Bioinformatics, 35(18), 2019, 3257–3262
doi: 10.1093/bioinformatics/btz096
Advance Access Publication Date: 12 February 2019
Original Paper

OXFORD

Genome analysis

mCSEA: detecting subtle differentially methylated regions

Jordi Martorell-Marugán[1,2], Víctor González-Rumayor[2] and Pedro Carmona-Sáez[1,*]

[1]Bioinformatics Unit. GENYO. Centre for Genomics and Oncological Research: Pfizer, University of Granada, Andalusian Regional Government, Granada, Spain and [2]Atrys Health, Barcelona, Spain

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on July 27, 2018; revised on January 9, 2019; editorial decision on February 6, 2019; accepted on February 10, 2019

Abstract

Motivation: The identification of differentially methylated regions (DMRs) among phenotypes is one of the main goals of epigenetic analysis. Although there are several methods developed to detect DMRs, most of them are focused on detecting relatively large differences in methylation levels and fail to detect moderate, but consistent, methylation changes that might be associated to complex disorders.

Results: We present *mCSEA*, an R package that implements a Gene Set Enrichment Analysis method to identify DMRs from Illumina450K and EPIC array data. It is especially useful for detecting subtle, but consistent, methylation differences in complex phenotypes. *mCSEA* also implements functions to integrate gene expression data and to detect genes with significant correlations among methylation and gene expression patterns. Using simulated datasets we show that *mCSEA* outperforms other tools in detecting DMRs. In addition, we applied *mCSEA* to a previously published dataset of sibling pairs discordant for intrauterine hyperglycemia exposure. We found several differentially methylated promoters in genes related to metabolic disorders like obesity and diabetes, demonstrating the potential of *mCSEA* to identify DMRs not detected by other methods.

Availability and implementation: *mCSEA* is freely available from the Bioconductor repository.

Contact: pedro.carmona@genyo.es

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

DNA methylation is by far the most studied epigenetic mark. It affects gene expression and has an important role in several disorders. Epigenome-wide association studies (EWASs) are performed to find associations between DNA methylation alterations and a given phenotype (Flanagan, 2015).

There are several methodologies to determine DNA methylation status, including high-throughput techniques such as whole-genome bisulfite sequencing (WGBS) or methylation arrays. WGBS is the one with the highest coverage but Illumina's BeadChip arrays (Infinium HumanMethylation450and InfiniumMethylationEPIC) are still much more affordable and simpler to analyze, and they are currently the most used platforms in human EWAS (Teh *et al.*, 2016).

EWAS are usually applied to find associations between individual CpG sites and outcomes. However, methylation patterns are not usually found in isolated CpGs, but clusters of proximal CpGs are hypermethylated or hypomethylated (Peters *et al.*, 2015). That is the reason why several methods have been designed to detect differentially methylated regions (DMRs) instead of differentially methylated positions (DMPs).In this context, some methods use predefined regions as candidates for DMRs identification [e.g. gene promoters or CpG Islands (CGIs)], whereas others do not rely on previous annotations and search *de novo* DMRs.

There are two different paradigms related to DNA methylation pointed out in a recent review by Leenen *et al.* (2016). The first one is that, in some disorders such as cancer, regulatory regions are

© The Author(s) 2019. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Downloaded from https://academic.oup.com/bioinformatics/article/35/18/3257/5316232 by Royal College of Surgeons Ireland user on 06 July 2022

clearly hypermethylated or hypomethylated, with methylation differences greater than 60% (see e.g. De Smet *et al.*, 1999; Mikeska and Craig, 2014). However, there is a second paradigm in which complex disorders are associated with very subtle differences in CpGs methylation, with methylation differences of 1–10% between phenotypes. As remarked by Leenen *et al.* these subtle methylation differences are relevant hallmarks associated to the diversity of many complex non-malignant diseases, such as Type 2 diabetes, major depression, schizophrenia, hypertension and cardiovascular diseases (see e.g. Guerrero-Bosagna *et al.*, 2014; Levenson, 2010).

