ENIGMA Epigenetics Working Group: Cortical EWAS 2024

DNA Methylation QC for LBC1936

Dr Eleanor Conole

Version of: January 20, 2024

20th January, 2024
Lothian Birth Chohorts group, University of Edinburgh

Contents

Chapter 1

Introduction

1.1 DNA methylation microarray processing and normalization procedure

The ENIGMA QC protocol is outlined here, which includes standardised quality control procedures and quantile normalization. These are performed using the `minfi Bioconductor` package in R.

In brief what the following code does is map red and green channels intensities to the methylated and unmethylated status, while average intensities are used to check for low quality samples. Initial quality assessment of methylation data is performed using the `preprocessIllumina` option.

Principal component analyses (PCA) are performed using the singular value decomposition method, to identify methylation outliers based on the first four components. Samples with intensities more than 3 standard deviations away from the median ought to be considered outliers should be removed. Intensities from the sex chromosomes are used to predict sex, and samples with predicted sex different from their recorded value are also to be removed.

Samples that are initially processed in batches get merged at this stage before further pre-processing. Stratified quantile normalization then gets applied across samples. The data are then normalized together using the minfi `preprocessQuantile` function [1].

A PCA of normalized beta values is then used to control for unknown structure in the methylation data.

Estimated cell counts are controlled for the 6 major cell types in blood (granulocytes, B cells, CD4+T cells, CD8+ T cells, monocytes and NK cells) for each individual by implementing the `estimateCellCounts` function in `minfi`, which gives sample specific estimates of cell proportions based on reference information on cell specific methylation signatures.

Further information about the ENIGMA QC protocol can be found in previous publications by the ENIGMA Epigenetics Working Group [2]; the code that follows has been adapted specifically for running the ENIGMA QC protocol on LBC1936 data.

1.2 Contact

For further queries relating to this protocol please contact the main team involved in the ENIGMA Epigenetics Working Group. The QC for LBC1936 was originally performed by Tianye Jia for the subcortical EWAS paper, but was performed externally and not stored on any Edinburgh servers. Eleanor Conole re-ran the QC in 2023, after email correspondence with Sylvane Desrivières and Xinyang Yu, who should be thanked for their patience with the Cortical EWAS analysis.

Chapter 2

DNA methylation quality control

2.1 Some notes on array design and `IDAT` files

Array Design

The 450k array has a very unusual design, which to some extent impacts analysis. It is really a mixture of a two-color array and two one-color arrays. There are two main types of probes (type I and type II), and the probe design affects the signal distribution of the probe.

Raw Data Format: IDAT

The raw data format for the 450k array is known as `IDAT`. Because the array is measured in two different colors, there are two files for each sample, typically with the extension `_Grn.idat` and `_Red.idat`.

Illumina's software suite for the analysis of this array is called `GenomeStudio`. It is not unusual for practitioners to only have access to processed data from GenomeStudio instead of the raw IDAT files. Still, note that there is information in the IDAT files beneficial to analysis (see minfi documentation) [3].

Further note on IDATs

In the context of DNA methylation analysis, IDAT files are associated with the Infinium DNA methylation microarray technology developed by Illumina. The term "IDAT" stands for "Intensity Data." These files contain the raw intensity data generated during the scanning of Infinium microarrays.

The Infinium DNA methylation microarray technology is widely used for profiling DNA methylation patterns at a single-nucleotide resolution. The arrays use two chemistries, Infinium I and Infinium II, to interrogate DNA methylation at specific CpG sites across the genome. Each array includes probes that target methylated and unmethylated states of CpG dinucleotides.

The IDAT files store the fluorescence intensity values for each probe on the microarray. There are two types of IDAT files associated with Infinium microarrays: one for the methylated channel (M) and one for the unmethylated channel (U). So, for each sample analyzed on an Infinium microarray, you typically have two corresponding IDAT files: one ending with "_M" and one ending with "_U."

Researchers use these IDAT files as input for bioinformatics tools and software packages designed for the analysis of DNA methylation data. The intensity values are processed to derive methylation values, often represented as beta values, which indicate the proportion of DNA methylation at each CpG site. This information can then be used for downstream analyses and interpretation of DNA methylation patterns in the studied samples.

The data used to generate these plots is taken from the `IDATs` folder in `"/CCACE_Shared/EleanorC/CorticalEWAS/Data"` and all scripts are documented online at [4].

2.2 Code required for package installation

Load Required Packages

```
##############################
### DNAm idats from LBC36 ###
#################''###########

install.packages("tidyverse")
library(tidyverse)

if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("minfi")

# To install the FlowSorted.Blood.450k package, which is necessary when
# using the function of "estimateCellCounts", enter:

if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("FlowSorted.Blood.450k")

# Highlighted portions of the instructions require you to make changes
# so that the commands work on your system and data.

## open R and copy the lines below:
require(minfi)
require(minfiData)

library(RColorBrewer)
library(minfi)
library(limma)
library(FlowSorted.Blood.450k)

## Set your local working directory
setwd("/CCACE_Shared/EleanorC/CorticalEWAS/Data")  # replace with your local
↪  working directory
```

2.3 Reading in DNA Methylation Data

Here you will specify your local directory, which contains both raw IDAT files and the sample sheets (i.e., a .csv file, which includes the path for each sample's IDAT file).

