Assessing the impact of batch correction using Scanorama and Harmony on a neuroblastoma cell atlas

S. Sarangi[1], J. Zahiri[1], N. Patikas[2], H. Yao[2], I. Korsunsky[2], C. Bragdon[1], M. Hemberg[2], Z. Pitluk[1]

1 – Paradigm4; 2 – Brigham and Women's Hospital

Background

- Reduced costs of single cell RNA sequencing (scRNA-seq) has enabled generation of 10-100 patient datasets
- Larger population-scale datasets necessary for a biologically-relevant impact on biomarker and drug discovery
- Often these datasets are compiled from multiple experiments with variations in the platform used, technicians, experimental conditions, among others
- These differences result in batch effects that must be corrected and/or understood quantitatively for meaningful analysis
- Numerous batch correction algorithms have been developed
- Here, we compare two popular algorithms – *Scanorama* & *Harmony* – profiling for memory usage, computational complexity, & batch correction efficacy
- We used k-Nearest Neighbor Batch Effect Test (*kBET*) and Average Silhouette Width (*ASW*) to measure efficacy of batch correction
- Data from a Neuroblastoma Cell Atlas (DOI: 10.1126/sciadv.abd3311) used in this study
- Batch correction also applied across neuroblastoma samples collected from different sequencing platforms – CEL-seq2 & Chromium 10x
- Batch correction applied across cells collected from multiple fetal adrenal gland samples (normal tissue) and multiple neuroblastoma samples (tumor tissue)
- Analysis performed in a scalable pipeline using REVEAL, a bioinformatics platform developed by Paradigm4
- Results provide deeper insights into impact of batch correction, and rate limiting steps and limitations to support refinement
- Study also presents a scalable and easily deployable workflow for assessing single-cell algorithms

Methods

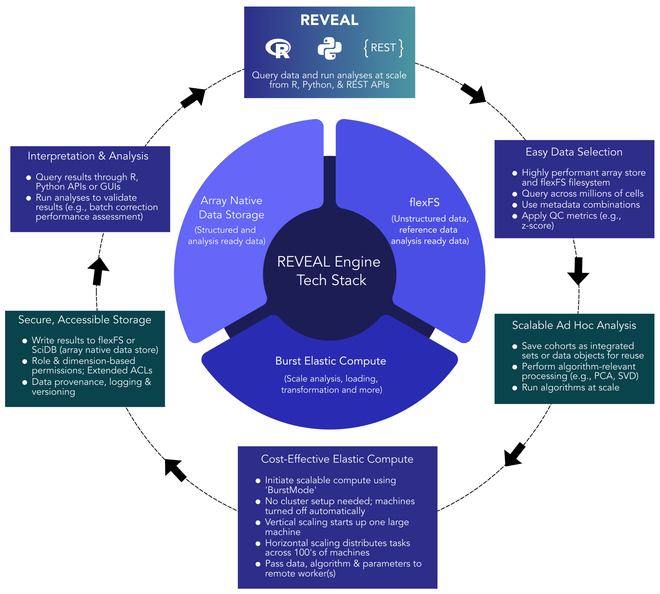

Figure 1: Using REVEAL to run scalable single cell analysis workflows

Batch Correction Algorithms

- Scanorama (DOI: 10.1038/s41587-019-0113-3) enables batch correction & integration of heterogenous scRNA-seq datasets
- Harmony (DOI: 10.1038/s41592-019-0619-0) allows fast, sensitive, & accurate integration of single cell data
- Python-based implementation of *Scanorama* used
- Sparse version of *Harmony* utilized in R
- Dimensionality reduction for *Harmony* performed using *fbpca (PCA)* & *scikit-learn (truncated SVD)* instead of default Seurat PCA for better scalability
- *m5.16xlarge* AWS EC2 instance type used for all analyses

Results

Neuroblastoma Cell Atlas Use Case

Dataset	# of Cells	# of Samples	Seq. Platform
Normal Adrenal Gland	57,972	7 samples from 5 fetuses	10x
Neuroblastoma	6,442	6 samples from 5 patients	10x
Neuroblastoma	13,281	16 patient samples	CEL-seq2

Batch Correction Performance Assessment

- kBET (k-Nearest Neighbor Batch Effect Test) package in R assessed batch effects by comparing local & global batch label distributions using a fixed k-NN matrix (DOI: 10.1038/s41592-018-0254-1)
- ASW (Avg. Silhouette Width) package in Python (*scikit-learn)* measured batch mixing performance across sequencing platforms (10x & CEL-seq2) ($ASW_{platform}$), samples from different individuals in a dataset (ASW_{sample}), & preservation of cell type purity (ASW_{cell})

Measuring Efficacy of Batch Correction Across Seq. Platforms & Samples

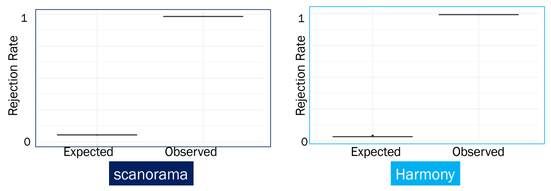

Figure 2: *kBET* rejection rate shows batch effect for both algorithms

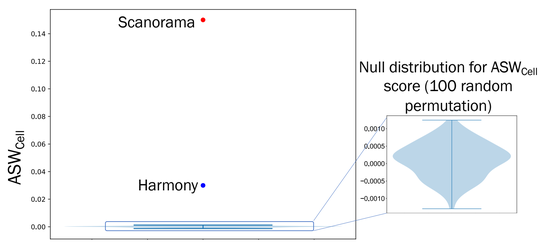

Figure 3: ASW_{cell} values suggest both algorithms preserve cell type purity based on cell annotations from source data

