Removing Batch Effects from Genomics Data

W. Evan Johnson, Ph.D.
Professor, Division of Infectious Disease
Director, Center for Data Science
Director, Center for Biomedical Informatics and Health AI
Rutgers University – New Jersey Medical School

2025-08-03

ComBat Intuition

Consider the following model:

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

where:

- α_g is the overall gene expression
- X is a design matrix
- β_g contains the regression coefficients
- The error terms $\epsilon_{ijg} \sim N(0, \sigma_g^2)$
- γ_{ig} and δ_{ig} are additive and multiplicative batch effects

(Johnson et al., Biostatistics, 2007)}}

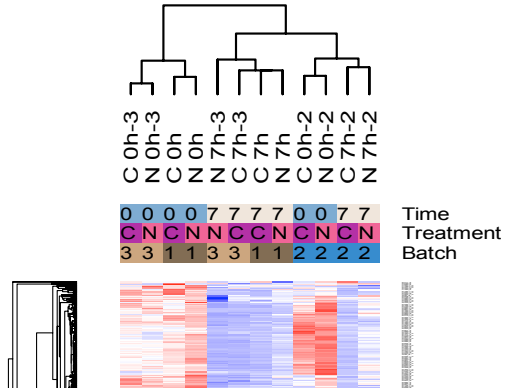
Batch effects

Batch Effect: Non-biological variation due to differences in batches of data that confound the relationships between covariates of interest.

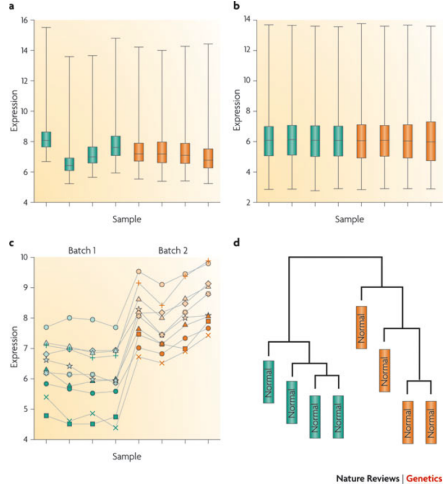
Caused by differences:

- Gene expression profiling platform
- Lab protocol or experimenter
- Time of day or hybridization
- Atmospheric ozone level (Rhodes et al. 2004)

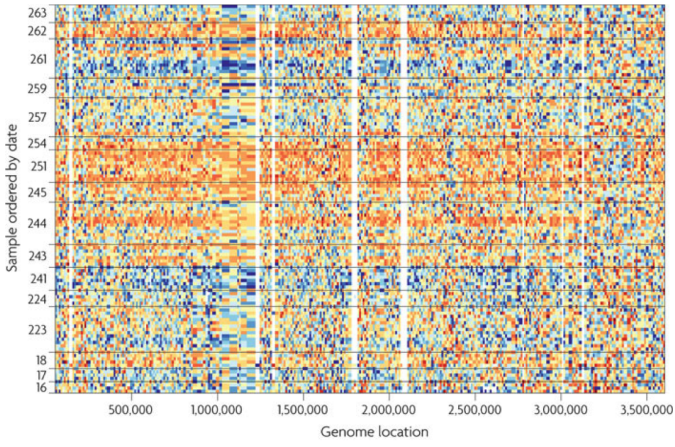
Batch effect examples: Nitric Oxide



Batch effect examples: Bladder cancer



Batch effect examples: 1000 genomes



Nature Reviews | Genetics

Batch effect examples: Proteomics

Proteomic data with batch effects:

- Proteomic markers to predict endometriosis (39 total)
- Single peptide predictors of disease (AUC): 0.82, 0.76, 0.74, 0.74, 0.70 (+12 more $>$0.6)
- Single peptide predictors of batch (AUC): 0.99, 0.94, 0.91, 0.86, 0.86, 0.84, 0.84, 0.84, 0.83, 0.82 (+7 more $>$0.6)
- Predict batch better than disease!

Normalization?

