

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

Supplementary material

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

2021-07-02

Introduction

Background

Rare diseases are pathologies with a low prevalence (< 1 per 2,000 people). Most of these pathologies have an onset during childhood and a strong genetic etiology. Consequently, rare disease diagnosis has relied on identifying genetic and genomic mutations that can cause the disease. Although these variants have provided a diagnosis for many patients and families, around 60% of the cases remained undiagnosed (Lionel et al. 2018). Aberrant methylation can be an underlying cause of undiagnosed patients, either as a primary event (a.k.a. epimutation) or as a functional consequence of chromatin dysregulation by genetic or environmental agents (a.k.a. episignature). Epimutations are the cause of some rare diseases, such as Prader-Willi, Angelman or Beckwith-Wiedemann syndromes (Aref-Eshghi et al. 2019) and some human malformations (Serra-Juhé et al. 2015). Syndrome-specific episignatures are increasingly defined as biomarkers for a growing number of disorders (Aref-Eshghi et al. 2019). Therefore, tools to detect epimutations and episignatures should be made available to the rare disease community and included in standardized analysis workflows.

The package `epimutacions` provides tools for analysing epivariants using multiple outlier detection approaches. In addition, functions to plot and annotate the identified epimutations. The full `epimutacions` user's guide is available in this vignette. The main function to estimate the epivariants is called `epimutations()`.

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.

Methodology

The `epimutacions` package computes a genome-wide DNA methylation analysis to detect the epigenetic variants to be considered as biomarkers for samples with rare diseases (epimutations). The method compares a case sample with suspected rare disease against a reference panel. The package focused on the detection of outliers in DNA methylation patterns associated with the diseases as proposed by (Aref-Eshghi et al. 2019).

The identification of relevant genomic methylation regions for a given sample having a rare disease will be driven by detecting differentially methylated CpG sites when comparing beta values of all control samples with the given proband. Firstly, bump-hunter (Jaffe et al. 2012) approach is used to identify the Differentially Methylated Regions (DMRs). After that, CpGs in the proband sample are tested in those DMRs in order to identify regions with CpGs being outliers when comparing with the reference panel. To this end, different anomaly detection statistical approaches are used. These include Multivariate Analysis of Variance (MANOVA) (Friedrich et al. 2017), Multivariate Linear Model (Martín 2020), isolation forest (Cortes and Cortes 2021) and robust mahalanobis distance (Maechler et al. 2021). However, Barbosa (M et al. 2018)

and Beta methods do not use bump-hunter output. Barbosa (M et al. 2018) checks for each CpG, if the proband's measurement is an outlier. Then, it calls an epimutation to those regions where 3 contiguous CpGs are outliers, and they are separated by less than 500 base pairs. Beta approach models the DNA methylation data using a beta distribution.

Input data

The package allows two different types of inputs:

- (1) IDAT files containing the case samples and `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.
- (2) `GenomicRatioSet` object containing case and control samples.

In the first case, normalization through `epi_preprocess()` function is highly recommended. In order to remove the unwanted variation caused by the batch effect when combining data from different sources. As a result, `GenomicRatioSet` object will be obtained.

Finally, these datasets are used as an input of `epimutations()` function. It should be mentioned that case samples and reference panel are introduced separately.

Candidate regions

Epimutations detection has two main steps: (1) definition of candidate regions and (2) evaluation of outlier significance. Although there are different algorithms to define epimutations regions, they share common features. In general, we define an epimutation as at least 3 contiguous CpGs with a maximum distance of 1kb between them.

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. So, we have computed a dataset containing the regions that are candidates to become an epimutation.

To define the candidate epimutations, we relied on the clustering from bumphunter. 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. As a result, we found 40408 candidate epimutations which are available in `candRegsGR` dataset. The code for generating these regions can be found in `epimutation` package.

In addition, we converted the candidate region from hg19 to hg38 coordinates, using NCBI remap. We selected regions that mapped to one region in hg38 with the same length. This yielded a total of 39944, the 98.85% of total hg19 regions. After converting to hg38, we can use these ranges to be annotated to ENCODE cREs. Overall, we mapped 30163 candidate regions to cREs, representing 74.65% of total candidate regions.

```
data("candRegsGR")
candRegsGR
```

GRanges object with 40408 ranges and 9 metadata columns:

	seqnames	ranges	strand	value	area
	<Rle>	<IRanges>	<Rle>	<numeric>	<numeric>
chr6_32128101	chr6	32128101-32173532	*	1	381
chr6_33156164	chr6	33156164-33181870	*	1	291
chr6_32034322	chr6	32034322-32059605	*	1	239
chr6_31618987	chr6	31618987-31639143	*	1	234
chr6_33279563	chr6	33279563-33292029	*	1	233
...
chr9_140652685	chr9	140652685-140652743	*	1	3
chr9_140656200	chr9	140656200-140657381	*	1	3
chr9_140680393	chr9	140680393-140681206	*	1	3

```

chr9_140732731      chr9 140732731-140733980      * |      1      3
chr9_141012312      chr9 141012312-141013537      * |      1      3

```

	cluster	indexStart	indexEnd	L	clusterL
	<numeric>	<integer>	<integer>	<numeric>	<integer>
chr6_32128101	133070	165174	165554	381	381
chr6_33156164	133204	167451	167741	291	291
chr6_32034322	133058	164512	164750	239	239
chr6_31618987	132987	162583	162816	234	234
chr6_33279563	133221	168282	168514	233	233
...
chr9_140652685	162642	247198	247200	3	3
chr9_140656200	162643	247201	247203	3	3
chr9_140680393	162649	247214	247216	3	3
chr9_140732731	162660	247252	247254	3	3
chr9_141012312	162683	247294	247296	3	3

```


```

	CRE	CRE_type
	<character>	<character>
chr6_32128101	EH38E2459822,EH38E24..	pELS,CTCF-bound;PLS;..
chr6_33156164	EH38E2460436,EH38E24..	PLS;pELS,CTCF-bound;..
chr6_32034322	EH38E2459711,EH38E24..	dELS;dELS;dELS;pELS;..
chr6_31618987	EH38E2459340,EH38E24..	pELS,CTCF-bound;PLS;..
chr6_33279563	EH38E2460551,EH38E24..	pELS,CTCF-bound;pELS;..
...
chr9_140652685		
chr9_140656200	EH38E2738315,EH38E27..	pELS,CTCF-bound;pELS;..
chr9_140680393	EH38E2738332,EH38E27..	dELS,CTCF-bound;dELS
chr9_140732731		
chr9_141012312		

```

-----
seqinfo: 22 sequences from an unspecified genome; no seqlengths

```

Setup

Installing the package

```

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

BiocManager::install("epimutations")

```

Loading libraries

```

library(epimutations)

```

Preparing dataset

The package includes a small `GenomicRatioSet` class dataset (`methy`) containing the DNA methylation profiles from a total of 51 individuals, 3 cases and 48 controls. The DNA methylation profiles were generated using the Illumina 450k Human Methylation BeadChip. The data were obtained from Gene Expression Omnibus (GEO) and adapted for the package usage.

```

data("methy")
methy

```

```

class: GenomicRatioSet
dim: 80731 51
metadata(0):
assays(3): Beta M CN
rownames(80731): cg00725145 cg16080333 ... cg07468397 cg08821909
rowData names(0):
colnames(51): GSM2808239 GSM2808240 ... GSM2562700 GSM2562701
colData names(4): sampleID age sex status
Annotation
  array: IlluminaHumanMethylation450k
  annotation: ilmn12.hg19
Preprocessing
  Method: NA
  minfi version: NA
  Manifest version: NA
table(methy$status)

```

```

  case control
    3       48

```

We are going to create two different datasets for further analysis, `case_samples` and `control_panel`:

```

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

```

Preprocessing

The preprocessing in `epimutations` package is done by `epi_preprocess()` function. It contains 6 preprocessing methods corresponding to minfi package that can be selected by the user:

Method	Function	Description
raw	<code>preprocessRaw</code>	Converts the Red/Green channel for an Illumina methylation array into methylation signal, without using any normalization
illumina	<code>preprocessIllumina</code>	Implements preprocessing for Illumina methylation microarrays as used in Genome Studio
swan	<code>preprocessSWAN</code>	Subset-quantile Within Array Normalisation (SWAN). It allows Infinium I and II type probes on a single array to be normalized together
quantile	<code>preprocessQuantile</code>	Implements stratified quantile normalization preprocessing for Illumina methylation microarrays
noob	<code>preprocessNoob</code>	Noob (normal-exponential out-of-band) is a background correction method with dye-bias normalization for Illumina Infinium methylation arrays
funnorm	<code>preprocessFunnorm</code>	Functional normalization (FunNorm) is a between-array normalization method for the Illumina Infinium HumanMethylation450 platform

In addition, the unique parameters for each normalization approach are defined through `norm_parameters()`:

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

The default settings for each method can be obtained 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, to modify the parameters related to a method you can do as the following example for `illumina` approach:

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

```
[1] FALSE
```

Epimutations

Epimutations detection

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

In the mentioned first 4 methods, firstly, Differentially Methylated Regions (DMRs) are identified using bump-hunter method (Jaffe et al. 2012). Then, those DMRs are tested to identify regions with CpGs being outliers when comparing with the reference panel. However, `barbosa` and `beta` do not identify outliers by filtering the DMRs. `barbosa` utilized a sliding window approach to individually compare the methylation value in each proband against the reference panel. `Beta` used beta distribution to identify epivariants in the case sample.

```
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")
```

```
epi_mcd <- epimutations(case_samples, control_samples, method = "mahdistmcd")
```

```
epi_brbr <- epimutations(case_samples, control_samples, method = "barbosa")
```

```
epi_beta <- epimutations(case_samples, control_samples, method = "beta")
```

Unique parameters

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

Method	Parameter	Description
manova mlm beta	pvalue_cutoff	The threshold p-value to select which CpG regions are outliers
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
barbosa	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

Invoking `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] 0.5
```

```
$isoforest$ntrees
```

```
[1] 100
```

```
$mahdistmcd
```

```
$mahdistmcd$nsamp
```

```
[1] "deterministic"
```

```
$barbosa
```

```
$barbosa$window_sz
```

```
[1] 10
```

```
$barbosa$offset_mean
[1] 0.15
```

```
$barbosa$offset_abs
[1] 0.1
```

```
$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 of p-value cut-off for **manova**:

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

```
[1] 0.01
```

Results description

The **epimutations** function returns a tibble containing all the epivariants identified in the given case sample. In case no epimutation is found, a row containing the case sample information and missing values for each argument is returned. The following table describes each argument in the result data frame:

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_n	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 barbosa and mahdistmcd it is filled with NA.
pvalue	For methods manova and mlm it provides the p-value obtained from the model. For method barbosa , 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 barbosa 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 barbosa , 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.

Column name	Description
CRE_type	Type of cREs (cis-Regulatory Elements) as defined by ENCODE.

Epimutations annotations

The `epimutations` package also includes the `annotate_epimutations` function dedicated to enriching the epimutations identified by the previously described methods:

```
rst_mvo <- annotate_epimutations(emi_mvo)

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

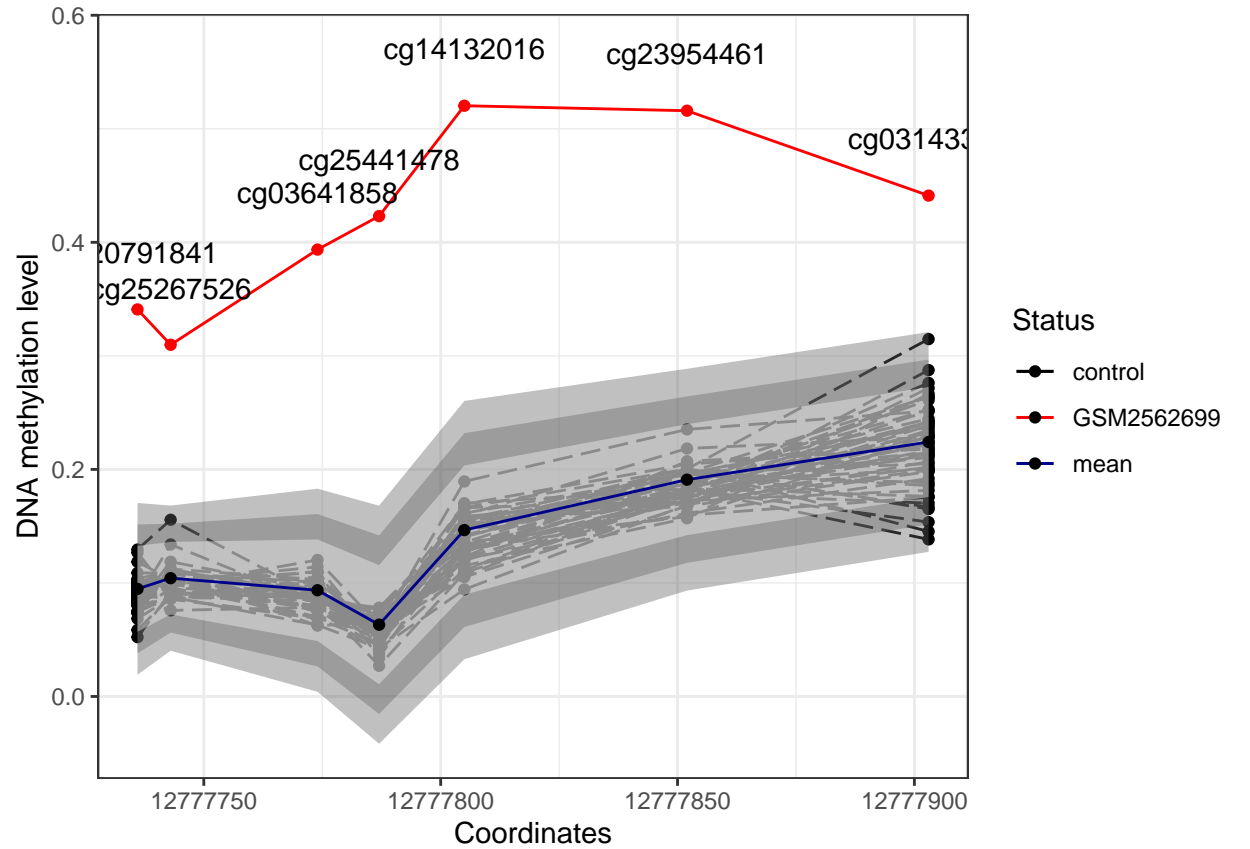
Table 5: epimutations annotation

emi_id	adj_pvalue	emi_region	CRE
emi_manova_1	2	chr19_127777750-127777800	EH38E1939817,EH38E1939818,EH38E1939819
emi_manova_3	3	chr7_9089136-9089146	EH38E2570884,EH38E2570885,EH38E2570886,EH38E2570887,EH38E2570888,EH38E2570889,

Epimutation visualization

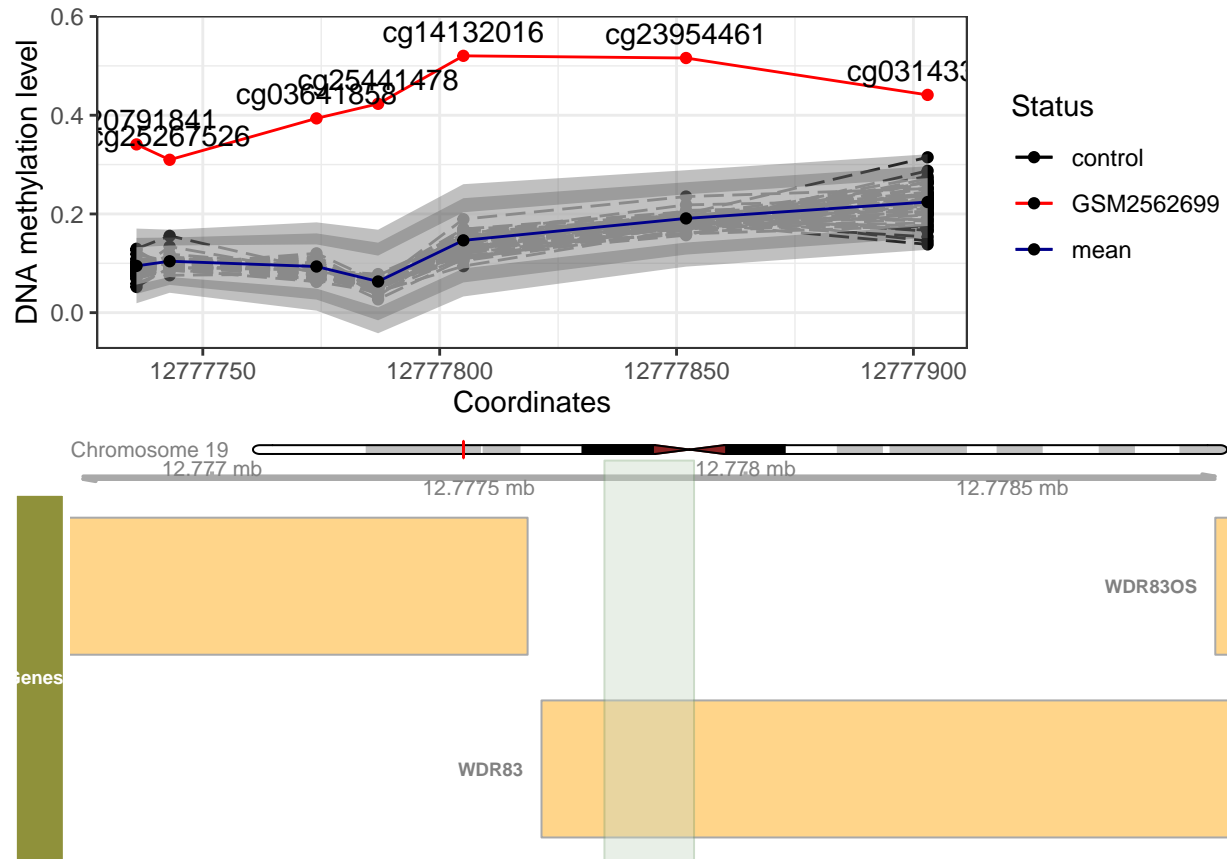
The visualization approach locates the epimutations along the genome. The function `plot_epimutations` plots the methylation values of the individual with the epimutation in red, the control samples in dashed black lines and population mean in blue:

```
plot_epimutations(as.data.frame(emi_mvo[1,]), methy)
```



Furthermore, it includes the gene annotations in the regions in which the epivariation is located. This can be achieved by using the argument `gene_annot == TRUE`:

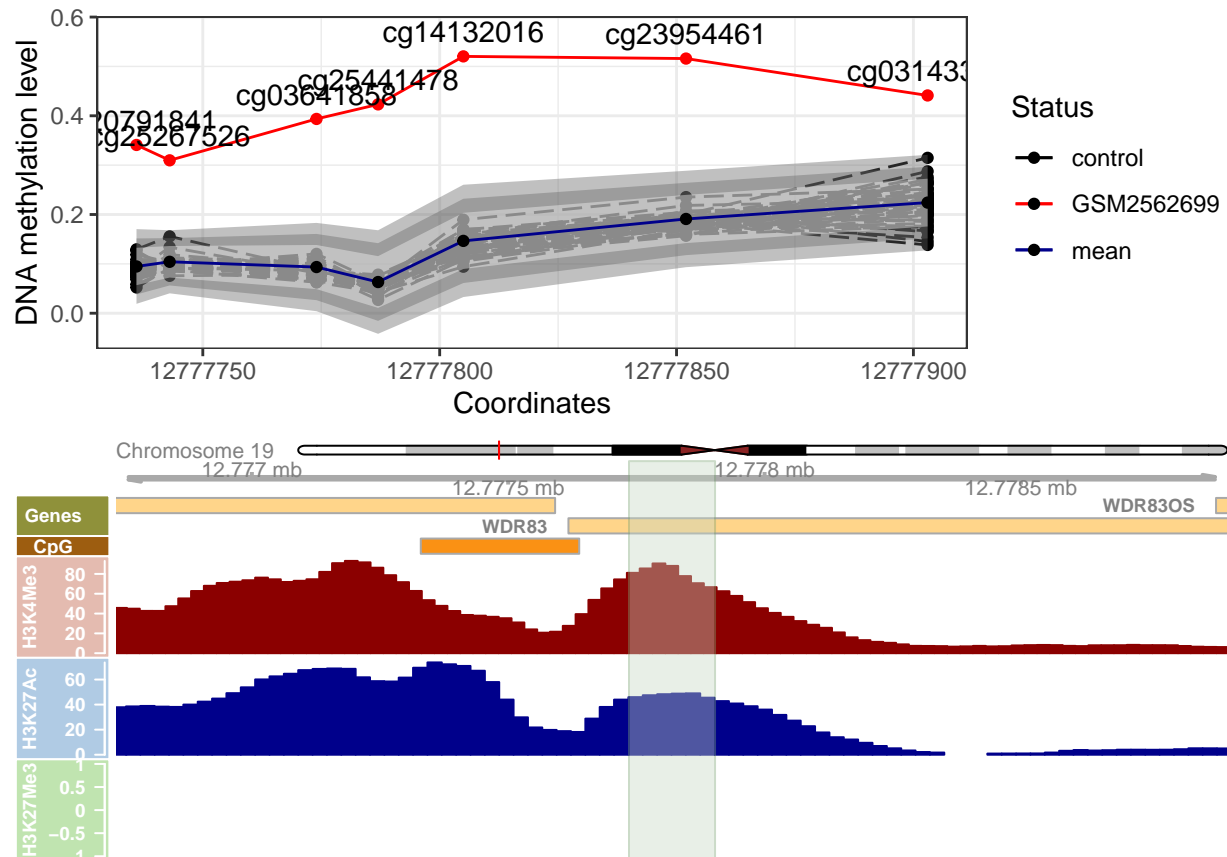
```
plot_epimutations(as.data.frame(epi_mvo[1,]), methy, genes_annot = TRUE)
```



Also, it is possible to plot the chromatin marks H3K4me3, H3K27me3 and H3K27ac by setting the argument `regulation = TRUE`:

- **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

