Estimating DNA methylation levels and finding differentially methylated regions using the charm package

Martin Aryee*, Rafael Irizarry

January, 2010

Johns Hopkins School of Medicine / Johns Hopkins School of Public Health
Baltimore, MD, USA

1 Introduction

The Bioconductor package charm can be used to analyze data from the Nimblegen McrBC/CHARM DNA methylation microarray platform [1].

Functions include:

- Quality control
- Percentage methylation estimates
- Identification of differentially methylated regions

As input we will need raw Nimblegen data (.xys) files.
This vignette uses the following packages:

- charm: contains the analysis functions
- charmData: an example dataset
- pd.feinberg.hg18.me.hx1: the annotation package for the human CHARM microarray

Each sample is represented by two xys files corresponding to the untreated (green) and methyl-depleted (red) channels. The 532.xys and 635.xys suffixes indicate the green and red channels respectively.

*aryee@jhu

2 Install annotation and example data

Install the CHARM array annotation package and example data (if not already installed.)

```
R> if (!require(charmData)) {
     install.packages("charmData", repos = "http://R-Forge.R-project.org")
 }
R> if (!require(pd.feinberg.hg18.me.hx1)) {
     install.packages("pd.feinberg.hg18.me.hx1",
         repos = "http://R-Forge.R-project.org")
 }
```

3 Read in raw data

Get the name of your data directory (in this case, the example data):

```
R> dataDir <- system.file("data", package = "charmData")
R> dataDir
```

```
[1] "/thumper/ctsa/genomicsR/install/R-2.10.0/lib64/R/site-library/charmData/data"
```

First we read in the sample description file:

```
R> pd <- read.delim(file.path(dataDir, "sample_description_file.txt"))
R> pd
```

```
        Filename            DNA Individual Tissue_Type
1 136413_532.xys      untreated        441       brain
2 136421_532.xys      untreated        441       liver
3 136593_532.xys      untreated        449       brain
4 186974_532.xys      untreated        432       liver
5 136413_635.xys methyldepleted        441       brain
6 136421_635.xys methyldepleted        441       liver
7 136593_635.xys methyldepleted        449       brain
8 186974_635.xys methyldepleted        432       liver
```

Now we load the charm package and read in the data. The `readCharm` command makes the assumption (unless told otherwise) that the two xys files for a sample have the same file name up to the suffixes 532.xys (untreated) and 635.xys (methyl-depleted).

```
R> library(charm)
R> rawData <- readCharm(files = pd$Filename, path = dataDir,
     sampleKey = pd)
```

```
Checking designs for each XYS file... Done.
Allocating memory... Done.
Reading /thumper/ctsa/genomicsR/install/R-2.10.0/lib64/R/site-library/charmData/data/136413_
Reading /thumper/ctsa/genomicsR/install/R-2.10.0/lib64/R/site-library/charmData/data/136421_
Reading /thumper/ctsa/genomicsR/install/R-2.10.0/lib64/R/site-library/charmData/data/136593_
Reading /thumper/ctsa/genomicsR/install/R-2.10.0/lib64/R/site-library/charmData/data/186974_
Checking designs for each XYS file... Done.
Allocating memory... Done.
Reading /thumper/ctsa/genomicsR/install/R-2.10.0/lib64/R/site-library/charmData/data/136413_
Reading /thumper/ctsa/genomicsR/install/R-2.10.0/lib64/R/site-library/charmData/data/136421_
Reading /thumper/ctsa/genomicsR/install/R-2.10.0/lib64/R/site-library/charmData/data/136593_
Reading /thumper/ctsa/genomicsR/install/R-2.10.0/lib64/R/site-library/charmData/data/186974_
```

```
R> rawData
```

```
TilingFeatureSet2 (storageMode: lockedEnvironment)
assayData: 2197815 features, 4 samples
  element names: channel1, channel2
phenoData
  sampleNames: 136413, 136421, 136593, 186974
  varLabels and varMetadata description:
    Individual: NA
    Tissue_Type: NA
    ...: ...
    channel2DateTime: date/time from raw files
    (6 total)
  additional varMetadata: channel
featureData
  featureNames: 1, 2, ..., 2197815  (2197815 total)
  fvarLabels and fvarMetadata description: none
experimentData: use 'experimentData(object)'
Annotation: pd.feinberg.hg18.me.hx1
```

4 Array quality assessment

We can calculate array quality scores and generate a pdf report with the qcReport command.

