BINFF-402 — Genomics, Proteomics and Evolution
Assignment 1

Robin Petit

December 11, 2017

1 Introduction

The Illumina Infinium 450K sequencer [1] was designed in order to sequence DNA methylation. It works by first applying bisulfite transformation to the DNA fragments to transform unmethylated Cytosines to Uracils, and leave methylated Cytosines unchanged, and then sequencing uses two different bead types: one to determine methylated loci and one to determine unmethylated loci. [2]

The Infinium HumanMethylation450 BeadChip contains over 4.5e5 methylation sites probes, and two different chemical assays: Infinium I and Infinium II. In this assignment, only Infinium I is considered. Yet, Infinium II should also be considered in analyses, but separately. [3]

In this document, methylation level is studied in the case of a DKO of genes DNMT1 and DNMT3b (respectively OMIM ids 126375 and 602900) [4] which are both methyltransferases. This double knockout induces a loss of hypermethylated CpG islands which is not present in case of single knockout of either of these genes. [5]

The study has been performed with the R language. Note that all code has been written from scratch even though there is a wide range of available packages for bioinformatics in R [6]. See for instance the package IMA that is made on purpose for Infinium data analysis [7].

All the R code used for this report can be found at `https://github.com/RobinPetit/BINF-F402`.

2 Analysis

2.1 β-value distribution

The β distribution is shown in Figure 1. For each probe, the β value is computed to be $\frac{M}{M+U+\alpha}$, where M is the methylated score, U is the unmethylated score, and α acts as a pseudo-count to avoid the case $M = U = 0$ to lead to a division by zero. This α value has been set to 100. [8]

It is clear that controls have a two-spikes β distribution: most probes have a methylation β-value around 0 or around 1. A high methylation β-value (close to 1) means that almost all of the sequenced cells had a methylated Cytosine at this position, whereas a low methylation β-value (close to 0) means that almost none of the sequenced cells had a methylated Cytosine at this position. Yet, as several different cells are sequenced at the same time, it is possible to have a Cytosine that is methylated in some cells, but not in others. This explains how β-value can take so many different values.

It is also clear that cases have a single-spike β distribution: most probes have a methylation β-value close to 0, but almost none has a β-value close to 1. Also, there is a higher density of intermediate β-values, i.e. between 0.2 and 0.8, which means that there is still methylation happening in the cases, but the methylation of DNA varies from one cell to another.

