Statistical Methods for the Analysisof Microarray and Short-Read Sequencing Data

R Project

Student: Damianos Melidis, Msc in Computational Biology and Bioinformatics
Student id: 12-945-416, dmelidis@student.ethz.ch

December 28, 2013

Abstract

The main goal of this project is to reproduce experimental results using *R* from a chosen publication [Lovisa E. Reinius et al.].

1. Introduction

The authors are motivated to study epigenetic marks for blood samples because of the limitations of extracting large number of (affected and non-affected) tissue samples. In this study they extract 10 different blood cell types from 6 male donors. The cell types are **whole blood**, peripheral blood mononuclear cells (**PBMC**) *CD4 T, CD8 T, CD56 NK, CD19 B, CD14 monocytes* and **granulocytes** *Neutrophils, Eosinophils*. The Illumina®Human Methylation 450k array used to measure the methylation status for each cell type in the majority of known genes (RefSeq®database). The researchers would like to investigate the differenece in methylation between each cell type. Besides trying to examine if only the whole blood methylation status is a sufficient epigenetic mark, the methylation of CpG sites comparing cell types with whole blood cell type. Interestingly authors after extracting the probes related to candidate genes for inflammatory diseases, they assess the similarity of differential methylation of the corresponding CpG sites between all cell types. In this project Figure 2 and 3a as well as table 1 and 2 from [Lovisa E. Reinius et al.] are reproduced using *R*.

3. Reproduced Results

In this section we will show the procedure that was followed to reproduce some results from [Lovisa E. Reinius et al.]. Following the bioinformatics analysis of *Materials and Methods* section the raw data files (.idat) were partitioned to distinct folder for *each* blood cell type, following the given sample_sheet_IDAT.xlsx, creating 10 different folders. Afterwards normalization and background correction was performed. Then the median for probe's methylation signal of the 6 donors is calculated and the principal component analysis and hierachical clustering was done resulting to Figure 1 and 2, respectively. Comparing these figures with figure 2 of the publication we can see some variation in the PCA (but the median signal and not all individual signals are used (insufficient workspace memory)) and the left clade of reproduced clustering tree is the same as the right clade of Figure 2a. Using *limma* we perform linear fitting for differential methylation between each pair of cells and tested by Bayesian moderated t-test on our normalized methylation signals. This gives us the table 1.
The authors would like to investigate if the whole blood is a sufficient cell type to study epigenetic marks. In this direction they compared the differentially methylated sites between each cell population *and whole blood*. In order to assign the methylation status to *unmethylated, marginal or methylated* they suggest to fit a gamma model which represents one *normal* distribution for

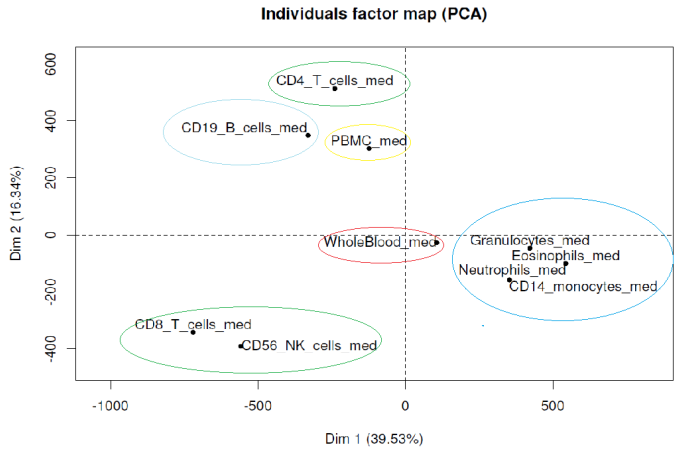

Figure 1: PCA for median signal of the 6 donors

marginal status (centered on 0) one gamma distribution (shifted normal to the positive values) representing *methylated* and one gamma shifted to the negative values *unmethylated*. Afterwards they extract the most significant differentially methylated sites depending on the Benjamini and hochberg corrected p-value. The same analysis is conducted leading to Figure 3 (Figure 3a in [Lovisa E. Reinius et al.]) and table 2. The *R* code used for the whole analysis is given in the next section.

