Multi-sample single-cell analysis

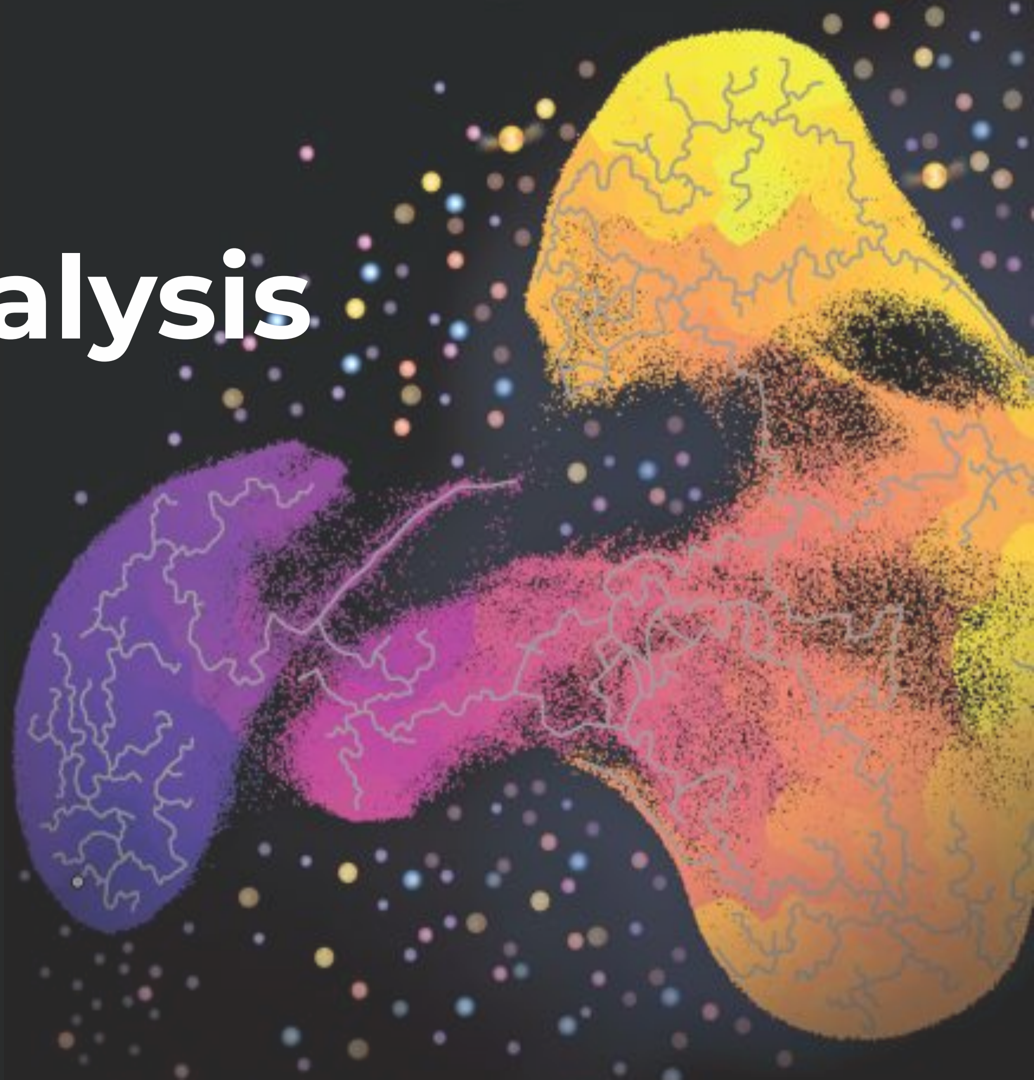
Lucia Ramirez Navarro

Wellcome Sanger Institute

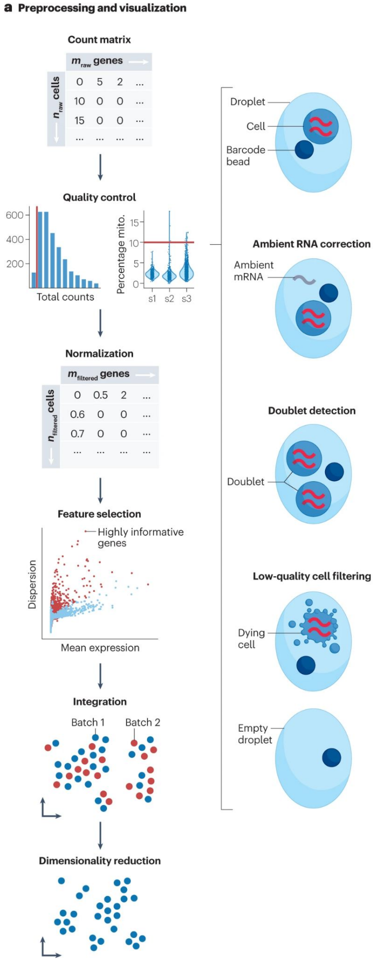
Single Cell Genomic Approaches to Study the Immune System

November 10th, 2024

lr23@sanger.ac.uk

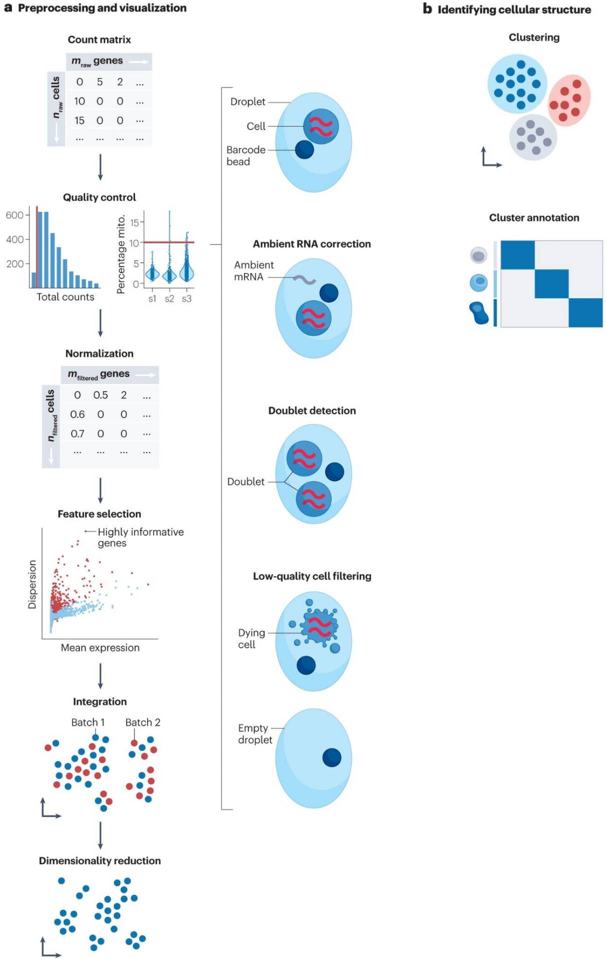


Single-cell workflow



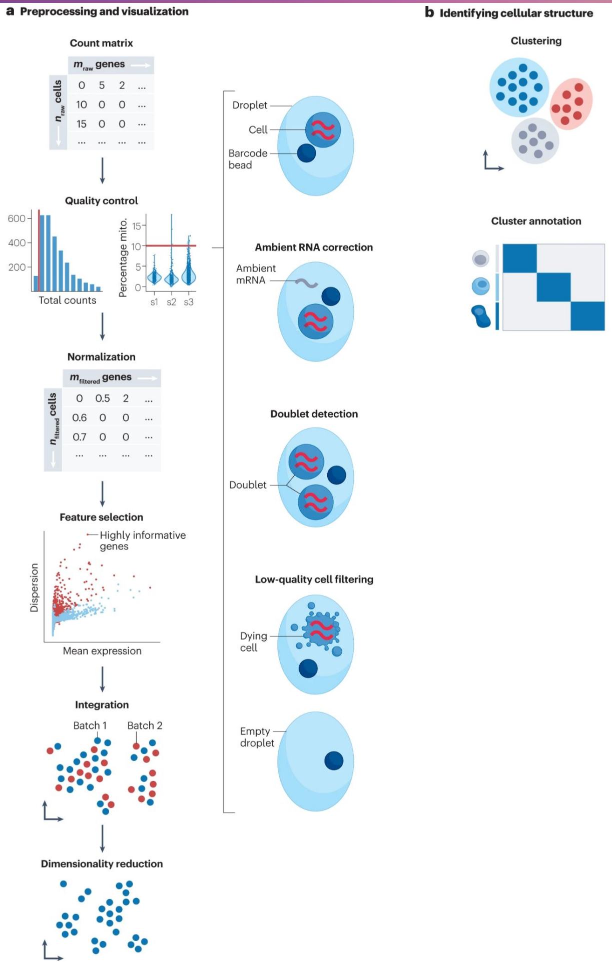
Modified from Heumos, L. et. al (2023)

Single-cell workflow



Modified from Heumos, L. et. al (2023)

Single-cell workflow



Count matrix

0	5	2	...
10	0	0	...
15	0	0	...
...	...	...	...

