



# Introduction to integration of scRNA-seq data

*The Hitchhiker's Guide to scRNA-seq, iMM*

09/07/2024

António Sousa - [aggode@utu.fi](mailto:aggode@utu.fi)

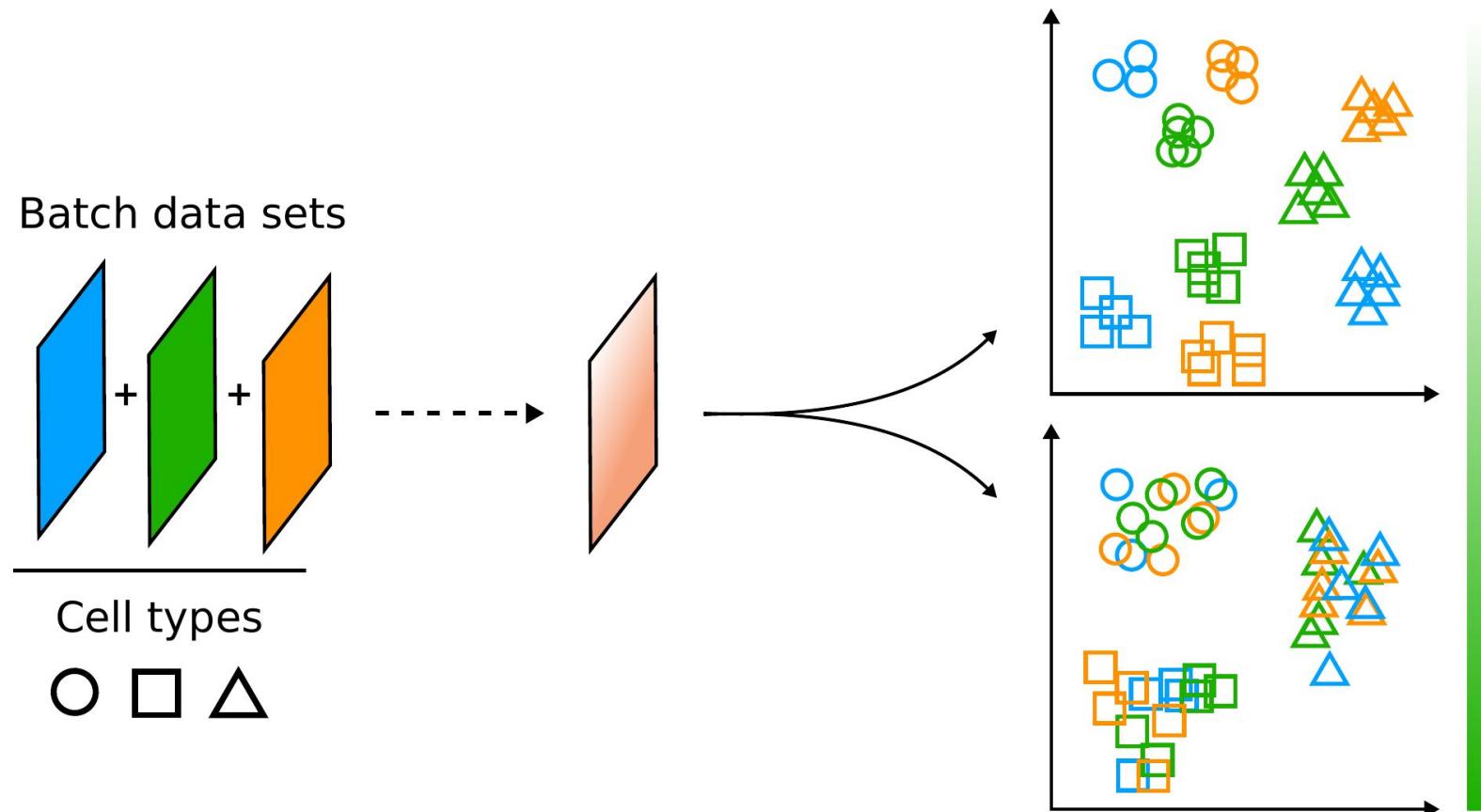


# Outline

- Integration: methods & benchmarks
- Strengths & weaknesses of the different methods
- Strategies to assess integration accuracy
- Integration: *de novo* versus reference-based

# What is integration?

- Identification of shared cell types/states across heterogeneous scRNA-seq datasets



# How do we know if integration is required or not?

- Three main approaches (can be combined):
  - Dimensional reduction techniques
  - Automatic cell annotation
  - Independent sample analysis: clustering → cluster markers → annotation
  - (cluster comparison between samples)

# Overview of integration strategies

REVIEW ARTICLE

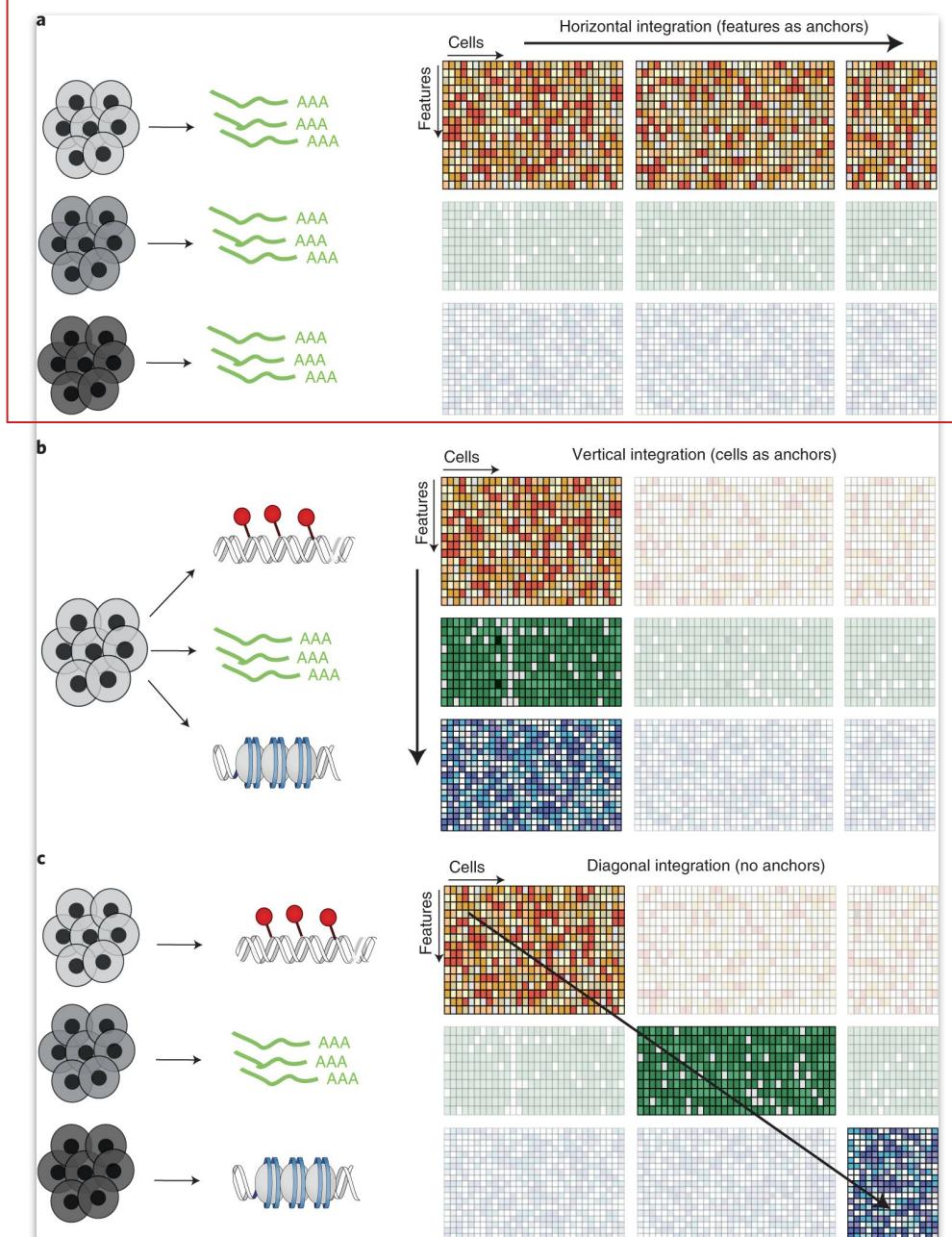
<https://doi.org/10.1038/s41587-021-00895-7>

