# 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

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. Due to their low prevalence, there is a lack of knowledge which causes a delay in diagnosis and a lack of effective treatment. Thus, affecting the life expectancy and quality of the patient. Current methodologies for identifying mutations related to rare diseases have relied on Whole Genomic Sequencing. Despite exhaustive assessments, in a large proportion of case subjects, the underlying genetic etiology is not identified or the clinical assessment does not indicate a diagnosis. In order to overcome this difficulty, genome-wide DNA methylation analysis has been proposed to facilitate the molecular diagnosis of unresolved clinical cases and be consider for routine clinical assessment. We developed `epimutacions`, a method that detects outliers in DNA methylation patterns associated with the diseases as proposed by (Aref-Eshghi et al. 2019). We validated our method by performing simulations based on the data and results obtained in the survey (Garg et al. 2020).

## 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 $\approx$ 20 kb containing $\geq$ 3 GpGs.

Table 1: 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.

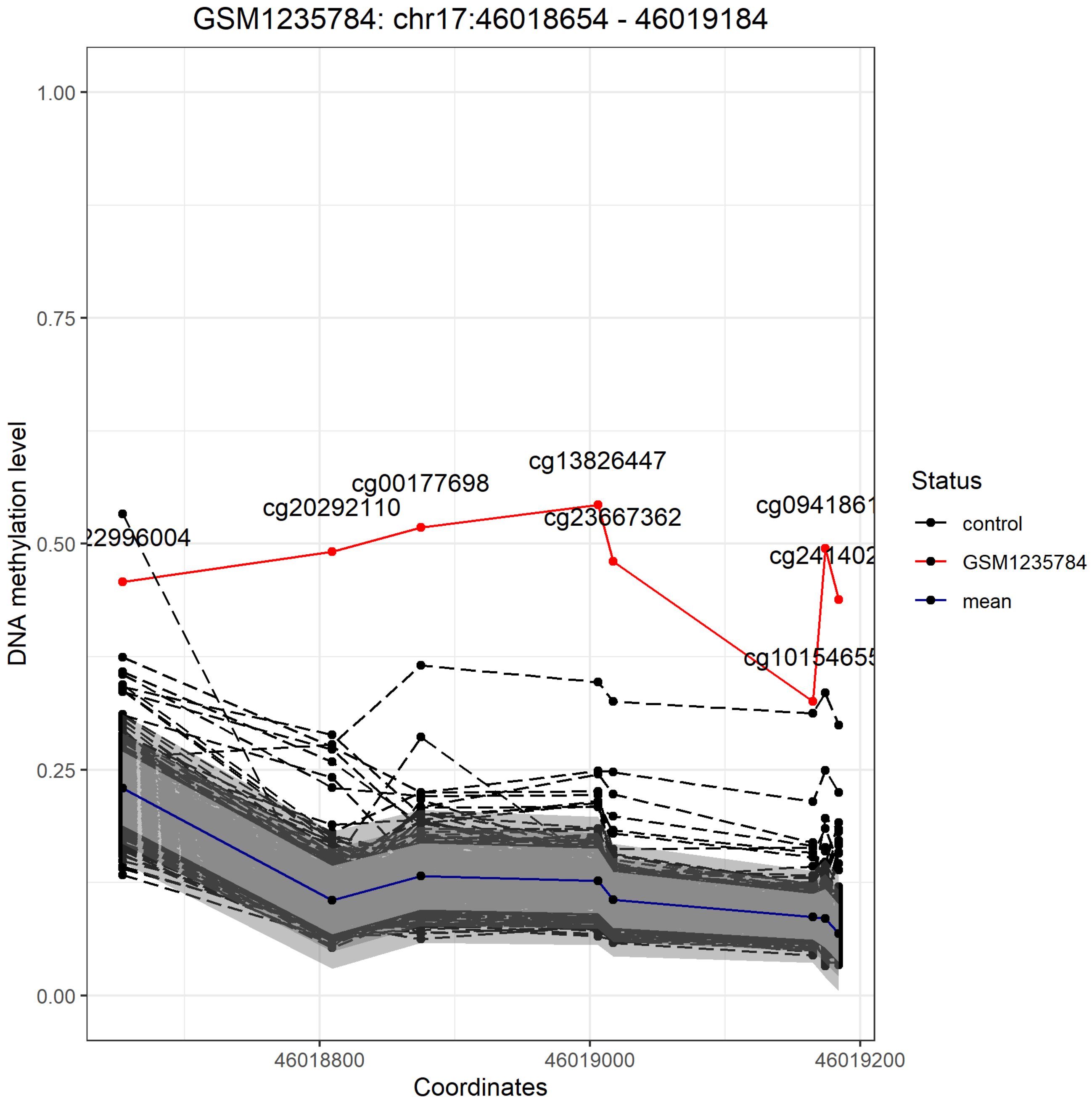

Figure 1: GSE51032 samples in the region chr17:46018654-46019184

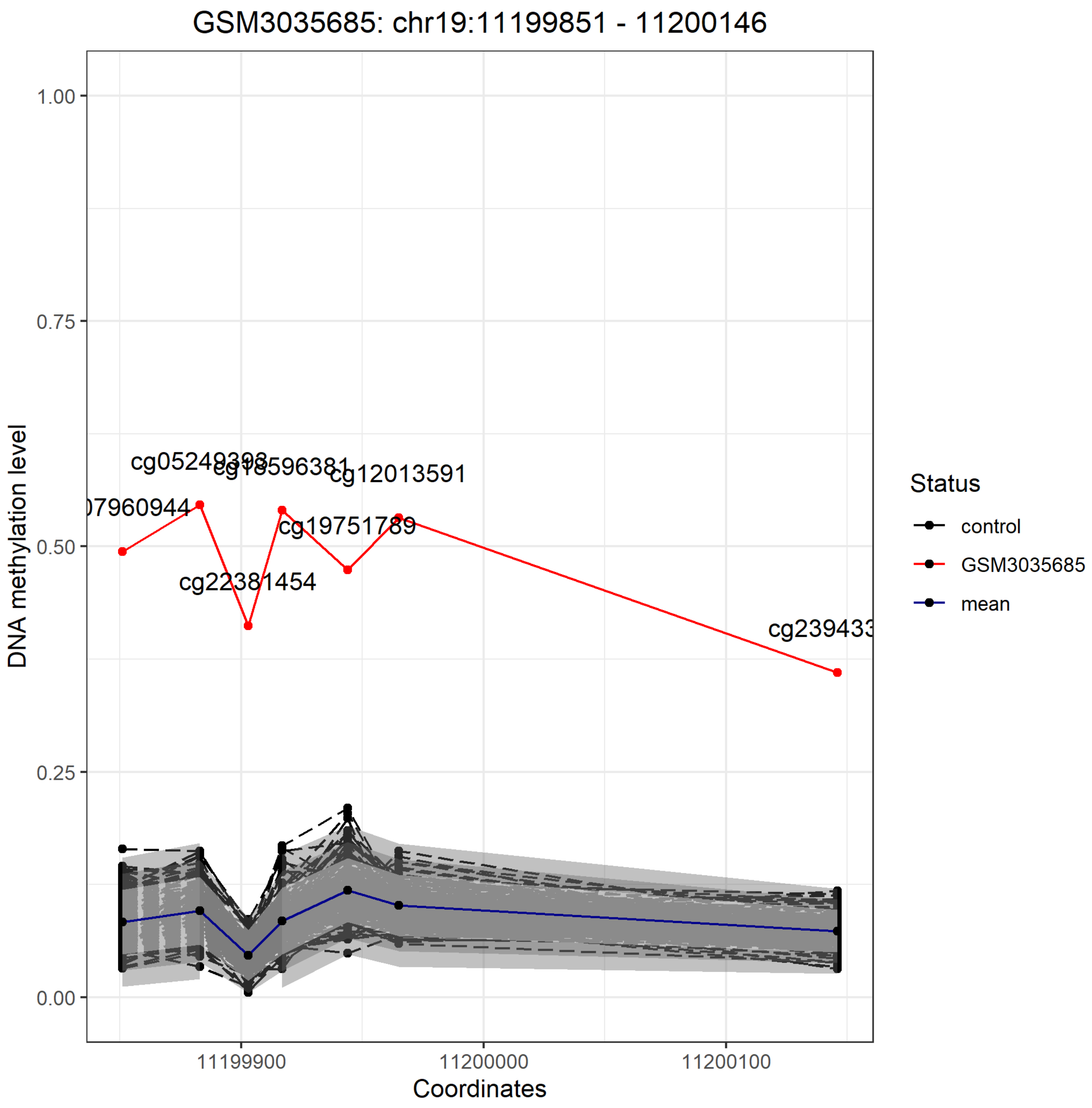