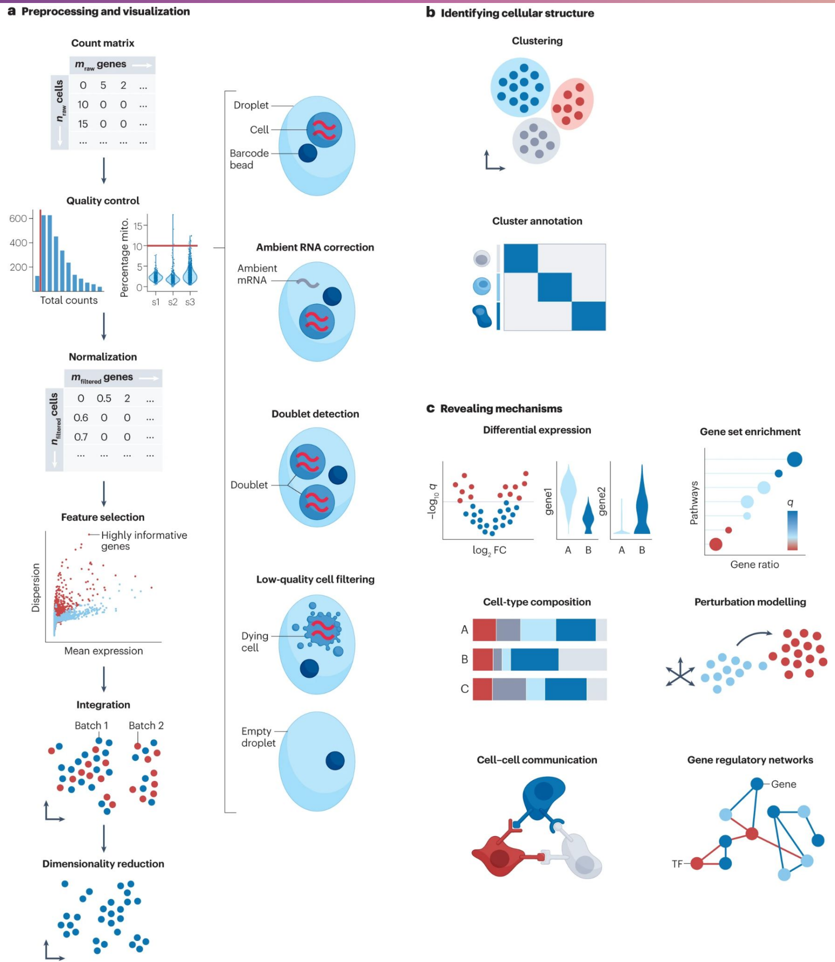
Normalization

0	0.5	2	...
0.6	0	0	...
0.7	0	0	...
...	...	...	...

What now?

Modified from Heumos, L. et. al (2023)

Single-cell workflow



Modified from Heumos, L. et. al (2023)

Topics

1. Dimensionality reduction and batch correction (integration)
2. Differential gene expression analysis (DGE)
3. Differential abundance analysis (DA)
4. Cell-cell communication

Topics

1. **Dimensionality reduction and batch correction (integration)**
2. Differential gene expression analysis (DGE)
3. Differential abundance analysis (DA)
4. Cell-cell communication

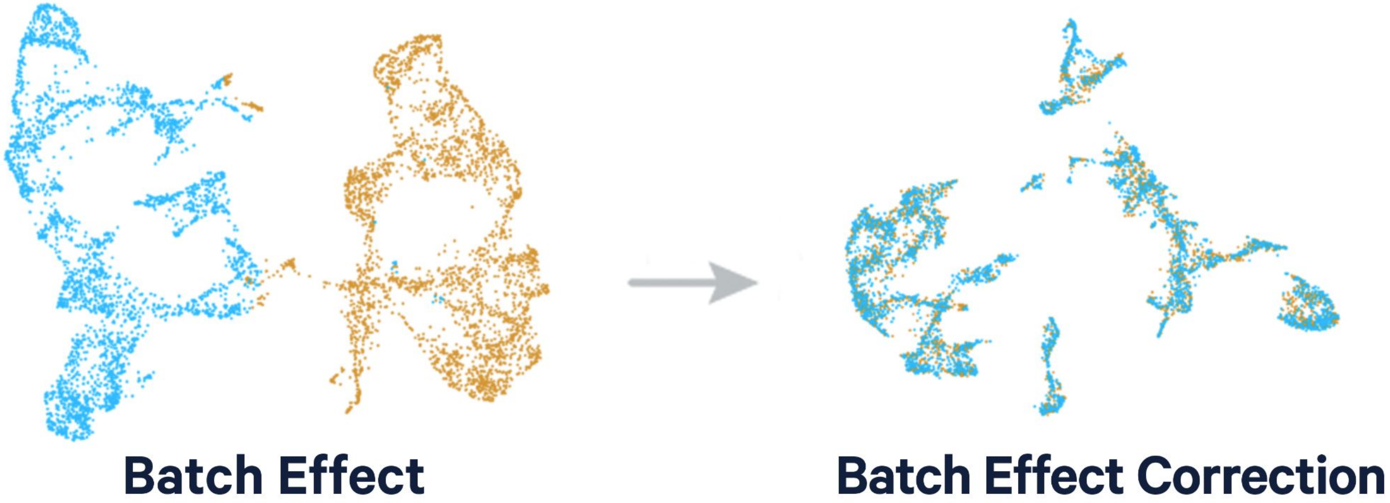
The challenge in analyzing single cell data

Batch effects are changes in gene expression levels that results from handling cells in different groups or “batches”.

- **Technical** : Sample collection and processing (eg. plate effects; protocols)
- **Biological**: Conditions (eg healthy vs disease), cell type differences, etc

The challenge in analyzing single cell data

Only by performing integration, you will be able to define common cell types across batches



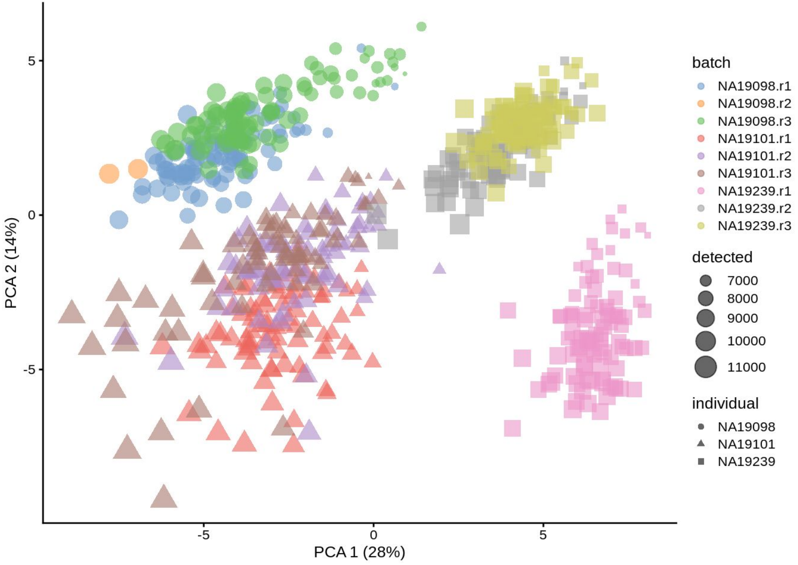
Batch Effect

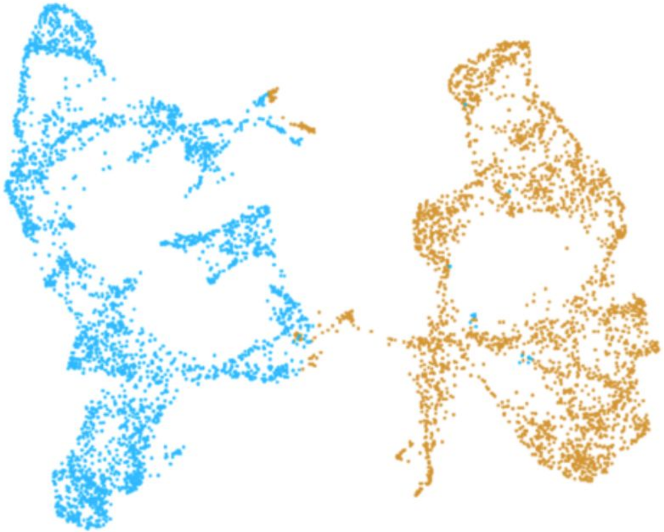
Batch Effect Correction

In general, batch effects are difficult to pin down and are dataset-dependent.

