The DMRcate package user's guide

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

June 29, 2015

Summary

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

```
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 significance testing via limma because of increased sensitivity, but we will retain the beta matrix for visualisation purposes later on.

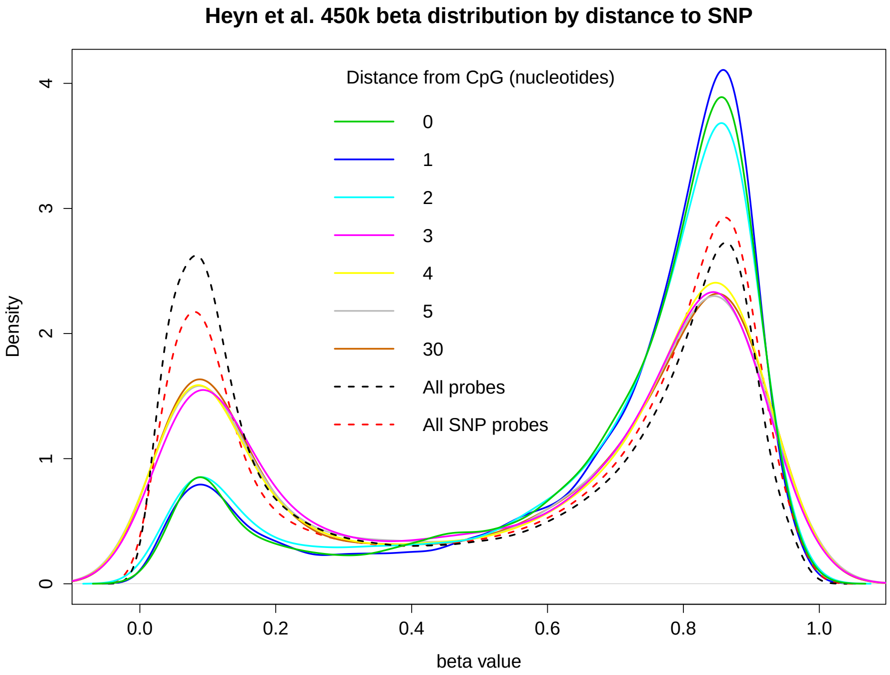
The TCGA (Cancer Genome Atlas - colorectal cancer) data in myBetas only comes from chromosome 20, but DMRcate will have no problem taking in the approximately half million probes as input for this pipeline either.

```
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).

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

Figure 1: Beta distribution of 450K probes from publically available data from blood samples of healthy individuals [2] 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.



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 t-statistics changes and, optionally, filter probes by their fdr-corrected p-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 `cpg.annotate` since this corresponds to the phenotype comparison in `design`.

```
patient <- factor(sub("-.*", "", colnames(myMs)))
type <- factor(sub(".*-", "", colnames(myMs)))
design <- model.matrix(~patient + type)
myannotation <- cpg.annotate(myMs.noSNPs, analysis.type="differential",
    design=design, coef=39)

## Your contrast returned 6101 individually significant probes.  We
recommend the default setting of pcutoff in dmrcate().
## Loading required package:  IlluminaHumanMethylation450kanno.ilmn12.hg19
```

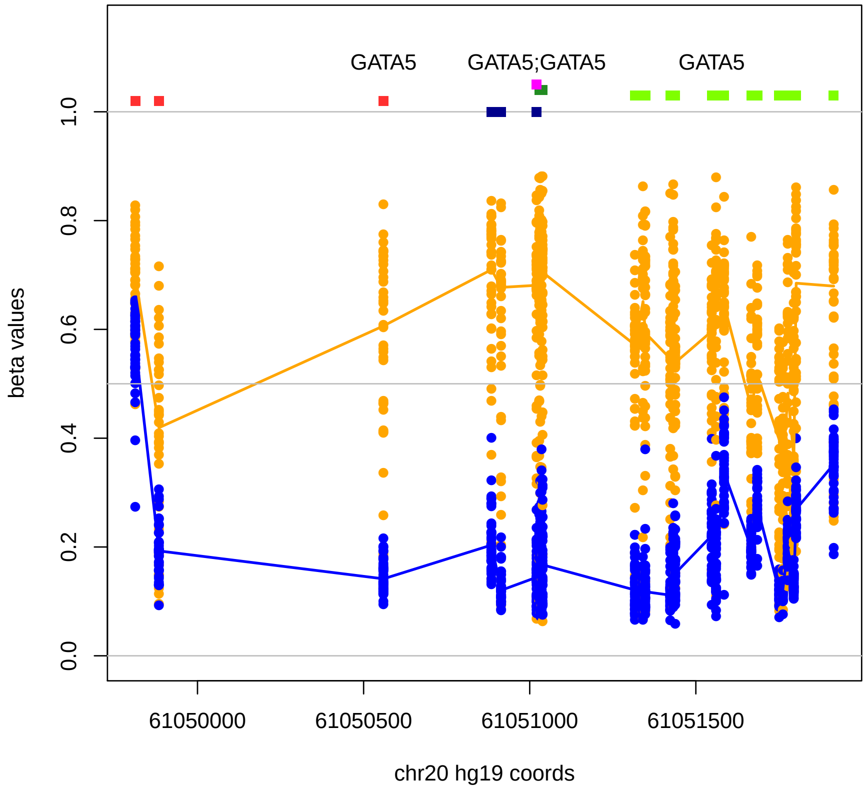
Now we can find our most differentially methylated regions with `dmrcate`.