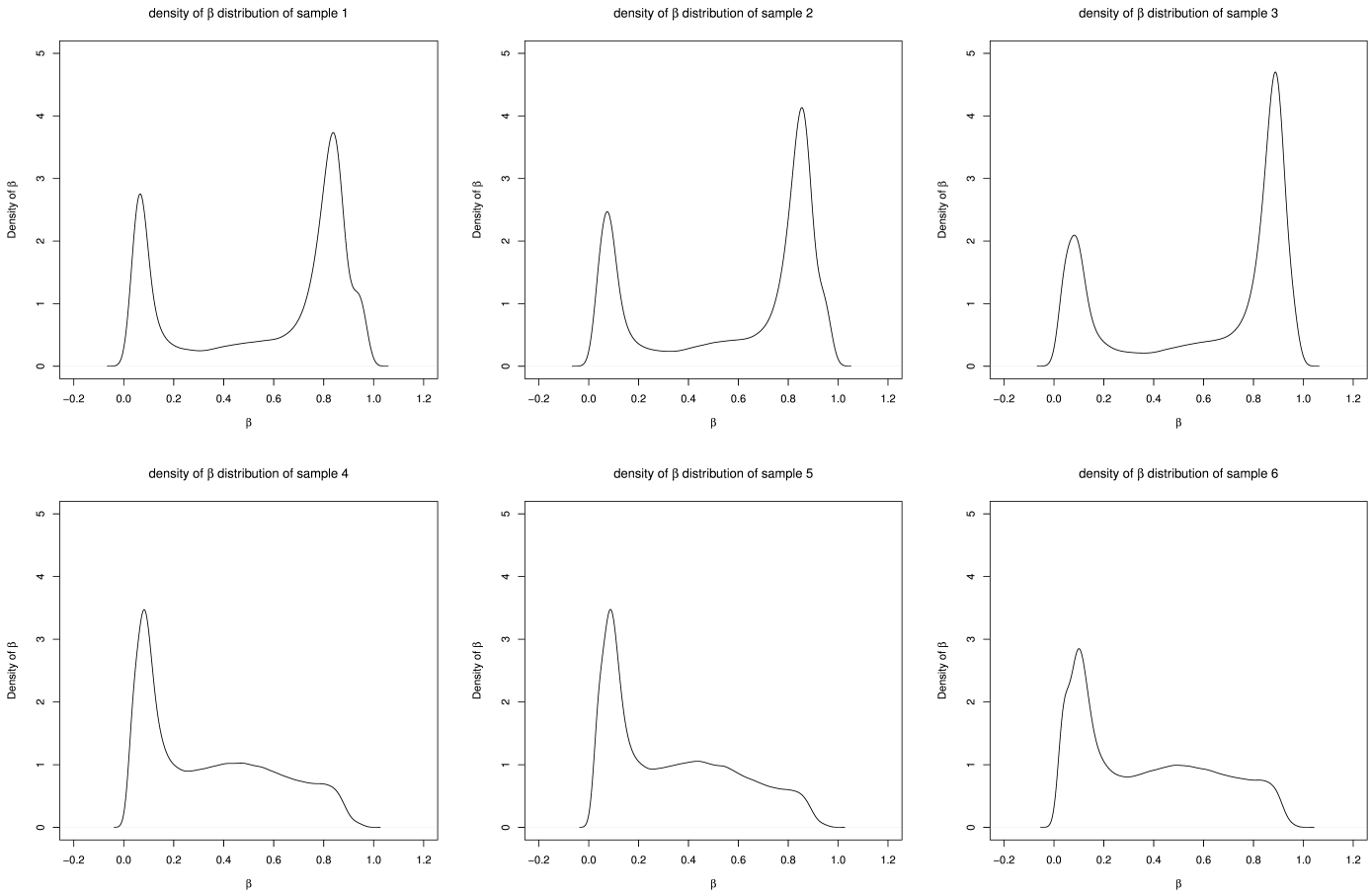


Figure 1: Distribution of the β value for each sample. Samples 1 to 3 are the controls, and samples 4 to 6 are the cases.

2.2 Global overview

An analysis of the methylation level of each chromosome is shown in Figure 2a. We can observe that the hypomethylation is present on each chromosome, with highest hypomethylation being on chromosomes 10, 16 and 20 and lowest hypomethylation is on chromosome Y, probably because the Y chromosome has a lower methylation level in control samples. Yet, chromosome X is also less hypomethylated than the other chromosomes, but sum of difference in methylation level for X and Y chromosomes reaches roughly the same value as the other chromosomes.

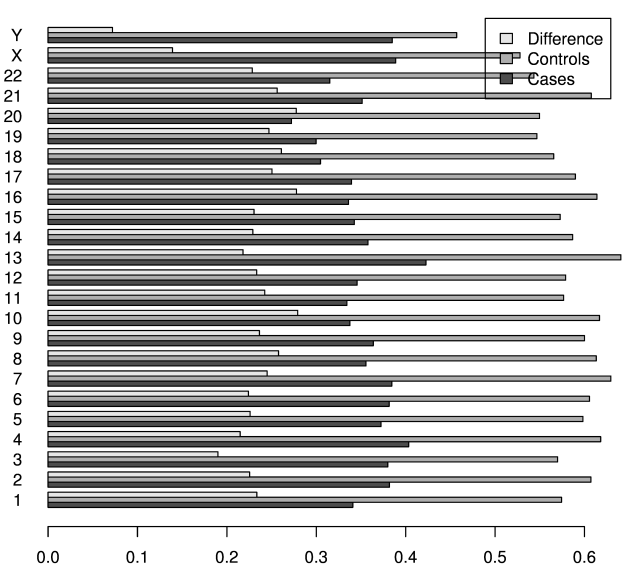
The distribution of the difference in methylation level per chromosome seems to be pretty uniform between 0.2 and 0.25.

Figure 2b shows the methylation level of chromosomes 19 to 21 as stated in the instructions.

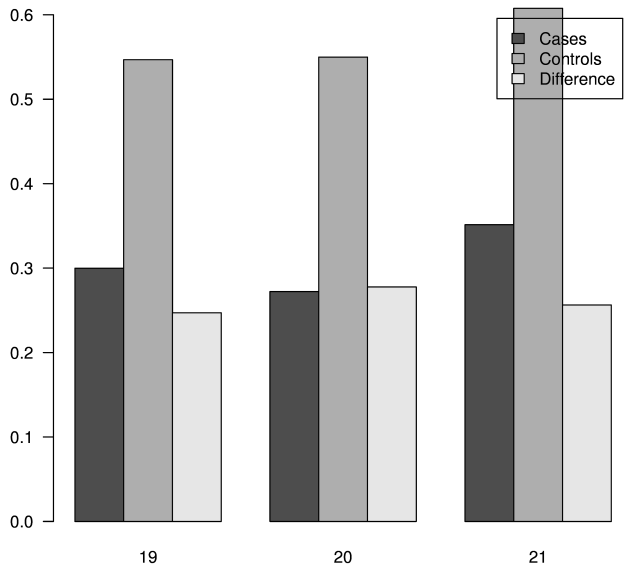
2.3 Number of probes per gene

The distribution of the number of probes per gene is shown in Figure 3: Figure 3a shows the distribution in which we observe that almost all of the genes have only a few probes related to them, so Figure 3 shows the same data in y log-scale.

Yet, as displayed on Figure 3b, the average number of probes per gene is around 32.36. This mean value is impacted by the very high values that are 2199, 2404, and 3897, but since only one such gene exist for each of these values, the impact is very low, due to the big amount of probes: 485577 (computed by either `nrow(infinium)` in the R code, or by `wc -l annotations.csv` minus 1 to remove the header).

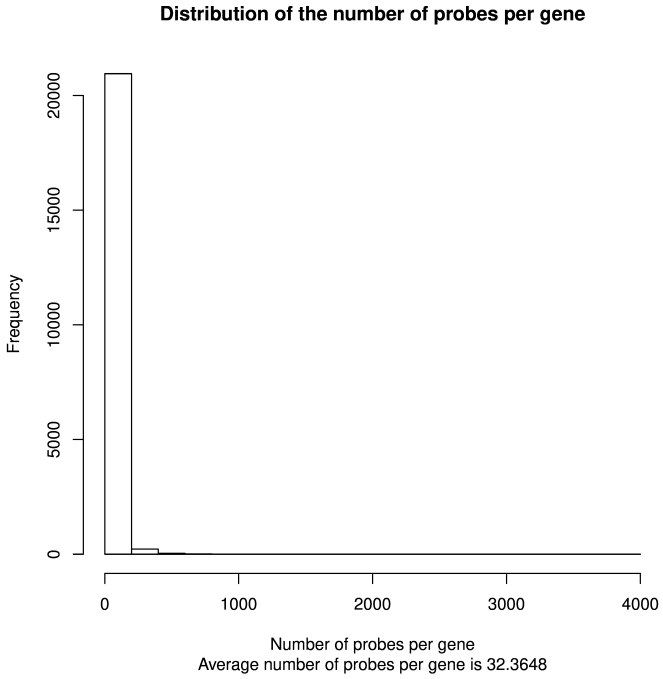


(a) Methylation level on each chromosome (1 to 22 plus X and Y).