How to detect batch effects?

1. Visualize your data! (PCA / UMAP)





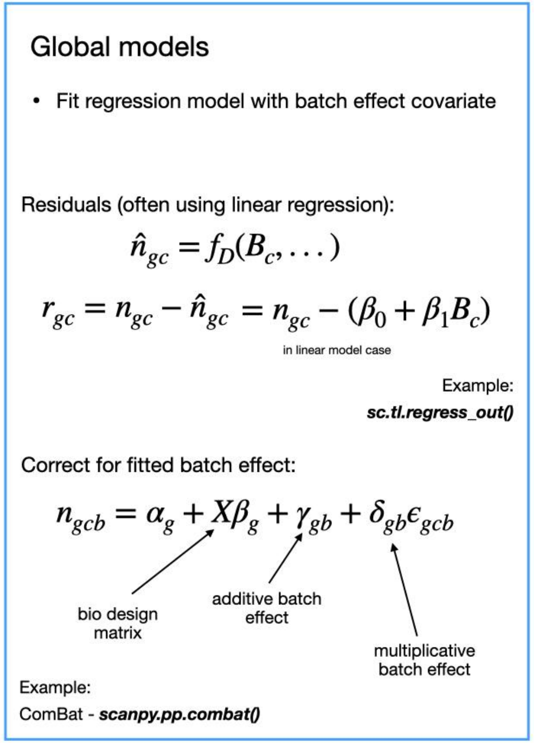
How to detect batch effects?

1. Visualize your data! (PCA / UMAP)
2. Calculate variance explained by X (scater::getExplanatoryPCs for R / scib_metrics.utils.principal_component_regression in python)

How to detect batch effects?

1. Visualize your data! (PCA / UMAP)
2. Calculate variance explained by X (scater::getExplanatoryPCs for R / scib_metrics.utils.principal_component_regression in python)
3. Post-annotation: does your clusters have similar abundance across all your samples?

Integration methods



Global models

- Fit regression model with batch effect covariate

Residuals (often using linear regression):

$$\hat{n}_{gc} = f_D(B_c, \ldots)$$

$$r_{gc} = n_{gc} - \hat{n}_{gc} = n_{gc} - (\beta_0 + \beta_1 B_c)$$

in linear model case

Example:
sc.tl.regress_out()

Correct for fitted batch effect:

$$n_{gcb} = \alpha_g + X\beta_g + \gamma_{gb} + \delta_{gb}\epsilon_{gcb}$$

bio design matrix
additive batch effect
multiplicative batch effect

Example:
ComBat - ***scanpy.pp.combat()***

Fig taken from Heumos, L. et. al (2023)

Integration methods

Global models

- Fit regression model with batch effect covariate

Residuals (often using linear regression):

$$\hat{n}_{gc} = f_D(B_c, \ldots)$$

$$r_{gc} = n_{gc} - \hat{n}_{gc} = \underbrace{n_{gc} - (\beta_0 + \beta_1 B_c)}_{\text{in linear model case}}$$

Example:
sc.tl.regress_out()

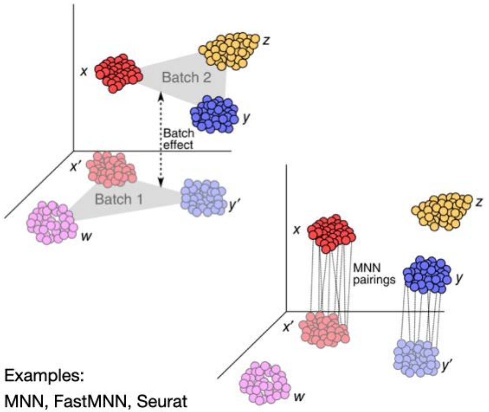
Correct for fitted batch effect:

$$n_{gcb} = \alpha_g + \underbrace{X\beta_g}_{\text{bio design matrix}} + \underbrace{\gamma_{gb}}_{\text{additive batch effect}} + \underbrace{\delta_{gb}}_{\text{multiplicative batch effect}}\epsilon_{gcb}$$

Example:
ComBat - ***scanpy.pp.combat()***

Linear embedding models

- Project cells into low dimensional embedding
- find most similar cells in other batch e.g., using mutual nearest neighbours (MNNs)
- Use MNNs as anchors to calculate a correction vector



Examples:
MNN, FastMNN, Seurat v3, Scanorama

Fig taken from Heumos, L. et. al (2023)

Integration methods

Global models

- Fit regression model with batch effect covariate

Residuals (often using linear regression):

$$\hat{n}_{gc} = f_D(B_c, \dots)$$

$$r_{gc} = n_{gc} - \hat{n}_{gc} = n_{gc} - (\beta_0 + \beta_1 B_c)$$

in linear model case

Example:

sc.tl.regress_out()

Correct for fitted batch effect:

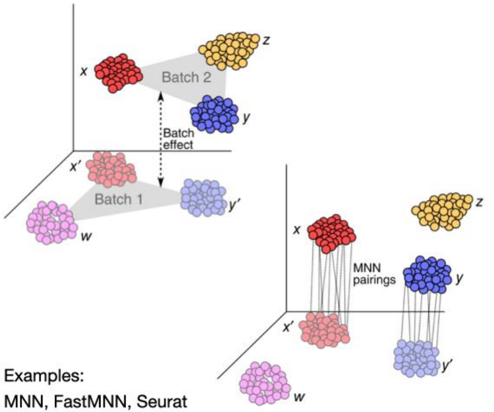
$$n_{gcb} = \alpha_g + \underbrace{X\beta_g}_{\text{bio design matrix}} + \underbrace{\gamma_{gb}}_{\text{additive batch effect}} + \underbrace{\delta_{gb}}_{\text{multiplicative batch effect}}\epsilon_{gcb}$$

Example:

ComBat - ***scanpy.pp.combat()***

Linear embedding models

- Project cells into low dimensional embedding
- find most similar cells in other batch e.g., using mutual nearest neighbours (MNNs)
- Use MNNs as anchors to calculate a correction vector

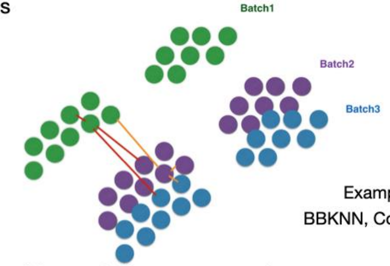


Examples:

MNN, FastMNN, Seurat v3, Scanorama

Graph-based methods & Deep learning

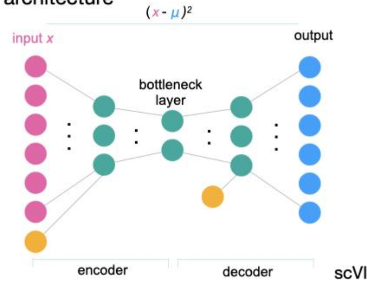
Enforce graph connections between different batches



Examples:

BBKNN, Conos

Add condition node into auto-encoder architecture

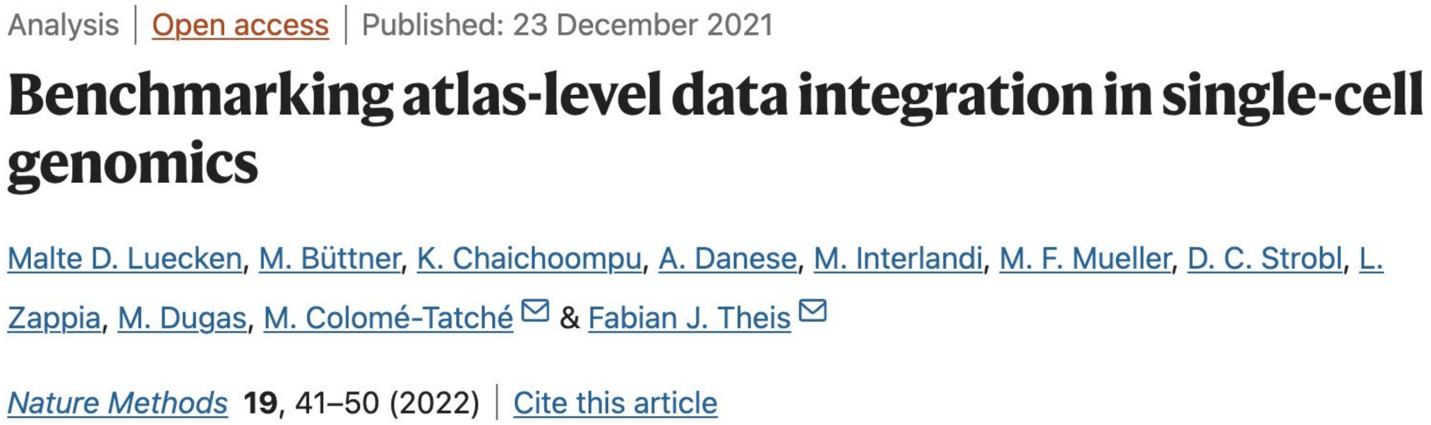


Examples:

scVI, trVAE, SAUCIE

Fig taken from Heumos, L. et. al (2023)

How to choose an integration method?



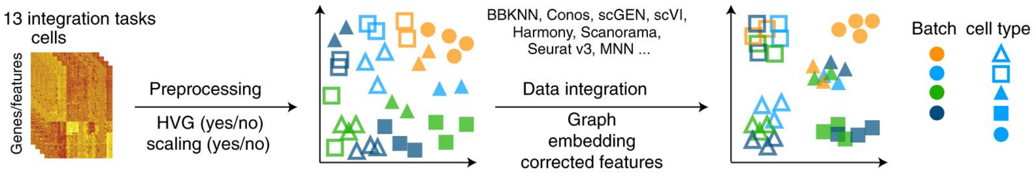
Analysis | Open access | Published: 23 December 2021

Benchmarking atlas-level data integration in single-cell genomics

Malte D. Luecken, M. Büttner, K. Chaichoompu, A. Danese, M. Interlandi, M. F. Mueller, D. C. Strobl, L. Zappia, M. Dugas, M. Colomé-Tatché ✉ & Fabian J. Theis ✉

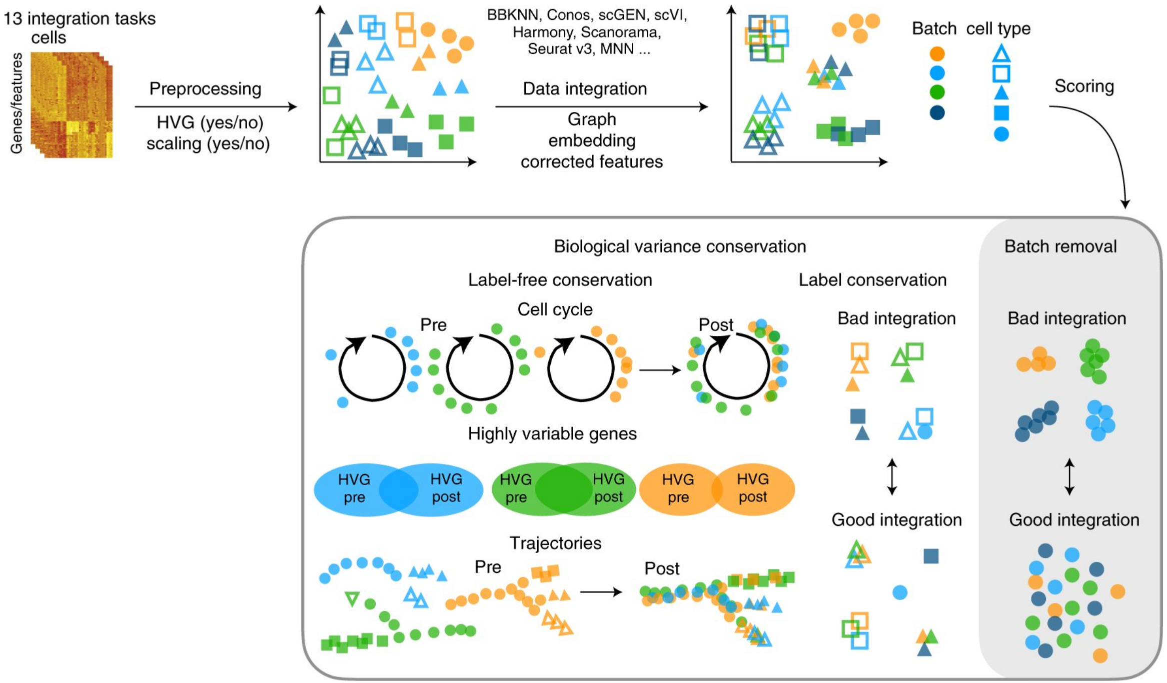
Nature Methods **19**, 41–50 (2022) | Cite this article

How to choose an integration method?



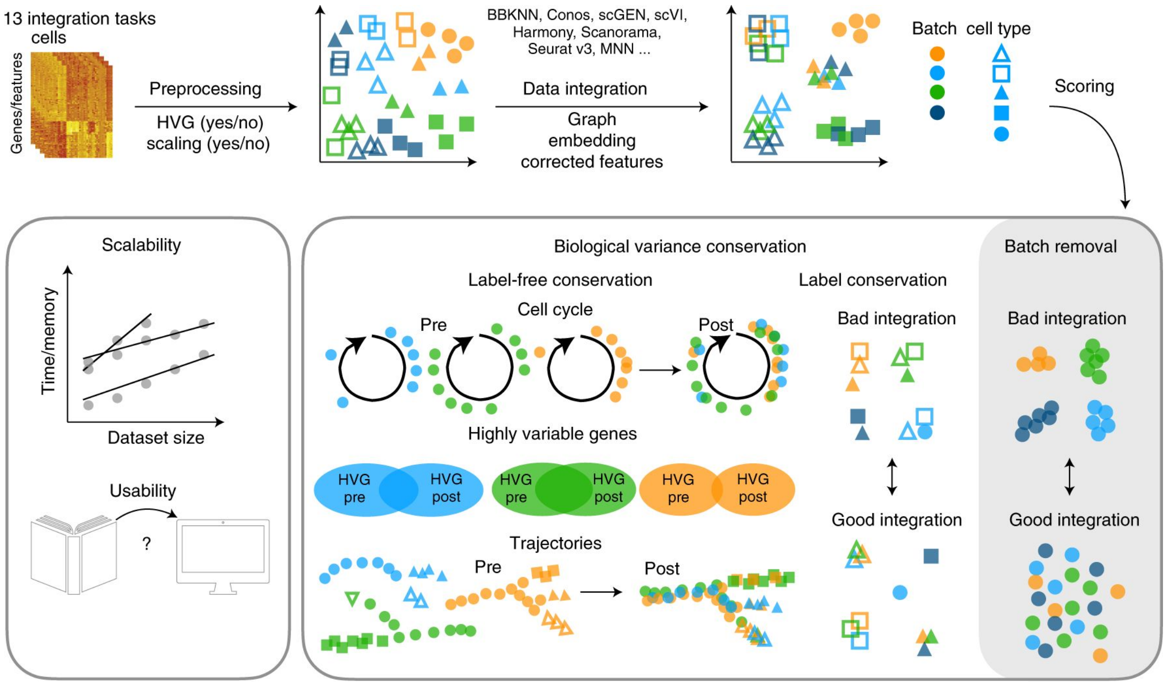
- 7 scRNA-seq (5 real and 2 simulated) and 6 scATAC-seq tasks
- For each task, they also also consider different combinations of pre-processing, including highly variable gene (HVG) selection and scaling
- 16 integration methods

How to choose an integration method?



