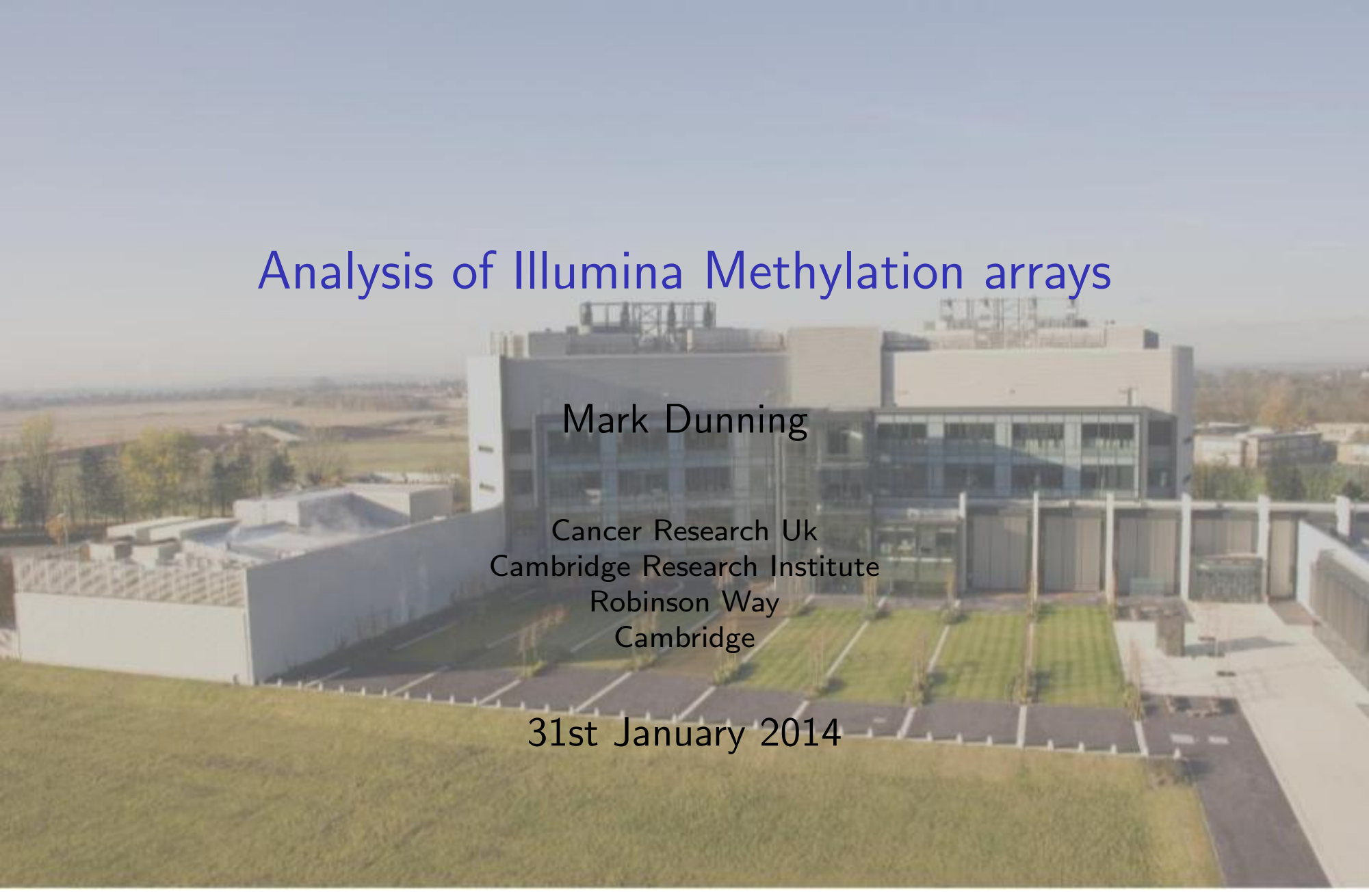


Analysis of Illumina Methylation arrays

Mark Dunning

Cancer Research Uk
Cambridge Research Institute
Robinson Way
Cambridge

31st January 2014

Outline

DNA Methylation

- A chemical modification of DNA structure that plays key role in regulating gene expression. Addition of *methyl* group.
- Occurs at *CpG* locations in the genome; C followed by G
- Areas with dense concentration of CpGs known as *CpG islands*
- Increased methylation known as *hypermethylation*. Decreased methylation known as *hypomethylation*
- Aberrant gains and losses of methylation reported in the progression of cancer.

Measuring DNA Methylation on microarrays

- Bisulphite-treat the sample to introduce mutations at unmethylated Cs
- Umethylated Cs converted to U
- Perform genotyping assay on two colour microarray.
- Compare methylated and unmethylated signal obtained.

Quantifying methylation

For each probe, we obtain measurements for the **M**ethylated and **U**nmethylated alleles. We define the methylation level of the probe, β, to be;

$$\beta = \frac{M}{U + M + 100} \tag{1}$$

β is the proportion of methylation for a given locus $0 < \beta < 1$

Technology evolution

Illumina only offer human methylation arrays

- GoldenGate
  - A cancer gene panel
  - 1,500 locations per sample, 96 samples per plate
  - GPL9183 4468 Samples 27 Datasets
- 27k
  - 27,000 locations per sample, 12 samples per chip
  - GPL8490 13965 Samples 238 Datasets
- 450k
  - 450,000 locations per sample, 12 samples per chip
  - GPL13534 8793 Samples 160 Datasets

TCGA also has over 6,000 samples

`http://cancergenome.nih.gov`

We will concentrate on the analysis of 450k data

- It is the currently available technology
- It is an area of active research
- Annual workshop
  - http://www2.cancer.ucl.ac.uk/medicalgenomics/tmorris/450k.html
- Online forum
  - http://tinyurl.com/o7lqty8
- At the moment, it is prefered to sequencing for methylation analysis