The following scripts expect the sample sheet to include column names as shown in the figure below, where the header information (all lines up to [Data]) could be omitted (Variables such as sex, age, site, and disease status can also be included as extra columns):

```
##################################
### Reading the methylation data ####
##################################

## Creating the initial object of the minfi analysis that contains
# the raw intensities in the green and red channels
# Set your data directory containing IDAT files
idat_dirs <- c("/CCACE_Shared/EleanorC/CorticalEWAS/Data/idats")

# Specify your local directory, which contains both raw IDAT files
# and the sample sheets (i.e., a .csv file, which includes the path
# for each sample specific IDAT file).

#####################################
### DNAm idats from LBC36 CREATED ###
#####################################

## Try creating RGset
#library(minfi)

targets=read.metharray.sheet(idat_dirs,
"Deary Meth450K_plates 9_21_7-SampleSheet1220413.csv",
recursive=T) #loads the corresponding .csv sample sheet file

targets=rbind(targets, read.metharray.sheet(idat_dirs,
"E11970_Meth450K_plates6-18-10_230413.csv",
recursive=T))
# ... (additional targets)
```

Run Some Checks

```
## 1. check for any duplicates in basenames
#targets <- targets[!duplicated(targets$Basename)]
targets <- targets %>% filter(Basename != "character(0)")
targets <- targets %>% filter(grepl("_W2", Sample_Name))

#### targets
## NOTE: later need sex variable included, so potentially use
# the targets_sex which contains cell types + sex info
# target_sex <- read.csv("targets_W2_sex.csv")

targets <- read.csv("new_targets_W2.csv")

RGset <- read.metharray.exp(
  base = idat_dirs,
  targets = targets,
  verbose = T
)
```

Save the RGset

Note that this will take around 20 minutes to run.

The 'RGset' is a large object file; it is a class called 'RGChannelSet' which represents two color data with a green and a red channel.

```
## NOTE this takes 20 minutes to run
save(RGset, file="/CCACE_Shared/EleanorC/CorticalEWAS/Data/Output/RGset.rda")
↪  #saves the object


\subsection{Create QC Plots from the RGset}

\begin{lstlisting}[language=R, caption={Create QC Plots from the RGset},
↪  label={lst:qc-plots}]
### If the initial object with raw methylation data had already been created
# as above, it can be directly loaded with the command below
# (if the file hasn't been generated yet or if you do not know what it means,
# please follow the procedure above):

# If you accidentally kill the session
load("/CCACE_Shared/EleanorC/CorticalEWAS/Data/Output/RGset.rda")

### Producing Quality Control plots
pd <- pData(RGset) #extracting the sample information (phenotype data)
# from the sample sheet

if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("IlluminaHumanMethylation450kmanifest")

## NOTE: This took 10 minutes to run
```

Create QC Report PDF

These plots are useful for identifying samples with data quality. They display summaries of signals from the array (e.g. density plots) as well as the values of several types of control probes included on the array. A good rule of thumb is to be wary of samples whose behavior deviates from that of others in the same or similar experiments.

In case outliers are detected, their individuals' Sample.ID should be registered into the file 'Outliers', which will be used to remove outliers in a later step.

You want to generate figures such as this to visualise the array signals:

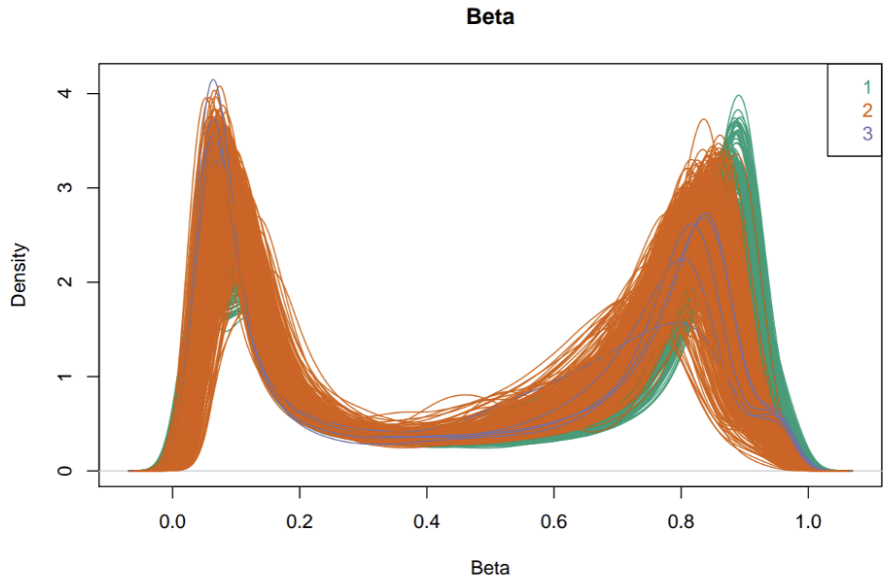

Figure 2.1: LBC1936 density plot of genome-wide methylation values

The code below produces a PDF QC report of common plots, colored by groups of samples. Check that the highlighted variables correspond to column names in your sample sheet, as illustrated above. In this example, the column named 'Experimenter' corresponds to the different waves by which our DNA samples were processed and hybridized.

```
###################################
### Creating a PDF report of QC ####
###################################

qcReport(RGset,
sampNames = pd$Sentrix,
sampGroups=pd$set,
pdf = "/CCACE_Shared/EleanorC/CorticalEWAS/Data/Output/qcReport.pdf")
# location of pdf report
```