Nevertheless, most of the available DMR methods have focused on detecting large methylation differences between phenotypes. In this context, they have worked properly and they have allowed the discovery of many epigenetic causes of several diseases (Lappalainen and Greally, 2017). However, these tools may fail to detect significant DMRs in complex diseases or heterogeneous phenotypes, wherethere might be small differences among methylation signals but consistent across the analyzed regions and samples. Therefore, no individual CpGs or regions may meet the threshold for statistical significance in many published studies, although there may be biologically meaningful differences (see e.g. Bohlin *et al.*, 2015; Chiavaroli *et al.*, 2015; Gervin *et al.*, 2012; Kim *et al.*, 2017; van Dongen *et al.*, 2015).

In addition, some of these tools average all sites in a given region, but if a significant pattern is associated to a subset of sites it may be underestimated if all sites are analyzed as a block.

This scenario motivated us to develop a new approach based on Gene Set Enrichment analysis (GSEA) (Subramanian *et al.*, 2005), a popular methodology for functional analysis that was specifically designed to avoid some related drawbacks in the field of gene expression. GSEA is able to detect significant gene sets that exhibit strong cross-correlation when differential expression of individual genes is modest from the statistical point of view. GSEA uses a given statistical metric to rank all genes of a genome and applies a weighted Kolmogorov–Smirnov (KS) statistic (Hollander and Wolfe, 1999) to calculate an Enrichment Score (ES). Basically, ES for each set is calculated running through the entire ranked list increasing the score when a gene in the set is encountered and decreasing the score when the gene encountered is not in the analyzed set. ES of this set is the maximum difference from 0. The significance of each ES is calculated permuting the sets and recomputing ES, getting a null distribution for the ES.

We have developed a new R package in which we have implemented a GSEA-based differential methylation analysis where gene sets are defined as sets of CpG sites in predefined regions. This new tool, named *mCSEA* (methylated CpGs Set Enrichment Analysis), is capable to detect subtle but consistent methylation differences in predefined genomic regions from 450 K and EPIC microarrays data. The R package is freely available in Bioconductor repository.

2 Materials and methods

2.1 mCSEA workflow

mCSEA R package consists of five main functions (Fig. 1). The first step is to rank all the CpG probes by differential methylation. As input, a presorted list can be used, but if a matrix of β-values or M-values is provided the rankProbes() function applies *limma* (Ritchie *et al.*, 2015) to fit a linear model and return the t-statistic assigned to each CpG site.

The main *mCSEA* function, mCSEATest(), evaluates the enrichment of CpG sites belonging to the same region in the top positions

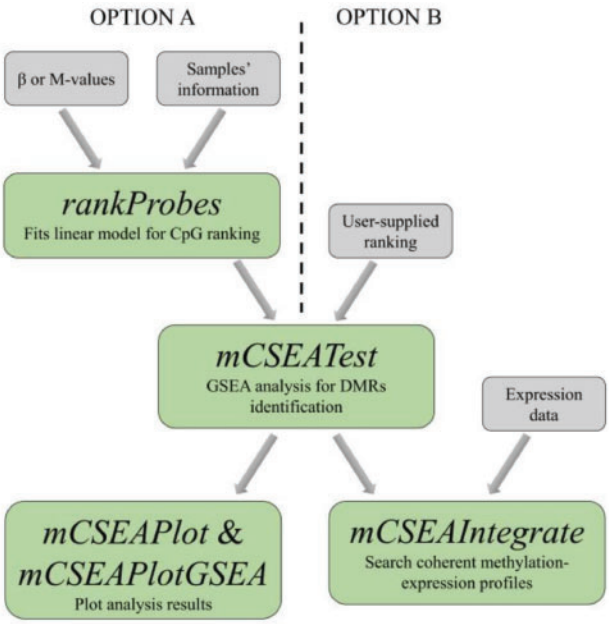

Fig. 1. *mCSEA* workflow. Gray boxes are input data and green boxes are *mCSEA*'s functions. The scheme also shows the order in which functions should be executed

Table 1. Terms from annotation data used for defining each type of region in mCSEA

Region type	Column from annotation data	Terms
Promoters	UCSC_RefGene_Group	TSS1500, TSS200, 5' untranslated region [UTR], 1stExon
Gene bodies	UCSC_RefGene_Group	Body
CGIs	Relation_to_Island	Island, N_Shore, S_Shore, N_Shelf, S_Shelf

of the ranked list by applying the GSEA implementation of the *fgsea* package (Sergushichev, 2016). Regions whose CpG sites are over-represented in the top or bottom of the list can be detected as differentially methylated.