- **Question:** Shouldn't normalization take care of this?
  - Answer: NO!
- Batch effects often impacts genes or sets of genes
- Batch effects often remain after normalization

WEJ LAB

Adjusting for batch effects

Center for Data Science

Early methods for batch effects

- Singular value decomposition (Alter et al., 2000)
- Distance weighed discrimination (Benito et al., 2004)
- ComBat (Johnson et al., 2007)
- Surrogate variable analysis (Leek and Storey, 2007)
- Cross Platform Normalization (Shabalin er al., 2008)
- Barcoding (Zilliox and Irizarry 2007; Piccolo et al. 2013)
- Single sample normalization (Hubbell et al., 2002; McCall et al., 2010; Piccolo et al. 2012)
- Removing unwanted variation (Risso et al., 2014; Jacob et al., 2016)

ComBat Intuition

Consider the following model:

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

where:

- α_g is the overall gene expression
- X is a design matrix
- β_g contains the regression coefficients
- The error terms $\epsilon_{ijg} \sim N(0, \sigma_g^2)$
- γ_{ig} and δ_{ig} are additive and multiplicative batch effects

ComBat Intuition

Adjust for batch effects:

$$Y^*_{ijg} = \frac{Y_{ig} - \hat{\alpha}_g - X\hat{\beta}_g - \hat{\gamma}_{ig}}{\hat{\delta}_{ig}} + \hat{\alpha}_g + X\hat{\beta}_g$$

Problem: How to robustly estimate the parameters?

Answer: Empirical Bayes!

Step 1: Standardize the data

Genes are on different scales, so first standardize the data:

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

where $\hat{\alpha}_g$, $\hat{\beta}_g$, and $\hat{\sigma}_g^2$ are estimated using gene-wise MLEs.

Step 2: EB batch effect estimates

Additive batch adjustment (γ_{ig}): Using Bayes theorem and a *Gaussian*(γ_i, τ_i^2) conjugate prior:

$$E[\gamma_{ig}|\mathbf{Z_{ig}}, \delta_{ig}^2] = \frac{\tau_i^2 \sum_j Z_{ijg} + \delta_{ig}^2 \gamma_i}{n_i \tau_i^2 + \delta_{ig}^2}.$$

which can therefore be estimated as

$$\gamma_{ig}^* = \hat{E}[\gamma_{ig}|\mathbf{Z_{ig}}, \delta_{ig}^2] = \frac{n_i \bar{\tau}_i^2 \hat{\gamma}_{ig} + \delta_{ig}^{2*} \bar{\gamma}_i}{n_i \bar{\tau}_i^2 + \delta_{ig}^{2*}}.$$

Step 2: EB batch effect estimates

Mulitplicative batch adjustment (δ_{ig}): Using Bayes theorem and an *Inverse Gamma*(λ_i, θ_i) conjugate prior:

$$E[\delta_{ig}^2 | \mathbf{Z_{ig}}, \gamma_{ig}] = \frac{\theta_i + \frac{1}{2}\sum_j (Z_{ijg} - \gamma_{ig})^2}{\frac{n_i}{2} + \lambda_i - 1}.$$

which can therefore be estimated as

$$\delta_{ig}^{2*} = \hat{E}[\delta_{ig}^2 | \mathbf{Z_{ig}}, \gamma_{ig}] = \frac{\bar{\theta}_i + \frac{1}{2}\sum_j (Z_{ijg} - \gamma_{ig}^*)^2}{\frac{n_i}{2} + \bar{\lambda}_i - 1}.$$

Step 2: EB batch effect estimates

Estimate hyperpriors using the Method of Moments across all genes:

For the additive batch adjustment:

$$\bar{\gamma}_i = \frac{1}{G}\sum_g \hat{\gamma}_{ig}, \text{ and } \bar{\tau}_i^2 = \frac{1}{G-1}\sum_g (\gamma_{ig} - \bar{\gamma}_i)^2.$$

