epimutacions: 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

2021-07-15

Contents

1 Introduction

1.1 Background

According to the European Commission, rare diseases are pathologies with a prevalence of less than 1 person in 2,000 people (European-Commission 2020). Most of these conditions have an onset during childhood and a strong genetic etiology (López-Bastida et al. 2016). Consequently, rare disease diagnosis has relied on identifying genetic and genomic mutations that can cause the disease (Aref-Eshghi et al. 2019). 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) (Aref-Eshghi et al. 2019). 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; Garg et al. 2020). Therefore, tools to detect epimutations and episignatures should be made available to the rare disease community and included in standardized analysis workflows.

This manual describes the tools available in `epimutacions` package to identify epivariants using multiple outlier detection approaches. It also includes functions to plot and annotate the 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.

1.2 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). It 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). Then, 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, quantile (Garg et al. 2020) and Beta methods do not use bump-hunter output. Quantile (Garg et al. 2020) 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.

1.3 Input data

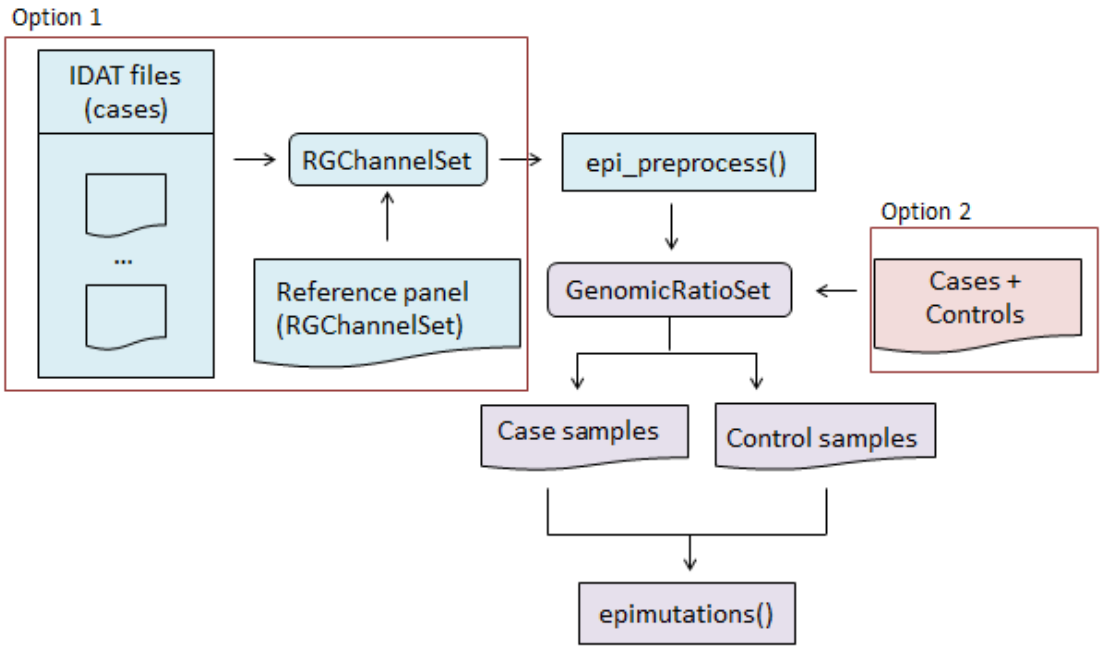
The package allows two different types of inputs:

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

The input data should contain information about β values of CpG sites, phenotype and feature data.

If you want to combine data from different studies, normalization is highly recommended. The preprocessing step removes the unwanted variation caused by the batch effect. In addition, it converts raw microarray intensities to `GenomicRatioSet` (`epimutations()` function input). Finally, the case and control samples are introduced separately in `epimutations()` function.

Figure 1: Allowed data formats, normalization and epimutations() function input



2 Getting started

The `epimutacions` package is available on GitHub and can be installed by executing the following command:

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

The package is loaded in R as usual:

```
library(epimutacions)
```

The document has the following dependencies:

```
library(Knitr)
library(kableExtra)
```

3 Datasets

3.1 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 (Aref-Eshghi et al. 2019; Garg et al. 2020).

In Illumina 450K array (Reproducibility 2012), 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 (Jaffe et al. 2012). 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.

Besides, we converted the candidate region from hg19 to hg38 coordinates, using NCBI remap (Holmes et al. 2020). 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 (Hon and Carninci 2020). Overall, we mapped 30163 candidate regions to cREs, representing 74.65% of total candidate regions.

```
#Candidate regions dataset
##load data
data("candRegsGR")
##data class
class(candRegsGR)
```

```
[1] "GRanges"
attr(,"package")
[1] "GenomicRanges"
```

```
##dataset
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
```

3.2 GenomicRatioSet

The package includes a small `GenomicRatioSet` class dataset (`methy`) containing the DNA methylation profiles from a total of individuals, 3 cases and 48 controls. The DNA methylation profiles were generated using the Illumina 450k Human Methylation BeadChip (Reproducibility 2012). The data corresponds to GSE104812 cohort from Gene Expression Omnibus (GEO). It has been adapted for 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
```

3.3 IDAT files and RGChannelSet

Additionally, as external data 4 case samples IDAT files from GSE131350 cohort are included.

```
baseDir <- system.file("extdata", package = "epimutacions")
list.files(baseDir)
```

```
[1] "GSM3770871_R04C01_Grn.idat" "GSM3770871_R04C01_Red.idat"
[3] "GSM3770882_R05C01_Grn.idat" "GSM3770882_R05C01_Red.idat"
[5] "GSM3770883_R08C01_Grn.idat" "GSM3770883_R08C01_Red.idat"
[7] "GSM3770899_R02C01_Grn.idat" "GSM3770899_R02C01_Red.idat"
[9] "SampleSheet.csv"
```

```
targets <- minfi::read.metharray.sheet(baseDir)
```

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

The reference panel (`RGChannelSet` object) contains DNA methylation profiles of 22 whole cord blood samples from healthy children born via caesarian (GEO: GSE127824). This dataset is available in the `ExperimentHub` package.

4 Preprocessing

The normalization removes the unwanted variation caused by the batch effect when combining data from different sources. The preprocessing in `epimutacions` package is done by `epi_preprocess()` function. It contains 6 preprocessing methods corresponding to minfi package (Aryee et al. 2014), which can be selected by the user:

Table 1: 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.
`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.

Method	Function	Description
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.

Those methods have unique parameters (table 2) that can be defined through `norm_parameters()` function:

Table 2: 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

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
```

5 Epimutations

5.1 Epimutations detection

The `epimutacions` 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) `quantile` 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, `quantile` and `beta` do not identify outliers by filtering the DMRs. `quantile` 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.

To ilustrate multiple examples we are going to use `methy` dataset. Fristly, we are going to split the dataset in two subset: (1) cases (`case_samples`) and (2) controls (`control_samples`):

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

Then, we are going to identify epimutations in each case sample using one statistical approach at a 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")
epi_mcd <- epimutations(case_samples, control_samples, method = "mahdistmcd")
epi_qtl <- epimutations(case_samples, control_samples, method = "quantile")
epi_beta <- epimutations(case_samples, control_samples, method = "beta")
```

5.2 Unique parameters

The `epi_parameters()` function is useful to set the individual parameters for each approach. The arguments are described in table 3:

Table 3: 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

Method	Parameter	Description
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
```

5.3 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. Table 4 describes each argument in the result data frame:

Table 4: 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_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 `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.

