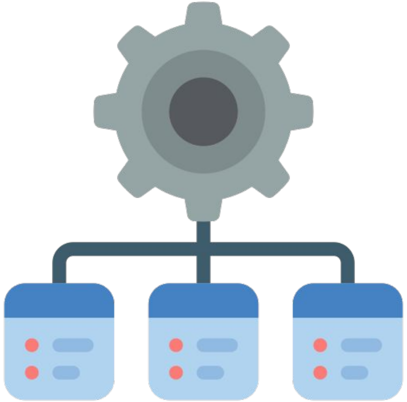
Batch effect and batch correction for image-based profiling

Fernanda Garcia Fossa



Summary

1

Batch effect

2

Empirical Bayes
&
Conditional
probability

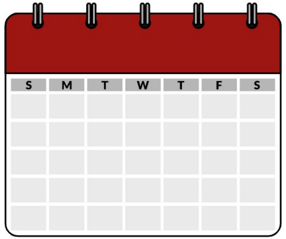
3

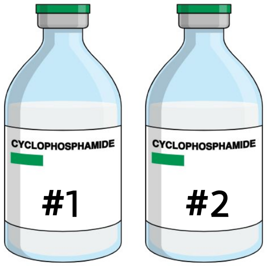
Baseball
example

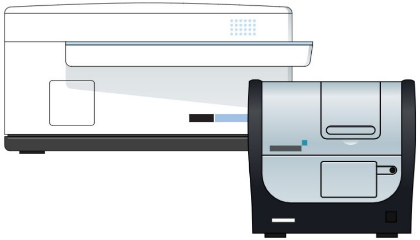
4

ComBat and
how it corrects
for batch effect

