Calling DMRs from EPICv1 and 450K data

Peters TJ

June 16, 2024

Summary

This vignette demonstrates how to call DMRs from older versions of Illumina arrays, namely 450K and EPICv1 (pre-2022).

```
if (!require("BiocManager"))
        install.packages("BiocManager")
BiocManager::install("DMRcate")
```

Load `DMRcate` into the workspace:

```
library(DMRcate)
```

For this vignette, we will demonstrate DMRcate's array utility using data from `ExperimentHub`, namely Illumina HumanMethylationEPIC data from the data packages `FlowSorted.Blood.EPIC`. Specifically, we are interested in the methylation differences between CD4+ and CD8+ T cells.

```
library(ExperimentHub)
eh <- ExperimentHub()
FlowSorted.Blood.EPIC <- eh[["EH1136"]]
tcell <- FlowSorted.Blood.EPIC[,colData(FlowSorted.Blood.EPIC)$CD4T==100 |
                                colData(FlowSorted.Blood.EPIC)$CD8T==100]
```

¡¡chr2, Firstly we filter out any probes where any sample has a failed position. Then we normalise using `minfi::preprocessFunnorm`. For this vignette, we will restrict the analysis to chromosome 2. After this, we extract the M-values from the GenomicRatioSet.

```
detP <- detectionP(tcell)

## Loading required package:  IlluminaHumanMethylationEPICmanifest

remove <- apply(detP, 1, function (x) any(x > 0.01))
tcell <- preprocessFunnorm(tcell)
```

```
## [preprocessFunnorm] Background and dye bias correction with noob
## Loading required package:  IlluminaHumanMethylationEPICanno.ilm10b4.hg19
## [preprocessFunnorm] Mapping to genome
## [preprocessFunnorm] Quantile extraction
## [preprocessFunnorm] Normalization

tcell <- tcell[seqnames(tcell) %in% "chr2",]
tcell <- tcell[!rownames(tcell) %in% names(which(remove)),]
tcellms <- getM(tcell)
```

M-values (logit-transform of beta) are preferable to beta values for significance testing via `limma` since they approximate normality, and provide greater sensitivity towards the extremes of the distribution, but we will use a beta matrix for visualisation purposes later on.

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, 2]. 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 48,000 [1, 2] cross-reactive probes on either 450K and/or EPIC, so as to reduce confounding. Here we use a combination of *in silico* analyses from [1, 2]. About 4,000 are removed from our M-matrix of 64,729 chromosome 2 probes:

```
nrow(tcellms)

## [1] 64729

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

## [1] 60445
```

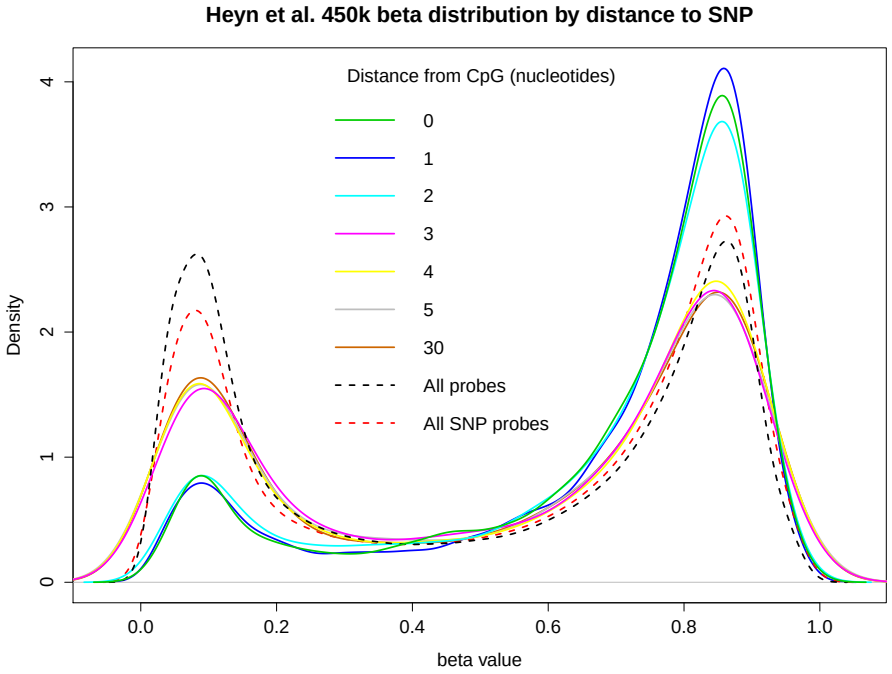
Here we have 6 CD8+ T cell assays, and 7 CD4+ T cell assays; we want to call DMRs between these groups. One of the CD4+ assays is a technical replicate, so we will average these two replicates like so:

```
tcell$Replicate

##  [1] ""          ""          ""          ""          ""          ""
##  [7] ""          ""          ""          "Th2535-1" "Th2535-1" ""
## [13] ""

tcell$Replicate[tcell$Replicate==""] <- tcell$Sample_Name[tcell$Replicate==""]
tcellms.noSNPs <- limma::avearrays(tcellms.noSNPs, tcell$Replicate)
tcell <- tcell[,!duplicated(tcell$Replicate)]
```