Unless specified otherwise, code will be from the `minfi` package

Different types of probe

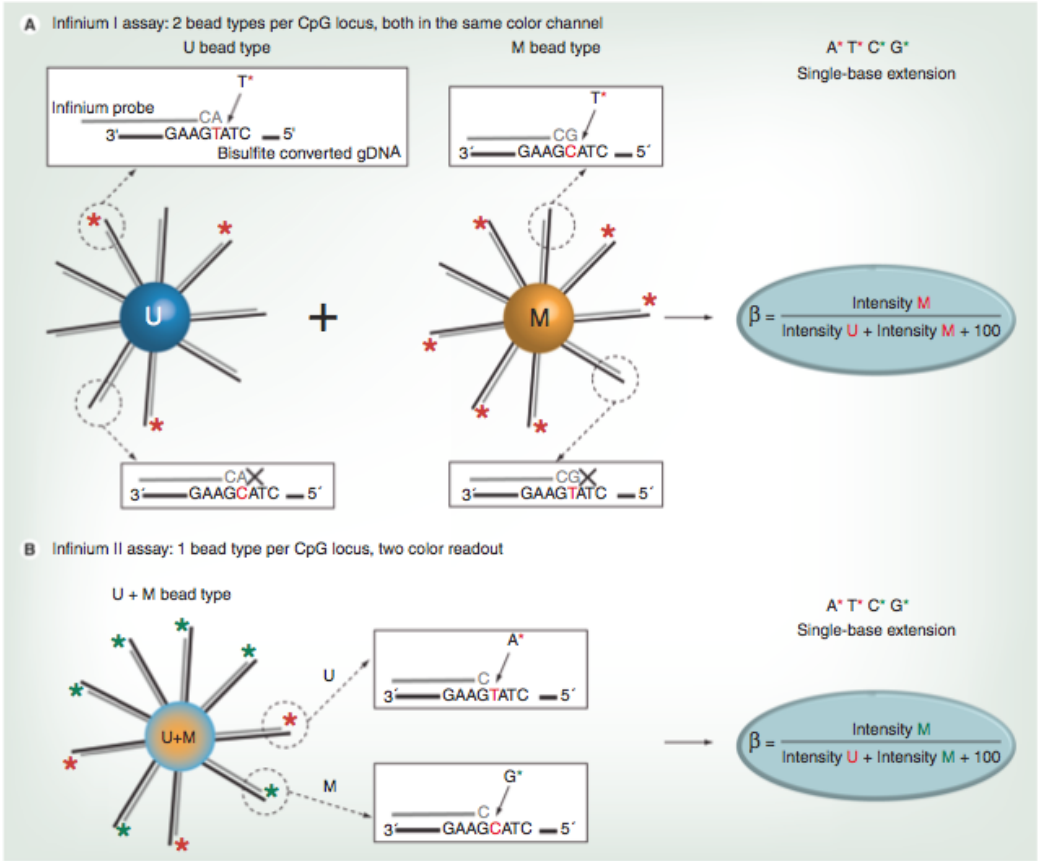


Figure 2. Overview of the Infinium I and Infinium II assays. (A) Infinium I and **(B)** Infinium II present on the Infinium Methylation 450K array.
M: Methylated; U: Unmethylated.

Dedeurwaerder et al. Evaluation of the Infinium Methylation 450K technology. Future Medicine

- Assigning methylation values to each loci is not a trivial task
- For Type I design a pair of probes measure methylated and unmethylated in the *either* the red or green channel
- For Type II, a single probe measures methylated in the red channel and unmethylated in the green.

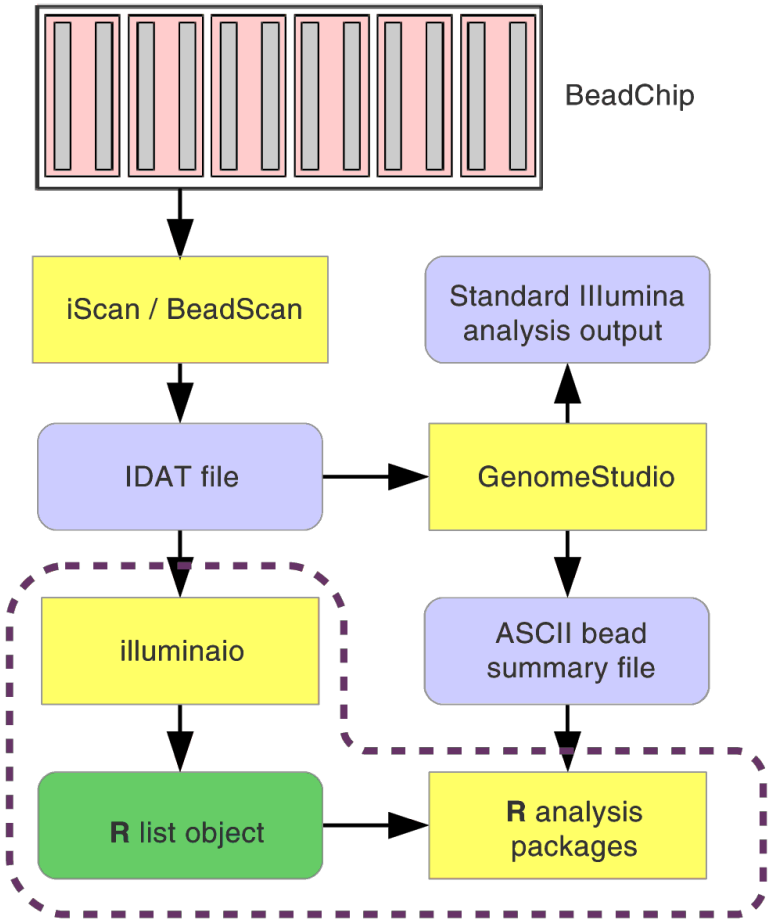
- Type II probes can only tolerate three CpGs within the probe
- Type I probes tolerate more, but assumes that all methylation loci have the same state. i.e. For a 'methylated' probe all CpGs in the probe assumed to be methylated.
- Type I probes used in regions of high CpG density. e.g .CpG islands.
- Earlier 27k technology used exclusively Type I probes.

	Probe design		
Region type	I	II	**Total**
CpG Island	77,674	72,580	150,254
CpG Island Shore	22,371	89,696	112,067
CpG Island Shelf	6,913	40,231	47,144
Open sea	28,518	147,529	176,047
Total	135,476	350,036	485,512

Aryee et al (2014) Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA Methylation microarrays

idat format

450k arrays are commonly analysed from *idat* files
http://f1000research.com/articles/2-264/v1



Each sample has a Red and Green idat file

```
## [1] "5723646052_R02C02_Grn.idat" "5723646052_R02C02_Red.idat"
## [3] "5723646052_R04C01_Grn.idat" "5723646052_R04C01_Red.idat"
## [5] "5723646052_R05C02_Grn.idat" "5723646052_R05C02_Red.idat"
```

A targets file / sample sheet is used to define the samples

```
## [read.450k.sheet] Found the following CSV files:
## [1] "/Users/dunnin01/Library/R/3.0/library/minfiData/extdata/SampleS
##   Sample_Name Sample_Well Sample_Plate Sample_Group Pool_ID person a
## 1    GroupA_3          H5           NA       GroupA      NA    id3
## 2    GroupA_2          D5           NA       GroupA      NA    id2
##   status  Array     Slide
## 1 normal R02C02 5.724e+09
## 2 normal R04C01 5.724e+09
##
## 1 /Users/dunnin01/Library/R/3.0/library/minfiData/extdata/5723646052
## 2 /Users/dunnin01/Library/R/3.0/library/minfiData/extdata/5723646052
```

