**ARIES data - Josine Min**

**Clean data:**

**Data processing:**

All data processing has been done using meffil (<https://github.com/perishky/meffil>). See below the workflow.

meffilworkflow.pdf

**Sample QC**

All 615 outliers from Table 1 are removed. In total there are 4854 samples in the clean dataset as indicated in the “Pass” column. 112/5469 samples failed genotype QC due to sample swaps, gender mismatches, high IBD or relatedness issues between mums and kids. In addition samples are not removed due to population stratification. See script below how you can remove these samples.

**Table1: Number of outliers**

|  |  |
| --- | --- |
| **Issue** | **N** |
| Control probe (dye.bias) | 14 |
| Detection p-value | 166 |
| failed GWAS QC | 112 |
| genotype concordance (>90%) between mum-kid | 22 |
| genotype discordance (80%) across duplicates | 200 |
| genotype discordance (80%) across mums duplicates | 10 |
| genotype discordance (80%) across timepoints | 10 |
| genotype discordance (80%) across timepoints with mums | 24 |
| Genotype mismatch | 411 |
| Low bead numbers | 2 |
| Methylated vs Unmethylated | 68 |
| Sex mismatch | 161 |
| X-Y ratio outlier | 30 |
| Post-Normalization PCA | 2 |
| Post-Normalization t-test | 11 |

**Normalization**

We used **functional normalization** implemented in the R package meffil to normalize the data[1](#_ENREF_1). Functional normalization is a between-array normalization method for the Illumina Infinium HumanMethylation450 platform and an extension to quantile normalisation. It removes unwanted *technical* variation by regressing out variability explained by the control probes present on the array. It is expected to be most useful in datasets with large differences between samples. For ALSPAC, it might be useful to remove the plate effects, chip row and slide (technical variation) but keep biological differences such as sample type and timepoint differences. The normalization procedure is applied to the Meth and Unmeth intensities separately, and to type I and type II signals separately. For the probes on the X and Y chromosomes, males and females are normalized separately using the gender information. For the Y chromosome, standard quantile normalization is used due to the small number of probes, which results in instability for functional normalization. As we observed dyebias and background signal based on the negative probes I used a background correction and a dyebias correction.

As the slide effects were huge even after normalization, we regressed out slide on the raw betas before normalization and on the controlmatrix. The idea is that technical artefacts are captured better by the PCs on the controlmatrix if slide has been regressed out first. We have tested this using the ~6073 significant CpGs from the maternal smoking EWAS from PACE. The sensitivity (we found more significant signals) was indeed increased after regressing out slide first.

**Technical replicates**

In the clean normalized data there are 71 replicates (between 9-25 for each timepoint). In the manifest file there is a column **remove.duplicates** which are 71 replicates that should be removed from the data if you don’t want to use replicates. We kept the sample with the highest number of detected probes.

**Table2 Number of replicates (N=71) by time point**

|  |  |  |
| --- | --- | --- |
| **Timepoint** | **Timecode** | **Number of replicates** |
| Cord | Cord | 9 |
| F7 | F7 | 10 |
| TF3 | 15up | 1 |
| F17 | 15up | 10 |
| antenatal | antenatal | 25 |
| FOM | FOM | 15 |
| TF1-3 | FOM | 1 |

**Cell counts**

Meffil has a function called meffil.cell.count.estimates. It normalizes your sample to a reference blood count dataset (blood gse35069) and then calculates cell counts using the Houseman et al. method[2](#_ENREF_2).  Meffil normalizes each sample individually to the cell type reference dataset rather than normalizing all samples in a dataset to the cell type reference in order to avoid having cell count estimates depend on the other samples being included in the normalization. In addition to Houseman, meffil has two cord blood references included. The function meffil.list.cell.type.references() gives you a list of implemented cell count references.

