The `DMRcate` package user's guide

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

October 13, 2014

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 use

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.

Heyn et al. 450k beta distribution by distance to SNP
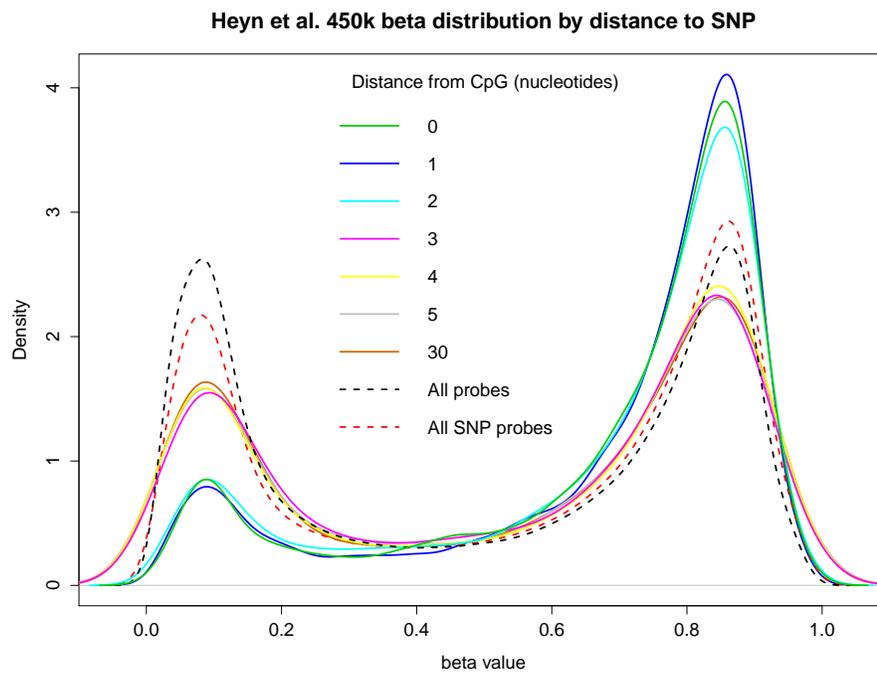
Illumina®'s database of approximately 150,000 potentially SNP-confounded probes, and an internally-loaded dataset of the probes from [1], to filter these probes out. About 600 are removed from our M-matrix of approximately 10,000:

```
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)

## 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
## 1568 GNASAS,GNAS Body,3'UTR,TSS200,TSS1500,5'UTR,1stExon
## 1756       GATA5         Body,5'UTR,1stExon,TSS200,TSS1500
## 1869    MIR124-3                       TSS1500,TSS200,Body
## 1050        TOX2        TSS1500,TSS200,Body,5'UTR,1stExon
## 538      TMEM90B             TSS1500,TSS200,5'UTR,1stExon
## 555         VSX1        Body,1stExon,5'UTR,TSS200,TSS1500
##                    hg19coord no.probes minpval       meanpval   maxbetafc
## 1568 chr20:57424521-57431303        77       0   3.002698e-29 -0.2084268
## 1756 chr20:61049813-61051915        27       0   2.924163e-74  0.4770680
## 1869 chr20:61806628-61810795        23       0   5.818155e-24  0.4182034
## 1050 chr20:42543034-42545099        22       0   1.177014e-33  0.3684618
## 538  chr20:24448859-24452131        21       0 3.740710e-113  0.4263522
## 555  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)
```
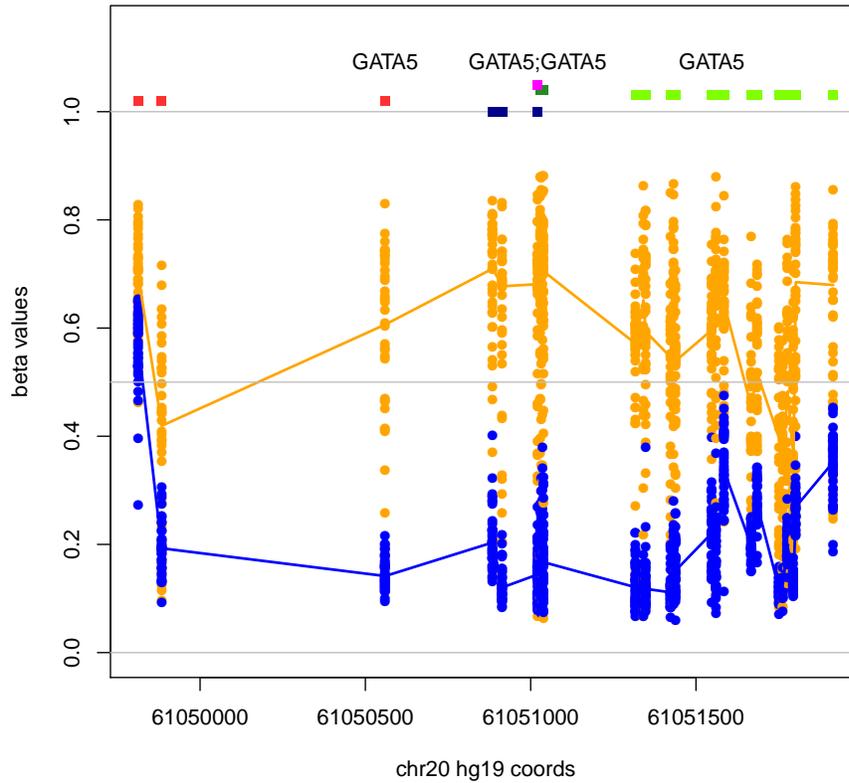
5

```
##
## other attached packages:
##  [1] IlluminaHumanMethylation450kanno.ilmn12.hg19_0.2.1
##  [2] DMRcate_1.2.0
##  [3] DMRcatedata_1.1.0
##  [4] minfi_1.12.0
##  [5] bumphunter_1.6.0
##  [6] locfit_1.5-9.1
##  [7] iterators_1.0.7
##  [8] foreach_1.4.2
##  [9] Biostrings_2.34.0
## [10] XVector_0.6.0
## [11] GenomicRanges_1.18.0
## [12] GenomeInfoDb_1.2.0
## [13] IRanges_2.0.0
## [14] S4Vectors_0.4.0
## [15] lattice_0.20-29
## [16] Biobase_2.26.0
## [17] BiocGenerics_0.12.0
## [18] limma_3.22.0
##
## loaded via a namespace (and not attached):
##  [1] AnnotationDbi_1.28.0  DBI_0.3.1             MASS_7.3-35
##  [4] R.methodsS3_1.6.1     RColorBrewer_1.0-5    RSQLite_0.11.4
##  [7] Rcpp_0.11.3           XML_3.98-1.1          annotate_1.44.0
## [10] base64_1.1            beanplot_1.2          codetools_0.2-9
## [13] digest_0.6.4          doRNG_1.6             evaluate_0.5.5
## [16] formatR_1.0           genefilter_1.48.0     grid_3.1.1
## [19] highr_0.3             illuminaio_0.8.0      knitr_1.7
## [22] matrixStats_0.10.0    mclust_4.4            multtest_2.22.0
## [25] nlme_3.1-118          nor1mix_1.2-0         pkgmaker_0.22
## [28] plyr_1.8.1            preprocessCore_1.28.0 quadprog_1.5-5
## [31] registry_0.2          reshape_0.8.5         rngtools_1.2.4
## [34] siggenes_1.40.0       splines_3.1.1         stringr_0.6.2
## [37] survival_2.37-7       tools_3.1.1           xtable_1.7-4
## [40] zlibbioc_1.12.0
```

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