5.4 Epimutations annotations

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

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

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

Table 5: epimutations annotation

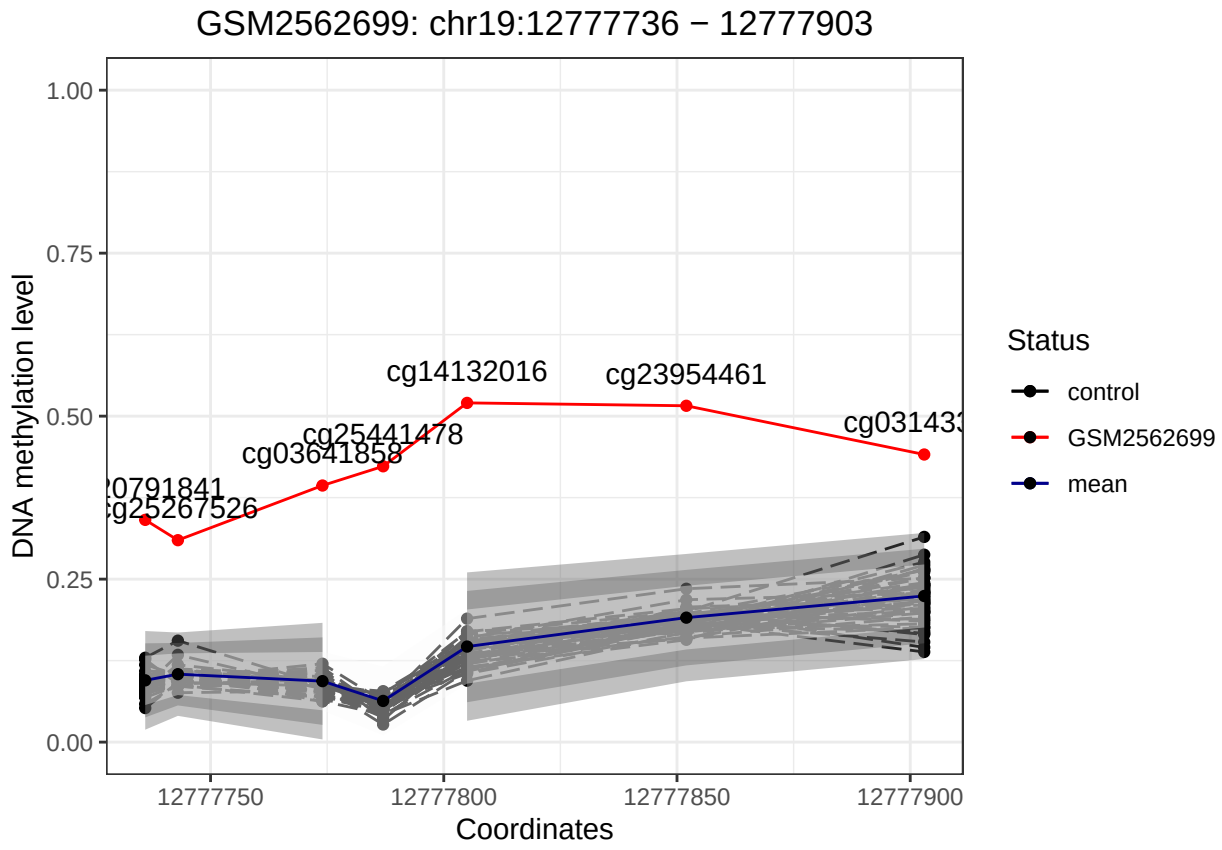
	epi_id	adj_pvalue	epi_region_id	CRE
1	epi_manova_1	0.0000000	chr19_12776725	EH38E1939817,EH38E1939818,EH38E1939819
29	epi_manova_47	0.0116991	chr7_73894573	EH38E2563868,EH38E2563869

5.5 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 (figure 2):

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

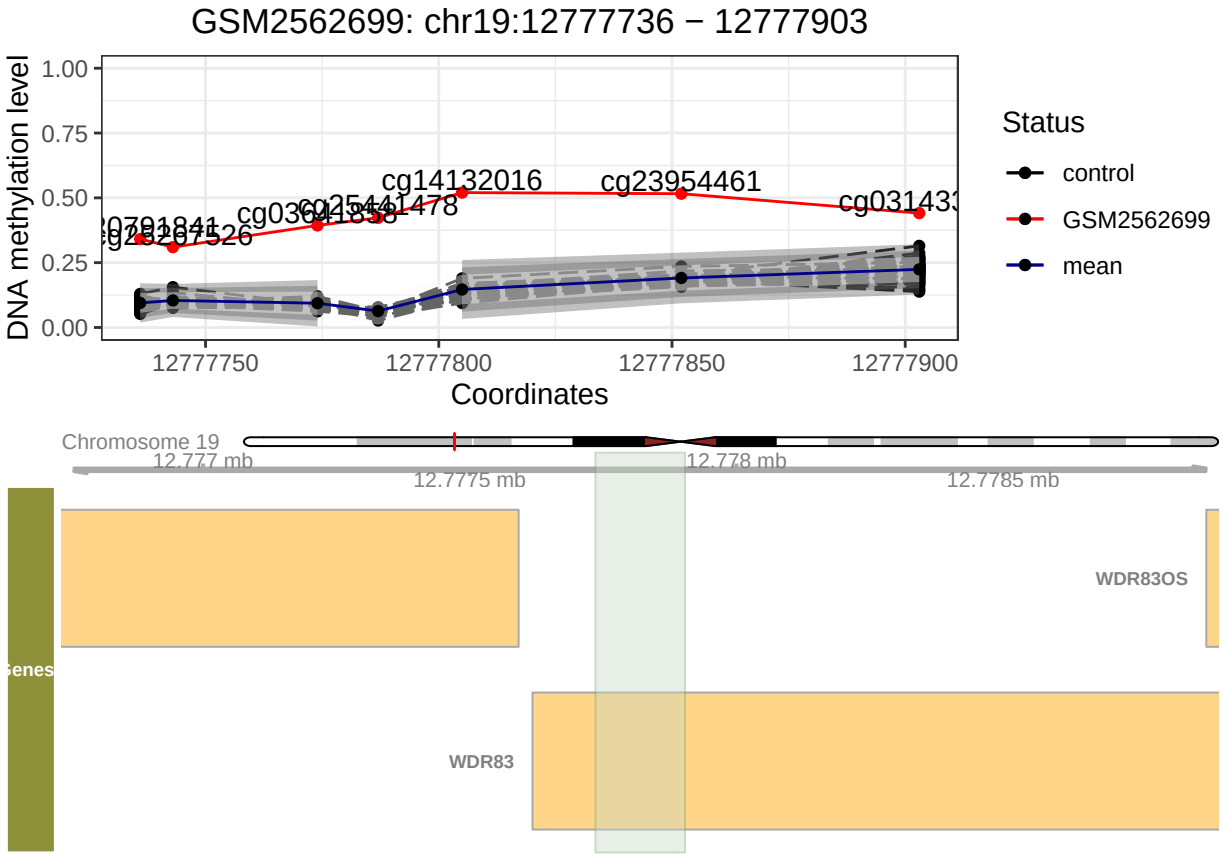
Figure 2: beta values of case sample againts control samples



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` (figure 3):

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

Figure 3: beta values of case sample againts control samples including UCSC gene annotation