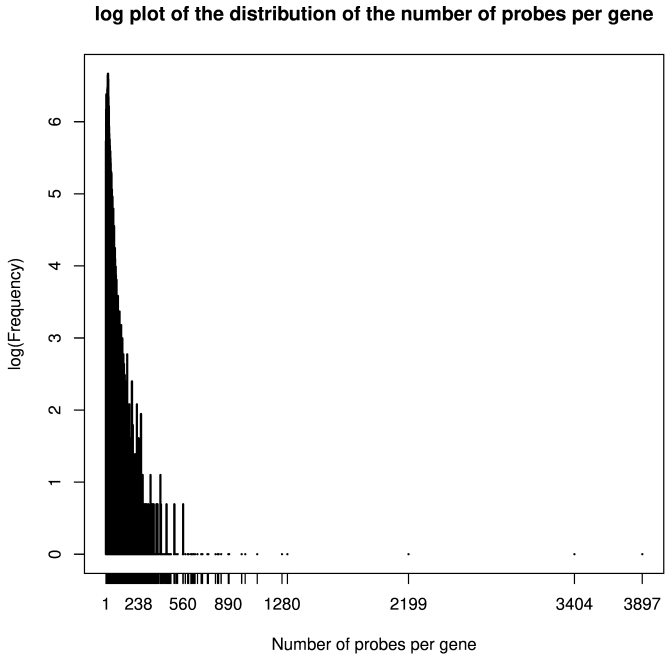


(b) Subplot of Figure 2 of chromosomes 19 to 21.

Figure 2: General analysis of the methylation level per chromosome for both controls and cases samples. Difference in methylation is defined as control minus case since cases are hypomethylated.



(a) Distribution of the number of probes per gene.



(b) Log-distribution of the number of probes per gene.

Figure 3: Distribution of the number of probes per gene in both linear and log scale.

2.4 $\Delta\beta$

The distribution of $\Delta\beta$ (control - case) is shown in Figure 4. The three first subplots represent the $\Delta\beta$ distribution of each couple (control, case), but the most interesting part is the last subplot showing the $\Delta\beta$ in average for each control and each case sample. We observe no hypermethylation ($\Delta\beta > 0$) but a clear hypomothylation ($\Delta\beta < 0$) as expected.

Still, most of the probability mass is around $\Delta\beta = 0$. This comes from the fact that even though case samples are hypomethylated, they still contain methlyated regions (see Figure 1 and Figure 2), for which $\Delta\beta$ is close to 0.