A useful quick way of assessing data quality is to examine the untreated channel where we expect every probe to have signal. Very low signal intensities on all or part of an array can indicate problems with hybridization or scanning. The CHARM array includes background probes that do not match any genomic sequence. Any signal at these probes can be assumed to be the result of optical noise or cross-hybridization. The array quality score is the average percentile rank of the signal robes among these background probes. A score of 100 means all signal probes rank above all background probes (the ideal scenario).

```
R> qual <- qcReport(rawData, file = "qcReport.pdf")
R> qual
```

```
       pmSignal       sd1       sd2
136413 77.68940 0.1728477 0.2001947
136421 78.66791 0.2020508 0.2042365
136593 77.69742 0.1362804 0.2326469
186974 84.02167 0.1497900 0.2497021
```

The PDF quality report is shown in Appendix A.

5 Percentage methylation estimates and Differentially methylated regions (DMRs)

Having determined that no arrays need to be thrown out due to hybridization quality issues we can go ahead and calculate probe-level percentage methylation estimates for each sample. The 'plotDensity' option of methp produces useful PDF diagnostic plots to help identify non-hybridization quality issues. The report is shown in Appendix B.

```
R> p <- methp(rawData, plotDensity = "density.pdf")
```

```
Spatial normalization
Background removal
Within sample normalization: loess
Between sample normalization: quantile
Estimating percentage methylation
```

```
R> head(p)
```

```
     136413    136421    136593     186974
1 0.6227355 0.6042311 0.5174794 0.04779634
2 0.7522372 0.8563471 0.6094113 0.93035737
3 0.7083703 0.6292665 0.3327768 0.58263832
4 0.8299732 0.5648182 0.3164507 0.37394479
5 0.7952075 0.5330259 0.1146673 0.43362372
6 0.2136278 0.2168371 0.0929393 0.66097041
```

We can also identify differentially methylated regions using dmrFinder:

```
R> grp <- pData(rawData)$Tissue_Type
R> grp
```

```
[1] brain liver brain liver
Levels: brain liver
```

```
R> dmr <- dmrFinder(rawData, p = p, groups = grp)
```

```
Computing group medians and SDs for 2 groups:
 1
 2
Done.
Smoothing.....................Done.
Finding DMRs for each pairwise comparison.
 brain-liver.....................
Done
```

```
R> names(dmr)
```

```
 [1] "tabs"         "p"            "l"
 [4] "chr"          "pos"          "pns"
 [7] "index"        "controlIndex" "gm"
[10] "groups"       "args"         "cutoff"
[13] "filter"       "ws"           "comps"
[16] "package"
```

```
R> names(dmr$tabs)
```

```
[1] "brain-liver"
```

```
R> head(dmr$tabs[[1]])
```