nature  
biotechnology

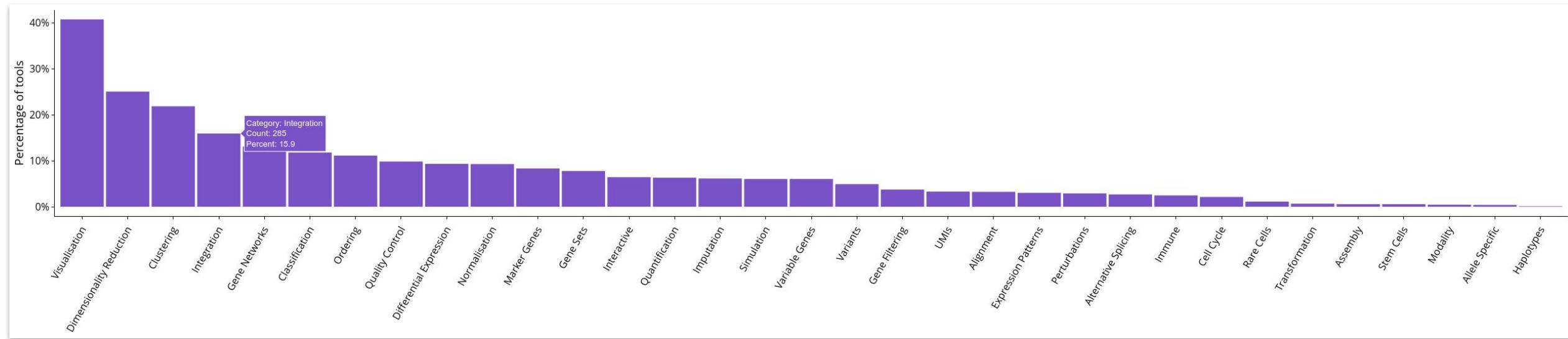
 Check for updates

## Computational principles and challenges in single-cell data integration

Ricard Argelaguet  <sup>1,2</sup>  , Anna S. E. Cuomo  <sup>1,3</sup>  , Oliver Stegle  <sup>3,4,5</sup>  and John C. Marioni  <sup>1,3,6</sup> 



# Overview of integration methods: [www.scrna-tools.org](http://www.scrna-tools.org) (20/06/24)



285 methods/tools!

# Overview of integration methods: [www.scrna-tools.org](http://www.scrna-tools.org) (20/06/24)

**Tools**  
View entries for individual tools

**Sorting & Filters**

**Sort by**  
Citations

**Filter by category**  
Select multiple categories and click FILTER below

- Cell Cycle
- Classification
- Clustering
- Differential Expression
- Dimensionality Reduction
- Gene Filtering
- Gene Networks
- Gene Sets
- Haplotypes
- Immune
- Imputation
- Integration**

**FILTER**   **RESET**

1. Seurat	CRAN 5.1.0	downloads 65k/month
2. Harmony	CRAN 1.2.0	downloads 6749/month
3. scran	in BioC 8 years	rank 87 / 2300
4. scvi-tools	pypi package 1.1.2	downloads/month 2k
5. batchelor	In BioC 5 years	rank 99 / 2300
6. RacelID	CRAN 0.3.5	downloads 447/month
7. MIMOSCA		
8. bseqsc		
9. MOFA	pypi package 1.2	downloads/month 99
10. LIGER	CRAN 2.0.1	downloads 1002/month
11. scanorama	pypi package 1.7.4	downloads/month 9k
12. scMC		
13. bbknn	pypi package 1.6.0	downloads/month 8k

- Shared nearest neighbors in the low dimensional space
- Dimensional reductions: Canonical Correlation Analysis (CCA) or Reciprocal PCA (RPCA)
- Integration or reference-mapping/transfer learning

## Comprehensive Integration of Single-Cell Data

Tim Stuart,<sup>1,4</sup> Andrew Butler,<sup>1,2,4</sup> Paul Hoffman,<sup>1</sup> Christoph Hafemeister,<sup>1</sup> Efthymia Papalexis,<sup>1,2</sup> William M. Mauck III,<sup>1,2</sup> Yuhan Hao,<sup>1,2</sup> Marlon Stoeckius,<sup>3</sup> Peter Smibert,<sup>3</sup> and Rahul Satija<sup>1,2,5,\*</sup>

<sup>1</sup>New York Genome Center, New York, NY, USA

<sup>2</sup>Center for Genomics and Systems Biology, New York University, New York, NY, USA

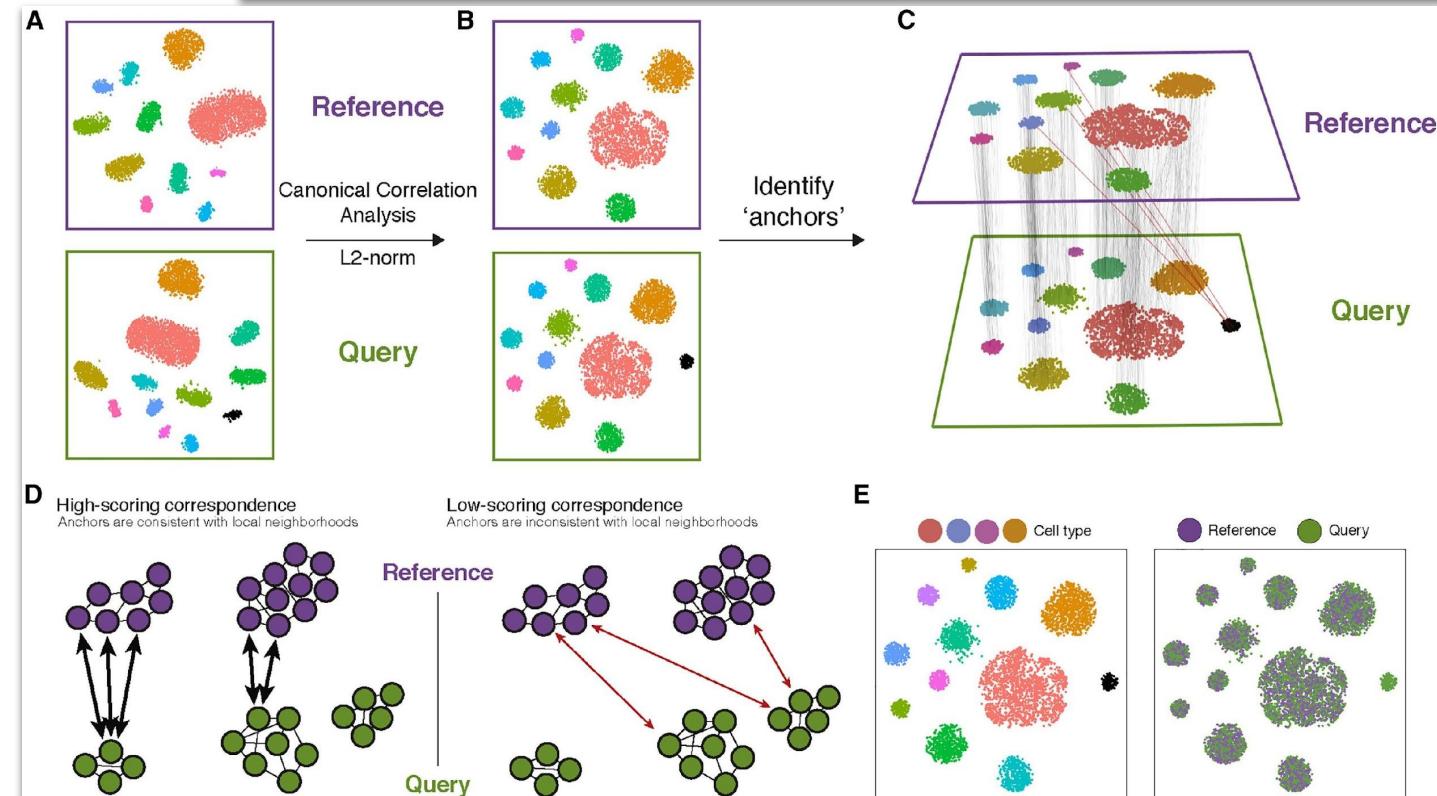
<sup>3</sup>Technology Innovation Lab, New York Genome Center, New York, NY, USA