For the multiplicative batch adjustment:

$$\bar{\lambda}_i = \frac{\bar{V}_i + 2\bar{S}_i^2}{\bar{S}_i^2} \text{ and } \bar{\theta}_i = \frac{\bar{V}_i^3 + \bar{V}_i\bar{S}_i^2}{\bar{S}_i^2}.$$

Step 2: EB batch effect estimates

Alternative non-parametric prior, assume:

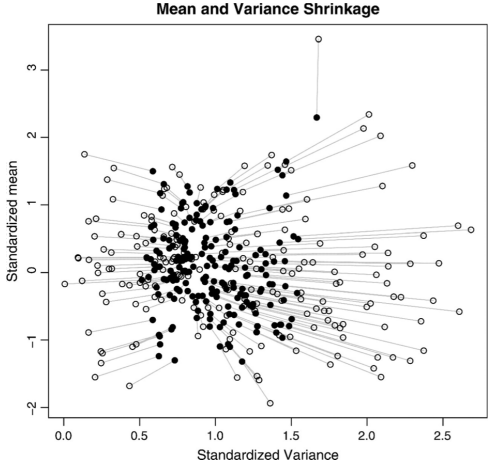
$$Z_{ijg} \sim N(\gamma_{ig}, \delta^2_{ig}), \text{ and let } w_{igk} = L(\mathbf{Z_{ig}}|\hat{\gamma}_{ik}, \hat{\delta}^2_{ik}).$$

The nonparametric EB batch adjustments $\gamma^*_{ig}, \delta^{2*}_{ig}$ are given by

$$\gamma^*_{ig} = \frac{\sum_k w_{igk}\hat{\gamma}_{ik}}{\sum_k w_{igk}} \text{ and } \delta^{2*}_{ig} = \frac{\sum_k w_{igk}\hat{\delta}^2_{ik}}{\sum_k w_{igk}},$$

i.e. Monte Carlo integration over an unspecified empirical prior.

Empirical Bayes Shrinkage

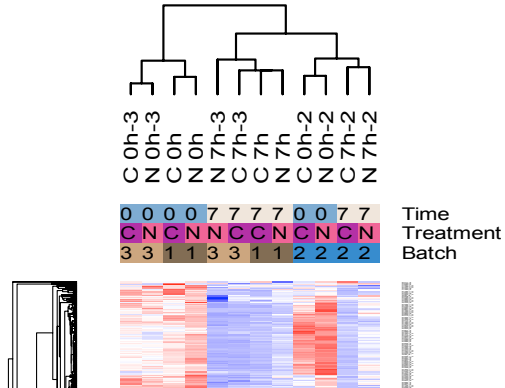


Step 3: Adjust the data for batch effects

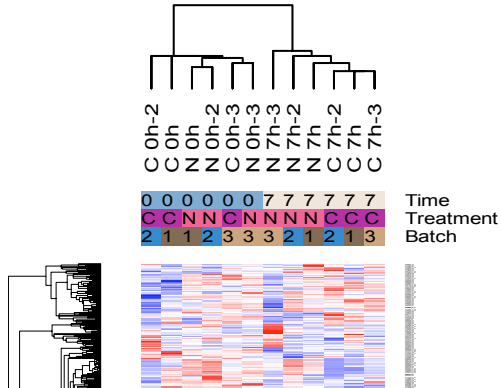
The EB batch adjusted data Y^*_{ijg} can be calculated using EB estimated batch effects:

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

Batch effect examples: Nitric Oxide



Batch effect examples: Nitric Oxide



Batch effect examples: Proteomics

Proteomics data with batch effects:

- Proteomic markers to predict endometriosis (39 total)
- Single peptide predictors of disease (AUC): 0.82, 0.76, 0.74, 0.74, 0.70 (+12 more $>$0.6)
- Single peptide predictors of batch (AUC): 0.99, 0.94, 0.91, 0.86, 0.86, 0.84, 0.84, 0.84, 0.83, 0.82 (+7 more $>$0.6)
- Predict batch better than disease!