**Sample type**

Sample\_type2 variable in the manifest is recoded from the samcode. There are differences in cell type composition for each sampletype and sampletype has been confounded with timecode (Table3) White cells are buffy coats (comprising lymphocytes, monocytes and granulocytes) and can be collected with heparin or EDTA tubes (coded in “additive”). PBLs are peripheral (mononuclear) blood lymphocytes and they comprise T cells, NK cells and B cells and collected with CPDA tubes. Blood spots are cord blood samples and are taken at birth. Blood spots are not taken from heel prick. Sampletype differences are explaining a lot of the variance in cord blood but not so much in the other timepoints after adjusting for cell counts. Please note sampletype in cord blood is confounded with slide in cord blood.

**Table 3**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **TimePoint** | **TimeCode** | **Gender** | **Mean Age** | **BloodSpots** | **PBL** | **Whitecells** | **Wholeblood** |
| cord | cord | M+F | 0 | 168 | 0 | 746 | 0 |
| F7 | F7 | M+F | 7.5 | 0 | 0 | 64 | 916 |
| TF3 | 15up | M+F | 15.4 | 0 | 0 | 254 | 0 |
| F17 | 15up | M+F | 17.7 | 0 | 0 | 727 | 0 |
| antenatal | antenatal | F | 28.7 | 0 | 0 | 436 | 551 |
| TF1-3 | FOM | F | 43 | 0 | 177 | 0 | 0 |
| FOM | FOM | F | 47.8 | 0 | 0 | 810 | 0 |

**Age**

Age are taken from the following variables: f7003c (F7), fh0011a (TF3), FJ003a (TF4), mz028a (antenatal), fm1a011 (FOM). This doesn’t include TF1-TF3 mums.

Ages for TF1-3 mums are taken from TF clinic files and compiled by Kate Northstone (file: ARIES\_age\_mums\_090315.xlsx). For TF1-3 mums, clinic and age of sampling is unknown. Kate Northstone: “There are a handful of duplicates”/ "Indicator vars have been created to show which clinic the sample MAY have been taken at (MAY because there was overlap between TF1 and TF2 - so we can't be sure which clinic she attended - match to actual clinic data to be sure)."

**Plate/Slide effects**

There are 134 batches of BCD plate (1-47 samples on each plate) and 500 batches of slides (1-12 samples on each slide). Samples from same plate ended on the same slide using a semi-randomisation procedure. Slide effects are confounded with sampletype in cord blood. eg. blood spots are grouped together on one slide and whitecells samples are all on the other slide. Although the technical probes capture some of the slide effects it doesn’t remove all of the slide effects and it is still necessary to regress out slide or plate effects.

**Slide row effects**

The staining probes which are part of the control probes tagging plate row effects. As the effect of the control probes are regressed out during the normalization, you don’t need to adjust for slide row anymore.

**Files:**

**samplesheet.release3.4854indiv.inclreplicates\_popstrat.Robj**: manifest file with columns extracted directly from LIMS and age, sex, aln, timepoint, timecode, sampletype, genotypeQC columns to remove population stratification samples, duplicate.rm column to remove duplicates.

**BetaMatrix:**

**Aries\_funnorm.randomeffect.pc10.160915.Robj:** normalized betas using functional normalization[1](#_ENREF_1). We used 10 PCs on the controlmatrix to regress out technical variation.

**Cellcounts:**

**cellcounts.4854.Houseman.txt:** Cell counts extracted using Houseman algorithm[2](#_ENREF_2) implemented in meffil. In addition, there are two files (“**cellcounts.cord.gervinandlyle.txt**” and “**cellcounts.cord.gse68456.txt**”) with predicted cellcounts for cord blood samples. The cord blood predictions are done using different reference datasets(gervinandlyle -unpublished and deGoede et al. [3](#_ENREF_3)).

**Controlmatrix:**

**control.matrix.4854.txt**: The 850 control probes are summarized in 42 control types. These probes can roughly be divided into negative control probes (613), probes intended for between array normalization (186) and the remainder (49), which are designed for quality control, including assessing the bisulfite conversion rate. None of these probes are designed to measure a biological signal. The summarized control probes can be used as surrogates for unwanted variation and are used for the functional normalization.