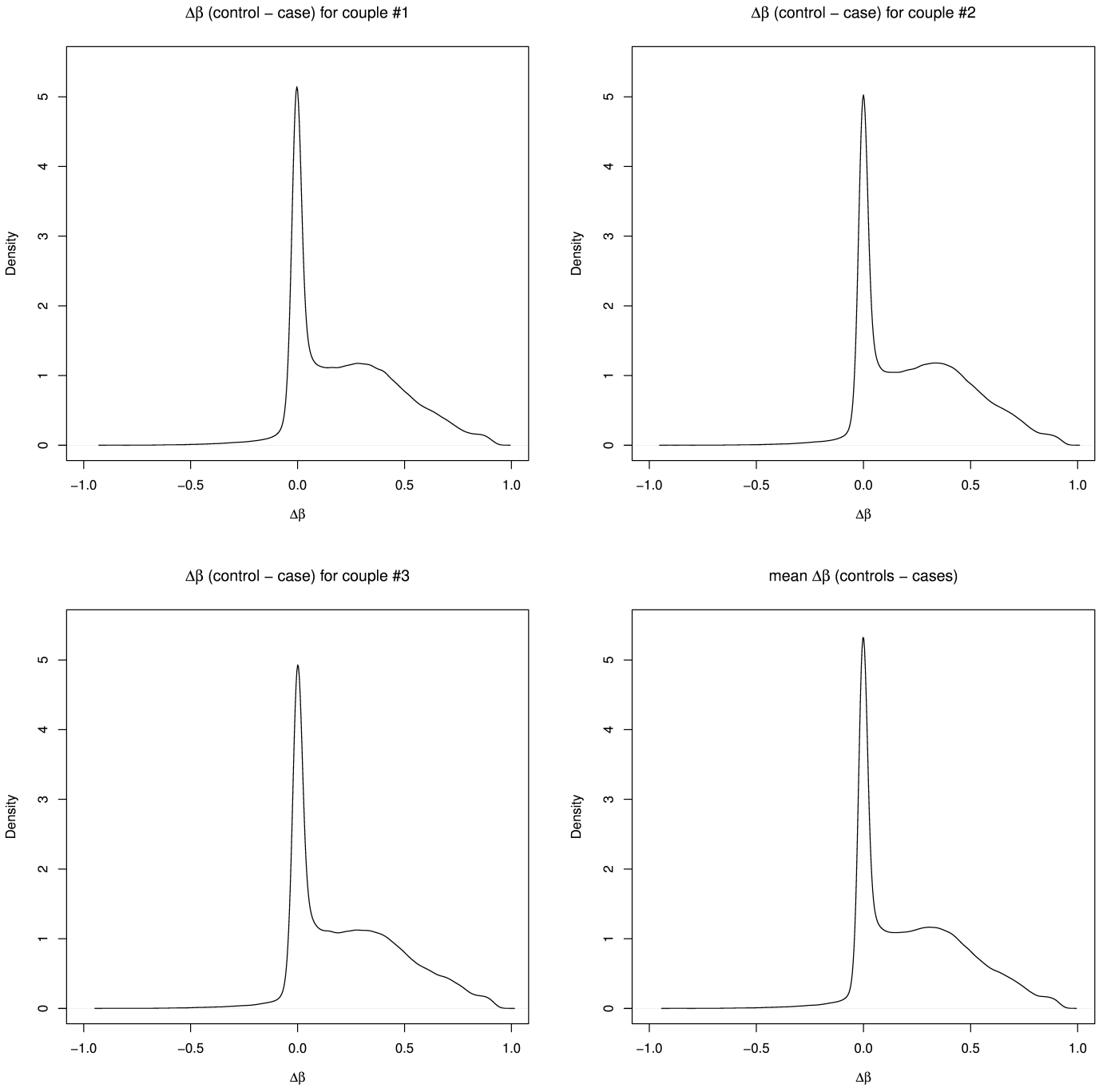
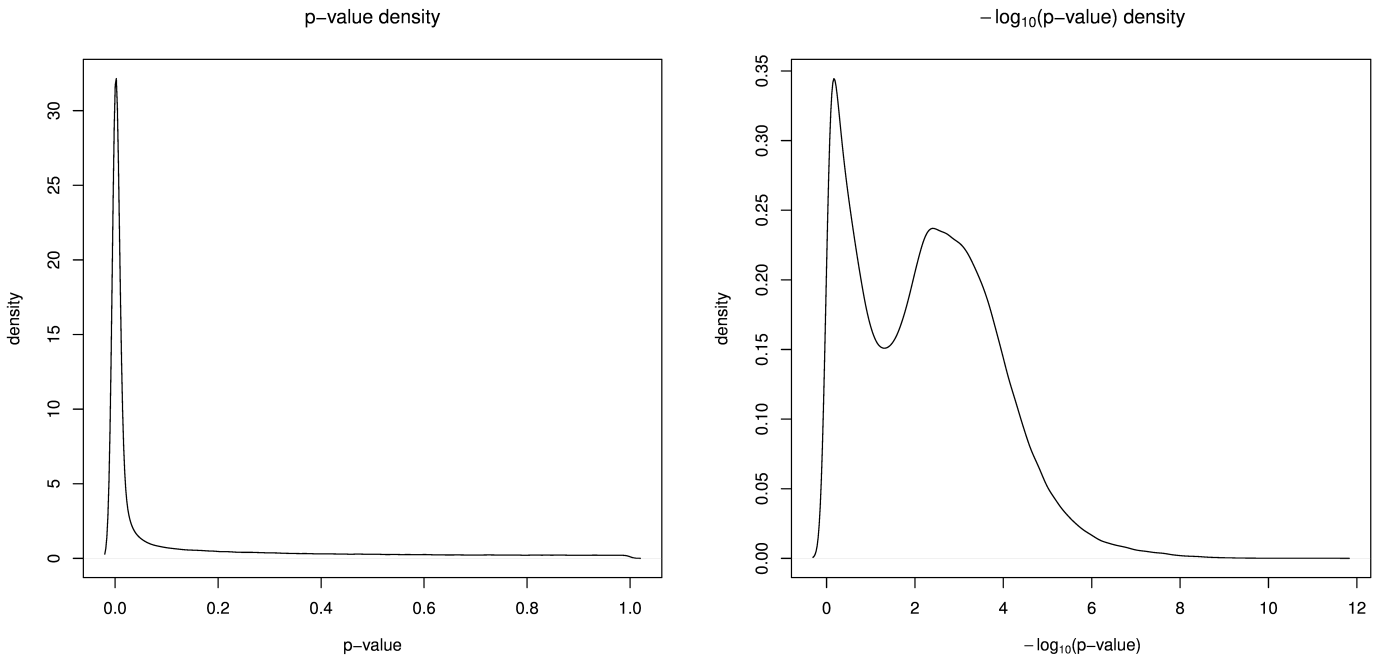


Figure 4: Distribution of $\Delta\beta$ for each couple control/case and in average.

2.5 p-values and volcano plot

p-value of the $\Delta\beta$ value for each probe have been computed by performing a Student t-test. Their distribution is shown in Figure 5. As can be seen on Figure 5b, many probes have high p-value (around $10^{\frac{-1}{2}}$) which is far from significant, but still a big part is below 10^{-3}.

Note that a t-test takes variance into consideration, meaning that a high $\Delta\beta$ might not be significant if variance is high inside the groups.



(a) Distribution of p-value.

(b) Distribution of base 10 log p-value.

Figure 5: Distribution of p-value and $\log_{10}(p\text{-value})$.

Figure 6 plots $\Delta\beta$ and p-values in a volcano plot. We can observe a big portion of the probes in the upper right area of the plot, meaning significant $\Delta\beta$ of hypomethylation, but we can also observe the upper left area of the plot showing significant $\Delta\beta$ of hypermethylation.

2.6 Heatmap

When taking the top 30 probes according to $|\Delta\beta|$, we can plot them in a heatmap (see Figure 7). Separate plots for methylated and unmethylated probes are shown in Figure 8.