```
plot_epimutations(as.data.frame(epi_mvo[1,]), methy, regulation = TRUE)
```



Method validation

Data collection

The data were obtained for the studies previously described (Garg et al. 2020). The datasets were downloaded from Gene Expression Omnibus (GEO). We accessed DNA methylation data from a total 1, 417 individuals from GSE51032 and GSE111629 cohorts. The DNA methylation profiles were generated using the Illumina 450k Human Methylation BeadChip.

The GSE51032 study analysed primary cancers samples: 424 cancer free, 235 primary breast cancer, 166 primary colorectal cancer and 20 other primary cancers. The GSE111629 cohort 335 Parkinson's disease and 237 control samples.

Validation

We evaluated the performance of the method using TPR (True Positive Rate), False Positive Rate (FPR) and accuracy. We use the TPR to measure the proportion of detected epivariations by the **epimutations** approach present in the validated (Garg et al. 2020). FPR to calculate the identified epimutations outside the once found in (Garg et al. 2020), whether validated or not. The accuracy measures the closeness of the detected epimutation to the validated regions.

We select samples differently depending on the study group and measure to compute. Control samples were selected randomly using different sample size: 20, 30, 40, 50, 60, 70, 80, 90 and 100. However, case samples were selected considering validated epimutations (for TPR and accuracy) or excluding epivariations found (for FPR) (Garg et al. 2020).

The validated epimutations in table 1 were only present on 5 individuals: GSM1235784 from GSE51032

cohort and GSM3035933, GSM3035791, GSM3035807 and GSM3035685 from GSE111629. Therefore, they were established as case samples when computing TPR and accuracy. Nevertheless, we compute FPR excluding the samples containing at least one epimutation found by (Garg et al. 2020). For the remaining case samples, 4 were selected randomly in each execution.