Batch effect are unrelated to biological variables

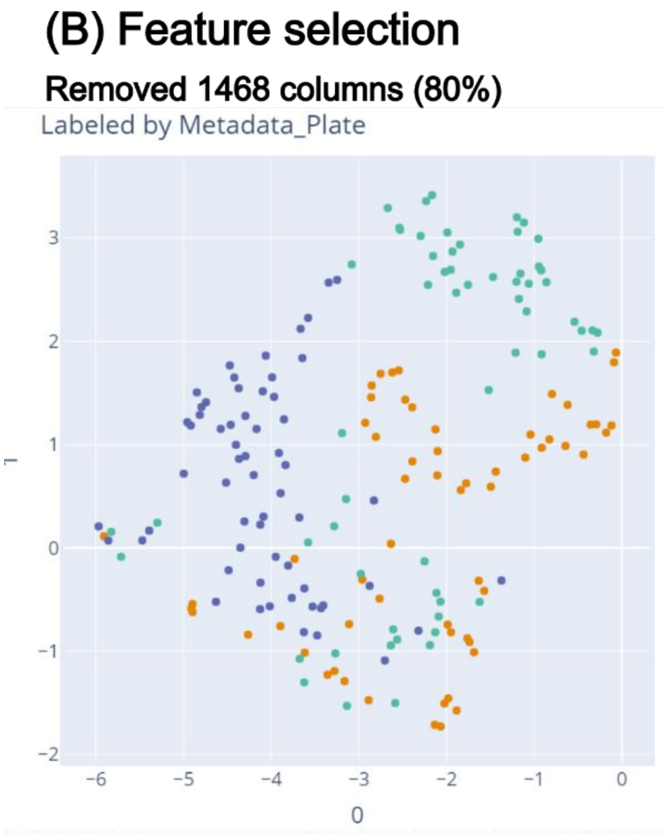


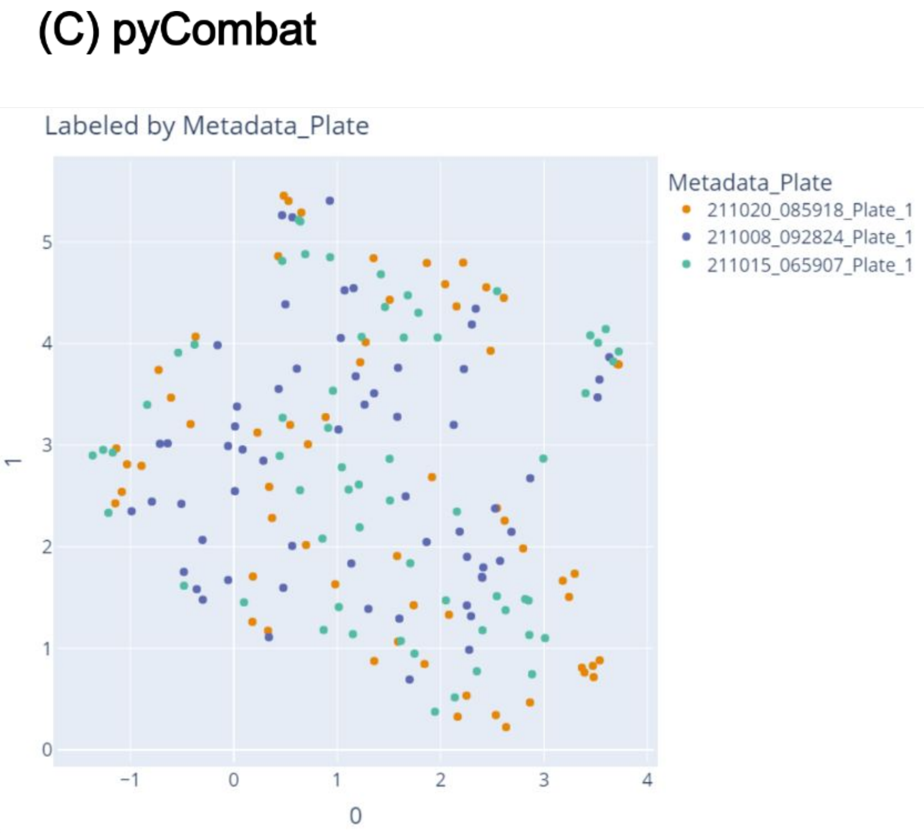




Leek et. al, 2010 Nature Reviews Genetics

Batch effects can be corrected after data acquisition

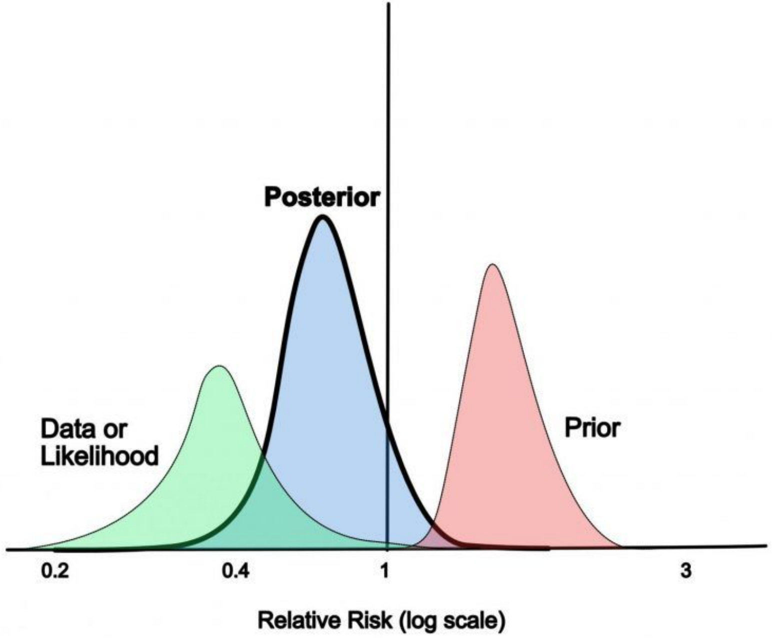




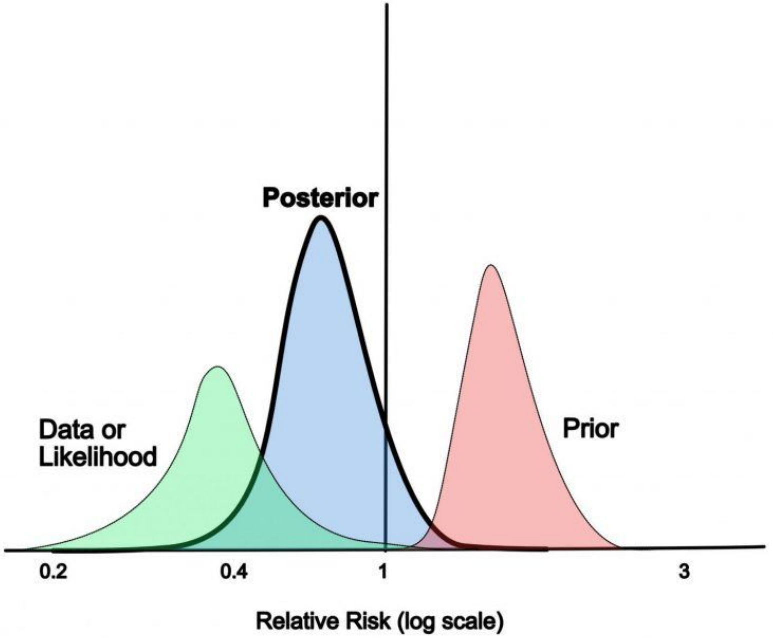
Conditional probability & Empirical Bayes