<sup>4</sup>These authors contributed equally

<sup>5</sup>Lead Contact

\*Correspondence: rsatija@nygenome.org

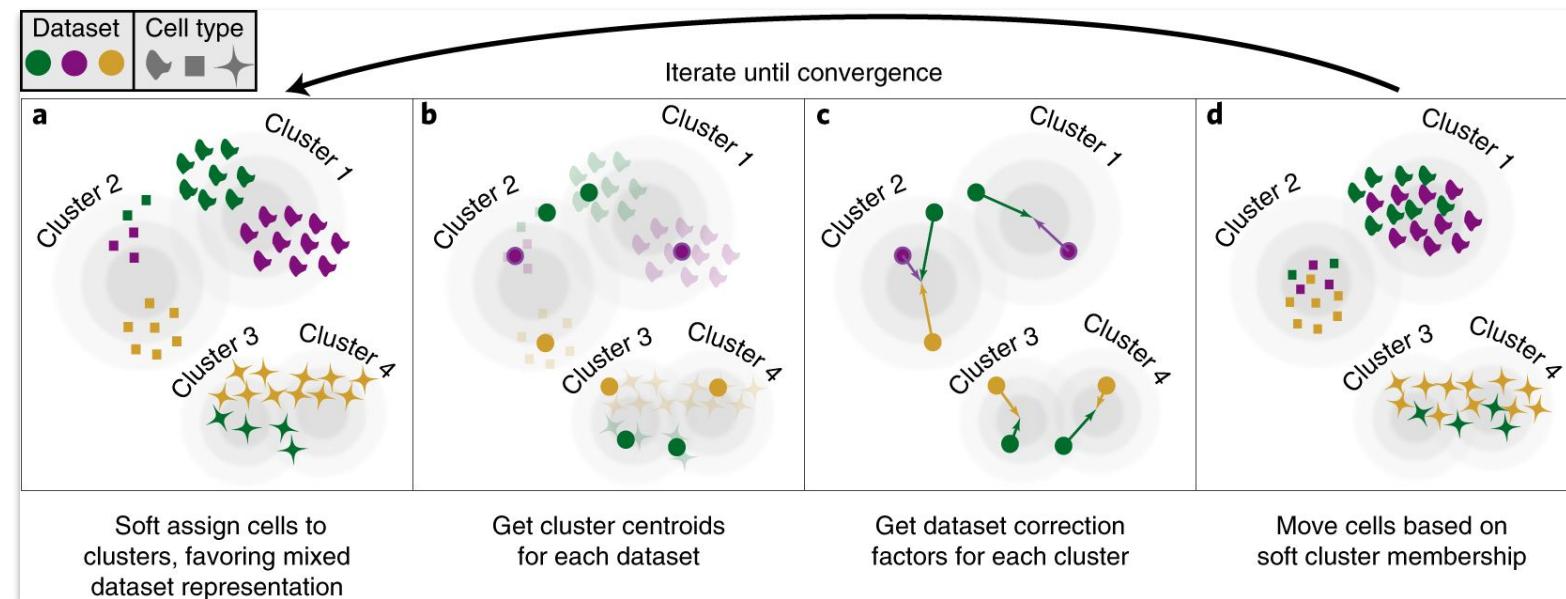
<https://doi.org/10.1016/j.cell.2019.05.031>



<https://satijalab.org/seurat>

## Fast, sensitive and accurate integration of single-cell data with Harmony

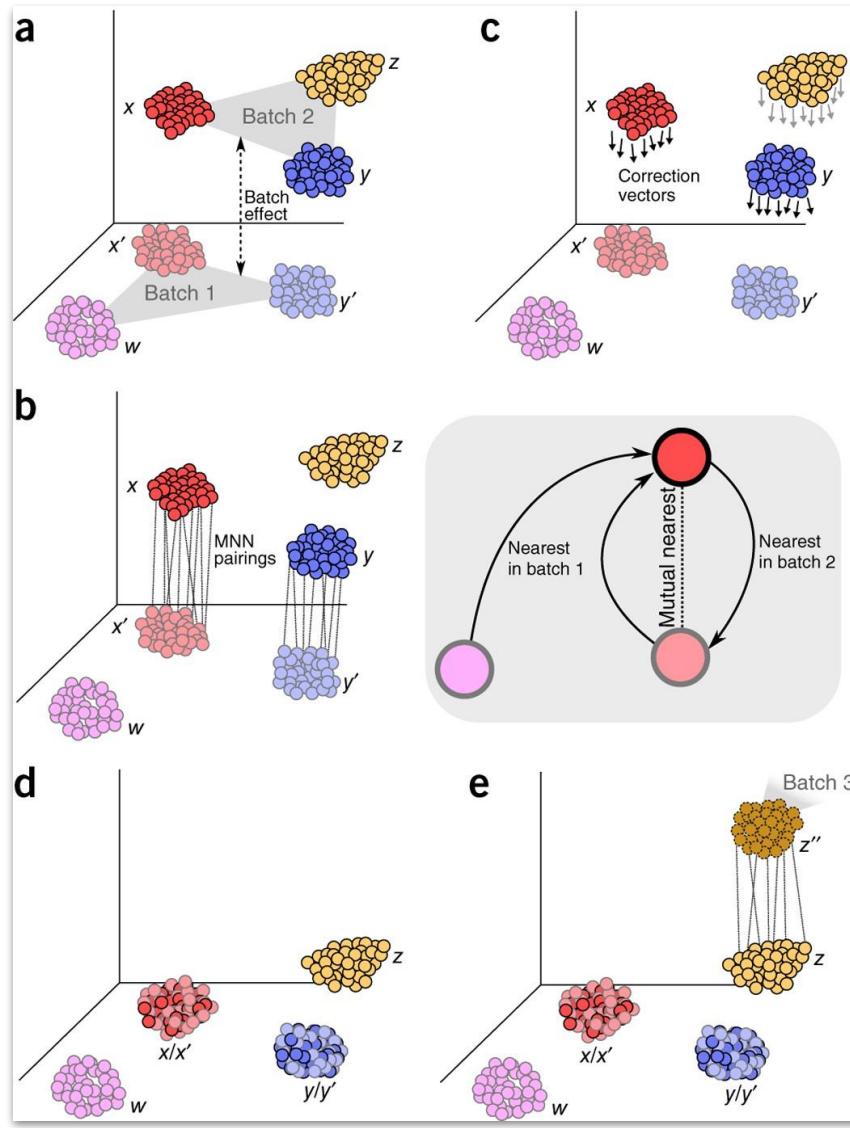
Ilya Korsunsky<sup>1,2,3,4</sup>, Nghia Millard<sup>1,2,3,4</sup>, Jean Fan<sup>1,5</sup>, Kamil Slowikowski<sup>1,2,3,4</sup>,  
 Fan Zhang<sup>1,2,3,4</sup>, Kevin Wei<sup>2</sup>, Yuriy Baglaenko<sup>1,2,3,4</sup>, Michael Brenner<sup>2</sup>, Po-ru Loh<sup>1,3,4</sup> and  
 Soumya Raychaudhuri<sup>1,2,3,4,6\*</sup>



- Approximates dataset-specific cluster centroids to global centroids in the PCA space

<https://portals.broadinstitute.org/harmony>

# fastMNN



Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors

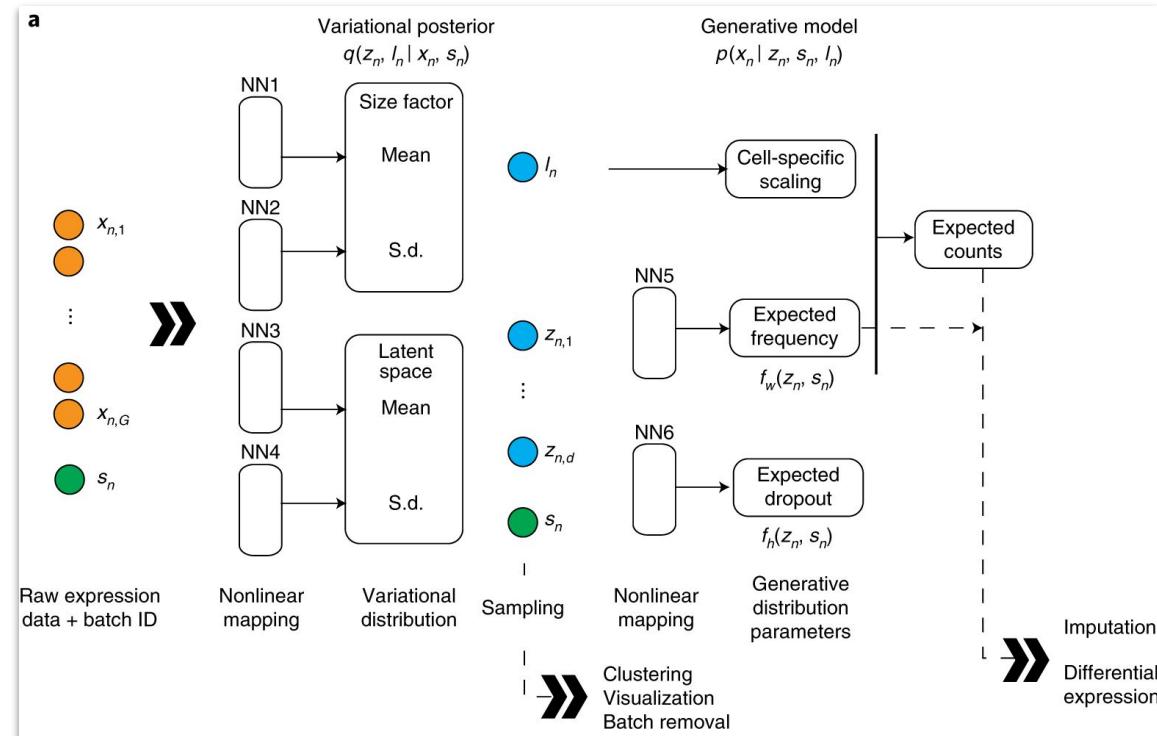
Laleh Haghverdi<sup>1,2</sup>, Aaron T L Lun<sup>3</sup> , Michael D Morgan<sup>4</sup> & John C Marioni<sup>1,3,4</sup>

- *Correction vectors applied between mutual nearest neighbors pairs*

<https://bioconductor.org/packages/devel/bioc/html/batchelor.html>

## Deep generative modeling for single-cell transcriptomics

Romain Lopez<sup>1</sup>, Jeffrey Regier<sup>1</sup>, Michael B. Cole<sup>2</sup>, Michael I. Jordan<sup>1,3</sup> and Nir Yosef<sup>1,4,5\*</sup>