Table 1: Number of differentially methylated CpG sites comparing each pair of cell types.

	Whole blood	PBMC	CD4	CD8	CD56	CD19	CD14	Granulocytes	Neutrophils	Eosinophils
Whole blood	0	126409	128806	115707	119812	124910	128369	130564	128346	128346
PBMC	0	0	130640	117666	121543	126428	128753	130764	128710	128710
CD4	0	0	0	120955	124171	129337	131042	132817	130931	130931
CD8	0	0	0	0	112702	116520	117051	119306	117212	117212
CD56	0	0	0	0	0	120199	121768	123798	121713	121713
CD19	0	0	0	0	0	0	127843	129459	127685	127685
CD14	0	0	0	0	0	0	0	135082	133381	133381
Granulocytes	0	0	0	0	0	0	0	0	135082	135082
Neutrophils	0	0	0	0	0	0	0	0	0	133800
Eosinophils	0	0	0	0	0	0	0	0	0	0

Table 2: Differential methylated probes for each cell type in comparison to whole blood

	M-value	Comparison on calls	Unmethylated	Marginal	Methylated
PBMC	126409	30000	0.76	0.011	0.23
CD4	128806	30000	0.74	0.02	0.26
CD8	115707	26706	0.73	0.02	0.25
CD56	119812	28745	0.74	0.02	0.24
CD19	124910	28630	0.72	0.02	0.26
CD14	128369	29774	0.72	0.04	0.24
Granulocytes	130564	30000	0.74	0.03	0.24
Neutrophils	128346	30000	0.74	0.03	0.24
Eosinophils	128346	30000	0.74	0.03	0.23