Figure 2: GSE111629 samples in the region chr19:11199851-11200146

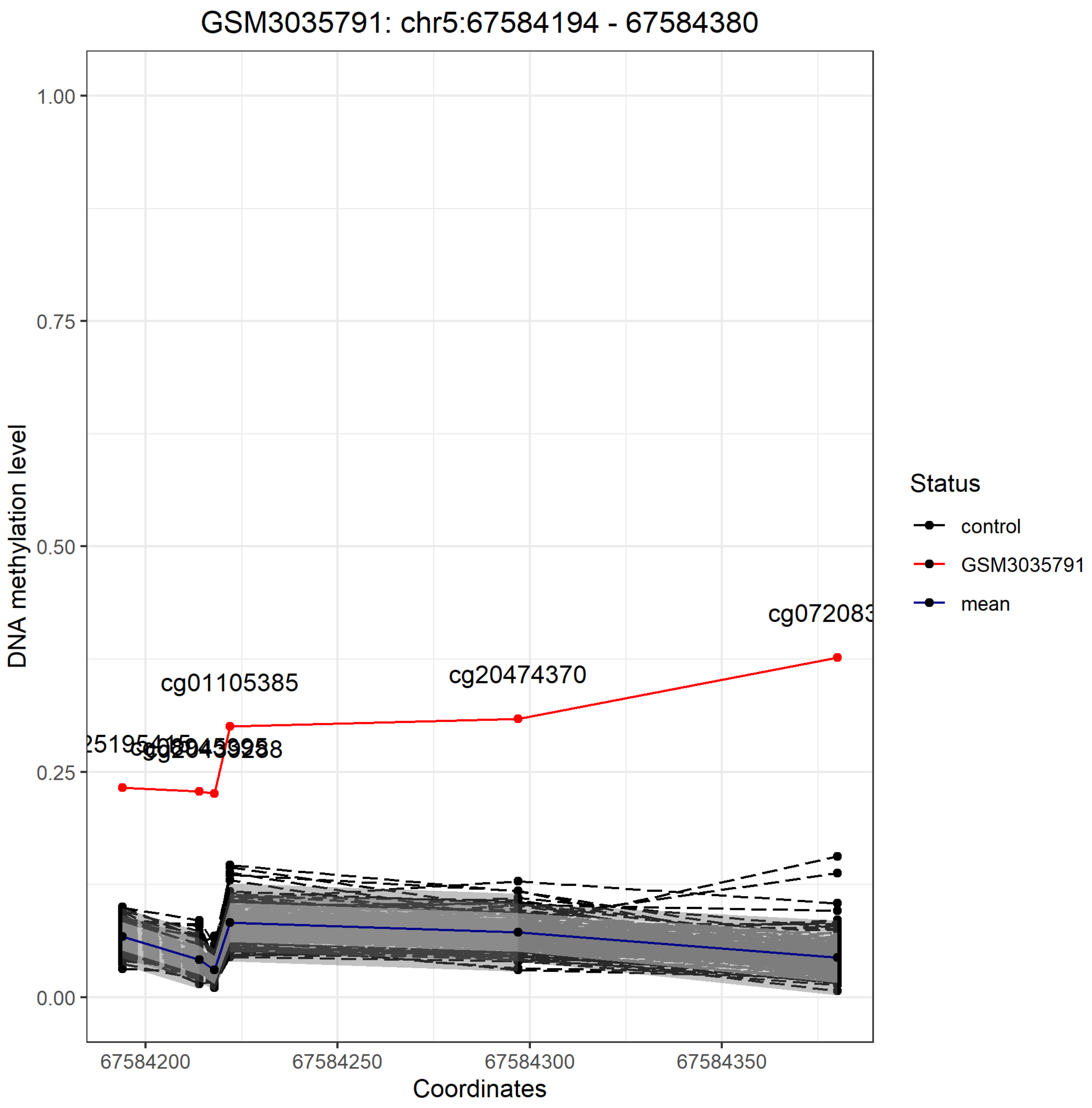

Figure 3: GSE111629 samples in the region chr5:67584194-67584380

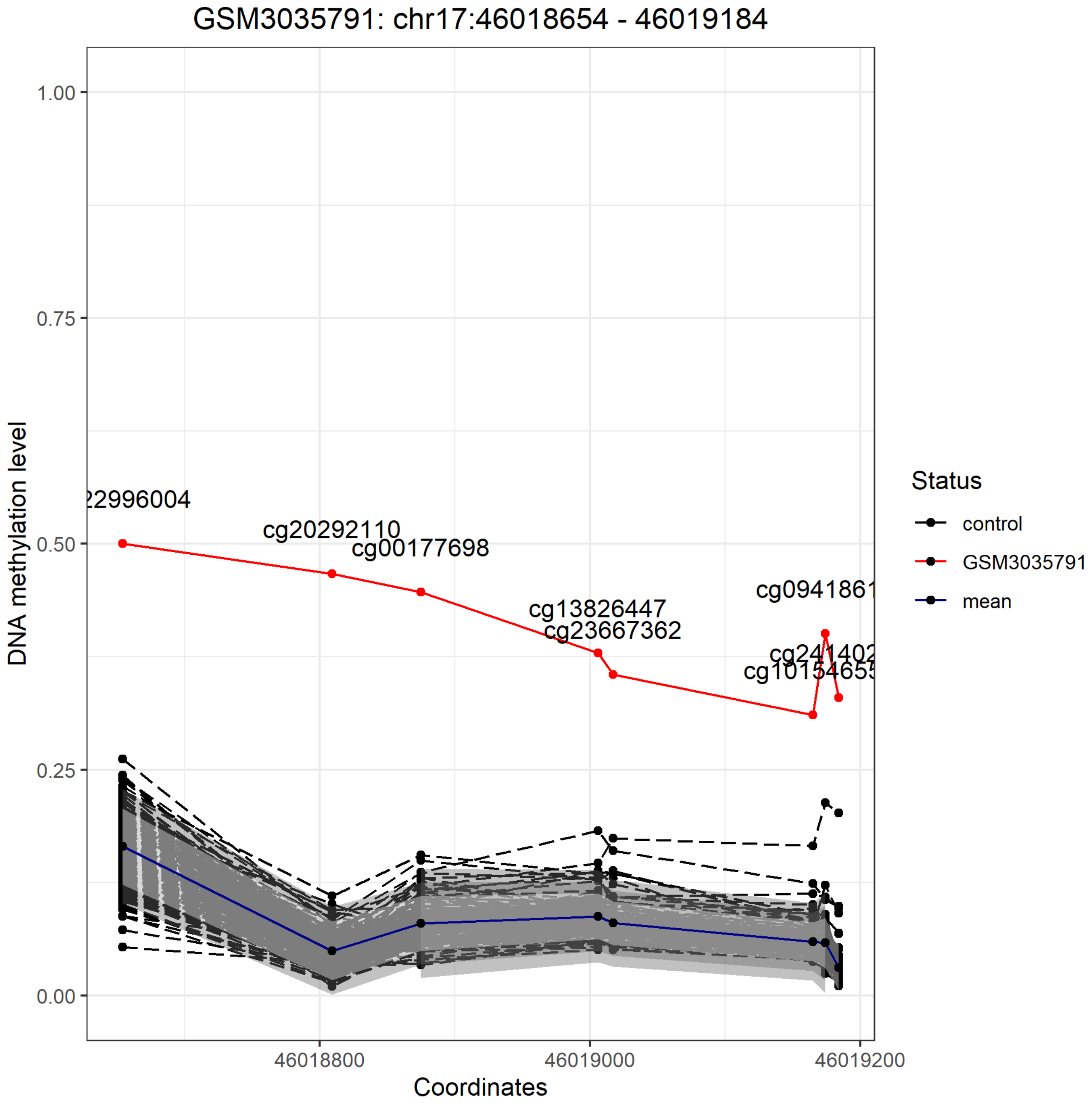

Figure 4: GSE111629 samples in the region chr17:46018654-46019184

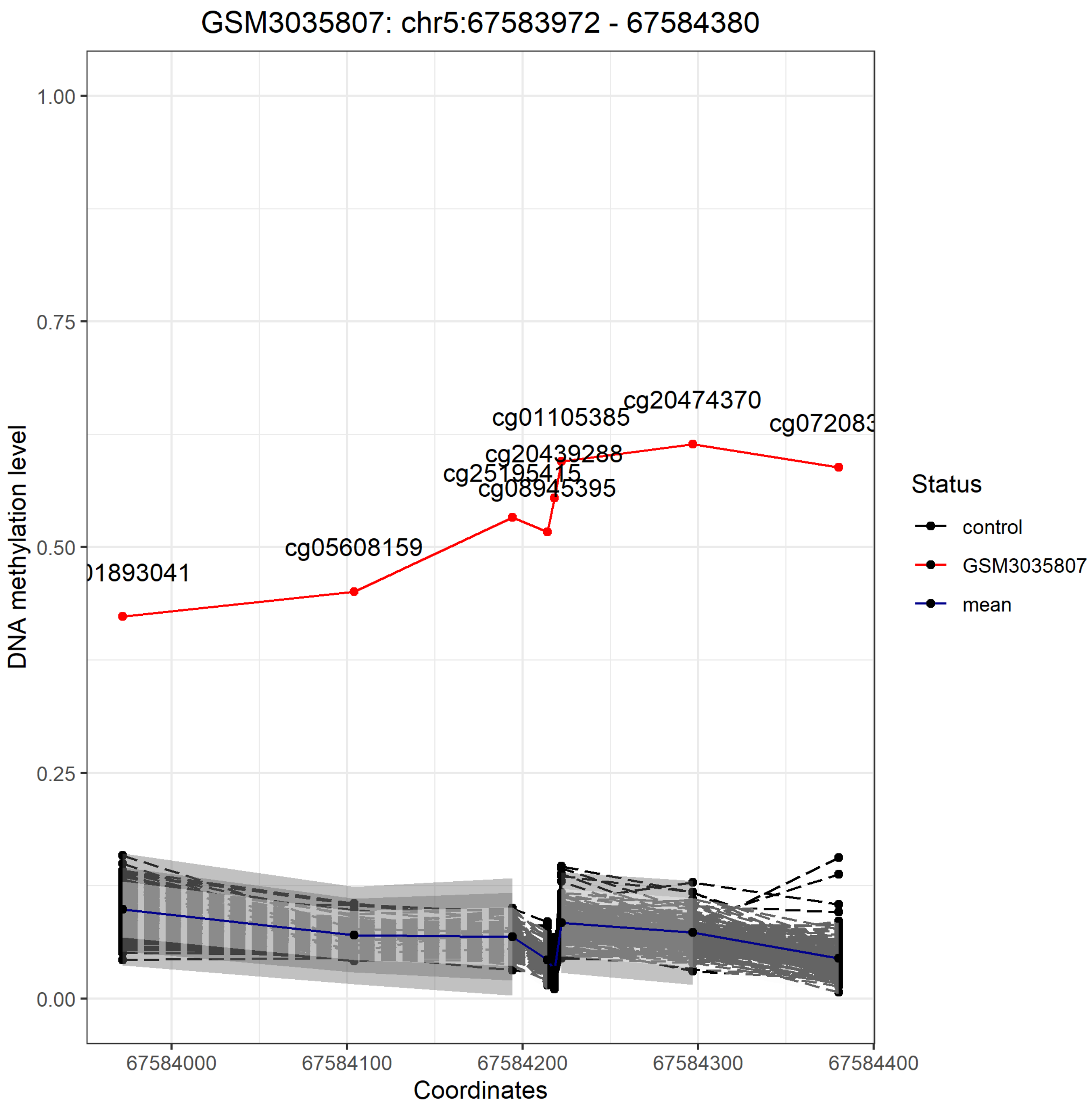

Figure 5: GSE111629 samples in the region chr5:67583972-67584380

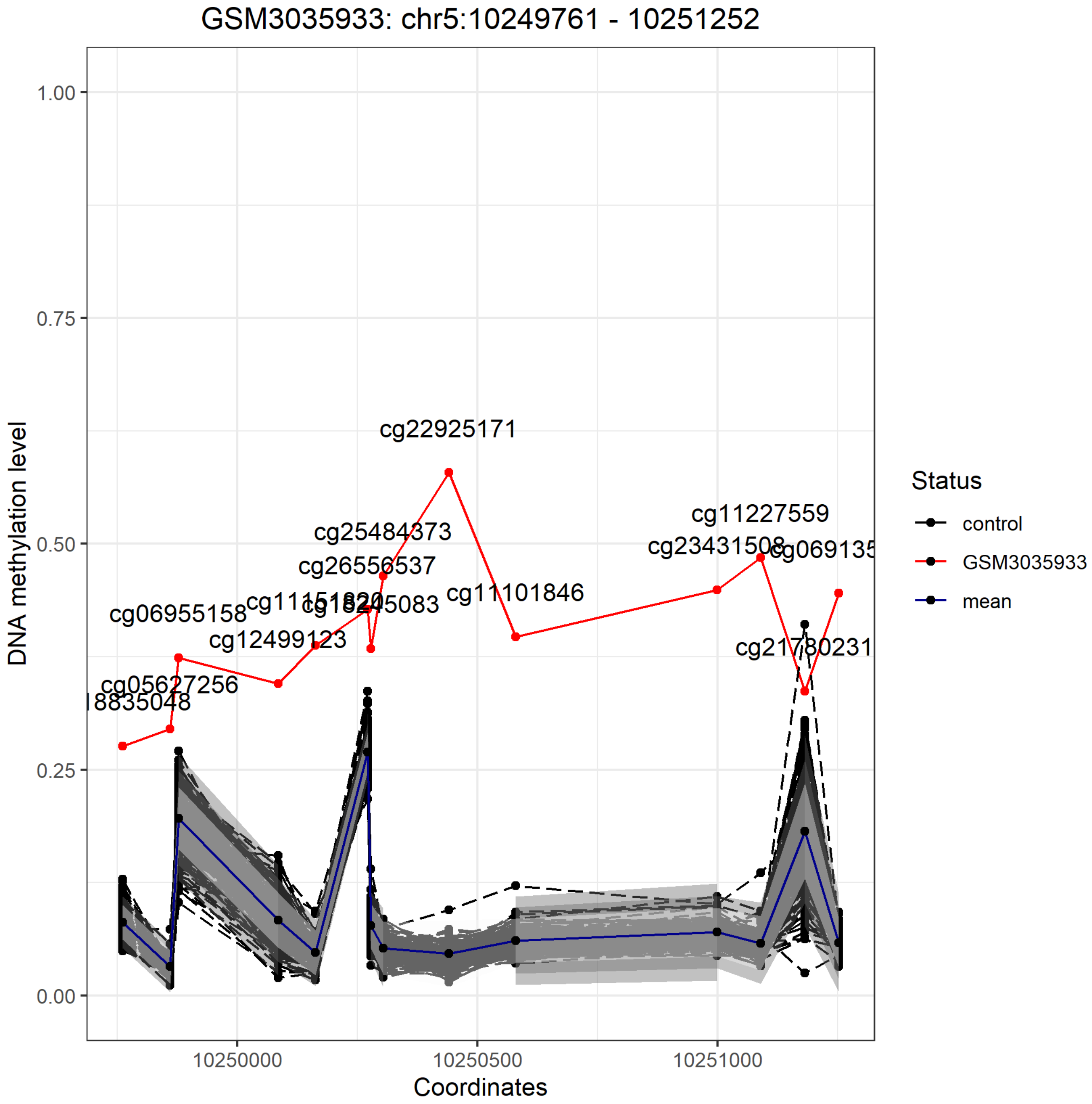

