Journal Club

Deepak Tanwar, Hana Parizkova

@JC-STA426-20181022

22 th October, 2018

Detection and accurate false discovery rate control of differentially methylated regions from whole genome bisulfite sequencing

Keegan Korthauer, Sutirtha Chakraborty, Yuval Benjamini, Rafael A Irizarry

Biostatistics, kxy007, https://doi.org/10.1093/biostatistics/kxy007

Published: 22 February 2018

Table of Contents

Epigenetics

A layer of **additional information over genetic information**.

Epigenetics

A layer of **additional information over genetic information**.

Study of stable heritable changes in gene expression that occur without changes in DNA sequence.

Epigenetics

A layer of **additional information over genetic information**.

Study of stable heritable changes in gene expression that occur without changes in DNA sequence.

Epigenetics modification/ marks = Punctuation marks

Epigenetics

A layer of **additional information over genetic information**.

Study of stable heritable changes in gene expression that occur without changes in DNA sequence.

Epigenetics modification/ marks = Punctuation marks

`bioinformaticiansarenotantisocialwearejustgenomefriendly`

Epigenetics

A layer of **additional information over genetic information**.

Study of stable heritable changes in gene expression that occur without changes in DNA sequence.

Epigenetics modification/ marks = Punctuation marks

`bioinformaticiansarenotantisocialwearejustgenomefriendly`

Bioinformaticians are not anti-social; we are just genome friendly.

Epigenetics

A layer of **additional information over genetic information**.

Study of stable heritable changes in gene expression that occur without changes in DNA sequence.

Epigenetics modification/ marks = Punctuation marks

`bioinformaticiansarenotantisocialwearejustgenomefriendly`

Bioinformaticians are not anti-social; we are just genome friendly.

`execute not let free`

Epigenetics

A layer of **additional information over genetic information**.

Study of stable heritable changes in gene expression that occur without changes in DNA sequence.

Epigenetics modification/ marks = Punctuation marks

`bioinformaticiansarenotantisocialwearejustgenomefriendly`

Bioinformaticians are not anti-social; we are just genome friendly.

`execute not let free`

Execute, not let free!

Epigenetics

A layer of **additional information over genetic information**.

Study of stable heritable changes in gene expression that occur without changes in DNA sequence.

Epigenetics modification/ marks = Punctuation marks

`bioinformaticiansarenotantisocialwearejustgenomefriendly`

Bioinformaticians are not anti-social; we are just genome friendly.

`execute not let free`

Execute not, let free!

Epigenetics

A layer of **additional information over genetic information**.

Study of stable heritable changes in gene expression that occur without changes in DNA sequence.