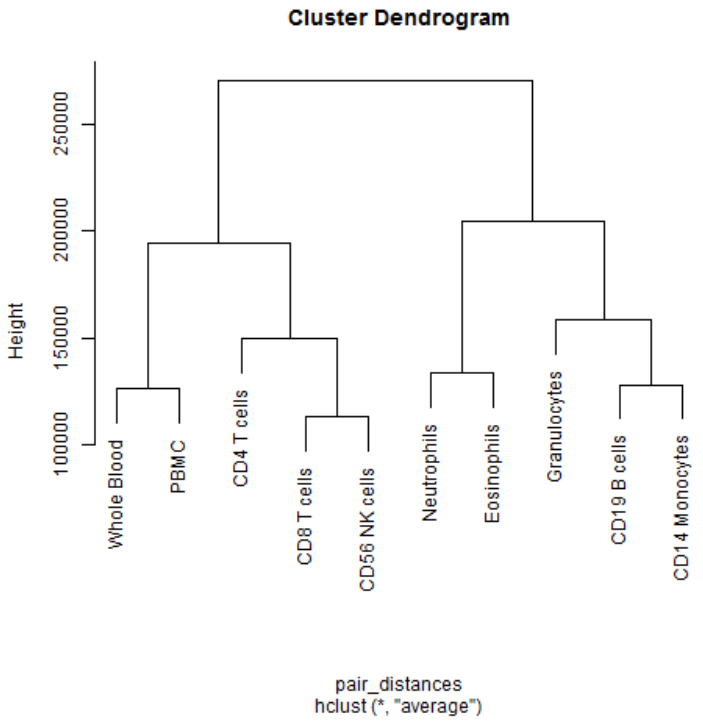

Figure 2: Hierachical clustering for median signal of the 6 donors

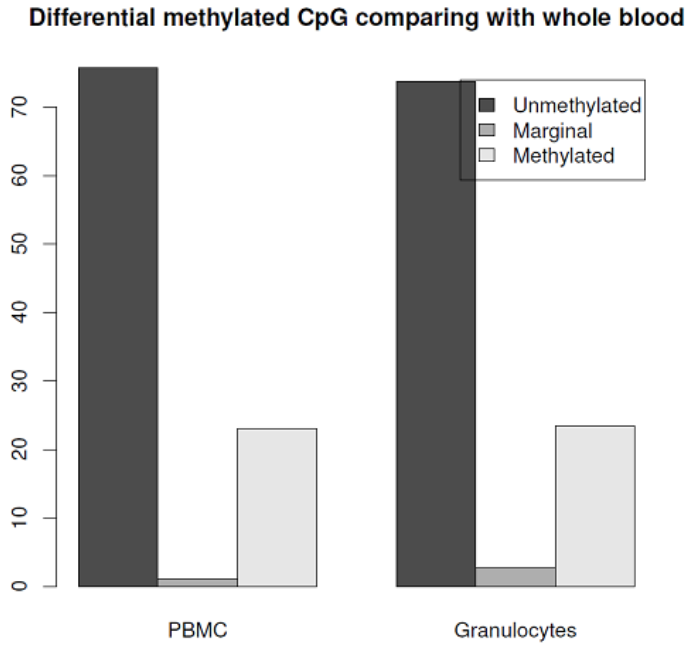

Figure 3: Distribution of methylation status for PBMC and Granulocytes comparing to whole blood.