- 7 scRNA-seq (5 real and 2 simulated) and 6 scATAC-seq tasks
- For each task, they also also consider different combinations of pre-processing, including highly variable gene (HVG) selection and scaling
- 16 integration methods
- 14 metrics

How to choose an integration method?



- 7 scRNA-seq (5 real and 2 simulated) and 6 scATAC-seq tasks
- For each task, they also also consider different combinations of pre-processing, including highly variable gene (HVG) selection and scaling
- 16 integration methods
- 14 metrics

How to choose an integration method?

- **Depends on your dataset** (how complex is the experimental design, do you have predefined labels?, size) **and capabilities** (language, GPU availability, etc)

How to choose an integration method?

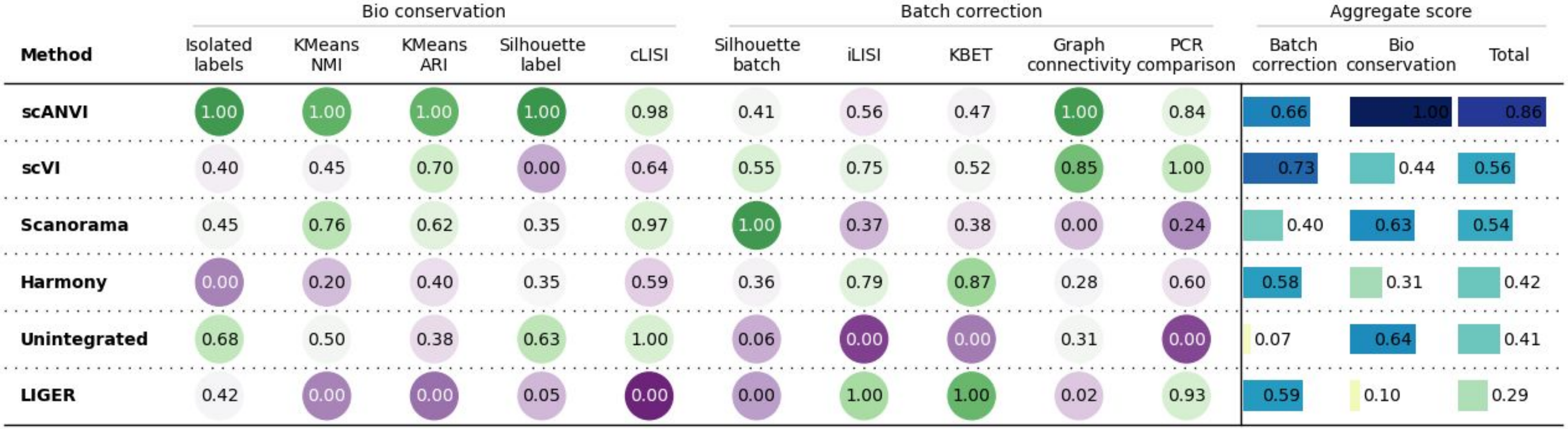
- **Depends on your dataset** (how complex is the experimental design, do you have predefined labels?, size) **and capabilities** (language, GPU availability, etc)
- Some general rules for scRNA-seq:
    - Linear-embedding models, particularly Harmony performs well for simple integration tasks
    - Deep-learning approaches such as scANVI, scVI and scGen as well as Scanorama (linear embedding method) can perform well for more complex tasks
        - scANVI and scGen can incorporate cell type labels
    - Deep-learning approaches often times allow the use of GPU for scalability

How to choose an integration method?

- **Depends on your dataset** (how complex is the experimental design, do you have predefined labels?, size) **and capabilities** (language, GPU availability, etc)
- Some general rules for scRNA-seq:
  - Linear-embedding models, particularly Harmony performs well for simple integration tasks
  - Deep-learning approaches such as scANVI, scVI and scGen as well as Scanorama (linear embedding method) can perform well for more complex tasks
    - scANVI and scGen can incorporate cell type labels
  - Deep-learning approaches often times allow the use of GPU for scalability
- Want to decide yourself? The scib metrics are available to benchmark but you need a proxy ground truth

Benchmarking integration methods

[scIB package](#) / [scib-metrics](#)



Method	Bio conservation					Batch correction					Aggregate score		
	Isolated labels	KMeans NMI	KMeans ARI	Silhouette label	cLISI	Silhouette batch	iLISI	KBET	Graph connectivity	PCR comparison	Batch correction	Bio conservation	Total
scANVI	1.00	1.00	1.00	1.00	0.98	0.41	0.56	0.47	1.00	0.84	0.66	1.00	0.86
scVI	0.40	0.45	0.70	0.00	0.64	0.55	0.75	0.52	0.85	1.00	0.73	0.44	0.56
Scanorama	0.45	0.76	0.62	0.35	0.97	1.00	0.37	0.38	0.00	0.24	0.40	0.63	0.54
Harmony	0.00	0.20	0.40	0.35	0.59	0.36	0.79	0.87	0.28	0.60	0.58	0.31	0.42
Unintegrated	0.68	0.50	0.38	0.63	1.00	0.06	0.00	0.00	0.31	0.00	0.07	0.64	0.41
LIGER	0.42	0.00	0.00	0.05	0.00	0.00	1.00	1.00	0.02	0.93	0.59	0.10	0.29

Summary I

- Integration methods are useful tool to remove unwanted variation in our single-cell data and find a common structure in the data across batches.
- There is no rule to define ‘unwanted variation’. It depends on your experimental design and the question being asked. Oftentimes, it is useful to visualize your data before attempting to correct for batch effects. Batch effect correction is not always required and it might mask the biological variation of interest.

Summary I

- Integration methods are useful tool to remove unwanted variation in our single-cell data and find a common structure in the data across batches.
- There is no rule to define ‘unwanted variation’. It depends on your experimental design and the question being asked. Oftentimes, it is useful to visualize your data before attempting to correct for batch effects. Batch effect correction is not always required and it might mask the biological variation of interest.
- Harmony is useful for simple design while deep-learning methods like scVI are better for complex designs. If cell labels are available, scANVI or scGen are preferred

Summary I

- Integration methods are useful tool to remove unwanted variation in our single-cell data and find a common structure in the data across batches.
- There is no rule to define ‘unwanted variation’. It depends on your experimental design and the question being asked. Oftentimes, it is useful to visualize your data before attempting to correct for batch effects. Batch effect correction is not always required and it might mask the biological variation of interest.
- Harmony is useful for simple design while deep-learning methods like scVI are better for complex designs. If cell labels are available, scANVI or scGen are preferred
- If possible, consider running several integration methods on your dataset and evaluate them with the scIB metrics

Topics

1. Dimensionality reduction and batch correction (integration)
2. **Differential gene expression analysis (DGE)**
3. Differential abundance analysis (DA)
4. Cell-cell communication

Differential gene expression (DGE) analysis

Having multi-sample and multi-condition experiments allow us to test the magnitude and significance of differences in gene expression patterns between these conditions of interest