Epigenetics modification/ marks = Punctuation marks

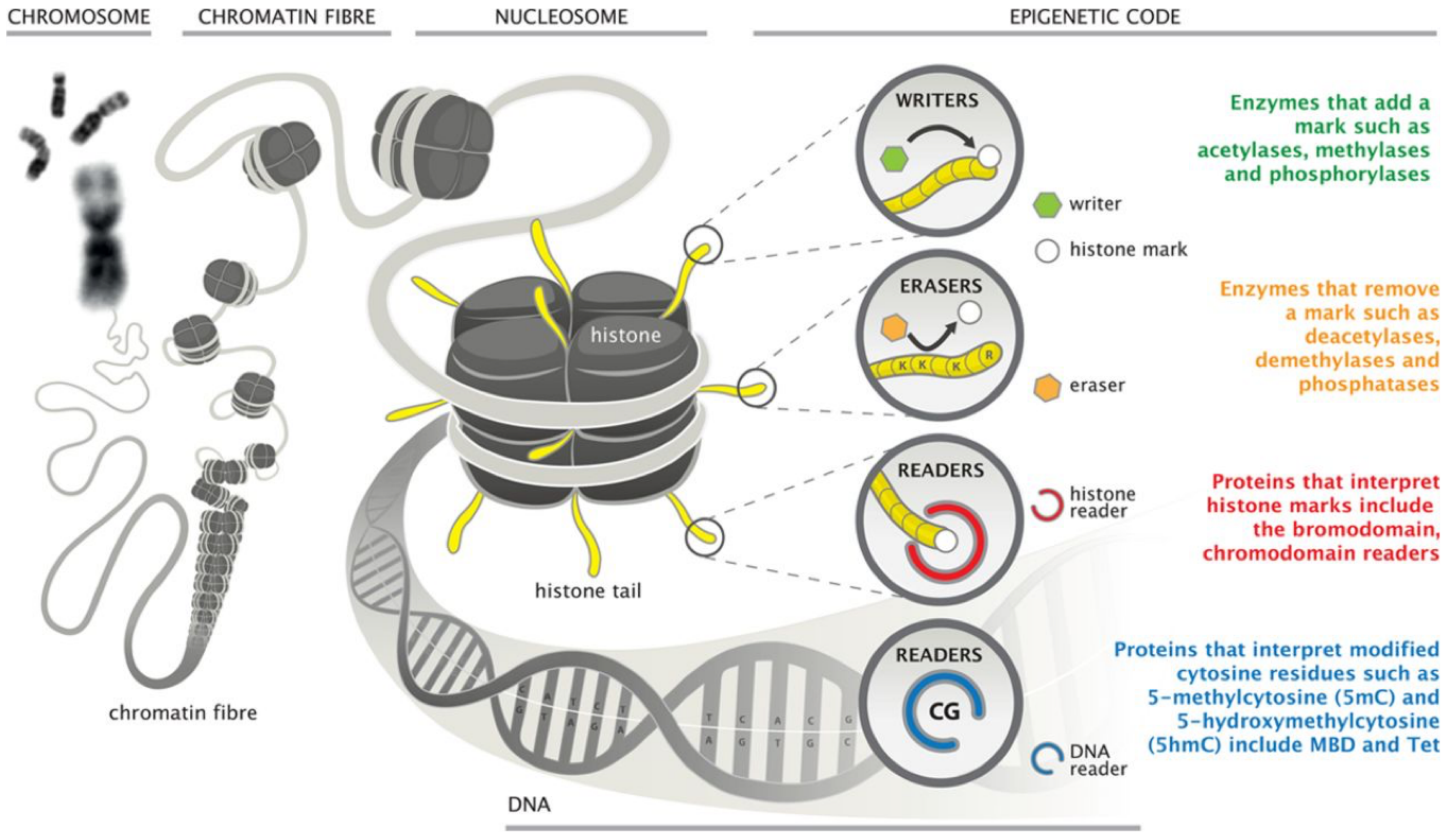
`bioinformaticiansarenotantisocialwearejustgenomefriendly`

Bioinformaticians are not anti-social; we are just genome friendly.

`execute not let free`

Execute not, let free!

Various Epigenetic Modifications

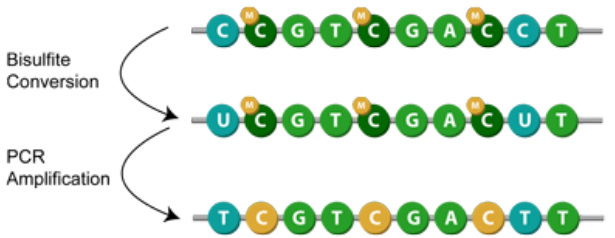


Samuel T. Keating et al. Circ Res. 2016;118:1706-1722

American Heart Association.

Whole Genome Bisulphite Sequencing (WGBS)

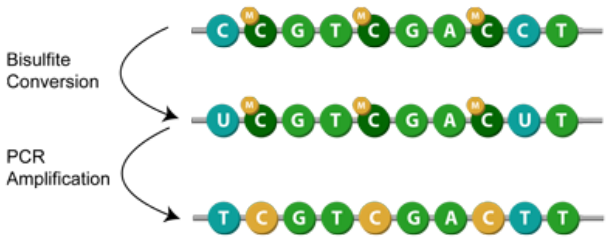
Bisulphite conversion



ref

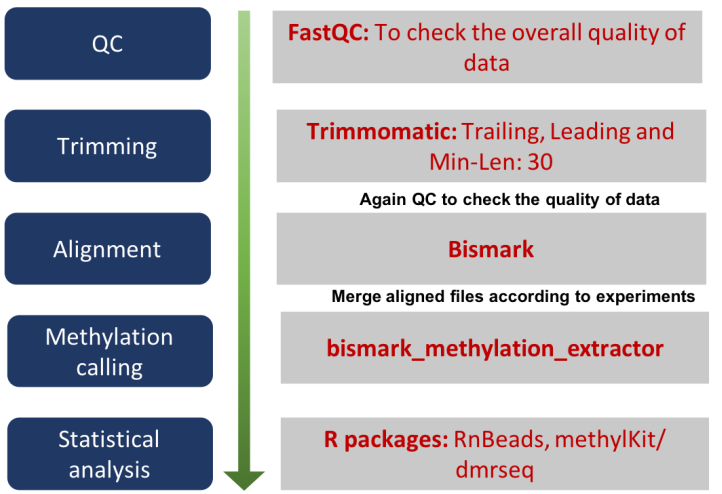
Whole Genome Bisulphite Sequencing (WGBS)

