BINFF-402 — Genomics, Proteomics and Evolution
Assignment 1

Robin Petit

December 10, 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.

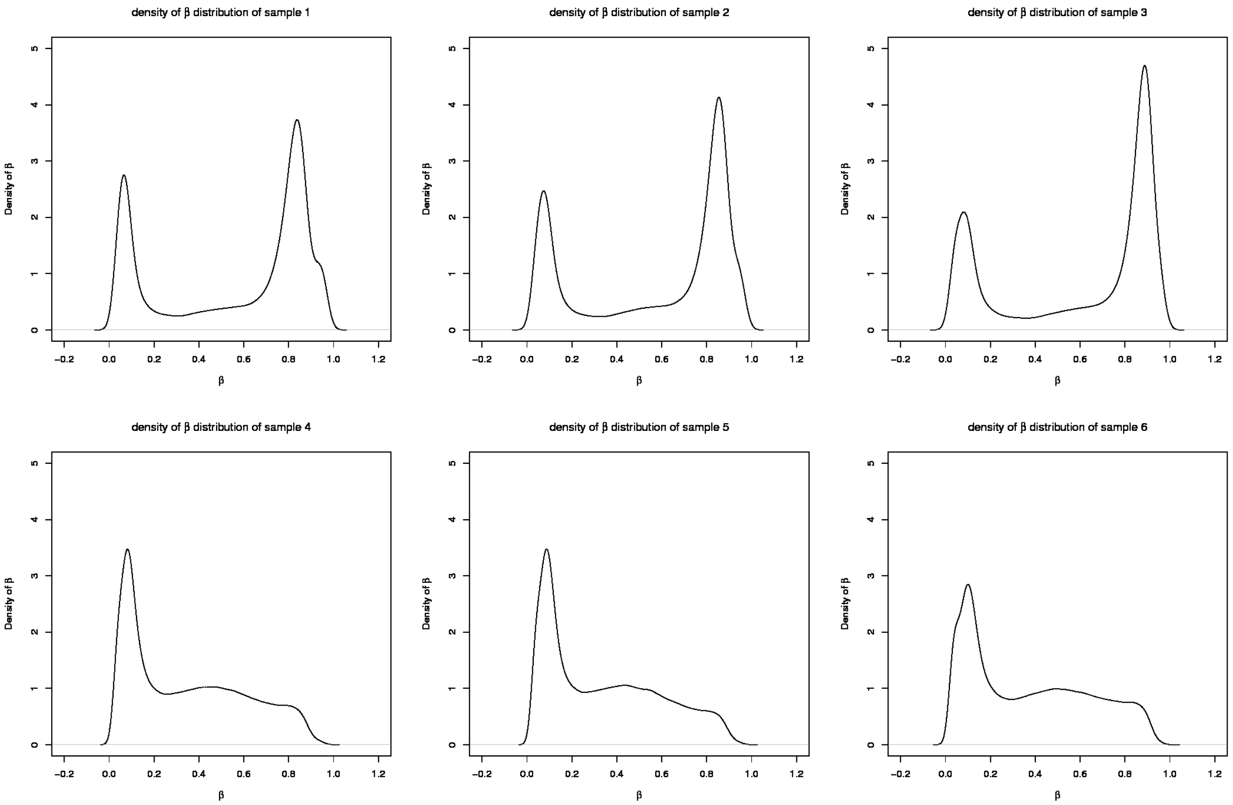

Figure 1: Distribution of the β value for each sample. Samples 1 to 3 are the controls, and samples 4 to 6 are the cases.

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.

2.2 Global overview

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.

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:

```r
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;

# 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 $\beta$ value of a sample, let's define a few functions:

```r
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:

```r
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);
```