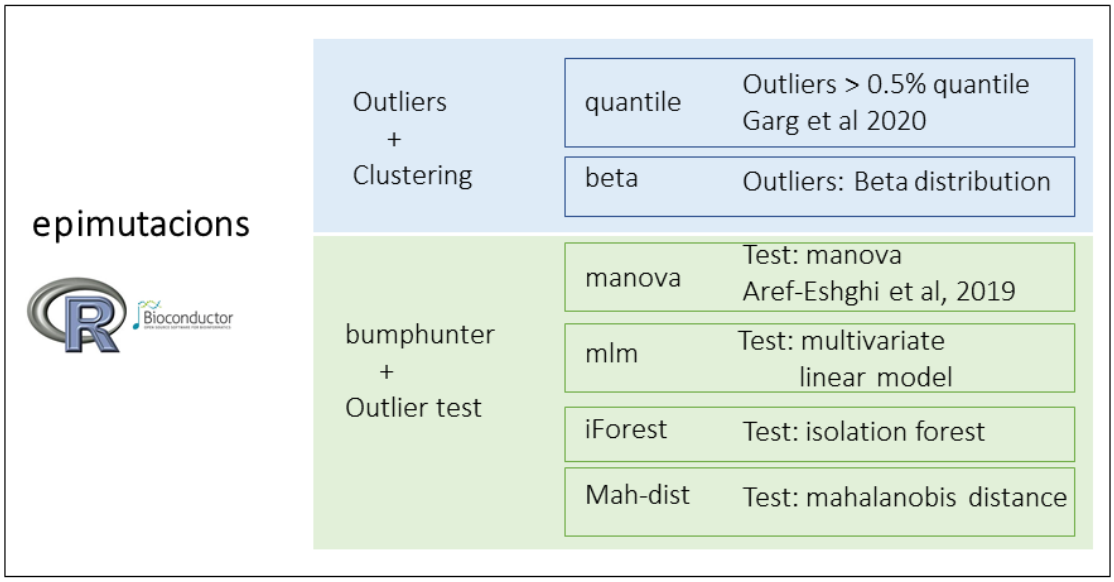
The `epimutacions` package computes a genome-wide DNA methylation analysis to identify the epimutations to be considered as biomarkers for case samples with a suspected genetic disease. The package includes 2 functions to infer epimutations: (1) `epimutations()` which uses a case-control design. It compares a case sample against a reference panel (healthy individuals); and (2) `epimutations_one_leave_out()` which compares individual methylation profiles of a single sample (regardless if they are cases or controls) against all other samples from the same cohort.

The `epimutacions` package integrates 6 outlier detection methods; an argument referred to as `method` allows choosing between `manova`, `mlm`, `isoforest` and `mahdistmcd` to define the differentially methylated regions (DMRs) using bumphunter. Then, those DMRs are tested to identify regions with CpGs being outliers when comparing with the reference panel. Two other values, `quantile` and `beta` defines the outlier CpGs and then group them into epimutations. The computations filter these intermediate results alongside candidate regions to be epimutations (section 3.3) dataset calculated with bumphunter; all the regions are selected with a minimum of 3 CpGs and a maximum distance of 1kb between them (Barbosa et al. 2018). The dataset is available in `epimutacionsData` package.

- `quantile`: defines outliers as values exceeding an extreme quantile Value of the reference population (by default $> 0.5\%$). Implementation based on (Garg et al. 2020).

- `beta`: fits values from reference population to a beta distribution. Values with low probability of belonging to the distribution (<1e-6 by default) are considered outliers.
- `manova`: applies multivariate analysis of variance to the regions identified by bumphunter. It selects the outlier regions using p-value significance (by default < 0.05).
- `mlm`: applies Multivariate Linear Model (Martín 2020) to regions identified with bumphunter. It selects the outlier regions using p-value significance (by default < 0.05).
- `isoforest`: defines the outlier regions based on isolation forest (Cortes and Cortes 2021). It uses outlier score to identify the outliers (by default > 0.7).
- `mahdistmcd`: computes robust mahalanobis distance (using Minimum Covariance Determinant) to define epimutations (Maechler et al. 2021). It uses chi-square distribution to identify outlier regions.

Figure S1: Implementation of each outlier detection method



2 Setup

2.1 Installing the package

```
devtools::install_github("isglobal-brge/epimutacions")
```

2.2 Loading libraries

```
library(epimutacions)
```

2.3 Quick start

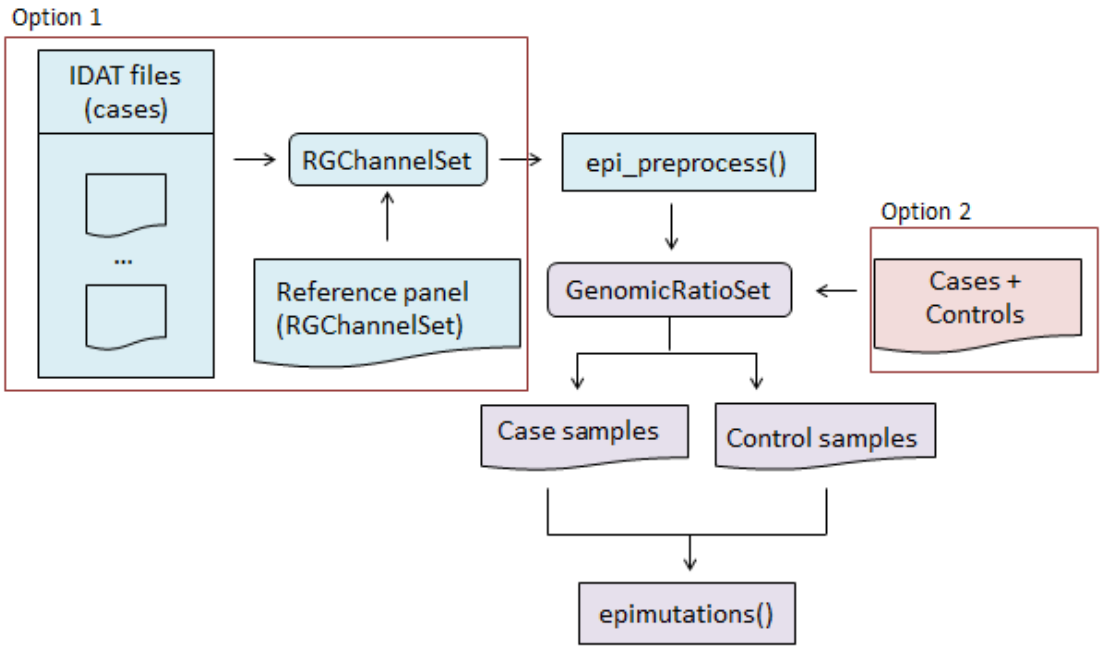
The workflow in figure S2 explains the main analysis in the `epimutacions` package.

The package allows two different types of inputs:

- (1) Case samples `IDAT` files (raw microarray intensities) together with `RGChannelSet` object as reference panel. The reference panel can be supplied by the user or can be selected through the example datasets that the package provides (section 3).
- (2) `GenomicRatioSet` object containing case and control samples.

The normalization (`epi_preprocess()`) converts the raw microarray intensities into usable methylation measurement (β values at CpG locus level). As a result, we obtain a `GenomicRatioSet` object, which can be used as `epimutations()` function input. The data should contain information about values of CpG sites, phenotype and feature data.

Figure S2: Allowed data formats, normalization and input types



3 Datasets

The package contains 3 example datasets adapted from Gene Expression Omnibus (GEO):

- (1) 4 case samples IDAT files (GEO: GSE131350)
- (2) `reference_panel`: a `RGChannelSet` class object containing 22 healthy individuals (GEO: GSE127824)
- (3) `methy`: a `GenomicRatioSet` object which includes 49 controls (GEO: GSE104812) and 3 cases (GEO: GSE97362).

We also included a dataset specifying the 40,408 candidate regions in Illumina 450K array which could be epimutations (see section 3.3).

We created the `epimutacionsData` package in `ExperimentHub`. It contains the reference panel, methy and the candidate epimutations datasets. The package includes the IDAT files as external data. To access the datasets we need to install the packages by running the following commands:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("ExperimentHub")
```

```
devtools::install_github("LeireAbarrategui/epimutacionsData")
```

Then, we need to load the package and create an `ExperimentHub` object:

```
library(ExperimentHub)
eh <- ExperimentHub()
query(eh, c("epimutacionsData"))
```

```
ExperimentHub with 3 records
# snapshotDate(): 2021-09-13
# $dataprovider: GEO, Illumina 450k array
# $species: Homo sapiens
# $rdataclass: RGChannelSet, GenomicRatioSet, GRanges
# additional mcols(): taxonomyid, genome, description, coordinate_1_based, maintainer, rdatadateadded,
#   rdatapath, sourceurl, sourcetype
# retrieve records with, e.g., 'object[["EH6690"]]'

           title
  EH6690 | Control and case samples
  EH6691 | Reference panel
  EH6692 | Candidate epimutations
```

3.1 IDAT files and reference panel

IDAT files directory in `epimutacionsData` package:

```
baseDir <- system.file("extdata", package = "epimutacionsData")
```

The reference panel can be found in `EH6691` record of the `eh` object:

```
reference_panel <- eh[["EH6691"]]
```

3.2 Methy

The `methy` is stored in `EH6690` record:

```
methy <- eh[["EH6690"]]
```

3.3 Candidate regions

In Illumina 450K array, probes are unequally distributed along the genome, limiting the number of regions that can fulfil the requirements to be considered an epimutation. Thus, we have computed a dataset containing all the regions that could be candidates to become an epimutation.

We used the clustering approach from bumphunter to define the candidate epimutations. We defined a primary dataset with all the CpGs from the Illumina 450K array. Then, we run bumphunter and selected those regions with at least 3 CpGs and a maximum distance of 1kb between them. As a result, we found 40,408 candidate epimutations. The function `epimutation()` filters the identified epimutations using these candidate regions.

The following is the code used to identify the candidate epimutations in Illumina 450K array:

```
library(minfi)
# Regions 450K
library(IlluminaHumanMethylation450kanno.ilmn12.hg19)
data(Locations)

### Select CpGs (names starting by cg) in autosomic chromosomes
locs.450 <- subset(Locations,
                   grepl("^cg", rownames(Locations)) & chr %in% paste0("chr", 1:22))
locs.450GR <- makeGRangesFromDataFrame(locs.450, start.field = "pos",
                                       end.field = "pos", strand = "*")
locs.450GR <- sort(locs.450GR)
mat <- matrix(0, nrow = length(locs.450GR), ncol = 2,
              dimnames = list(names(locs.450GR), c("A", "B")))

## Set sample B to all 1
mat[, 2] <- 1

## Define model matrix
pheno <- data.frame(var = c(0, 1))
model <- model.matrix(~ var, pheno)

## Run bumphunter
bumps <- bumphunter(mat, design = model, pos = start(locs.450GR),
                    chr = as.character(seqnames(locs.450GR)),
                    cutoff = 0.05)$table
bumps.fil <- subset(bumps, L >= 3)
```

The candidate regions can be found in `EH6692` record of the `eh` object:

```
candRegsGR <- eh[["EH6692"]]
```

4 Preprocessing

The `epi_preprocess()` function allows calling the 6 preprocessing methods from `minfi` package (Aryee et al. 2014):

Table S1: Preprocessing methods description

Method	Function	Description
`raw`	`preprocessRaw`	Converts the Red/Green channel for an Illumina methylation array into methylation signal. This method does not normalize the data.

Method	Function	Description
illumina	preprocessIllumina	Implements preprocessing for Illumina methylation microarrays as used in Genome Studio.
swan	preprocessSWAN	Subset-quantile Within Array Normalisation (SWAN). It allows Infinium I and II type probes on a single array to be normalized together.
quantile	preprocessQuantile	Implements stratified quantile normalization preprocessing for Illumina methylation microarrays.
noob	preprocessNoob	Noob (normal-exponential out-of-band) is a background correction method with dye-bias normalization for Illumina Infinium methylation arrays.
funnorm	preprocessFunnorm	Functional normalization (FunNorm) is a between-array normalization method for the Illumina Infinium HumanMethylation450 platform.

Each normalization approach has some unique parameters (table S2) which can be modified through `norm_parameters()` function:

Table S2: Preprocessing methods unique parameters

Method	Parameters	Description
illumina	bg.correct	Performs background correction
	normalize	Performs controls normalization
	reference	The reference array for control normalization
quantile	fixOutliers	Low outlier Meth and Unmeth signals will be fixed
	removeBadSamples	Remove bad samples
	badSampleCutoff	The cutoff to label samples as 'bad'
	quantileNormalize	Performs quantile normalization
	stratified	Performs quantile normalization within region strata
	mergeManifest	Merged to the output the information in the associated manifest package
	sex	Sex of the samples
noob	offset	Offset for the normexp background correct
	dyeCorr	Performs dye normalization
	dyeMethod	Dye bias correction to be done
funnorm	nPCs	The number of principal components from the control probes
	sex	Sex of the samples
	bgCorr	Performs NOOB background correction before functional normalization
	dyeCorr	Performs dye normalization
	keepCN	Keeps copy number estimates

We can obtain the default settings for each method by invoking the function `norm_parameters()` with no arguments:

```
norm_parameters()
```

```
$illumina
$illumina$bg.correct
[1] TRUE

$illumina$normalize
```

```
[1] "controls" "no"

$illumina$reference
[1] 1


$quantile
$quantile$fixOutliers
[1] TRUE

$quantile$removeBadSamples
[1] FALSE

$quantile$badSampleCutoff
[1] 10.5

$quantile$quantileNormalize
[1] TRUE

$quantile$stratified
[1] TRUE

$quantile$mergeManifest
[1] FALSE

$quantile$sex
NULL


$noob
$noob$offset
[1] 15

$noob$dyeCorr
[1] TRUE

$noob$dyeMethod
[1] "single"    "reference"


$funnorm
$funnorm$nPCs
[1] 2

$funnorm$sex
NULL

$funnorm$bgCorr
[1] TRUE

$funnorm$dyeCorr
[1] TRUE

$funnorm$keepCN
```

```
[1] FALSE
```

However, we can modify the parameter(s) as the following example for `illumina` approach:

```
parameters <- norm_parameters(illumina = list("bg.correct" = FALSE))
parameters$illumina$bg.correct
```

```
[1] FALSE
```

We are going to preprocess the IDAT files and reference panel (3). We need to specify the IDAT files directory and the reference panel in `RGChannelSet` format. As a result, we will obtain a `GenomicRatioSet` object containing the control and case samples:

```
GRset <- epi_preprocess(baseDir, reference_panel, pattern = "SampleSheet.csv")
```

```
[1] "C:/Users/nla94/Documents/R/win-library/4.1/epimutacionsData/extdata/SampleSheet.csv"
```

```
GRset
```

```
class: GenomicRatioSet
dim: 452567 26
metadata(0):
assays(1): Beta
rownames(452567): cg13869341 cg14008030 ... cg08265308 cg14273923
rowData names(0):
colnames(26): GSM3639453_R03C01 GSM3639454_R02C01 ... GSM3770882_R05C01 GSM3770871_R04C01
colData names(17): X sample ... filenames ArrayTypes
Annotation
  array: IlluminaHumanMethylation450k
  annotation: ilmn12.hg19
Preprocessing
  Method: Raw (no normalization or bg correction)
  minfi version: 1.39.1
  Manifest version: 0.4.0
```

5 Epimutations

5.1 Epimutations detection

The function `epimutations()` includes 6 methods for epimutation identification: (1) Multivariate Analysis of variance (`manova`), (2) Multivariate Linear Model (`mlm`), (3) isolation forest (`isoforest`), (4) robust mahalanobis distance (`mahdistmcd`) (5) `quantile` and (6) `beta`.

To illustrate the following examples we are going to use the dataset `methy` (section 3.2). We need to separate the case and control samples:

```
case_samples <- methy[,methy$status == "case"]
control_samples <- methy[,methy$status == "control"]
```

We can specify the chromosome or region to analyze which helps to reduce the execution time:

```
epi_mvo <- epimutations(case_samples, control_samples, method = "manova")
```

```
epi_ml <- epimutations(case_samples, control_samples, method = "mlm")
epi_iso <- epimutations(case_samples, control_samples, method = "isoforest",
                        chr = "chr10", start = 100, end  = 10000)
epi_mcd <- epimutations(case_samples, control_samples, method = "mahdistmcd", chr = "chr12")
epi_qtl <- epimutations(case_samples, control_samples, method = "quantile", chr = "chr22")
epi_beta <- epimutations(case_samples, control_samples, method = "beta", chr = "chr17")
```

The function `epimutations_one_leave_out()` compared individual methylation profiles of a single sample (regardless if they are cases or controls) against all other samples from the same cohort. To use this function we do not need to split the dataset:

```
#manova (default method)
epi_mva_one_leave_out<- epimutations_one_leave_out(methy)
```

5.2 Unique parameters

The `epi_parameters()` function is useful to set the individual parameters for each approach. The arguments are described in table S3:

Table S3: epimutation function approaches unique parameters

Method	Parameter	Description
manova	pvalue_cutoff	The threshold p-value to select which CpG regions are outliers
mlm		
beta		
iso.forest	outlier_score_cutoff	The threshold to select which CpG regions are outliers
	ntrees	The number of binary trees to build for the model
mahdist.mcd	nsamp	The number of subsets used for initial estimates in the MCD
quantile	window_sz	The maximum distance between CpGs to be considered in the same DMR
	offset_mean/offset_abs	The upper and lower threshold to consider a CpG an outlier
beta	pvalue_cutoff	The minimum p-value to consider a CpG an outlier
	diff_threshold	The minimum methylation difference between the CpG and the mean methylation to consider a position an outlier

`epi_parameters()` with no arguments, returns a list of the default settings for each method:

```
epi_parameters()
```

```
$manova
$manova$pvalue_cutoff
[1] 0.05


$mlm
$mlm$pvalue_cutoff
[1] 0.05
```

```
$isoforest
$isoforest$outlier_score_cutoff
[1] 0.7

$isoforest$ntrees
[1] 100


$mahdistmcd
$mahdistmcd$nsamp
[1] "deterministic"


$quantile
$quantile$window_sz
[1] 1000

$quantile$offset_abs
[1] 0.15

$quantile$qsup
[1] 0.995

$quantile$qinf
[1] 0.005


$beta
$beta$pvalue_cutoff
[1] 1e-06

$beta$diff_threshold
[1] 0.1
```

The set up of any parameter can be done as the following example for `manova`:

```
parameters <- epi_parameters(manova = list("pvalue_cutoff" = 0.01))
parameters$manova$pvalue_cutoff
```

```
[1] 0.01
```

5.3 Results description

The `epimutations` function returns a data frame (tibble) containing all the epimutations identified in the given case sample. If no epimutation is found, the function returns a row containing the case sample information and missing values for all other arguments. The table S4 describes each argument:

Table S4: epimutation function output arguments description

Column name	Description
`epi_id`	Systematic name for each epimutation identified
`sample`	The name of the sample containing that epimutation
`chromosome`	The location of the epimutation
`start end`	
`sz`	The window's size of the event
`cpg_n`	The number of CpGs in the epimutation
`cpg_ids`	The names of CpGs in the epimutation
`outlier_score`	For method `manova` it provides the approximation to F-test and the Pillai score, separated by `/` For method `mlm` it provides the approximation to F-test and the R2 of the model, separated by `/` For method `isoforest` it provides the magnitude of the outlier score. For method `beta` it provides the mean p-value of all GpGs in that DMR For methods `quantile` and `mahdistmcd` it is filled with `NA`.
`pvalue`	For methods `manova` and `mlm` it provides the p-value obtained from the model. For method `quantile`, `isoforest`, `beta` and `mahdistmcd` it is filled with `NA`.
`outlier_direction`	Indicates the direction of the outlier with "hypomethylation" and "hypermethylation." For `manova`, `mlm`, `isoforest`, and `mahdistmcd` it is computed from the values obtained from `bumphunter`. For `beta` is computed from the p value for each CpG using `diff_threshold` and `pvalue_threshold` arguments. For `quantile` it is computed from the location of the sample in the reference distribution (left vs. right outlier).
`adj_pvalue`	For methods `manova` and `mlm` it provides the adjusted p-value with Benjamini-Hochberg based on the total number of regions detected by Bumphunter. For method `quantile`, `isoforest`, `mahdistmcd` and `beta` it is filled with `NA`.
`epi_region_id`	Name of the epimutation region as defined in `candRegsGR`.
`CRE`	cREs (cis-Regulatory Elements) as defined by ENCODE overlapping the epimutation region.
`CRE_type`	Type of cREs (cis-Regulatory Elements) as defined by ENCODE.

As an example we are going to visualize the obtained results with MANOVA method (`epi_mvo`):

```
dim(epi_mvo)
```

```
[1] 51 15
```

```
class(epi_mvo)
```

```
[1] "tbl_df"     "tbl"        "data.frame"
```

```
colnames(epi_mvo)
```

```
 [1] "epi_id"            "sample"            "chromosome"        "start"             "end"
 [7] "cpg_n"             "cpg_ids"           "outlier_score"     "outlier_direction" "pvalue"
[13] "epi_region_id"     "CRE"               "CRE_type"
```

```
head(as.data.frame(epi_mvo[,1:6]), 1)
```

```
        epi_id     sample chromosome    start      end  sz
1 epi_manova_1 GSM2562699      chr19 12777736 12777903 167
```

5.4 Epimutations annotations

The `annotate_epimutations()` function enriches the identified epimutations. It includes information about GENECODE gene names, description of the regulatory feature provided by methylation consortium, the location of the CpG relative to the CpG island, OMIM accession and description number and Ensembl region id, coordinates, type and tissue:

```
rst_mvo <- annotate_epimutations(epi_mvo)
```

```
colnames(rst_mvo[2:3, c(1, 12:14)])
```

```
[1] "epi_id"        "adj_pvalue"    "epi_region_id" "CRE"
```

```
rst_mvo[c(1,29), c(1, 12:14)]```
```

epimutations annotation

epi_id

adj_pvalue

epi_region_id

CRE

1

epi_manova_1

0.0000000

chr19_12776725

EH38E1939817,EH38E1939818,EH38E1939819

29

epi_manova_47

0.0116991

chr7_73894573

EH38E2563868,EH38E2563869

Column name	Description
`GencodeBasicV12_NAME`	Gene names from the basic GENECODE build
`Regulatory_Feature_Group`	Description of the regulatory feature provided by the Methylation Consortium
`Relation_to_Island`	The location of the CpG relative to the CpG island
`OMIM_ACC`	OMIM accession and description number
`OMIM_DESC`	

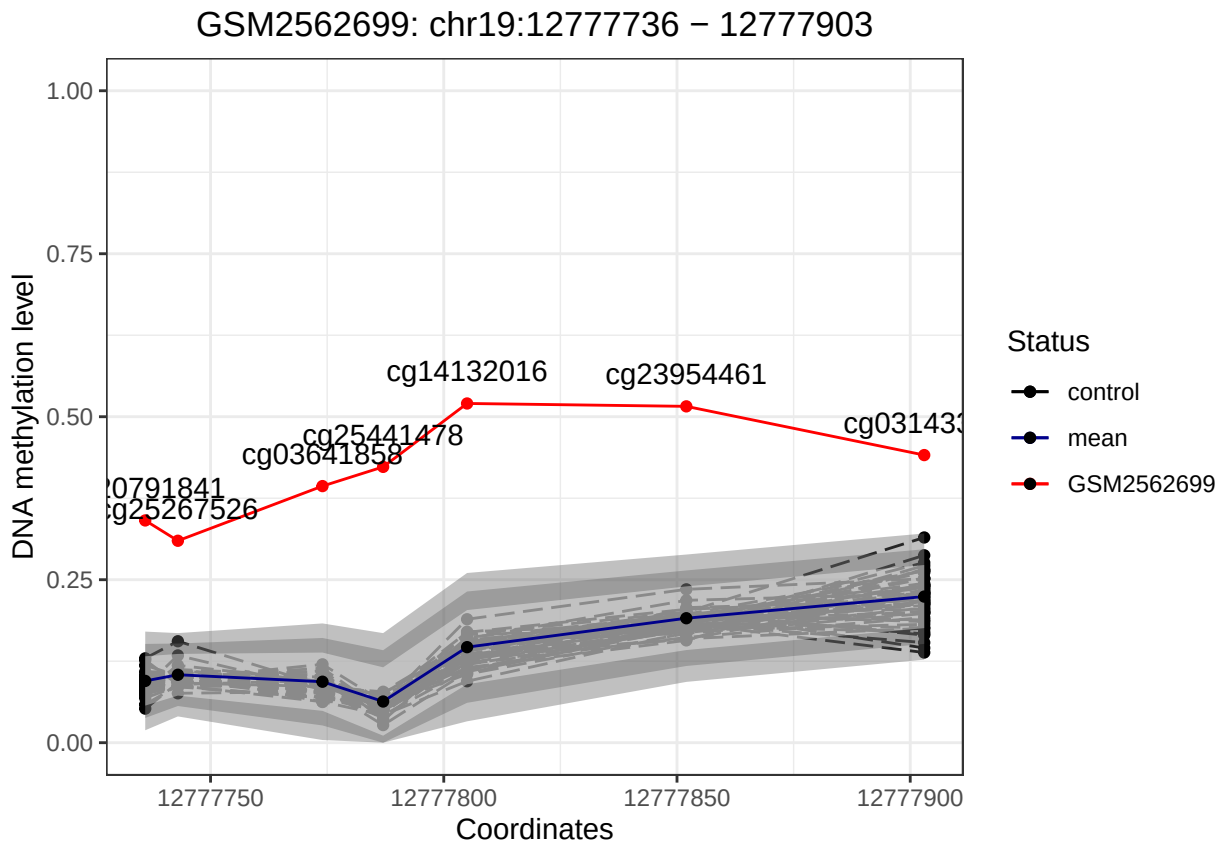
Column name	Description
`ensembl_reg_id`	The Ensembl region id, coordinates, type and tissue
`ensembl_reg_coordinates`	
`ensembl_reg_type`	
`ensembl_reg_tissues`	

5.5 Epimutation visualization

The `plot_epimutations()` function locates the epimutations along the genome. It plots the methylation values of the case sample in red, the control samples in dashed black lines and population mean in blue(figure S3).