Bisulphite conversion



ref

Basic Data analysis pipeline



Compared to methods developed for RNA-Seq, there are not that many statistical methods available for studying WGBS

Applications of WGBS

- Detection of Differentially Methylated Regions (**DMRs)**
- Detection of Differentially Methylated Loci (**DMLs)**
- Detection of Copy Number Variations (CNVs)
- Detection of Single Nucleotide Polymorphisms (SNPs)
- Detection of Cytosine methylation levels of Transcription Factor Binding Sites (TFBSs)

....cont

Limitations of current methods

- Methods are greatly hindered by low sample size. Most of the methods depend on large sample sizes.

Limitations of current methods

- Methods are greatly hindered by low sample size. Most of the methods depend on large sample sizes.
- The number of tests performed = the number of loci analyzed. **Example:** Human genome has ~30 million CpGs [1].

Limitations of current methods

- Methods are greatly hindered by low sample size. Most of the methods depend on large sample sizes.
- The number of tests performed = the number of loci analyzed. **Example:** Human genome has ~30 million CpGs [1].
- Measurements are spatially correlated across the genome [2], but in most methods measurements from all loci are treated independently.

Limitations of current methods

- Methods are greatly hindered by low sample size. Most of the methods depend on large sample sizes.
- The number of tests performed = the number of loci analyzed. **Example:** Human genome has ~30 million CpGs [1].
- Measurements are spatially correlated across the genome [2], but in most methods measurements from all loci are treated independently.
- Muliple testing corrections without considering the spatial correlation can result in a loss of power.

Limitations of current methods

- Methods are greatly hindered by low sample size. Most of the methods depend on large sample sizes.
- The number of tests performed = the number of loci analyzed. **Example:** Human genome has ~30 million CpGs [1].
- Measurements are spatially correlated across the genome [2], but in most methods measurements from all loci are treated independently.
- Muliple testing corrections without considering the spatial correlation can result in a loss of power.
- DMRs more biologically relevant than DMLs.

....cont.

- DML approaches may construct DMRs by chaining together neighboring significant loci, but this type of approach will not yield a proper assessment of the statistical significance of the constructed regions, nor will the FDR be properly controlled [4]. Controlling the FDR at the level of individual loci is not the same as controlling FDR of regions.

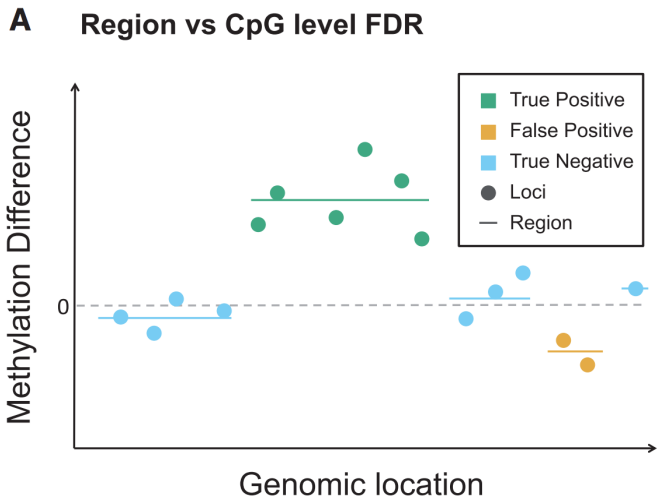
....cont.

- DML approaches may construct DMRs by chaining together neighboring significant loci, but this type of approach will not yield a proper assessment of the statistical significance of the constructed regions, nor will the FDR be properly controlled [4]. Controlling the FDR at the level of individual loci is not the same as controlling FDR of regions.
- Methylation data cannot be modeled neither by Gaussian models (due to low coverage), nor by Binomial models (do not account for biological variability). Beta-Binomial models are computationaly difficult.

....cont.