Figure 1: Beta distribution of 450K probes from publicly 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.



```
tcell <- tcell[rownames(tcellms.noSNPs),]
colnames(tcellms.noSNPs) <- colnames(tcell)
assays(tcell)[["M"]] <- tcellms.noSNPs
assays(tcell)[["Beta"]] <- ilogit2(tcellms.noSNPs)
```

Next we want to annotate our matrix of M-values with relevant information. We also use the backbone of the `limma` pipeline for differential array analysis. We want to compare within patients across tissue samples, so we set up our variables for a standard limma pipeline, and set `coef=2` in `cpg.annotate()` since this corresponds to the phenotype comparison in `design`.

`cpg.annotate()` takes either a data matrix with Illumina probe IDs, or an already prepared GenomicRatioSet from `minfi`.

```
type <- factor(tcell$CellType)
design <- model.matrix(~type)
myannotation <- cpg.annotate("array", tcell, arraytype = "EPICv1",
                             analysis.type="differential", design=design, coef=2)
```

```
myannotation

## CpGannotated object describing 60445 CpG sites, with independent
## CpG threshold indexed at fdr=0.05 and 2710 significant CpG sites.
```

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

For each chromosome, two smoothed estimates are computed: one weighted with per-CpG t-statistics and one not, for a null comparison. The two estimates are compared via a Satterthwaite approximation[4], and a significance test is calculated at all hg19 coordinates that an input probe maps to. After fdr-correction, regions are then aggregated from groups of post-smoothed significant probes where the distance to the next consecutive probe is less than `lambda` nucleotides.

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

## Fitting chr2...
## Demarcating regions...
## Done!

dmrcoutput

## DMResults object with 439 DMRs.
## Use extractRanges() to produce a GRanges object of these.
```

We can convert our DMR list to a GRanges object, which uses the `genome` argument to annotate overlapping gene loci.

```
results.ranges <- extractRanges(dmrcoutput, genome = "hg19")
results.ranges
```