We execute 100 times the same process for each control sample size. We define for the analysis regions of ≈ 20 kb containing ≥ 3 GpGs.

Table 6: 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

Additionally, we have plotted the methylation values of the samples in the regions where the validated epimutations were found.

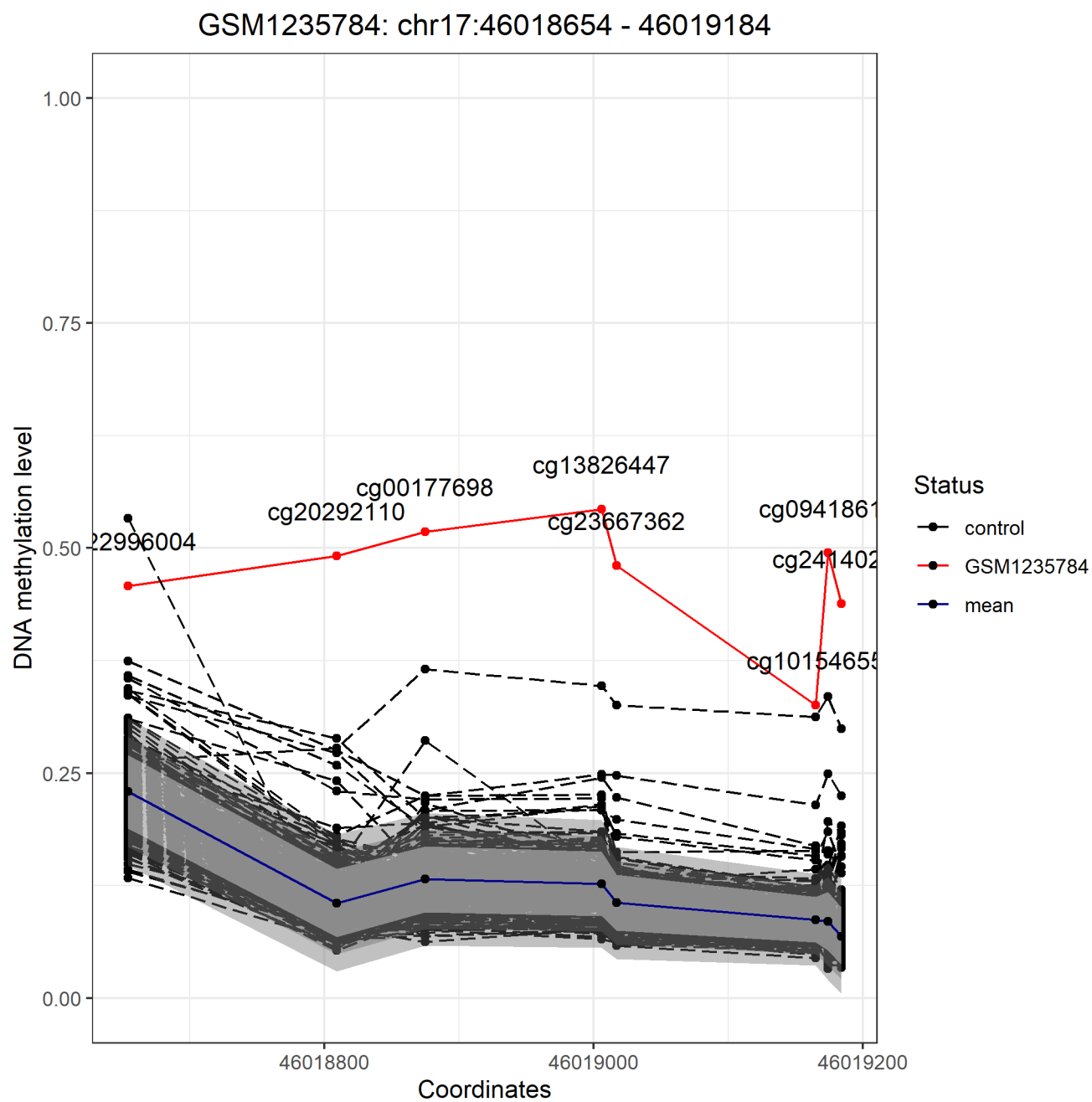


Figure 1: GSE51032 samples in the region chr17:46018654-46019184

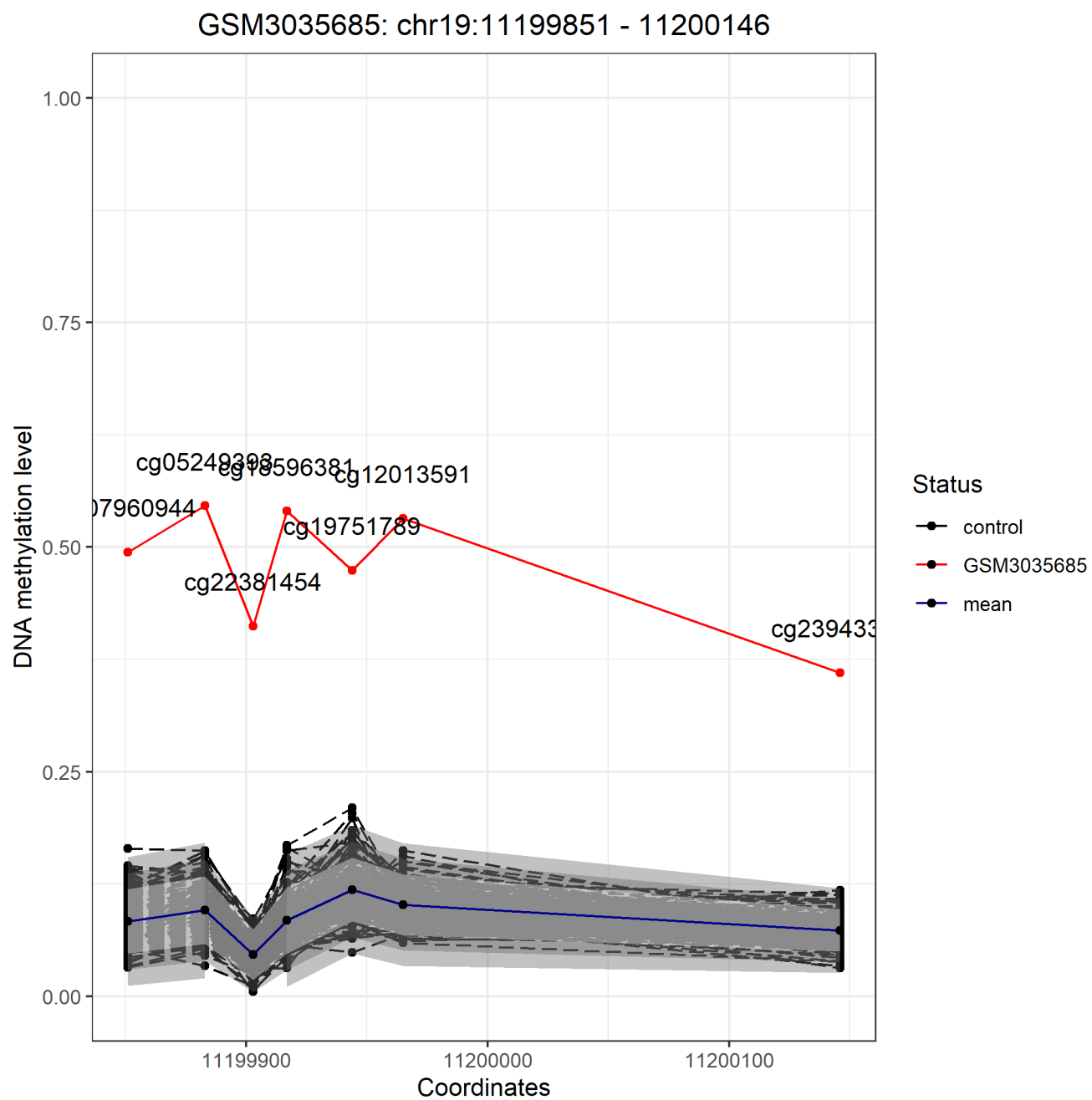


Figure 2: GSE111629 samples in the region chr19:11199851-11200146

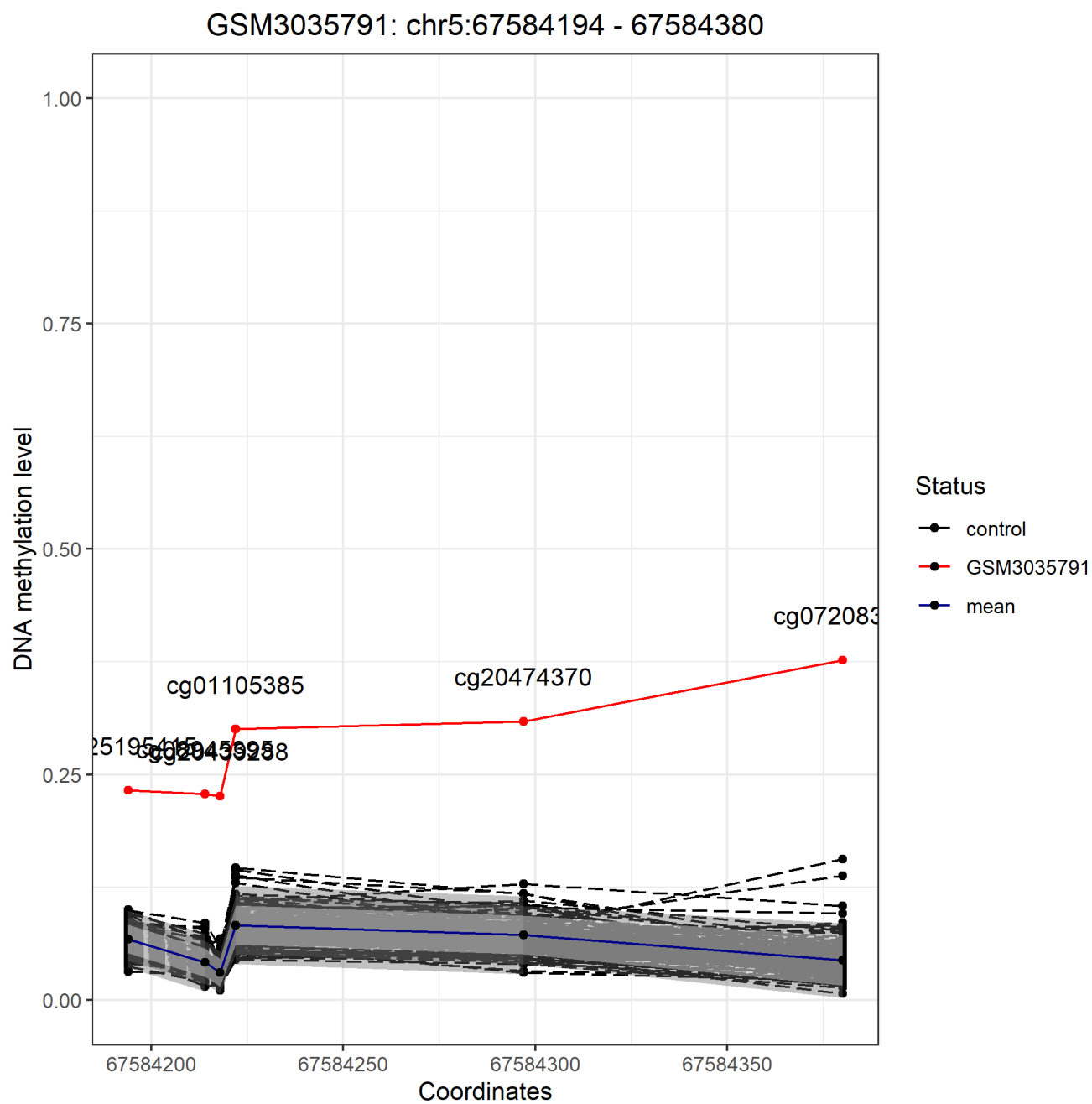


Figure 3: GSE111629 samples in the region chr5:67584194-67584380

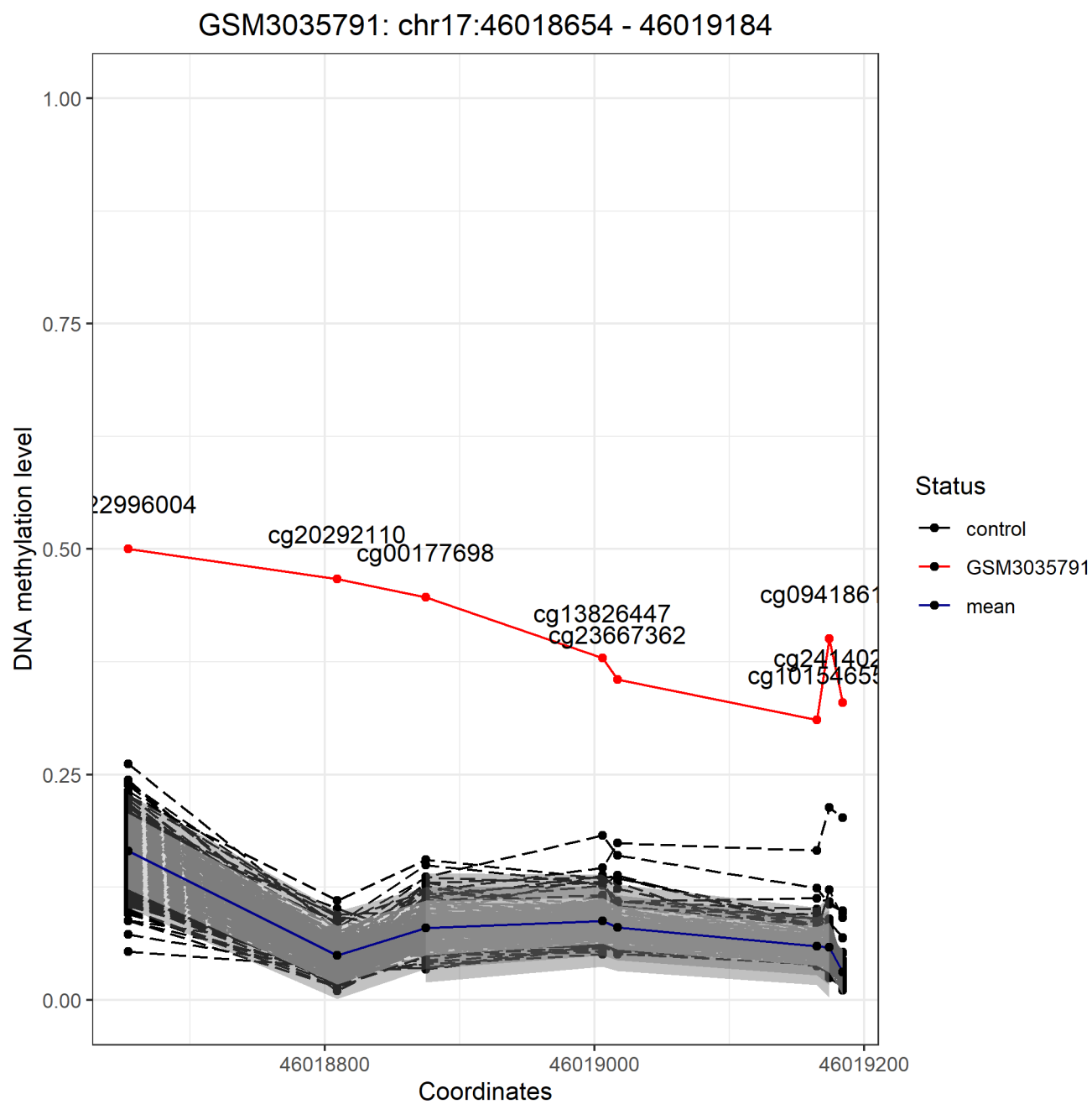


Figure 4: GSE111629 samples in the region chr17:46018654-46019184

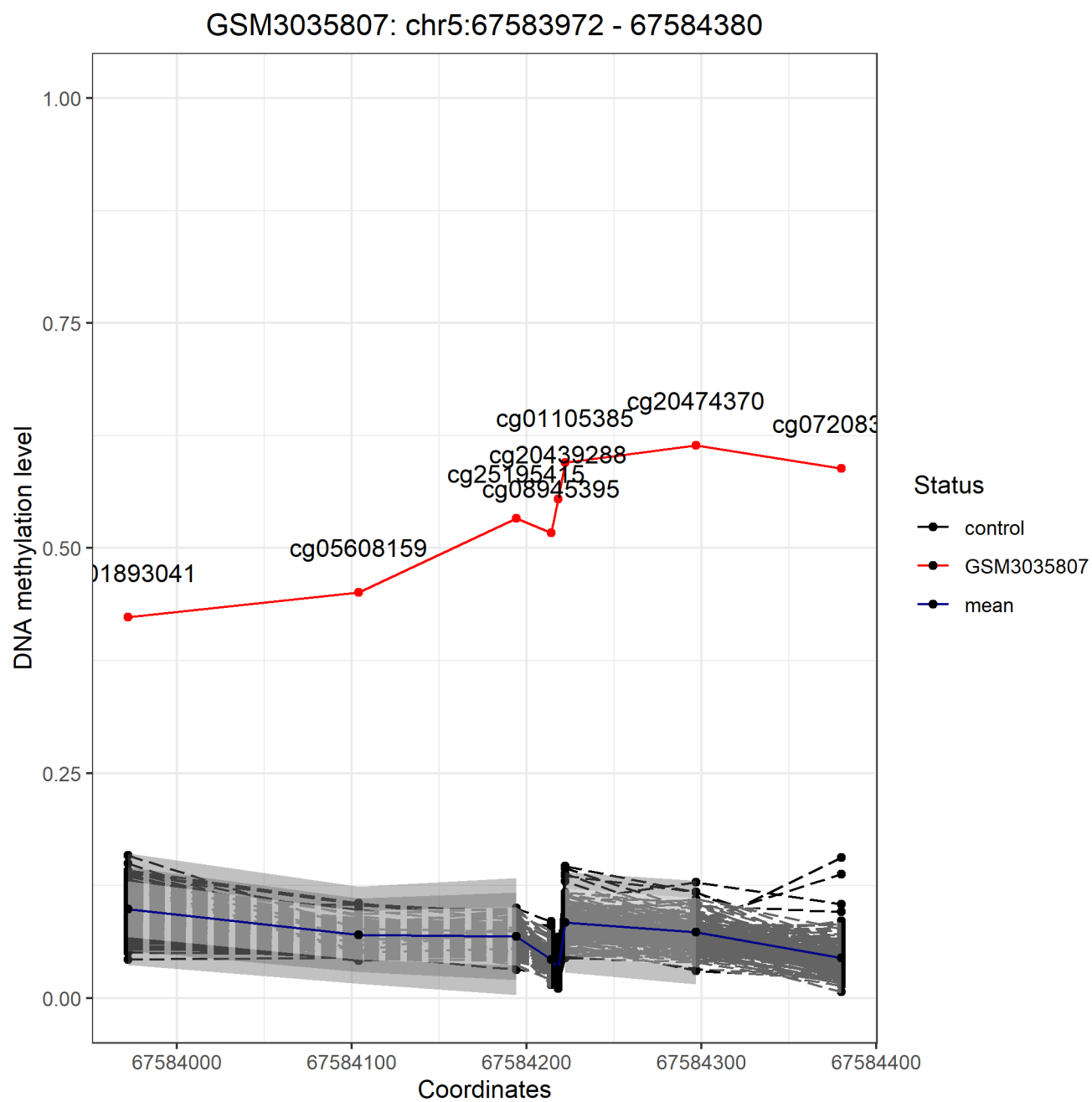


Figure 5: GSE111629 samples in the region chr5:67583972-67584380

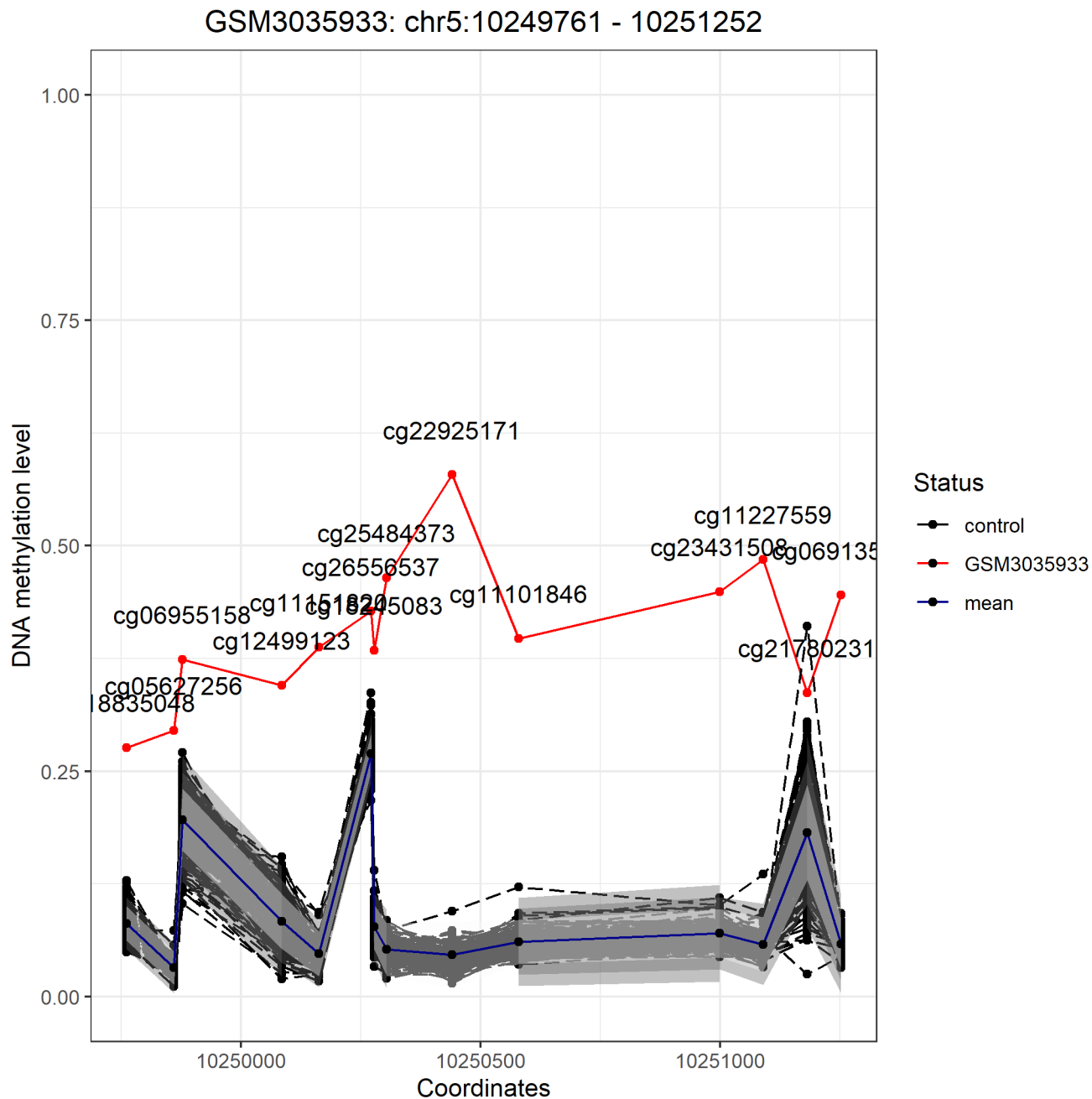


Figure 6: GSE111629 samples in the region chr5:10249761-10251252

Results

We compare GSM1235784 case sample against randomly selected control samples from GSE51032 and GSM3035933, GSM3035791, GSM3035807 and GSM3035685 case samples against controls from GSE111629 specifying a region of 20 kb and ≥ 3 GpGs.

We obtained similar results in both cohorts. We observed that the methods manova, mahalanobis distance and multivariate linear models identified the validated epimutations with a TPR of $> 99\%$ even if the control sample is small. However, the TPR in isolation forest increases together with the number of control samples obtaining a TPR ≥ 75 with 50 control samples or more. The TPR in barbosa and beta approaches for