```
knitr::opts_chunk$set(fig.pos = 'H')
plot_epimutations(as.data.frame(epi_mvo[1,]), methy)
```

Figure S3: beta values of case sample againts control samples

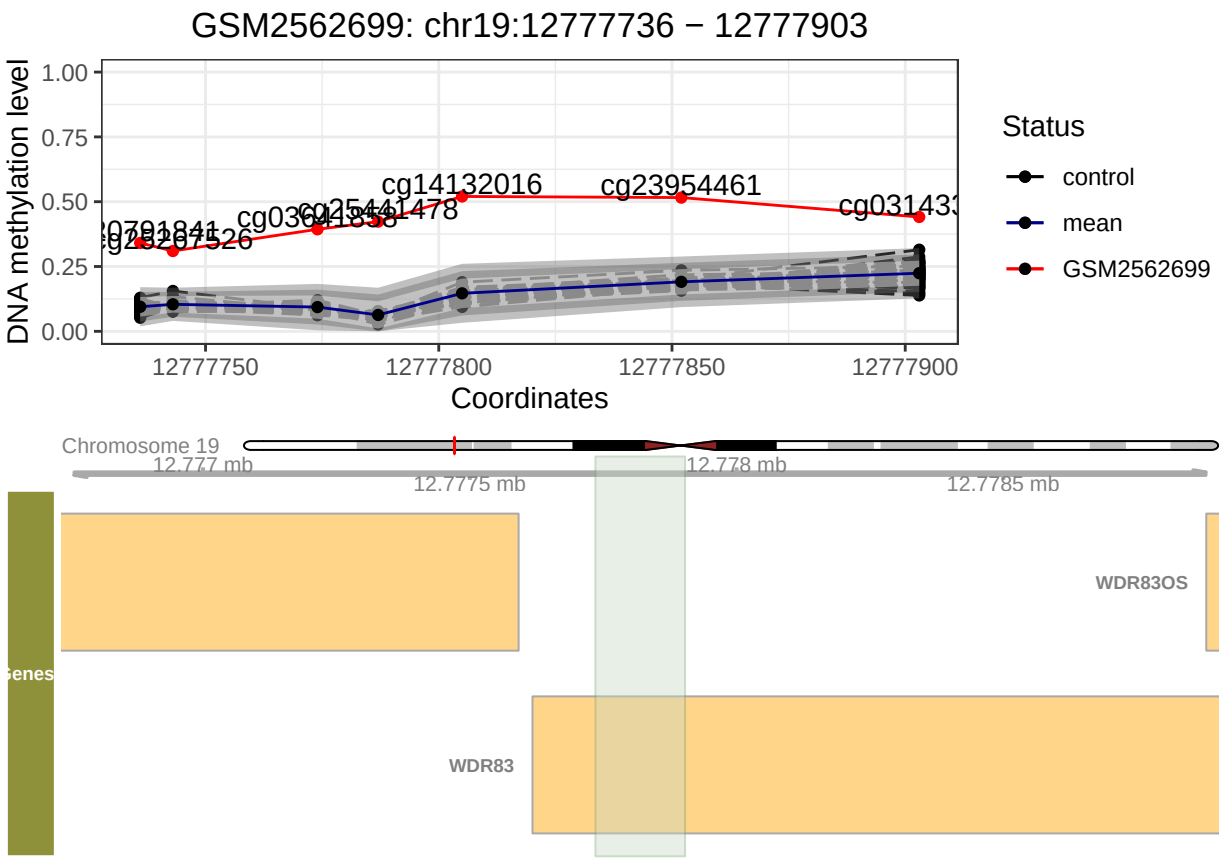


If we set the argument `gene_annot == TRUE` the plot includes the gene annotations: (figure S4):

```
knitr::opts_chunk$set(fig.pos = 'H')
plot_epimutations(as.data.frame(epi_mvo[1,]), methy, genes_annot = TRUE)
```

To plot the chromatin marks H3K4me3, H3K27me3 and H3K27ac we need to specify the argument `regulation = TRUE`(figure S5):

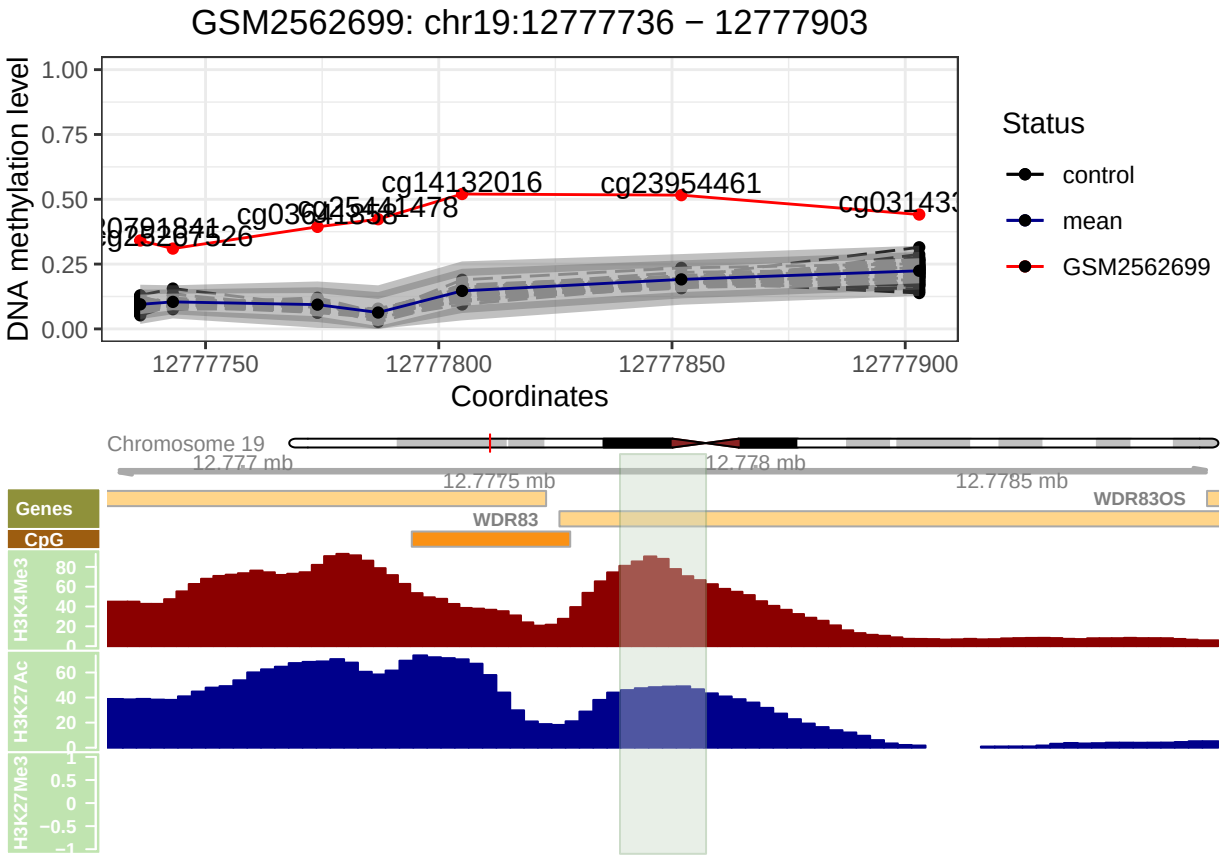
Figure S4: beta values of case sample againts control samples including UCSC gene annotation



- **H3K4me3**: commonly associated with the activation of transcription of nearby genes.
- **H3K27me3**: is used in epigenetics to look for inactive genes.
- **H3K27ac**: is associated with the higher activation of transcription and therefore defined as an active enhancer mark