For the 3 first samples (i.e. the controls), we can see that the unmethylated probes received way less signals than the methylated ones, and that this tendency is reversed for the 3 last samples (i.e. the cases). This is as expected since the probes that are present in this plot are the ones with the highest $\Delta\beta$ value, meaning that the difference between the methylation level in the controls and in the cases is the highest. Therefore, it is expected to have highly methylated controls and highly unmethylated cases signals.

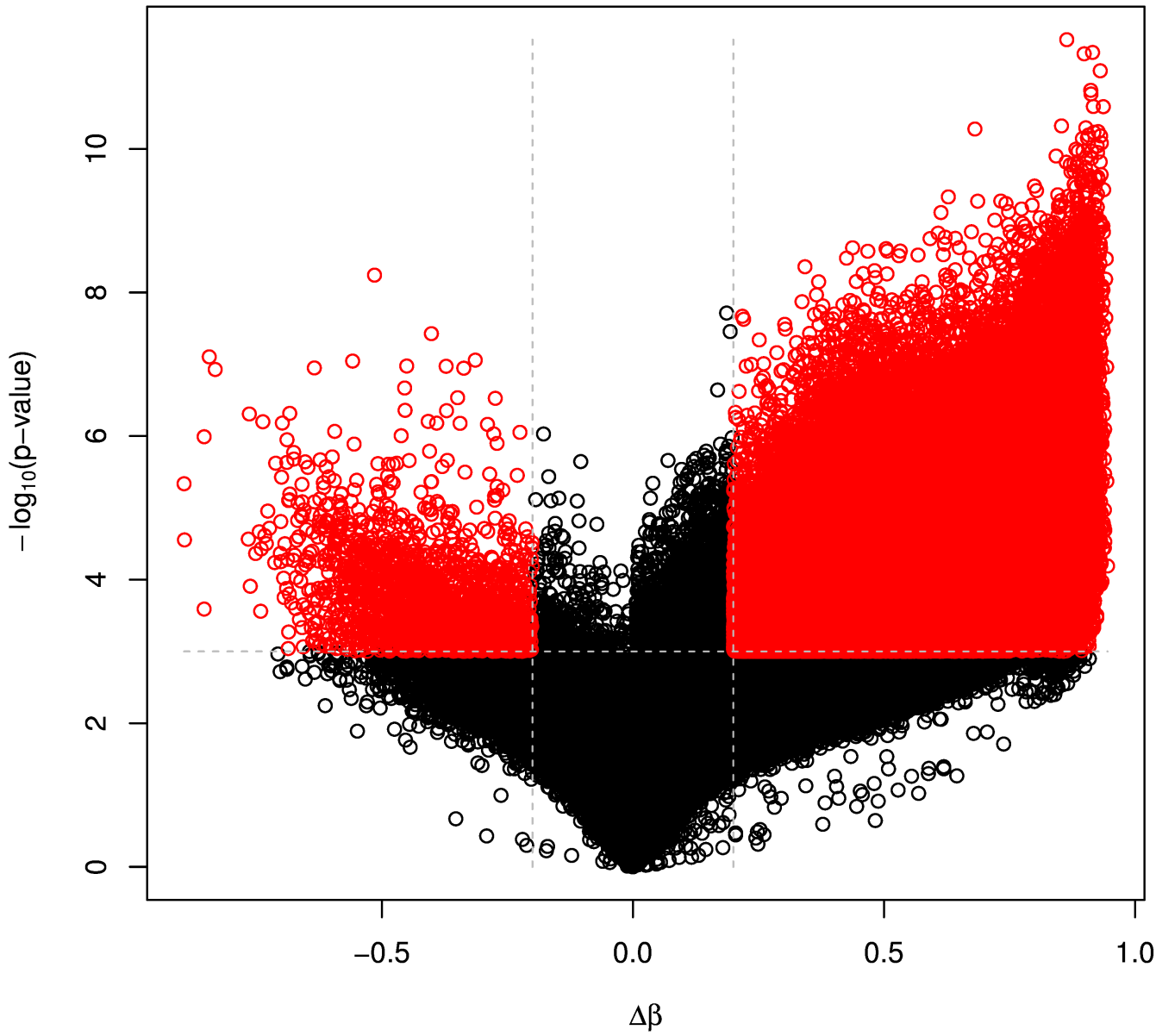


Figure 6: Volcano plot of p-value according to $\Delta\beta$. Significance threshold for p-values is set to $p = 10^{-3}$, and for $|\Delta\beta| = .2$. Note that even though the study here is about hypomethylated cases ($\Delta\beta > 0$), several probes lie before $\Delta\beta = -.2$, meaning that some regions are hypermethylated (even *significantly* hypermethylated).