First, we retrieve the red and green intensities from the idat files

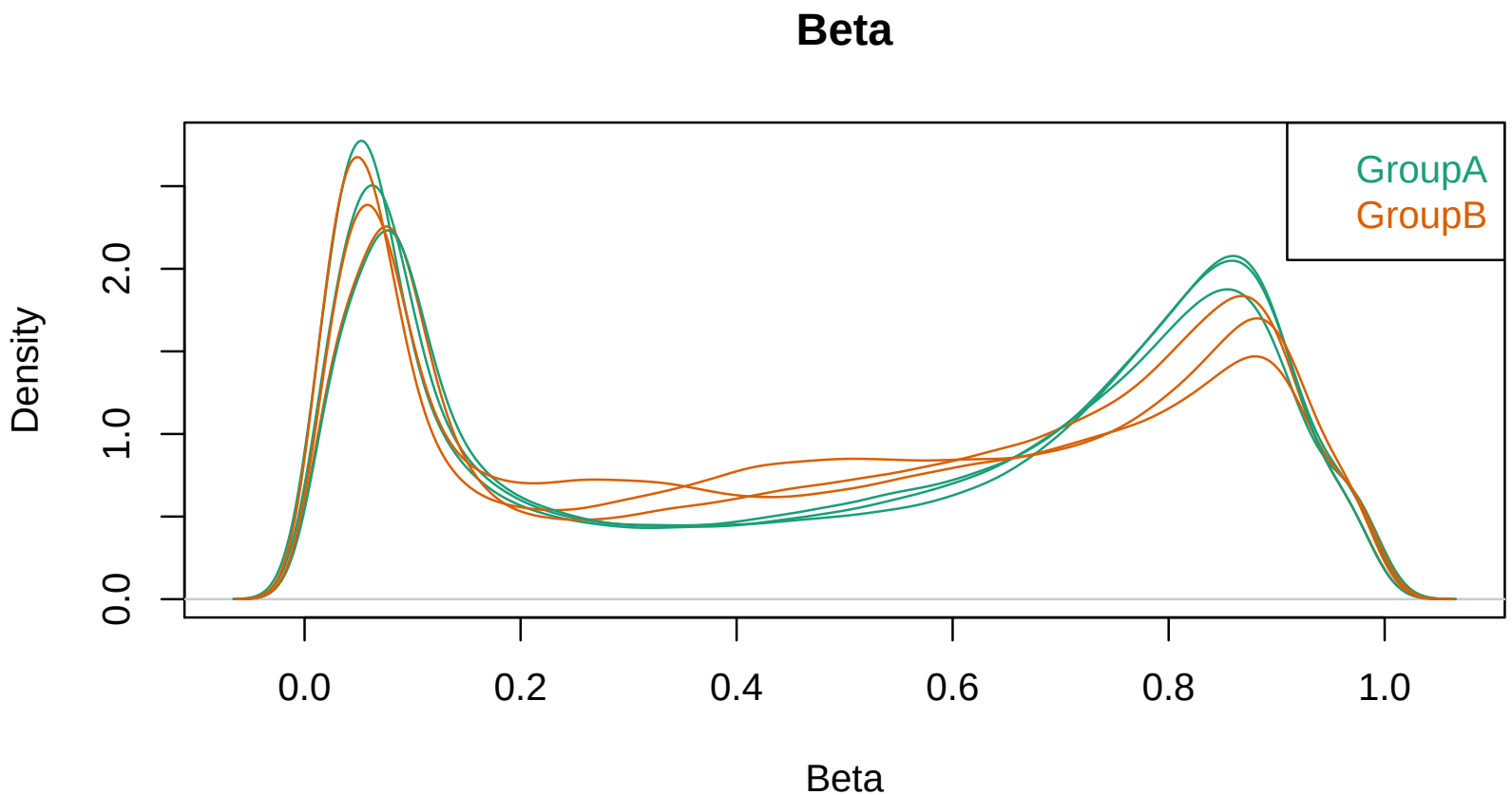
```
RGset <- read.450k.exp(base = baseDir, targets = targets)
```

```
##          5723646052_R02C02 5723646052_R04C01 5723646052_R05C02
## 10600313               415               394               272
## 10600322              9685             11737             11343
## 10600328              1647              1953              1998
## 10600336              3680              6290             16109
## 10600345              3616              4730              2904
## 10600353              4578              5399              4958
##          5723646053_R04C02 5723646053_R05C02 5723646053_R06C02
## 10600313               356               455               356
## 10600322              9262             12883              7176
## 10600328              2022              2451              1938
## 10600336             16020              7650              3621
## 10600345              3150              4579              2075
## 10600353              4499              4955              4134
```

QC of samples

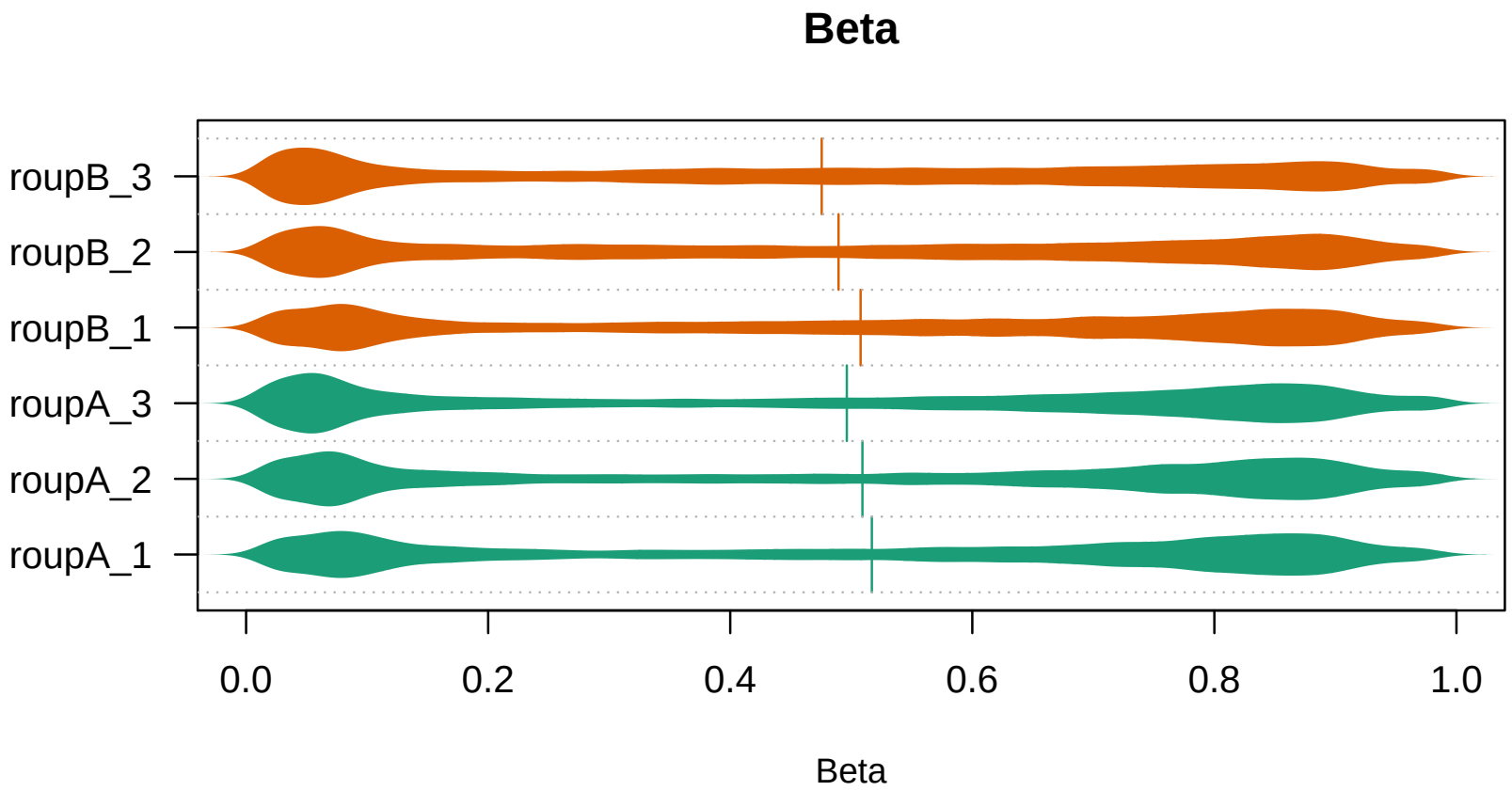
```
pd <- pData(RGset)
densityPlot(RGset, sampGroups = pd$Sample_Group, main = "Beta")

## Loading required package:
IlluminaHumanMethylation450kmanifest
```