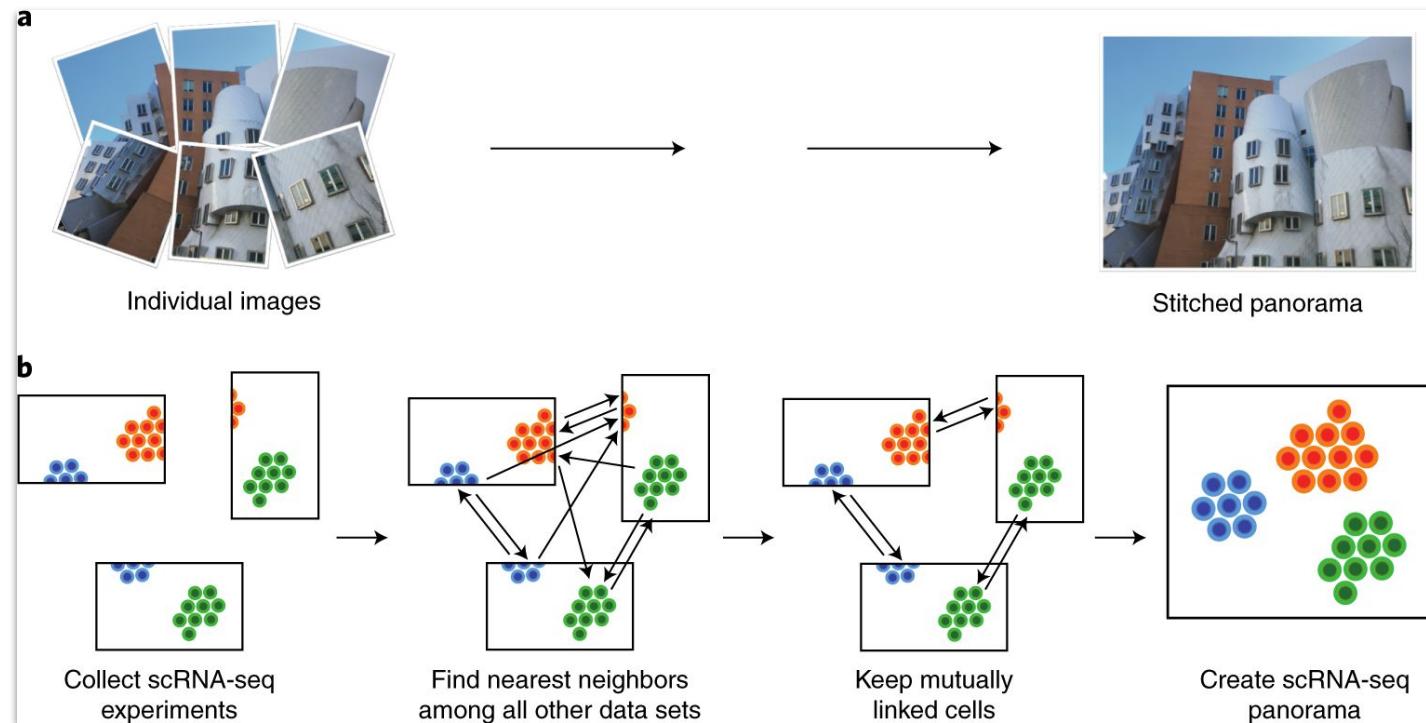
- “stochastic optimization and deep neural networks to aggregate information across similar cells and genes and to approximate the distributions that underlie observed expression values, while accounting for batch effects”

<https://scvi-tools.org>

# Scanorama

## Efficient integration of heterogeneous single-cell transcriptomes using Scanorama

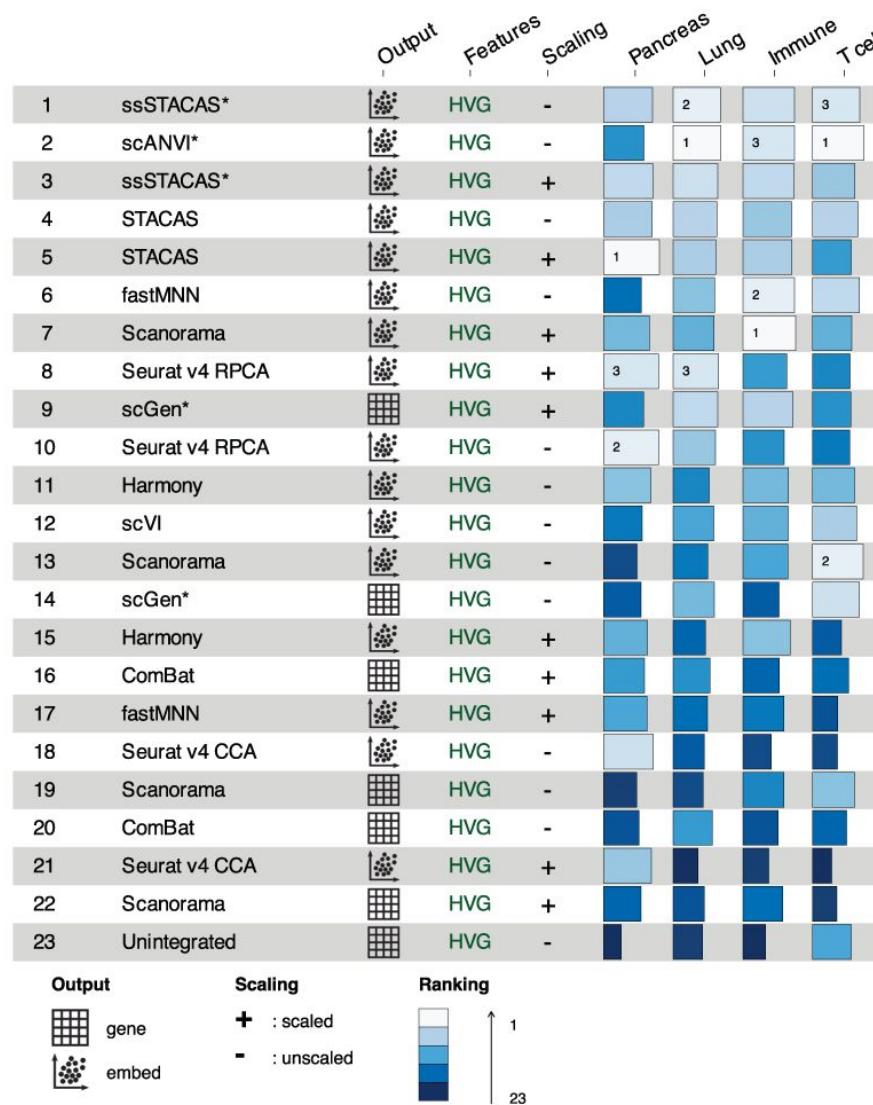
Brian Hie<sup>1</sup>, Bryan Bryson<sup>2\*</sup> and Bonnie Berger<sup>1,3\*</sup>



- “Mutually linked cells form matches that can be leveraged to correct for batch effects and merge experiments together; whereby the datasets forming connected components on the basis of these matches become a scRNA-seq ‘panorama’”

<https://github.com/brianhie/scanorama>

# Semi-supervised integration



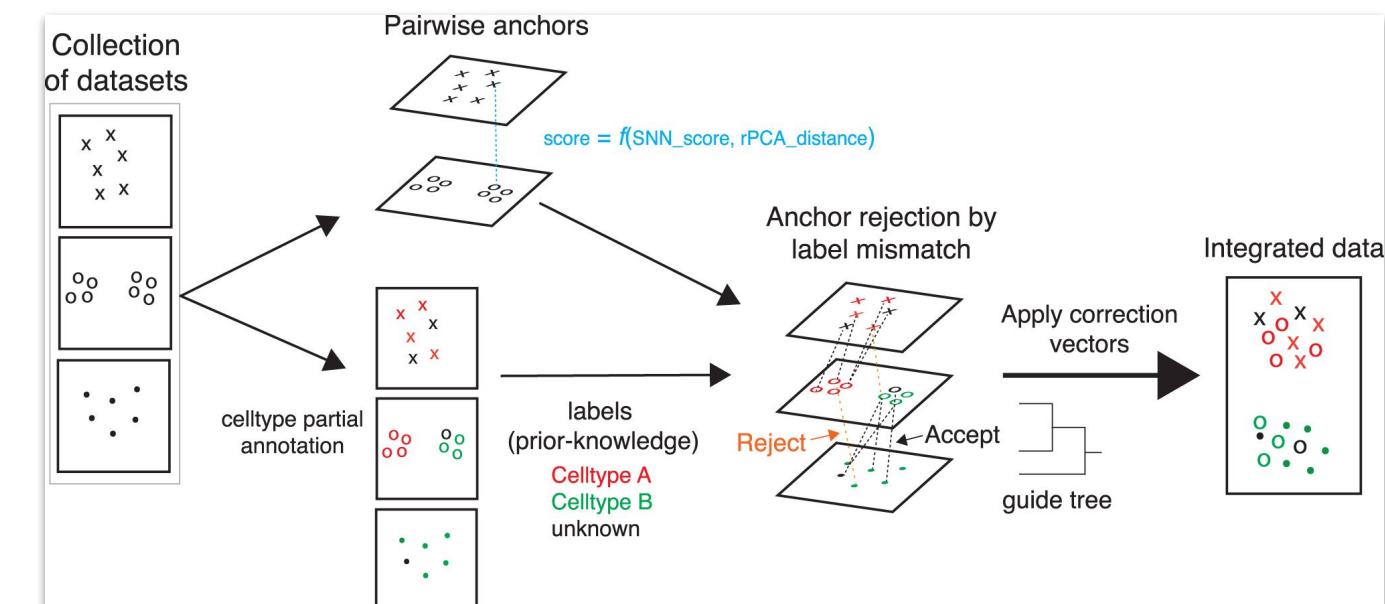
## Semi-supervised integration of single-cell transcriptomics data