For each chromosome, two smoothed estimates are computed: one weighted with `myannotation$weights` and one not, for a null comparison. The two estimates are compared via a Satterthwaite approximation[3], and a significance test is calculated at all hg19 coordinates that an input probe maps to. After fdr-correction, regions are then agglomerated from groups of significant probes where the distance to the next consecutive probe is less than `lambda` nucleotides.

```
dmrcoutput <- dmrcate(myannotation, lambda=1000, C=2)

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

Now we can plot a significant DMR. We'll choose one associated with the GATA5 locus.

```
head(dmrcoutput$results)

##      gene_assoc                                   group
## 839 GNASAS,GNAS Body,3'UTR,TSS200,TSS1500,5'UTR,1stExon
## 934       GATA5       Body,5'UTR,1stExon,TSS200,TSS1500
## 993    MIR124-3                     TSS1500,TSS200,Body
## 554        TOX2       TSS1500,TSS200,Body,5'UTR,1stExon
## 275     TMEM90B            TSS1500,TSS200,5'UTR,1stExon
## 283        VSX1       Body,1stExon,5'UTR,TSS200,TSS1500
##                   hg19coord no.probes minpval      meanpval  maxbetafc
## 839 chr20:57424521-57431303        77       0  3.002698e-29 -0.2084268
## 934 chr20:61049813-61051915        27       0  2.924163e-74  0.4770680
## 993 chr20:61806628-61810795        23       0  5.818155e-24  0.4182034
## 554 chr20:42543034-42545099        22       0  1.177014e-33  0.3684618
## 275 chr20:24448859-24452131        21       0 3.740710e-113  0.4263522
## 283 chr20:25061762-25065553        20       0  1.070921e-45  0.4679376

DMR.plot(dmrcoutput=dmrcoutput, dmr=2, betas=myBetas,
         phen.col=c(rep("orange", 38), rep("blue", 38)),
         pch=16, toscale=TRUE, plotmedians=TRUE)
```



```
sessionInfo()

## R version 3.2.1 (2015-06-18)
## Platform: x86_64-unknown-linux-gnu (64-bit)
## Running under: Ubuntu 14.04.2 LTS
##
## 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] stats4    parallel  stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] IlluminaHumanMethylation450kanno.ilmn12.hg19_0.2.1
##  [2] DMRcate_1.4.2
##  [3] DMRcatedata_1.4.0
##  [4] minfi_1.14.0
##  [5] bumphunter_1.8.0
##  [6] locfit_1.5-9.1
##  [7] iterators_1.0.7
##  [8] foreach_1.4.2
##  [9] Biostrings_2.36.1
## [10] XVector_0.8.0
## [11] GenomicRanges_1.20.5
## [12] GenomeInfoDb_1.4.1
## [13] IRanges_2.2.5
## [14] S4Vectors_0.6.1
## [15] lattice_0.20-31
## [16] Biobase_2.28.0
## [17] BiocGenerics_0.14.0
## [18] limma_3.24.12
##
## loaded via a namespace (and not attached):
##  [1] genefilter_1.50.0       splines_3.2.1
##  [3] beanplot_1.2            rtracklayer_1.28.5
##  [5] GenomicFeatures_1.20.1  XML_3.98-1.2
##  [7] survival_2.38-2         DBI_0.3.1
##  [9] BiocParallel_1.2.6      RColorBrewer_1.1-2
## [11] registry_0.2            rngtools_1.2.4
## [13] lambda.r_1.1.7          doRNG_1.6
## [15] matrixStats_0.14.2      plyr_1.8.3
## [17] pkgmaker_0.22           stringr_1.0.0
## [19] zlibbioc_1.14.0         futile.logger_1.4.1
## [21] codetools_0.2-11        evaluate_0.7
## [23] knitr_1.10.5            biomaRt_2.24.0
## [25] AnnotationDbi_1.30.1    illuminaio_0.10.0
## [27] preprocessCore_1.30.0   highr_0.5
## [29] Rcpp_0.11.6             xtable_1.7-4
## [31] formatR_1.2             base64_1.1
## [33] annotate_1.46.0         Rsamtools_1.20.4
## [35] digest_0.6.8            stringi_0.5-5
## [37] nor1mix_1.2-0           grid_3.2.1
## [39] GEOquery_2.34.0         quadprog_1.5-5
## [41] tools_3.2.1             bitops_1.0-6
## [43] magrittr_1.5            siggenes_1.42.0
## [45] RCurl_1.95-4.6          RSQLite_1.0.0
## [47] futile.options_1.0.0    MASS_7.3-41
## [49] reshape_0.8.5           mclust_5.0.1
## [51] nlme_3.1-120            GenomicAlignments_1.4.1
## [53] multtest_2.24.0
```

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] 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.

[3] Satterthwaite, F. E. (1946), An Approximate Distribution of Estimates of Variance Components., *Biometrics Bulletin.* 1946 **2**: 110-114