```
knitr::opts_chunk$set(fig.pos = 'H')
plot_epimutations(as.data.frame(epi_mvo[1,]), methy, regulation = TRUE)
```

Figure S5: beta values of case sample againts control samples including CpG islands



6 Results

6.1 Simulations

We explored the effect of sample size on the identification of 4 epimutations experimentally validated by Grag and colleagues (Garg et al. 2020). We downloaded the data from GEO. We accessed DNA methylation data from a total of 1, 417 individuals from GSE51032 and GSE111629 cohorts.

We evaluated the performance of the methods using True Positive Rate (TPR), False Positive Rate (FPR) and accuracy. We used the TPR to measure the proportion of detected epimutations by our method present in the validated epimutations (table S6). The FPR computes the identified epimutations by our method that were not found by Grag and colleagues. The accuracy measures the closeness of the epimutations detected by our methods to the validated epimutations.

We selected control samples randomly using different sample size: 20, 30, 40, 50, 60, 70, 80, 90 and 100. We set as case samples the individuals containing the validated epimutations (table S6): GSM1235784 sample from GSE51032 cohort; and GSM3035933, GSM3035791, GSM3035807 and GSM3035685 samples from GSE111629. We utilized those case samples to compute TPR and accuracy. However, we measured the FPR selecting all the case samples with no epimutations identified by Grag and colleagues.

We analyzed regions of $\approx$ 20 kb containing $\geq$ 3 GpGs. We executed 100 times the same process for each control sample size.

Table S6: validated epimutations (Garg et al. 2020).

Chromosome	Start	End	Width	Strand	Samples
chr17	46018653	46019185	533	*	GSM1235784/GSM3035791
chr19	11199850	11200147	298	*	GSM3035685
chr5	10249760	10251253	1494	*	GSM3035933
chr5	67583971	67584381	411	*	GSM3035791/GSM3035807

Figures S6, S7, S8, S9, S10 and S11 illustrate the methylation values for each validated epimutation. The y-axis represents the methylation value and the x-axis the location of the epimutation in the genome. We represented the case sample in red, the control samples in dashed black lines and population mean in blue:

Figure S6: GSE51032 cohort samples in the region chr17:46018654-46019184

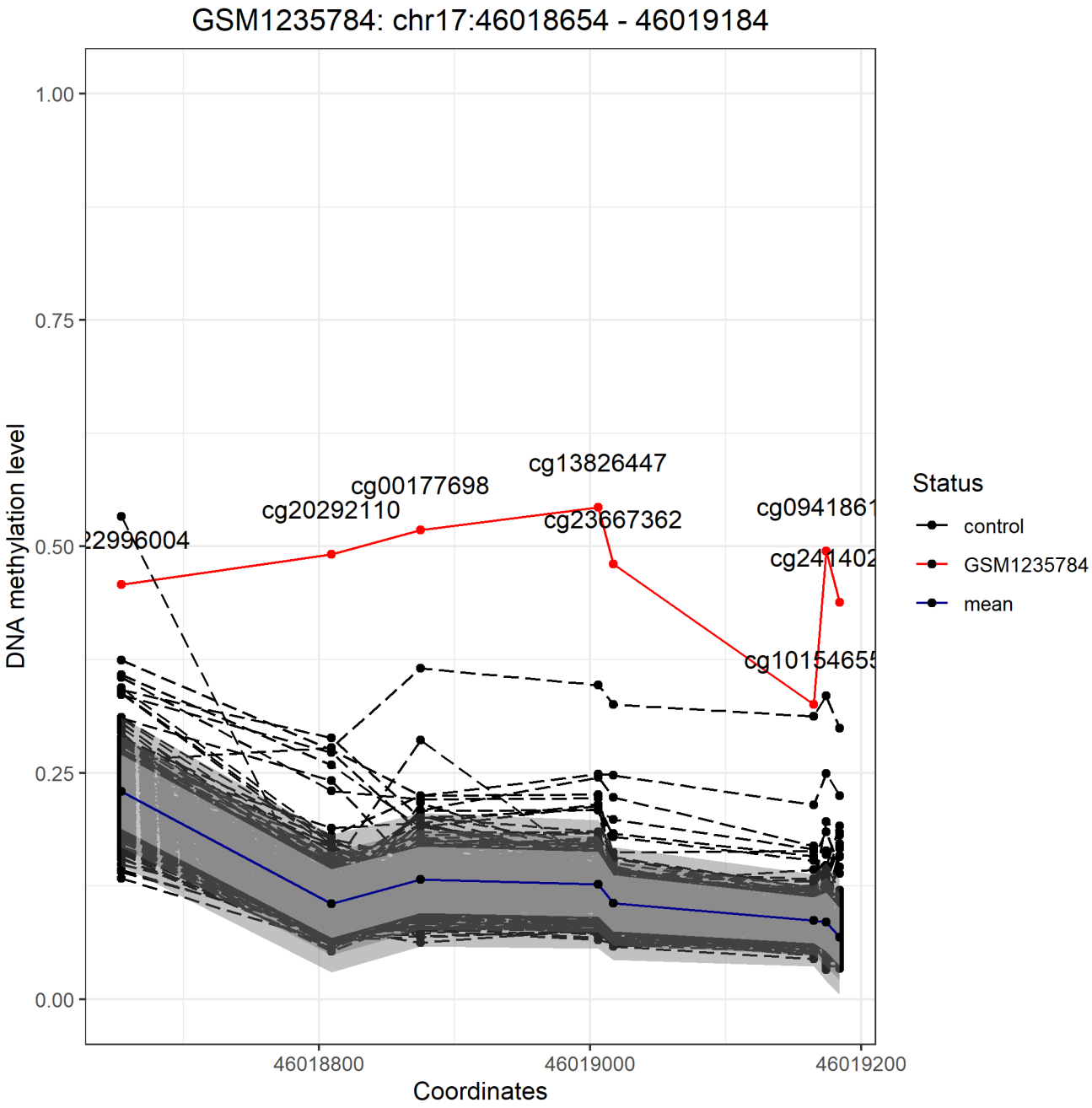


Figure S7: GSE111629 cohort samples in the region chr19:11199851-11200146

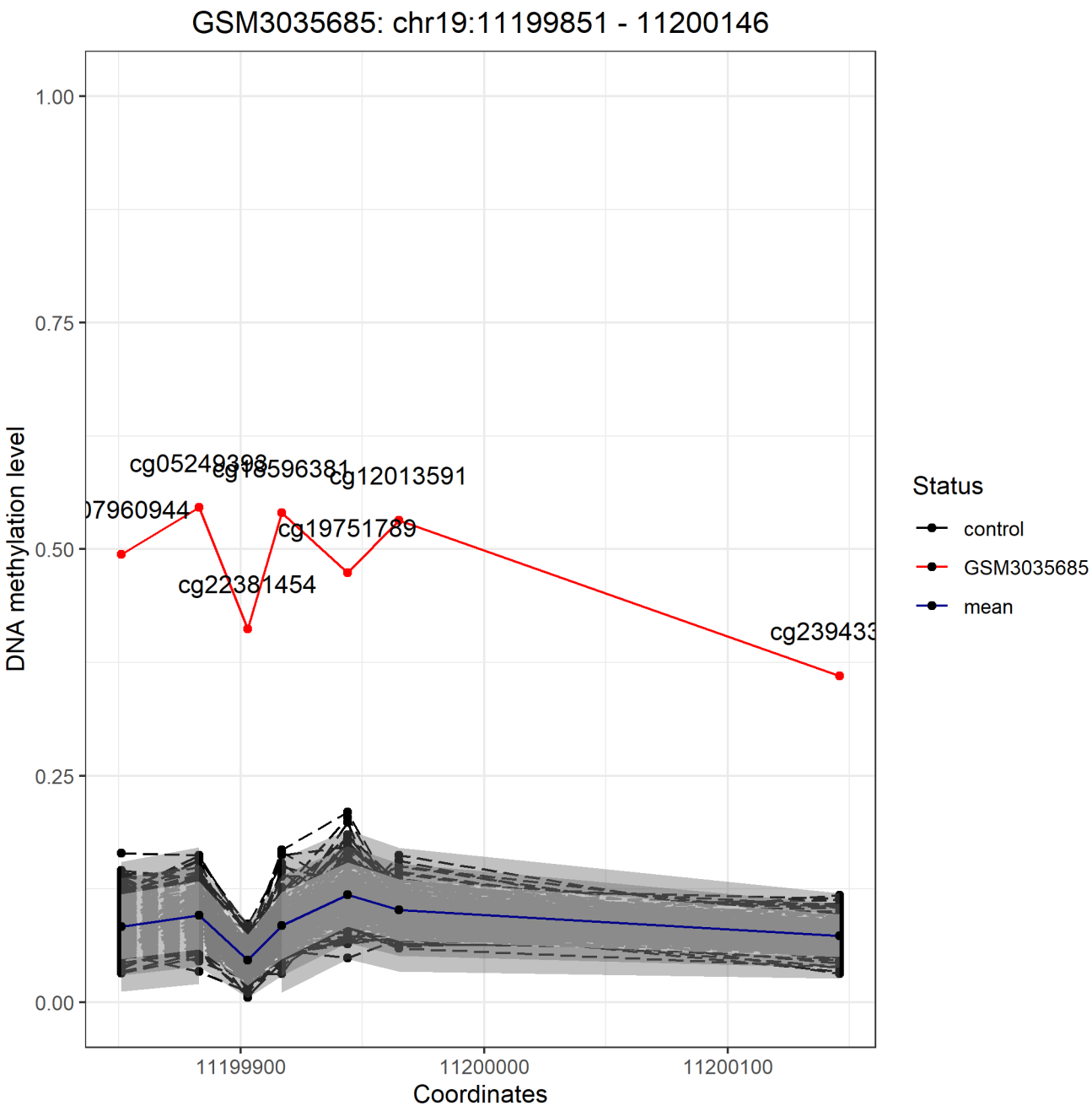


Figure S8: GSE111629 cohort samples in the region chr5:67584194-67584380

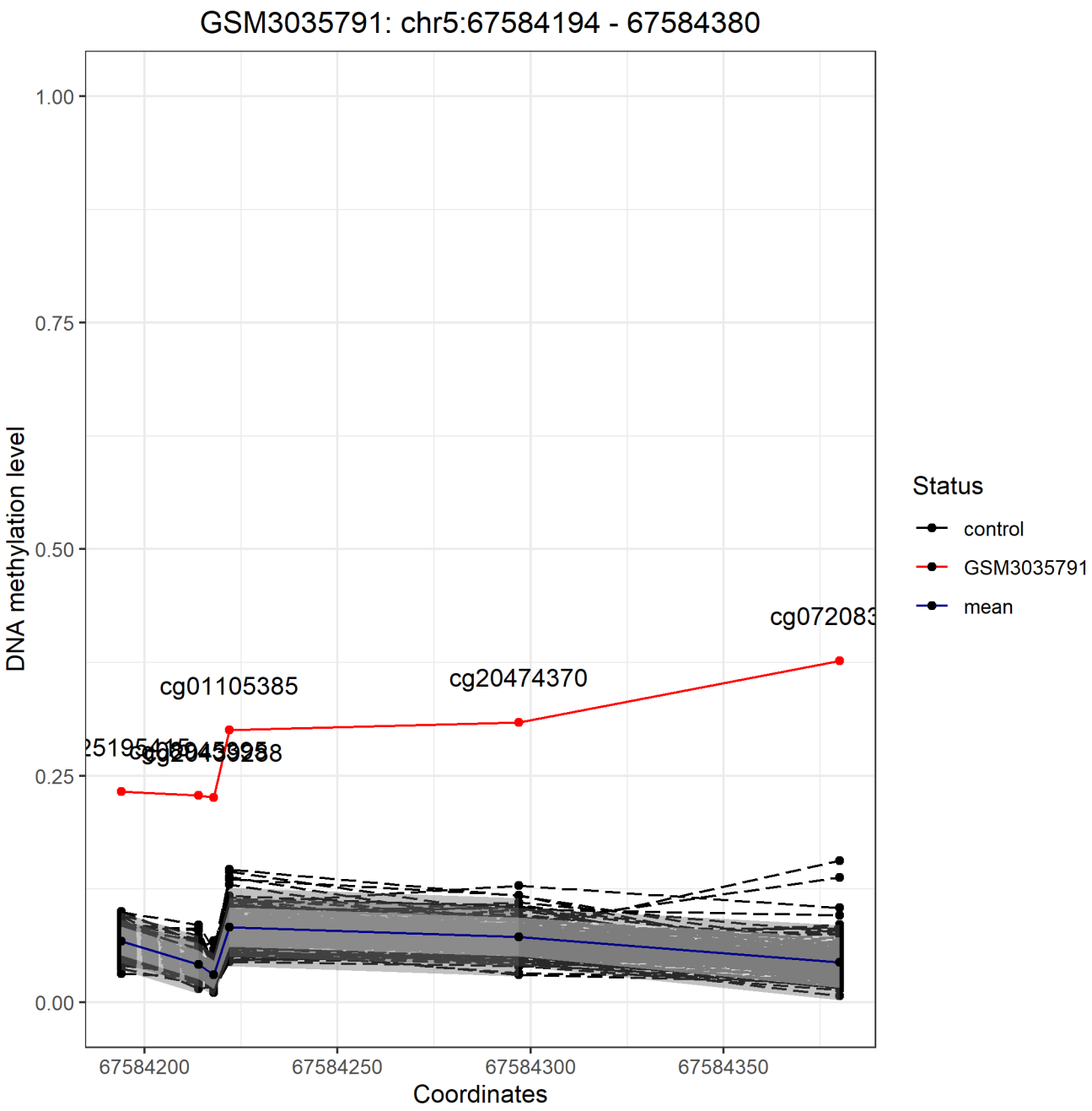


Figure S9: GSE111629 cohort samples in the region chr17:46018654-46019184

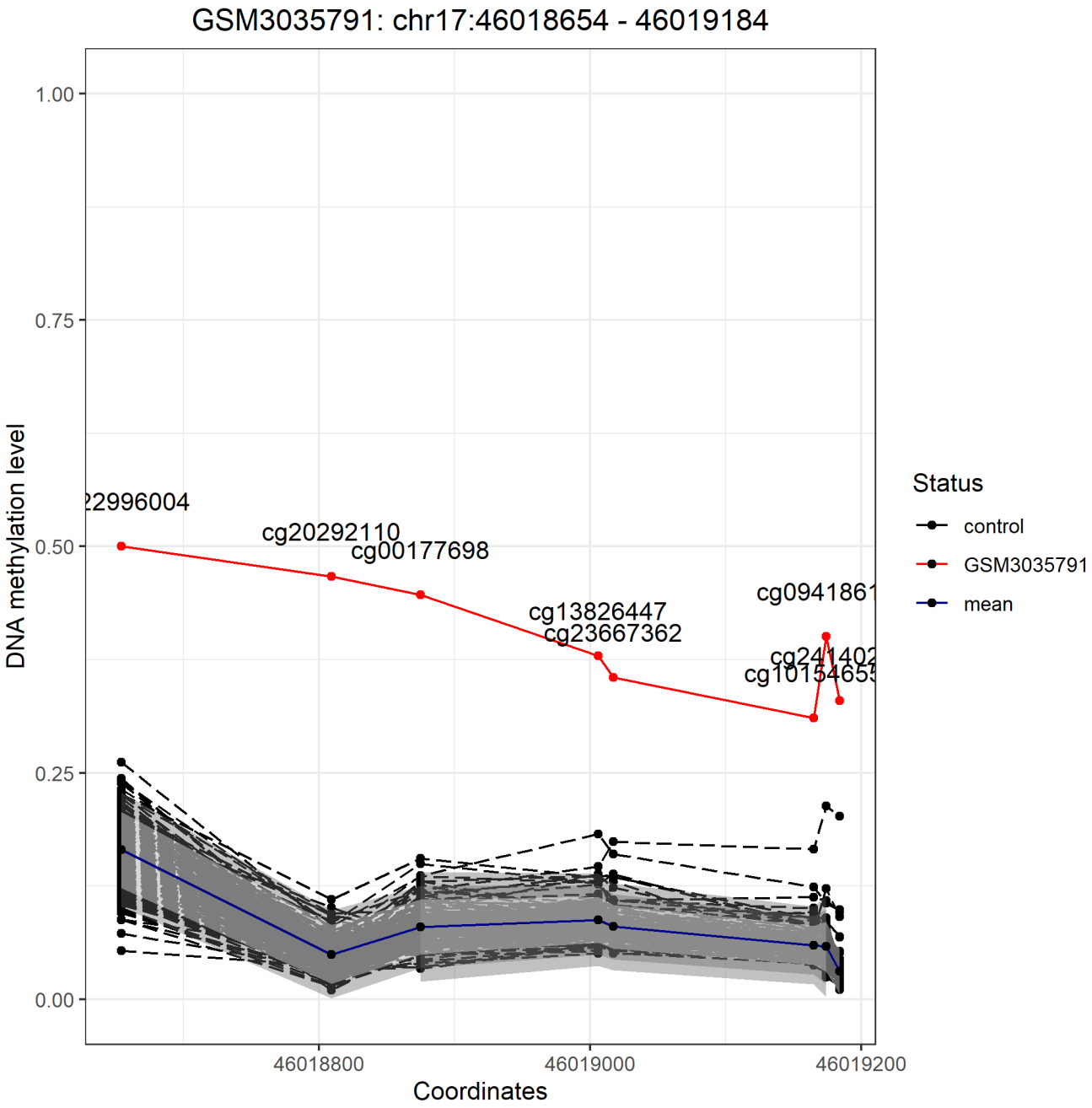


Figure S10: GSE111629 cohort samples in the region chr5:67583972-67584380

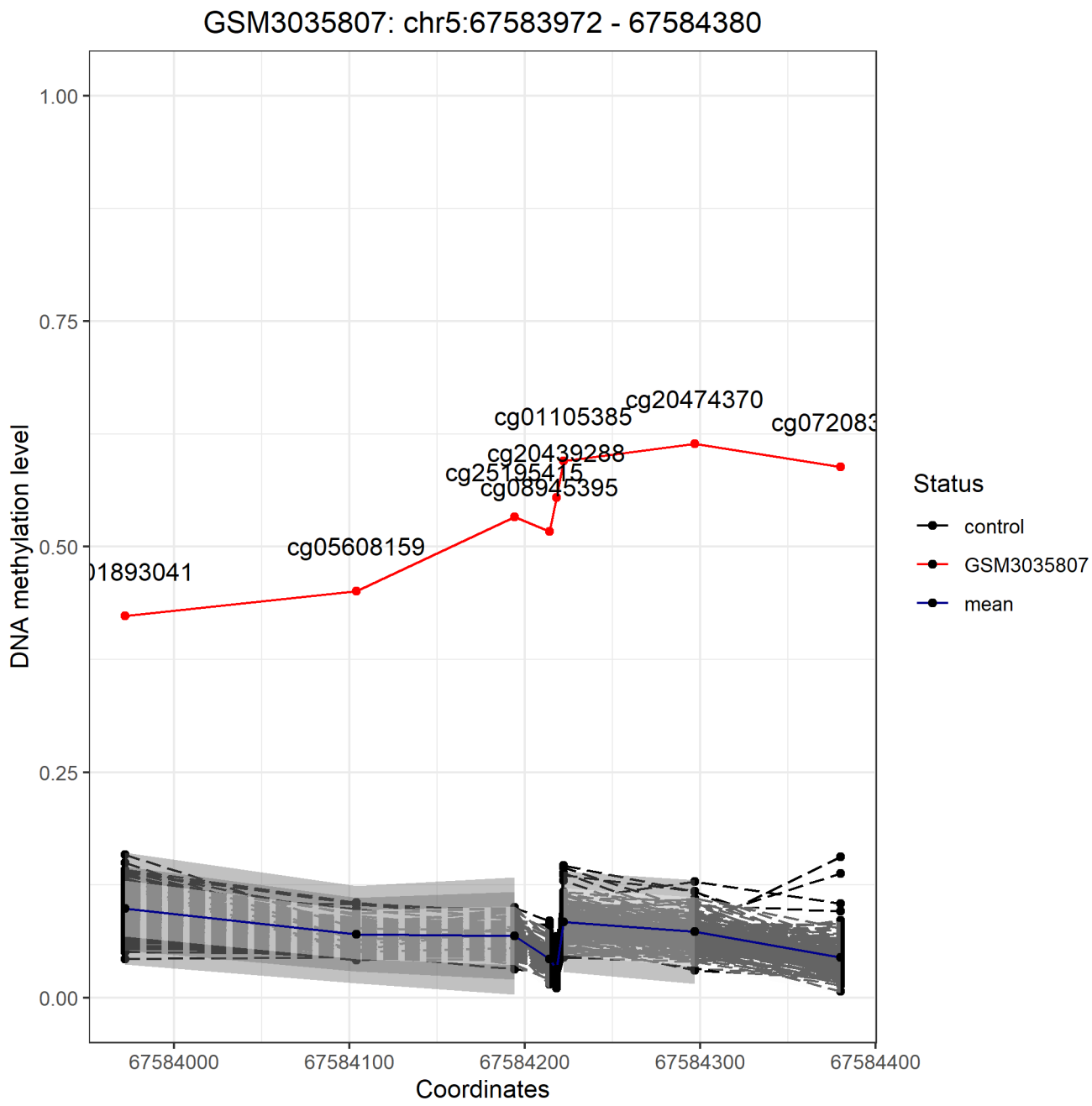
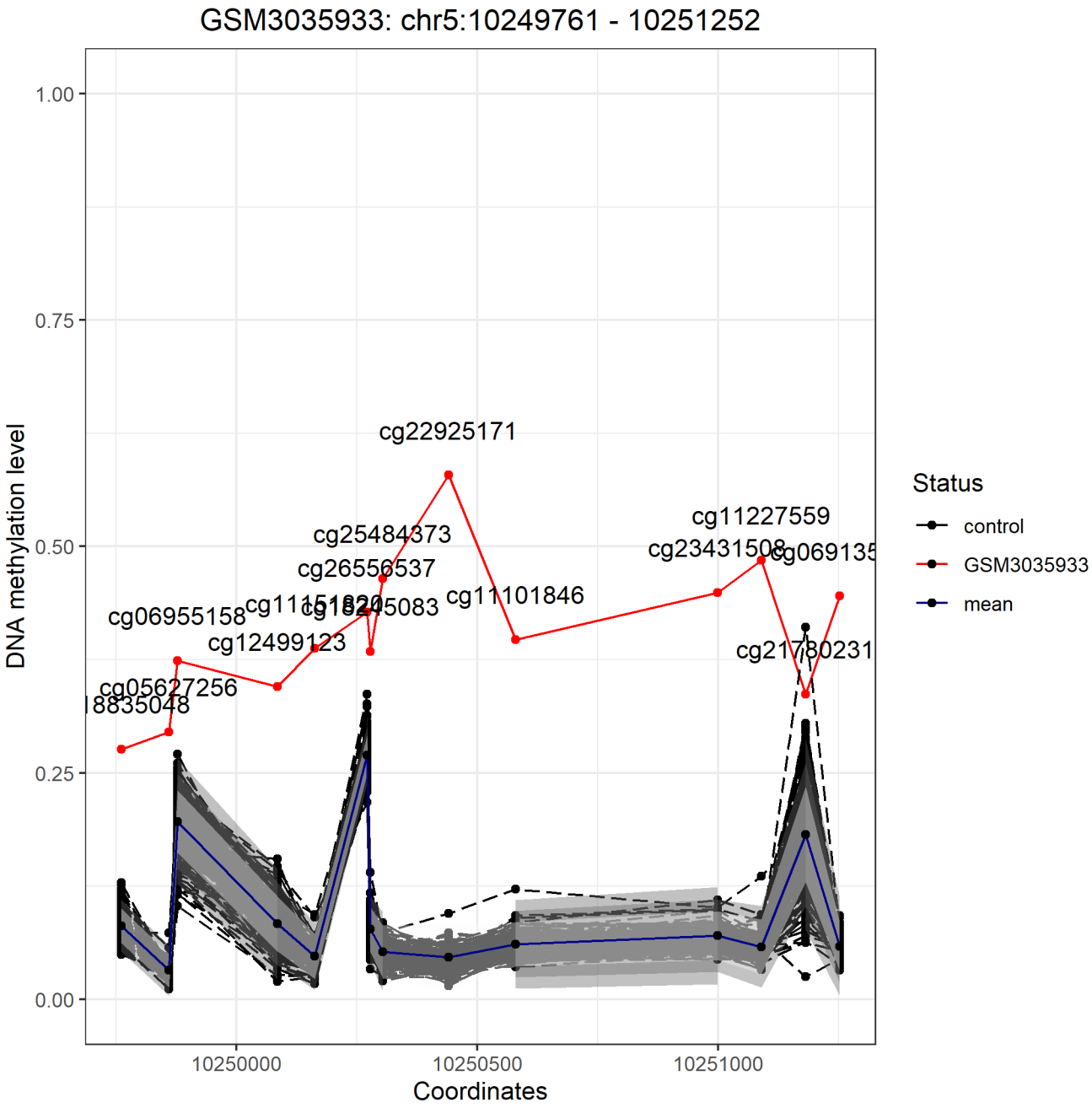


Figure S11: GSE111629 cohort samples in the region chr5:10249761-10251252