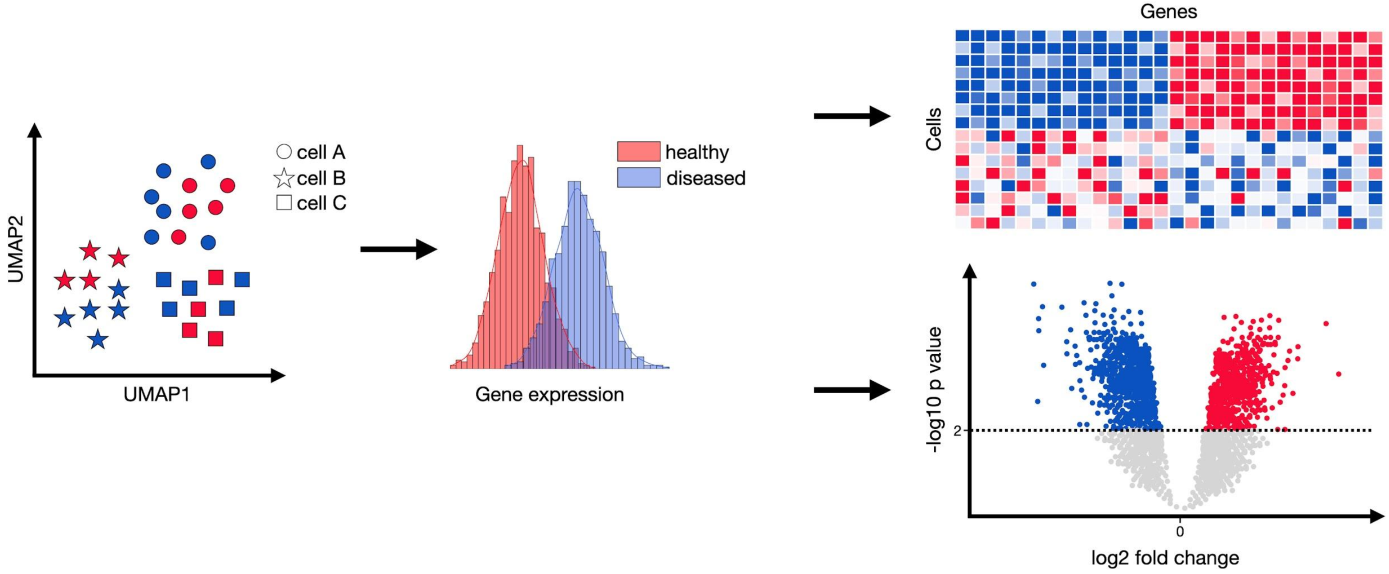


Fig taken from Heumos, L. et. al (2023)

A cautious tale of using t-test / wilcoxon-test for DGE

- Implies each cell is an independent biological replicate, resulting in a higher false discovery rate (FDR) (1,2,3,4,5,6).
- Doesn’t take into account the excess number of zeros.
- Usually, we don’t have enough samples to accurately estimate the variance without pooling information across genes.
- If using raw counts, we are not considering that gene expression levels depends on the library preparation and sequencing depth.

1. https://doi.org/10.1038/s41467-021-21038-1
2. https://doi.org/10.1038/s41467-022-35519-4
3. https://doi.org/10.1038/s41467-022-35520-x
4. https://doi.org/10.1093/bib/bbac286
5. https://doi.org/10.1038/s41467-021-25960-2
6. https://doi.org/10.1038/s41467-020-19894-4

Methods to model gene expression

Pseudobulk: aggregate cells per sample and cluster (mean/median/sum) and then analyse the data with methods originally designated for bulk expression such as edgeR, DEseq2 and limma.

Collapsing cells into samples reflects the fact that our biological replication occurs at the sample level and avoids problems from modelling correlations between samples.

Methods to model gene expression

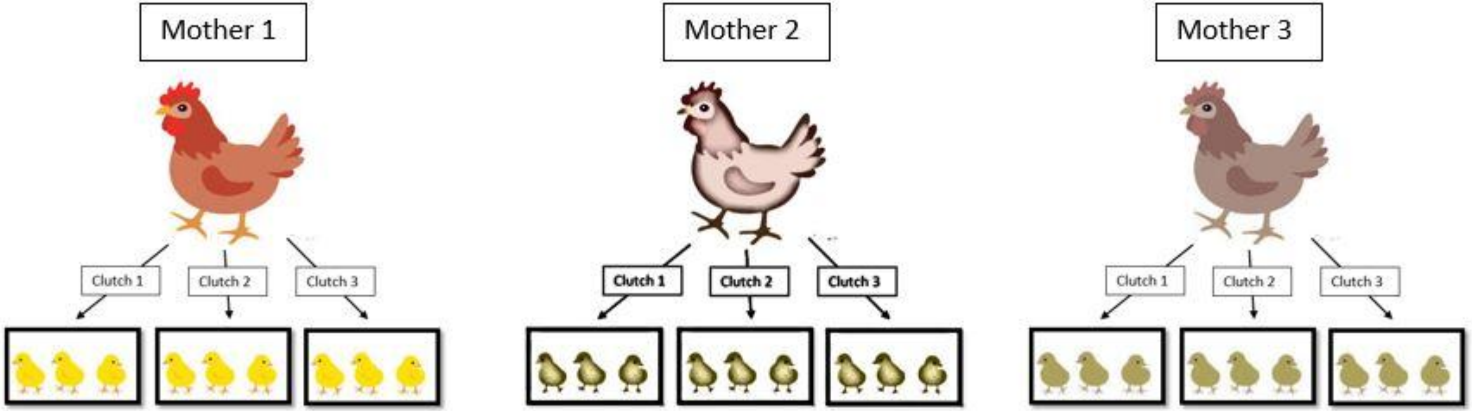
Pseudobulk: aggregate cells per sample and cluster (mean/median/sum) and then analyse the data with methods originally designated for bulk expression such as edgeR, DEseq2 and limma.

Collapsing cells into samples reflects the fact that our biological replication occurs at the sample level and avoids problems from modelling correlations between samples.

Single-cell: cells are modeled individually using mixed models to account for batch effects and the sample correlation. The list includes MAST and nebula.

What are mixed models?

Class of generalized linear models that are very useful when working with a within-subjects design.



Types of effects

- **Random effects:** discrete groupings which are variable or the source of random variability within the dataset. These are often factors that represent a random factor sampled of a larger population (such as field sites, genotype, temporal blocks used in a study).
- **Fixed effects:** are constant and tend to be directly measured in the experiment (but can have multiple levels).

Expression ~ sex + age + … + (1|donor)

DGE workflow

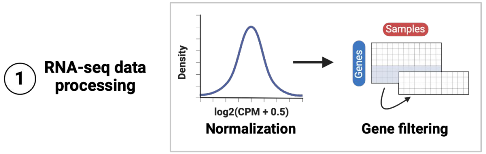
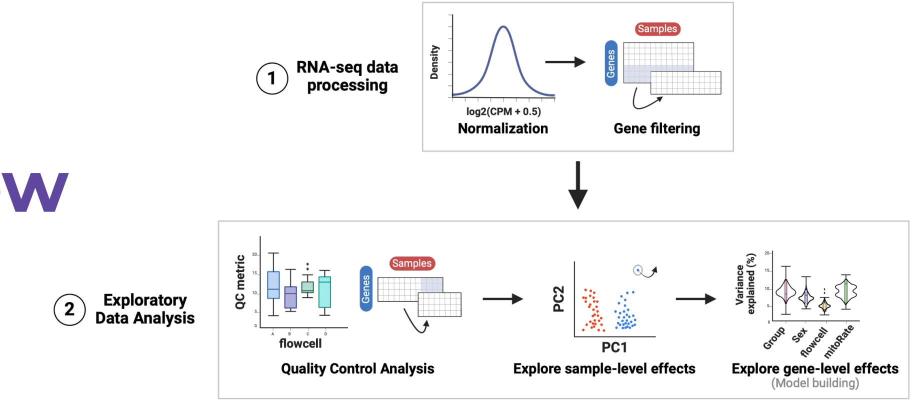