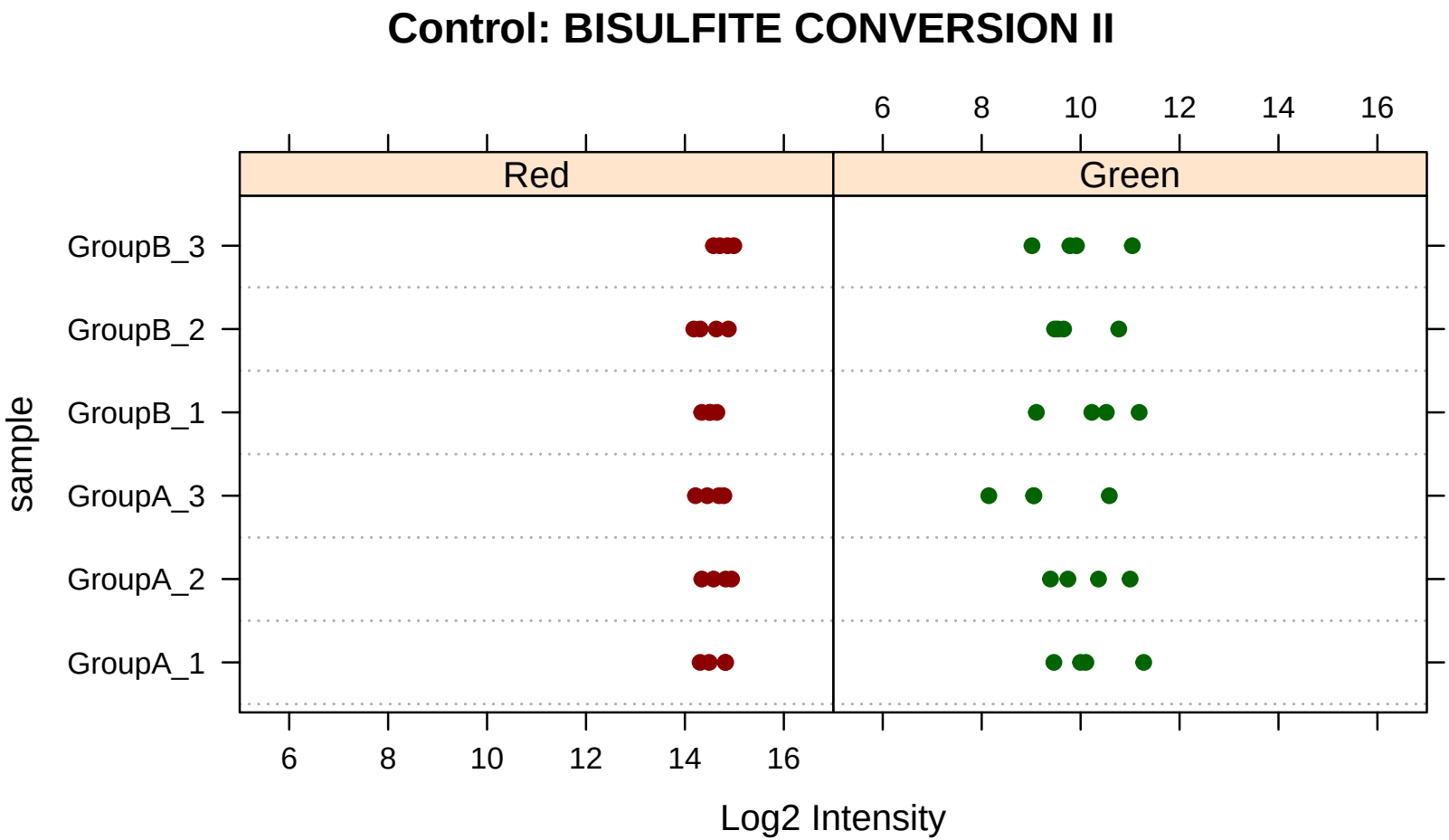
QC of samples

```
densityBeanPlot(RGset, sampGroups = pd$Sample_Group, sampNames = pd$Sam
```



QC of samples

```
controlStripPlot(RGset, controls = "BISULFITE CONVERSION II", sampNames
```



minfi QA

Aryee et al (2014) Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA Methylation microarrays

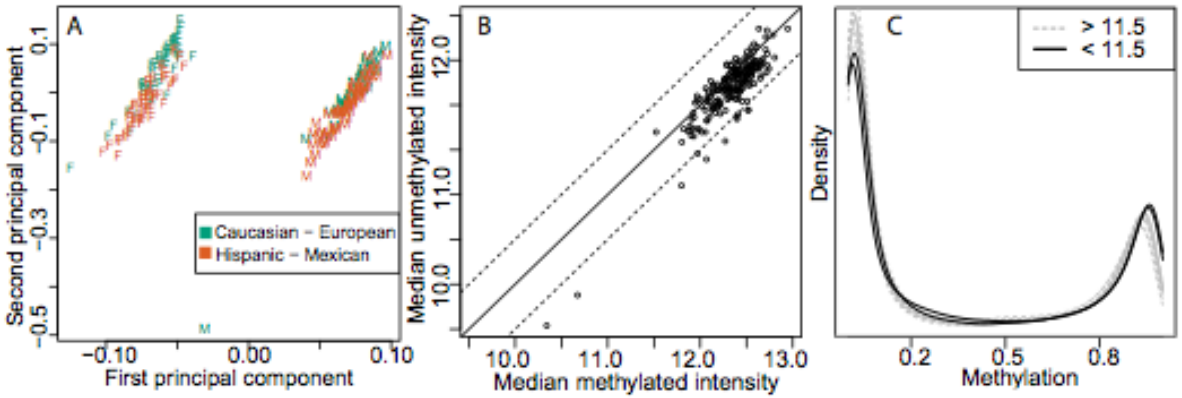


Figure 4: Quality assessment plots based on the blood sample dataset. A) A multidimensional scaling (MDS) plot. Color represents reported ethnicity. B) Scatter plot of median Unmeth signal vs median Meth signal value for each sample. Points outside the dashed lines represent cases were the differences are larger than 0.5. C) Beta density plots for all samples with black curves representing samples where the average of the median Unmeth and Meth is below 11.5.

