DNA methylation

Data generation and pre-processing

University of BRISTOL
BBSRC bioscience for the future
E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
MRC Integrative Epidemiology Unit

DNA methylation dataset

	sample1	sample2	sample3
cg07881041	0.81	0.83	0.80
cg18478105	0.02	0.02	0.02
cg23229610	0.91	0.92	0.90

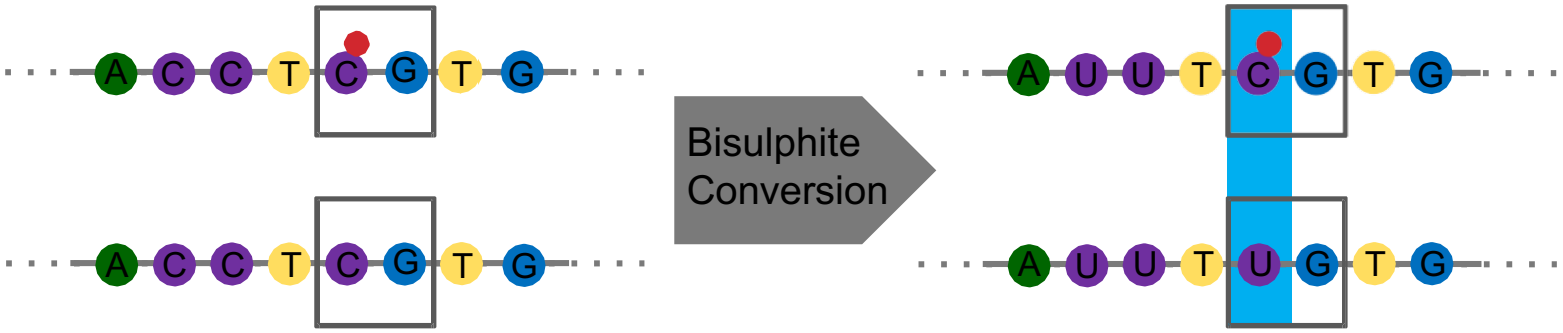
90% of cells from sample3 have a methyl group at CpG cg23229610.

What is an EWAS?

An Epigenome-Wide Association Study (EWAS) is what is sounds like:

- Epi = Epigenetic marks *(specifically DNA methylation).*
- Genome-Wide = Measured at specific locations across the genome.
- Association Study = Test associations between these marks at each genome location and some trait or exposure.

Bisulfite conversion



(During amplification, the U's are replaced with T's.)

We typically follow bisulfite conversion with analysis by:

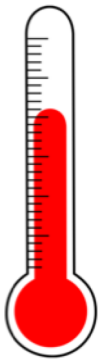
1. Illumina Bead Chip
2. DNA sequencing

Batch effects

Batch effects are systematic variation in the data due to technical factors in how the data was generated

e.g.

- Chips that were run on separate days
- Bisulfite modifications that were performed in different batches



QC and pre-processing solutions

There are a variety of R packages available for performing quality control and preprocessing datasets:

1. Normalization and quality control
   e.g. meffil, minfi, wateRmelon, missMethyl, methylumi
2. Post-normalization batch removal
   e.g. ComBAT (Johnson *et al. Biostatistics,* 2007)

When effects are unknown, including the top principal components in regression models is possible but could introduce colliders.

An uncooperative genome

- Probes for sites on the sex chromosomes
- Probes that are ‘non-specific’
- Probes whose functioning affected by SNPs

Published lists:

- *Naeem et al. BMC Genomics. 2014*
- *Chen et al. Epigenetics. 2013*

Outliers

- Methylation data outliers can drive associations
- Possibly due to technical errors or rare genetic variants