- DML approaches may construct DMRs by chaining together neighboring significant loci, but this type of approach will not yield a proper assessment of the statistical significance of the constructed regions, nor will the FDR be properly controlled [4]. Controlling the FDR at the level of individual loci is not the same as controlling FDR of regions.
- Methylation data cannot be modeled neither by Gaussian models (due to low coverage), nor by Binomial models (do not account for biological variability). Beta-Binomial models are computationaly difficult.
- Challenges for assessing DMRs:
    - defining region boundaries
    - methods ignore correlation across loci
    - biological variability from sample to sample

FDR at the loci level is not the same as FDR at the region level



Genomic location versus methylation difference estimates at several neighboring loci.

The individual CpGs (points) are shaded by whether they are a true or false positive. Regions are denoted by lines.

The loci FDR is:

FDRloci = (#False Positive Loci) / (Total # of Significant Loci)

The region FDR is:

FDRregion=(#False Positive Regions) / (Total # of Significant Regions)

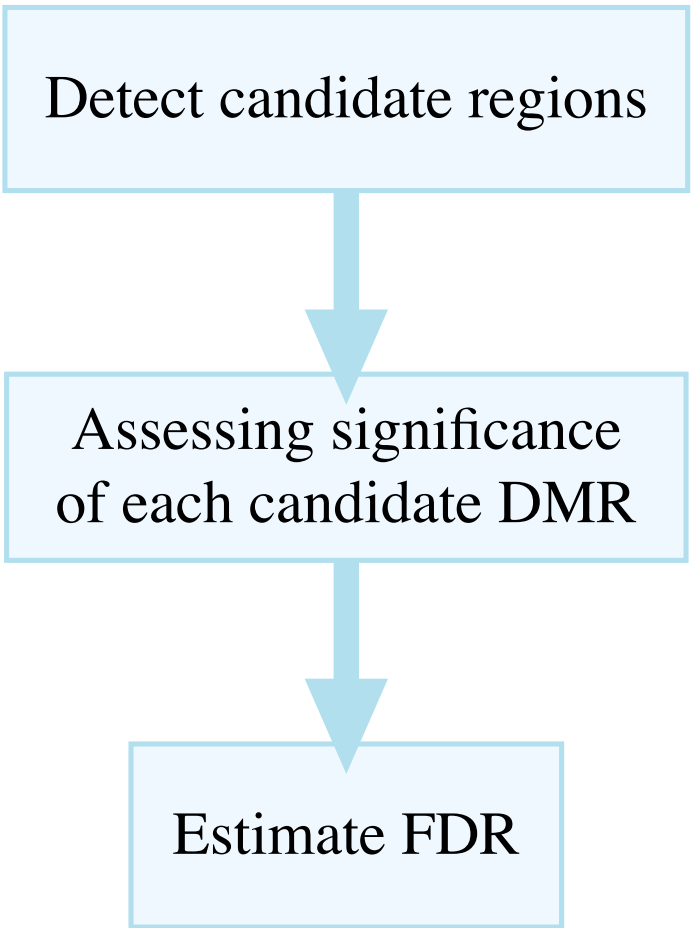
dmrseq

Aim:

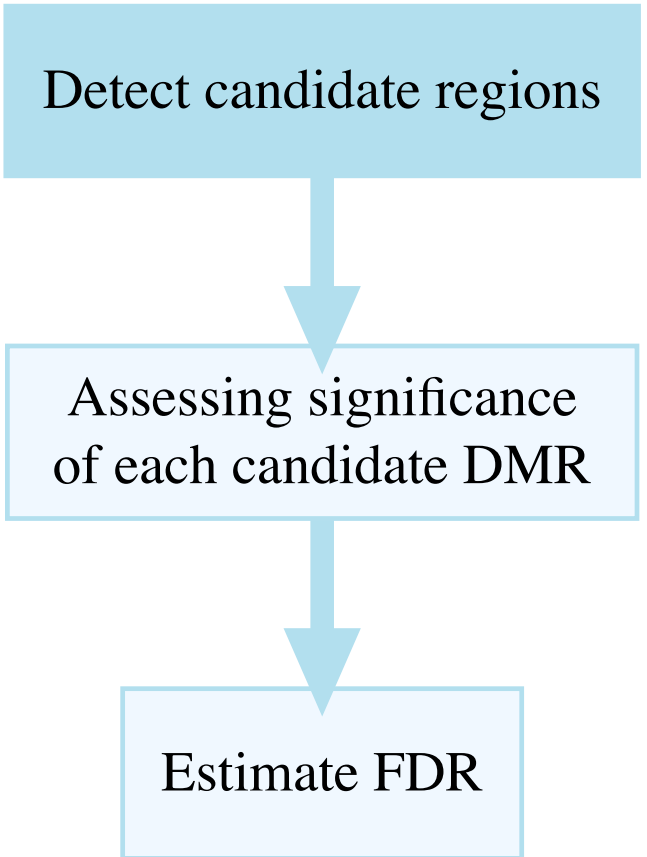
1. To **maximize power** while **controlling the FDR** even with sample sizes as small as **two samples per condition**.

