Identifying differentially methylated regions using dmrff

## Download and prepare and prepare an example DNA methylation dataset

We’ll use a small cord blood DNA methylation dataset available on GEO: [GSE69633](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE69633).

Download data files:

geo.url <- "ftp://ftp.ncbi.nlm.nih.gov/geo/series"  
series.file <- file.path(geo.url, "GSE69nnn/GSE69633/matrix/GSE69633\_series\_matrix.txt.gz")  
if (!file.exists(basename(series.file)))  
 download.file(series.file, destfile=basename(series.file))  
data.file <- file.path(geo.url, "GSE69nnn/GSE69633/suppl/GSE69633\_processed\_betas\_UCB\_HM450K.txt.gz")  
if (!file.exists(basename(data.file)))  
 download.file(data.file, destfile=basename(data.file))

Retrieve sample information:

samples <- readLines(gzfile(basename(series.file)))  
start <- grep("Sample\_title", samples)  
end <- grep("series\_matrix\_table\_begin", samples)-1  
samples <- read.table(textConnection(samples[start:end]), sep="\t", header=F)

samples <- t(samples)  
colnames(samples) <- sub("!Sample\_", "", samples[1,])  
samples <- as.data.frame(samples[-1,], stringsAsFactors=F)  
samples$id <- sub("^[^(]+\\((.\*)\\)$", "\\1", samples$title)

Format sample characteristics:

idx <- which(colnames(samples) == "characteristics\_ch1")  
characteristics <- samples[,idx]  
colnames(characteristics) <- sub("([^:]+):.\*", "\\1", characteristics[1,])  
rownames(characteristics) <- samples$geo\_accession  
characteristics <- apply(characteristics, 2, function(x) sub("[^:]+: (.\*)", "\\1", x))  
characteristics <- as.data.frame(characteristics, stringsAsFactors=F)  
for (name in c("socioeconomic score", "gestational age", "birth weight", "pbconc (ng/dl)"))  
 characteristics[[name]] <- as.numeric(characteristics[[name]])  
colnames(characteristics) <- sub(" ", ".", colnames(characteristics))  
colnames(characteristics)[grep("pbconc", colnames(characteristics))] <- "pbconc"

Load DNA methylation data.

meth <- read.table(gzfile(basename(data.file)), header=T)  
meth <- meth[,-grep("Detection.Pval", colnames(meth))]  
meth <- as.matrix(meth)

Match between samples in samples, characteristics and meth.

idx <- match(colnames(meth), paste0("X", sub("-", ".", samples$id)))  
samples <- samples[idx,]  
characteristics <- characteristics[idx,]  
colnames(meth) <- rownames(characteristics)

## Test DNA methylation associations

Prepare surrogate variables to handle unknown confounding.

library(sva)

mod <- model.matrix(~ gender + socioeconomic.score + gestational.age + smoke.ever + birth.weight + pbconc, characteristics)  
mod0 <- mod[,1]  
set.seed(20190410)  
random.idx <- sample(1:nrow(meth), 5000)  
sva.fit <- sva(meth[random.idx,], mod=mod, mod0=mod0)

## Number of significant surrogate variables is: 13   
## Iteration (out of 5 ):1 2 3 4 5

Add surrogate variables to the model.

design <- cbind(mod, sva.fit$sv)

Test the associations using limma.

library(limma)  
fit <- lmFit(meth, design)  
fit <- eBayes(fit)

Save the summary statistics for gestational age.

stats <- data.frame(estimate=fit$coefficients[,"gestational.age"],  
 se=sqrt(fit$s2.post) \* fit$stdev.unscaled[,"gestational.age"],  
 p.value=fit$p.value[,"gestational.age"])

## Add genomic coordinates for CpG sites (Illumina Beadchips)

dmrff and all other methods for identifying differentially methylated regions require information about the genomic locations of all CpG sites.

If your data was generated using one of the Illumina BeadChips, then it is possible to annotate your EWAS summary statistics using the appropriate [Bioconductor](https://www.bioconductor.org/) annotation package.

For the Illumina 450K microaray, we use the IlluminaHumanMethylation450kanno.ilmn12.hg19 package.

if (!require(IlluminaHumanMethylation450kanno.ilmn12.hg19)) {  
 install.packages("BiocManager")  
 BiocManager::install("IlluminaHumanMethylation450kanno.ilmn12.hg19")  
 library(IlluminaHumanMethylation450kanno.ilmn12.hg19)  
}

If we were using the more recent Illumina MethylationEPIC BeadChip, then we would load the IlluminaHumanMethylationEPICanno.ilm10b2.hg19 package instead.