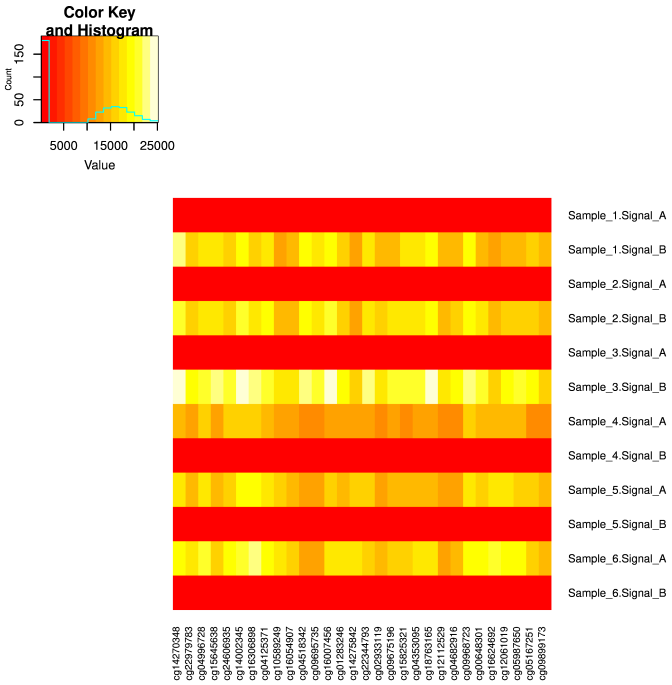
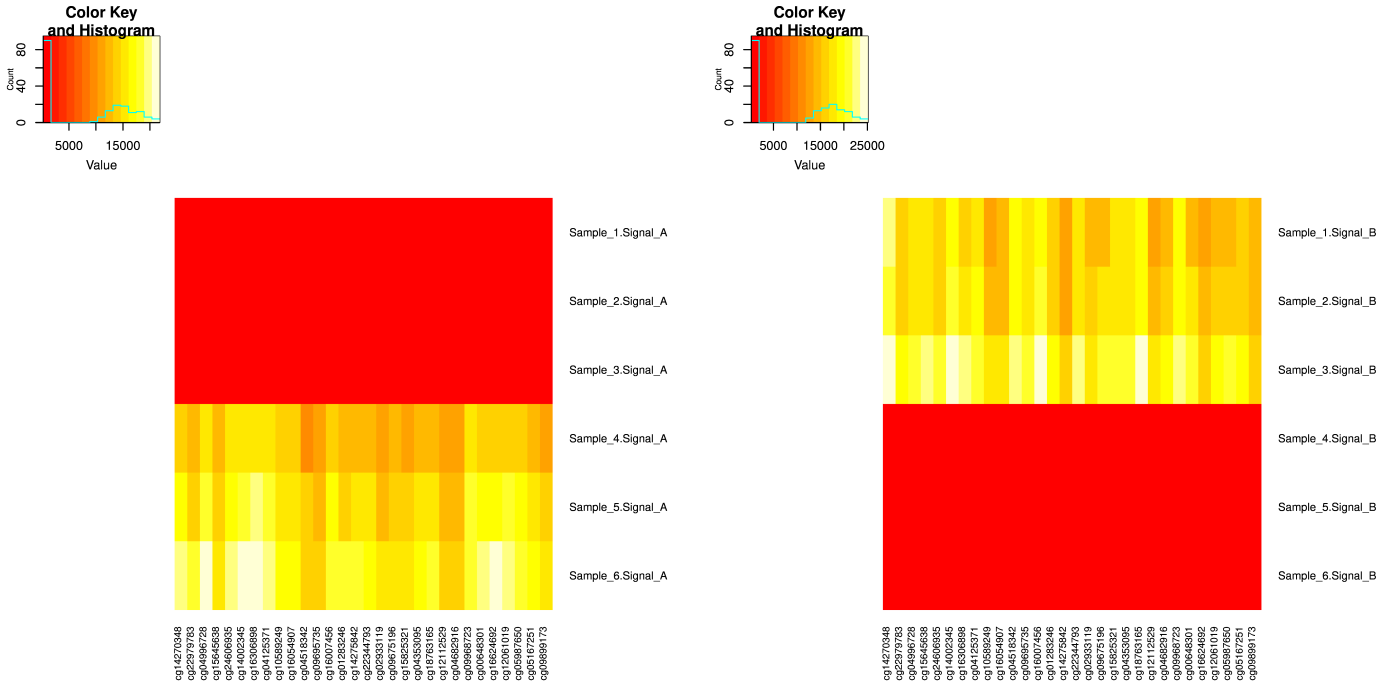


Figure 7: Heatmap of the probes value of the 30 probes having the highest average $\Delta\beta$.



(a) Heatmap of the unmethylated probes value of the 30 probes having the highest $\Delta\beta$.

(b) Heatmap of the methylated probes value of the 30 probes having the highest $\Delta\beta$.

Figure 8: Separate heatmaps for methylated and unmethylated probes of the top 30 probes.

3 R source code

First of all, libraries must be imported and constants must be defined. Only two libraries have been used files must be loaded into R:

```
library('latex2exp');  # Lib necessary to use LaTeX in plots
library('gplots');     # for heatmap.2

INFINIUM_PATH  <- 'Infinium450k_raw_data.txt';
ANNO_MINI_PATH <- 'HumanMethylation450_anno_mini.csv';
NB_CASES <- 3;
NB_CONTROLS <- NB_CASES;
CONTROLS <- 1:NB_CONTROLS;
CASES <- (1:NB_CASES) + NB_CONTROLS;
ALPHA_PSEUDO_COUNT <- 100;
ALL_CHROMOSOMES <- c(1:22, 'X', 'Y');
CHROMOSOMES_TO_TEST <- 19:21;

LOG_P_VALUE_THRESHOLD <- -3;
DELTA_BETA_THRESHOLD <- .2;

# Use first column as row names
infinium <- read.table(INFINIUM_PATH, header=T, dec=',', row.names=1);
annotations <- read.table(ANNO_MINI_PATH, header=T, sep=',', row.names=1);
```