```
## GRanges object with 439 ranges and 8 metadata columns:
##         seqnames              ranges strand |   no.cpgs min_smoothed_fdr
##            <Rle>           <IRanges>  <Rle> | <integer>        <numeric>
##     [1]     chr2   87014979-87021117      * |        26      0.00000e+00
##     [2]     chr2 234294036-234297039      * |        14     1.31193e-103
##     [3]     chr2   86991846-86992657      * |         3     2.82449e-203
##     [4]     chr2 112939119-112941244      * |         6      3.49350e-93
##     [5]     chr2 197124443-197125372      * |         4     1.36472e-135
##     ...      ...                 ...    ... .       ...              ...
##   [435]     chr2   43454761-43455773      * |        15      2.70362e-14
##   [436]     chr2 177001256-177001263      * |         3      6.91077e-10
##   [437]     chr2 121200209-121200256      * |         2      7.67192e-10
##   [438]     chr2 177052527-177053018      * |         5      2.38880e-10
##   [439]     chr2 173940027-173940121      * |         2      7.61333e-10
##           Stouffer       HMFDR      Fisher    maxdiff   meandiff
##          <numeric>   <numeric>   <numeric>  <numeric>  <numeric>
##     [1] 1.72294e-54 3.18747e-07 5.43315e-68  -0.733427  -0.236312
##     [2] 2.28737e-17 8.93582e-06 1.78156e-19  -0.385476  -0.139055
##     [3] 6.13508e-18 2.11325e-07 5.99765e-17  -0.530621  -0.395736
##     [4] 5.62342e-17 2.22940e-06 8.23449e-17   0.476714   0.336035
##     [5] 8.10953e-17 1.43019e-06 6.91821e-16   0.380658   0.287335
##     ...         ...         ...         ...        ...        ...
##   [435]    0.489338   0.0335443    0.205729 -0.0748315 -0.0133684
##   [436]    0.130045   0.2290845    0.211291  0.0456072  0.0272108
##   [437]    0.217200   0.1787023    0.226449  0.0408411  0.0400848
##   [438]    0.463773   0.1395391    0.250201  0.0408253  0.0215292
##   [439]    0.357791   0.1553425    0.253507  0.0526885  0.0306643
##         overlapping.genes
##               <character>
##     [1]              CD8A
##     [2]  DGKD, AC019221.4
##     [3]            RMND5A
##     [4]             FBLN7
##     [5] AC020571.3, HECW2
##     ...               ...
##   [435]            THADA
##   [436]          HOXD-AS2
##   [437]              <NA>
##   [438]          HOXD-AS1
##   [439]              <NA>
##   -------
##   seqinfo: 1 sequence from an unspecified genome; no seqlengths
```

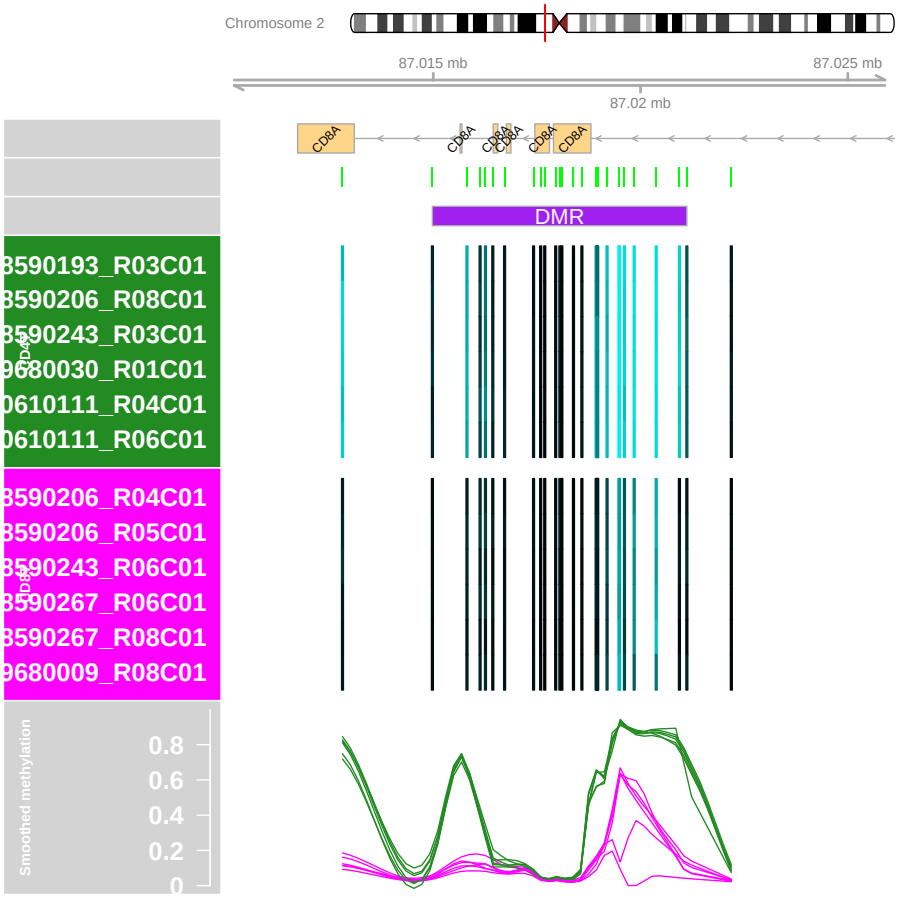
DMRs are ranked by Fisher's multiple comparison statistic, but `Stouffer` scores and the harmonic mean of the individual component FDRs (`HMFDR`) are also given in this object as alternative options for ranking DMR significance.

We can then pass this GRanges object to `DMR.plot()`, which uses the `Gviz` package as a backend for contextualising each DMR.