GSE51032 dataset is small ($< 50\%$). Nonetheless, for GSE111629 the TPR value increases considerably $> 99\%$. Regarding the accuracy, all the statistical approaches detect the epivariants with $> 80\%$ of closeness to the validated epimutations.

We detected possible epivariations outside the epimutations found by (Garg et al. 2020) selecting control and case samples randomly. For the analysis, we selected regions of 20 kb and ≥ 3 GpGs. We compared each case sample individually against control samples. We observed that in both cohorts and for every approach the FPR value is very small $< 0.01\%$.

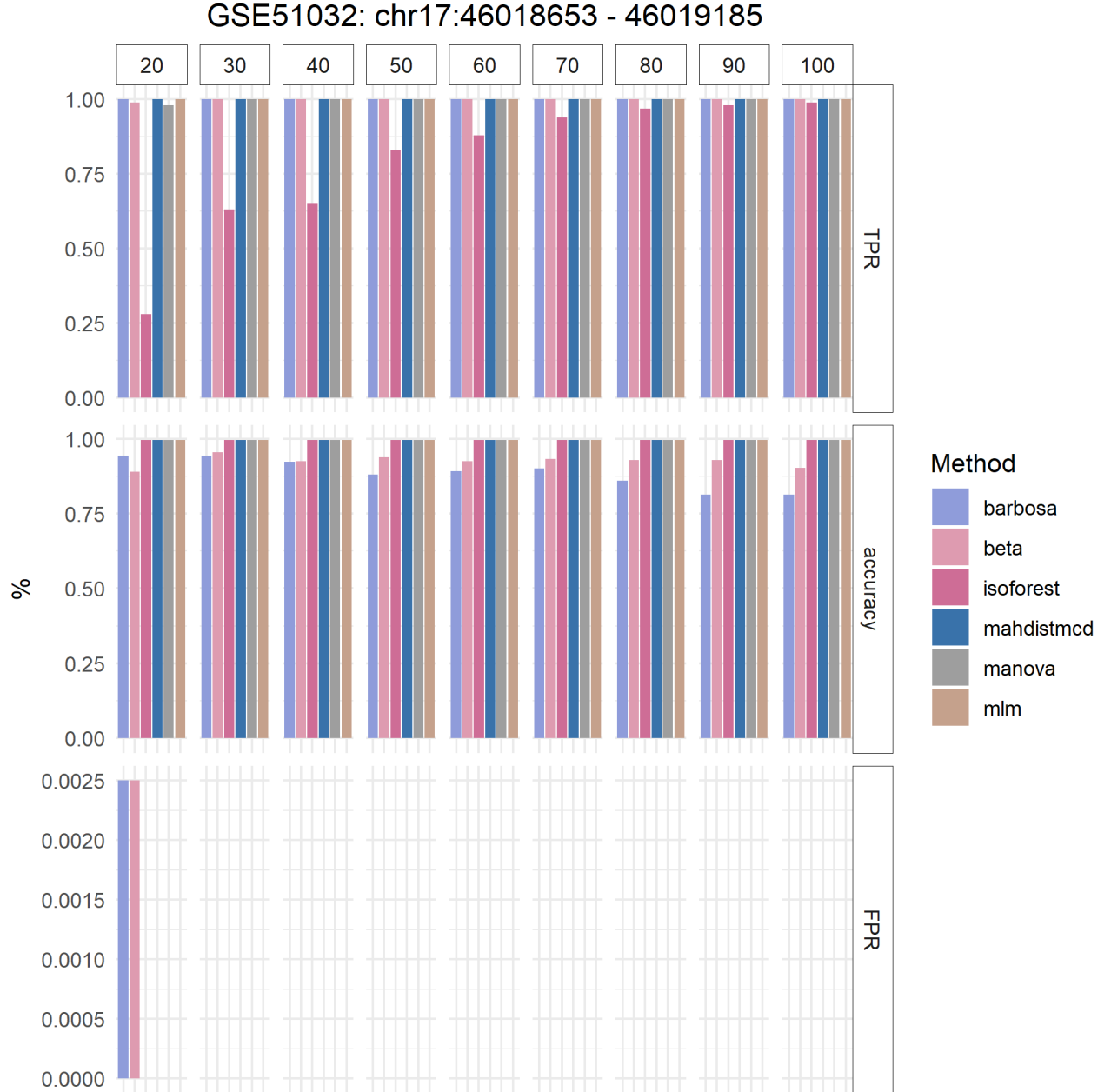


Figure 7: epimutations performance for GSE51032 cohort detecting the epivariation located in chr5:10249760-10251253

method	TPR	accuracy	FPR
n20			
manova	98	99.6	0.00
mlm	100	99.6	0.00
mahdistmcd	100	99.6	0.00
isoforest	28	99.6	0.00
barbosa	100	94.4	0.25
beta	99	89.0	0.25
n30			
manova	100	99.6	0.00
mlm	100	99.6	0.00
mahdistmcd	100	99.6	0.00
isoforest	63	99.6	0.00
barbosa	100	94.4	0.00
beta	100	95.4	0.00
n40			
manova	100	99.6	0.00
mlm	100	99.6	0.00
mahdistmcd	100	99.6	0.00
isoforest	65	99.6	0.00
barbosa	100	92.4	0.00
beta	100	92.6	0.00
n50			
manova	100	99.6	0.00
mlm	100	99.6	0.00
mahdistmcd	100	99.6	0.00
isoforest	83	99.6	0.00
barbosa	100	88.0	0.00
beta	100	93.8	0.00
n60			
manova	100	99.6	0.00
mlm	100	99.6	0.00
mahdistmcd	100	99.6	0.00
isoforest	88	99.6	0.00
barbosa	100	89.2	0.00
beta	100	92.6	0.00
n70			
manova	100	99.6	0.00
mlm	100	99.6	0.00
mahdistmcd	100	99.6	0.00
isoforest	94	99.6	0.00
barbosa	100	90.0	0.00
beta	100	93.2	0.00
n80			
manova	100	99.6	0.00
mlm	100	99.6	0.00
mahdistmcd	100	99.6	0.00
isoforest	97	99.6	0.00
barbosa	100	86.0	0.00
beta	100	92.9	0.00
n90			
manova	100	99.6	0.00
mlm	100	99.6	0.00
mahdistmcd	100	99.6	0.00
isoforest	99	99.6	0.00
barbosa	100	81.3	0.00