In order to determine the β value of a sample, let's define a few functions:

```
compute.beta <- function(signalA, signalB) {
  # Note the pseudo count $\alpha = 100$ to avoid dividing by 0
  return (signalB / (signalA + signalB + ALPHA_PSEUDO_COUNT));
}

# Get the column corresponding to the required sample
get.sample.signal <- function(table, sample.id, signalA=T) {
  if(signalA) {
    return (as.vector(table[[2*sample.id-1]]));
  } else {
    return (as.vector(table[[2*sample.id]]));
  }
}

# Returns the $\beta$ values of a given sample
get.beta <- function(infinium, sample.id) {
  return (compute.beta(
    get.sample.signal(infinium, sample.id, T),
    get.sample.signal(infinium, sample.id, F)
  ));
}
```

so that plotting becomes:

```
# Plot the beta distribution of each sample from 1 to 6
plot.beta.distribution.infinium <- function(infinium) {
  for(sample.id in 1:(NB_CASES+NB_CONTROLS)) {
    plot(
      density(get.beta(infinium, sample.id)),
      xlim=c(-.2, 1.2),
      ylim=c(0, 5),
      xlab=TeX('$\\beta$'),
      ylab=TeX('Density of $\\beta$'),
      main=TeX(paste('density of $\\beta$ distribution of sample', sample.id)),
    );
  }
}
plot.beta.distribution.infinium(infinium);
```

The methylation level per chromosome is performed as follows:

```
# Returns the subtable corresponding to the required chromosome
extract.chromosome <- function(annotations, chr) {
  return (annotations[annotations$CHR == chr,]);
}

# Make the barplots of the average methylation level of each chromosome
plot.mean.methylation.level <- function(infinium, annotations) {
  means <- matrix(rep(0, 3*length(ALL_CHROMOSOMES)), nrow=3, ncol=length(ALL_CHROMOSOMES));
  idx <- 1;
  for(chromosome in ALL_CHROMOSOMES) {
    intersection <- intersect(rownames(infinium), rownames(extract.chromosome(annotations, chromosome)));
    chromosome.infinium.subtable <- infinium[intersection,];
    for(case.sample.id in CASES)
      means[1,idx] <- means[1,idx] + mean(get.beta(chromosome.infinium.subtable, case.sample.id))/NB_CASES;
    for(control.sample.id in CONTROLS)
      means[2,idx] <- means[2,idx] + mean(get.beta(chromosome.infinium.subtable, control.sample.id))/NB_CONTROLS;
    # Take difference control - case to see methylation difference per chromosome
    means[3,] <- means[2,] - means[1,];
    idx <- idx+1;
  }
  barplot(means, names.arg=ALL_CHROMOSOMES, horiz=T, beside=T, legend=c('Cases', 'Controls', 'Difference'), las=1);
  barplot(means[,CHROMOSOMES_TO_TEST], names.arg=CHROMOSOMES_TO_TEST, beside=T, legend=c('Cases', 'Controls', 'Difference'), las=1);
}


plot.mean.methylation.level(infinium, annotations);
```

To find the number of probes per gene, only annotations are required since all ids are common to the infinium file and the annotations file:

```
# Plot the histograms of the distribution and log-distribution of the number of probes per gene
plot.number.of.genes.per.probe <- function(annotations) {
  non.null.genes.associations <- as.vector(annotations[annotations$UCSC_RefGene_Name != '',]$UCSC_RefGene_Name);
  genes.counter <- table(unlist(sapply(non.null.genes.associations, function(s) {unique(strsplit(s, ';'))})));
  hist(genes.counter, main='Distribution of the number of probes per gene',
    sub=sprintf('Average number of probes per gene is %g', mean(genes.counter)),
    xlab='Number of probes per gene',
    ylab='Frequency');
  plot(log(table(genes.counter)), main='log plot of the distribution of the number of probes per gene',
    xlab='Number of probes per gene',
    ylab='log(Frequency)');
}
```

Note that `function(s) unique(strsplit(s, ';'))` allows to retrieve all the genes associated to a probe, but not taking the genes that come several times, thanks to the `unique` function. Making a table of the unlisted results (since `sapply` returns a list) makes R count the number of probes related to each gene on its own. Therefore, making a histogram of the table is sufficient to plot the distribution.

$\Delta\beta$ distribution is computed and plotted as follows:

```
# Get the $\Delta\beta$ vector of a (control, sample) couple
get.delta.beta <- function(infinium, case.sample.id, control.sample.id) {
  return (get.beta(infinium, control.sample.id) - get.beta(infinium, case.sample.id));
}

# Plots the density of $\Delta\beta$ for each (control, sample) couple
plot.delta.beta.distribution <- function(infinium) {
  # plot \Delta \beta distribution for each couple control/case
  for(control.sample.id in CONTROLS) {
    case.sample.id <- control.sample.id + NB_CONTROLS;
```

```
    infinium$de
    plot(
      density(get.delta.beta(infinium, case.sample.id, control.sample.id)),
      main=TeX(sprintf('$\\Delta\\beta$ (control - case) for couple #%d', control.sample.id)),
      xlab=TeX('$\\Delta\\beta$'),
      ylab='Density',
      xlim=c(-1, 1),
      ylim=c(0, 5.5)
    );
  }
  # Plot \Delta \beta for the mean of all samples
  delta.betas <- matrix(0, nrow=nrow(infinium), ncol=NB_CONTROLS+NB_CASES);
  for(control.sample.id in CONTROLS) {
    delta.betas[,control.sample.id] <- get.beta(infinium, control.sample.id);
    case.sample.id <- control.sample.id + NB_CONTROLS;
    delta.betas[,case.sample.id] <- get.beta(infinium, case.sample.id);
  }
  infinium$delta.beta <- apply(delta.betas[,CONTROLS], 1, mean) - apply(delta.betas[,CASES], 1, mean);
  plot(
    density(infinium$delta.beta),
    main=TeX('mean $\\Delta\\beta$ (controls - cases)'),
    xlab=TeX('$\\Delta\\beta$'),
    ylab='Density',
    xlim=c(-1, 1),
    ylim=c(0, 5.5)
  );
  infinium$p.values <- apply(delta.betas, 1,
    function(row) {
      return (t.test(row[CONTROLS], row[CASES])$p.value);
    }
  );
  infinium$log.p.value <- log(infinium$p.value, 10);
  plot(
    density(infinium$p.values),
    main=TeX('$p$-value density'),
    xlab=TeX('$p$-value'),
    ylab='density'
  );
  plot(
    density(-infinium$log.p.value),
    main=TeX('$-\\log_{10}(p$-value$)$ density'),
    xlab=TeX('$-\\log_{10}(p$-value$)$'),
    ylab='density'
  );
  return (infinium);
}

infinium <- plot.delta.beta.distribution(infinium);
```