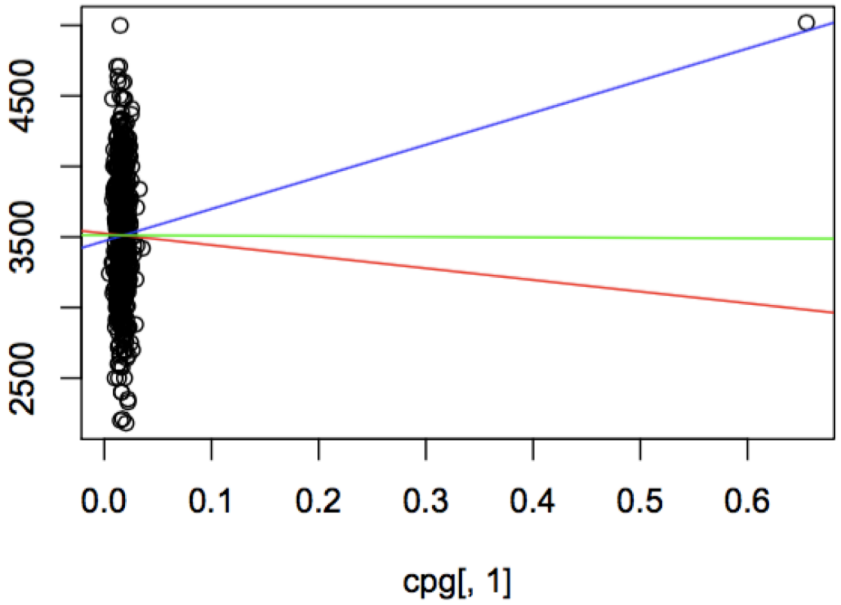
- Handling outliers

  Winsorizing: setting values outside percentiles 0.05-0.95 to the closest value within the percentiles

  Tukey method: remove any value

  < first quartile – 3IQR, or

  > third quartile + 3IQR



Confounding

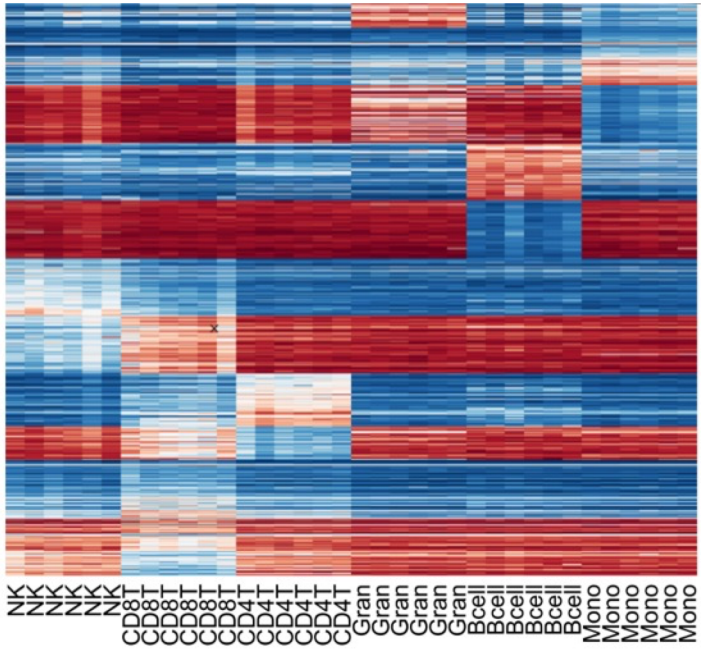
DNA methylation can be influenced by environmental exposures, disease, other phenotypes and genetic factors.

Can be handled by either:

- Removing effects from data prior to analysis
- Including effects in statistical models
- Stratifying

Cellular heterogeneity

Blood (and many other tissues) contains many cell types. Cell types are differently methylated. In fact, DNA methylation is a key mechanism for initiating and maintaining cell identity.

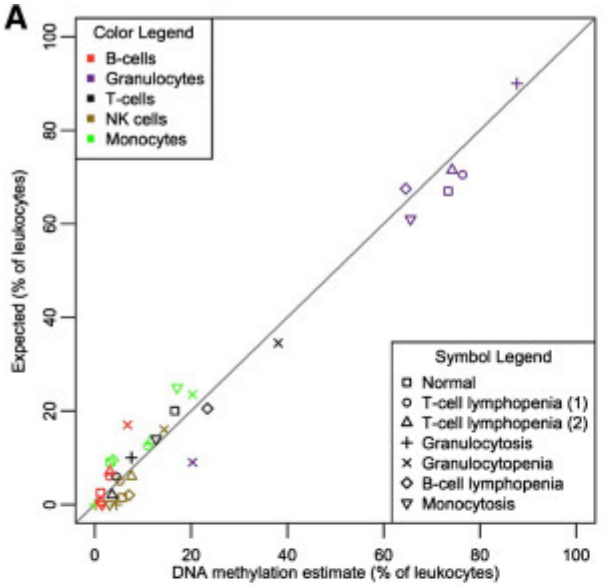


Heatmap of cell sorted 450k data

(Jaffe & Irizarry *Genome Biology*, 2014)

Solutions (in order of effectiveness)

- Flow cytometry Adjust for **measured** cell proportions or restrict analysis to one subtype (expensive and impractical for prospectively collected samples)
- Houseman method (Houseman *et al. BMC Bioinformatics,* 2012) Uses data from purified cells to build a DNAm model to **estimate** cell population in unknown DNAm profiles.
- Reference-free methods (e.g. Houseman *et al., Bioinformatics,* 2014, Zou *et al., Nature Methods,* 2014).