Figure 6: GSE111629 samples in the region chr5:10249761-10251252

## 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 $\geq 3$ 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 $\geq$ 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 $\geq$ 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 $< 0.01\%$.

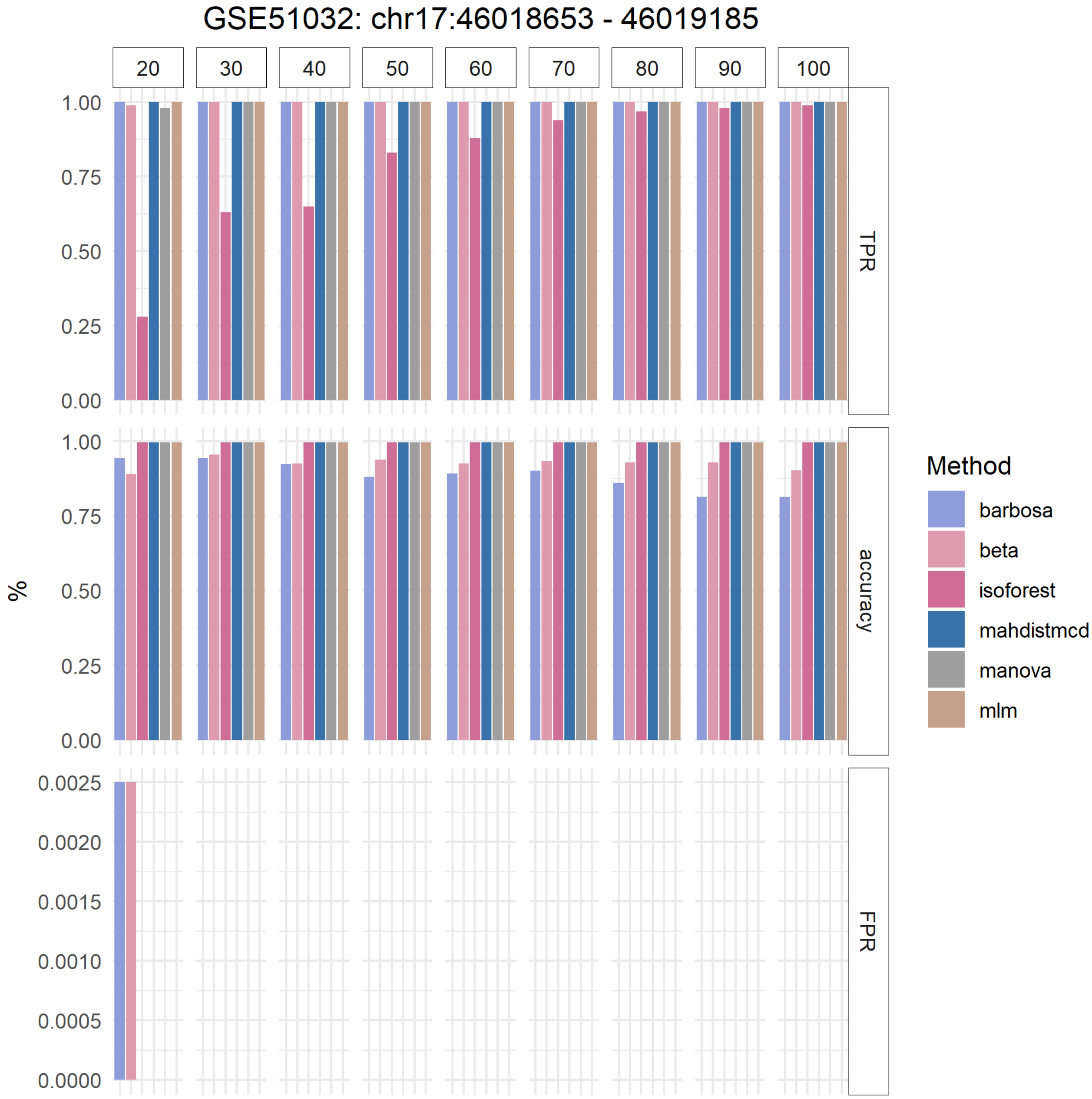

Figure 7: epimutations performance for GSE51032 cohort detecting the epivariation located in chr5:10249760-10251253

