DNA methylation pre-processing in Airwave data

Maryam Karimi, Marc Chadeau-Hyam

December 18, 2018

1 Preprocessing

The Infinium HumanMethylation BeadChip for the measurement of DNA methylation levels is based on the bisulphite conversion of DNA at genomic locations at near-single-ncleotide resolution. It also includes 59 highly-polymorphics SNPs that were included to confirm the identity of samples from the same individual, that will be used as a level of quality control. Two different assay types coexist, Infinium I and Infinium II.

The first one has been used in genomic regions of high CpG density. It requires two types per CpG locus, a *methylated* bead and an *unmethylated* bead. Beads of the two types fluoresence in the same colour and DNA methylation levels are calculated as the ratio of these fluorescent signals. The second assay type beads fluoresce simultaneously in two colours and the ratio of these signals can be used to estimate the DNA methylation levels.

The Infinium HumanMethylation technology has also included other type of beads that can be used for different quality control purposes. These control beads are targeting sequences that do not contain CpG dinucleotides and thus do not depend on the methylation status. They are classified into *sample-independent* (4 categories: Hybridisation, Extension, Target removal and Staining) and *sample-dependent* (5 categories: Bisulphite conversion, Negative, Non-polymorphic, Specificity and Normalisation) control beads.

The Infinium laboratory protocol yields two idat files for each sample containing summary statistics of fluorescence intensities in the two green and red colours. Following are the pre-processing steps leading to the final set of DNA methylation measurements:

1. Mean intensities from non-control beads are rearranged into two matrices with intensity data for each of the two alleles at a specific genomic location. One matrix corresponds to absence of DNA methylation (A) and the other matrix corresponds to presence of DNA methylation (B).

2. Censoring values below detection limits:

- Detection thresholds are estimated from the 600 'negative' control beads as follows:

$$\begin{aligned} 2\bar{x}_G + z_\alpha \sqrt{2} s_G \quad & \text{Infinium I measuered in green} \\ 2\bar{x}_R + z_\alpha \sqrt{2} s_R \quad & \text{Infinium I measuered in red} \\ \bar{x}_G + \bar{x}_R + z_\alpha \sqrt{s_G^2 + s_R^2} \quad & \text{Infinium II measuered in green} \end{aligned} \tag{1}$$

  where $\bar{x}_G$ and $\bar{x}_R$ are the means, s_G and s_R are the standard deviations of the background noise in each colour separately obtained from the 'negative' control beads and $z_\alpha = \Phi^{-1}(1-\alpha)$.

- Elements of A and B are censored if the total intensity I=A+B is below these thresholds.

3. To control for the background noise and since the 'negative' control beads should not hybridise, it is recommended to perform the background substraction in addition to censoring. This is done by substracting mean intensities of 'negative' beads from A and B, depending on the colour each is measured in, and censoring negative values.

4. Dye bias which refers to the differences in fluorescence between the two colours is another source of bias in the measurements. This bias will not affect Infinium I assay since the two beads types fluoresce in the same colour. But in Infinium II assay, since we are using the same bead type that fluoresces in two colours the DNA methylation measurements is affected by this bias. For this purpose, normalisation control beads, which consists of 85 pairs of intensities in two green and red colors were used. The Dye bias correction has been performed using an in-house Rscript mimicking the same procedure as GenomeStudio software, with the difference of computing two multiplicative correction factors independently for each sample rather than with respect to an arbitrary reference. The dye bias correction constant are computed as follows and then applied to the A and B matrices restricted to the Infinium II assays.

$$\begin{aligned} k/\bar{r} \quad & \text{for red intensities} \\ k/\bar{g} \quad & \text{for green intensities} \end{aligned} \tag{2}$$

   with $\bar{g} = \sum_i g_i/85$ and $\bar{r} = \sum_i r_i/85$, the mean intensities from normalisation control beads in green and red respectively and $k = (\bar{g} + \bar{r})/2$.

5. Finally, DNA methylation levels are calculated as B (methylated) over total (A+B).

In the same program, we extracted intensities for SNP beads using theta/r format (as in GenomeStudio). Also, all control probes data has been extracted, and summary statistics were added to samples table. At the end, **raw intensities(A and B as above) for all probes** (including controls) (1), **raw intensities for control probes** (2), **DNAm values** (ratios) (3), **SNPs** (theta/r format) (4), and **samples table** (5) were stored in .rds format.

2 Preparation of DNA methylation data for statistical analysis

After the preprocessing steps, some additional manipulations should be performed on data before moving to statistical analysis. In this part, we read all .rds files created in the preprocessing process. The DNA methylation data has 1142 rows (samples) and 834 012 colomns (CpGs).

1. First from samples data (4) we remove individuals with more that 10% missing data (1 individual).

2. The mean intensities of BeadChip controls that were calculated for each individual for Infinium I red, green and Infinium II assays (variables bc1.red, bc1.grn and bc2 in samples data (4)) are extracted and a multivariate outlier identification was performed using the *pcout()* function of the *mvoutlier* package. Individuals with potential outliers were then removed (182 individuals).

3. Intensities of normalisation controls (170 beads) were extracted from data (2), principal component analysis were performed and PCs were added to the samples dataset using the *pca()* function of the *pcaMethods* package with 50 PCs.

4. Two variables were calculated for each individual in the preprocessing section using the DNA methylation intensities: the median intensity of the CpGs located on chromsome X and the mean of the number of times that DNA methlation values are missing for CpGs located on chromosome Y. Using these two values the data were clsutered into two classes with *cmeans()* function of the *e1071* package. If the membership value of an individual in a class is greater than 0.95 then that individual belongs to that class which represents gender here, and if the membership is not greater than 0.95 for both classes, NA is considered for the gender. The inferred sex from DNA methylation data is then added to the samples data.

5. Next step comprises of calling SNPs (59 SNPs), merging them with samples data and removing samples with more than 10% missing data (15 individuals).

6. We merged the DNAm with cohort data and we obtain 932 common individuals (12 individuals were not found).

7. We then compare the cohort's sex variable and the inferred sex and remove individuals with discordant DNAm sex (2 individuals).

8. We compute the similarity matrix and identify duplicates from SNPs. We also identify duplicates from annotation. We check the similarity of these duplicates. We remove individuals

with the same annotated ID, but different genotype (0 individual).

9. We extract DNAm data of the remaining individuals in the samples data (930 individuals) and we stored these data into **AIRWAVE_dnam.rds** and **AIRWAVE_samples.rds**.
10. Suggestion : It is better to use M-values of DNA methylation values instead of β values. This can be done by applying an M-transformation on the DNAm delta : $log2(dnam/(1-dnam))$.