- Given that we know one thing about an event can be derived from knowing the other thing about the event

- Bayesian statistics - knowing how to take a guess



Conditional probability & Empirical Bayes



B = data
A = model to describe the data (ideal outcome)

$$P(A/B) = \frac{P(B/A).P(A)}{P(B)}$$

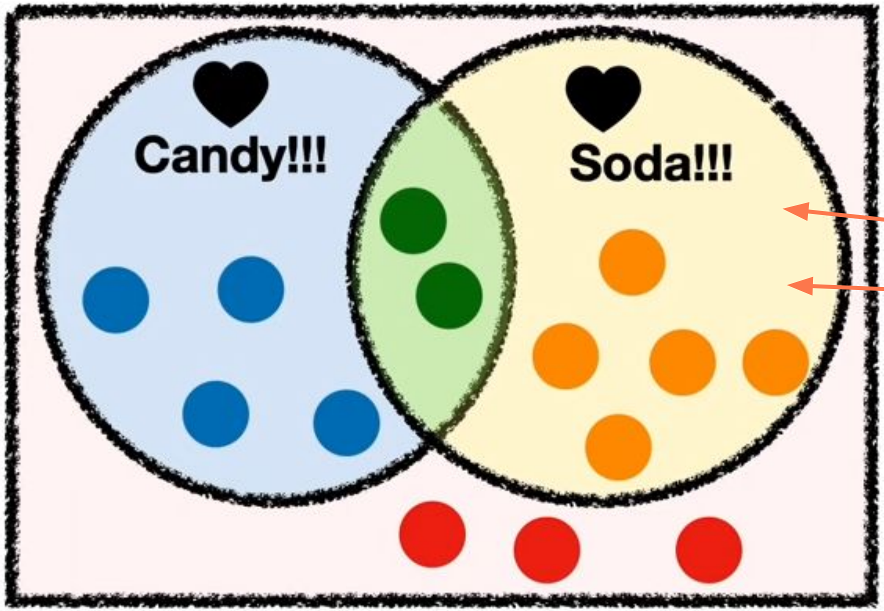
P (B/A): likelihood (making the measurement B given that the model A is correct)

