The DMRcate package user's guide

Peters TJ, Buckley MJ, Pidsley R, Clark SJ, Molloy PL

March 11, 2014

Summary

DMRcate extracts the most differentially methylated regions (DMRs), variably methylated regions (VMRs) and hypermethylated regions (HMRs) from Illumina®Infinium HumanMethylation450 BeadChip (hereby referred to as the 450k array) samples via kernel density modelling. We provide clean, transparent code and highly interpretable and exportable results.

DMRcate requires R version 3.0.2 or higher.

```
source("http://bioconductor.org/biocLite.R")
biocLite("DMRcate")
```

Load DMRcate into the workspace:

```
library(DMRcate)
```

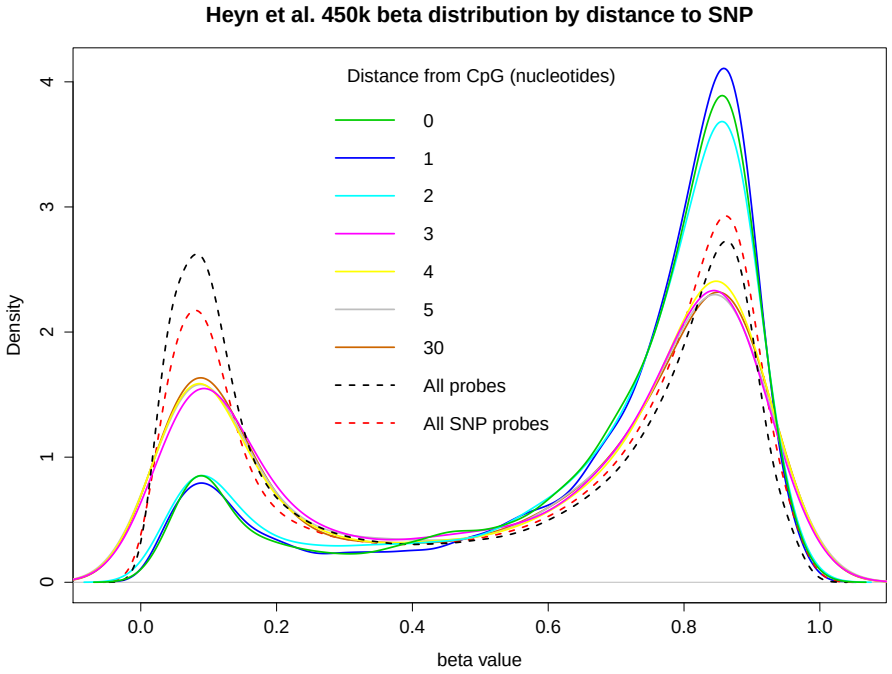
We now can load in the test data set of beta values. We assume at this point that normalisation and filtering out bad-quality probes via their detection p-values have already been done. Many packages are available for these purposes, including minfi, wateRmelon and methylumi. M-values (logit-transform of beta) are preferable to beta values for analysis because of increased sensitivity, but we will retain the beta matrix for visualisation purposes later on.

The TCGA (Cancer Genome Atlas - colorectal cancer) data in myBetas only comes from chromosome 20, but DMRcate will have no problem taking in the approximately half million probes as input for this pipeline either.

```
data(dmrcatedata)
myMs <- logit2(myBetas)
```

Some of the methylation measurements on the array may be confounded by proximity to SNPs, and cross-hybridisation to other areas of the genome[1]. In particular, probes that are 0, 1, or 2 nucleotides from the methylcytosine of interest show a markedly different distribution to those farther away, in healthy tissue (Figure 1).

Figure 1: Beta distribution of 450K probes from publically available data from blood samples of healthy individuals [3] by their proximity to a SNP. "All SNP probes" refers to the 153 113 probes listed by Illumina® whose values may potentially be confounded by a SNP.