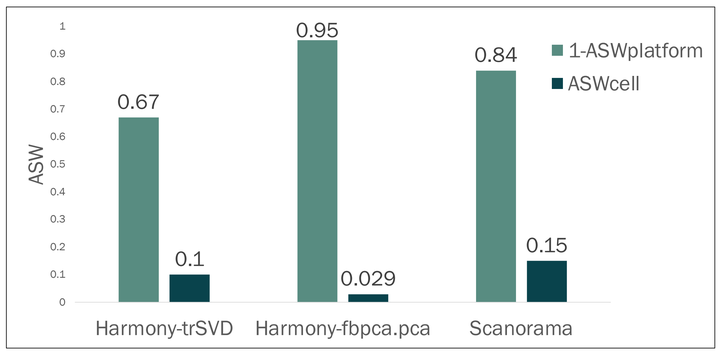

Figure 4: Plot show impact of dim. red. method & algorithm used in removing batch effects between neuroblastoma cells from different seq. platforms

	Dataset	ASW_{cell}	$1\text{-}ASW_{sample}$
Harmony	Neuroblastoma 10x	0.19	0.97
	Neuroblastoma CEL-seq2	0.28	1.18
	Normal Adrenal Gland 10x	0.63	1.14
Scanorama	Neuroblastoma 10x	0.12	0.95
	Neuroblastoma CEL-seq2	0.02	1.06
	Normal Adrenal Gland 10x	0.27	1.14
Uncorrected data	Neuroblastoma 10x	0.18	0.82
	Neuroblastoma CEL-seq2	0.2	0.8
	Normal Adrenal Gland 10x	0.33	1.01

Figure 5: Table shows ASW values for batch effect removal performance between samples of a dataset and preservation of cell type purity

Results

Memory & Compute Time Profiling of Batch Correction Algorithms

Dataset	# of data sets to be integrated	Dataset Size (GB)	# of cells	Density	# of genes	# of common genes
1	10	7.6	100K	10%	32,738	1000
2	10	16	200K	10%	32,738	1000
3	10	39	500K	10%	32,738	1000
4	10	176	1 Million	10%	32,738	1000
5	10	300	2 Million	10%	32,738	1000

Figure 6: Table shows the design of experiment (DOE) space using synthetic data generated by sampling *Human Cell Atlas* datasets

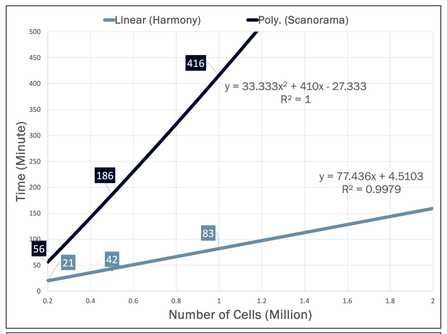

Figure 7: Plot shows relationship between computation time for batch correction and number of cells using *Scanorama* & *Harmony*

Scanorama run failed on 2M cells during merging of panoramas on the test EC2 instance

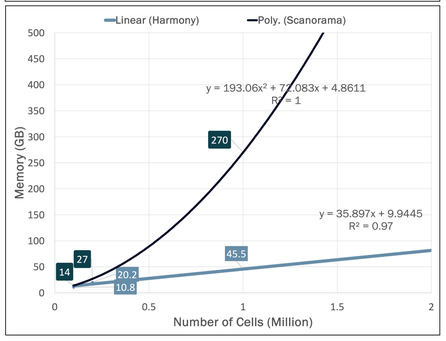

Figure 8: Plot shows relationship between memory usage during batch correction and number of cells using *Scanorama* & *Harmony*

Note: Running *Scanorama* on 2M cells (or more) could have been achieved by launching a larger EC2 machine through 'Burst'

Conclusions

- Sensitive *kBET* analysis suggests batch bias exists after batch correction across platforms & samples in a dataset (not shown) using both algorithms
- *ASW* values show both algorithms preserve cell type information after batch correction across seq. platforms with *Scanorama* having a better score
- Using a truncated SVD instead of the *fbpca* PCA improves *Harmony's* performance at maintaining cell type information, but reduces the batch mixing efficacy when correcting across seq. platforms
- Results across samples within a normal or neuroblastoma dataset suggest *Harmony* doing better at maintaining cell type information, & slightly better performance at batch mixing
- Memory & computation time profiling results show *Harmony* significantly outperforming *Scanorama* for the datasets tested
- REVEAL allows bioinformaticians & computational biologists to run compute intensive algorithms at scale with minimal time spent on setup
- Ad hoc hypothesis testing with secure access to 100's of datasets, easy cohort selection, algorithms of choice, & cost-effective elastic compute
- Assessment & enablement of scalability is critical for drug & biomarker discovery using single cell data – e.g., 2 million cells ~ 200 patients
- Study provides quantitative metrics to assess impact of batch correction in heterogenous scRNA-seq datasets
- Further work will include testing on more datasets & investigating the effect of different dimensionality red. methods on batch correction

paradigm4