Fig taken from https://lcolladotor.github.io/cshl_rstats_genome_scale_2024

DGE workflow



1 RNA-seq data processing

Normalization

Gene filtering

2 Exploratory Data Analysis

Quality Control Analysis

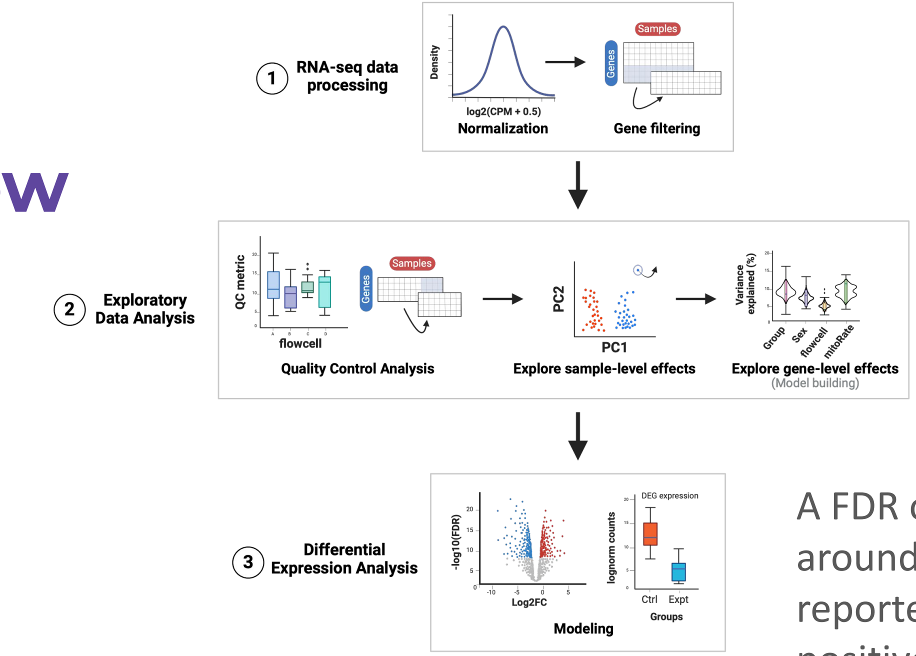
Explore sample-level effects

Explore gene-level effects

(Model building)

Fig taken from https://lcolladotor.github.io/cshl_rstats_genome_scale_2024

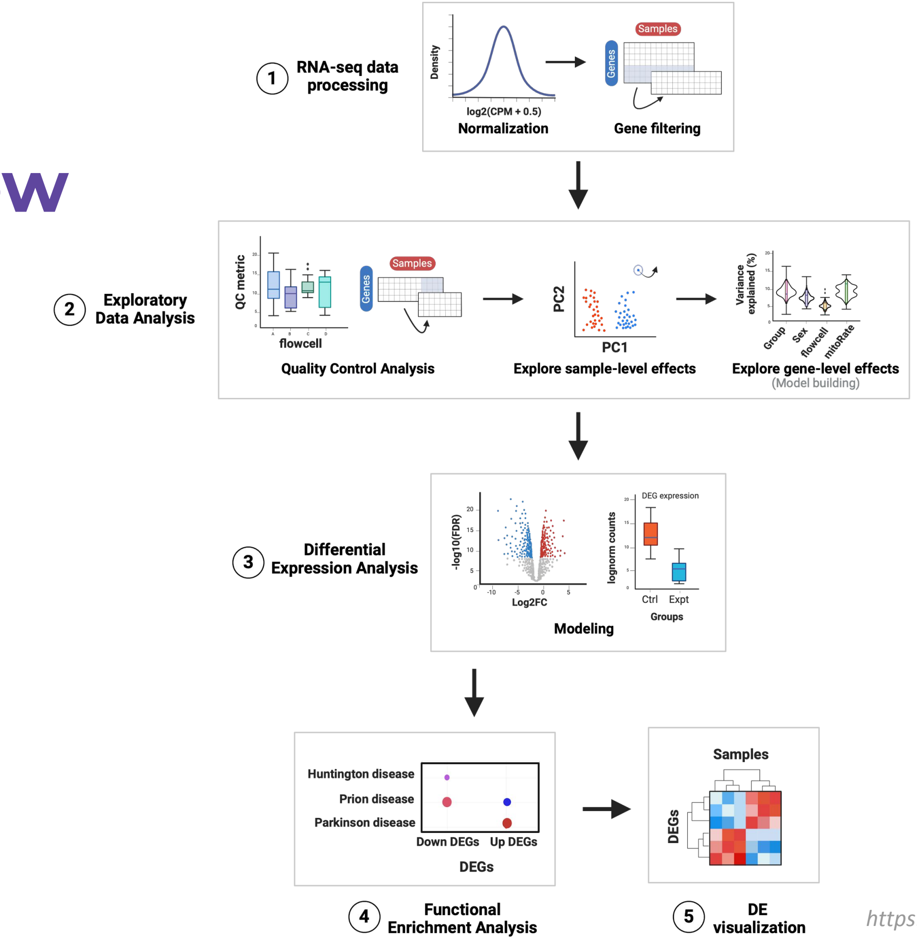
DGE workflow



A FDR of 0.05 indicates that around 5% of significantly reported results are actually false positives

Fig taken from https://lcolladotor.github.io/cshl_rstats_genome_scale_2024

DGE workflow



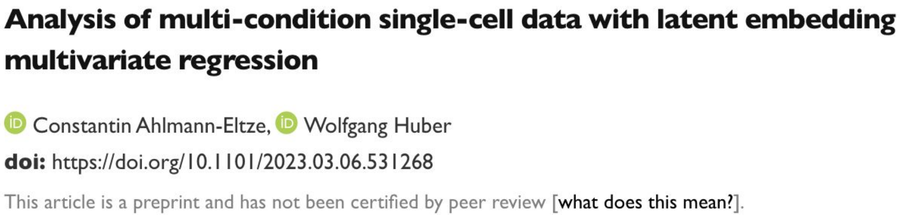
Gene set enrichment analysis

The aim is to identify gene programs, such as biological processes, gene ontologies or regulatory pathways that are over-represented in an experimental condition compared to control or other conditions, on the basis of differentially expressed (DE) genes.

MSigDB database

Fig taken from https://lcolladotor.github.io/cshl_rstats_genome_scale_2024

Beyond cell type labels

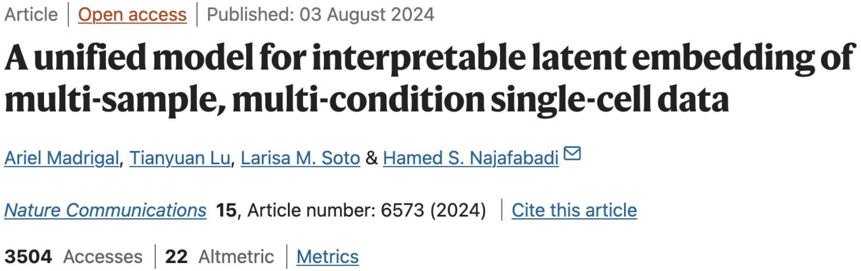


Analysis of multi-condition single-cell data with latent embedding multivariate regression

Constantin Ahlmann-Eltze, Wolfgang Huber

doi: https://doi.org/10.1101/2023.03.06.531268

This article is a preprint and has not been certified by peer review [what does this mean?].



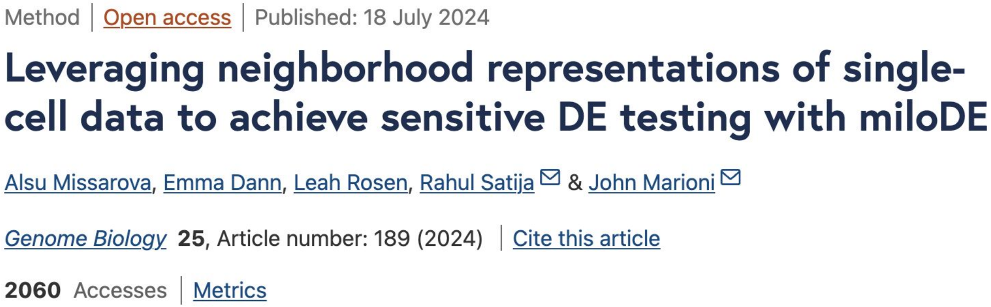
Article | Open access | Published: 03 August 2024

A unified model for interpretable latent embedding of multi-sample, multi-condition single-cell data

Ariel Madrigal, Tianyuan Lu, Larisa M. Soto & Hamed S. Najafabadi

Nature Communications **15**, Article number: 6573 (2024) | Cite this article

3504 Accesses | **22** Altmetric | Metrics



Method | Open access | Published: 18 July 2024

Leveraging neighborhood representations of single-cell data to achieve sensitive DE testing with miloDE

Alsu Missarova, Emma Dann, Leah Rosen, Rahul Satija & John Marioni

Genome Biology **25**, Article number: 189 (2024) | Cite this article

2060 Accesses | Metrics

Robustness?

Interpretability?

Speed?

Summary II

- Single-cell data contains repeated measurements (cells) from the same individual in scRNA-seq. Failing to account for that results in lack of sensitivity and specificity in the DGE analysis.