Received: 25 July 2023

Massimo Andreatta 1,2,3, Léonard Héault 1,2,3, Paul Gueguen 1,2,3,

Accepted: 16 January 2024

David Gfeller 1,2,3, Ariel J. Berenstein 4 &amp; Santiago J. Carmona 1,2,3



<https://github.com/carmonalab/STACAS>

# Integration scRNA-seq benchmarks

Tran et al. *Genome Biology* (2020) 21:12  
https://doi.org/10.1186/s13059-019-1850-9

Genome Biology

RESEARCH

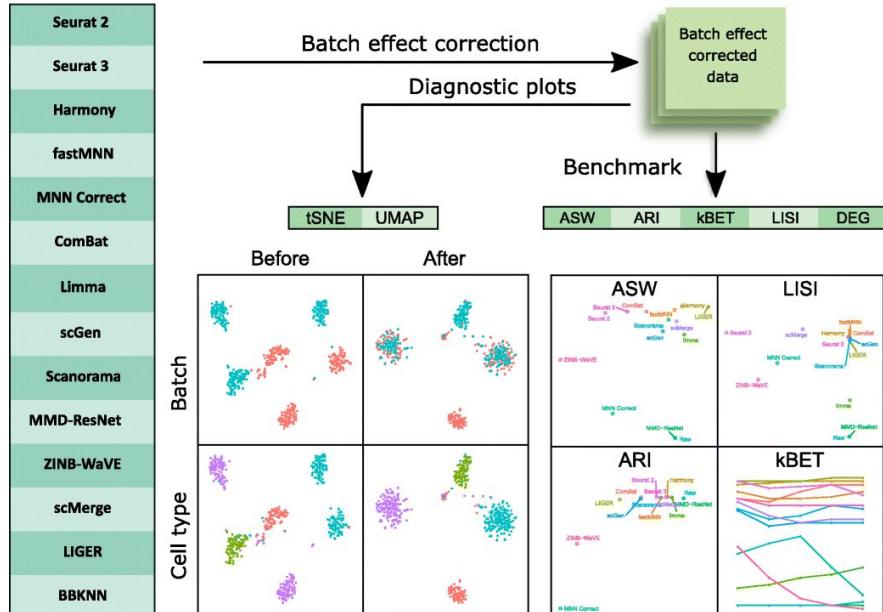
Open Access



## A benchmark of batch-effect correction methods for single-cell RNA sequencing data

Hoa Thi Nhu Tran<sup>†</sup>, Kok Siong Ang<sup>†</sup>, Marion Chevrier<sup>†</sup>, Xiaomeng Zhang<sup>†</sup>, Nicole Yee Shin Lee, Michelle Goh and Jinmiao Chen<sup>\*</sup>

A



## Abstract

**Background:** Large-scale single-cell transcriptomic datasets generated using different technologies contain batch-specific systematic variations that present a challenge to batch-effect removal and data integration. With continued growth expected in scRNA-seq data, achieving effective batch integration with available computational resources is crucial. Here, we perform an in-depth benchmark study on available batch correction methods to determine the most suitable method for batch-effect removal.

**Results:** We compare 14 methods in terms of computational runtime, the ability to handle large datasets, and batch-effect correction efficacy while preserving cell type purity. Five scenarios are designed for the study: identical cell types with different technologies, non-identical cell types, multiple batches, big data, and simulated data. Performance is evaluated using four benchmarking metrics including kBET, LISI, ASW, and ARI. We also investigate the use of batch-corrected data to study differential gene expression.

**Conclusion:** Based on our results, Harmony, LIGER, and Seurat 3 are the recommended methods for batch integration. Due to its significantly shorter runtime, Harmony is recommended as the first method to try, with the other methods as viable alternatives.

**Keywords:** Single-cell RNA-seq, Batch correction, Batch effect, Integration, Differential gene expression

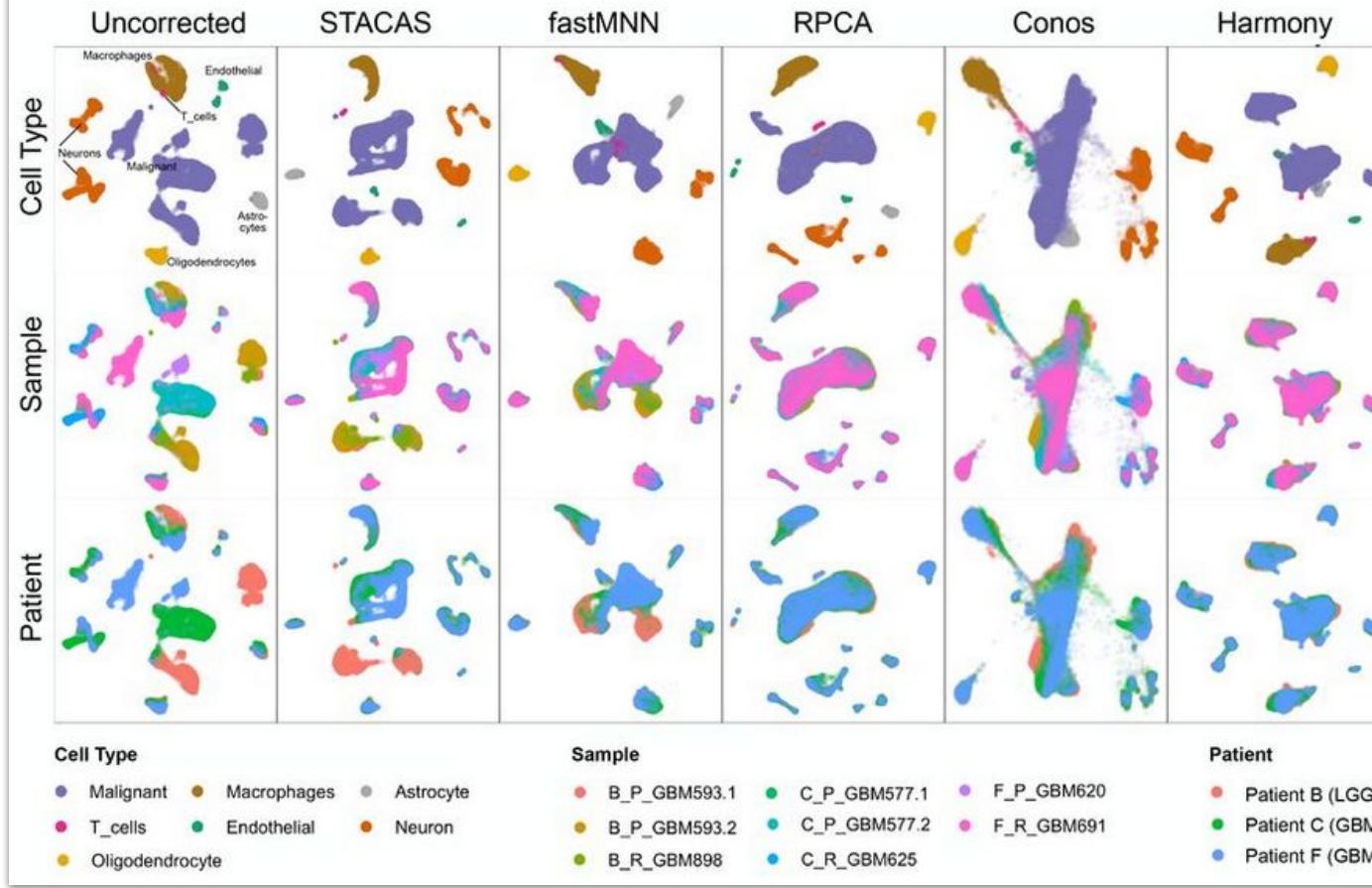
B

Dataset	Description	Number of batches	Total cell number	Technologies
1	Human Dendritic Cells	2	576	Smart-Seq2
2	Mouse Cell Atlas	2	6,954	Microwell-Seq Smart-Seq2
3	Simulation			
4	Human Pancreas	5	14,767	inDrop CEL-Seq2 Smart-Seq2 SMARTer SMARTer
5	Human Peripheral Blood Mononuclear Cell	2	15,476	10x 3' 10x 5'
6	Cell line	3	9,530	10x
7	Mouse Retina	2	71,638	Drop-seq
8	Mouse Brain	2	833,206	Drop-seq SPLiT-seq
9	Human Cell Atlas	2	621,466	10x
10	Mouse Haematopoietic Stem and Progenitor Cells	2	4,649	MARS-seq Smart-Seq2