It is with this in mind that we filter out probes 2 nucleotides or closer to a SNP that have a minor allele frequency greater than 0.05, and the approximately 30,000 [1] cross-reactive probes, so as to reduce confounding. Here we use Illumina®'s database of approximately 150,000 potentially SNP-confounded probes, and an internally-loaded dataset of the probes from [1], to filter these probes out. About 600 are removed from our M-matrix of approximately 10,000:

```
nrow(illuminaSNPs)

## [1] 153113

nrow(myMs)

## [1] 10042

myMs.noSNPs <- rmSNPandCH(myMs, dist = 2, mafcut = 0.05)
nrow(myMs.noSNPs)

## [1] 9403
```

Next we want to annotate our matrix of M-values with relevant information. The default is the `ilmn12.hg19` annotation, but this can be substituted for any argument compatible with the interface provided by the `minfi` package. We also use the backbone of the `limma` pipeline for differential array analysis to get fold changes and, optionally, filter probes by their fdr-corrected textitp-value. Here we have 38 patients with 2 tissue samples each taken from them. We want to compare within patients across tissue samples, so we set up our variables for a standard limma pipeline, and set `coef=39` in `annotate` since this corresponds to the phenotype comparison in `design`. We also only want to retain probes that are signficant in this comparison, post-fdr-correction.

```
patient <- factor(sub("-.*", "", colnames(myMs)))
type <- factor(sub(".*-", "", colnames(myMs)))
design <- model.matrix(~patient + type)
myannotation <- annotate(myMs.noSNPs, analysis.type = "differential", design = design,
    coef = 39, diff.metric = "FC", paired = TRUE, pcutoff = 0.01)

## Loading required package:  IlluminaHumanMethylation450kanno.ilmn12.hg19
```

We can see that, after using an adjusted p-value threshold of 0.01, we are left with about 5000 CpG sites to fit on chromosome 20:

```
length(myannotation$ID)

## [1] 5025
```

Now we can find our most differentially methylated regions with `dmrcate`. Mathematically, we formulate the underlying density f from the vector X of hg19 coordinates of each CpG used as input like this:

$$\hat{f}_{nH}(x) = n^{-1} \sum_{i=1}^{n} w(X_i) K_H(x - X_i) \quad (1)$$

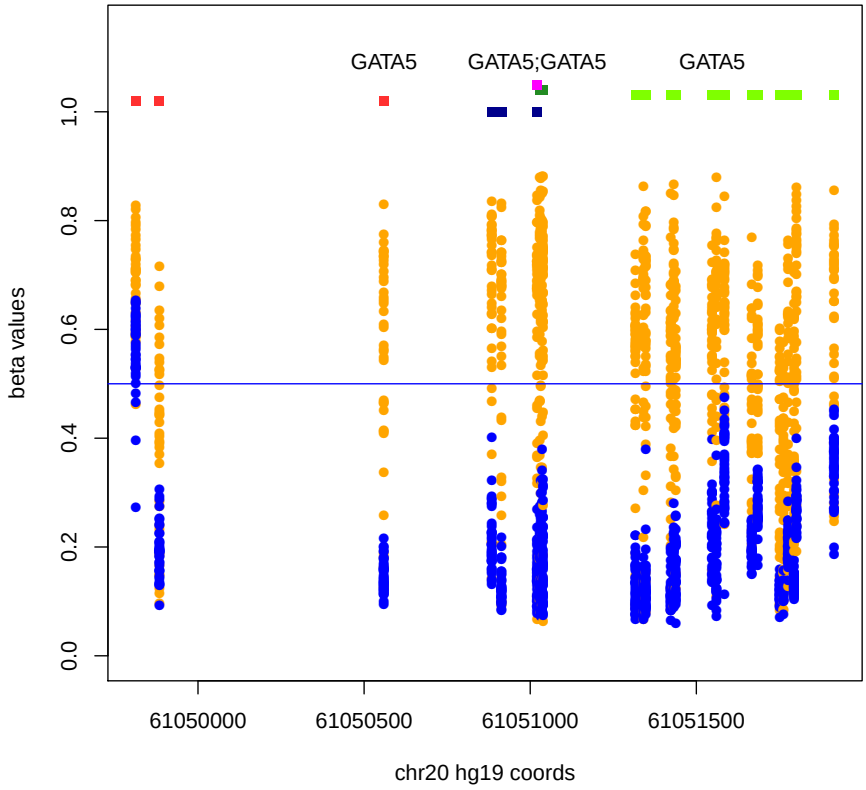
where $K(x)$ is the kernel function, H the bandwidth (needs to be specified, either in nucleotides or in probes when `consec=TRUE`) and w the weight vector containing the absolute fold change values for each probe.