Sample identity can also be confirmed using 65 SNP probes and gender inferred from the data

- Probes that are not consistently *detected* can be discarded.
  - e.g. $> 25\%$ with detection p-value < 0.05
- Annotation considerations
  - Cross-hybridisation
  - Autosomal effects
  - Probes that include SNPs may be influenced by sample genotype

Detection filtering

```
detP <- detectionP(RGset)
failed <- detP > 0.05
```

Fraction of failed positions per sample

```
colMeans(failed)

## 5723646052_R02C02 5723646052_R04C01 5723646052_R05C02 5723646053_R04
##         0.0006406         0.0021132         0.0056353         0.0026
## 5723646053_R05C02 5723646053_R06C02
##         0.0022759         0.0227595

colMeans(failed) > 0.01

## 5723646052_R02C02 5723646052_R04C01 5723646052_R05C02 5723646053_R04
##             FALSE             FALSE             FALSE              FA
## 5723646053_R05C02 5723646053_R06C02
##             FALSE              TRUE
```

How many positions failed in 50% of samples?

```
sum(rowMeans(failed) > 0.5)
```

Preprocess the intensities

Convert the red and green intensities into Methylated and Unmethylated values using a *manifest* package

```
MSet.raw <- preprocessRaw(RGset)
MSet.raw

## MethylSet (storageMode: lockedEnvironment)
## assayData: 485512 features, 6 samples
##   element names: Meth, Unmeth
## phenoData
##   sampleNames: 5723646052_R02C02 5723646052_R04C01 ...
##     5723646053_R06C02 (6 total)
##   varLabels: Sample_Name Sample_Well ... filenames (13 total)
##   varMetadata: labelDescription
## Annotation
##   array: IlluminaHumanMethylation450k
##   annotation: ilmn12.hg19
## Preprocessing
##   Method: Raw (no normalization or bg correction)
##   minfi version: 1.8.9
##   Manifest version: 0.4.0
```

Background correction

Can mimic the background correction steps used by Illumina in GenomeStudio

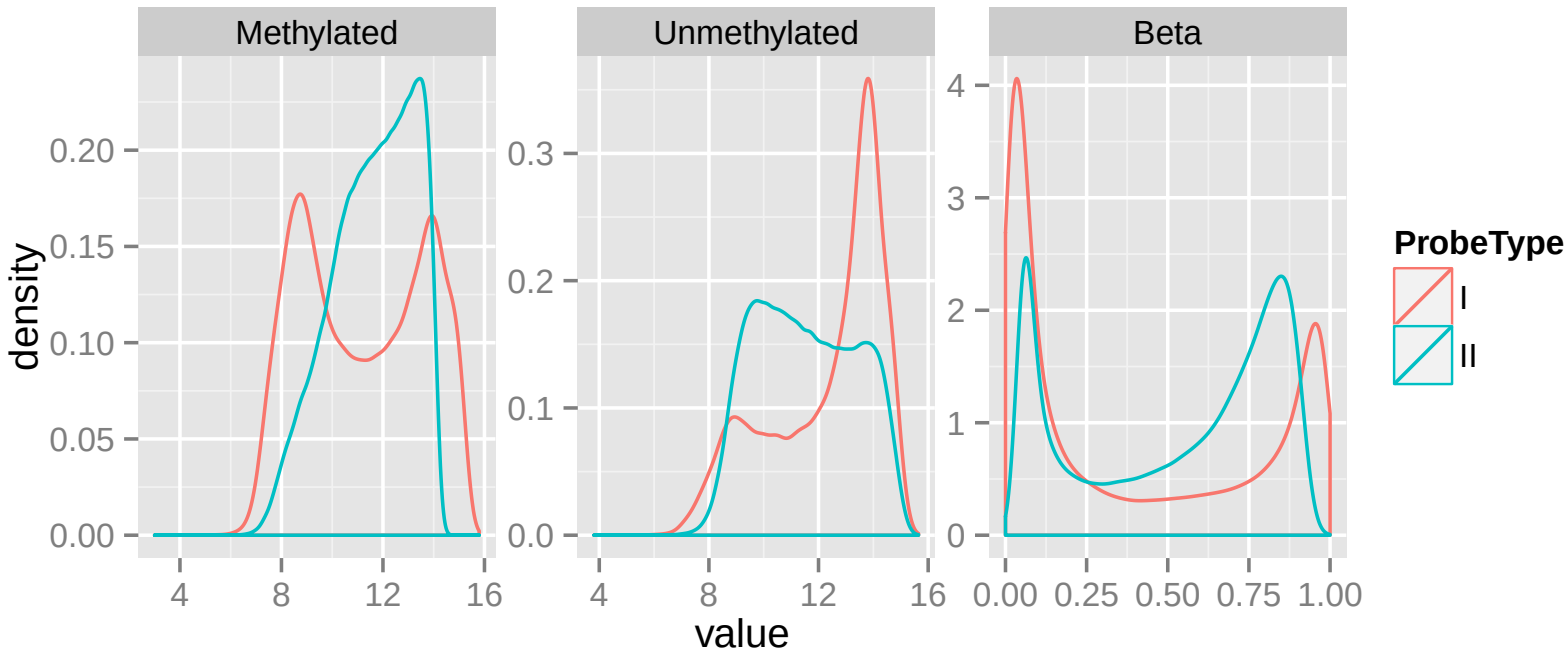
```
MSet.bg <- preprocessIllumina(RGset)
args(preprocessIllumina)

## function (rgSet, bg.correct = TRUE, normalize = c("controls",
##     "no"), reference = 1)
## NULL
```