```
groups <- c(CD8T="magenta", CD4T="forestgreen")
cols <- groups[as.character(type)]
cols
```

```
##          CD4T          CD8T          CD8T          CD4T          CD4T
## "forestgreen"     "magenta"     "magenta" "forestgreen" "forestgreen"
##          CD8T          CD8T          CD8T          CD8T          CD4T
##     "magenta"     "magenta"     "magenta"     "magenta" "forestgreen"
##          CD4T          CD4T
## "forestgreen" "forestgreen"
```

```
DMR.plot(ranges=results.ranges, dmr=1, CpGs=getBeta(tcell), what="Beta",
         arraytype = "EPICv1", phen.col=cols, genome="hg19")
```



Consonant with the expected biology, our top DMR shows the CD8+ T cells hypomethylated across parts of the CD8A locus. The two distinct hypomethylated sections have been merged because they are less than 1000 bp apart - specified by `lambda` in the call to `dmrcate()`. To call these as separate DMRs, make `lambda` smaller.

```
sessionInfo()

## R version 4.4.0 (2024-04-24)
## Platform: x86_64-pc-linux-gnu
## Running under: Ubuntu 22.04.4 LTS
##
## Matrix products: default
## BLAS:   /home/biocbuild/bbs-3.19-bioc/R/lib/libRblas.so
## LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.10.0
##
```

```
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_GB              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
##
## time zone: America/New_York
## tzcode source: system (glibc)
##
## attached base packages:
## [1] parallel  stats4    stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] DMRcatedata_2.22.0
##  [2] IlluminaHumanMethylationEPICanno.ilm10b4.hg19_0.6.0
##  [3] IlluminaHumanMethylationEPICmanifest_0.3.0
##  [4] FlowSorted.Blood.EPIC_2.8.0
##  [5] minfi_1.50.0
##  [6] bumphunter_1.46.0
##  [7] locfit_1.5-9.9
##  [8] iterators_1.0.14
##  [9] foreach_1.5.2
## [10] Biostrings_2.72.1
## [11] XVector_0.44.0
## [12] SummarizedExperiment_1.34.0
## [13] Biobase_2.64.0
## [14] MatrixGenerics_1.16.0
## [15] matrixStats_1.3.0
## [16] GenomicRanges_1.56.1
## [17] GenomeInfoDb_1.40.1
## [18] IRanges_2.38.0
## [19] S4Vectors_0.42.0
## [20] ExperimentHub_2.12.0
## [21] AnnotationHub_3.12.0
## [22] BiocFileCache_2.12.0
## [23] dbplyr_2.5.0
## [24] BiocGenerics_0.50.0
## [25] DMRcate_3.0.2
##
## loaded via a namespace (and not attached):
##   [1] splines_4.4.0
##   [2] BiocIO_1.14.0
```

```
##   [3] bitops_1.0-7
##   [4] filelock_1.0.3
##   [5] cellranger_1.1.0
##   [6] tibble_3.2.1
##   [7] R.oo_1.26.0
##   [8] preprocessCore_1.66.0
##   [9] XML_3.99-0.16.1
##  [10] rpart_4.1.23
##  [11] lifecycle_1.0.4
##  [12] httr2_1.0.1
##  [13] edgeR_4.2.0
##  [14] base64_2.0.1
##  [15] MASS_7.3-61
##  [16] lattice_0.22-6
##  [17] ensembldb_2.28.0
##  [18] scrime_1.3.5
##  [19] backports_1.5.0
##  [20] magrittr_2.0.3
##  [21] limma_3.60.3
##  [22] Hmisc_5.1-3
##  [23] rmarkdown_2.27
##  [24] yaml_2.3.8
##  [25] doRNG_1.8.6
##  [26] askpass_1.2.0
##  [27] Gviz_1.48.0
##  [28] DBI_1.2.3
##  [29] RColorBrewer_1.1-3
##  [30] abind_1.4-5
##  [31] zlibbioc_1.50.0
##  [32] quadprog_1.5-8
##  [33] purrr_1.0.2
##  [34] R.utils_2.12.3
##  [35] AnnotationFilter_1.28.0
##  [36] biovizBase_1.52.0
##  [37] RCurl_1.98-1.14
##  [38] nnet_7.3-19
##  [39] VariantAnnotation_1.50.0
##  [40] rappdirs_0.3.3
##  [41] GenomeInfoDbData_1.2.12
##  [42] genefilter_1.86.0
##  [43] annotate_1.82.0
##  [44] permute_0.9-7
##  [45] DelayedMatrixStats_1.26.0
##  [46] codetools_0.2-20
##  [47] DelayedArray_0.30.1
```