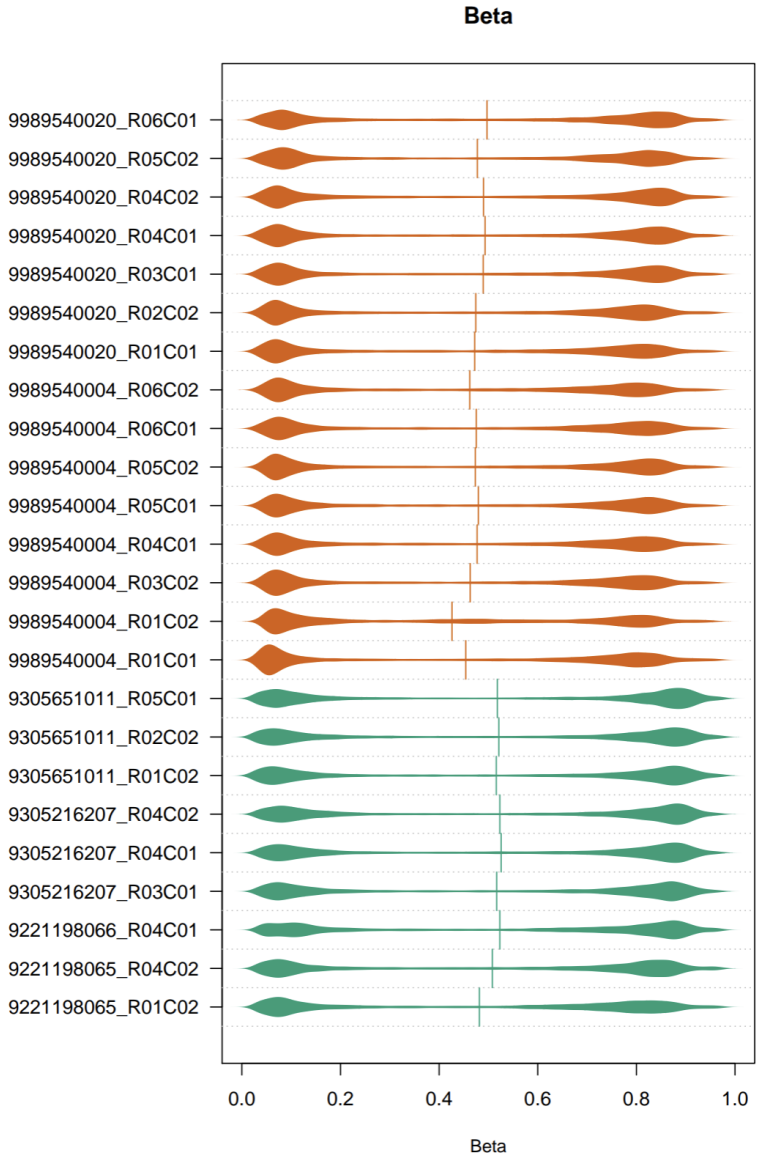

Figure 2.2: Truncated violin plot of genome-wide methylation values per LBC1936 sample (12 participants)

2.4 QC Assessment

```
###################################
### Quality assessment of methylation data ####
###################################

object=preprocessIllumina(RGset) #preprocessIllumina considers the background
↪  correction as well as normalization to internal controls; this minimizes the
↪  amount of variation between arrays.
object<-mapToGenome(object) #assign probes to physical location on the genome
object=ratioConvert(object, type="Illumina") #convert raw methylation data
beta <- getBeta(object) #get the beta value for each probe
dat <- object #rename 'object' to 'dat'
pd=pData(dat) #get phenotypes of methylation data
```

Bibliography

[1] Touleimat N, Tost J., *Complete pipeline for Infinium((R)) Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. Epigenomics 2012; 4(3): 325-341*. [Online]. Available: https://www.futuremedicine.com/doi/10.2217/epi.12.21

[2] Jia, T., Chu, C., Liu, Y. et al., *pigenome-wide meta-analysis of blood DNA methylation and its association with subcortical volumes: findings from the ENIGMA Epigenetics Working Group. Mol Psychiatry 26, 3884–3895 (2021)*. [Online]. Available: https://doi.org/10.1038/s41380-019-0605-z

[3] Kasper Daniel Hansen, *minfi Documentation*. [Online]. Available: https://kasperdanielhansen.github.io/genbioconductor/html/minfi.html

[4] Eleanor L.S. Conole, *ENIGMA DNAm QC*. [Online]. Available: https://github.com/EleanorSC/ENIGMA_Cortical_EWAS/tree/main