Distribution of probe types

```
##
##      I     II
## 135476 350036
```

Differences in signal



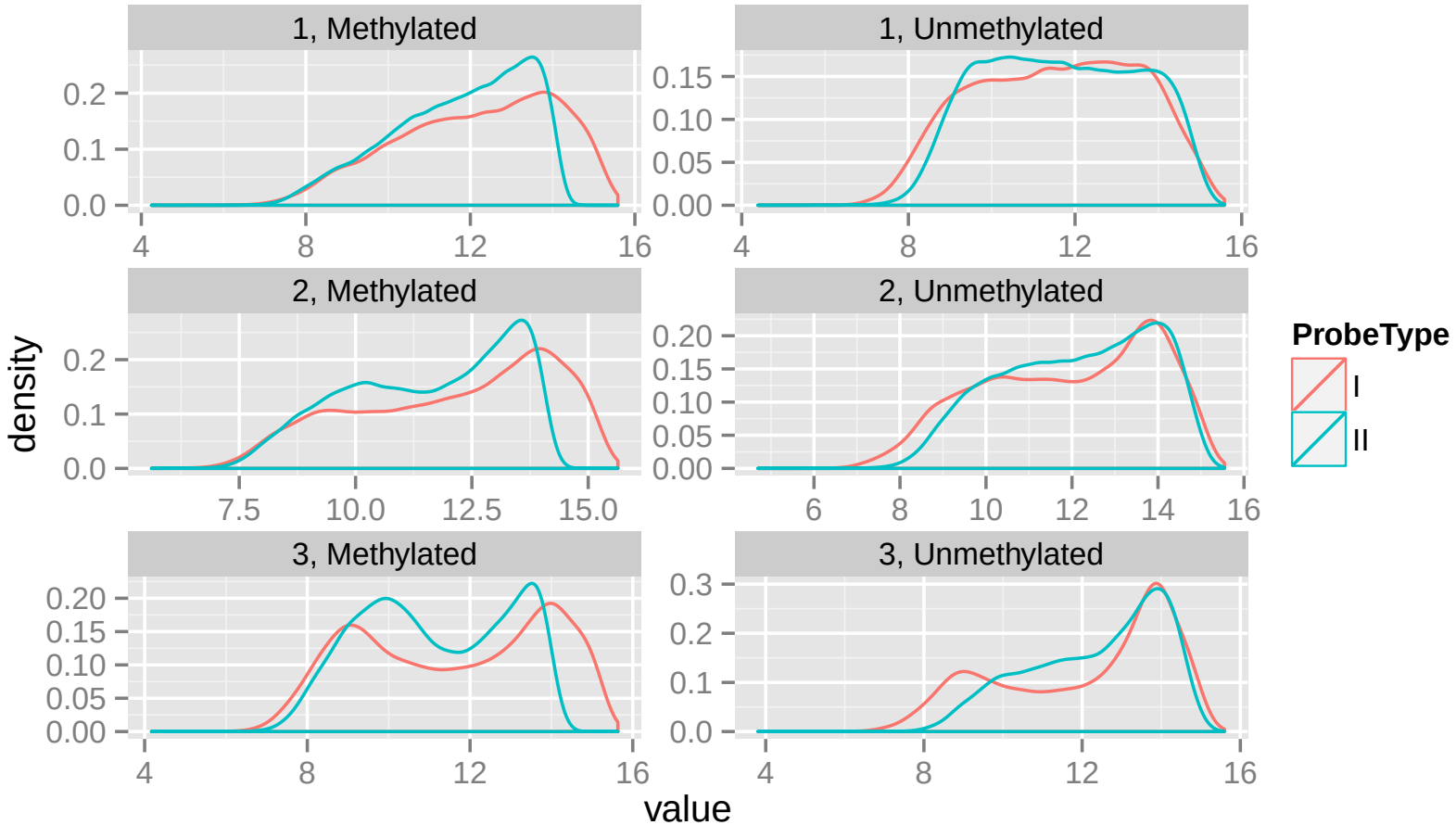
Problem definition

- TypeI and TypeII probes are shown to have different distributions
- They also target different genomic regions
- Need to use separate analysis and normalisation

Dedeurwaerder et al

SWAN

Maksimovic et al. Genome Biology 2012, 13:R44
Observed that for specific number of CpGs within the probe, distributions are comparable



Hence, they apply quantile normalisation to each subset.

Other methods

- BMIQ
  - Beta Mixture Quantile Dilation
  - Does not use assumptions about biological characteristics to select subsets of the data
  - Implemented in wateRmelon
- Peak-based correction
  - Rescale TypeII probes based on TypeI assuming a bimodal shape
  - Implemented in IMA package
  - Does not work well when distribution does not exhibit well-defined peaks

Quantify Methylation

The standard is to use β values

```
M <- getMeth(MSet.raw)
U <- getUnmeth(MSet.raw)
beta <- getBeta(MSet.raw)
beta[1:5, 1:2]
```

```
##            5723646052_R02C02 5723646052_R04C01
## cg00050873           0.91891            0.5759
## cg00212031           0.09371            0.6548
## cg00213748           0.80838            0.4773
## cg00214611           0.08443            0.7704
## cg00455876           0.78155            0.3345
```

```
M[1:5, 1:2]/(M[1:5, 1:2] + U[1:5, 1:2])
```

```
##            5723646052_R02C02 5723646052_R04C01
## cg00050873           0.91891            0.5759
## cg00212031           0.09371            0.6548
## cg00213748           0.80838            0.4773
## cg00214611           0.08443            0.7704
## cg00455876           0.78155            0.3345
```

To be(ta) or not to be(ta)

- beta has a more natural interpretation
- log-ratios are more-ameanable for analysis

Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. Du et al. BMC Bioinformatics

$$LR = log_2 \frac{M}{U} \tag{2}$$

Sometimes (confusingly) called *M-values*

Conversion

The two measures can be converted easily

$$\beta = \frac{2^M}{2^M + 1} \tag{3}$$

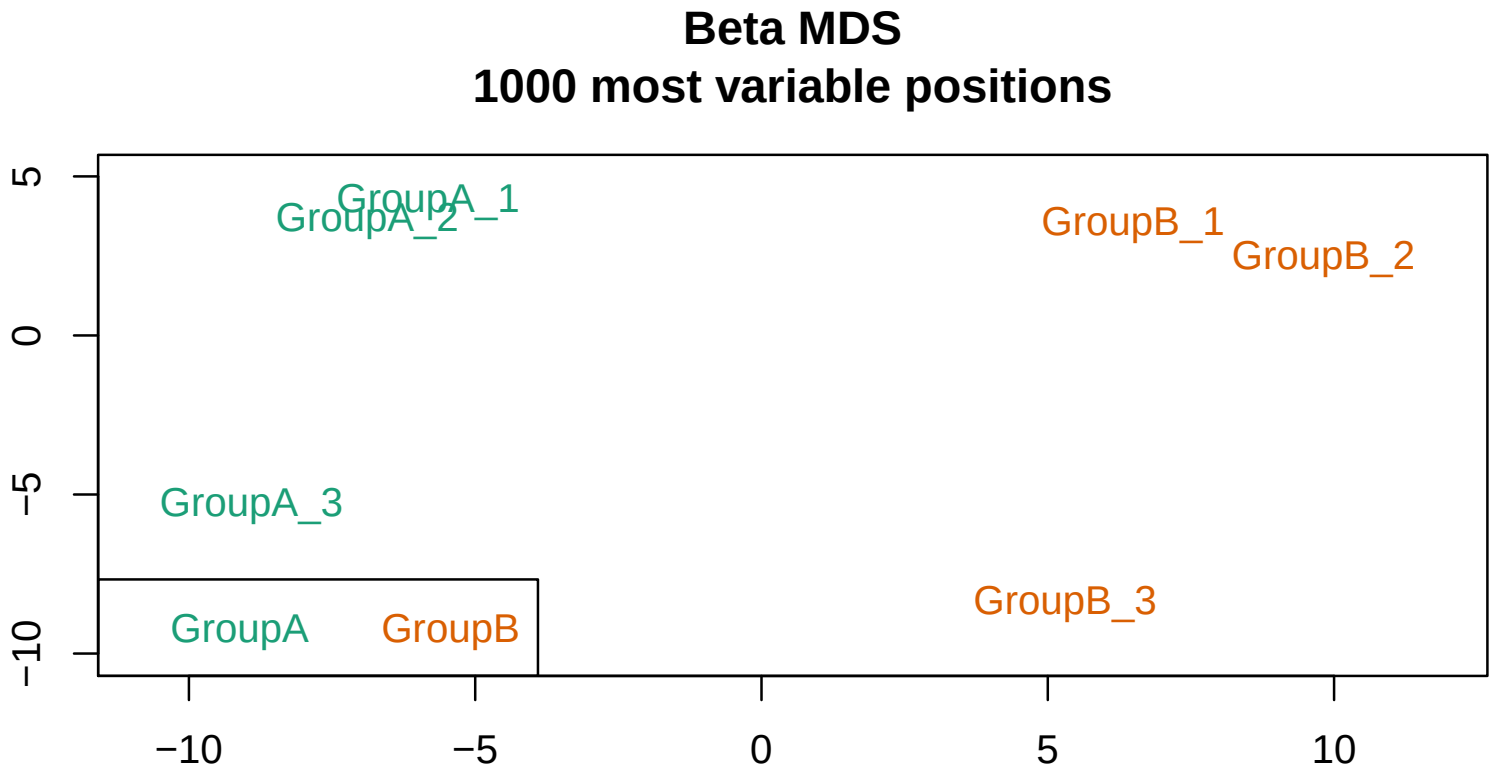
$$M = log_2 \frac{\beta}{1 - \beta} \tag{4}$$