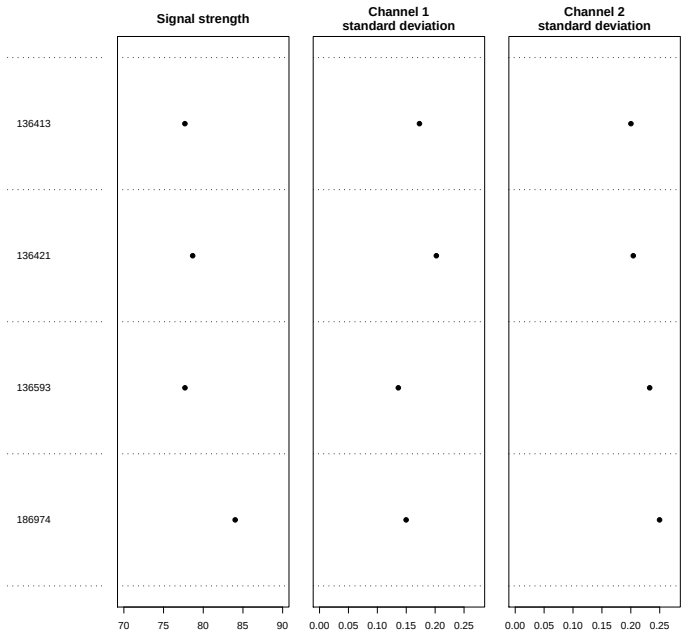
```
        chr     start       end        p1        p2
6738  chr15  91163239  91164505 0.1202958 0.7918648
16077  chr4  99796292  99797778 0.1324501 0.6269146
14405  chr3 149898776 149899872 0.7876629 0.1605199
18633  chr7 130439092 130440100 0.1667484 0.7236147
11473 chr20  55267807  55268617 0.1683895 0.8333141
18230  chr6  52637822  52638747 0.7805848 0.1488053
                   regionName indexStart indexEnd     area
6738   chr15:91150286-91166158     642449   642484 24.17648
16077   chr4:99796289-99800666    1572828  1572869 20.76751
14405 chr3:149897739-149899949    1494860  1494891 20.06858
18633 chr7:130437932-130444273    1870692  1870720 16.14912
11473  chr20:55266497-55276390    1288942  1288965 15.95819
18230   chr6:52635302-52638967    1733591  1733613 14.53093
```

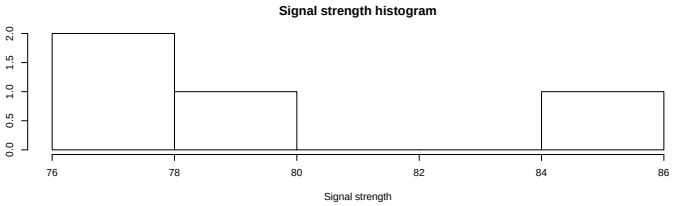
— UNDER CONSTRUCTION —- One of the samples has consistent signal strength across its array while the second has clear spatial artifacts. The `methp` function we ran earlier includes a step that can correct such artifacts if they are not too severe. We can check to see how successful the correction was by running this step manually:

References

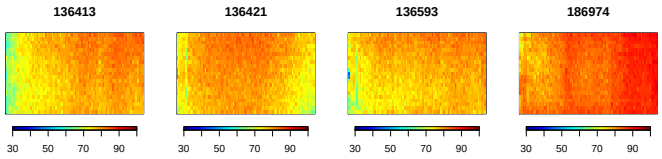
[1] Irizarry et al. Comprehensive high-throughput arrays for relative methylation (charm). *Genome Research*, 18(5):780–790, 2008.

6 Appendix A: Quality report

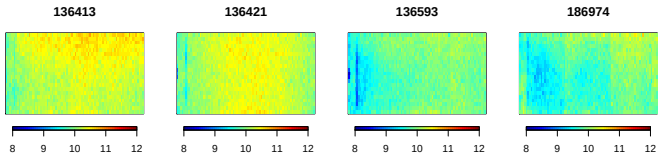

Signal strength
Channel 1
standard deviation
Channel 2
standard deviation
136413
136421
136593
186974
70
75
80
85
90
0.00
0.05
0.10
0.15
0.20
0.25
0.00
0.05
0.10
0.15
0.20
0.25