```
##  [48] xml2_1.3.6
##  [49] tidyselect_1.2.1
##  [50] UCSC.utils_1.0.0
##  [51] beanplot_1.3.1
##  [52] base64enc_0.1-3
##  [53] illuminaio_0.46.0
##  [54] GenomicAlignments_1.40.0
##  [55] jsonlite_1.8.8
##  [56] multtest_2.60.0
##  [57] Formula_1.2-5
##  [58] survival_3.7-0
##  [59] missMethyl_1.38.0
##  [60] tools_4.4.0
##  [61] progress_1.2.3
##  [62] Rcpp_1.0.12
##  [63] glue_1.7.0
##  [64] gridExtra_2.3
##  [65] SparseArray_1.4.8
##  [66] xfun_0.44
##  [67] dplyr_1.1.4
##  [68] HDF5Array_1.32.0
##  [69] withr_3.0.0
##  [70] IlluminaHumanMethylation450kanno.ilmn12.hg19_0.6.1
##  [71] BiocManager_1.30.23
##  [72] fastmap_1.2.0
##  [73] latticeExtra_0.6-30
##  [74] rhdf5filters_1.16.0
##  [75] fansi_1.0.6
##  [76] openssl_2.2.0
##  [77] digest_0.6.35
##  [78] mime_0.12
##  [79] R6_2.5.1
##  [80] colorspace_2.1-0
##  [81] gtools_3.9.5
##  [82] jpeg_0.1-10
##  [83] dichromat_2.0-0.1
##  [84] biomaRt_2.60.0
##  [85] RSQLite_2.3.7
##  [86] R.methodsS3_1.8.2
##  [87] tidyr_1.3.1
##  [88] utf8_1.2.4
##  [89] generics_0.1.3
##  [90] data.table_1.15.4
##  [91] rtracklayer_1.64.0
##  [92] prettyunits_1.2.0
```

```
##  [93] httr_1.4.7
##  [94] htmlwidgets_1.6.4
##  [95] S4Arrays_1.4.1
##  [96] pkgconfig_2.0.3
##  [97] gtable_0.3.5
##  [98] blob_1.2.4
##  [99] siggenes_1.78.0
## [100] htmltools_0.5.8.1
## [101] ProtGenerics_1.36.0
## [102] scales_1.3.0
## [103] png_0.1-8
## [104] knitr_1.47
## [105] rstudioapi_0.16.0
## [106] tzdb_0.4.0
## [107] rjson_0.2.21
## [108] nlme_3.1-165
## [109] checkmate_2.3.1
## [110] curl_5.2.1
## [111] org.Hs.eg.db_3.19.1
## [112] cachem_1.1.0
## [113] rhdf5_2.48.0
## [114] stringr_1.5.1
## [115] BiocVersion_3.19.1
## [116] foreign_0.8-86
## [117] AnnotationDbi_1.66.0
## [118] restfulr_0.0.15
## [119] GEOquery_2.72.0
## [120] pillar_1.9.0
## [121] grid_4.4.0
## [122] reshape_0.8.9
## [123] vctrs_0.6.5
## [124] xtable_1.8-4
## [125] cluster_2.1.6
## [126] htmlTable_2.4.2
## [127] evaluate_0.24.0
## [128] bsseq_1.40.0
## [129] readr_2.1.5
## [130] GenomicFeatures_1.56.0
## [131] cli_3.6.2
## [132] compiler_4.4.0
## [133] Rsamtools_2.20.0
## [134] rngtools_1.5.2
## [135] rlang_1.1.4
## [136] crayon_1.5.2
## [137] nor1mix_1.3-3
```

```
## [138] mclust_6.1.1
## [139] interp_1.1-6
## [140] plyr_1.8.9
## [141] stringi_1.8.4
## [142] deldir_2.0-4
## [143] BiocParallel_1.38.0
## [144] munsell_0.5.1
## [145] lazyeval_0.2.2
## [146] Matrix_1.7-0
## [147] BSgenome_1.72.0
## [148] hms_1.1.3
## [149] sparseMatrixStats_1.16.0
## [150] bit64_4.0.5
## [151] ggplot2_3.5.1
## [152] Rhdf5lib_1.26.0
## [153] KEGGREST_1.44.0
## [154] statmod_1.5.0
## [155] highr_0.11
## [156] memoise_2.0.1
## [157] bit_4.0.5
## [158] readxl_1.4.3
```

References

[1] Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, Van Dijk S, Muhlhausler B, Stirzaker C, Clark SJ. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biology*. 2016 17(1), 208.

[2] 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).

[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.

[4] Satterthwaite FE. An Approximate Distribution of Estimates of Variance Components., *Biometrics Bulletin*. 1946 **2**: 110-114