Batch effect examples: Proteomics

After ComBat batch adjustment:

- Single peptide predictors of disease (AUC): 0.80, 0.79, 0.75, 0.70, 0.70 ($+7$ more \$$>$\$0.6)
- Single peptide predictors of batch (AUC): 0.64, 0.60

WEJ LAB

Confounded Designs

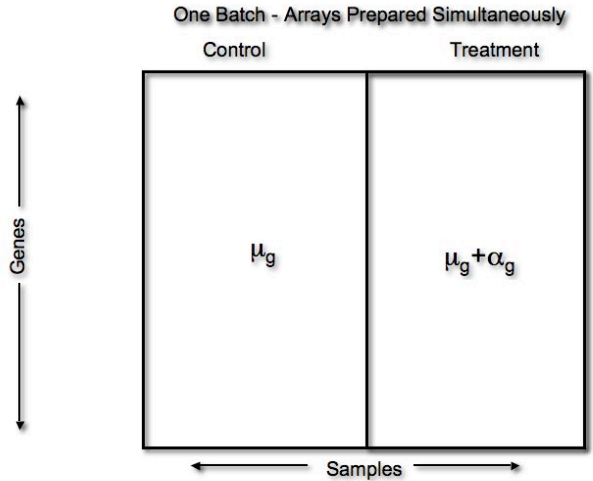
Center for Data Science

Confounded Designs

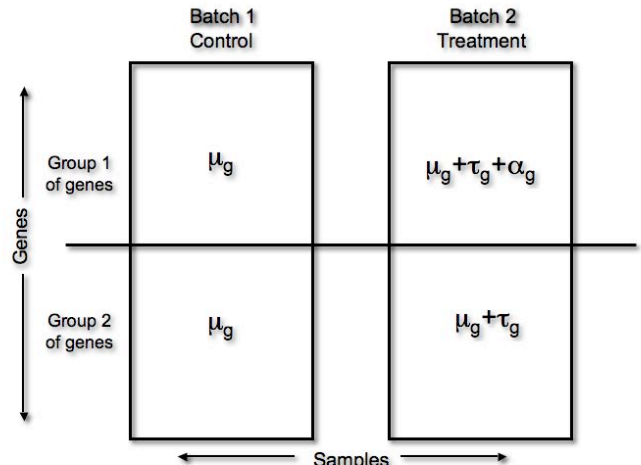
What do you do in the following cases?

- Treatments in one batch, controls in another
- Treatment 1 in Batch 1, Treatment 2 in Batch 2
- Same experiment with different tissues or cell lines
- Some samples are in their own batch?
- Can I do with anything with these cases?

Confounded Designs



Confounded Designs—Solution!

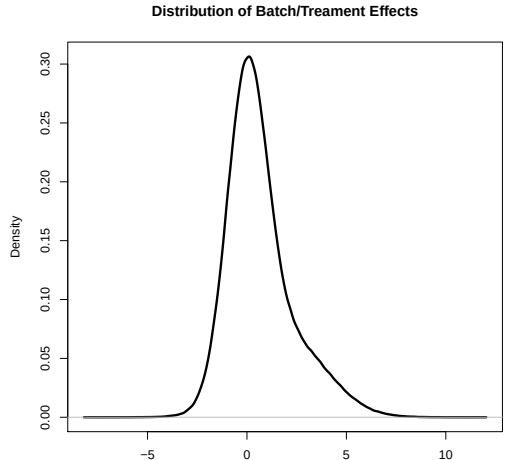


Confounded Designs—Solution!

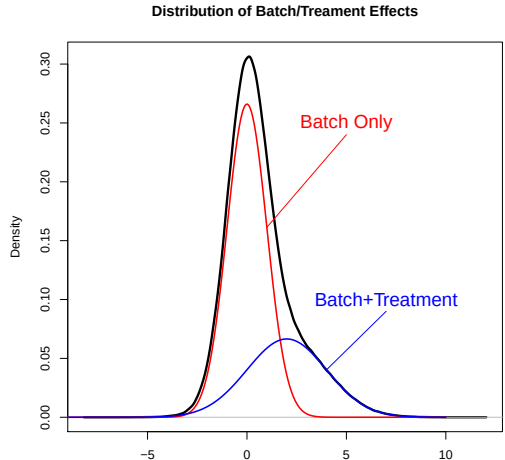
Assumptions for parametric ComBat: 1. Gaussian data: $Z_{ijg} \sim N(\gamma_{ig}, \delta_{ig}^2)$ 2. Systematic batch effects: $\gamma_{ig} \sim N(\gamma_i, \tau_i^2)$, $\delta_{ig}^2 \sim IG(\lambda_i, \theta_i)$