We construct an annotation data frame as follows:

data(list="IlluminaHumanMethylation450kanno.ilmn12.hg19")  
data(Locations)  
data(Other)  
annotation <- cbind(as.data.frame(Locations), as.data.frame(Other))

We then add the annotation to the stats object (**Warning**: we assume that the row names of the stats object are the Illumina CpG site identifiers, e.g. cg05775921).

annotation <- annotation[match(rownames(stats), rownames(annotation)),]  
stats <- cbind(stats, annotation)

## Apply dmrff to summary statistics

If dmrff is not already installed, we install it and load it.

if (!require(dmrff)) {  
 library(devtools)  
 install\_github("perishky/dmrff")  
 library(dmrff)  
}

Below we assume that you have already performed an epigenome-wide association analysis and have loaded your summary statistics in R as a data frame stats which has the following columns: - estimate (regression coefficient), - se (standard error of the coefficient), - p.value, - chr (chromosome of the CpG site), - pos (position of the CpG site on the chromosome).

We also assume that you have loaded your DNA methylation dataset in R as matrix meth for which rows correspond to CpG sites and columns to samples. The DNA methylation levels are necessary for dmrff to calculate and adjust for dependencies between CpG sites.

Before running dmrff, ensure that the dataset is ordered by CpG site genomic position.

idx <- order(stats$chr, stats$pos)  
stats <- stats[idx,]  
meth <- meth[idx,]

dmrff can then be applied as follows.

dmrs <- dmrff(estimate=stats$estimate,  
 se=stats$se,  
 p.value=stats$p.value,  
 methylation=meth,  
 chr=stats$chr,  
 pos=stats$pos,  
 maxgap=500,  
 verbose=T)

## [dmrff.candidates] Wed Apr 10 12:11:26 2019 Found 27414 candidate bumps.

The algorithm will then identify differentially methylated regions by:

* Identifying regions composed of CpG sites with EWAS p-values < 0.05 and consist direction of effect (at most 500bp between CpG sites – see maxgap parameter).
* Meta-analysing EWAS statistics within each region as well as sub-regions to identify the sub-region with lowest meta-analysed p-value.
* Adjusting meta-analysed p-values for multiple tests (all EWAS tests + sub-region meta-analysis tests).
* Finally, returning the DMR results.

Our output dmrs is a data frame listing all genomic regions tested along with summary statistics for each.

For example, the first 10 regions might look like the following:

kable(dmrs[1:10,c("chr","start","end","n","estimate","se","z","p.value","p.adjust")])

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| chr | start | end | n | estimate | se | z | p.value | p.adjust |
| chr6 | 31744831 | 31744927 | 3 | -0.0052637 | 0.0011350 | -4.6378024 | 0.0000035 | 1 |
| chr6 | 31744636 | 31744636 | 1 | -0.0132362 | 0.0037004 | -3.5769566 | 0.0003476 | 1 |
| chr6 | 31744612 | 31744612 | 1 | -0.0150106 | 0.0044921 | -3.3415449 | 0.0008331 | 1 |
| chr6 | 31744391 | 31744391 | 1 | -0.0098522 | 0.0029805 | -3.3055030 | 0.0009481 | 1 |
| chr6 | 31743769 | 31743952 | 5 | -0.0028035 | 0.0009087 | -3.0852055 | 0.0020341 | 1 |
| chr6 | 31744037 | 31744339 | 3 | -0.0016838 | 0.0013148 | -1.2806828 | 0.2003051 | 1 |
| chr6 | 31744033 | 31744033 | 1 | -0.0006374 | 0.0009404 | -0.6777688 | 0.4979183 | 1 |
| chr6 | 31743986 | 31743986 | 1 | 0.0000130 | 0.0010535 | 0.0122943 | 0.9901908 | 1 |
| chr6 | 31744398 | 31744524 | 2 | -0.0038149 | 0.0015648 | -2.4378715 | 0.0147740 | 1 |
| chr6 | 31744545 | 31744545 | 1 | -0.0123068 | 0.0051349 | -2.3966699 | 0.0165448 | 1 |

Some regions contain only one CpG site. EWAS is just fine for identifying associations with single CpG sites, so we might just remove these regions from the output.

dmrs <- dmrs[which(dmrs$n > 1),]

1769 regions with at least 2 CpG sites were tested.

All genomic regions with p-value < 0.05 (Bonferroni adjusted for multiple tests) can be listed as follows:

kable(dmrs[which(dmrs$p.adjust < 0.05),  
 c("chr","start","end","n","estimate","se","z","p.value","p.adjust")])

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | chr | start | end | n | estimate | se | z | p.value | p.adjust |
| 2151 | chr13 | 29148952 | 29149132 | 2 | 0.0085360 | 0.0015125 | 5.643680 | 0 | 0.0070554 |
| 2454 | chr4 | 6695614 | 6695698 | 2 | 0.0104099 | 0.0018918 | 5.502625 | 0 | 0.0158602 |
| 2812 | chr1 | 67519155 | 67519474 | 2 | -0.0054861 | 0.0009986 | -5.493554 | 0 | 0.0166972 |

We can also list the statistics for each CpG site in these regions.

sites.idx <- sapply(which(dmrs$p.adjust < 0.05),  
 function(dmr.idx) dmrs$start.idx[dmr.idx]:dmrs$end.idx[dmr.idx])  
sites.idx <- unlist(sites.idx)  
kable(stats[sites.idx, c("chr","pos","UCSC\_RefGene\_Name","estimate","se","p.value")])