For each chromosome, two of these density estimates are computed: one weighted with `myannotation$weights` and one with a vector of all 1s, for a null comparison. The two estimates are compared, and a significance test[2] is calculated at all hg19 coordinates that an input probe maps to (underlying function is `kde.local.test`). After fdr-correction, regions are then agglomerated from groups of significant probes where the distance to the next consecutive probe is less than `bw` nucleotides.

```
dmrcoutput <- dmrcate(myannotation, bw = 1000)

## Fitting chr20...
## Demarcating regions...
## Done!
```

Now we can plot our most significant differentially methylated region. It is associated with the GATA5 locus.

```
dmrcoutput$results[1, ]

##   gene_assoc                                group               hg19coord
## 5      GATA5 Body,5'UTR,1stExon,TSS200,TSS1500 chr20:61049813-61051915
##   no.probes   minpval meanpval
## 5        27 8.151e-07 0.001609

DMR.plot(dmrcoutput = dmrcoutput, dmr = 1, betas = myBetas, phen.col = c(rep("orange",
    38), rep("blue", 38)), pch = 16, toscale = TRUE)
```




```
sessionInfo()

## R Under development (unstable) (2014-02-17 r65021)
## Platform: x86_64-unknown-linux-gnu (64-bit)
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] parallel  stats     graphics  grDevices utils     datasets  methods
## [8] base
```

```
##
## other attached packages:
##  [1] IlluminaHumanMethylation450kanno.ilmn12.hg19_0.2.1
##  [2] DMRcate_0.99.6
##  [3] DMRcatedata_0.99.2
##  [4] minfi_1.9.13
##  [5] bumphunter_1.3.8
##  [6] locfit_1.5-9.1
##  [7] iterators_1.0.6
##  [8] foreach_1.4.1
##  [9] Biostrings_2.31.14
## [10] XVector_0.3.7
## [11] GenomicRanges_1.15.38
## [12] GenomeInfoDb_0.99.19
## [13] IRanges_1.21.34
## [14] lattice_0.20-27
## [15] Biobase_2.23.6
## [16] BiocGenerics_0.9.3
## [17] limma_3.19.23
## [18] ks_1.8.13
## [19] misc3d_0.8-4
## [20] rgl_0.93.996
## [21] mvtnorm_0.9-9997
## [22] KernSmooth_2.23-10
##
## loaded via a namespace (and not attached):
##  [1] AnnotationDbi_1.25.14 DBI_0.2-7             MASS_7.3-30
##  [4] R.methodsS3_1.6.1     RColorBrewer_1.0-5    RSQLite_0.11.4
##  [7] XML_3.98-1.1          annotate_1.41.2       base64_1.1
## [10] beanplot_1.1          codetools_0.2-8       digest_0.6.4
## [13] doRNG_1.6             evaluate_0.5.1        formatR_0.10
## [16] genefilter_1.45.2     grid_3.1.0            highr_0.3
## [19] illuminaio_0.5.6      knitr_1.5             matrixStats_0.8.14
## [22] mclust_4.2            multtest_2.19.2       nlme_3.1-115
## [25] nor1mix_1.1-4         pkgmaker_0.20         preprocessCore_1.25.5
## [28] registry_0.2          reshape_0.8.4         rngtools_1.2.4
## [31] siggenes_1.37.2       splines_3.1.0         stats4_3.1.0
## [34] stringr_0.6.2         survival_2.37-7       tools_3.1.0
## [37] xtable_1.7-3
```

References

[1] Chen YA, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, Gallinger S, Hudson TJ, Weksberg R. Discovery of cross-reactive probes and

polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics*. 2013 Jan 11;8(2).

[2] Duong T. Local significant differences from nonparametric two-sample tests. *Journal of Nonparametric Statistics*. 2013 **25**(3), 635-645.

[3] Heyn H, Li N, Ferreira HJ, Moran S, Pisano DG, Gomez A, Esteller M. Distinct DNA methylomes of newborns and centenarians. *Proceedings of the National Academy of Sciences*. 2012 **109**(26), 10522-7.