1. To develop a procedure to control FDR at the region level and provide an accurate measure of statistical significance for each region.

Identification of DMRs by dmrseq



Detecting candidate regions



- mean methylation proportion for locus i in condition s:

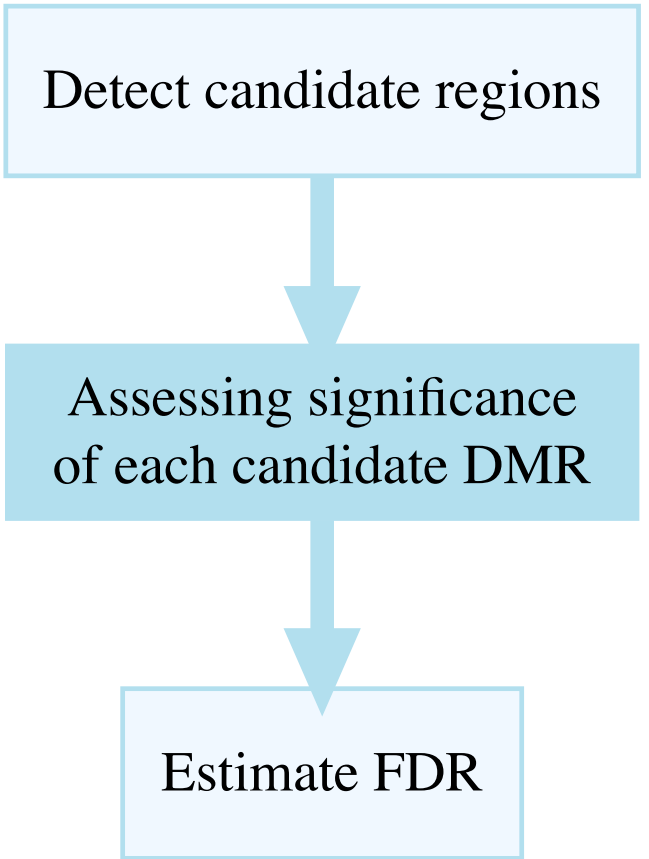
$$\pi_{is} = M_{is}/N_{is}$$

- methylation difference of locus i between biological conditions s and s':

$$\beta_i = \pi_{is} - \pi_{is'}$$

- β_i's are smoothed
- candidate regions: segments of genomes with smoothed β_i's in the same direction and in absolute value greater than a threshold (e.g. 0.1)

Assessing significance of candidate DMRs



- methylation counts modeled by Beta-Binomial distribution
- to stabilize the dependence of variance on mean, data are transformed:

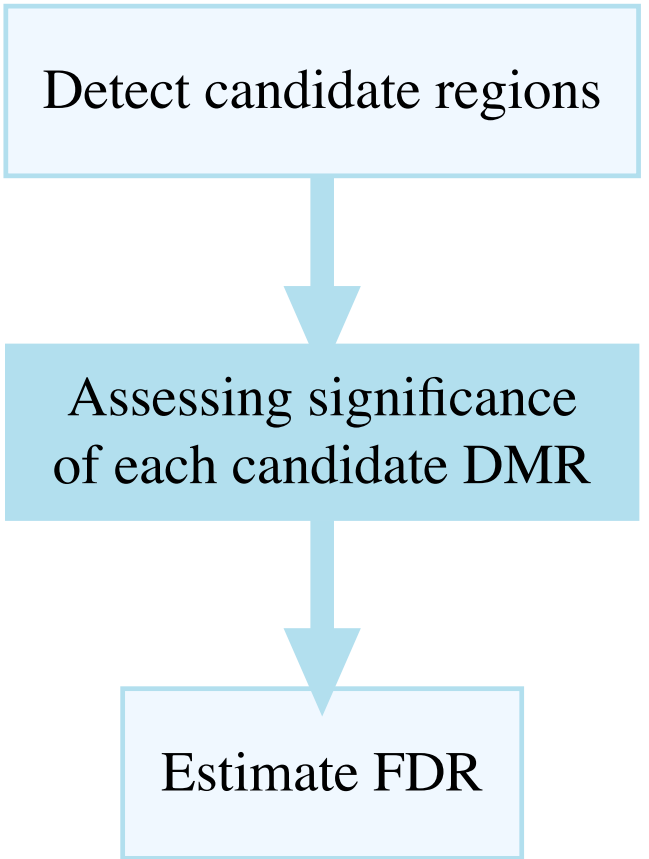
$$Z_{ijr} = \arcsin\left(2M_{ijr}/N_{ijr} - 1\right)$$