| 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 | 98 | 99.6 | 0.00 |
| barbosa | 100 | 81.3 | 0.00 |
| beta | 100 | 92.9 | 0.00 |
| **n100** | | | |
| | 100 | 99.6 | 0.00 |

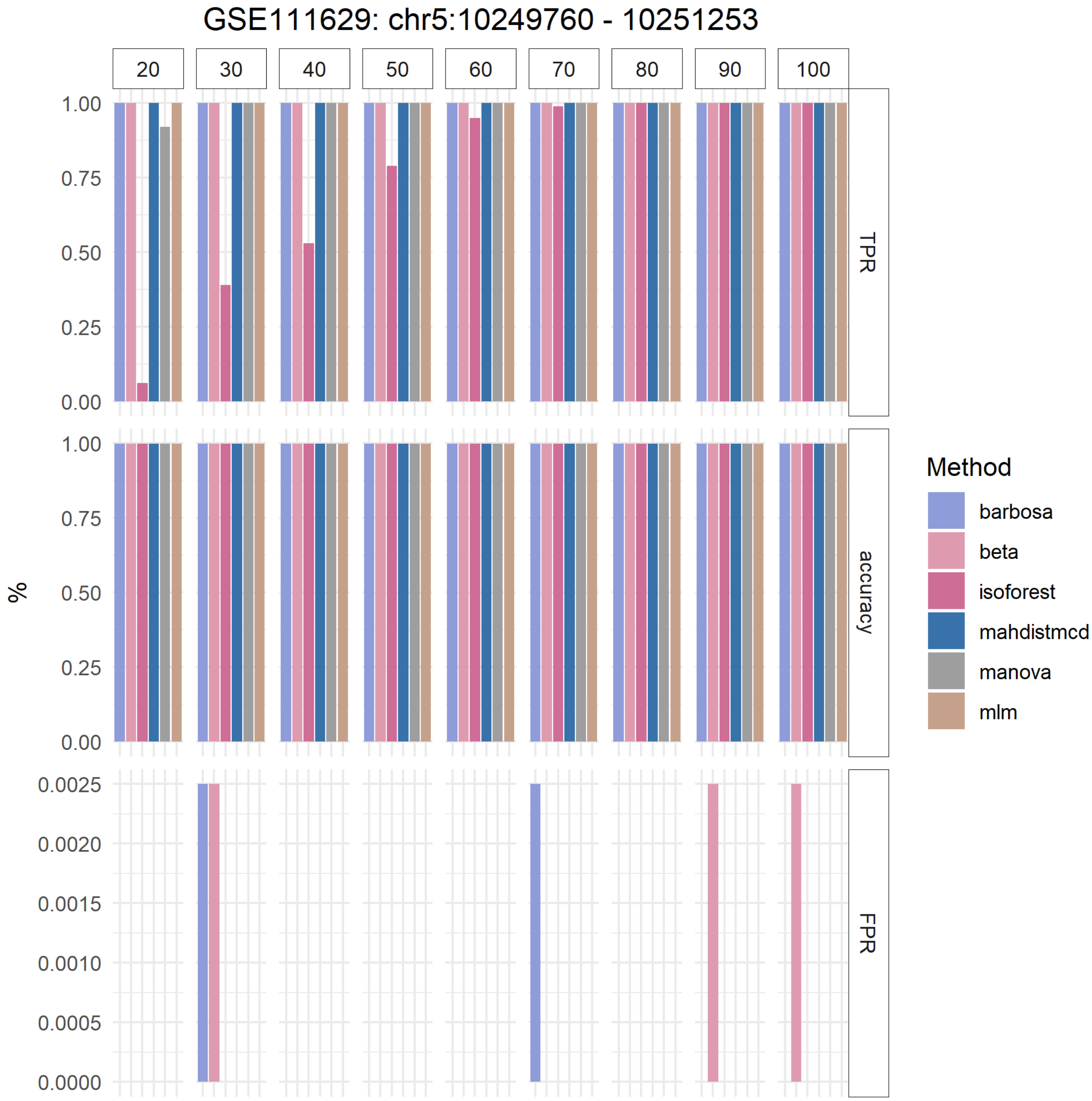

Figure 8: epimutations performance using GSE111629 cohort to detect the epivariation located in chr5:10249760-10251253