4. Appendix

In the following the actual code is given:

```r
setwd("C:/Users/damian/Desktop/R_project/idat_files")
library("limma")
library("lumni")
library("minfi")
library("FactoMineR")
IlluminaHumanMethylation450k("IlluminaHumanMethylation450k.db")
#read idat files for each cell type
WholeBlood_Int <- read.450k.exp("WholeBlood")
PBMC_Int <- read.450k.exp("PBMC")
CD19_Int <- read.450k.exp("CD19")
CD4_Int <- read.450k.exp("CD4")
CD56_Int <- read.450k.exp("CD56")
CD8_Int <- read.450k.exp("CD8")
CD14_Int <- read.450k.exp("CD14")
Eosinophils_Int <- read.450k.exp("Eosinophils")
Granulocytes_Int <- read.450k.exp("Granulocytes")
Neutrophils_Int <- read.450k.exp("Neutrophils")

#normalization of data with background correction ##needs IlluminaHumanMethylation450kmanifest
WholeBlood_Meth.norm <- preprocessIllumina(WholeBlood_Int, bg.correct = TRUE, normalize = "controls",
    reference = 2)
PBMC_Meth.norm <- preprocessIllumina(PBMC_Int, bg.correct = TRUE, normalize = "controls", reference = 2)
CD19_Meth.norm <- preprocessIllumina(CD19_Int, bg.correct = TRUE, normalize = "controls", reference = 2)
CD4_Meth.norm <- preprocessIllumina(CD4_Int, bg.correct = TRUE, normalize = "controls", reference = 2)
CD56_Meth.norm <- preprocessIllumina(CD56_Int, bg.correct = TRUE, normalize = "controls", reference = 2)
CD8_Meth.norm <- preprocessIllumina(CD8_Int, bg.correct = TRUE, normalize = "controls", reference = 2)
CD14_Meth.norm <- preprocessIllumina(CD14_Int, bg.correct = TRUE, normalize = "controls", reference = 2)
Eosinophils_Meth.norm <- preprocessIllumina(Eosinophils_Int, bg.correct = TRUE, normalize = "controls",
    reference = 2)
Granulocytes_Meth.norm <- preprocessIllumina(Granulocytes_Int, bg.correct = TRUE, normalize = "controls",
    reference = 2)
Neutrophils_Meth.norm <- preprocessIllumina(Neutrophils_Int, bg.correct = TRUE, normalize = "controls",
    reference = 2)

#get the M values
WholeBlood_Mvalue <- getM(WholeBlood_Meth.norm)
PBMC_Mvalue <- getM(PBMC_Meth.norm)
CD19_Mvalue <- getM(CD19_Meth.norm)
CD4_Mvalue <- getM(CD4_Meth.norm)
CD56_Mvalue <- getM(CD56_Meth.norm)
CD8_Mvalue <- getM(CD8_Meth.norm)
CD14_Mvalue <- getM(CD14_Meth.norm)
Eosinophils_Mvalue <- getM(Eosinophils_Meth.norm)
Granulocytes_Mvalue <- getM(Granulocytes_Meth.norm)
Neutrophils_Mvalue <- getM(Neutrophils_Meth.norm)

#create a matrix with rows the probes and columns the all cell types for each donor
mat_Mvalue <- rbind(WholeBlood_Mvalue, PBMC_Mvalue, CD19_Mvalue, CD4_Mvalue, CD56_Mvalue, CD8_Mvalue, CD14
    _Mvalue, Eosinophils_Mvalue, Granulocytes_Mvalue, Neutrophils_Mvalue)
#give name to columns
WB_donors <- paste("WB_Donor_",1:6)
PBMC_donors <- paste("PBMC_Donor_",1:6)
CD19_donors <- paste("CD19_Donor_",1:6)
CD4_donors <- paste("CD4_Donor_",1:6)
CD56_donors <- paste("CD56_Donor_",1:6)
CD8_donors <- paste("CD8_Donor_",1:6)
CD14_donors <- paste("CD14_Donor_",1:6)
Eosinophils_donors <- paste("Eosinophils_Donor_",1:6)
Granulocytes_donors <- paste("Granulocytes_Donor_",1:6)
Neutrophils_donors <- paste("Neutrophils_Donor_", 1:6)
colnames(mat_Mvalue) <- c(WB_donors, PMBC_donors, CD19_donors, CD4_donors, CD56_donors, CD8_donors, CD14_
    donors, Eosinophils_donors, Granulocytes_donors, Neutrophils_donors)

#get the median values as representative for the six individuals
GenomicRegions <- rownames(WholeBlood_Mvalue)
WholeBlood_med <- apply(WholeBlood_Mvalue, 1, median, na.rm = T)
PBMC_med <- apply(PBMC_Mvalue, 1, median, na.rm = T)
CD19_B_cells_med <- apply(CD19_Mvalue, 1, median, na.rm = T)
CD4_T_cells_med <- apply(CD4_Mvalue, 1, median, na.rm = T)
CD56_NK_cells_med <- apply(CD56_Mvalue, 1, median, na.rm = T)
CD8_T_cells_med <- apply(CD8_Mvalue, 1, median, na.rm = T)
CD14_monocytes_med <- apply(CD14_Mvalue, 1, median, na.rm = T)
Eosinophils_med <- apply(Eosinophils_Mvalue, 1, median, na.rm = T)
Granulocytes_med <- apply(Granulocytes_Mvalue, 1, median, na.rm = T)
Neutrophils_med <- apply(Neutrophils_Mvalue, 1, median, na.rm = T)

#####################################Analysis for Figure 2 and Table 1#####################################
#create a matrix n * m, where (n) rows are the number of different cells and (m) columns are the genomic
    regions that measured for methylation
mat_Mvalue_med <- rbind(WholeBlood_med, PBMC_med, CD4_T_cells_med, CD8_T_cells_med, CD56_NK_cells_med,
    CD19_B_cells_med, CD14_monocytes_med, Granulocytes_med, Neutrophils_med, Eosinophils_med)
#replace the -Inf (totally unmethylated) and Inf (totally methylated) with 1.0 and 0.0 respectively