- mean region methylation level can be then estimated using Generalized Least Squares:

$$Z_r = X\beta_{0r} + \beta_{1r},$$

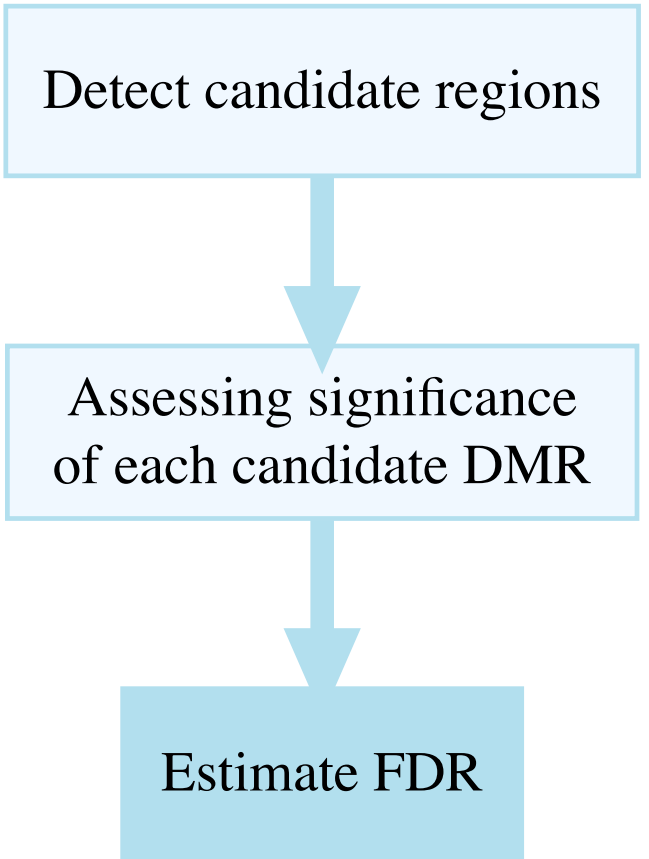
with β_{1r} being the effect of biological group

Assessing significance of candidate DMRs



- Estimation of covariance matrix V_r:
  - variance dependent on coverage
    $\rightarrow Var(Z_{ijr}) \approx \sigma_r^2 / N_{ijr}$
  - correlation of nearby loci $\rightarrow$ modeled correlation by coefficients dependent of the lenght of the interval between the two loci

Estimating FDR



- permutation to get an approximate null distribution
- p-value computed by comparing the observed test statistics to the null distribution

Data used for comparison

- six samples of normal human dendritic cells, partitioned into groups of two or three $\rightarrow$ no DMRs, **negative control**