Batch effects

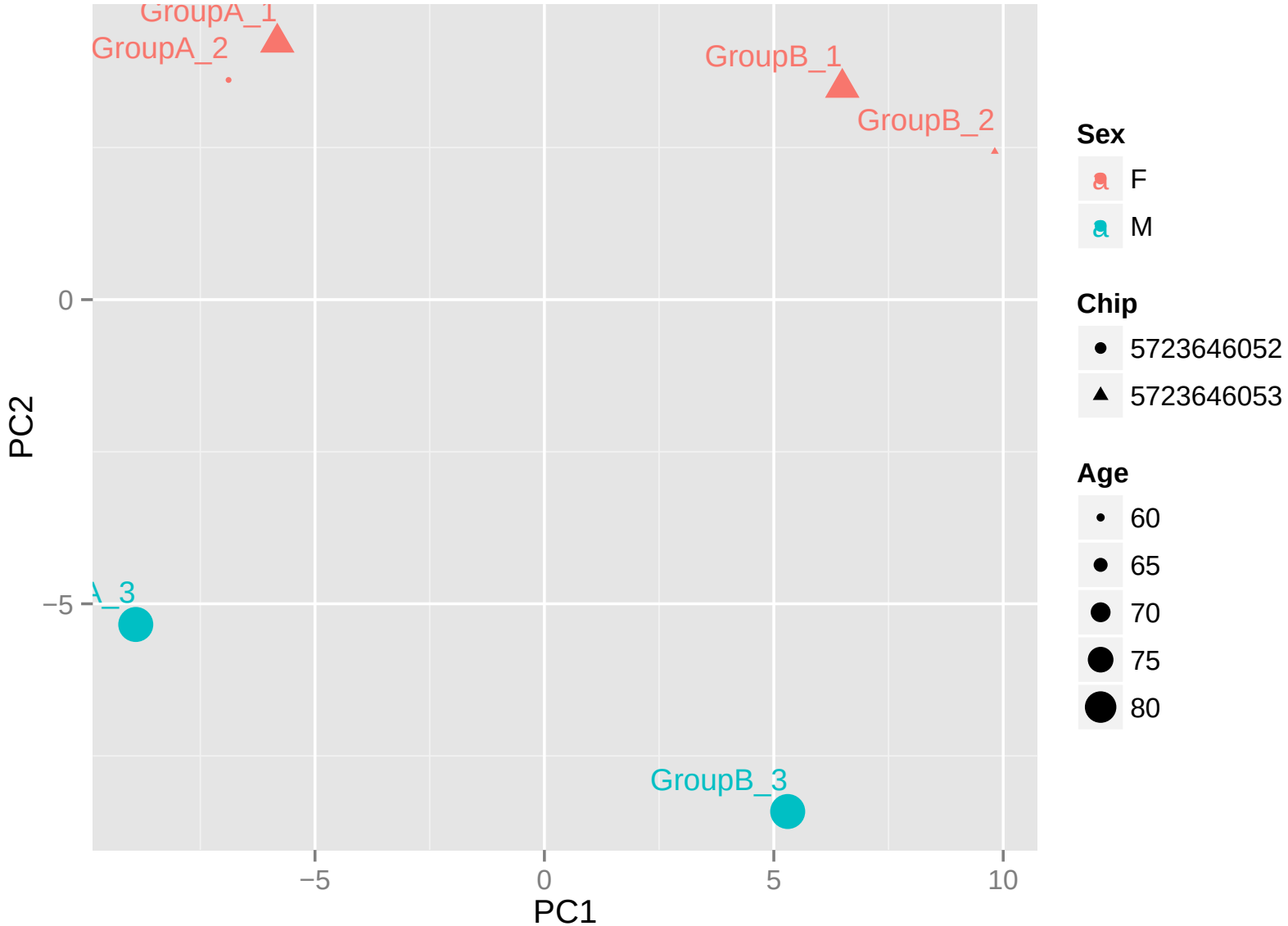
'it seems that batch effects are almost always present in large-scale Infinium data sets, and they can introduce severe bias during subsequent analysis steps if no adequate countermeasures are taken.' Bock. Analysing and interpreting DNA methylation data. Nature Reviews Genetics

Visualisation

Multi-dimensional scaling (*MDS*) plots are useful for assesing sample relations in a similar way to PCA.