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | chr | pos | UCSC\_RefGene\_Name | estimate | se | p.value |
| cg15887927 | chr13 | 29148952 |  | 0.0062306 | 0.0022104 | 0.0087764 |
| cg24561305 | chr13 | 29149132 |  | 0.0111855 | 0.0023987 | 0.0000706 |
| cg26233331 | chr4 | 6695614 | S100P;S100P | 0.0092628 | 0.0019054 | 0.0000411 |
| cg22266967 | chr4 | 6695698 | S100P | 0.0065167 | 0.0024427 | 0.0125828 |
| cg01802975 | chr1 | 67519155 | SLC35D1 | -0.0054169 | 0.0010328 | 0.0000144 |
| cg17930550 | chr1 | 67519474 | SLC35D1 | -0.0059922 | 0.0026819 | 0.0336787 |

## Annotating differentially methylated regions (Illumina Beadchips)

Using the CpG site annotations described earlier, we can also annotate differentially methylated regions.

To save time, we will only annotate the top 50 differentially methylated regions.

dmrs50 <- dmrs[order(dmrs$p.value)[1:50],]

For annotation, we will use the following function.

annotate.regions <- function(dmrs, stats, package) {  
 if (!require(package, character.only=T)) {  
 install.packages("BiocManager")  
 BiocManager::install(package)  
 library(package, character.only=T)  
 }   
 data(list=package)  
 data(Locations)  
 data(Other)  
 annotation <- cbind(as.data.frame(Locations), as.data.frame(Other))  
  
 stopifnot(all(rownames(stats) %in% rownames(annotation)))  
   
 annotation <- annotation[match(rownames(stats), rownames(annotation)),]  
 stats <- cbind(stats, annotation)  
   
 dmrs.annot <- lapply(1:nrow(dmrs), function(dmr.idx) {  
 annotation$Forward\_Sequence <- annotation$SourceSeq <- annotation$pos <- annotation$chr <- NULL  
 site.idx <- dmrs$start.idx[dmr.idx]:dmrs$end.idx[dmr.idx]  
 annotation <- annotation[site.idx,]  
 multi.idx <- grep("UCSC\_RefGene", colnames(annotation))  
 for (idx in multi.idx)  
 annotation[[idx]] <- strsplit(annotation[[idx]], ";")  
 n <- sapply(annotation[[multi.idx[1]]], length)  
 annotation <- c(lapply(annotation[multi.idx], unlist),  
 lapply(annotation[-multi.idx], rep, n))  
 annotation <- do.call(data.frame, c(annotation, list(stringsAsFactors=F)))  
 sapply(annotation, function(x) paste(setdiff(unique(x),""), collapse=";"))  
 })  
 dmrs.annot <- do.call(rbind, dmrs.annot)  
  
 cbind(dmrs, dmrs.annot)  
}

We annotate the top 50 differentially methylated regions.

dmrs50 <- annotate.regions(dmrs50, stats, "IlluminaHumanMethylation450kanno.ilmn12.hg19")

If we were using the more recent Illumina MethylationEPIC BeadChip, then we would use the IlluminaHumanMethylationEPICanno.ilm10b2.hg19 package name instead.

Here we can see the information added for the regions with Bonferonni adjusted p-values less than 0.05.

kable(dmrs50[which(dmrs50$p.adjust < 0.05),])

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | chr | start | end | n | start.idx | end.idx | start.orig | end.orig | z.orig | p.orig | B | S | estimate | se | z | p.value | p.adjust | UCSC\_RefGene\_Name | UCSC\_RefGene\_Accession | UCSC\_RefGene\_Group | strand | Random\_Loci | Methyl27\_Loci | Phantom | DMR | Enhancer | HMM\_Island | Regulatory\_Feature\_Name | Regulatory\_Feature\_Group | DHS |
| 2151 | chr13 | 29148952 | 29149132 | 2 | 102277 | 102278 | 102277 | 102278 | 5.643680 | 0 | 0.0085360 | 0.0015125 | 0.0085360 | 0.0015125 | 5.643680 | 0 | 0.0070554 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 2454 | chr4 | 6695614 | 6695698 | 2 | 271174 | 271175 | 271174 | 271175 | 5.502625 | 0 | 0.0104099 | 0.0018918 | 0.0104099 | 0.0018918 | 5.502625 | 0 | 0.0158602 | S100P | NM\_005980 | 1stExon;5’UTR | - |  | TRUE |  |  |  |  |  |  |  |
| 2812 | chr1 | 67519155 | 67519474 | 2 | 17843 | 17844 | 17843 | 17844 | -5.493554 | 0 | -0.0054861 | 0.0009986 | -0.0054861 | 0.0009986 | -5.493554 | 0 | 0.0166972 | SLC35D1 | NM\_015139 | Body | + |  |  | high-CpG:67292032-67292659 |  |  | 1:67291641-67292781 |  |  | TRUE |