| 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 |

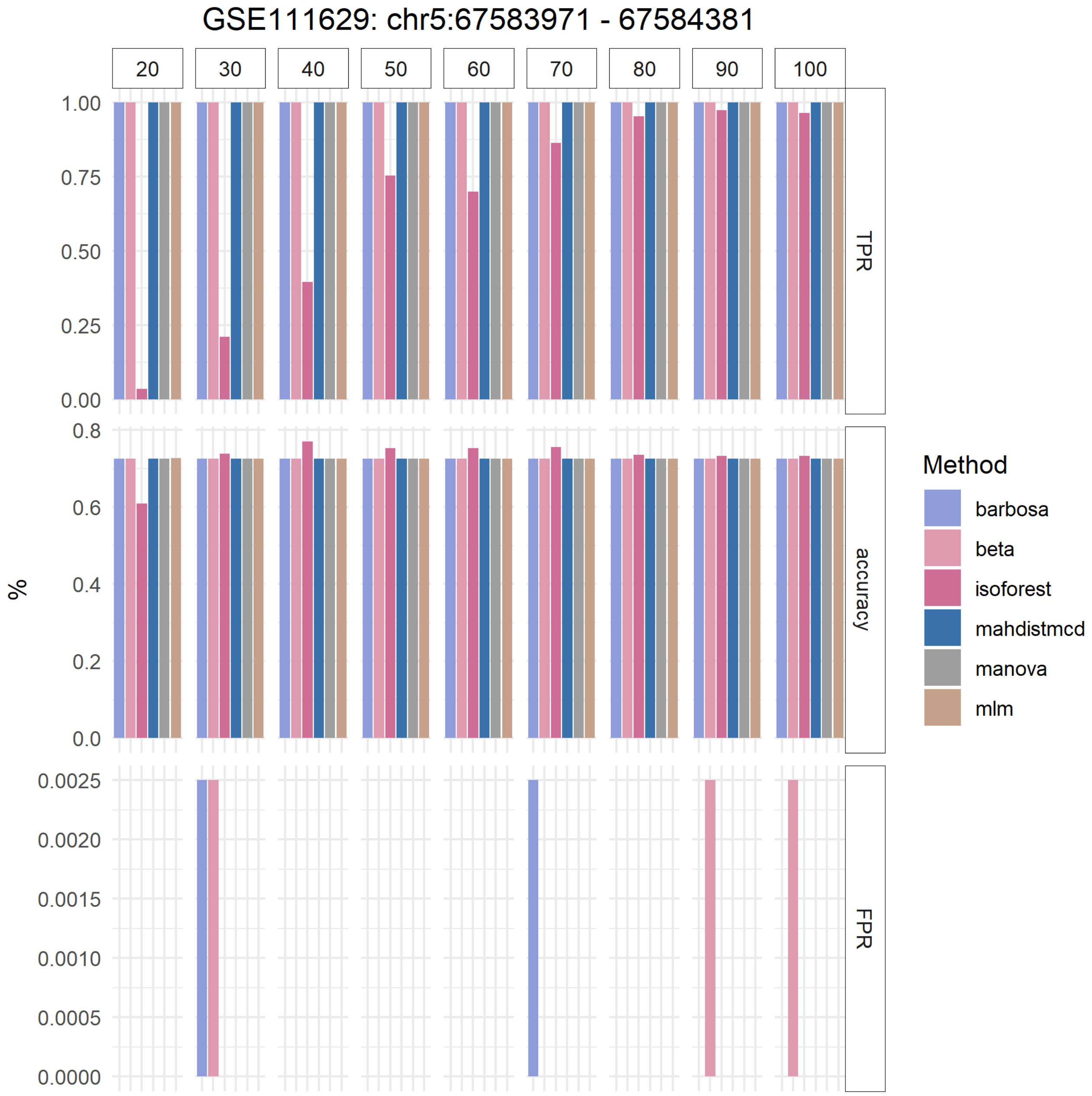

Figure 9: epimutations performance using GSE111629 cohort to detect the epivariation located in chr5:67583971-67584381