Also, it is possible to plot (figure 4) CpG islands and 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)
```

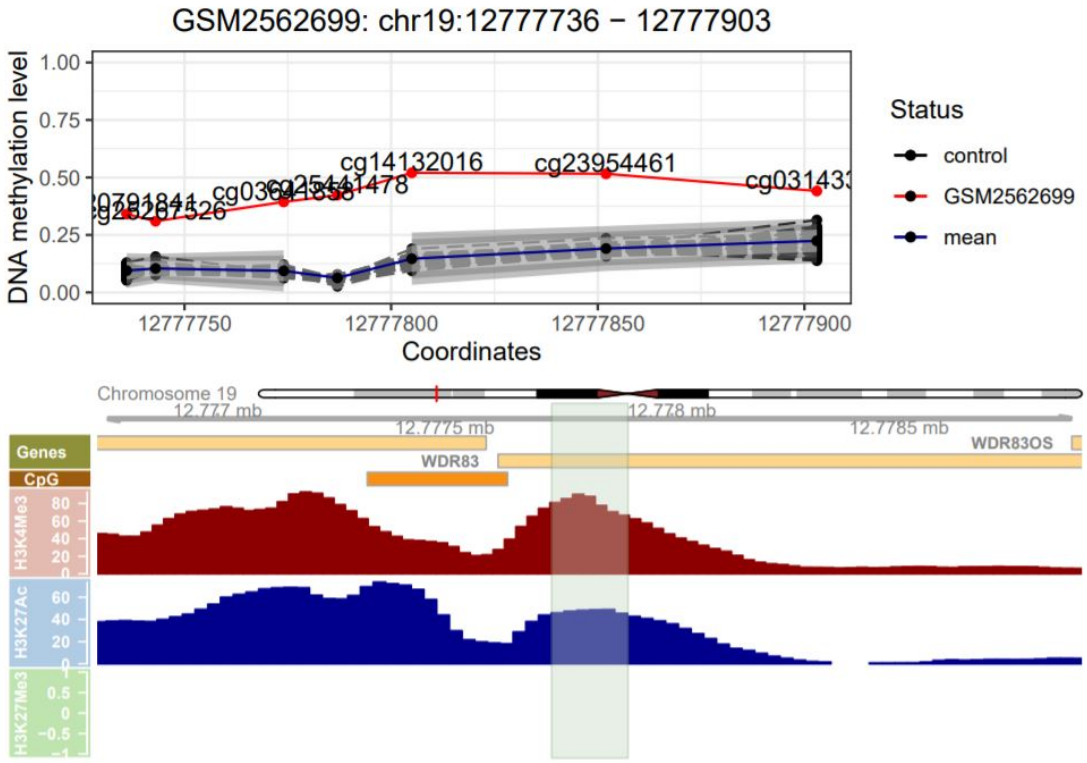
6 Method validation

We validate our method (`epimutacions`) reproducing the results in the study (Garg et al. 2020).

6.1 Data collection

The data were obtained from the study 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

Figure 4: beta values of case sample againts control samples including CpG islands



individuals from GSE51032 and GSE111629 cohorts. The DNA methylation profiles were generated using the Illumina 450k Human Methylation BeadChip.

The GSE51032 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.

6.2 Validation

We evaluated the performance of the method using True Positive Rate (TPR), False Positive Rate (FPR) and accuracy. We used the TPR to measure the proportion of detected epivariations by the `epimutations` approach present in the validated (table 6) by (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 6 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 executed 100 times the same process for each control sample size. We defined for the analysis regions of $\approx$ 20 kb containing $\geq$ 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 5: GSE51032 cohort samples in the region chr17:46018654-46019184