P(A): prior, belief that the model is true before measurements are made

P(B): probability of collecting the dataset B

P(A/B): probability of the model after the data has been collected

Bayes' theorem



A = does not love candy

B = loves soda

$$p(A \,\&\, B \mid B) = \frac{p(A \,\&\, B \mid A) \times p(A)}{p(B)}$$

$$p(A \,\&\, B \mid A) = \frac{p(A \,\&\, B \mid B) \times p(B)}{p(A)}$$

StatQuest https://www.youtube.com/watch?v=9wCnvr7Xw4E&t=324s

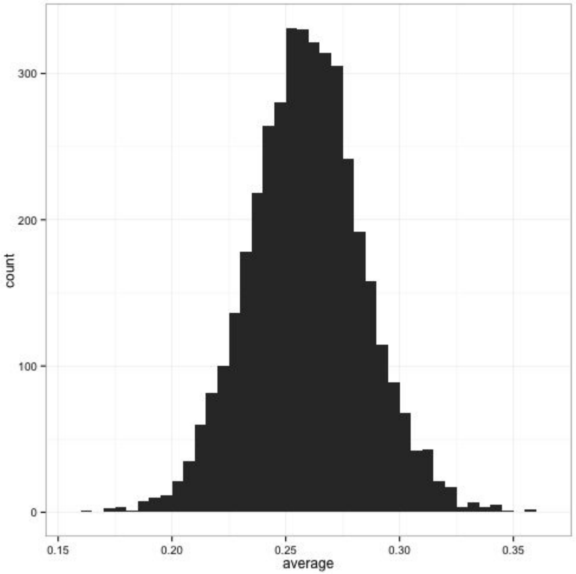
Applying empirical Bayes - Baseball

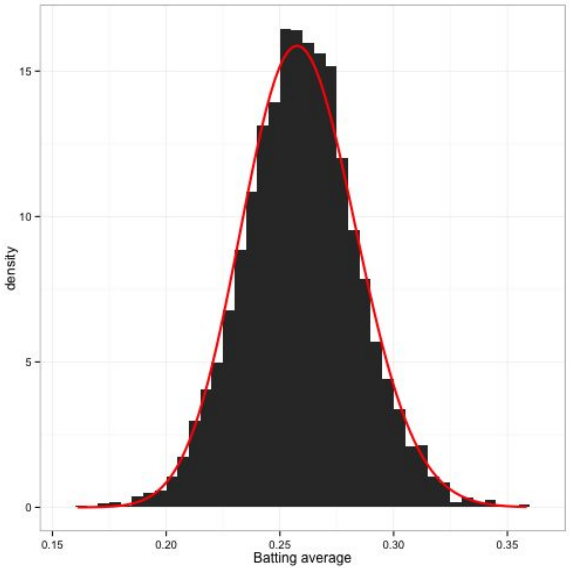
- Best hitters (H) in history of baseball;

name	H	AB	average
Jeff Banister	1	1	1
Doc Bass	1	1	1
Steve Biras	2	2	1
C. B. Burns	1	1	1
Jackie Gallagher	1	1	1
	4	10	0.4
	300	1000	0.3

