

# IMM Batch adjustment of metabolite levels

Feb 10, 2016

## 1. Adjustment method

For each batch, we calculate median level of specific metabolite in three master controls. Then we calculate median of all the medians for every batches. Using these two types of the median values, we do batch adjustment.

In particular, we divide the raw metabolite level in the actual sample by median of the metabolite levels in the corresponding batch and then multiply it by the median of all the medians for every batches.

So the idea is that, after the batch adjustment, the median of master controls in each batch will be identical for all the batches while keeping the ratio of metabolite level in the actual sample and master controls the same before and after the batch adjustment.

More formally, let  $L_{m,s,b}$  be level of metabolite  $m$  level in sample  $s$  measured in batch  $b$ . First, calculate median metabolite level  $Q_2$  of quality control samples  $C_b$  in batch  $b$  ( $W_{m,b}$ ):

$$W_{m,b} = Q_{2_{s \in C_b}}[L_{m,s,b}]$$

Then calculate median of  $W_{m,b}$  over all batches  $B$  in the project ( $W_m$ ):

$$W_m = Q_{2_{b \in B}}[W_{m,b}]$$

Then calculate ratio between  $W_m$  and  $W_{m,b}$  ( $\alpha_{m,b}$ ):

$$\alpha_{m,b} = \frac{W_m}{W_{m,b}}$$

$\alpha_{m,b}$  is the signal drift correction factor for batch  $b$  for metabolite  $m$ . Finally multiply level of metabolite  $m$  ( $L_{m,s,b}$ ) by  $L_{m,s,b}$  to obtain its batch-adjusted metabolite level

$L_{m,s,b}^{\text{adj}}$ :

$$L_{m,s,b}^{\text{adj}} = \alpha_{m,b} \cdot L_{m,s,b}$$

## 2. Assigning small quantity

Before or after batch adjustment, if metabolite level is below certain level ( $\epsilon^*$ ) or if the value of the metabolite is missing (possibly meaning that the metabolite is below detection level), we assign small quantity ( $\epsilon$ ) to that metabolite level. This can be done to both control or actual samples. Currently  $\epsilon^* = 0.001$  and  $\epsilon = 0.1$  are used.

## 3. Batch discarding

Data for specific metabolite in specific batch may be discarded if one of the following conditions is met:

- (1) Median metabolite level of control samples in a batch is  $\epsilon$  or below.
- (2) Median metabolite level of control samples in a batch is outlier with respect to those in other batches
- (3) Large variation of metabolite levels of control samples within a batch is observed

For (2), we calculate first quantile ( $Q_1$ ), median ( $Q_2$ ), third quantile ( $Q_3$ ) and interquartile range ( $\varphi = Q_3 - Q_1$ ) of  $\{ \log_2(W_{m,b}) \}$ ,  $b=b_1, b_2, \dots, b_p \in B$ . Then if one of the following conditions is satisfied, all data for the specific metabolite  $m$  from batch  $b$  will be discarded.

$$\log_2(W_{m,b}) \geq Q_3 + \lambda\varphi$$

$$\log_2(W_{m,b}) \leq Q_1 - \lambda\varphi,$$

Currently,  $\lambda = 1.5$ .

For (3), variance of metabolite level of control samples within a batch can be calculated with modified coefficient of variation (mCV) based on Absolute Deviation of Median (ADM) which is analogous to standard deviation calculated based on the mean ( $E$ ) of the samples  $s_1, s_2, \dots, s_n \in S$  ( $n = 3$ ,  $s_i \in C_b$  in our case). The definition of the standard deviation ( $\sigma$ ) is:

$$\sigma = \sqrt{E_i[(s_i - E[S])^2]}$$

Replacing  $E$  with  $Q_2$  gives ADM ( $\sigma_*$ ):

$$\sigma_* = \sqrt{Q_{2_i}[(s_i - Q_2[S])^2]}$$

There exists  $k$  such that:

$$Q_{2_i}[(s_i - Q_2[S])^2] = (s_k - Q_2[S])^2$$

and therefore ADM can be rewritten by as:

$$\sigma_* = \sqrt{(s_k - Q_2[S])^2} = |s_k - Q_2[S]| = Q_{2_i}[|s_i - Q_2[S]|]$$

Thus, mCV can be calculated by  $\sigma_* / Q_2(S)$ . Currently, we set threshold of mCV to 1.0.

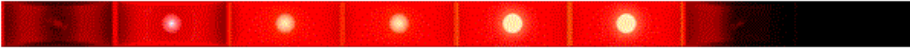
#### 4. Other possible factors to be considered

- Differences between metabolite levels of controls and actual samples within a batch
- Variations of actual samples

## 5. The system for batch adjustment

Web accessible, password-protected.

← → ↻ 🏠  ☆



### Batch Adjustment of Metabolome Data

♣ For single metabolome data table file

---

Control sample name prefix:

Representative value calculation

Small quantity to assign to metabolite levels below that quantity including 0 and no value:

Threshold of coefficient of variance (CV) of controls within each batch:

Factor to identify median of controls within a batch that is outlier with respect to other control medians in different batches:

Comma in number representation: ☒ Regard comma "," as a decimal point

Select metabolome data file:

☐ **metabsinglefile\_sample1\_5.tsv** - Test [ Jan. 26, 2016, 5:37 a.m. ]  
Test metabolome data file

☐ **CRIC1to405\_per\_umolcreat\_H2Osubtr2\_Ben1.tsv** - CRIC [ Jan. 24, 2016, 7:33 p.m. ]  
CRIC samples 1 to 405 normalized by creatinine, H2O subtracted, updated by Ben

☐ **CRIC1to405\_per\_umolcreat\_H2Osubtr1\_1.tsv** - CRIC [ Jan. 10, 2016, 4:01 p.m. ]  
CRIC 405 case samples, batch 1 to 26, per creatinine, H2O subtracted.