The figures S12, S13, S14, S15 and S16 represent the TPR, accuracy and FPR obtained for GSE51032 and GSE111629 cohorts respectively. We separated the results by cohort, case sample (containing the validated epimutation), method and sample size.

We observed similar results for TPR and FPR in both cohorts. The methods manova, mahalanobis distance, mlm, quantile and beta identified the epimutations even if the sample size is small. The TPR in isolation forest increases together with the sample size. However, the accuracy is high even if the sample size is small in this method as well as in manova, mahalanobis distance and mlm. In beta and quantile the accuracy changes depending on the cohort and the analyzed region. Isolation forest, mlm, manova and mahalanobis distance identified in GSE111629 cohort respectively $\approx$ 10%, 7%, 5% and 3% epimutations that were not

present in Grag and colleagues. However, in GSE51032 dataset 40%, 30%, 20% and 10%. The method quantile did not found new epimutations in GSE51032. For GSE111629 cohort the proportion (< 2%) of new epimutations in beta and quantile is lower than in the other methods.

Figure S12: epimutations performance for GSE51032 cohort detecting the epivariation located in chr5:10249760-10251253

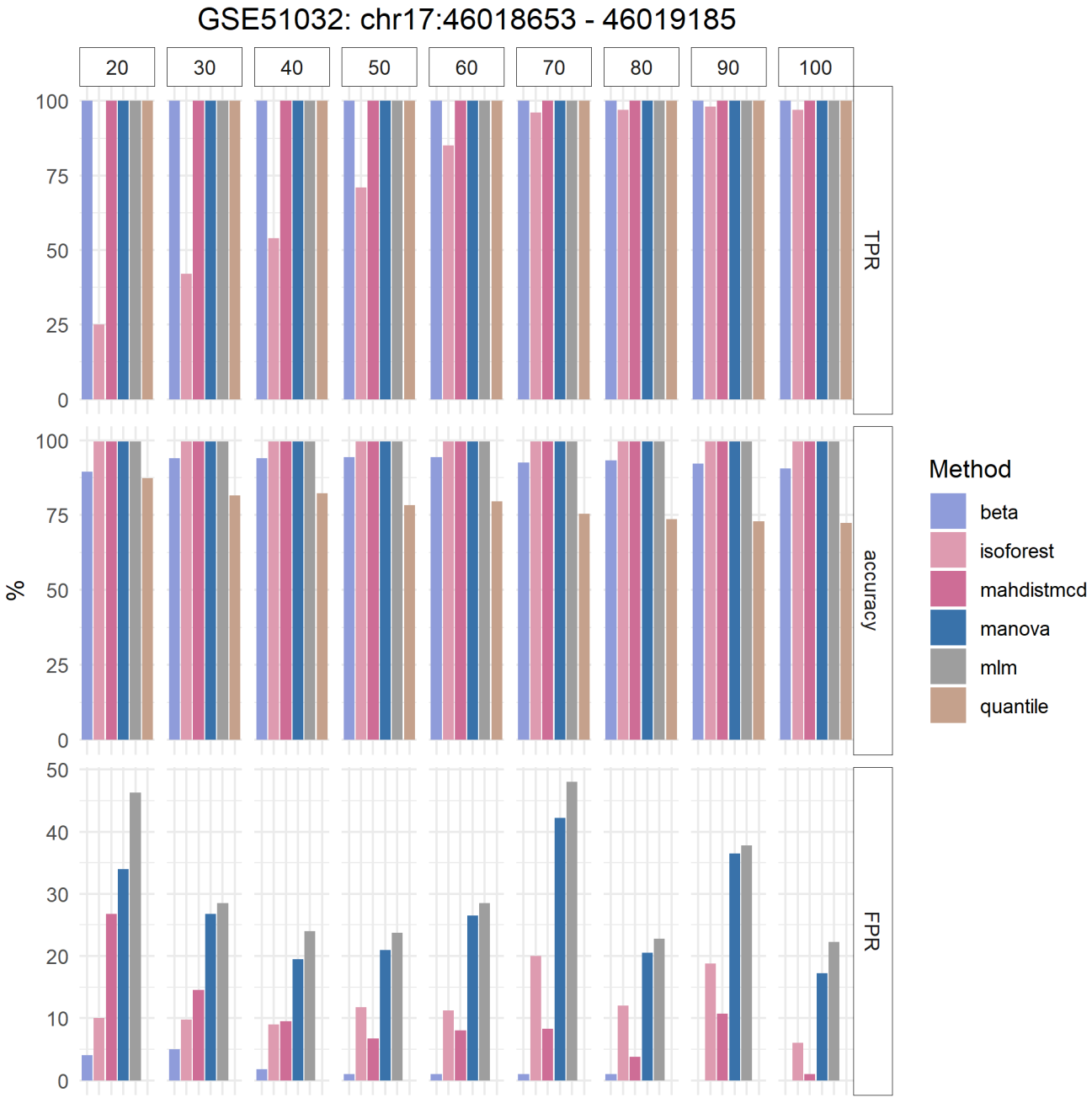


Figure S13: epimutations performance using GSE111629 cohort to detect the epivariation located in chr5:10249760-10251253

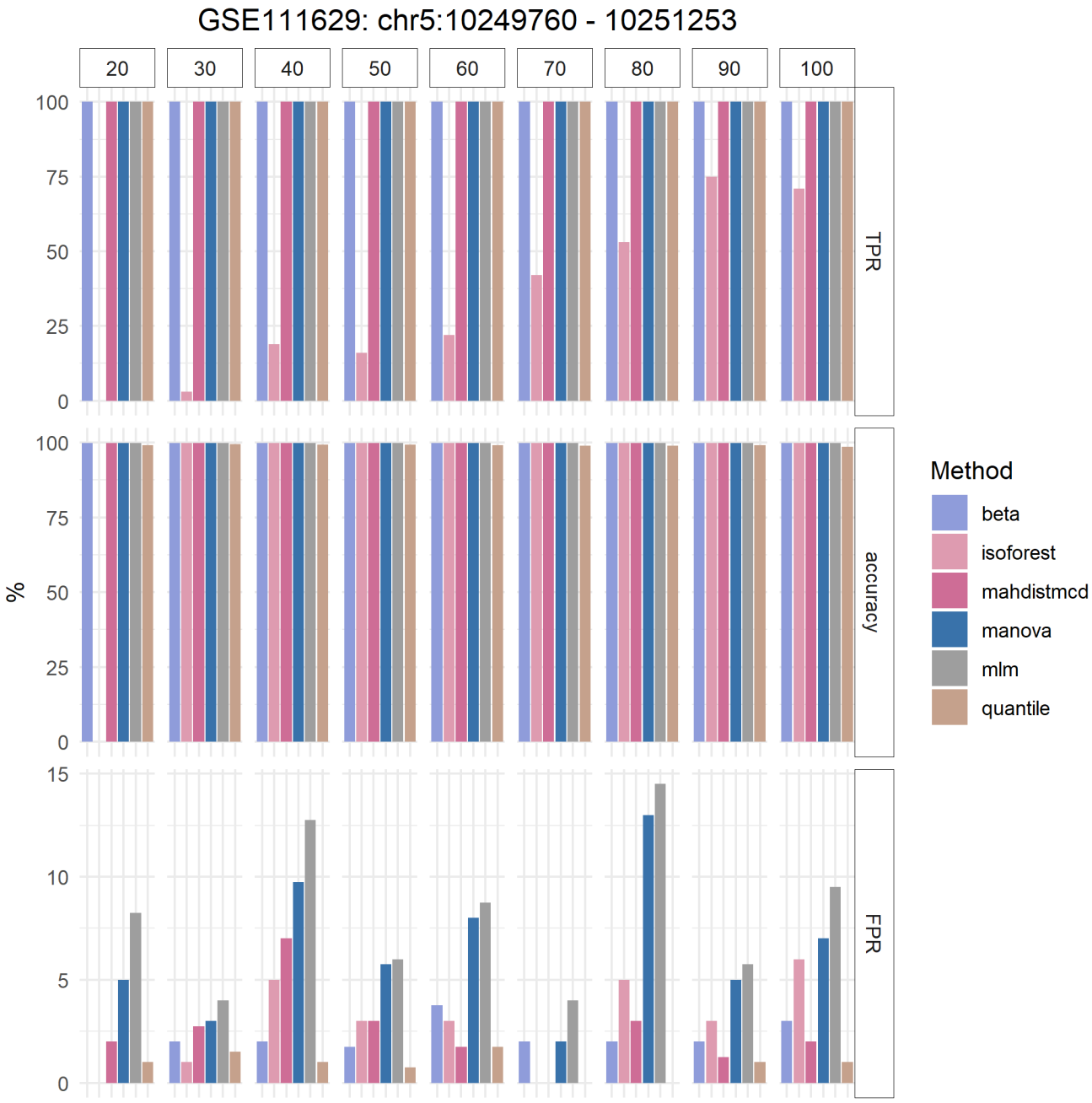


Figure S14: epimutations performance using GSE111629 cohort to detect the epivariation located in chr5:67583971-67584381

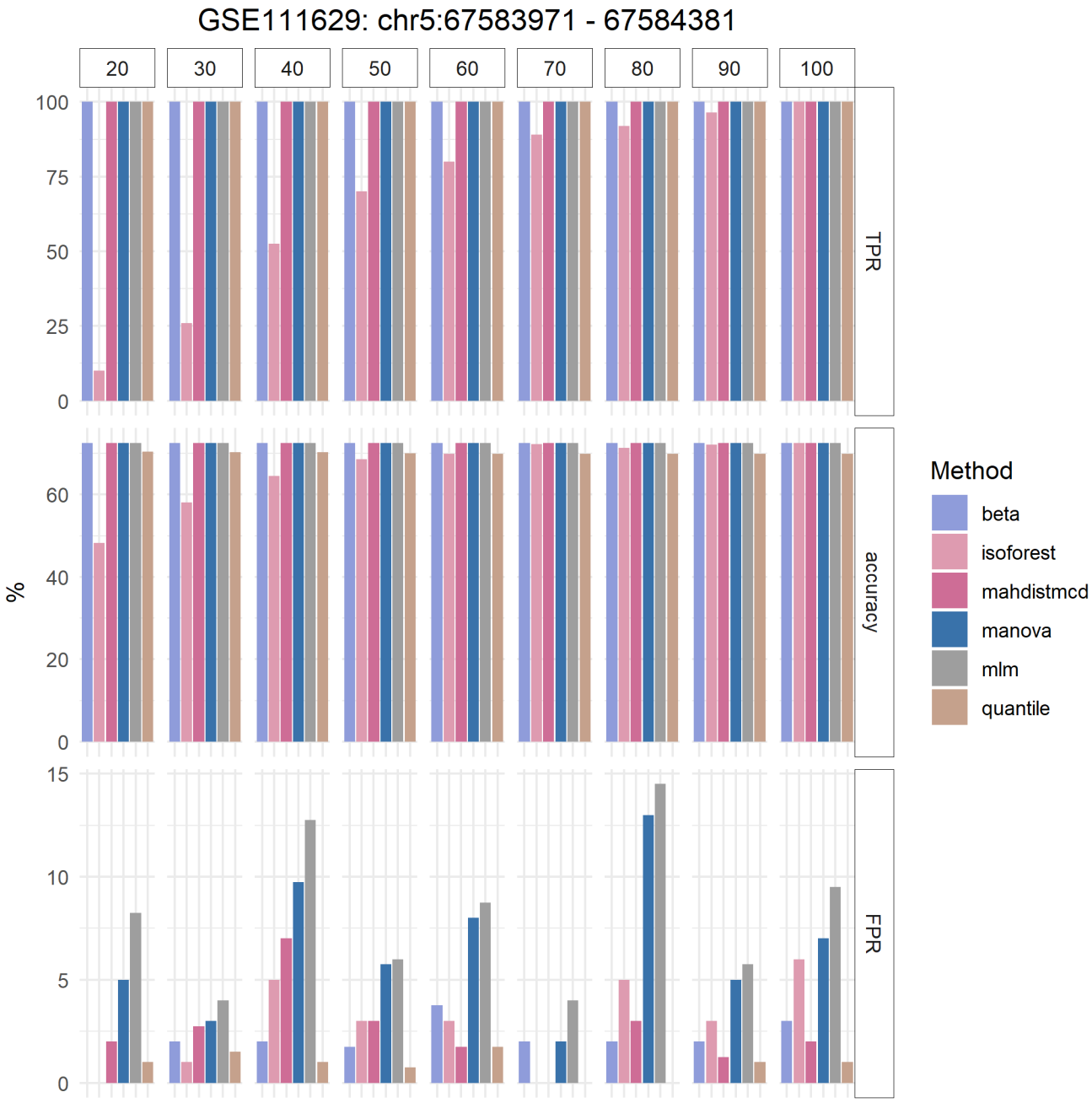


Figure S15: epimutations performance using GSE111629 cohort to detect the epivariation located in chr17:46018653-46019185

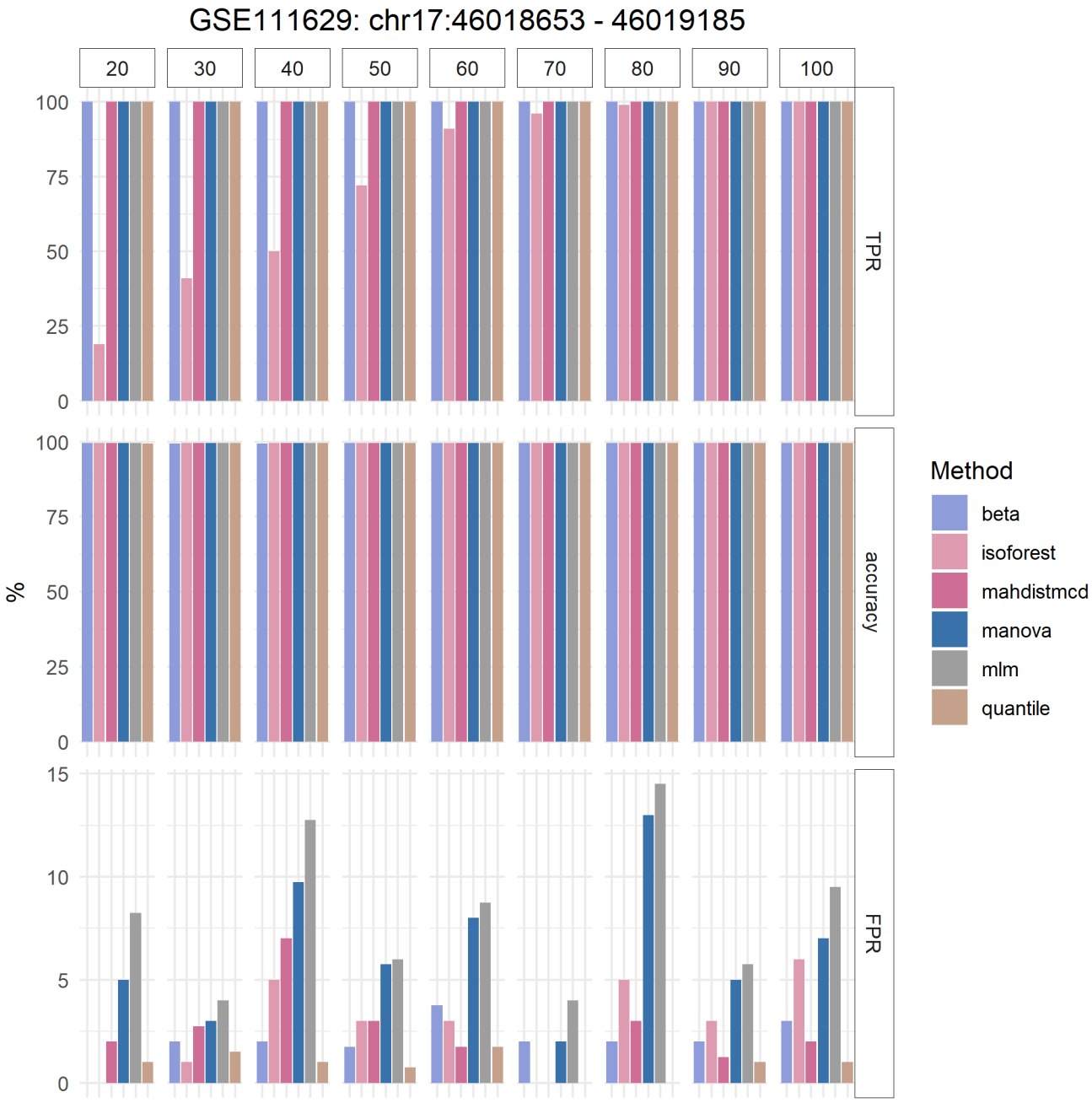
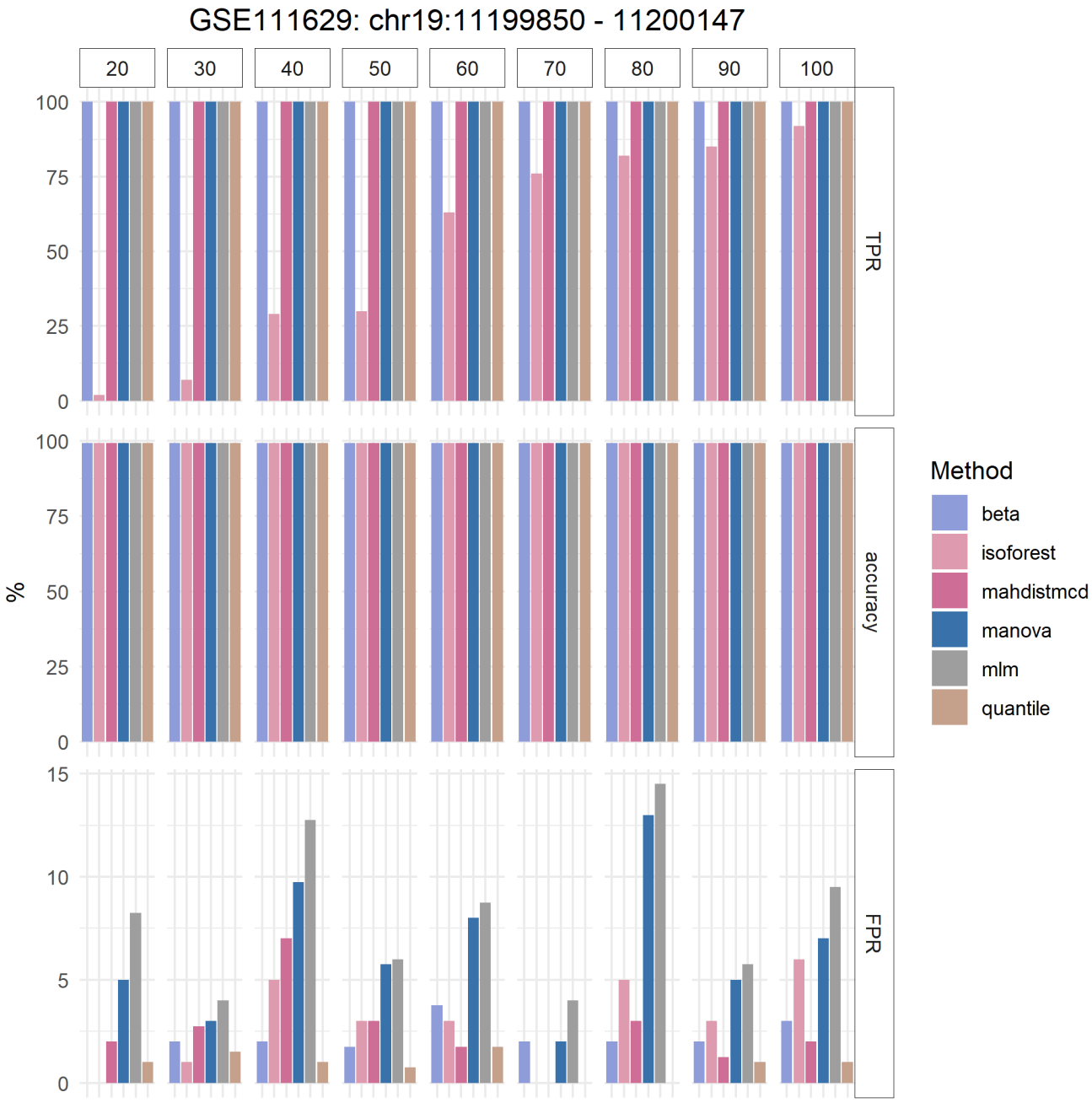


Figure S16: epimutations performance using GSE111629 cohort to detect the epivariation located in chr5:11199850-11200147



method	n	TPR	accuracy	FPR
manova	20	100	99.6	34.00
mlm	20	100	99.6	46.25
mahdistmcd	20	100	99.6	26.75
isoforest	20	25	99.6	10.00
quantile	20	100	87.4	0.00
beta	20	100	89.4	4.00
manova	30	100	99.6	26.75
mlm	30	100	99.6	28.50
mahdistmcd	30	100	99.6	14.50
isoforest	30	42	99.6	9.75
quantile	30	100	81.6	0.00
beta	30	100	94.0	5.00
manova	40	100	99.6	19.50
mlm	40	100	99.6	24.00
mahdistmcd	40	100	99.6	9.50
isoforest	40	54	99.6	9.00
quantile	40	100	82.2	0.00
beta	40	100	93.9	1.75
manova	50	100	99.6	21.00
mlm	50	100	99.6	23.75
mahdistmcd	50	100	99.6	6.75
isoforest	50	71	99.6	11.75
quantile	50	100	78.4	0.00
beta	50	100	94.4	1.00
manova	60	100	99.6	26.50
mlm	60	100	99.6	28.50
mahdistmcd	60	100	99.6	8.00
isoforest	60	85	99.6	11.25
quantile	60	100	79.6	0.00
beta	60	100	94.4	1.00

	method	n	TPR	accuracy	FPR
31	manova	70	100	99.6	42.25
32	mlm	70	100	99.6	48.00
33	mahdistmcd	70	100	99.6	8.25
34	isoforest	70	96	99.6	20.00