Data used for comparison

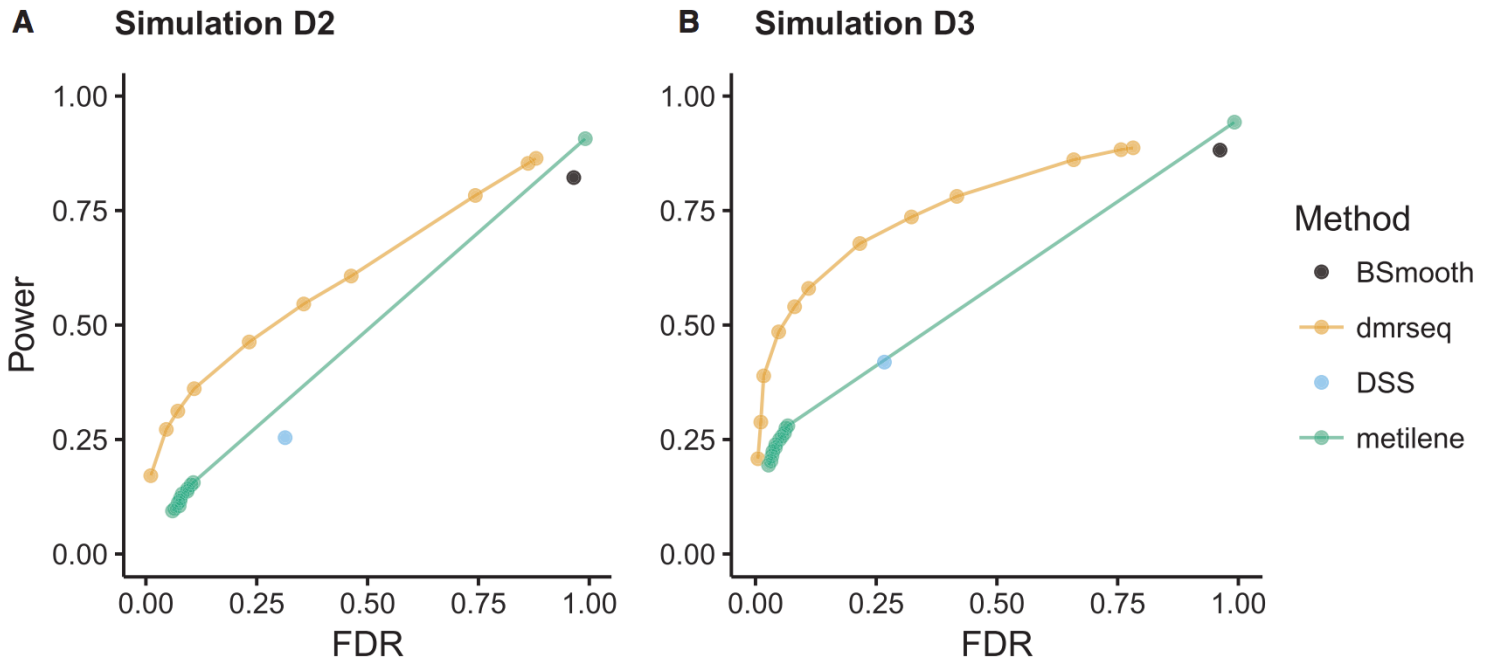
- six samples of normal human dendritic cells, partitioned into groups of two or three $\rightarrow$ no DMRs, **negative control**
- **simulated data:** 3000 simulated DMRs added to each data set
  - DMR is created by choosing a cluster of CpGs and simulating the number of methylated reads from a binomial distribution
  - binomial probabilities equal to the observed methylation proportions $\pm$ effect size (approx. 0.1 to 0.5)

Data used for comparison

- six samples of normal human dendritic cells, partitioned into groups of two or three $\rightarrow$ no DMRs, **negative control**
- **simulated data:** 3000 simulated DMRs added to each data set
  - DMR is created by choosing a cluster of CpGs and simulating the number of methylated reads from a binomial distribution
  - binomial probabilities equal to the observed methylation proportions $\pm$ effect size (approx. 0.1 to 0.5)
- **real data:** methylation data from several human tissues and from two mouse strains