http://varianceexplained.org/r/empirical_bayes_baseball/

Plot all the averages of hitters





α and β are the priors

Used to calculate a new corrected average

$$X \sim \mathrm{Beta}(\alpha_0, \beta_0)^*$$

* It can be mean and variance

http://varianceexplained.org/r/empirical_bayes_baseball/

It corrects the averages using empirical Bayes

Previous averages

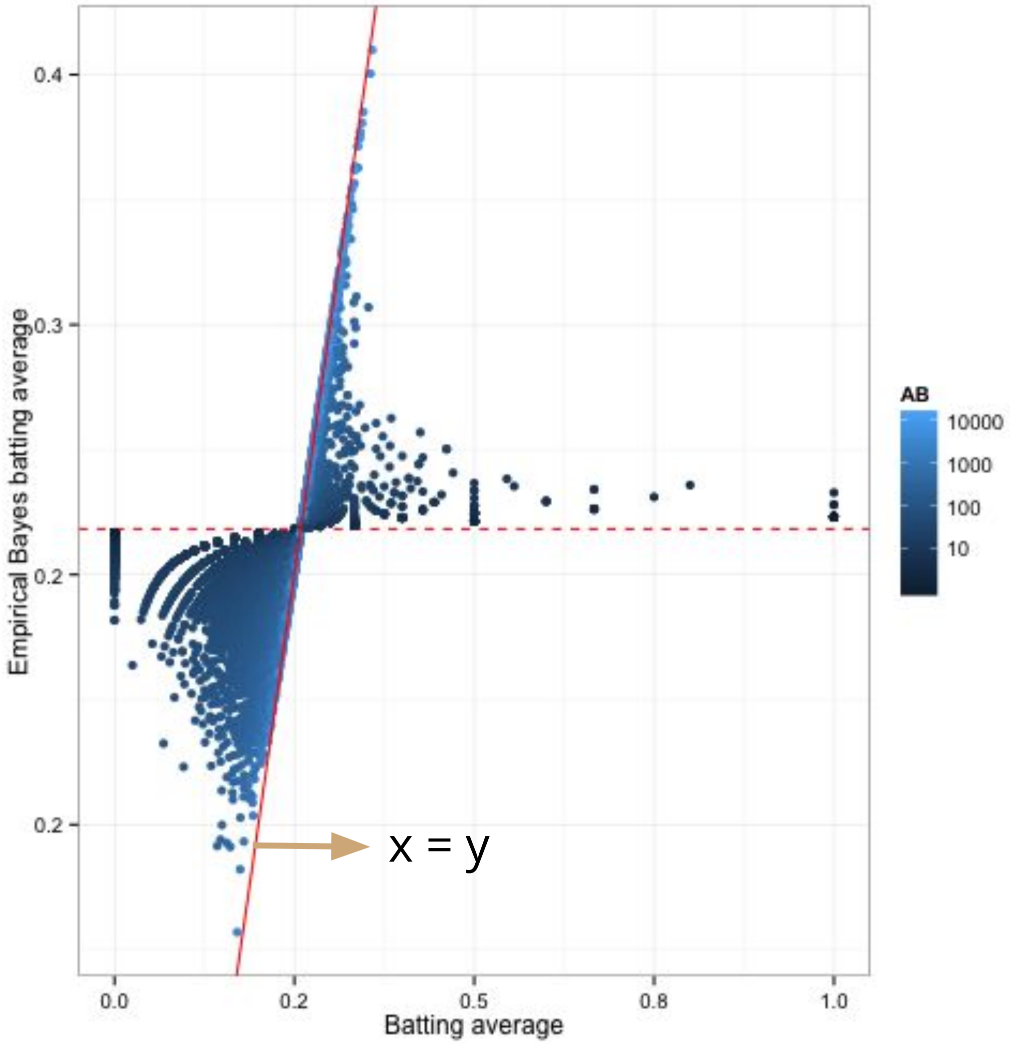
$$\frac{300+\alpha_0}{1000+\alpha_0+\beta_0}=\frac{300+78.7}{1000+78.7+224.9}=0.29$$ 0.3

$$\frac{4+\alpha_0}{10+\alpha_0+\beta_0}=\frac{4+78.7}{10+78.7+224.9}=0.264$$ 0.4

http://varianceexplained.org/r/empirical_bayes_baseball/

With **less** observations, the **more** the point moves;

With **more** observations, the **less** the point moves



http://varianceexplained.org/r/empirical_bayes_baseball/

pyComBat: adaptation of ComBat to Python

Biostatistics (2007), **8**, 1, *pp.* 118–127
doi:10.1093/biostatistics/kxj037
Advance Access publication on April 21, 2006

Adjusting batch effects in microarray expression data using empirical Bayes methods

W. EVAN JOHNSON, CHENG LI*

Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA, USA and Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA
cli@hsph.harvard.edu

ARIEL RABINOVIC

Department of Genetics and Complex Diseases, Harvard School of Public Health, Boston, MA, USA

ComBat concepts

1. Information we have - **what are the batches** and the feature values

2. Instead of making a random guess, we use the data we have to make a guess and get a **prior**

3. For each feature in each batch, two priors are calculated by fitting linear models

4. Priors are used to correct the data to what it should be (shrinkage)

ComBat premisses

1. Data must be **scaled/normalized** beforehand (unnormalized could bias the batch effect prior estimation);

2. Location and scale adjustments (L/S) - a model for the location (**mean**) and scale (**variance**) of the data WITHIN BATCHES.

3. Batch is **modeled/factored out** by standardizing means and variances across batches.

Location/Scale model

- Mean center and standardize the variance of each batch for **each** gene/feature **independently**;

PRIOR: Additive
Batch effect

PRIOR: Multiplicative
Batch effect

Feature Y of sample j in batch i

$$Y_{ijg} = \alpha_g + X\beta_g + \gamma_{ig} + \delta_{ig}\varepsilon_{ijg}$$

Design matrix (X) + regression coefficient Beta

Overall feature values

1st step: standardization

- Data MUST be normalized/standardized before applying the correction
- If not, it could bias the estimation of the parameters

Considers

$$Z_{ijg} = \frac{Y_{ijg} - \widehat{\alpha}_g - X\widehat{\beta}_g}{\widehat{\sigma}_g}$$

- Mean, variance,
  and size of dataset

2nd step: estimate empirical priors

- The two parameters are estimated empirically from standardized data using the **method of moments** = **mean and variance of the data**