35	quantile	70	100	75.5	0.00
36	beta	70	100	92.6	1.00
37	manova	80	100	99.6	20.50
38	mlm	80	100	99.6	22.75
39	mahdistmcd	80	100	99.6	3.75
40	isoforest	80	97	99.6	12.00
41	quantile	80	100	73.7	0.00
42	beta	80	100	93.2	1.00
43	manova	90	100	99.6	36.50
44	mlm	90	100	99.6	37.75
45	mahdistmcd	90	100	99.6	10.75
46	isoforest	90	98	99.6	18.75
47	quantile	90	100	72.9	0.00
48	beta	90	100	92.1	0.00
49	manova	100	100	99.6	17.25
50	mlm	100	100	99.6	22.25
51	mahdistmcd	100	100	99.6	1.00
52	isoforest	100	97	99.6	6.00
53	quantile	100	100	72.3	0.00
54	beta	100	100	90.6	0.00

method	n	TPR	accuracy	FPR
beta	100	100.00000	92.82500	3.00
isoforest	100	90.75000	92.82500	6.00
mahdistmcd	100	100.00000	92.82500	2.00
manova	100	100.00000	92.82500	7.00
mlm	100	100.00000	92.82500	9.50
quantile	100	100.00000	91.85000	1.00
beta	20	100.00000	92.80000	0.00
isoforest	20	10.33333	82.36667	0.00
mahdistmcd	20	100.00000	92.82500	2.00
manova	20	100.00000	92.82500	5.00
mlm	20	100.00000	92.82500	8.25
quantile	20	100.00000	92.05000	1.00
beta	30	100.00000	92.77500	2.00
isoforest	30	19.25000	89.20000	1.00
mahdistmcd	30	100.00000	92.82500	2.75
manova	30	100.00000	92.82500	3.00
mlm	30	100.00000	92.82500	4.00
quantile	30	100.00000	92.15000	1.50
beta	40	100.00000	92.77500	2.00
isoforest	40	37.62500	90.82500	5.00
mahdistmcd	40	100.00000	92.82500	7.00
manova	40	100.00000	92.82500	9.75
mlm	40	100.00000	92.82500	12.75
quantile	40	100.00000	92.10000	1.00
beta	50	100.00000	92.82500	1.75
isoforest	50	47.00000	91.85000	3.00
mahdistmcd	50	100.00000	92.82500	3.00
manova	50	100.00000	92.82500	5.75
mlm	50	100.00000	92.82500	6.00
quantile	50	100.00000	92.07500	0.75

	method	n	TPR	accuracy	FPR
31	beta	60	100.000	92.825	3.75
32	isoforest	60	64.000	92.150	3.00
33	mahdistmcd	60	100.000	92.825	1.75
34	manova	60	100.000	92.825	8.00
35	mlm	60	100.000	92.825	8.75
36	quantile	60	100.000	91.975	1.75
37	beta	70	100.000	92.825	2.00
38	isoforest	70	75.750	92.750	0.00
39	mahdistmcd	70	100.000	92.825	0.00
40	manova	70	100.000	92.825	2.00
41	mlm	70	100.000	92.825	4.00
42	quantile	70	100.000	91.950	0.00
43	beta	80	100.000	92.825	2.00
44	isoforest	80	81.500	92.525	5.00
45	mahdistmcd	80	100.000	92.825	3.00
46	manova	80	100.000	92.825	13.00
47	mlm	80	100.000	92.825	14.50
48	quantile	80	100.000	91.950	0.00
49	beta	90	100.000	92.825	2.00
50	isoforest	90	89.125	92.725	3.00
51	mahdistmcd	90	100.000	92.825	1.25
52	manova	90	100.000	92.825	5.00
53	mlm	90	100.000	92.825	5.75
54	quantile	90	100.000	91.975	1.00

6.2 Methods testing

We run the six implemented methods in `epimutacions` and the Perl script (quantile-perl) from Garg and colleagues. We implemented that script in R (Quantile-R) with a small difference. Quantile-R excludes the target sample when computing the methylation quantiles while in quantile-perl the quantiles are computed including the target sample. We used GSE84727 dataset from GEO which contains a total of 847 whole blood adults samples, 414 schizophrenia cases and 433 controls. This dataset have been previously analyzed by Garg and colleagues.