dmrseq is more powerful

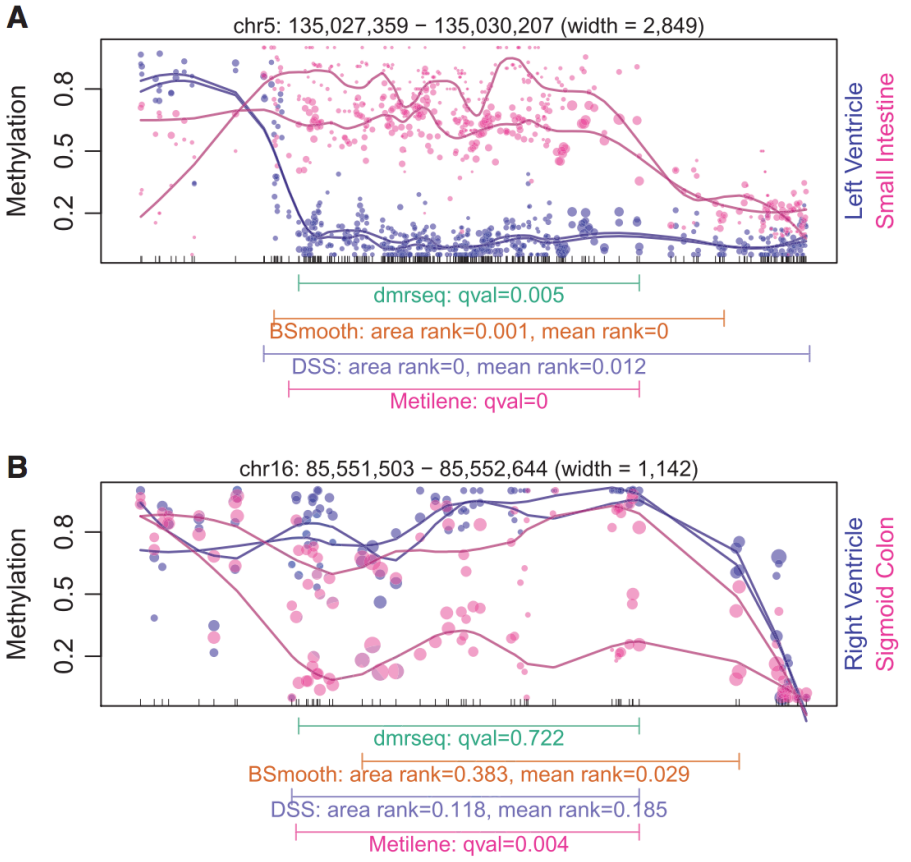
power vs. FDR on simulated data



Power is calculated as the proportion of simulated DMRs overlapped by at least one identified DMR. FDR is calculated as the proportion of DMRs identified that do not overlap with any of the simulated DMRs.

dmrseq performs better

Examples of real biological data and their ranking by different DMRs identification methods.



dmrseq availability

- GitHub: kdkorthauer/dmrseq
- Bioconductor: dmrseq

Summary

- Powerful even in the case of 2 samples per group.

Summary

- Powerful even in the case of 2 samples per group.
- Consideration of all possible shorcomings of the available methods for DMRs detection.

Summary

- Powerful even in the case of 2 samples per group.
- Consideration of all possible shorcomings of the available methods for DMRs detection.
- Evaluation of the software using experimental data and Monte Carlo simulations.

Summary

- Powerful even in the case of 2 samples per group.
- Consideration of all possible shorcomings of the available methods for DMRs detection.
- Evaluation of the software using experimental data and Monte Carlo simulations.
- Implementation in **R**.

Summary

- Powerful even in the case of 2 samples per group.
- Consideration of all possible shorcomings of the available methods for DMRs detection.
- Evaluation of the software using experimental data and Monte Carlo simulations.
- Implementation in **R**.
- Comparison of the new method with the "famous" methods.

Summary

- Powerful even in the case of 2 samples per group.
- Consideration of all possible shorcomings of the available methods for DMRs detection.
- Evaluation of the software using experimental data and Monte Carlo simulations.
- Implementation in **R**.
- Comparison of the new method with the "famous" methods.
- Comparison is not done with **RnBeads**. Probably because RnBeads performs differential methylation analysis (only) at single nucleotide level.
- Comparison is not done with **methylKit**, although methylKit performs differential methylation analysis at regional level.

References

[1] Smith Z. D. and Meissner A. (2013). DNA methylation: roles in mammalian development. Nature Reviews Genetics 14, 204–220.

[2] Leek J. T., Scharpf R. B., Bravo H. C., Simcha D., Langmead B., Johnson W. E., Geman D., Baggerly K. and Irizarry R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics 11, 733–739.

[3] Aryee M. J., Jaffe A. E., Corrada-Bravo H., Ladd-Acosta C., Feinberg A. P., Hansen K. D. and Irizarry R. A. (2014). Minfi: a flexible and comprehensive Bioconductor package for the analysis of infinium DNA methylation microarrays. Bioinformatics 30, 1363–1369.

[4] Robinson M. D., Kahraman A., Law C. W., Lindsay H., Nowicka M., Weber L. M. and Zhou X. (2014). Statistical methods for detecting differentially methylated loci and regions. Frontiers in Genetics 5, 324.

Thank you!

Slides are prepared using xaringan R package