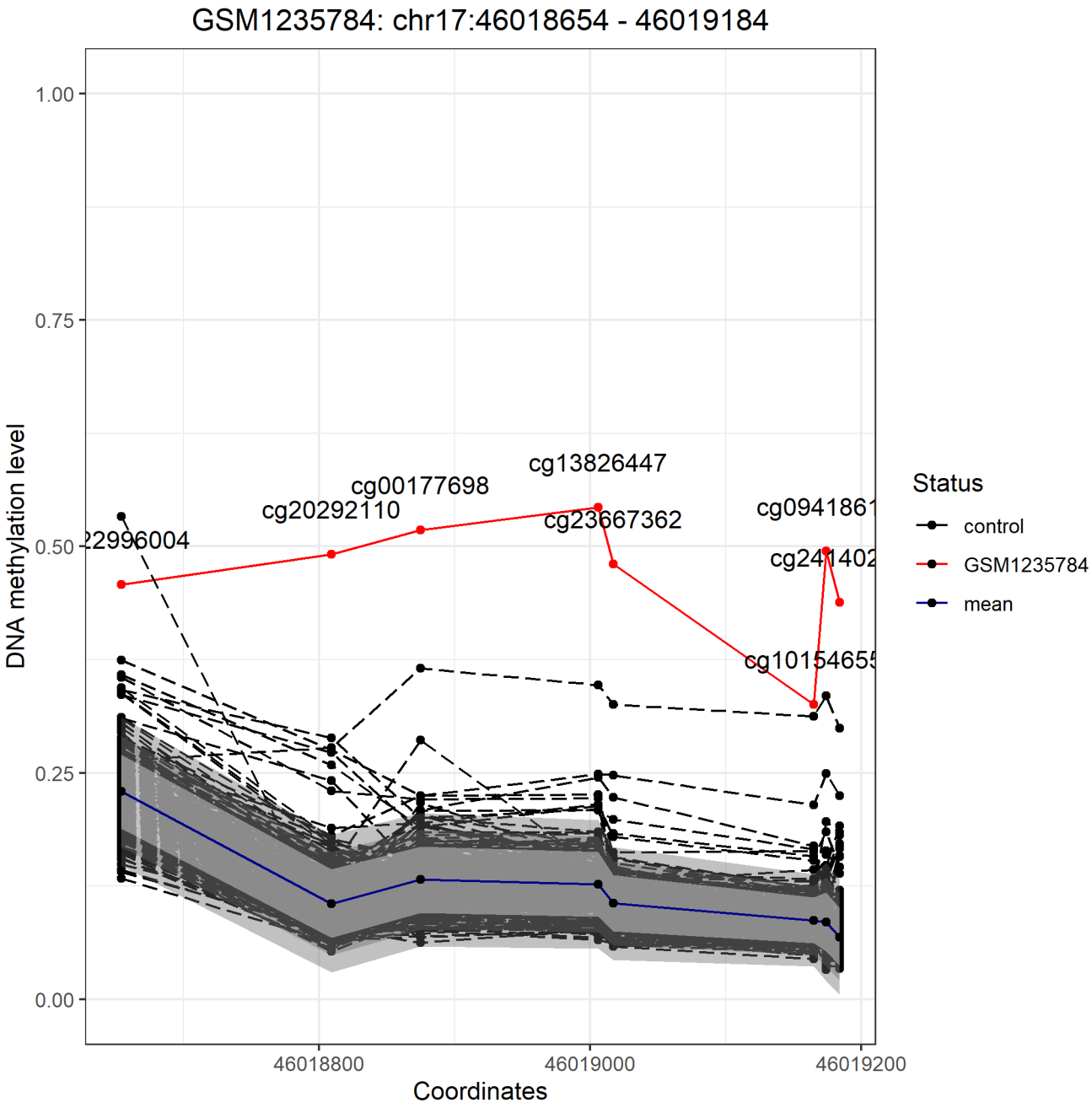


Figure 6: GSE111629 cohort samples in the region chr19:11199851-11200146

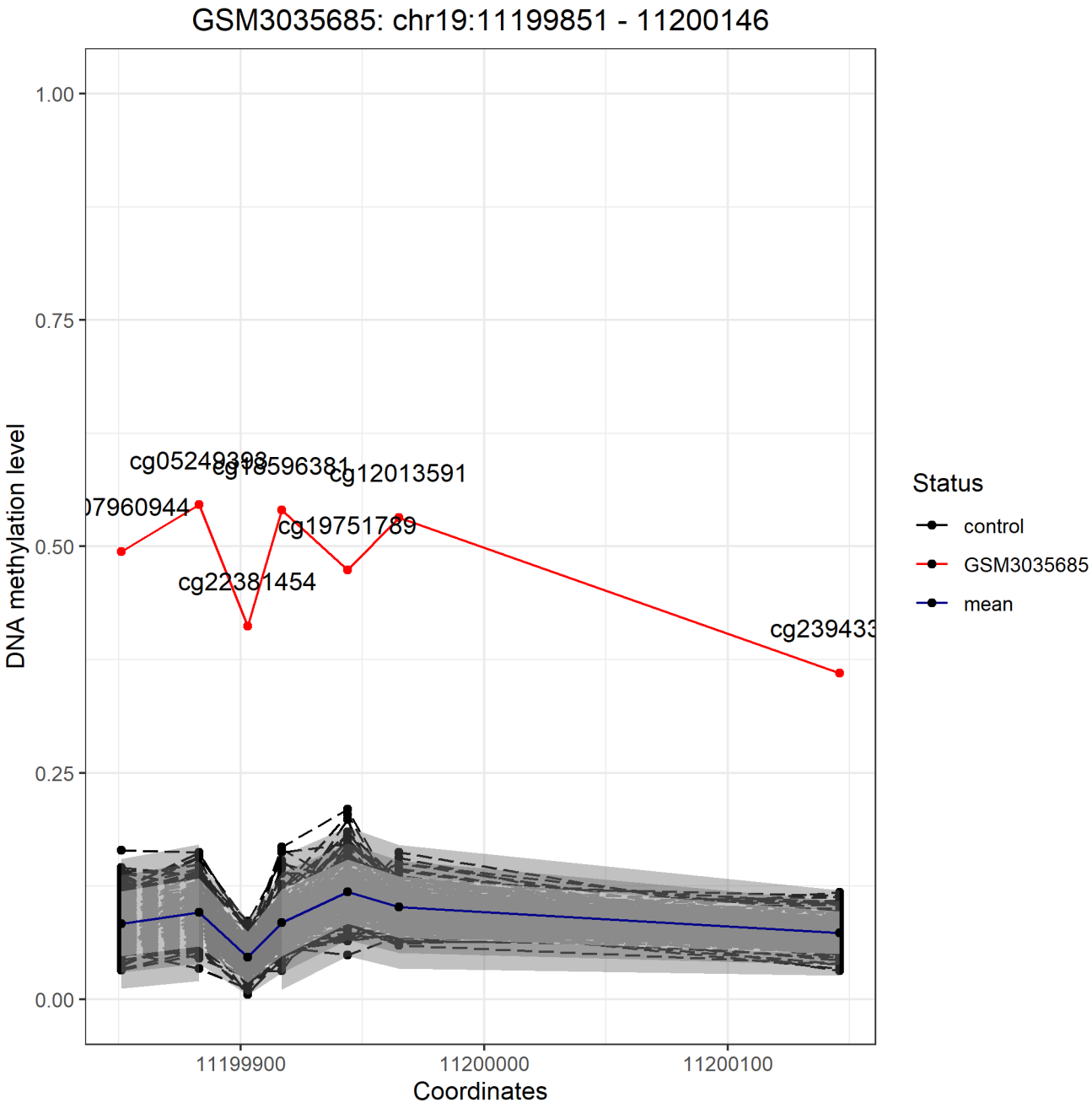


Figure 7: GSE111629 cohort samples in the region chr5:67584194-67584380

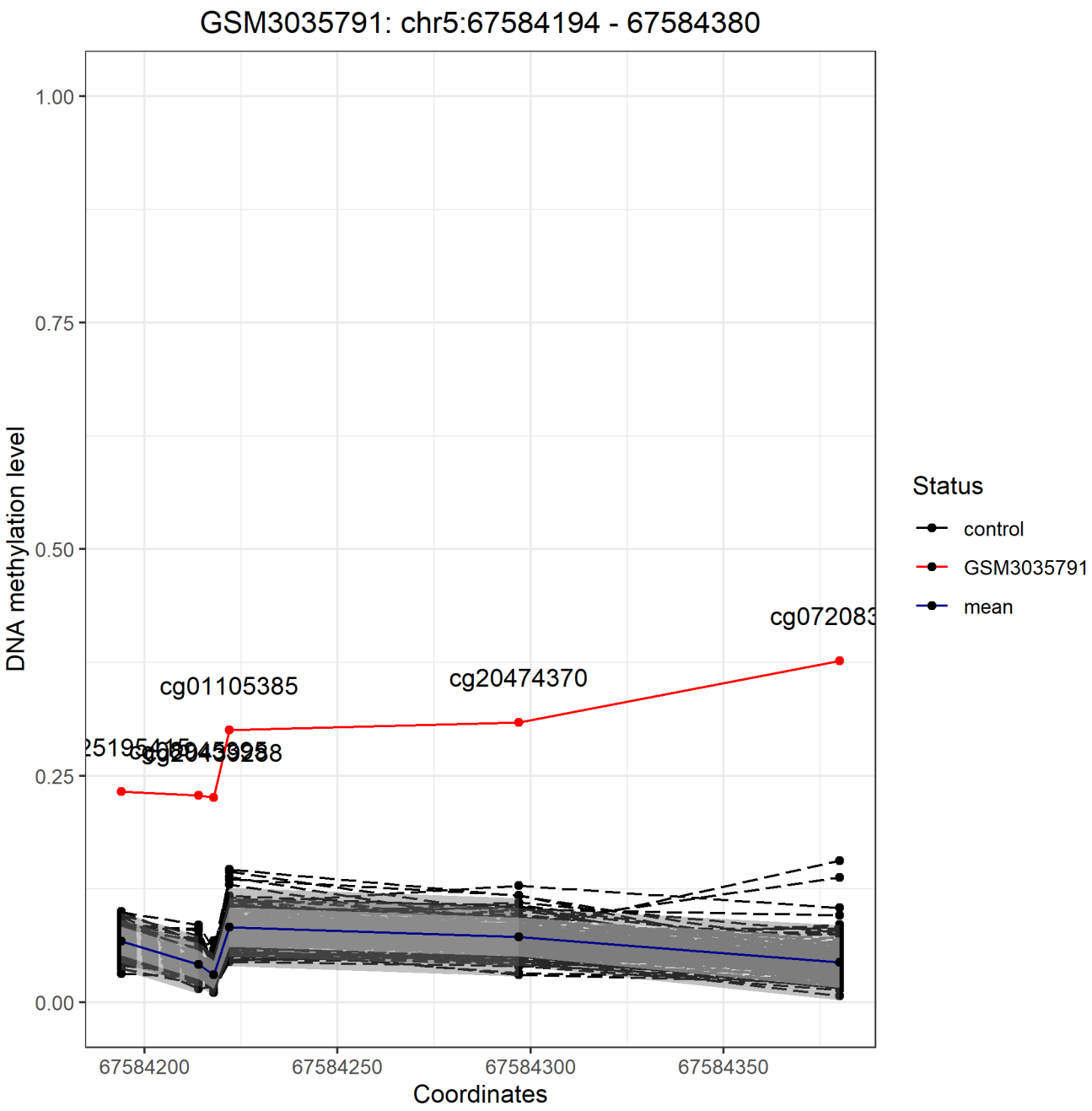


Figure 8: GSE111629 cohort samples in the region chr17:46018654-46019184

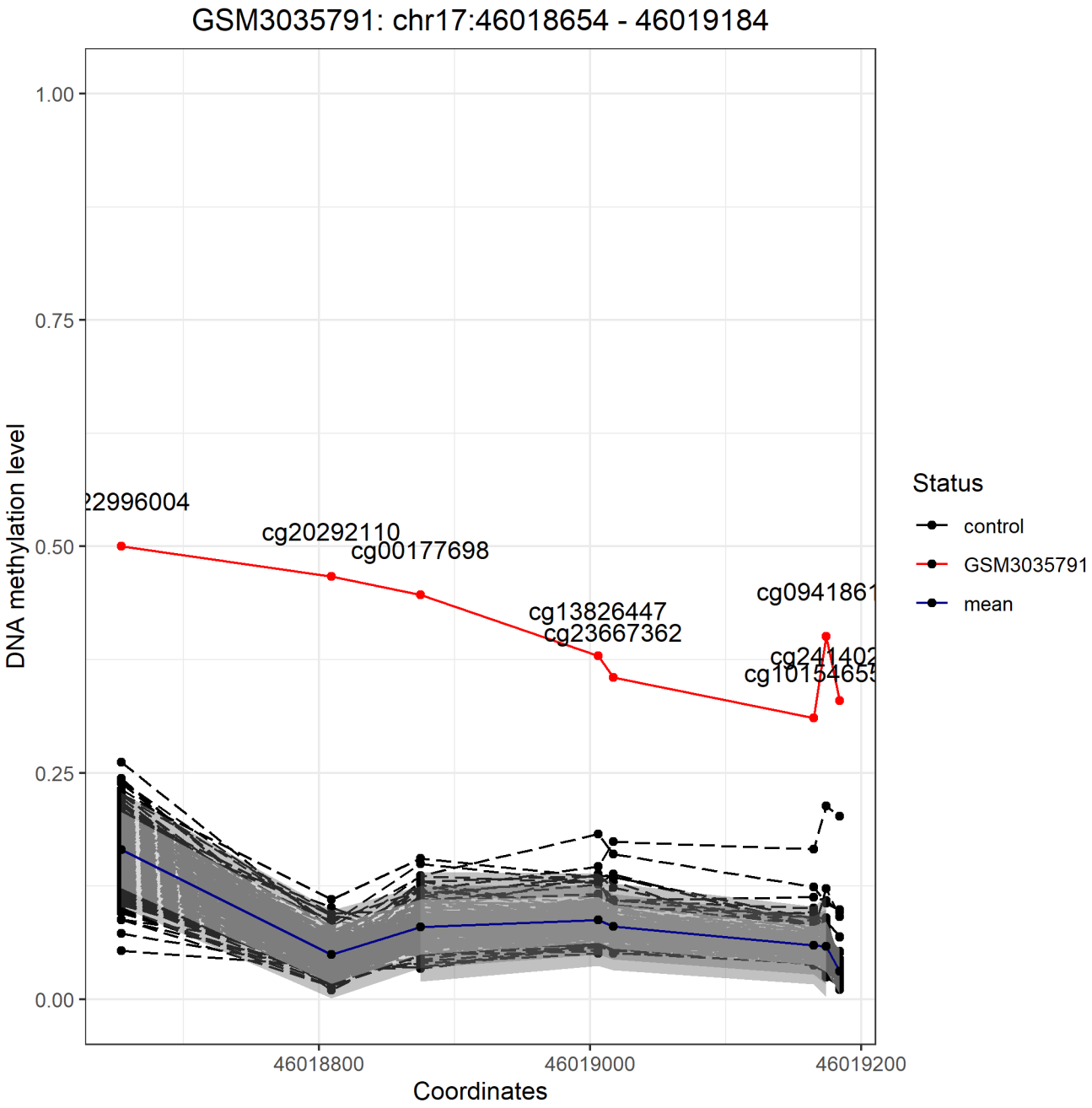


Figure 9: GSE111629 cohort samples in the region chr5:67583972-67584380

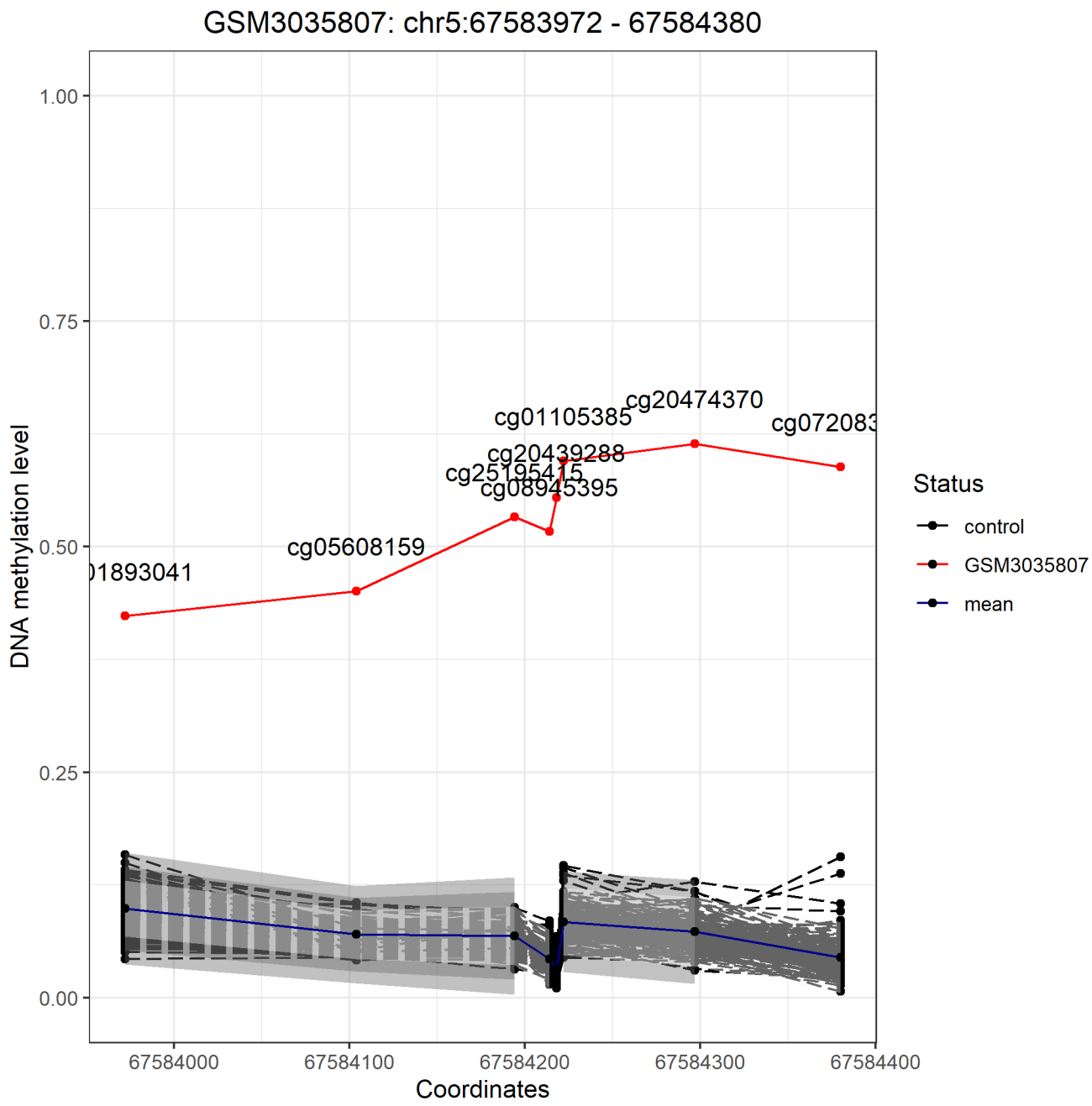
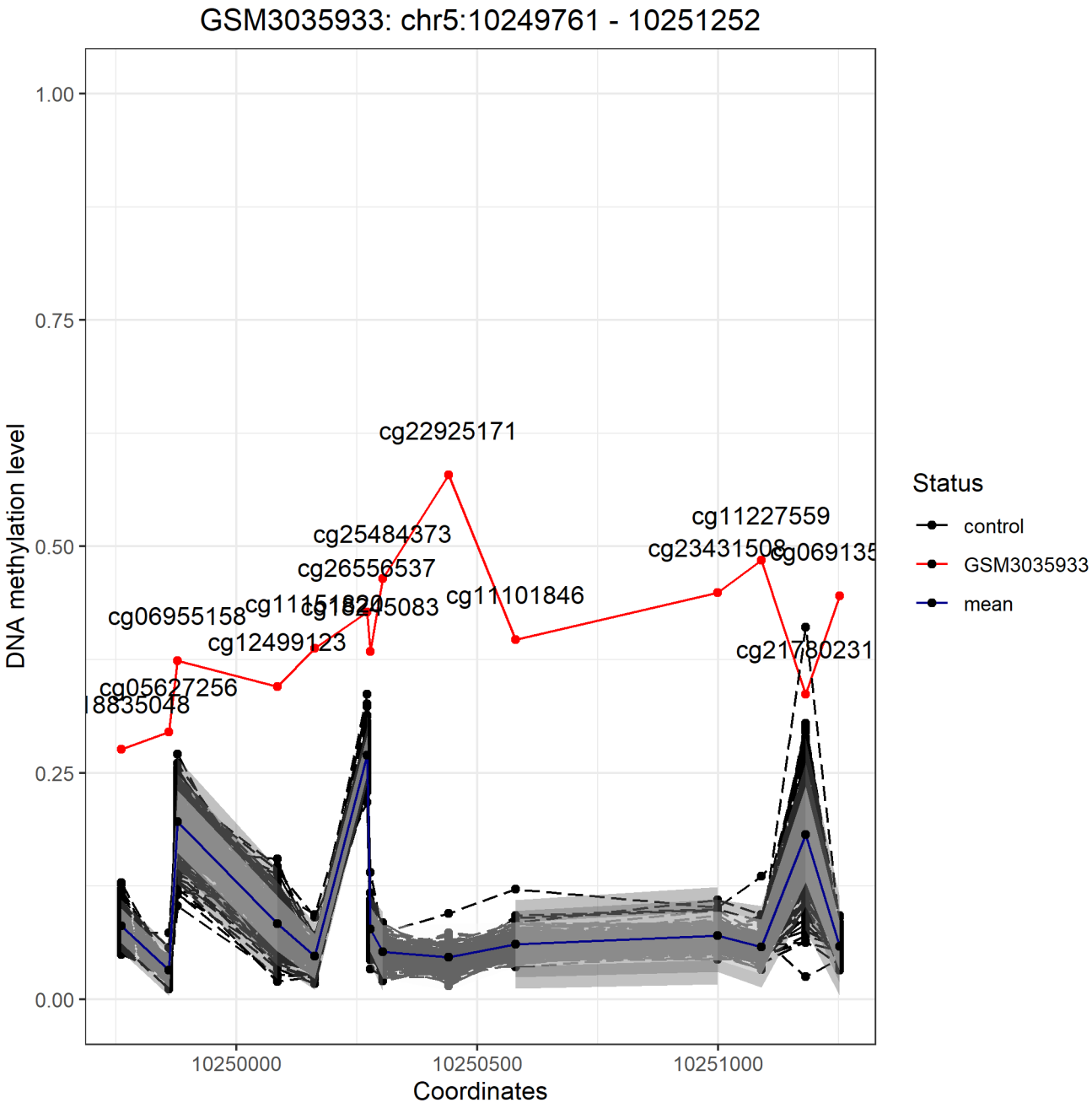


Figure 10: GSE111629 cohort samples in the region chr5:10249761-10251252



6.3 Results

We compared GSM1235784 case sample against randomly selected control samples from GSE51032. furthermore, GSM3035933, GSM3035791, GSM3035807 and GSM3035685 case samples were studied against controls from GSE111629 specifying a region of 20 kb and $\geq$ 3 GpGs.

We obtained similar results in both cohorts. We observed that the methods manova, mahalanobis distance, multivariate linear models, quantile and beta 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