mat_Mvalue_med[ mat_Mvalue_med == Inf ] <-  max(mat_Mvalue_med[mat_Mvalue_med < Inf])
mat_Mvalue_med[ mat_Mvalue_med == -Inf ] <- min(mat_Mvalue_med[mat_Mvalue_med > -Inf ])
#~~~~~~~~Figure 2.B~~~~~~~~#
#do PCA to find the most important genomic regions for this dataset (e.g. eigenvectors of previous matrix)
PCA(mat_Mvalue_med)
#~~~~~~~~Figure 2.B~~~~~~~~#
#find differential methylation between each pair of cells
mat_Mvalue_med <- t(mat_Mvalue_med) #transpose matrix in order to get the cell population in the rows and
    the CpG sites in the columns
diff_exp_probes <- matrix(0,10,10) # create a matrix to save the differential expression betweeen each
    pair of cells (e.g diagonial matrix)
```

```
for (i in 1:9){#loop through cell and compute the differential methylated CpG sites with all other cells
  c <- i+1
  for(j in c:10){
    ExtractedMatrix <- cbind(mat_Mvalue_med[,i],mat_Mvalue_med[,j]) #get the CpG sites for the two
     currently compared cells types
    #show(ExtractedMatrix)
    fit <- lmFit(ExtractedMatrix)
    fit <- eBayes(fit)
    diff_exp_probes[i,j] = nrow(topTable(fit, coef=1, number=300000, lfc = 4))
  }
}
cell_names <- c("Whole Blood", "PBMC", "CD4 T cells", "CD8 T cells", "CD56 NK cells", "CD19 B cells", "
    CD14 Monocytes", "Granulocytes", "Neutrophils", "Eosinophils")
rownames(diff_exp_probes) <- cell_names
colnames(diff_exp_probes) <- cell_names
#~~~~~~~~Table 1~~~~~~~~#
write.table(diff_exp_probes,"clipboard",sep="\t", col.names = T, row.names = T)
#~~~~~~~~Table 1~~~~~~~~#
#~~~~~~~~Figure 2.A~~~~~~~~#
#perform hierachical clustering for cell populations (using euclidean distance)
pair_distances <- dist(diff_exp_probes, method = "euclidean")
png('CellPopulation_Clustering_med22.png')
plot(hclust(pair_distances, method="average"))
dev.off()
#~~~~~~~~Figure 2.A~~~~~~~~#
#####################################Analysis for Figure 2 and Table 1#####################################
#####################################Analysis for Figure 3 and Table 2#####################################
table2 <- matrix(0,9,5)
table2[,1] = t(diff_exp_probes[1,2:10]) #get the first column showing differentially methylated sites
    between cell types and whole blood
for (i in 2:10){
  ExtractedMatrix <- cbind(mat_Mvalue_med[,1],mat_Mvalue_med[,i]) #compare each cell type with whole blood
  fit <- lmFit(ExtractedMatrix)
  fit <- eBayes(fit)
  significant_meth_sites <- topTable(fit,coef=NULL,number=30000,adjust.method="BH",sort.by="B",resort.by=
    NULL,p.value=0.01,lfc=6)
  table2[i-1,2] = nrow(significant_meth_sites) #get the number of significant methylation sites
  methylation_selected_sites <- mat_Mvalue_med[rownames(significant_meth_sites),2] #get the significantly
    methylated sites
  fittedGamma <- gammaFitEM(methylation_selected_sites,initialFit=NULL,  maxIteration = 150, tol = 1e-04,
    plotMode = FALSE, verbose = FALSE)
  #plotGammaFit(methylation_selected_sites, gammaFit=fittedGamma)
  status <- methylationCall(fittedGamma)
  methyl_status <- table(status)
  margin_proportion <- methyl_status[1] / table2[i-1,2]
  methyl_proportion <- methyl_status[2] / table2[i-1,2]
  unmethyl_proportion <- methyl_status[3] / table2[i-1,2]
  table2[i-1,3] <- unmethyl_proportion
  table2[i-1,4] <- margin_proportion
  table2[i-1,5] <- methyl_proportion
}
table2
rownames(table2) <- cell_names(2:10)
colnames(table2) <- c("M-value comparison", "Comparison on calls", "Unmethylated", "Marginal", "Methylated
    ")
#~~~~~~~~Table 2~~~~~~~~#
write.table(table2,"clipboard",sep="\t", col.names = T, row.names = T)
#~~~~~~~~Table 2~~~~~~~~#
#~~~~~~~~Figure 3.A~~~~~~~~#
PBMC_Granulocytes_Meth_status <- 100 * rbind(table2[1,3:5],table2[7,3:5])
rownames(PBMC_Granulocytes_Meth_status) <- c("PBMC", "Granulocytes")
barplot(PBMC_Granulocytes_Meth_status,legend=T,beside=T,main='Differential methylated CpG comparing with
    whole blood')
#~~~~~~~~Figure 3.A~~~~~~~~#
#####################################Analysis for Figure 3 and Table 2#####################################
#####Incomplete part#####
x <- IlluminaHumanMethylation450kENTREZID
# Get the probe identifiers that are mapped to an ENTREZ Gene ID
mapped_probes <- mappedkeys(x)
# Convert to a list
xx <- as.list(x[mapped_probes])
```

References

[Lovisa E. Reinius et al.] Lovisa E. Reinius et al. Differential DNA methylation in Purified Human Blood Cells: Implications for Cell Lineage and Studies on Disease Susceptibility Plos One, vol. 7 no. 7 (2012).