As predefined regions, *mCSEA* allows users to perform analysis based on promoters, gene bodies and CGIs. These predefined regions were defined based on R annotation packages IlluminaHumanMethylation450kanno.ilmn12.hg19 and IlluminaHumanMethylationEPICanno.ilm10b2.hg19 for 450 K and EPIC arrays, respectively. We defined each region as shown in Table 1, following previous works (Sandoval *et al.*, 2011). In addition, researchers can provide a set of defined regions in the analysis by providing a file with genomic positions.

mCSEATest() function provides different statistics for each analyzed region, including a *P*-value of the regions to be differentially methylated, a *P*-value adjusted by false discovery rate (FDR) and the ES. In addition, a Normalized Enrichment Score (NES) is calculated in order to correct the bias for the different region sizes. The necessity and implementation of NES was explained in the original GSEA's paper (Subramanian *et al.*, 2005).

mCSEA package include two functions to visualize the results: mCSEAPlot() and mCSEAPlotGSEA(). The former represents

Downloaded from https://academic.oup.com/bioinformatics/article/35/18/3257/5316232 by Royal College of Surgeons Ireland user on 06 July 2022

methylation values of a given region in its genomic context (see e.g. Fig. 3A). The latter generates GSEA's enrichment plot (see e.g. Fig. 3C), showing the positions of the CpG in a determined region along the entire ranked list.

Finally, the package implements a function, mCSEAIntegrate(), which integrates gene expression data in the analysis. For that purpose, the leading edge CpGs of each region is first defined. The leading edge is the set of CpGs that contributes to the ES of the region, so these CpGs are the most differentially methylated ones. These sites are averaged for each region in each sample. Then, Pearson's correlation coefficient is calculated between each region's methylation and the proximal gene(s)' expression (i.e. genes within 1500 base pairs upstream and downstream from the region). If the integration is performed with promoters, significant negative correlations are returned, due to it has been observed an inverse correlation between promoters' methylation and gene expression (Jones and Baylin, 2002). On the contrary, if the integration is performed in gene bodies, significant positive correlations are returned instead, due to a positive correlation between gene body methylation and expression has been observed (Aran *et al.*, 2011). If the integration is performed in CGIs, both positive and negative significant correlations are returned, due to CGIs can be located in both promoters and gene bodies.

2.2 Methods comparison

In order to test our method, we used both simulated and real data. We simulated 450Kβ-values for 20 samples using the same approach as Peters *et al.* (2015). We randomly selected 714 promoters to be hypermethylated and another 714 promoters to be hypomethylated in 10 samples (cases) compared with the other 10 (controls). Only promoters with at least five associated CpGs were selected. We simulated datasets with a β-value mode differences among phenotypes (Δβ) ranging from 0.9 to 0.05 across promoter CpG sites. We compared *mCSEA*'s performance with state-of-the-art solutions, both predefined [*IMA* (Wang *et al.*, 2012) and *RnBeads* (Assenov *et al.*, 2014)] and *de novo* [*DMRcate* (Peters *et al.*, 2015), *bumphunter* (Jaffe *et al.*, 2012) and *Probe Lasso*(Butcher and Beck, 2015)] algorithms. *IMA* package uses as input raw idat files and not a β-values matrix. Therefore, to compare its approach using the simulated data we implemented the method that is applied by IMA, that is to calculate the median of the methylation values for each predefined region and to apply *limma* to these averaged values. We did not included *COHCAP* package (Warden *et al.*, 2013) due to it restricts the analysis to CGIs. For all methods we used default parameters with the exceptions compiled in Supplementary Table S1.