Assumptions for confounded ComBat: 1. $Z_{ijg} \sim N(\gamma_{ig}, \delta_{ig}^2)$ 2. Some (but not all) have treatment effects: $\phi_g \sim Bern(\pi)$ 3. Systematic: $\gamma_{ig} \sim N(\gamma_i + \phi_g \eta_i, \tau_i^2 + \phi_g \rho_i^2)$, $\delta_{ig}^2 \sim IG(\lambda_i, \theta_i)$

Confounded Designs—Solution!



Confounded Designs—Solution!



Confounded Designs—Solution!

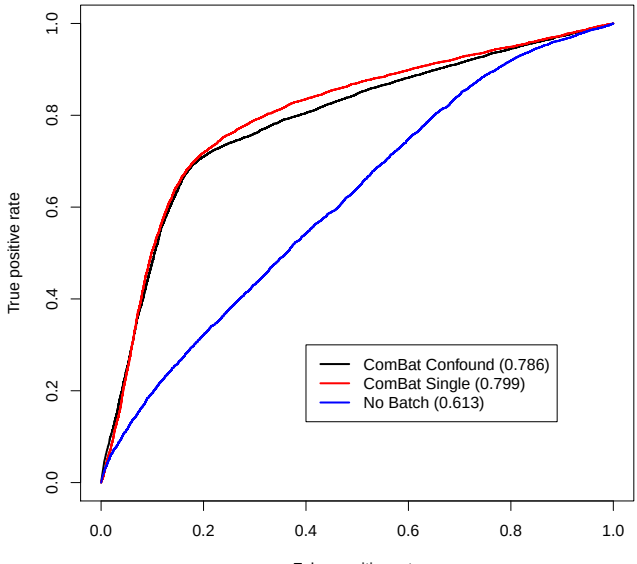
Confounded ComBat steps:

1. Standardize the data
2. Estimation of the hyper priors via the EM algorithm
3. EB batch/treatment estimates using maximum posterior estimates
4. Adjust the data and restore original scaling

Note: This works on single samples as well!

RAS and PI3K Pathway Activation

- Human primary mammary epithelial cells
- Transfect with an adenovirus expressing RAS or PI3K
- Observe change compared to control
- RAS and PI3K profiled at different times on different platforms
- Control cells are the same!


PI3K vs. RAS
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
ComBat Confound (0.786)
ComBat Single (0.799)
No Batch (0.613)

RAS and PI3K Pathway Activation

Results with p-value cutoff of 0.001:

	PI3K vs. RAS		
	Sens	Spec	AUC
ComBat Confound	0.747	0.730	0.786
ComBat Single	0.764	0.740	0.799
No Batch	0.787	0.361	0.613

Confounding: decrease in sensitivity (25-29%) and specificity (22-26%)

Session Info

```
sessionInfo()
```

```
## R version 4.5.1 (2025-06-13)
## Platform: aarch64-apple-darwin20
## Running under: macOS Sequoia 15.5
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.5-arm64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.5-arm64/Resources/lib/libRlapack.dylib;  LAPACK version 3.12.1
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## time zone: America/New_York
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## loaded via a namespace (and not attached):
##  [1] compiler_4.5.1    fastmap_1.2.0     cli_3.6.5         tools_4.5.1
##  [5] htmltools_0.5.8.1 rstudioapi_0.17.1 yaml_2.3.10       rmarkdown_2.29
##  [9] knitr_1.50        xfun_0.52         digest_0.6.37     rlang_1.1.6
## [13] evaluate_1.0.4
```