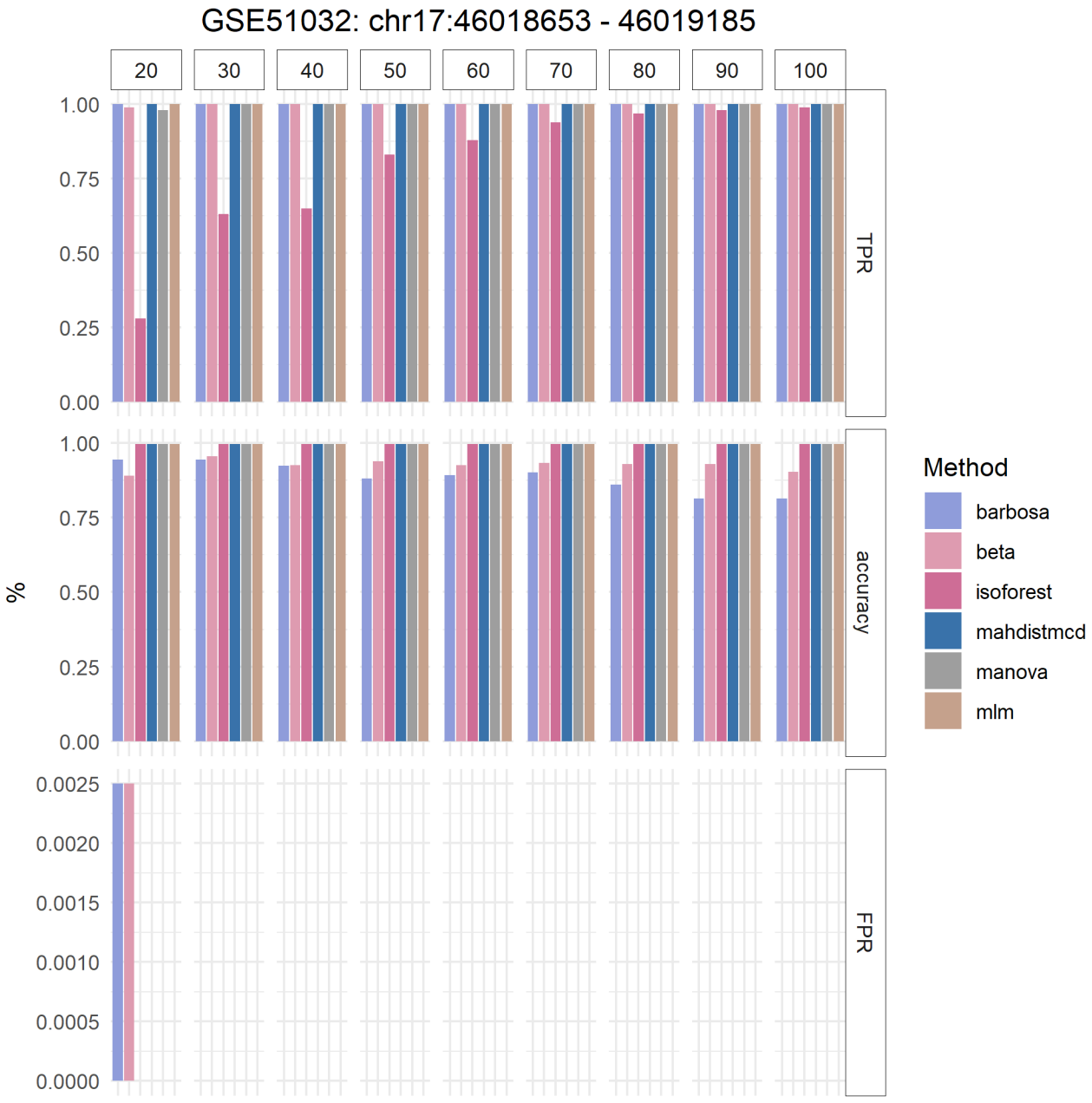
method	n	TPR	accuracy	FPR
manova	20	98	99.6	0.00
mlm	20	100	99.6	0.00
mahdistmcd	20	100	99.6	0.00
isoforest	20	28	99.6	0.00
barbosa	20	100	94.4	0.25
beta	20	99	89.0	0.25
manova	30	100	99.6	0.00
mlm	30	100	99.6	0.00
mahdistmcd	30	100	99.6	0.00
isoforest	30	63	99.6	0.00
barbosa	30	100	94.4	0.00
beta	30	100	95.4	0.00
manova	40	100	99.6	0.00
mlm	40	100	99.6	0.00
mahdistmcd	40	100	99.6	0.00
isoforest	40	65	99.6	0.00
barbosa	40	100	92.4	0.00
beta	40	100	92.6	0.00
manova	50	100	99.6	0.00
mlm	50	100	99.6	0.00
mahdistmcd	50	100	99.6	0.00
isoforest	50	83	99.6	0.00
barbosa	50	100	88.0	0.00
beta	50	100	93.8	0.00
manova	60	100	99.6	0.00
mlm	60	100	99.6	0.00
mahdistmcd	60	100	99.6	0.00
isoforest	60	88	99.6	0.00
barbosa	60	100	89.2	0.00
beta	60	100	92.6	0.00

	method	n	TPR	accuracy	FPR
31	manova	70	100	99.6	0
32	mlm	70	100	99.6	0
33	mahdistmcd	70	100	99.6	0
34	isoforest	70	94	99.6	0
35	barbosa	70	100	90.0	0
36	beta	70	100	93.2	0
37	manova	80	100	99.6	0
38	mlm	80	100	99.6	0
39	mahdistmcd	80	100	99.6	0
40	isoforest	80	97	99.6	0
41	barbosa	80	100	86.0	0
42	beta	80	100	92.9	0
43	manova	90	100	99.6	0
44	mlm	90	100	99.6	0
45	mahdistmcd	90	100	99.6	0
46	isoforest	90	98	99.6	0
47	barbosa	90	100	81.3	0