All results were considered significant using *P*-value adjusted by FDR < 0.05 threshold. For *IMA* and *RnBeads*, we searched for DMRs in promoter regions and we considered as true positives (TPs) those promoters annotated with the actual differentially methylated promoters, and as false positives (FPs) the called regions not annotated with the actual DMRs. For the rest of the methods, due to they return *de novo* DMRs, we considered as TP those actual DMRs overlapping at least one called region, and as FP the called regions not overlapping any actual DMR. For all methods we considered as false negatives (FNs) the actual DMRs not called by the corresponding method.

For each method and Δβ we calculated the sensitivity (Equation 1) and the precision or positive predictive value (PPV) (Equation 2).

$$\text{sensitivity} = \frac{TP}{TP + FN} \times 100 \quad (1)$$

$$\text{precision} = PPV = \frac{TP}{TP + FP} \times 100 \quad (2)$$

We also tested the performance of the proposed method in the methylation datasets previously published by Kim *et al.* (2017). This dataset contains Illumina 450 K methylation data from 18 sibling pairs discordant for intrauterine exposure to maternal gestational diabetes mellitus (GDM). This data are publicly available from GEO database (GEO ID: GSE102177).We reanalyzed the data with *IMA, RnBeads, DMRcate*, *Probe Lasso, bumphunter* and *mCSEA*. We selected these methods because all of them are popular tools for DMRs analysis and allow complex experimental designs with paired samples and covariates, as was our case. *Probe Lasso* does not directly allow paired analysis but we adapted its functions to include it in the comparison.

3 Results

3.1 Comparison of DMRs analysis packages

We performed a functional comparison of *mCSEA* and the most popular R packages used to DMRs analysis from Illumina microarrays data (Table 2). An essential function of this kind of software is the capability to analyze data from complex experimental designs, due to methylation data is very sensitive to environmental factors (Marsit, 2015) and it is important to take into account sex, age, ethnicity and other confounding factors. In addition, some experiments require a paired analysis (e.g. when normal and cancer cells are extracted from the same patient). *mCSEA* can handle with both, covariates adjusting and paired analysis. Other important features compared were the type of regions that can be included in the analysis and the capacity of integrating gene expression data. Our method and *COHCAP* are the only tools capable to integrate gene expression data in the analysis to define genes that show strong correlation in gene expression and methylation data, which is a very relevant feature.

3.2 Simulated data results

We calculated the number of TP, FP and FN returned by each tested method for each Δβ interval, in addition to sensitivity and PPV (Supplementary Table S2). As can be noted in Figure 2, *mCSEA* yielded a 100% of sensitivity detecting methylation differences ranging from Δβ = 0.9 to Δβ = 0.2 and it outperformed the rest of methods when the methylation differences were especially small (0.05). In addition, *mCSEA* returns a low number of FP, resulting in a high PPV for all Δβ (Supplementary Table S2). Only *DMRcate* and *Probe Lasso* overcome *mCSEA* in PPV, but at the cost of having a significantly lower sensitivity for all Δβ.

3.3 DMRs in maternal diabetes exposure discordant siblings

To demonstrate the *mCSEA*'s functionality, we analyzed the data reported by Kim *et al.* (2017). This is a methylation dataset from child sibling pairs: one of the siblings was exposed to maternal diabetes during their gestation, while the other was not. This intrauterine hyperglycemia exposure is associated with an increased risk of obesity and diabetes. Authors collected data from discordant siblings for maternal diabetes exposure in order to get insight into possible epigenetic aberrations in the exposed sibling. Methylation differences in such type of experiment were expected to be very subtle and, in fact, the authors did not report any significant result from the statistical point of view (FDR < 0.05), but they focused in the most differentially methylated genes and discussed their biological relevance.

