Empirical Bayes Method for scRNA-Seq Batch Effects

Poster 2490 M. Colange[1], G. Appé[1], L. Meunier[1], S. Weill[1], A. Nordor[1], A. Behdenna[1]. [1]Epigene Labs, Paris, France

AACR American Association for Cancer Research

Batch effect correction adapted to the unique distributions of single-cell transcriptomes

Contact
Akpéli Nordor, PharmD, PhD
(akpeli@epigenelabs.com).

CHECK US OUT!

The authors have no conflict of interest to declare.

Research question

Technical variability across datasets

- Datasets from different sources suffer from **technical effects**, caused by differences during sequencing such as protocol, or equipment.

Impact on single-cell data

- Single-cell RNA-Seq data are particularly sensitive to batch effects due to their high sparsity, lower RNA content, and cell-to-cell variability.

Consequences for analysis

- Without correction, batch effects can obscure true biological signals, misleading downstream analysis.

Cell type
- B cell
- Dentritic cell
- Epithelial cell
- Fibroblast
- Innate lymphoid cell
- Malignant cell
- Muscle cell
- Myeloid cell
- Pericyte
- Plasma cell
- T cell

Raw data with uncorrected batch effects

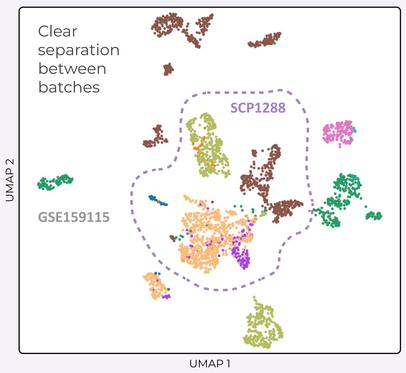
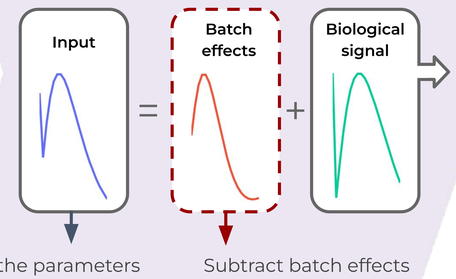


Figure 1: UMAP before batch effect correction.
2 Kidney datasets GSE159115, SCP1288 (See Results part for cohort details)

Batch effect correction

- Reduces the technical effects and makes sure the biological signal dominates.
  - Decompose the input data in 2 components



Corrected data showing biological structure

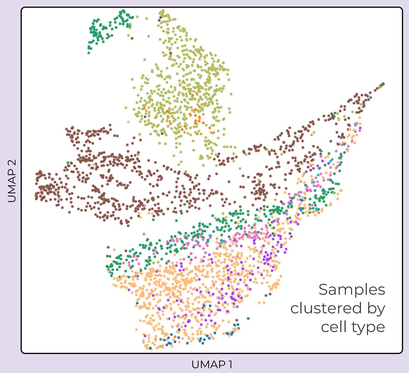


Figure 1: UMAP after batch effect correction.
2 Kidney datasets GSE159115, SCP1288 (See Results part for cohort details)

References

1. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007 Jan 1;8(1):118–27.
2. Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genomics Bioinforma. 2020 Sep;2(3):lqaa078.
3. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021 Jun 24;184(13):3573-3587.e29.

Results

Cohorts creation

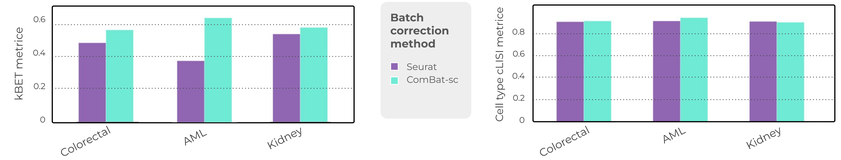
Colorectal cohort	AML cohort	Kidney cohort
3 datasets	2 datasets	2 datasets
14 cell types	7 cell types	11 cell types

For each dataset, **3,000 cells were selected to preserve the original cell type composition**. In each cohort, **4,000 highly variable genes were used**, capturing varying levels of integration complexity (i.e., number of datasets and cell types).

Comparison with Seurat

pyComBat-sc improves batch mixing by **31,11%** compared to Seurat

pyComBat-sc improves cell type clustering by **1,04%** compared to Seurat



kBET measures the bias of a batch variable in the kNN graph.

Cell-type LISI (cLISI) measures the quality of the clustering of cells by cell type.

All quality metrics are scaled between 0 and 1, with higher values indicating better performance.

Approach

Limitations of Existing Methods

- Methods developed for bulk RNA-Seq often fail to account for the sparsity and variability of single-cell data.
- Many single-cell batch correction tools return only low-dimensional embeddings, without gene-level corrected matrices.
  → limiting downstream analyses and reducing data interpretability.

We adapt the **Empirical Bayes framework to better fit the distributional complexity of single-cell data**, enabling effective batch correction while preserving biological signal.

Specificity of the batch effect correction tools

Empirical Bayes methods are state-of-the-art for microarray and bulk RNA-Seq data. → **Apply the methods to scRNA-Seq** Use a formal model and leverage observed data to infer parameters.

Microarray (ComBat, 2007)	RNA-Seq (ComBat-Seq, 2020)	scRNA-Seq (ComBat-sc, 2025)
$y \sim Normal(\mu ; \varphi)$	$y \sim NegBinom(\mu ; \varphi)$	$y \sim \pi\, \delta_0 + (1-\pi)\, NegBinom(\mu ; \varphi)$
ComBat uses a Generalized Linear Model (GLM) to correct batch effects in microarray data.	ComBat-Seq tweaks the distribution of the GLM, and adapts the model to RNA-Seq data.	Our approach further tweaks the distribution to **account for dropouts**, and better fit scRNA-Seq data.

Unique features

- handles **arbitrary experimental designs**, accounts for **covariates** to better preserve the biological signal
- supports selecting a batch as **reference**
- works in **full dimension**, raw count matrices in input and output
  - more **precise**, more **versatile** for downstream analyses: differential expression, signature for deconvolution...