# Integration scRNA-seq benchmarks

A

**Richards-GBM-LGG**  
(35,549 nuclei from 8 samples and 3 patients)



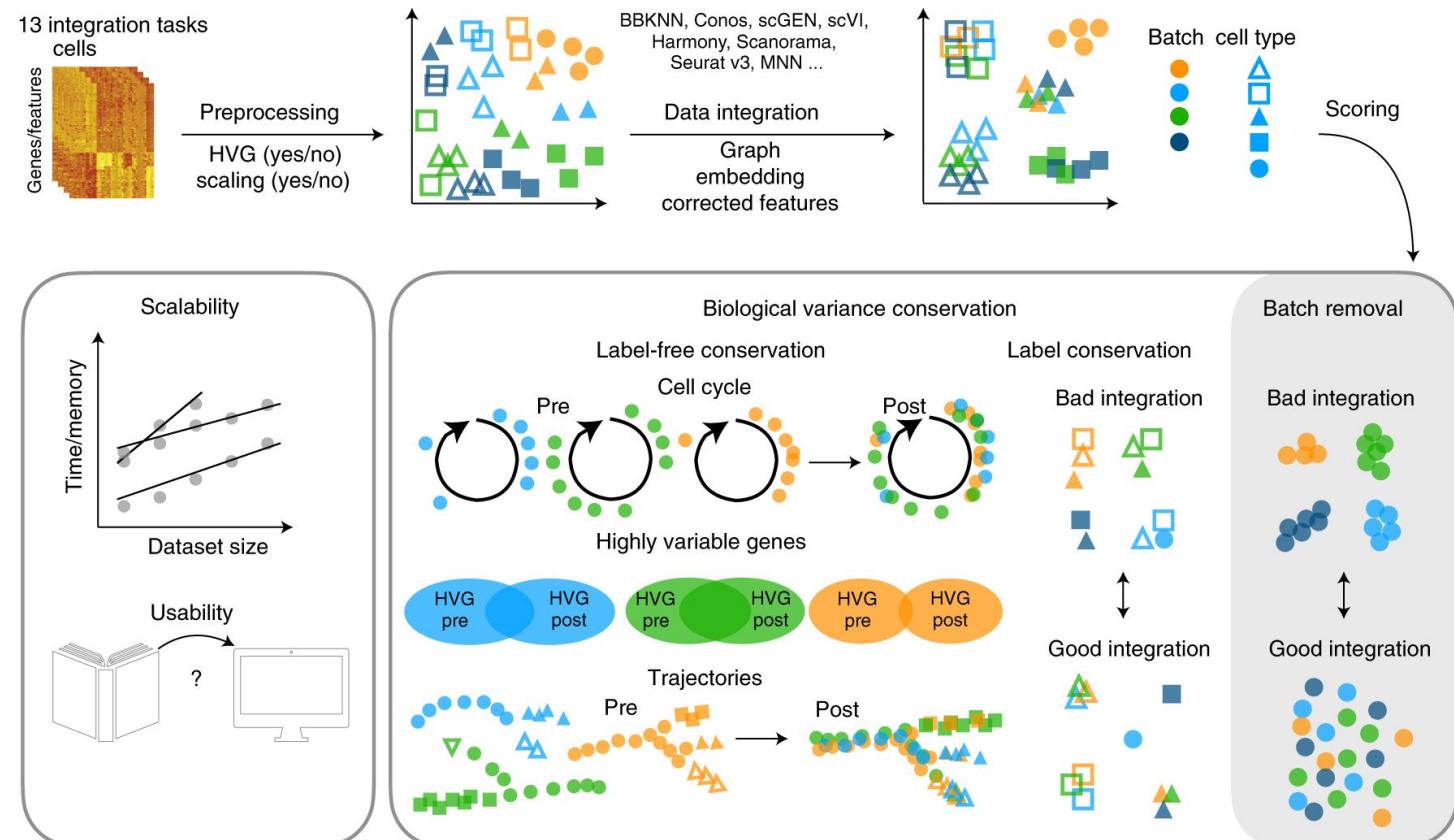
## A comparison of data integration methods for single-cell RNA sequencing of cancer samples

Laura M. Richards<sup>1,2</sup>, Mazdak Riverin<sup>2</sup>, Suluxan Mohanraj<sup>2</sup>, Shamini Ayyadhyury<sup>2,3</sup>, Danielle C. Croucher<sup>1,2</sup>, J. Javier Díaz-Mejía<sup>2</sup>, Fiona J. Coutinho<sup>5</sup>, Peter B. Dirks<sup>4,5,6</sup>, Trevor J. Pugh<sup>1,2,7,8</sup>

Tumours are routinely profiled with single-cell RNA sequencing (scRNA-seq) to characterize their diverse cellular ecosystems of malignant, immune, and stromal cell types. When combining data from multiple samples or studies, batch-specific technical variation can confound biological signals. However, scRNA-seq batch integration methods are often not designed for, or benchmarked, on datasets containing cancer cells. Here, we compare 5 data integration tools applied to 171,206 cells from 5 tumour scRNA-seq datasets. Based on our results, STACAS and fastMNN are the most suitable methods for integrating tumour datasets, demonstrating robust batch effect correction while preserving relevant biological variability in the malignant compartment. This comparison provides a framework for evaluating how well single-cell integration methods correct for technical variability while preserving biological heterogeneity of malignant and non-malignant cell populations.



# Integration scRNA-seq benchmarks



OPEN

## Benchmarking atlas-level data integration in single-cell genomics

Malte D. Luecken <sup>1</sup>, M. Büttner <sup>1</sup>, K. Chaichoompu <sup>1</sup>, A. Danese<sup>1</sup>, M. Interlandi<sup>2</sup>, M. F. Mueller<sup>1</sup>, D. C. Strobl<sup>1</sup>, L. Zappia<sup>1,3</sup>, M. Dugas<sup>4</sup>, M. Colomé-Tatché<sup>1,5,6</sup> and Fabian J. Theis <sup>1,3,5</sup>

Method	RNA	Simulations			Usability			Scalability				
		Pancreas	Lung	Immune (human)	Immune (mouse)	Mouse brain	Sim 1	Sim 2	Package	Paper	Time	Memory
1 scANVI*	HVG	-	2	3	1	1	2	2	3	1	3	3
2 Scanorama	HVG	+	3	1	2	2	2	2	2	2	1	1
3 scVI	HVG	-	3	3	3	3	3	3	3	3	3	1
4 fastMNN	HVG	-	2	2	2	2	2	2	2	2	2	2
5 scGen*	HVG	-	3	1	1	1	1	1	1	1	1	1
6 Harmony	HVG	-	1	1	1	1	1	1	1	1	1	1
7 fastMNN	HVG	-	2	2	2	2	2	2	2	2	2	2
8 Seurat v3 RPCA	HVG	+	2	2	2	2	2	2	2	2	2	2
9 BBKNN	HVG	-	3	3	3	3	3	3	3	3	3	2
10 Scanorama	HVG	+	2	2	2	2	2	2	2	2	2	1
11 ComBat	HVG	-	3	3	3	3	3	3	3	3	3	1
12 MNN	HVG	+	2	2	2	2	2	2	2	2	2	1
13 Seurat v3 CCA	HVG	-	2	2	2	2	2	2	2	2	2	1
14 trVAE	HVG	-	2	2	2	2	2	2	2	2	2	1
15 Conos	HVG	-	2	2	2	2	2	2	2	2	2	1
16 DESC	FULL	-	2	2	2	2	2	2	2	2	2	1
17 LIGER	HVG	-	2	2	2	2	2	2	2	2	2	1
18 SAUCIE	HVG	+	2	2	2	2	2	2	2	2	2	1
19 Unintegrated	FULL	-	2	2	2	2	2	2	2	2	2	1
20 SAUCIE	HVG	+	2	2	2	2	2	2	2	2	2	1