Downloaded from https://academic.oup.com/bioinformatics/article/35/18/3257/5316232 by Royal College of Surgeons Ireland user on 06 July 2022

Table 2. Comparison of available R packages for DMRs analysis using Illumina's microarray data

	IMA	RnBeads	DMRcate	Bumphunter	COHCAP	Probe Lasso[a]	mCSEA
References	Wang *et al.* (2012)	Assenov *et al.* (2014)	Peters *et al.* (2015)	Jaffe *et al.* (2012)	Warden *et al.* (2013)	Butcher and Beck (2015)	—
DMRs analyzed	Predefined	Predefined	*De novo*	*De novo*	Predefined	*De novo*	Predefined
Platforms	27K and 450K	27K and 450K	450K and EPIC	27K, 450K and EPIC	27K and 450K[2]	450K and EPIC	450K and EPIC[b]
Statistical test	Wilcoxon rank-sum, *t*-test and empirical Bayes	CpG-level *P*-values aggregation with Fisher's method	Kernel smoothing	Bumphunter algorithm	Analysis of variance	Probe Lasso algorithm	GSEA
Accepts methylation matrix as input	No	Yes	Yes	Yes	Yes	Yes	Yes
Adjusting for covariates	Yes	Yes	Yes	Yes	Only one[c]	No	Yes
Paired analysis	Yes	Yes	Yes	Yes	Yes[c]	No	Yes
Implemented parallelization	No	Yes	Yes	Yes	No	Yes	Yes
Integration of Gene Expression Data	No	No	No	No	Yes	No	Yes
Predefined Regions	UCSC-defined regions (TSS1500, 5' UTR, gene body...)	Promoters, gene bodies, CGIs, tilling regions, user-defined regions	—	—	CGIs	—	Promoters, gene bodies, CGIs, user-defined regions

[a]Implemented in ChAMP package (Morris *et al.*, 2014).
[b]Other platforms can be analyzed introducing custom annotations.
[c]It is only possible to adjust for one covariate or to perform a paired analysis, but not both.

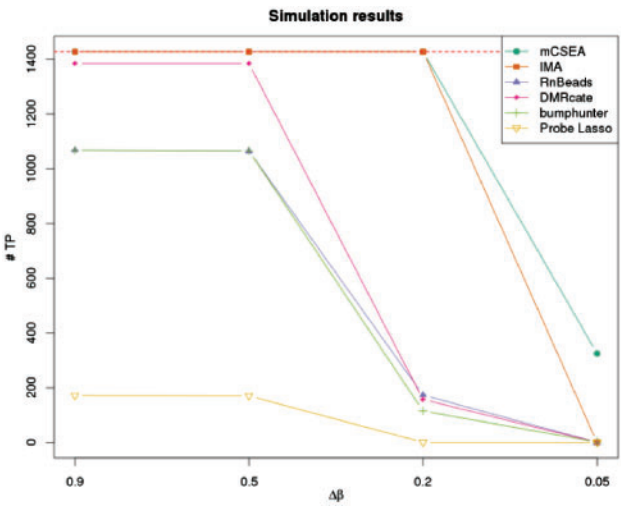

Fig. 2. Performance with simulated data. Each line represents results from different methods. The Y-axis represents the number of TP for each Δβ. Red line represents the total number of TP included in the dataset (1428)

In this dataset, *DMRcate* and *Probe Lasso* did not return any significant DMR. These methods applied *limma* to detect significant DMPs and call DMRs based on them. Although they work properly when methylation differences are high they did not reveal any significant result for slight methylation differences.

RnBeads is also based on *limma* or detecting DMRs, but it combines the results by region types (promoters, CGIs and so on) aggregating the *P*-values obtained by the linear modeling, so, even if there are not any significant DMPs, *RnBeads* is potentially capable to find significant DMRs. However, this was not the case. This method did not return any significant DMR (FDR < 0.05).*IMA* approach did not return any significant DMR neither.

Bumphunter yielded one significant DMR (FDR = 0.03, Family-wise error rate (FWER) = 0.01) located at the promoter of SDHAP3 pseudogene. Up to our knowledge, there is not any known relationship between SDHAP3 and development or metabolic disorders.

mCSEA yielded 1055 significant DMRs (FDR < 0.05) in gene promoters: 228 hypermethylated and 827 hypomethylated promoters in cases compared with controls (Supplementary Table S3).