Note that as `plot.delta.beta.distribution` returns the `infinium` table since it has been modified and R does not allow passing parameters by reference. What is added to the table is the p-value of $\Delta\beta$ for each probe (computed by a t-test), that is also plotted.

With these p-values, the volcano plot is performed using:

```
plot.volcano <- function(infinium) {
  significant.idx <- as.logical((abs(infinium$delta.beta) >= .2) * (infinium$log.p.value <= -3));
  significant <- infinium[significant.idx,];
  non.significant <- infinium[!significant.idx,];
  plot(
    c(non.significant$delta.beta, significant$delta.beta),
    -c(non.significant$log.p.value, significant$log.p.value),
    col=c(rep('black', nrow(non.significant)), rep('red', nrow(significant))),
    xlab=TeX('$\\Delta\\beta$'),
    ylab=TeX('$-\\log_{10}(p$-value$)$')
  )
  lines(c(min(infinium$delta.beta), max(infinium$delta.beta)), rep(-LOG_P_VALUE_THRESHOLD, 2), col='
   grey', lty=2)
  lines(rep(DELTA_BETA_THRESHOLD, 2), c(min(-infinium$log.p.value), max(-infinium$log.p.value)), col='
   grey', lty=2)
```

```
    lines(rep(-DELTA_BETA_THRESHOLD, 2), c(min(-infinium$log.p.value), max(-infinium$log.p.value)), col='grey', lty=2)
}

plot.volcano(infinium);
```

Only the heatmaps are still missing. These are plotted as follows:

```
get.top.delta.beta <- function(infinium, n=30) {
  return (infinium[order(abs(infinium$delta.beta), decreasing=T)[1:n],]);
}

plot.heatmap <- function(infinium) {
  top <- get.top.delta.beta(infinium);
  par(mar=c(0, 1, 3, 1))
  heatmap.2(t(as.matrix(top[,1:(2*(NB_CONTROLS+NB_CASES))])), Rowv=NA, Colv=NA, scale='none', cexRow=.8, cexCol=.75,
    margins=c(6, 8), dendogram='none', trace='none');
  for(i in 1:2) {
    heatmap.2(t(as.matrix(top[,seq(i, 2*(NB_CONTROLS+NB_CASES),2)])), Rowv=NA, Colv=NA, scale='none', cexRow=.8, cexCol=.75,
      margins=c(6, 8), dendogram='none', trace='none');
  }
}

plot.heatmap(infinium);
```

Note that `heatmap.2` is used, which explains the `library('gplots');` above.

References

[1] "Infinium® humanmethylation450 beadchip." [Online]. Available: https://cancergenome.nih.gov/abouttcga/aboutdata/platformdesign/illuminamethylation450

[2] D. Weisenberger, D. Van Den Berg, F. Pan, B. Berman, and P. Laird, "Comprehensive dna methylation analysis on the illumina infinium assay platform," *Illumina, San Diego*, 2008.

[3] S. Dedeurwaerder, M. Defrance, E. Calonne, H. Denis, C. Sotiriou, and F. Fuks, "Evaluation of the infinium methylation 450k technology," *Epigenomics*, vol. 3, no. 6, pp. 771–784, 2011.

[4] J. Amberger, C. A. Bocchini, A. F. Scott, and A. Hamosh, "Mckusick's online mendelian inheritance in man (omim®)," *Nucleic acids research*, vol. 37, no. suppl_1, pp. D793–D796, 2008.

[5] M. F. Paz, S. Wei, J. C. Cigudosa, S. Rodriguez-Perales, M. A. Peinado, T. H.-M. Huang, and M. Esteller, "Genetic unmasking of epigenetically silenced tumor suppressor genes in colon cancer cells deficient in dna methyltransferases," *Human Molecular Genetics*, vol. 12, no. 17, pp. 2209–2219, 2003.

[6] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry *et al.*, "Bioconductor: open software development for computational biology and bioinformatics," *Genome biology*, vol. 5, no. 10, p. R80, 2004.

[7] D. Wang, L. Yan, Q. Hu, L. E. Sucheston, M. J. Higgins, C. B. Ambrosone, C. S. Johnson, D. J. Smiraglia, and S. Liu, "Ima: an r package for high-throughput analysis of illumina's 450k infinium methylation data," *Bioinformatics*, vol. 28, no. 5, pp. 729–730, 2012.

[8] P. Du, X. Zhang, C.-C. Huang, N. Jafari, W. A. Kibbe, L. Hou, and S. M. Lin, "Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis," *BMC bioinformatics*, vol. 11, no. 1, p. 587, 2010.