**Detection pvalues:**

**detpvalues.4854.Robj:** This matrix shows the detection pvalues for each sample and each CpG.

**Norm.objects:**

**norm.objects.clean160915.4854.Robj**: All data processing has been conducted using Meffil. Meffil uses *illuminaio* R package to parse Illumina IDAT files into a *meffil* object called *norm.objects.* All meffil functions, QC summary, functional normalization and post-normalization QC summary operate on the norm.objects. Specifically, the norm.objects contain raw control probe intensities, quantile distributions of the raw intensities, poor quality probes based on detection Pvalues and number of beads, predicted sex, predicted cellcounts and a samplesheet with batch variables. In addition, copy number variation can be extracted.

**QCobjects:**

**ARIESall\_qc.Robj**: This objects contain all 5469 samples extracted from LIMS and is not cleaned up. This object has been used to do the data cleaning. All data processing has been conducted using Meffil. Meffil uses *illuminaio* R package to parse Illumina IDAT files into a *meffil* object called *qc.objects.* All meffil functions, QC summary, functional normalization and post-normalization QC summary operate on the qc or norm.objects. Specifically, the qc.objects contain raw control probe intensities, poor quality probes based on detection Pvalues and number of beads, predicted sex, predicted cellcounts and a samplesheet with batch variables. In addition, copy number variation can be extracted.

**In docs:**

**QCreports:**

Using meffil on **ARIESall\_qc.Robj** we have generated a qc.report: **ARIESall\_qc-report.html.** Based on the plots here we have cleaned the data.

**Normalization reports:**

**{normalisation-report.pc\*160915.html}:** Using meffil on Aries\_funnorm.randomeffect.pc10.160915.Robj and norm.objects.clean160915.4854.Robj generates these reports. It shows PCA plots on control probes and normalized betas. In addition it shows associations between PCs and batch variables.

**#Remove duplicate and population stratification samples**

#load manifest file

load("samplesheet.release3.4854indiv.inclreplicates\_popstrat.Robj")

dim(samplesheet) #4854

table(samplesheet$duplicate.rm)

#Remove

# 71

table(samplesheet$genotypeQCmums)

# /highhet /IBD10per /Miss /strat

# 2 159 4 59

#NotInMothersGWAS Y

# 113 1642

table(samplesheet$genotypeQCkids)

#CR ETHNICITY HZT;ETHNICITY IMISS;HZT NotInKidsGWAS

# 325 76 56 6 34

# Y

# 2378

sample.rm<-which(samplesheet$duplicate.rm=="Remove"|samplesheet$genotypeQCkids=="ETHNICITY"|samplesheet$genotypeQCkids=="HZT;ETHNICITY"|samplesheet$genotypeQCmums=="/strat")

#261 samples are removed

length(sample.rm)

# 261

samplesheet\_filtered<-samplesheet[-sample.rm,]

dim(samplesheet\_filtered) # 4593

#remove these samples from beta matrix

pc=10

load(aries\_funnorm.randomeffect.pc10.160915.Robj")

m<-match(samplesheet$Sample\_Name,colnames(norm.beta))

norm.beta<-norm.beta[,m]

dim(norm.beta) #4593

**References**

1. Fortin, J.P., Labbe, A., Lemire, M., Zanke, B.W., Hudson, T.J., Fertig, E.J., Greenwood, C.M., and Hansen, K.D. (2014). Functional normalization of 450k methylation array data improves replication in large cancer studies. Genome biology 15, 503.

2. Houseman, E.A., Accomando, W.P., Koestler, D.C., Christensen, B.C., Marsit, C.J., Nelson, H.H., Wiencke, J.K., and Kelsey, K.T. (2012). DNA methylation arrays as surrogate measures of cell mixture distribution. BMC bioinformatics 13, 86.

3. de Goede, O.M., Razzaghian, H.R., Price, E.M., Jones, M.J., Kobor, M.S., Robinson, W.P., and Lavoie, P.M. (2015). Nucleated red blood cells impact DNA methylation and expression analyses of cord blood hematopoietic cells. Clinical epigenetics 7, 95.