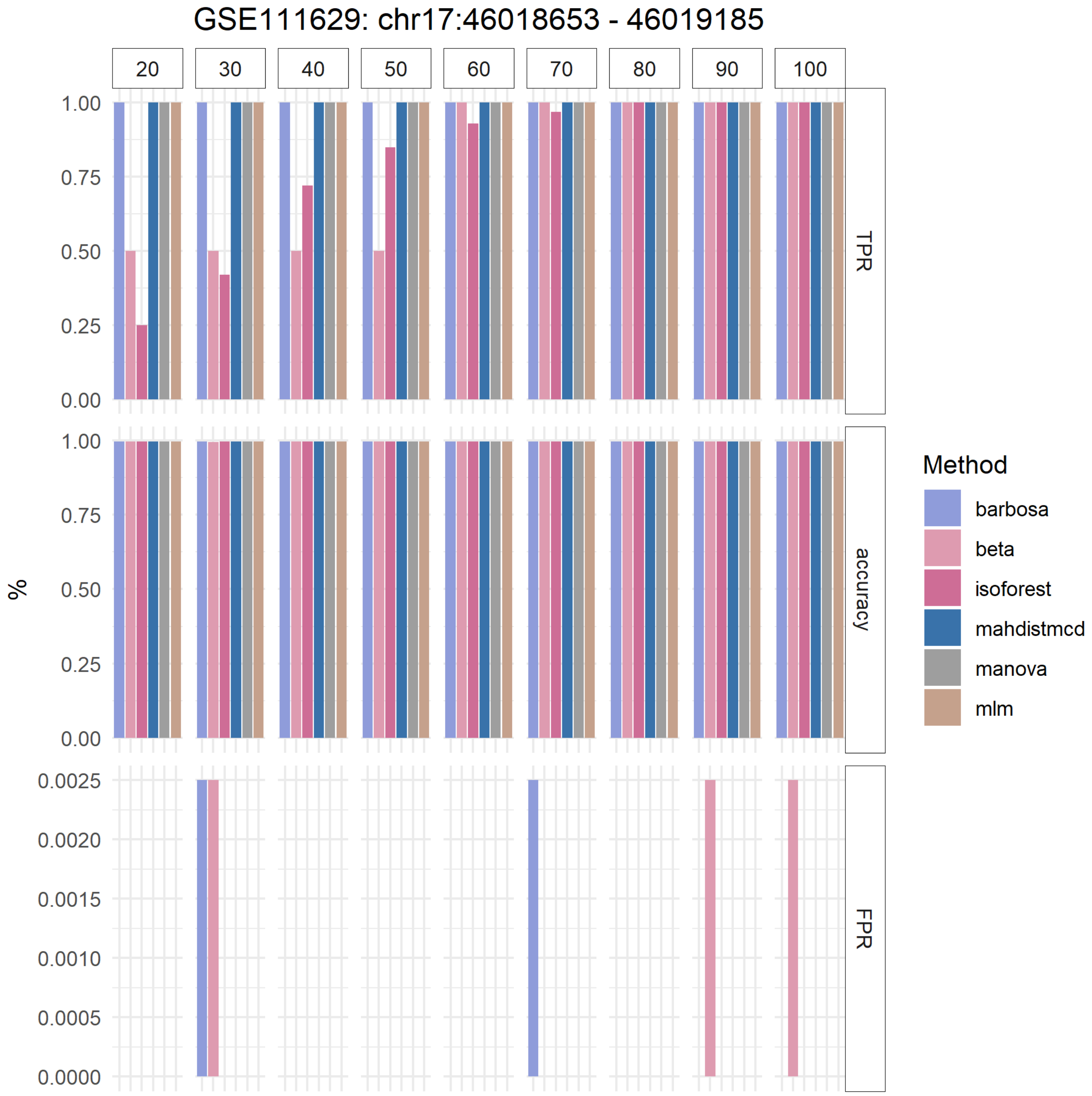

Figure 10: epimutations performance using GSE111629 cohort to detect the epivariation located in chr17:46018653-46019185

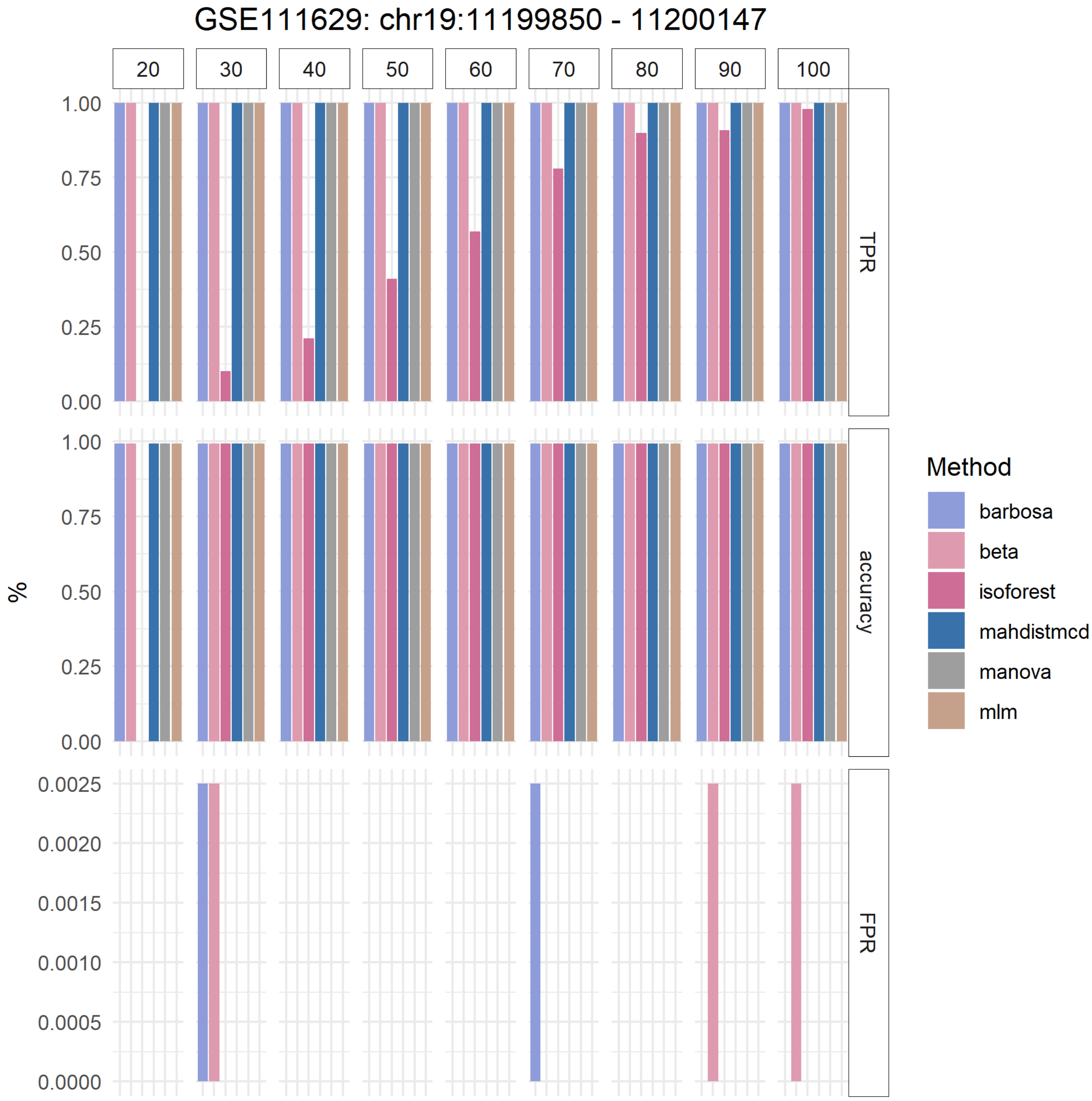

Figure 11: epimutations performance using GSE111629 cohort to detect the epivariation located in chr5:11199850-11200147

# 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.

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.