Check other covariates



Correction - ComBAT

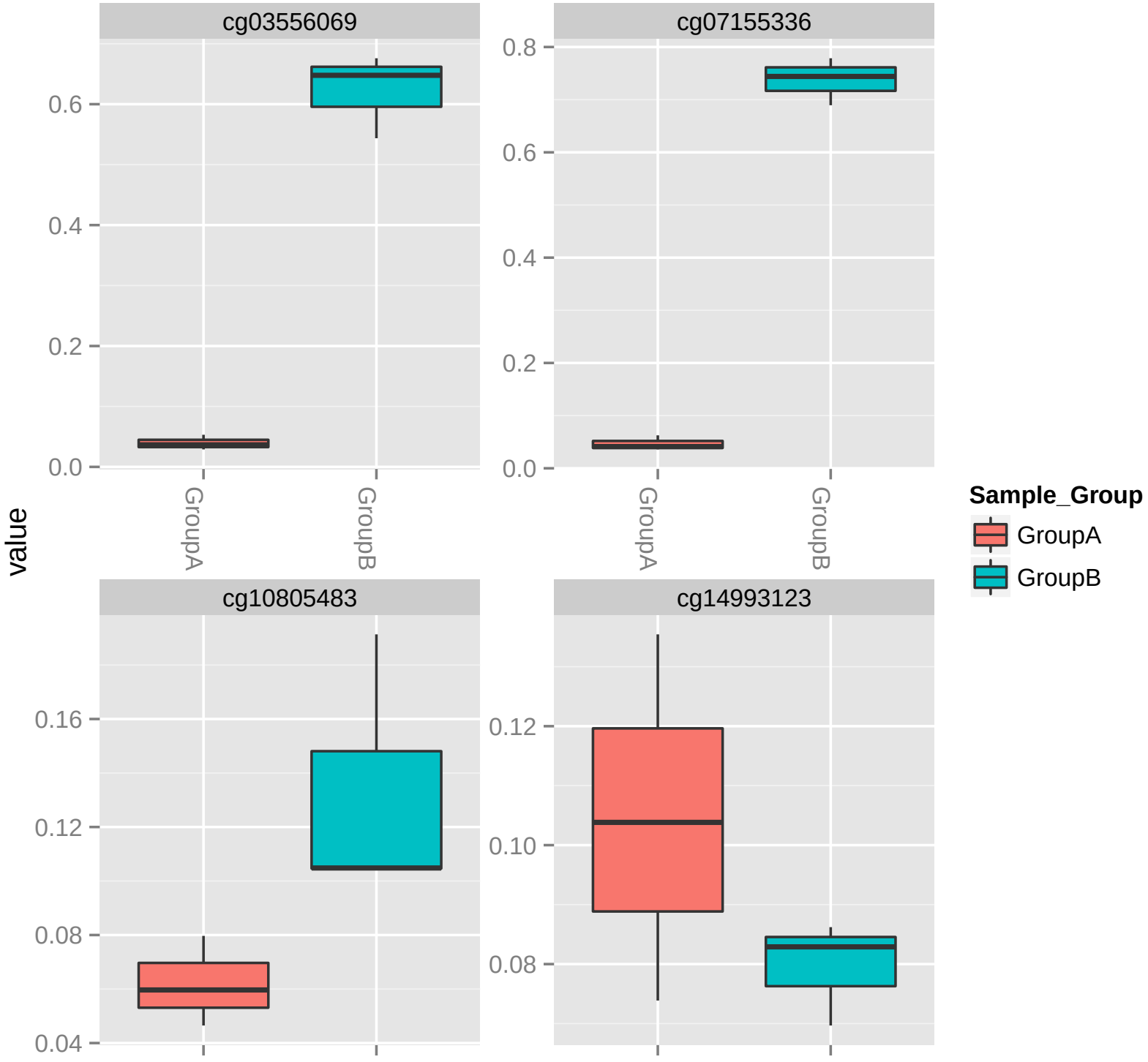
- The current favourite method seems to be *ComBAT*; implemented in the `sva` package.
- Uses an *empirical bayes* framework
- Requires that you have an adjustment variable (e.g. processing data) that you want to correct
- Also specify the variable of interest in the experiment (e.g. tumour vs normal)
- **No substitute for poorly-designed experiments!**

The `dmpFinder` method in `minfi` uses an *F-test* to find differentially methylated postions between groups.

- Log-ratios are used in the analysis
- The model is fitted using limma
- Variance-shrinkage recommended for small sample-size

```
dmp <- dmpFinder(M, pheno = pd$Sample_Group, type = "categorical",
    shrinkVar = T)
head(dmp)
```

```
##            intercept     f      pval    qval
## cg07155336     7.637 295.8 3.248e-06 0.04222
## cg03556069     7.068 231.3 6.555e-06 0.04222
## cg10805483     5.234 202.2 9.620e-06 0.04222
## cg14993123    -5.643 174.9 1.452e-05 0.04222
## cg08474164    -7.219 171.1 1.543e-05 0.04222
## cg20386875    -4.530 152.9 2.122e-05 0.04222
```


cg03556069
cg07155336
cg10805483
cg14993123
value
0.0
0.2
0.4
0.6
0.8
0.04
0.08
0.12
0.16
0.10
GroupA
GroupB
Sample_Group
GroupA
GroupB

'bump-hunting'

Larger Differentially Methylated Regions can be identified

Published by Oxford University Press on behalf of the International Epidemiological Association

International Journal of Epidemiology 2
doi:10

Bump hunting to identify differentially methylated regions in epigenetic epidemio studies

Andrew E Jaffe,[1,2,3] Peter Murakami,[3] Hwajin Lee,[3] Jeffrey T Leek,[1] M Daniele Fallin,[1,2,3] Andrew P Feinberg[1,3,4] and Rafael A Irizarry[1,3]*

[1]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA, [2]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA, [3]Center for Epigenetics, Johns Hopkins School of Medicine, Baltimore, MD, USA and [4]Department of Medicine, Johns Hopkins School of M Baltimore, MD, USA

*Corresponding author. Department of Biostatistics, 615 N. Wolfe St E3620, Baltimore, MD 21205. E-mail: rafa@jhu

also available in minfi

Available software

Processing and analysis of DNA methylation data.
Wilhelm-Benartzi et al. (2013)

Table 1. R/Bioconductor packages for the processing and analysis of array-based DNA methylation data

DNA methylation processing/analysis step	R/Bioconductor packages
Quality control samples	IMA, HumMethQCReport, methylkit, MethyLumi, preprocessing and analysis pipeline, minfi
Quality control probes	IMA, HumMethQCReport, lumi, LumiWCluster, preprocessing and analysis pipeline, wateRmelon
Background correction	Limma, lumi, MethyLumi, minfi, preprocessing and analysis pipeline
Normalisation	Combat[a], HumMethQCReport, lumi, minfi, TurboNorm, MethyLumi, wateRmelon
Type 1 and 2 probe scaling	IMA, minfi, wateRmelon
Batch/plate/chip/confounder adjustment	Combat[a], CpGassoc, ISVA, MethLAB
Data dimension reduction	MethyLumi
Differential methylation analysis/region-based analysis	CpGAssoc, IMA, limma, methylkit, MethLAB, MethVisual, minfi, EVORA
Clustering/profile analysis	Lumi, ISVA, HumMeth27QCReport, methylkit, RPMM, SS-RPMM[b]
Multiple testing correction	CpGAssoc, methylkit, MethLAB, NHMMfdr

[a]Freely available for download: http://www.bu.edu/jlab/wp-assets/ComBat/Abstract.html.
[b]Freely available for download: http://bio-epi.hitchcock.org/faculty/koestler.html.

ChAMP

- Automated workflow
- Data read using minfi from idat files
- Choice of normalization
- Visualisation and correction of batch effects
- CNA analysis