48	beta	90	100	92.9	0
49	manova	100	100	99.6	0
50	mlm	100	100	99.6	0
51	mahdistmcd	100	100	99.6	0
52	isoforest	100	99	99.6	0
53	barbosa	100	100	81.3	0
54	beta	100	100	90.3	0

of control samples obtaining a TPR ≥ 75 with 50 control samples or more. 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 random regions of 20 kb and ≥ 3 GpGs for both, control and case samples. We compared individual methylation profiles of a single case sample against control samples. We observed that in both cohorts and for every approach the FPR value was minimum $< 0.01\%$.

Figure 11: epimutations performance for GSE51032 cohort detecting the epivariation located in chr5:10249760-10251253



method	n	TPR	accuracy	FPR
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
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
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
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
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

	method	n	TPR	accuracy	FPR
31	barbosa	60	100.000	92.825	0.00
32	beta	60	100.000	92.825	0.00
33	isoforest	60	78.750	93.500	0.00
34	mahdistmcd	60	100.000	92.825	0.00
35	manova	60	100.000	92.825	0.00
36	mlm	60	100.000	92.825	0.00
37	barbosa	70	100.000	92.825	0.25
38	beta	70	100.000	92.825	0.00
39	isoforest	70	90.125	93.575	0.00
40	mahdistmcd	70	100.000	92.825	0.00
41	manova	70	100.000	92.825	0.00
42	mlm	70	100.000	92.825	0.00
43	barbosa	80	100.000	92.825	0.00
44	beta	80	100.000	92.825	0.00
45	isoforest	80	96.375	93.075	0.00
46	mahdistmcd	80	100.000	92.825	0.00
47	manova	80	100.000	92.825	0.00
48	mlm	80	100.000	92.825	0.00
49	barbosa	90	100.000	92.825	0.00
50	beta	90	100.000	92.825	0.25
51	isoforest	90	97.125	93.000	0.00
52	mahdistmcd	90	100.000	92.825	0.00
53	manova	90	100.000	92.825	0.00
54	mlm	90	100.000	92.825	0.00

Figure 12: epimutations performance using GSE111629 cohort to detect the epivariation located in chr5:10249760-10251253

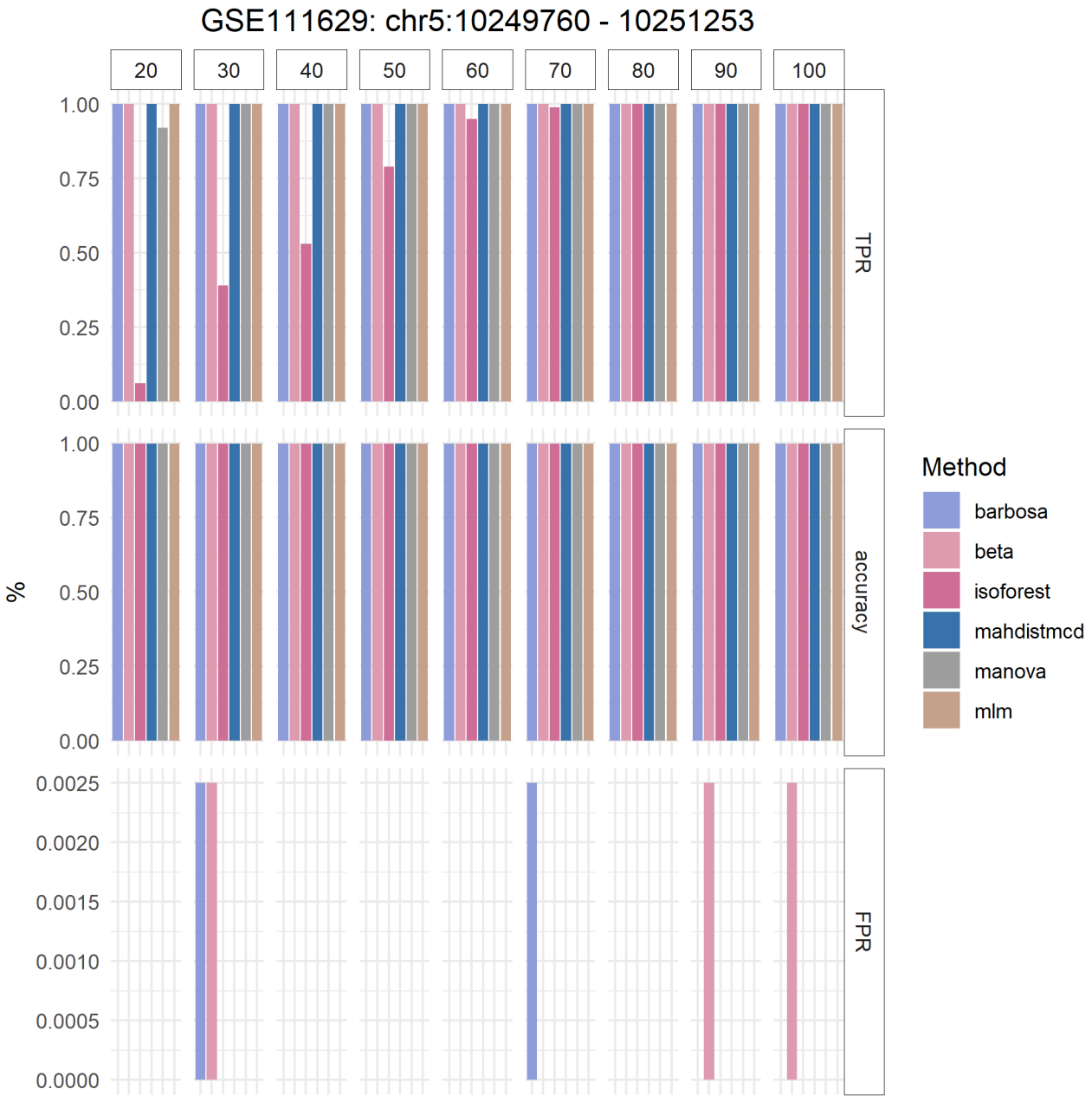


Figure 13: epimutations performance using GSE111629 cohort to detect the epivariation located in chr5:67583971-67584381

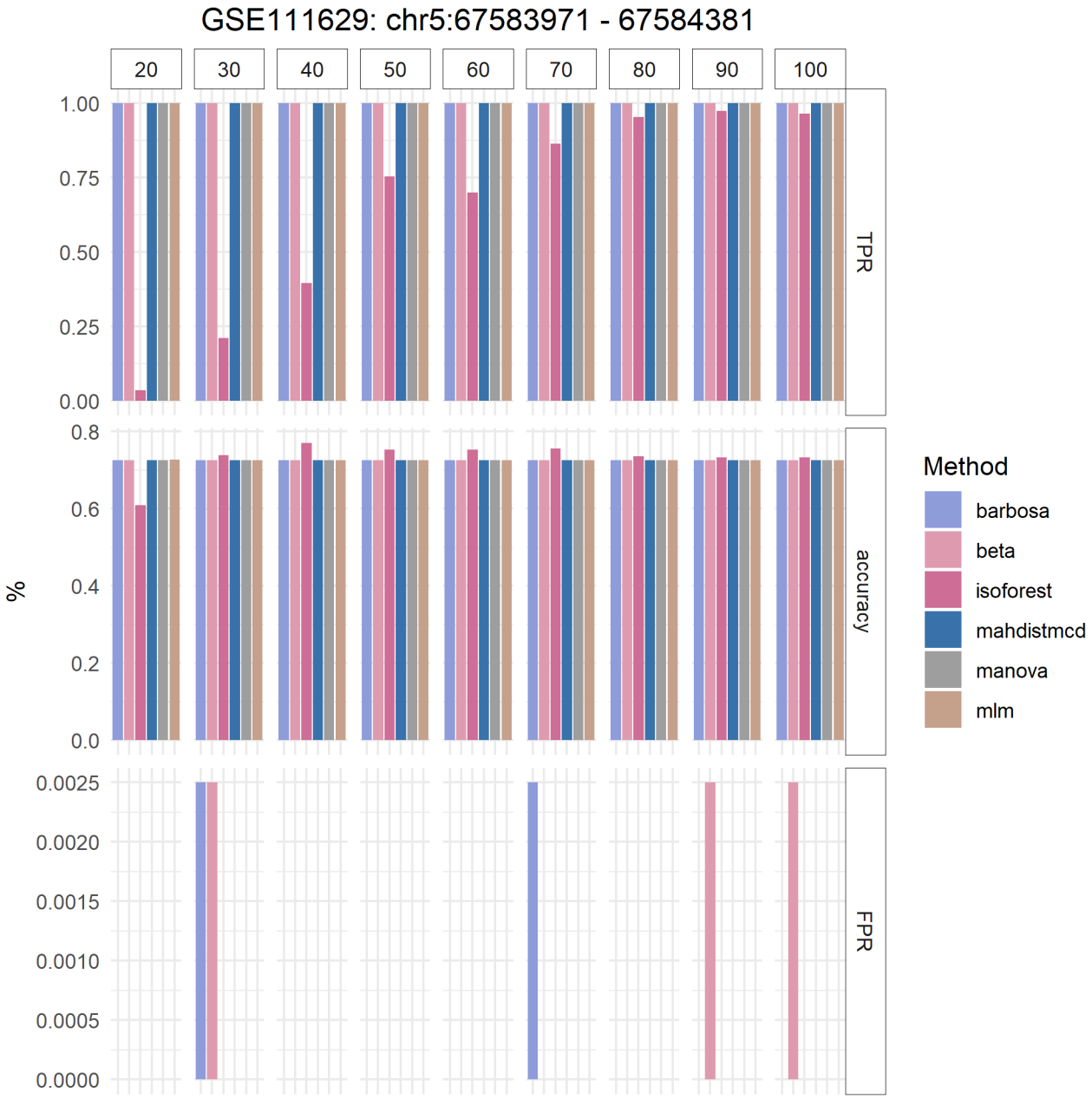


Figure 14: epimutations performance using GSE111629 cohort to detect the epivariation located in chr17:46018653-46019185

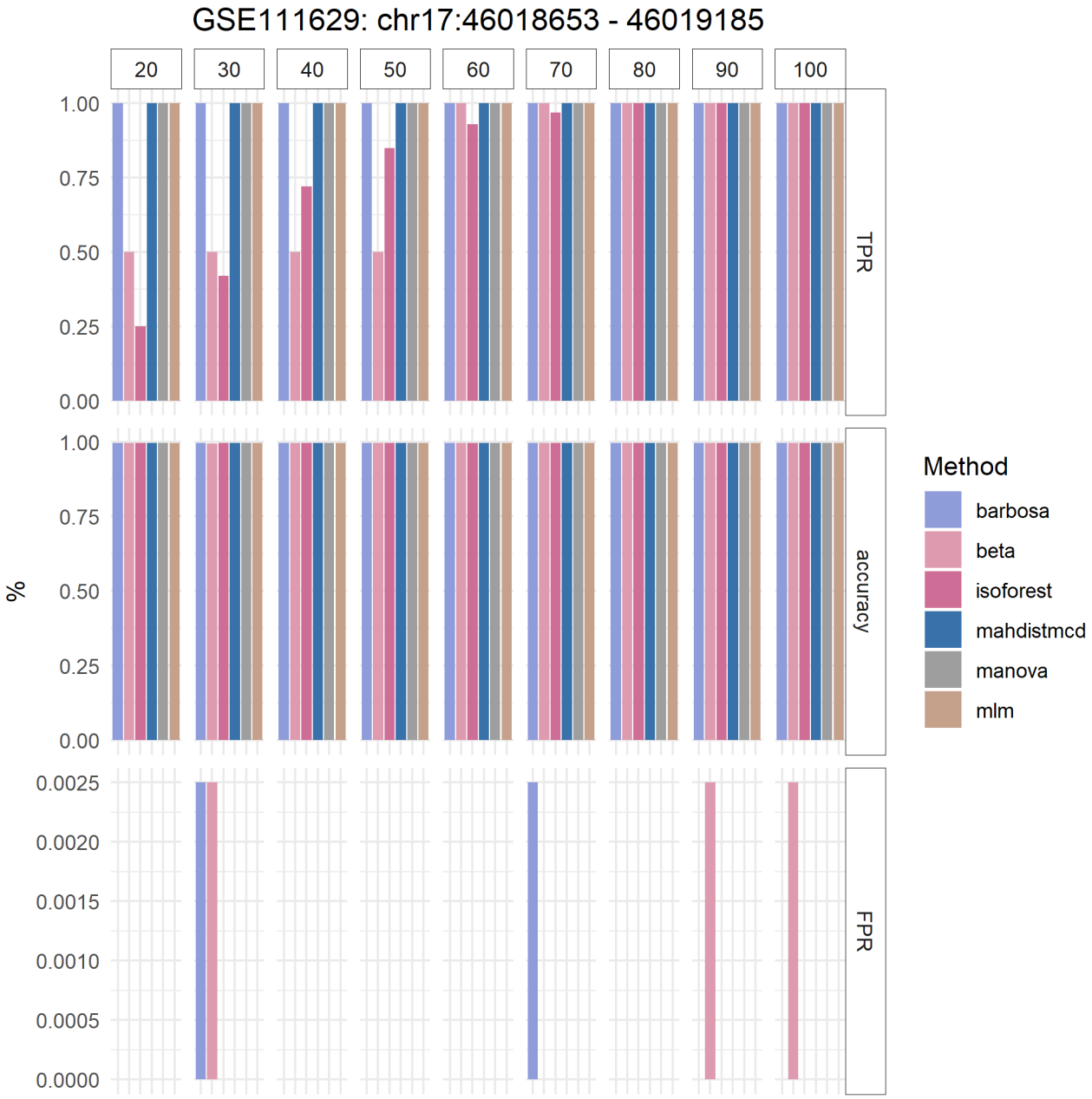
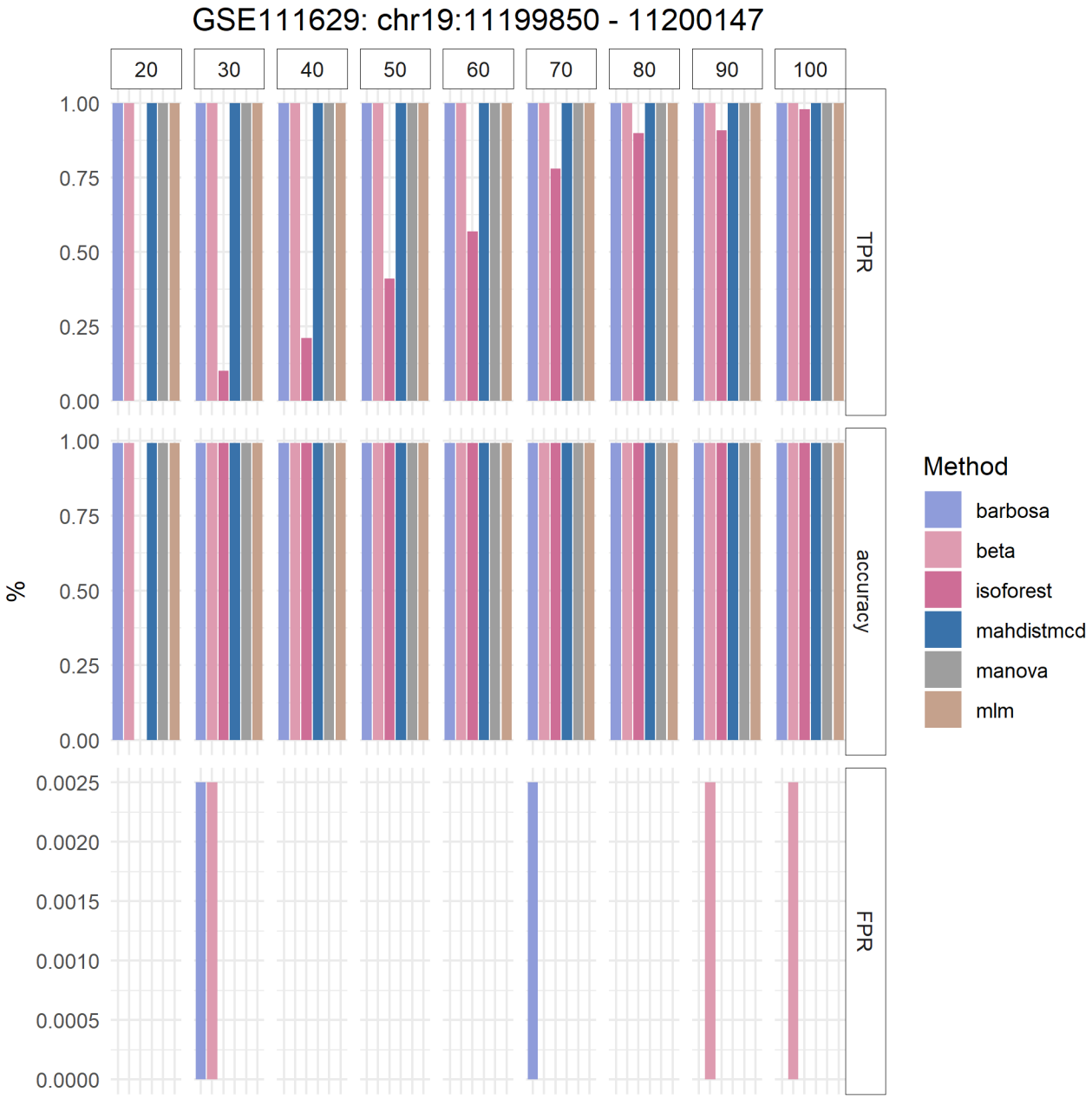


Figure 15: epimutations performance using GSE111629 cohort to detect the epivariation located in chr5:11199850-11200147



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
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.

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.

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.

Friedrich, Sarah, Frank Konietschke, 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.

Holmes, J Bradley, Eric Moyer, Lon Phan, Donna Maglott, and Brandi Kattman. 2020. "SPDI: Data Model for Variants and Applications at NCBI." *Bioinformatics* 36 (6): 1902–7.

Hon, Chung-Chau, and Piero Carninci. 2020. "Expanded ENCODE Delivers Invaluable Genomic Encyclopedia." Nature Publishing Group.

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.

López-Bastida, Julio, Juan Oliva-Moreno, Renata Linertová, and Pedro Serrano-Aguilar. 2016. "Social/Economic Costs and Health-Related Quality of Life in Patients with Rare Diseases in Europe." Springer.

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.

Reproducibility, Unrivaled Assay. 2012. "Infinium HumanMethylation450 BeadChip."

Serra-Juhé, Clara, Ivon Cuscó, Aïda Homs, Raquel Flores, Núria Torán, and Luis A Pérez-Jurado. 2015. "DNA Methylation Abnormalities in Congenital Heart Disease." *Epigenetics* 10 (2): 167–77.