beta	100	92.9	0.00
n100			
manova	100	99.6	0.00

method	n	TPR	accuracy	FPR
n20				
barbosa	100	100.000	92.825	0.00
beta	100	100.000	92.825	0.25
isoforest	100	98.625	93.000	0.00
mahdistmcd	100	100.000	92.825	0.00
manova	100	100.000	92.825	0.00
mlm	100	100.000	92.825	0.00
n30				
barbosa	20	100.000	92.825	0.00
beta	20	87.500	92.800	0.00
isoforest	20	11.500	86.800	0.00
mahdistmcd	20	100.000	92.825	0.00
manova	20	98.000	92.825	0.00
mlm	20	100.000	92.850	0.00
n40				
barbosa	30	100.000	92.825	0.25
beta	30	87.500	92.775	0.25
isoforest	30	28.000	93.150	0.00
mahdistmcd	30	100.000	92.825	0.00
manova	30	100.000	92.825	0.00
mlm	30	100.000	92.825	0.00
n50				
barbosa	40	100.000	92.825	0.00
beta	40	87.500	92.800	0.00
isoforest	40	46.375	93.950	0.00
mahdistmcd	40	100.000	92.825	0.00
manova	40	100.000	92.825	0.00
mlm	40	100.000	92.825	0.00
n60				
barbosa	50	100.000	92.825	0.00
beta	50	87.500	92.825	0.00
isoforest	50	70.125	93.500	0.00
mahdistmcd	50	100.000	92.825	0.00
manova	50	100.000	92.825	0.00
mlm	50	100.000	92.825	0.00
n70				
barbosa	60	100.000	92.825	0.00
beta	60	100.000	92.825	0.00
isoforest	60	78.750	93.500	0.00
mahdistmcd	60	100.000	92.825	0.00
manova	60	100.000	92.825	0.00
mlm	60	100.000	92.825	0.00
n80				
barbosa	70	100.000	92.825	0.25
beta	70	100.000	92.825	0.00
isoforest	70	90.125	93.575	0.00
mahdistmcd	70	100.000	92.825	0.00
manova	70	100.000	92.825	0.00
mlm	70	100.000	92.825	0.00
n90				
barbosa	80	100.000	92.825	0.00
beta	80	100.000	92.825	0.00
isoforest	80	96.375	93.075	0.00
mahdistmcd	80	100.000	92.825	0.00
manova	80	100.000	92.825	0.00
mlm	80	100.000	92.825	0.00
n100				
barbosa	90	100.000	92.825	0.00

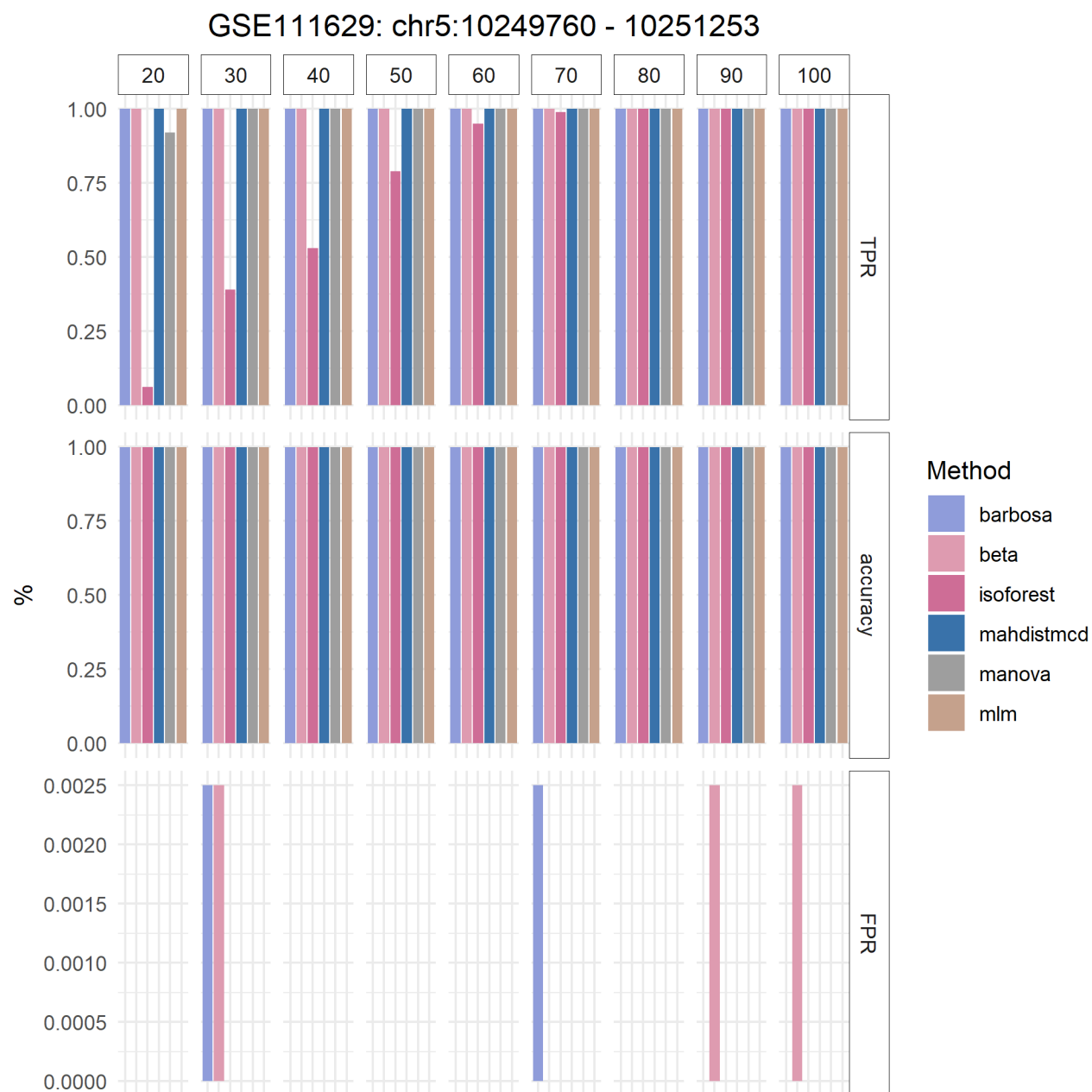


Figure 8: epimutations performance using GSE111629 cohort to detect the epivariation located in chr5:10249760-10251253

