QC on Aug. 21

GQ

08/21/2024

### Background

Data used in this report were first processed in two files,

*QC\_dataWranglingJuly30.Rmd* and *ReadinFeacureCountsAnnotation.R*.

* Data wrangling performed in *QC\_dataWranglingJuly30.Rmd*
  + organize all meta data together
  + filter data by RIN and delivery status
  + remove 3 samples with RIN equal to or less than 4
  + remove 3 samples with AD
* Gene counts processed in *ReadinFeacureCountsAnnotation.R*
  + Correct sample swaps

### Outline

-Filter lowly expressed genes

-Cut OMG into factors

-Regress out batch effects significantly correlated with PC1 rather than 4 batch effects

-Revisit the PCA to see which variables we should include in our models

### Pooled data set (110 samples)

* apc\_PC1/2/3 — ancestry PC1/2/3

To better know our data, variable names, summary statistics of numerical variables, and sample size for categorical variables shown as below.

Summary statistics of numerical variables

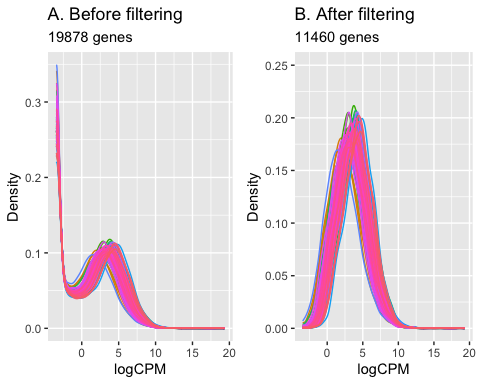
Sample size for sex (F-female, M-male) in each visit (v1/2/3/4 – 1st/2nd/3rd/4th visit) as shown below.

### Removing lowly expressed genes

Before filtering, I first annotated genes using the gene reference (basic gene annotation for GRCh38.p14 downloaded from GENCODE website), then filtered by Class (keeping “protein\_coding”), and Chromosomes (removing “chrY” and “chrM”). At last, we have **19,878** genes left.

**Keeping genes that have a count per million (CPM) > 1 in at least 25% of samples**

**11,000** genes were left after filtering lowly expressed genes.





**Fig. 1 The density of log-CPM values for raw pre-filtered data (A) and post-filtered data (B) shown for each sample**

Plot A shows lots of lowly expressed genes before filtering.

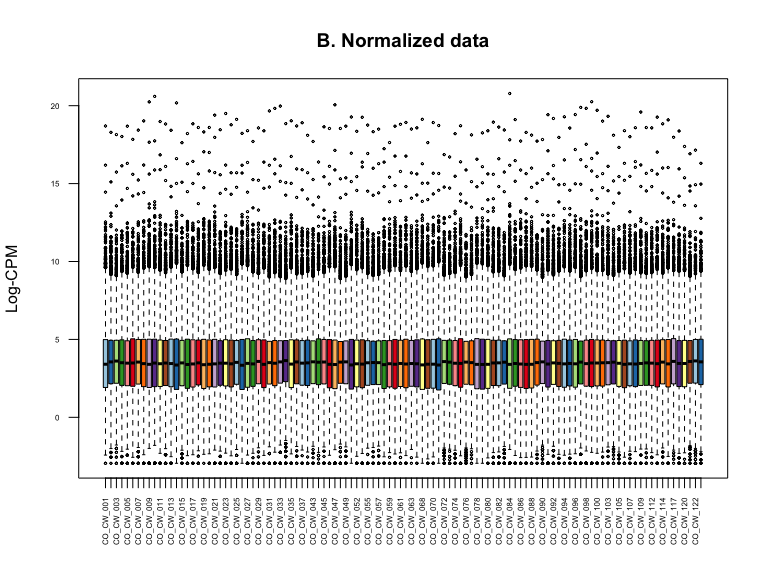
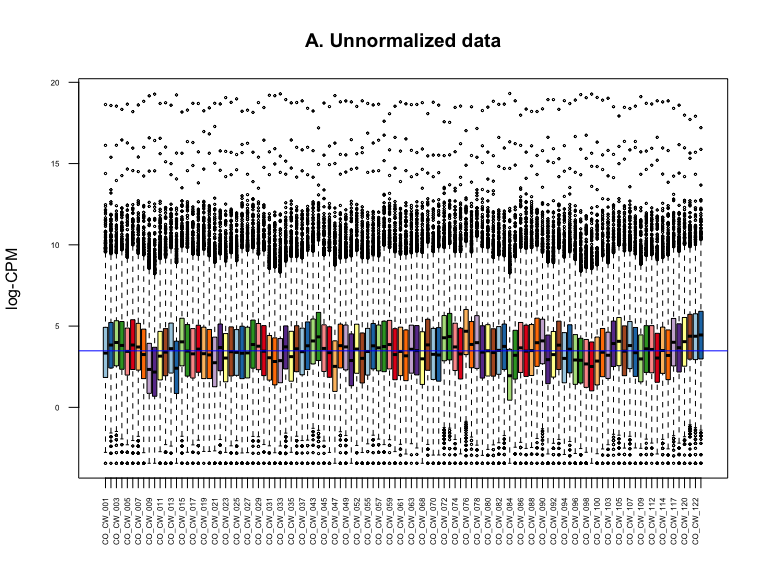
Plot B shows peaks of lowly expressed genes were removed after filtering.

##### Normalizing gene expression distributions

Using the method of **trimmed mean of M-values (TMM)** method to normalize and then remove batch effects.

TMM normalizes the library sizes to produce effective library sizes.

CPM values are counts normalized by the effective library sizes.



**Fig. 2 Boxplots of log-CPM values showing expression distributions for unnormalised data (A) and normalised data (B) for each sample**

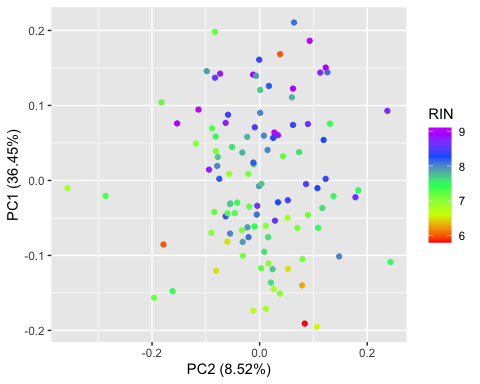
Distributions before normalization are noticeably different. Blue line indicates the median of log-CPM.

Distributions after normalization are similar.

### PCA

**Before batch effects removed**

**Fig. 3 PCA plots of first two PCs**



**Fig. 4 PCA plots of first two PCs with *RIN* colored by RIN levels**

**Fig. 5 PCA plots of first two PCs with *uniquely\_mapped\_pct* colored by the proportion of uniquely mapped reads**

**Fig. 6 PCA plots of first two PCs with *DHA* colored by DHA levels**

##### Assess correlation between all pairs of variables