Accomando *et al. Genome Biol*. 2014.

Bioinformatics, 34(23), 2018, 3983–3989
doi: 10.1093/bioinformatics/bty476
Advance Access Publication Date: 21 June 2018
Original Paper

OXFORD

Genome analysis

Meffil: efficient normalization and analysis of very large DNA methylation datasets

J. L. Min[1,2,*,†], G. Hemani[1,2,†], G. Davey Smith[1,2], C. Relton[1,2] and M. Suderman[1,2,*]

[1]MRC Integrative Epidemiology Unit and [2]Bristol Medical School, University of Bristol, Bristol, BS8 2BN, UK

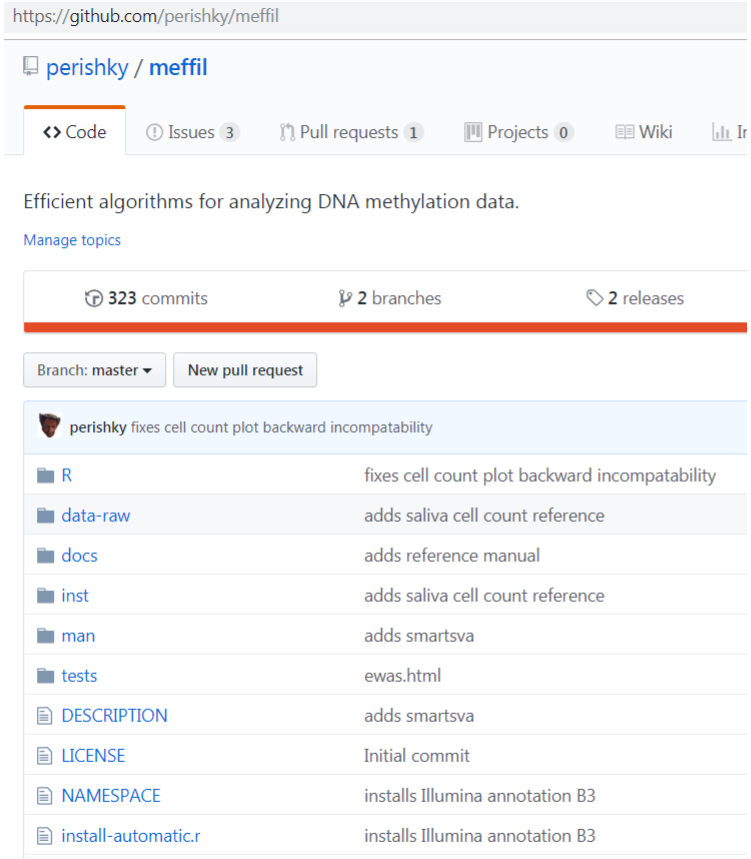
*To whom correspondence should be addressed.
†The authors wish it to be known that these authors contributed equally.
Associate Editor: John Hancock

Received on January 9, 2018; revised on May 6, 2018; editorial decision on June 11, 2018; accepted on June 18, 2018

Abstract

Motivation: DNA methylation datasets are growing ever larger both in sample size and genome coverage. Novel computational solutions are required to efficiently handle these data.
Results: We have developed *meffil*, an R package designed for efficient quality control, normalization and epigenome-wide association studies of large samples of Illumina Methylation BeadChip microarrays. A complete re-implementation of functional normalization minimizes computational memory without increasing running time. Incorporating fixed and random effects within functional normalization, and automated estimation of functional normalization parameters reduces technical variation in DNA methylation levels, thus reducing false positive rates and improving power. Support for normalization of datasets distributed across physically different locations without needing to share biologically-based individual-level data means that *meffil* can be used to reduce heterogeneity in meta-analyses of epigenome-wide association studies.
Availability and implementation: https://github.com/perishky/meffil/
Contact: josine.min@bristol.ac.uk or matthew.suderman@bristol.ac.uk
Supplementary information: Supplementary data are available at *Bioinformatics* online.

https://github.com/perishky/meffil



- Very large datasets
- Support from raw data to EWAS including QC, normalization, cell count variation, outliers, unknown confounders, EWAS

University of BRISTOL

BBSRC bioscience for the future

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

MRC Integrative Epidemiology Unit