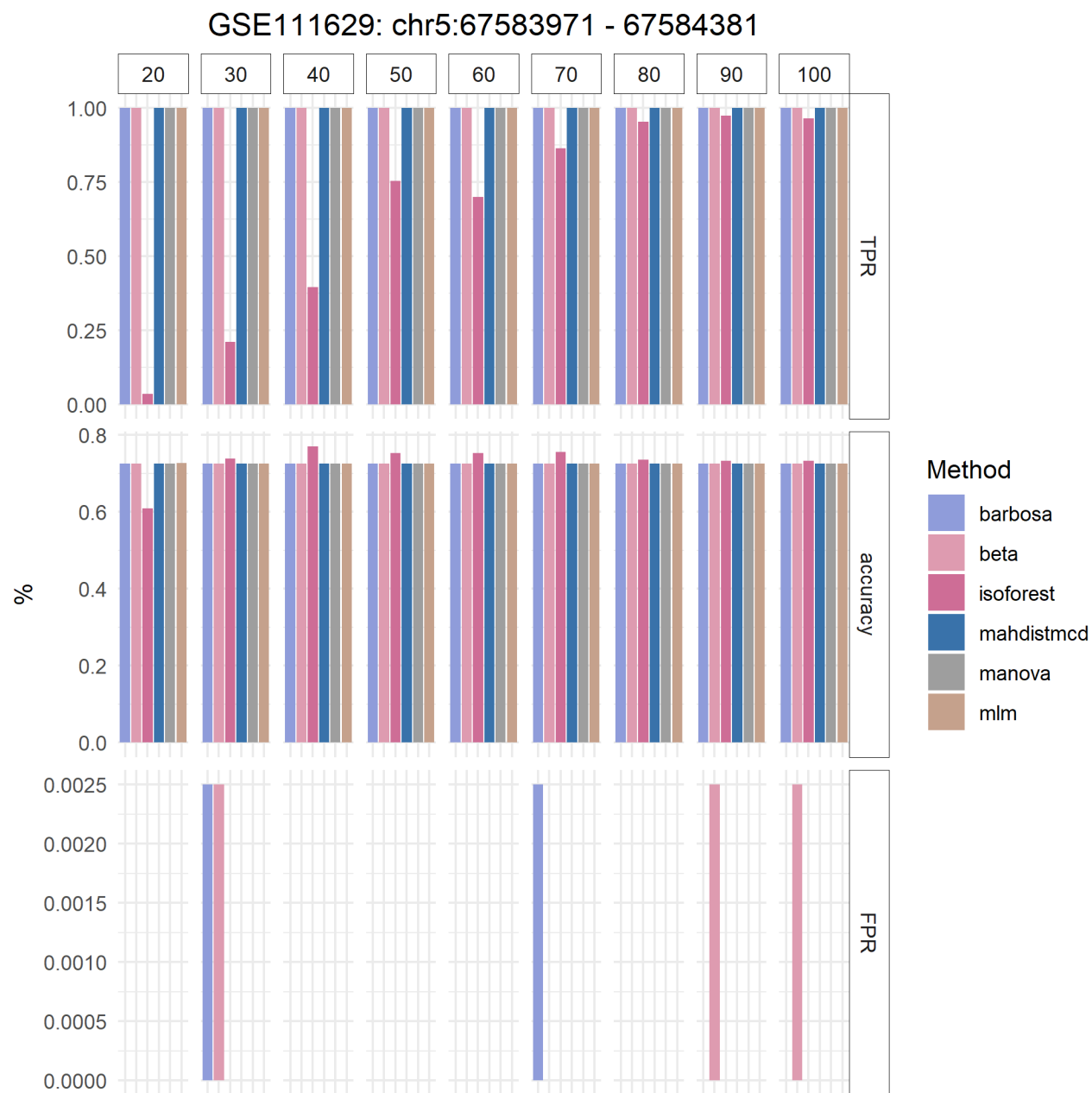


Figure 9: epimutations performance using GSE111629 cohort to detect the epivariation located in chr5:67583971-67584381

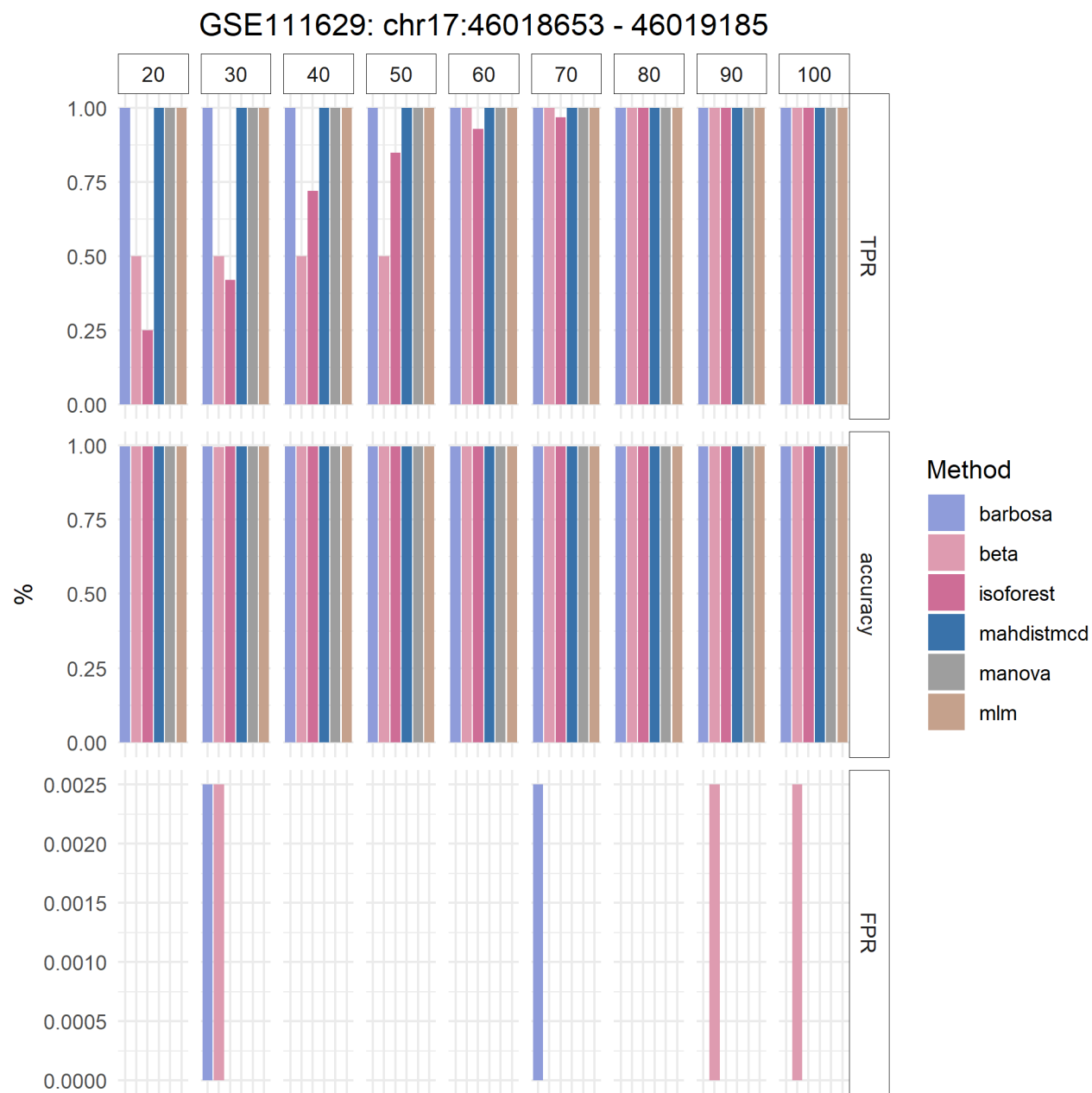


Figure 10: epimutations performance using GSE111629 cohort to detect the epivariation located in chr17:46018653-46019185

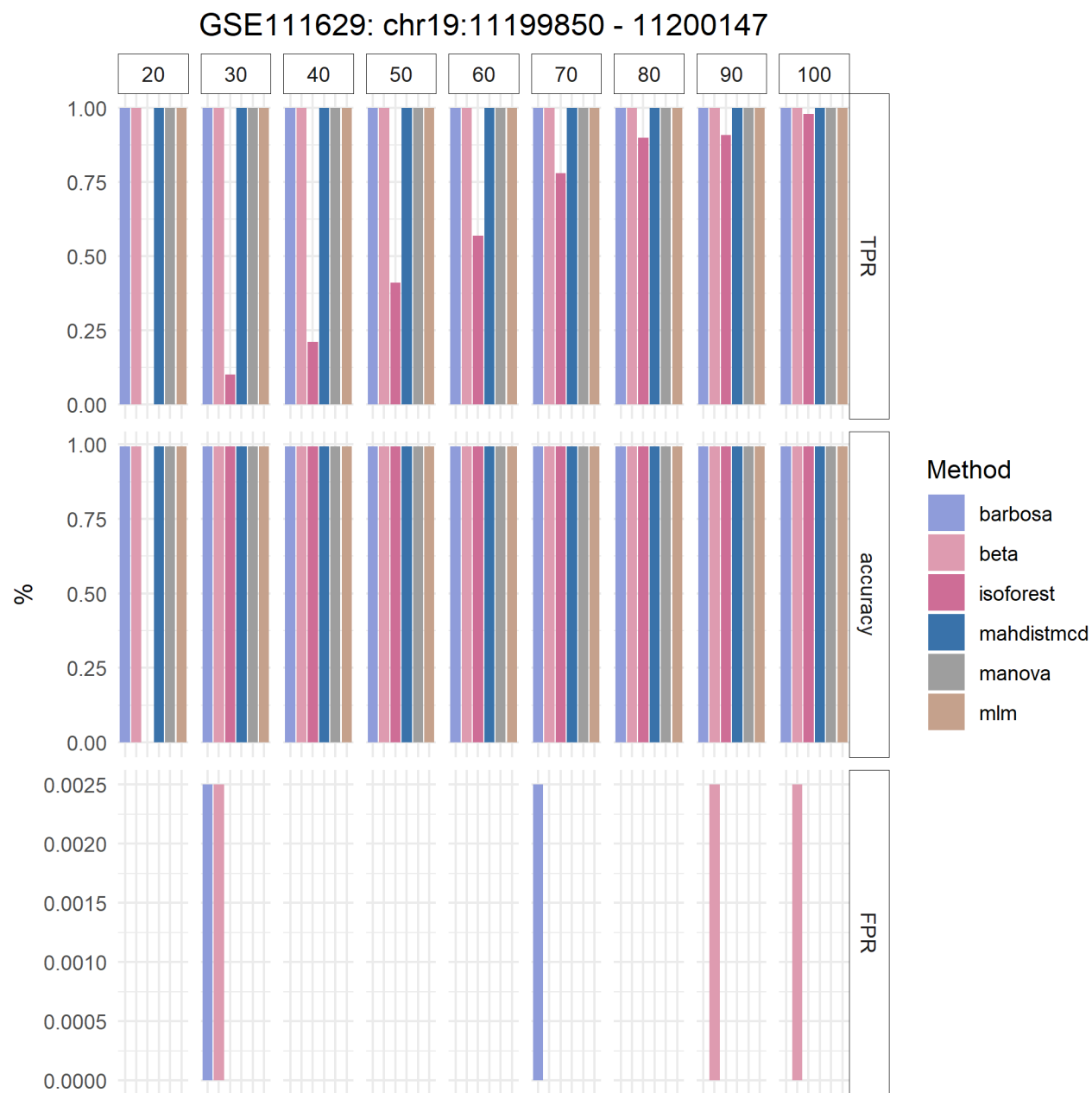


Figure 11: epimutations performance using GSE111629 cohort to detect the epivariation located in chr5:11199850-11200147

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
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

Name	Surname	ORCID	Affiliation	Team
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

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>.
- Cortes, David, and Maintainer David Cortes. 2021. “Package ‘Isotree’.”
- Friedrich, Sarah, Frank Konietzschke, Markus Pauly, and Maintainer Sarah Friedrich. 2017. “Package ‘MANOVA.RM’.”
- 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.
- Jaffe, Andrew E, Peter Murakami, Hwajin Lee, Jeffrey T Leek, M Daniele Fallin, Andrew P Feinberg, and Rafael A Irizarry. 2012. “Bump Hunting to Identify Differentially Methylated Regions in Epigenetic Epidemiology Studies.” *International Journal of Epidemiology* 41 (1): 200–209.
- 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.
- M, Barbosa, Joshi RS, Garg R, Martin-Trujillo A, Patel N, Jadhav B, Watson CT, et al. 2018. “Identification of Rare de Novo Epigenetic Variations in Congenital Disorders.” *Nature Communications*. <https://doi.org/https://doi.org/10.1038/s41467-018-04540-x>.
- 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.
- Serra-Juhé, Clara, Ivon Cuscó, Aida Homs, Raquel Flores, Núria Torán, and Luis A Pérez-Jurado. 2015. “DNA Methylation Abnormalities in Congenital Heart Disease.” *Epigenetics* 10 (2): 167–77.