The methods based in bumphunter detected epimutations in almost all samples. However, beta identified epimutations in $\approx$ 75% of the samples and quantile in $\approx$ 33% (figure S17). All the epimutations detected by Garg and colleagues are found at least by one of our methods. Isolation forest, mlm and manova shared >80% of the epimutations detected. They also identified most of the epimutations found by beta, being only 10% of the epimutations specific to beta. Nevertheless, mahalanobis distance presented more divergent results, the $\approx$ 38% of the detected epimutations are not present in the other methods.

Manova, mlm and isolation forest detected several epimutations per individual. Quantile found $\leq$ 1 epimutation per individual. However, beta and mahalanobis methods identified few epimutations (figure S18).

Figure S17: Overlap between methods

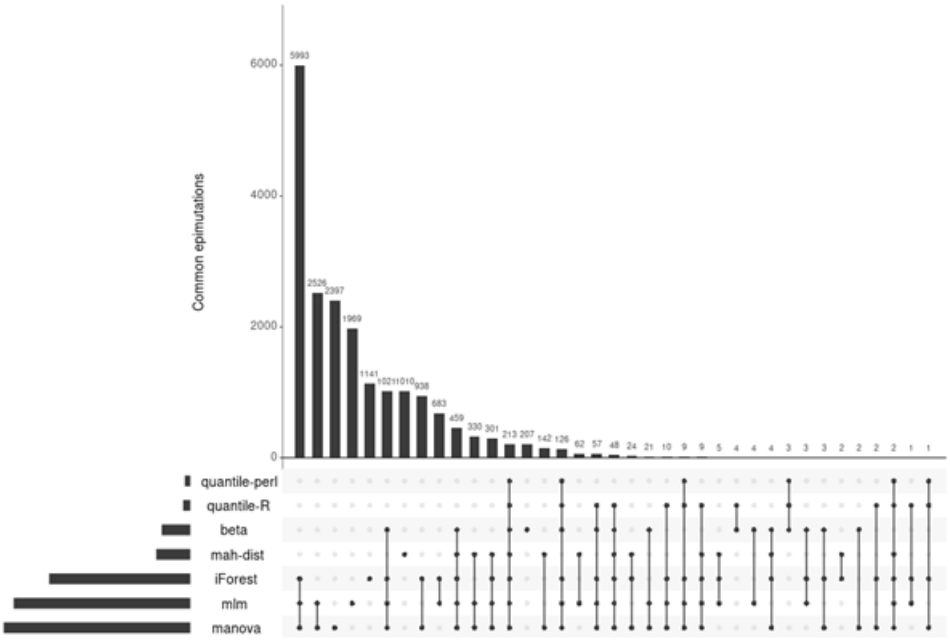
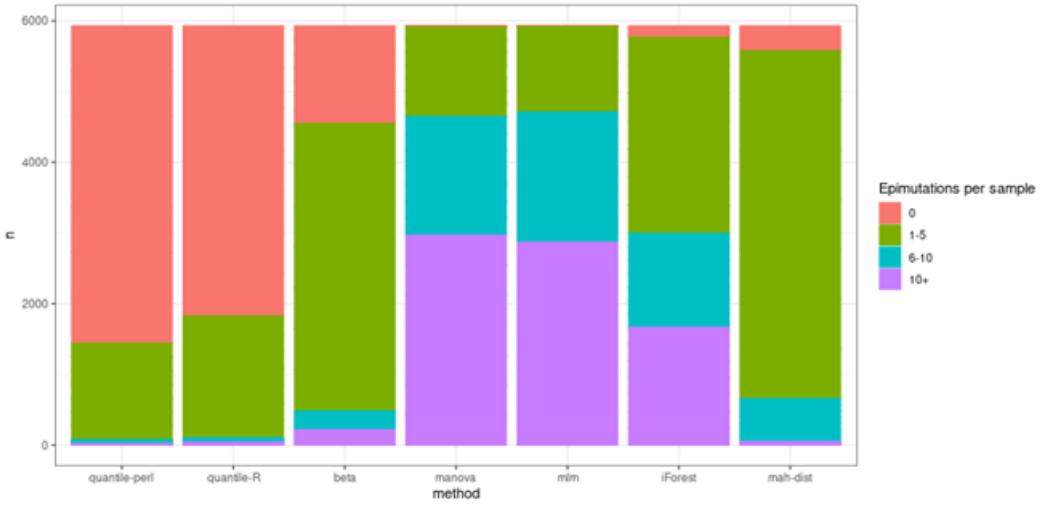


Figure S18: Proportion of individuals with epimutations



7 Acknowledgements

We acknowledge the organizers of the European BioHackathon 2020 for their support.

All the team members of *Project #5* for the contribution to this package:

Name	Surname	ORCID	Affiliation	Team
Leire	Abarrategui	0000-0002-1175-038X	Faculty of Medical Sciences, Newcastle University, Newcastle-Upon-Tyne, UK; Autonomous University of Barcelona (UAB), Barcelona, Spain	Development
Lordstrong	Akano	0000-0002-1404-0295	College of Medicine, University of Ibadan	Development
James	Baye	0000-0002-0078-3688	Wellcome/MRC Cambridge Stem Cell Institute, University of Cambridge, Cambridge CB2 0AW, UK; Department of Physics, University of Cambridge, Cambridge CB2 3DY, UK	Development
Alejandro	Caceres	-	ISGlobal, Barcelona Institute for Global Health, Dr Aiguader 88, 08003 Barcelona, Spain; Centro de Investigación Biomédica en Red en Epidemiología y Salud Pública (CIBERESP), Madrid, Spain	Development

Name	Surname	ORCID	Affiliation	Team
Carles	Hernandez-Ferrer	0000-0002-8029-7160	Centro Nacional de Análisis Genómico (CNAG-CRG), Center for Genomic, Regulation; Barcelona Institute of Science and Technology (BIST), Barcelona, Catalonia, Spain	Development
Pavlo	Hrab	0000-0002-0742-8478	Department of Genetics and Biotechnology, Biology faculty, Ivan Franko National University of Lviv	Validation
Raquel	Manzano	0000-0002-5124-8992	Cancer Research UK Cambridge Institute; University of Cambridge, Cambridge, United Kingdom	Reporting
Margherita	Mutarelli	0000-0002-2168-5059	Institute of Applied Sciences and Intelligent Systems (ISASI-CNR)	Validation
Carlos	Ruiz-Arenas	0000-0002-6014-3498	Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Barcelona, Spain; Universitat Pompeu Fabra (UPF), Barcelona, Spain	Reporting

8 Session Info