**Rank**

**Name**

**Output**

- Genes
- Embedding
- Graph

**Scaling**

- + Scaled
- Unscaled

**Ranking**

- 1
- 20

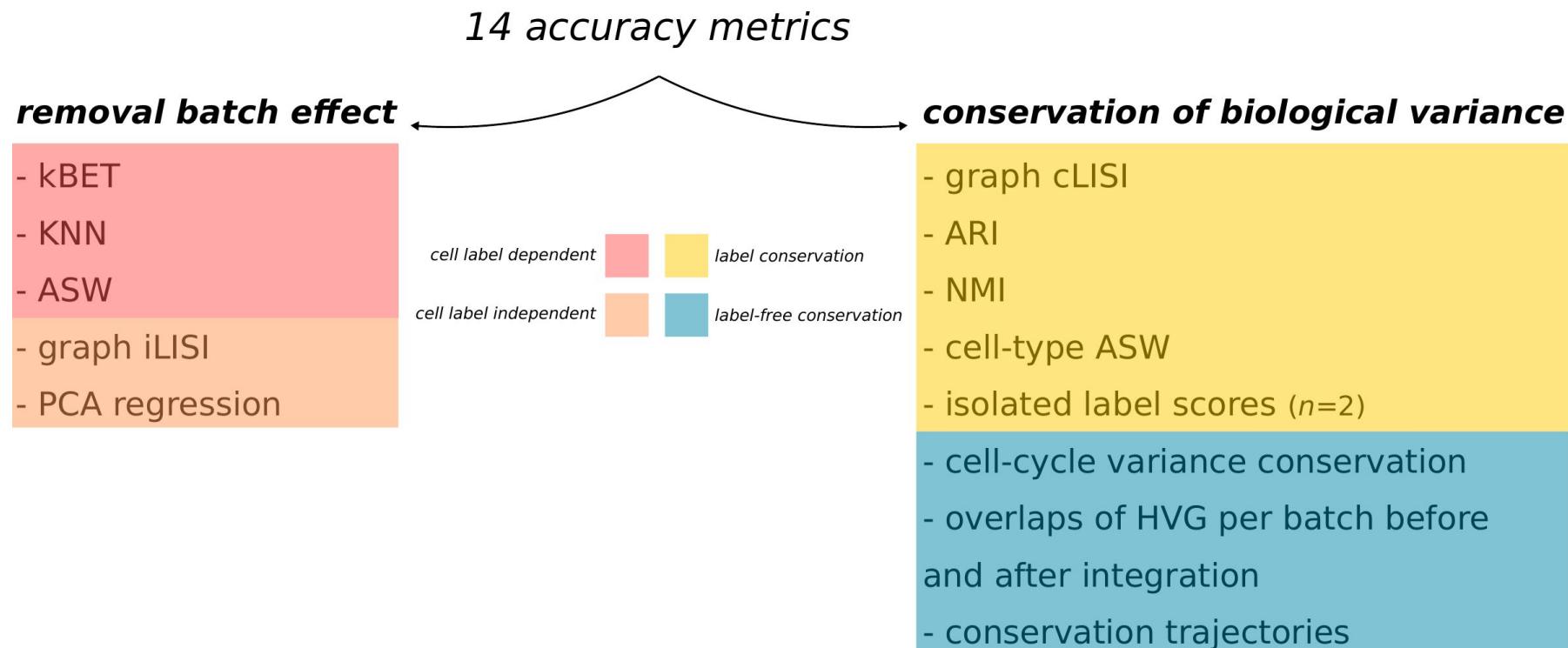
# Conclusions: strengths & weaknesses of the different methods

- Integration: trade-off between biological signal preservation and batch correction
- Inconsistent results across benchmarks: different methods, integration tasks, performance metrics
- Benchmarks: showed that performance of integration is inherently dependent on the complexity of the integration task, highlighting that there is no single method fitting all the tasks
- Methods more suitable to correct strong batch effects (prioritize batch correction): Seurat CCA, Harmony and fastMNN
- Methods more suitable to correct mild to moderate batch effects (prioritize biological conservation): Seurat RPCA, Scanorama and scVI

# Strategies to assess integration accuracy

- Two types of assessment:
  - qualitative (dimensional reduction visualizations)
  - quantitative (objective metrics)

# Measuring integration performance (quantitative): scib package



Input: *integrated joint embedding or feature corr. matrix*

Paper: <https://www.nature.com/articles/s41592-021-01336-8>

scIB github repo: <https://github.com/theislab/scib>

scib-pipeline github repo: <https://github.com/theislab/scib-pipeline>

# Measuring integration performance

## Research Article



## CellMixS: quantifying and visualizing batch effects in single-cell RNA-seq data

Almut Lütge<sup>1,2</sup>, Joanna Zyprych-Walczak<sup>3</sup>, Urszula Bryczynska Kunzmann<sup>4</sup>, Helena L Crowell<sup>1,2</sup>, Daniela Calini<sup>5</sup>, Dheeraj Malhotra<sup>5</sup>, Charlotte Soneson<sup>2,4</sup>, Mark D Robinson<sup>1,2</sup>

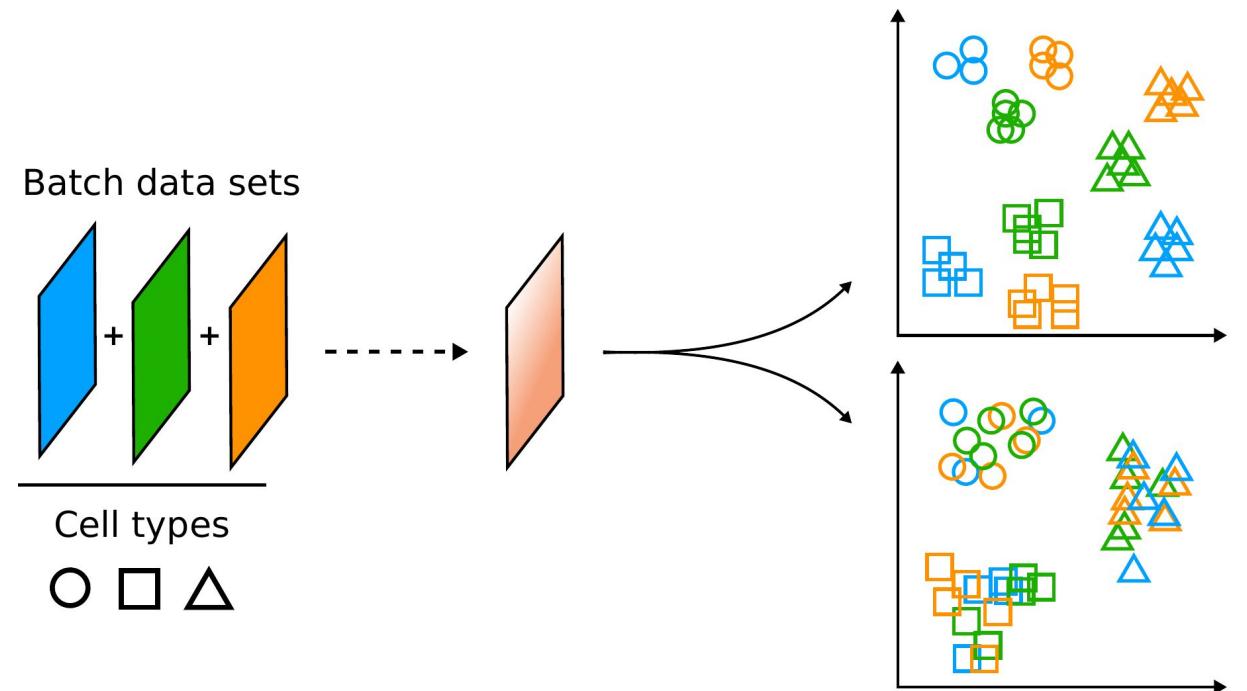
**Table 1.** Batch mixing metrics: short summaries of metrics included in the benchmark.

Metric	Level	Basis	Short description	Interpretation
Cell-specific Mixing Score (cms)	Cell	knn, pca	Test for whether distance distributions from a neighbourhood are batch specific	P-value: Probability to observe as large differences in distance distributions assuming the same underlying distribution
Local Inverse Simpson Index (lisi)	Cell	knn	Inverse of the sum of batch probabilities within weighted knn	Effective number of batches in neighbourhood
Entropy	Cell	knn	Sum of the products of the batch probabilities and their log within each cell's knn	Randomness in the data according to the batch variable
Mixing metric (mm)	Cell	knn	Median position of the fifth cell from each batch within its knn	Number of cells within knn until each batch is represented by five cells
Graph connectivity (graph)	Cell type	knn-graph	Fraction of directly connected cells within cell type graphs	Proportion of non-distorted cell type relationships
k-nearest neighbour Batch effect test (kBET)	Cell type	knn	Test for equal batch proportions within a random cell's knn	P-value: Probability to observe as large differences in batch proportions assuming the same underlying proportions
Average silhouette width (asw)	Cell type	pca	Average relationship of within and between batch-cluster distances for each cell type	Indication of how well clusters are separated
Principal component regression (pcr)	Global	pca	Correlation of the batch variable with principal components weighted by their variance attributes	Proportion of variance attributed to batch

<https://github.com/almutlue/CellMixS>

# Integration (*de novo*) versus reference-mapping

- Integration (*de novo*): integration → clustering → cluster markers → annotation