To assess the biological significance of these results, we performed an enrichment analysis using Enrichr (Chen *et al.*, 2013). The most significant enriched pathway in KEGG database (Kanehisa and Goto, 2000) is 'Maturity onset diabetes of the young' (hsa04950) pathway (adjusted *P*-value = 0.0011) (Supplementary Table S4). This pathway is related with a type of diabetes characterized to appear in patients younger than 25-years old and to be non-insulin dependent. Promoter regions of 9 out of the 26 genes associated to this pathway were identified as significantly differential methylated regions, including PDX1, FOXA2, PAX6 or INS. INS gene, which we found to be hypermethylated in cases, is an important gene that has been previously associated to diabetes in several works and it has been reported as a silenced gene with a fully methylated promoter associated to diabetes development (Yang *et al.*, 2011). In addition, it has been observed that high levels of glucose increase the INS methylation level (Yang *et al.*, 2011), so this hypermethylation could be induced during gestation. Methylation differences in INS promoter between children exposed and non-exposed to intrauterine hyperglycemia are subtle, but consistent across all CpG sites of the promoter (Fig. 3). Such small methylation

Downloaded from https://academic.oup.com/bioinformatics/article/35/18/3257/5316232 by Royal College of Surgeons Ireland user on 06 July 2022

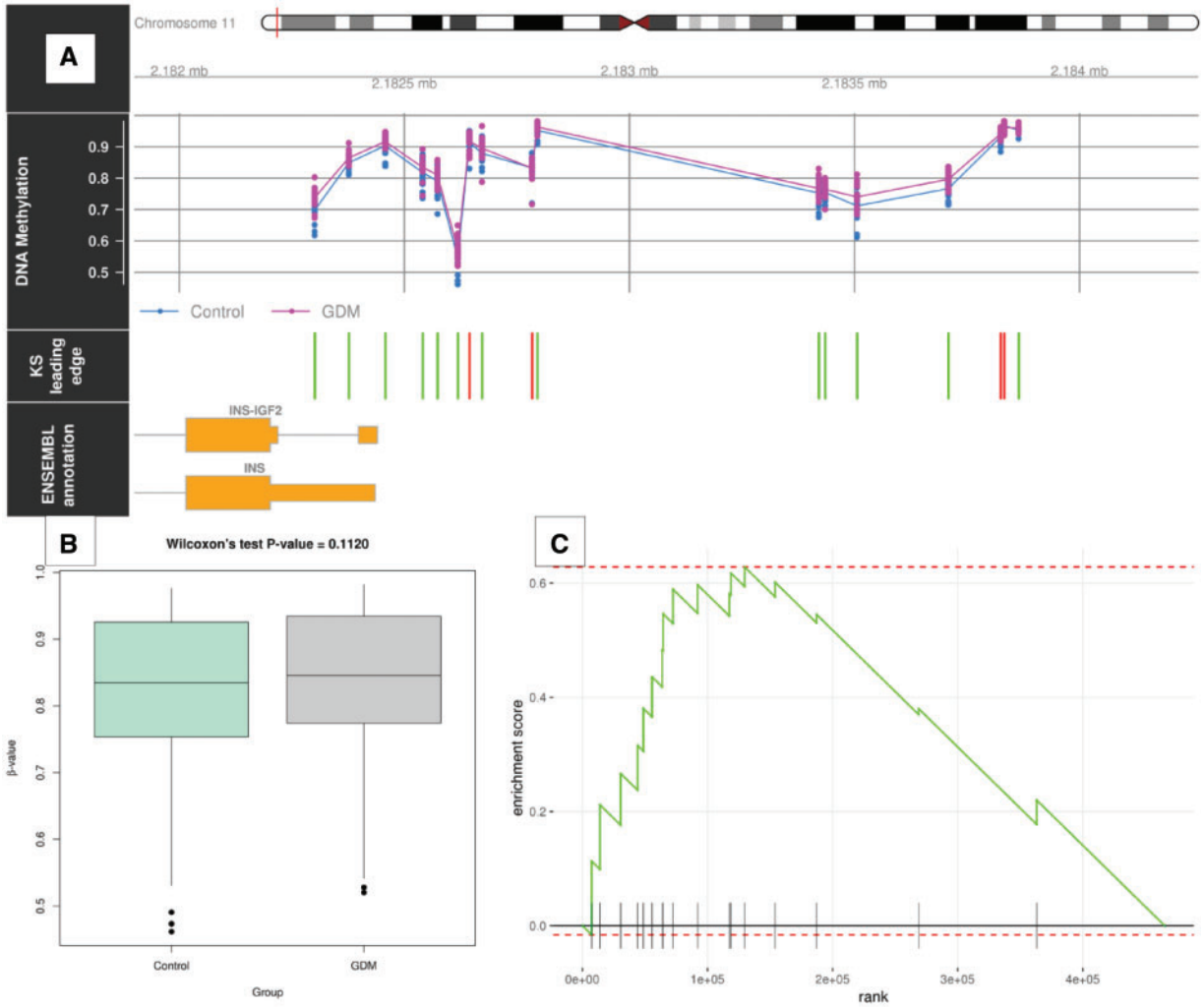

Fig. 3. INS gene promoter methylation in GDM and control samples. Methylation is quantified with β-values. **(A)** Genomic context of INS promoter. Each point represents the methylation of each sample. Lines link the mean methylation of each group. KS leading edge panel marks with green bars those CpGs contributing to the ES and with red bars the rest of them. This plot was obtained with mCSEAPlot() function, implemented in *mCSEA* package. **(B)** Boxplot showing the subtle difference in INS promoter methylation status between controls and GDM samples. **(C)** GSEA plot for INS promoter. Vertical lines mark the location of INS-associated CpGs along the entire ranked list of analyzed CpGs (horizontal black line). Red lines represent the maximum and minimum ES. This plot was obtained with mCSEAPlotGSEA() function, implemented in *mCSEA* package