1. **Additive prior γ**: This assumes that the impact of the batch is consistent across all values of the feature.
2. **Multiplicative prior δ**: This assumes that the impact of the batch is proportional to the original values of the feature.

$$\gamma_{ig}^{*} = \frac{n_i \bar{\tau}_i^2 \hat{\gamma}_{ig} + \delta_{ig}^{2*} \bar{\gamma}_i}{n_i \bar{\tau}_i^2 + \delta_{ig}^{2*}} \quad \text{and} \quad \delta_{ig}^{2*} = \frac{\bar{\theta}_i + \frac{1}{2} \sum_j (Z_{ijg} - \gamma_{ig}^{*})^2}{\frac{n_j}{2} + \bar{\lambda}_i - 1}.$$

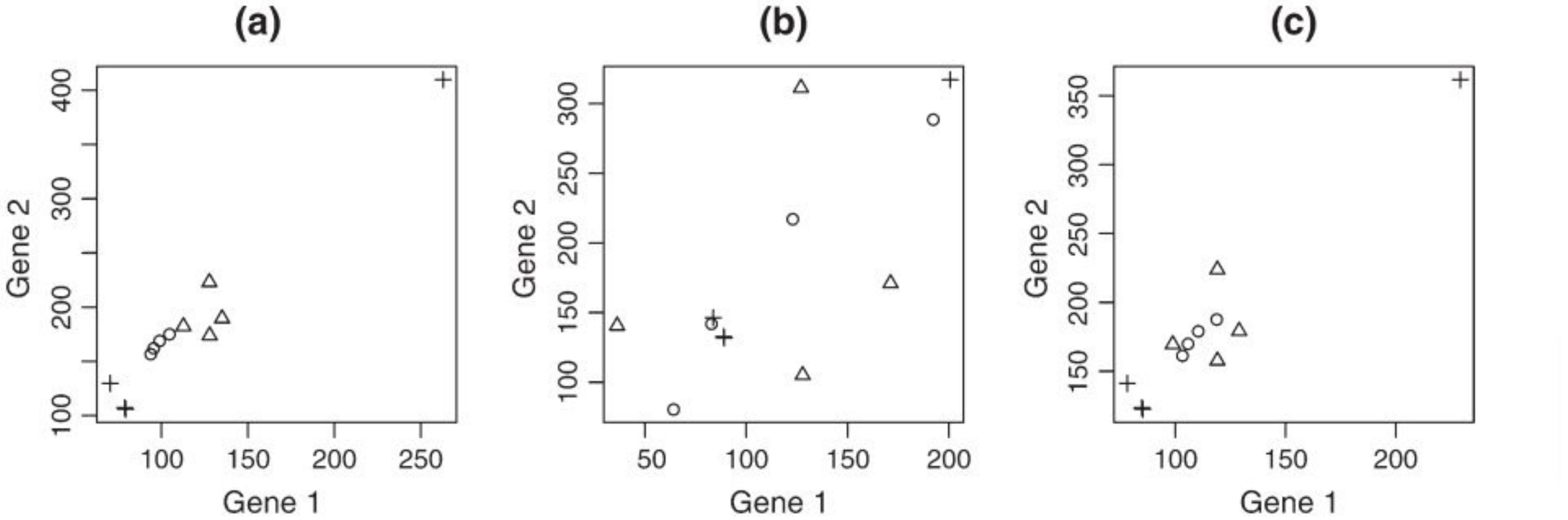
Finally, adjust for batch effects

Empirical Bayes batch adjusted data:

$$\gamma^{*}_{ijg} = \frac{\widehat{\sigma}_g}{\widehat{\delta}^{*}_{ig}}(Z_{ijg} - \widehat{\gamma}^{*}_{ig}) + \widehat{\alpha}_g + X\widehat{\beta}_g.$$

$$Y_{ijg} = \alpha_g + X\beta_g + \gamma_{ig} + \delta_{ig}\varepsilon_{ijg} \longrightarrow \text{L/S}$$

Empirical Bayes is robust to outliers



(a) Raw
(b) Only L/S corrected
(c) Empirical Bayes corrected

Before

Standardization pyCombat

After EB correction