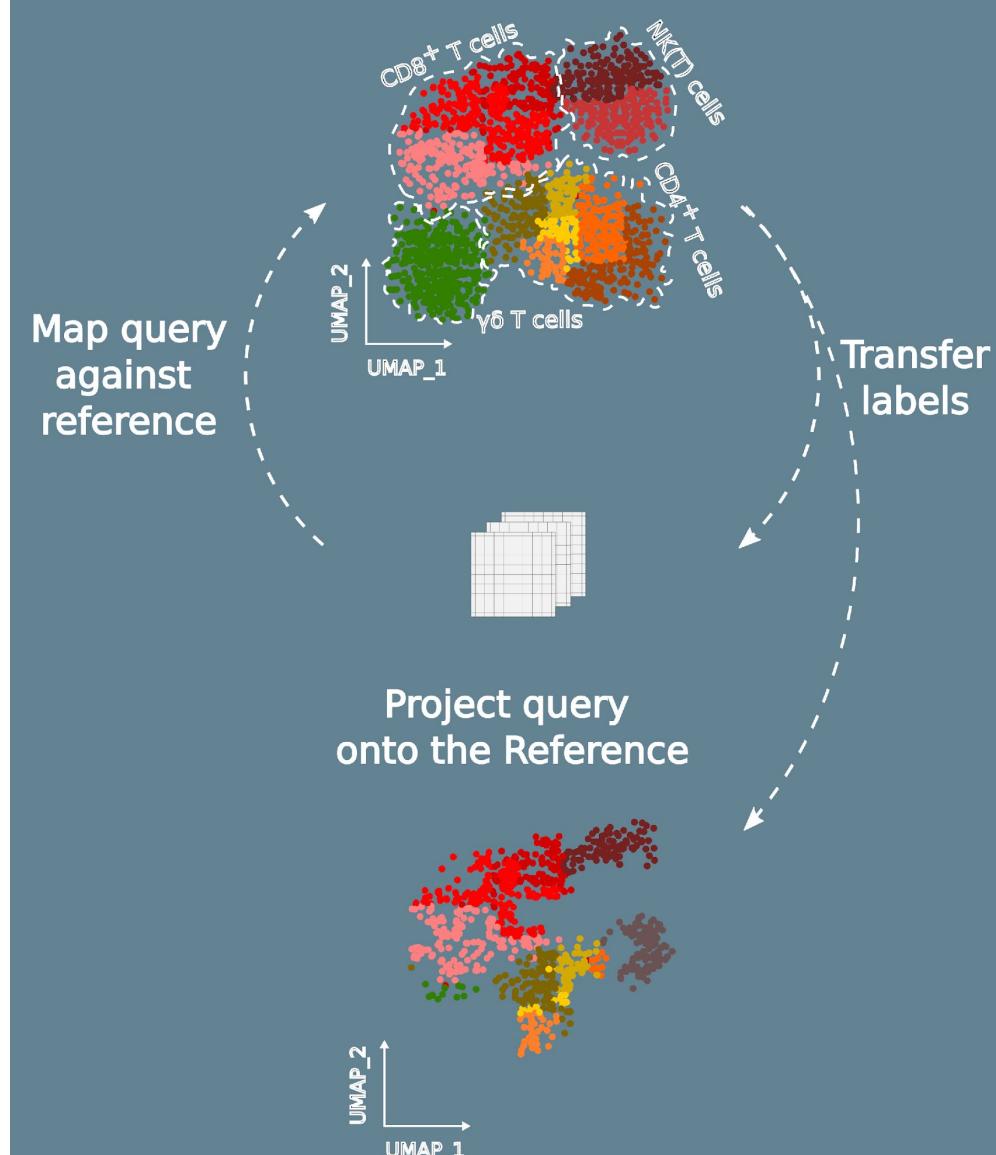


- Application: newly poorly described single-cell transcriptomic data

# Integration (*de novo*) versus reference-mapping

- **Reference-mapping:** projecting & transferring labels from a previously annotated reference to a new query data set
- **Application:** experiments generating single-cell transcriptomic data with well-known identities for which are references available

## Reference Mapping



# Reference-mapping

**Cell Resource**

## Comprehensive Integration of Single-Cell Data

Tim Stuart,<sup>1,4</sup> Andrew Butler,<sup>1,2,4</sup> Paul Hoffman,<sup>1</sup> Christoph Hafemeister,<sup>1</sup> Eftymia Papalexi,<sup>1,2</sup> William M. Mauck III,<sup>1,2</sup> Yuhan Hao,<sup>1,2</sup> Marlon Stoeckius,<sup>3</sup> Peter Smibert,<sup>3</sup> and Rahul Satija<sup>1,2,5,\*</sup>

<sup>1</sup>New York Genome Center, New York, NY, USA  
<sup>2</sup>Center for Genomics and Systems Biology, New York University, New York, NY, USA  
<sup>3</sup>Technology Innovation Lab, New York Genome Center, New York, NY, USA  
<sup>4</sup>These authors contributed equally  
<sup>5</sup>Lead Contact  
<sup>\*</sup>Correspondence: rsatija@nygenome.org  
<https://doi.org/10.1016/j.cell.2019.05.031>

**nature biotechnology** ANALYSIS  
<https://doi.org/10.1038/s41587-021-01001-7>

**OPEN**  
**Mapping single-cell data to reference atlases by transfer learning**

Mohammad Lotfollahi<sup>1,2</sup>, Mohsen Naghipourfar<sup>2</sup>, Malte D. Luecken<sup>2</sup>, Matin Khajavi<sup>1</sup>, Maren Büttner<sup>1</sup>, Marco Wagenstetter<sup>1</sup>, Žiga Avsec<sup>3</sup>, Adam Gayoso<sup>2</sup>, Nir Yosef<sup>2,4,5,6,7</sup>, Marta Interlandi<sup>8</sup>, Sergei Rybakov<sup>1,9</sup>, Alexander V. Misharin<sup>2,10</sup> and Fabian J. Theis<sup>1,2,9</sup>

## Additional resources about integration

- ***Analysis of single cell RNA-seq data*** Sanger course:

<https://www.singlecellcourse.org/scrna-seq-dataset-integration.html>

- ***Single-cell best practices*** website:

[https://www.sc-best-practices.org/cellular\\_structure/integration.html](https://www.sc-best-practices.org/cellular_structure/integration.html)



UNIVERSITY  
OF TURKU

ÅA  
Åbo Akademi

BF  
Biocenter Finland



InFLAMES  
Solution is in Immunity



Enlightten<sup>+</sup>

TURKU  
BIOSCIENCE

elixir

This project has received funding from the  
European Union's Horizon 2020 research and innovation programme  
under the Marie Skłodowska-Curie grant agreement No.: 955321



Enlightten<sup>+</sup>

## sclB: kBET, kNN & ASW (*dependent on cell label identity*)

### kBET (*k-nearest-neighbor batch effect test*):

- compares the batch label distribution of a local neighborhood of a (kNN) graph for a subset of a given cell label against the global neighborhood – Pearson’s Chi-squared test. Null hypothesis is that the distributions are similar, *i.e.*, batches are well mixed. K-BET result is the average rejection rate. Lower value means no/weak batch effect.

### kNN (*k-nearest-neighbor graph connectivity*):

- measures if all the cells from a given cell identity label are all connected (1).

### ASW (*average silhouette width*):

- measures the within- and between-cluster distances (to the closest cluster). The average of all widths corresponds to the ASW. The ASW ranges between -1 - 0 - 1 (misclassification - overlapping clusters - well-separated dense clusters). Computed on the integrated embedding (or PCA). Used for assessing batch effects as well as cell label conservation (some differences).

$$GC = \frac{1}{|C|} \sum_{c \in C} \frac{|\text{LCC}(G(N_c; E_c))|}{|N_c|}.$$

## sclB: iLISI & PCA (*independent on cell label identity*)

Graph iLISI (*graph integration local inverse Simpson's Index*):

- “*the inverse Simpson's index is used to determine the number of cells that can be drawn from a neighbor list before one batch is observed twice*”. It is also used for cell-identity, known as cLISI. A value of 0 corresponds to low batch integration or cell-type separation.

PCA regression:

- the  $R$ -squared of the linear regression between each principal component and the batch variable (covariate) is calculated. The variance contribution of the batch effect by each PC is calculated as the product of the variance and the  $R$ -squared. The sum of the individual variances gives the total variance explained by the batch variable.

$$\text{Var}(C|B) = \sum_{i=1}^G \text{Var}(C|\text{PC}_i) \times R^2(\text{PC}_i|B),$$

Paper: <https://www.nature.com/articles/s41592-021-01336-8>

sclB github repo: <https://github.com/theislab/scib>

scib-pipeline github repo: <https://github.com/theislab/scib-pipeline>

## sclB: ARI & NMI (*label conservation*)

### ARI (*Adjusted Rand Index*):

- it compares two clustering results, i.e., cell-labels vs louvain clustering results on the integrated data set, accounting for both, matches and mismatches. Values of 0 or 1 correspond to a random classification or to a total match.

### NMI (*Normalized Mutual Information*):

- it compares the overlap between cell-type labels with the louvain clustering results on the integrated data set. The overlap is scaled (0-1) *“using the mean of the entropy terms for cell-type and cluster labels”*. A value of 0 means uncorrelated clustering and a value of 1 perfect match.

Paper: <https://www.nature.com/articles/s41592-021-01336-8>

sclB github repo: <https://github.com/theislab/scib>

scib-pipeline github repo: <https://github.com/theislab/scib-pipeline>