difference is the cause why this DMR remains undetected by all the other tested methods. The same may be occurring in many other genomic regions.

On the other hand, the most significant enriched pathway from Online Mendelian Inheritance in Man Disease database is obesity (adjusted P-value = 0.0085) (Supplementary Table S5). Eight out of fifteen genes related to this disease contained significant DMRs, including UCP1, UCP3, GHRL or PCSK1. So, we found methylation alterations in genes related to diabetes and obesity, the two main diseases associated to intrauterine GDM exposure.

4 Conclusions

Here we present *mCSEA*, a novel R package for predefined DMRs detection based on GSEA method. We compared *mCSEA* with the most widely used methods to detect DMRs. Our method outperformed the rest of solutions for detecting small methylation differences in the simulated dataset. It is especially remarkable the capability of *mCSEA* to find DMRs even with the methylation difference as small as 0.05 between groups, but consistent along a relatively large region. We re-analyzed a previously published dataset, obtaining barely any significant results with other methods. However, *mCSEA* yielded several significant DMRs in promoters for genes associated to relevant biological pathways.

We think that *mCSEA* will provide researchers with a useful tool to detect DMRs in datasets from complex diseases in which the methylation differences among phenotypes are small but consistent.

Acknowledgements

This work is part of the JMM's PhD results. J.M.M. is enrolled in the PhD program in Biomedicine at the University of Granada, Spain.

Funding

J.M.M. is partially funded by Ministerio de Economía, Industria y Competitividad. This work is partially funded by Consejería de Salud, Junta de Andalucía (Grant PI-0152–2017).

Conflict of Interest: none declared.

Downloaded from https://academic.oup.com/bioinformatics/article/35/18/3257/5316232 by Royal College of Surgeons Ireland user on 06 July 2022

References

Aran,D. *et al.* (2011) Replication timing-related and gene body-specific methylation of active human genes. *Hum. Mol. Genet.*, **20**, 670–680.

Assenov,Y. *et al.* (2014) Comprehensive analysis of DNA methylation data with RnBeads. *Nat. Methods*, **11**, 1138–1140.

Bohlin,J. *et al.* (2015) Effect of maternal gestational weight gain on offspring DNA methylation: a follow-up to the ALSPAC cohort study. *BMC Res. Notes*, **8**, 321.