```
sessionInfo()
```

```
R version 4.1.0 (2021-05-18)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 18363)

Matrix products: default

locale:
[1] LC_COLLATE=English_United Kingdom.1252  LC_CTYPE=English_United Kingdom.1252
[3] LC_MONETARY=English_United Kingdom.1252 LC_NUMERIC=C
[5] LC_TIME=English_United Kingdom.1252

attached base packages:
[1] stats4    parallel  stats     graphics  grDevices utils     datasets  methods
[9] base

other attached packages:
 [1] IlluminaHumanMethylation450kanno.ilmn12.hg19_0.6.0
 [2] IlluminaHumanMethylationEPICmanifest_0.3.0
 [3] IlluminaHumanMethylation450kmanifest_0.4.0
```

```
 [4] epimutacionsData_0.9.3
 [5] ExperimentHubData_1.19.0
 [6] AnnotationHubData_1.23.0
 [7] ExperimentHub_2.0.0
 [8] AnnotationHub_3.1.4
 [9] BiocFileCache_2.0.0
[10] dbplyr_2.1.1
[11] BSgenome.Hsapiens.UCSC.hg19_1.4.3
[12] BSgenome_1.61.0
[13] dplyr_1.0.7
[14] rtracklayer_1.52.0
[15] knitr_1.33
[16] IlluminaHumanMethylationEPICanno.ilm10b2.hg19_0.6.0
[17] minfi_1.39.1
[18] bumphunter_1.35.0
[19] locfit_1.5-9.4
[20] iterators_1.0.13
[21] foreach_1.5.1
[22] Biostrings_2.60.1
[23] XVector_0.32.0
[24] SummarizedExperiment_1.23.1
[25] MatrixGenerics_1.5.1
[26] matrixStats_0.59.0
[27] GenomicRanges_1.44.0
[28] GenomeInfoDb_1.29.3
[29] IRanges_2.26.0
[30] S4Vectors_0.30.0
[31] epimutacions_0.1.0
[32] Biobase_2.52.0
[33] BiocGenerics_0.38.0
[34] futile.logger_1.4.3

loaded via a namespace (and not attached):
  [1] utf8_1.2.1                    RUnit_0.4.32
  [3] tidyselect_1.1.1              RSQLite_2.2.7
  [5] AnnotationDbi_1.55.1          grid_4.1.0
  [7] BiocParallel_1.26.1           munsell_0.5.0
  [9] codetools_0.2-18              preprocessCore_1.54.0
 [11] withr_2.4.2                   colorspace_2.0-2
 [13] filelock_1.0.2                OrganismDbi_1.35.0
 [15] highr_0.9                     rstudioapi_0.13
 [17] optparse_1.6.6                GenomeInfoDbData_1.2.6
 [19] bit64_4.0.5                   rhdf5_2.36.0
 [21] vctrs_0.3.8                   generics_0.1.0
 [23] lambda.r_1.2.4                xfun_0.24
 [25] R6_2.5.0                      illuminaio_0.34.0
 [27] bitops_1.0-7                  rhdf5filters_1.4.0
 [29] cachem_1.0.5                  reshape_0.8.8
 [31] DelayedArray_0.18.0           assertthat_0.2.1
 [33] promises_1.2.0.1              BiocIO_1.3.0
 [35] scales_1.1.1                  biocViews_1.61.0
 [37] rlang_0.4.11                  genefilter_1.74.0
 [39] systemfonts_1.0.2             splines_4.1.0
 [41] GEOquery_2.61.0               BiocManager_1.30.16
```

```
 [43] yaml_2.2.1                    GenomicFeatures_1.45.0
 [45] httpuv_1.6.1                  RBGL_1.69.0
 [47] tools_4.1.0                   nor1mix_1.3-0
 [49] ellipsis_0.3.2                kableExtra_1.3.4
 [51] RColorBrewer_1.1-2            siggenes_1.67.0
 [53] Rcpp_1.0.7                    plyr_1.8.6
 [55] sparseMatrixStats_1.4.0       progress_1.2.2
 [57] zlibbioc_1.38.0               purrr_0.3.4
 [59] RCurl_1.98-1.3                prettyunits_1.1.1
 [61] openssl_1.4.4                 magrittr_2.0.1
 [63] data.table_1.14.0             futile.options_1.0.1
 [65] hms_1.1.0                     mime_0.11
 [67] evaluate_0.14                 xtable_1.8-4
 [69] XML_3.99-0.6                  mclust_5.4.7
 [71] compiler_4.1.0                biomaRt_2.49.2
 [73] tibble_3.1.2                  crayon_1.4.1
 [75] htmltools_0.5.1.1             later_1.2.0
 [77] tidyr_1.1.3                   DBI_1.1.1
 [79] formatR_1.11                  MASS_7.3-54
 [81] rappdirs_0.3.3                Matrix_1.3-4
 [83] getopt_1.20.3                 readr_1.4.0
 [85] cli_3.0.1                     quadprog_1.5-8
 [87] pkgconfig_2.0.3               GenomicAlignments_1.28.0
 [89] xml2_1.3.2                    svglite_2.0.0
 [91] annotate_1.71.0               rngtools_1.5
 [93] stringdist_0.9.6.3            multtest_2.48.0
 [95] beanplot_1.2                  webshot_0.5.2
 [97] AnnotationForge_1.35.0        rvest_1.0.1
 [99] BiocCheck_1.29.10             doRNG_1.8.2
[101] scrime_1.3.5                  stringr_1.4.0
[103] digest_0.6.27                 graph_1.71.2
[105] rmarkdown_2.9                 base64_2.0
[107] DelayedMatrixStats_1.15.0     restfulr_0.0.13
[109] curl_4.3.2                    shiny_1.6.0
[111] Rsamtools_2.8.0               rjson_0.2.20
[113] lifecycle_1.0.0               nlme_3.1-152
[115] jsonlite_1.7.2                Rhdf5lib_1.14.2
[117] viridisLite_0.4.0             askpass_1.1
[119] limma_3.48.1                  fansi_0.5.0
[121] pillar_1.6.1                  lattice_0.20-44
[123] KEGGREST_1.33.0               fastmap_1.1.0
[125] httr_1.4.2                    survival_3.2-11
[127] interactiveDisplayBase_1.31.0 glue_1.4.2
[129] png_0.1-7                     BiocVersion_3.14.0
[131] bit_4.0.4                     stringi_1.6.2
[133] HDF5Array_1.20.0              blob_1.2.2
[135] memoise_2.0.0
```

References

Aref-Eshghi, Erfan, Eric G. Bend, Samantha Colaiacovo, Michelle Caudle, Rana Chakrabarti, Melanie Napier, Lauren Brick, et al. 2019. "Diagnostic Utility of Genome-Wide DNA Methylation Testing

in Genetically Unsolved Individuals with Suspected Hereditary Conditions." *The American Journal of Human Genetics*. https://doi.org/https://doi.org/10.1016/j.ajhg.2019.03.008.

Aryee, Martin J, Andrew E Jaffe, Hector Corrada-Bravo, Christine Ladd-Acosta, Andrew P Feinberg, Kasper D Hansen, and Rafael A Irizarry. 2014. "Minfi: A Flexible and Comprehensive Bioconductor Package for the Analysis of Infinium DNA Methylation Microarrays." *Bioinformatics* 30 (10): 1363–69.

Barbosa, Mafalda, Ricky S Joshi, Paras Garg, Alejandro Martin-Trujillo, Nihir Patel, Bharati Jadhav, Corey T Watson, et al. 2018. "Identification of Rare de Novo Epigenetic Variations in Congenital Disorders." *Nature Communications* 9 (1): 1–11.

Cortes, David, and Maintainer David Cortes. 2021. "Package 'Isotree'."

European-Commission. 2020. "EU Research on Rare Diseases." https://ec.europa.eu/info/research-and-innovation/research-area/health-research-and-innovation/rare-diseases_en.

Garg, Paras, Bharati Jadhav, Oscar L Rodriguez, Nihir Patel, Alejandro Martin-Trujillo, Miten Jain, Sofie Metsu, et al. 2020. "A Survey of Rare Epigenetic Variation in 23,116 Human Genomes Identifies Disease-Relevant Epivariations and CGG Expansions." *The American Journal of Human Genetics* 107 (4): 654–69.

Lionel, Anath C, Gregory Costain, Nasim Monfared, Susan Walker, Miriam S Reuter, S Mohsen Hosseini, Bhooma Thiruvahindrapuram, et al. 2018. "Improved Diagnostic Yield Compared with Targeted Gene Sequencing Panels Suggests a Role for Whole-Genome Sequencing as a First-Tier Genetic Test." *Genetics in Medicine* 20 (4): 435–43.

Maechler, Martin, Peter Rousseeuw, Christophe Croux, Valentin Todorov, Andreas Ruckstuhl, Matias Salibian-Barrera, Tobias Verbeke, Manuel Koller, Eduardo LT Conceicao, and Maria Anna di Palma. 2021. "Package 'Robustbase'." *Basic Robust Statistics*.

Martín, Diego Garrido. 2020. "A Multivariate Approach to Study the Genetic Determinants of Phenotypic Traits." PhD thesis, Universitat Pompeu Fabra.