Signal strength histogram
0.0
0.5
1.0
1.5
2.0
76
78
80
82
84
86
Signal strength

Untreated Channel: PM probe quality

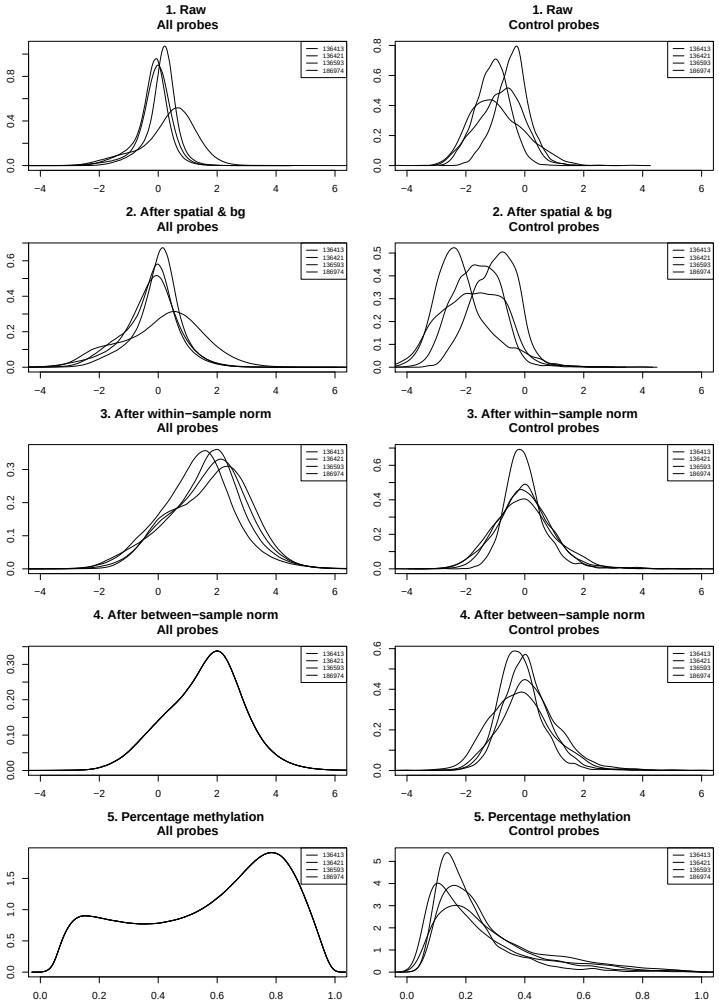


Enriched Channel: PM signal intensity



7 Appendix B: Density plots

Each row shows one stage of preprocessing. The left plot shows all probes while the right plot shows control probes.


1. Raw
All probes
1. Raw
Control probes
2. After spatial & bg
All probes
2. After spatial & bg
Control probes
3. After within−sample norm
All probes
3. After within−sample norm
Control probes
4. After between−sample norm
All probes
4. After between−sample norm
Control probes
5. Percentage methylation
All probes
5. Percentage methylation
Control probes
136413
136421
136503
186974

8 Details

This document was written using:

```
R> sessionInfo()
```

```
R version 2.10.0 (2009-10-26)
x86_64-unknown-linux-gnu

locale:
 [1] LC_CTYPE=en_US.iso885915
 [2] LC_NUMERIC=C
 [3] LC_TIME=en_US.iso885915
 [4] LC_COLLATE=en_US.iso885915
 [5] LC_MONETARY=C
 [6] LC_MESSAGES=en_US.iso885915
 [7] LC_PAPER=en_US.iso885915
 [8] LC_NAME=C
 [9] LC_ADDRESS=C
[10] LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.iso885915
[12] LC_IDENTIFICATION=C

attached base packages:
[1] tools     stats     graphics  grDevices utils
[6] datasets  methods   base

other attached packages:
 [1] genefilter_1.28.1
 [2] charmData_0.9.0
 [3] pd.feinberg.hg18.me.hx1_2.6.2
 [4] RSQLite_0.7-3
 [5] DBI_0.2-4
 [6] charm_0.9.27
 [7] snow_0.3-3
 [8] SQN_1.0
 [9] nor1mix_1.1-1
[10] mclust_3.3.2
[11] fields_6.01
[12] spam_0.15-5
[13] oligo_1.10.1
[14] preprocessCore_1.8.0
[15] oligoClasses_1.8.0
[16] Biobase_2.6.0

loaded via a namespace (and not attached):
 [1] ACME_2.2.0          affxparser_1.18.0
```

```
 [3] affyio_1.14.0       annotate_1.24.0
 [5] AnnotationDbi_1.8.1 Biostrings_2.14.8
 [7] bit_1.1-3           ff_2.1-1
 [9] gtools_2.6.1        IRanges_1.4.8
[11] MASS_7.3-3          multtest_2.2.0
[13] siggenes_1.20.0     spatial_7.3-1
[15] splines_2.10.0      survival_2.35-7
[17] xtable_1.5-6
```