Butcher,L.M. and Beck,S. (2015) Probe Lasso: a novel method to rope in differentially methylated regions with 450K DNA methylation data. *Methods*, **72**, 21–28.

Chen,E.Y. *et al.* (2013) Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, **14**, 128.

Chiavaroli,V. *et al.* (2015) Infants born large-for-gestational-age display slower growth in early infancy, but no epigenetic changes at birth. *Sci. Rep.*, **5**, 14540.

De Smet,C. *et al.* (1999) DNA methylation is the primary silencing mechanism for a set of germ line- and tumor-specific genes with a CpG-rich promoter. *Mol. Cell Biol.*, **19**, 7327–7335.

Flanagan,J.M. (2015) Epigenome-wide association studies (EWAS): past, present, and future. *Methods Mol. Biol.*, **1238**, 51–63.

Gervin,K. *et al.* (2012) DNA methylation and gene expression changes in monozygotic twins discordant for psoriasis: identification of epigenetically dysregulated genes. *PLoS Genet.*, **8**, e1002454.

Guerrero-Bosagna,C. *et al.* (2014) Identification of genomic features in environmentally induced epigenetic transgenerational inherited sperm epimutations. *PLoS One*, **9**, e100194.

Hollander,M. and Wolfe,D.A. (1999) *Nonparametric Statistical Methods*. Wiley, New York.

Jaffe,A.E. *et al.* (2012) Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int. J. Epidemiol.*, **41**, 200–209.

Jones,P.A. and Baylin,S.B. (2002) The fundamental role of epigenetic events in cancer. *Nat. Rev. Genet.*, **3**, 415–428.

Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Kim,E. *et al.* (2017) DNA methylation profiles in sibling pairs discordant for intrauterine exposure to maternal gestational diabetes. *Epigenetics*, **12**, 825–832.

Lappalainen,T. and Greally,J.M. (2017) Associating cellular epigenetic models with human phenotypes. *Nat. Rev. Genet.*, **18**, 441–451.

Leenen,F.A.D. *et al.* (2016) DNA methylation: conducting the orchestra from exposure to phenotype? *Clin. Epigenetics*, **8**, 92.

Levenson,V.V. (2010) DNA methylation as a universal biomarker. *Expert Rev. Mol. Diagn.*, **10**, 481–488.

Marsit,C.J. (2015) Influence of environmental exposure on human epigenetic regulation. *J. Exp. Biol.*, **218**, 71–79.

Mikeska,T. and Craig,J.M. (2014) DNA methylation biomarkers: cancer and beyond. *Genes (Basel)*, **5**, 821–864.

Peters,T.J. *et al.* (2015) De novo identification of differentially methylated regions in the human genome. *Epigenetics Chromatin*, **8**, 6.

Ritchie,M.E. *et al.* (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.

Sandoval,J. *et al.* (2011) Validation of a DNA methylation microarray for 450, 000 CpG sites in the human genome. *Epigenetics*, **6**, 692–702.

Sergushichev,A. (2016) An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. https://doi.org/10.1101/060012.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, **102**, 15545–15550.

Teh,A.L. *et al.* (2016) Comparison of methyl-capture sequencing vs. infinium 450K methylation array for methylome analysis in clinical samples. *Epigenetics*, **11**, 36–48.

van Dongen,J. *et al.* (2015) Epigenome-wide association study of aggressive behavior. *Twin Res. Hum. Genet.*, **18**, 686–698.

Wang,D. *et al.* (2012) IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data. *Bioinformatics*, **28**, 729–730.

Warden,C.D. *et al.* (2013) COHCAP: an integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis. *Nucleic Acids Res.*, **41**, e117.

Yang,B.T. *et al.* (2011) Insulin promoter DNA methylation correlates negatively with insulin gene expression and positively with HbA1c levels in human pancreatic islets. *Diabetologia*, **54**, 360–367.

Downloaded from https://academic.oup.com/bioinformatics/article/35/18/3257/5316232 by Royal College of Surgeons Ireland user on 06 July 2022