Summary II

- Single-cell data contains repeated measurements (cells) from the same individual in scRNA-seq. Failing to account for that results in lack of sensitivity and specificity in the DGE analysis.
- We can account for that by aggregating cells via a pseudobulk analysis or using a mixed model and account for individual as a fixed effect. (Both methods have similar performances).

Summary II

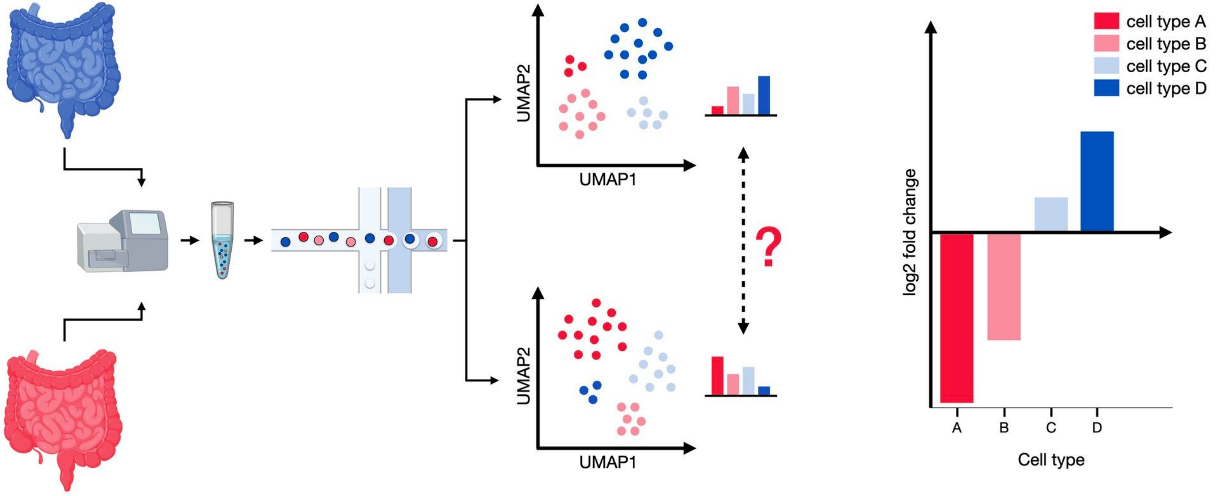
- Single-cell data contains repeated measurements (cells) from the same individual in scRNA-seq. Failing to account for that results in lack of sensitivity and specificity in the DGE analysis.
- We can account for that by aggregating cells via a pseudobulk analysis or using a mixed model and account for individual as a fixed effect. (Both methods have similar performances).
- After DGE, we can use gene set enrichment analysis to determine the biological relevance of our DE results.

Topics

1. Dimensionality reduction and batch correction (integration)
2. Differential gene expression analysis (DGE)
3. **Differential abundance analysis (DA)**
4. Cell-cell communication

Differential abundance (DA) analysis

Test for significant changes in cell type abundances across conditions.



Single-cell data represents a snapshot in time and it’s limited to the number of samples we sequence (compositional data - proportions)

Fig taken from Heumos, L. et. al (2023)

How to model cell type abundances?

- edgeR : allow us to take advantage of the NB GLM method to model overdispersed count data. In this case, the counts are not reads per genes, but cells per label.
- The advantage of using edgeR over simple statistics such as t-test/wilcoxon test is that we can share information across cell types to improve our estimates of the biological variability in cell abundance between replicates. Additionally, we can account for batch effects (eg age, sex, etc)

Other methods: scCODA

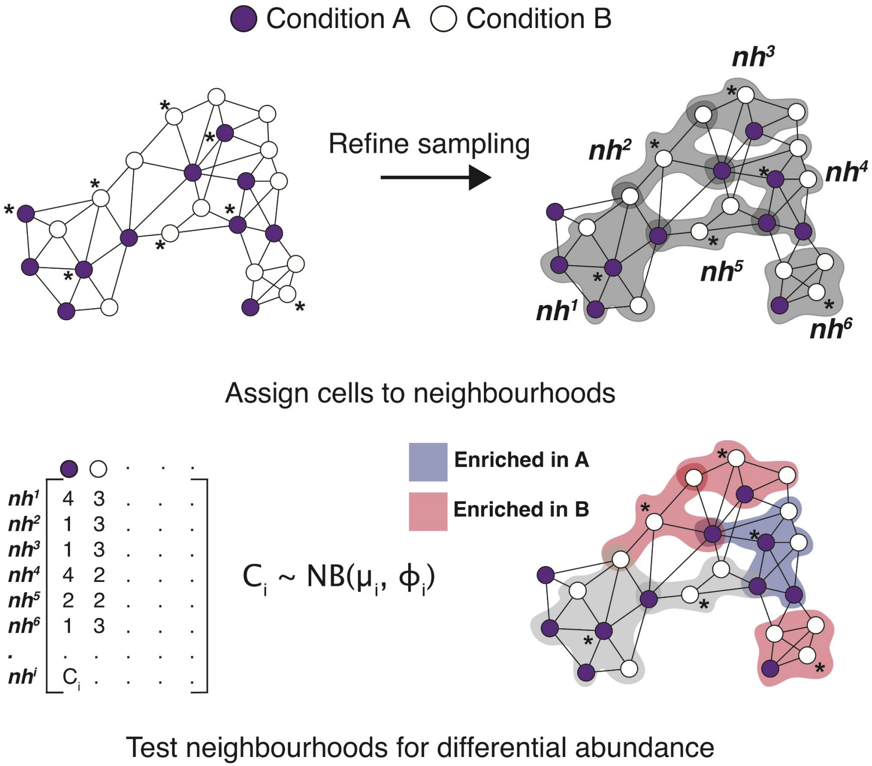
Bayesian approach approach that uses a hierarchical Dirichlet-Multinomial model (from microbiome analysis) which accounts for uncertainty in cell-type proportions and the negative correlative bias via joint modeling of all measured cell-type proportions.

Cons: any detected compositional changes by scCODA always have to be viewed in relation to the selected cell type of reference.

Beyond cell type labels

miloR assigns cells to partially overlapping neighborhoods on the KNN graph, then differential abundance (DA) testing is performed by modelling cell counts with a generalized linear model (GLM) (edgeR)

Cons: KNN graph is limited by the integration method and cells in a neighborhood may not necessarily represent a specific, unique biological subpopulation, because a cellular state may span over multiple neighborhoods.



Assign cells to neighbourhoods

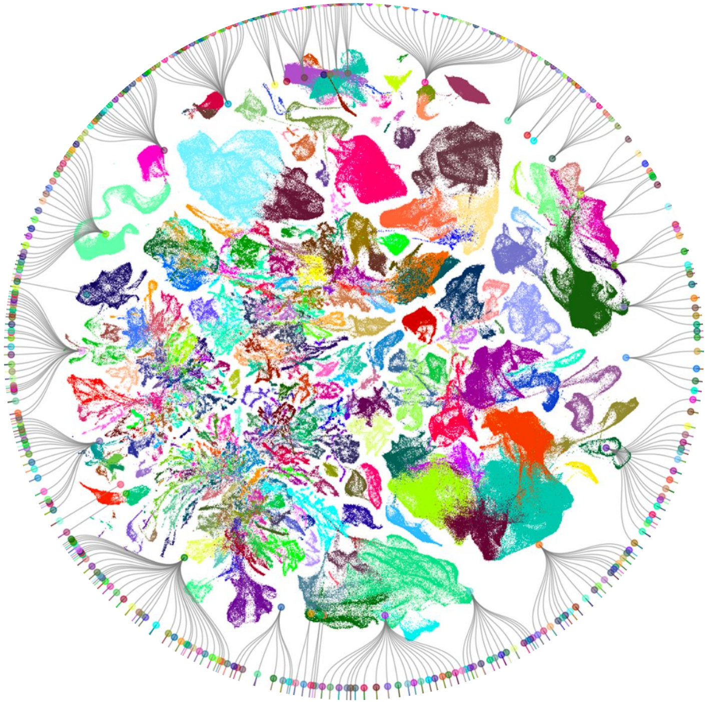
$$\begin{array}{c|ccccc} & \bullet & \circ & \cdot & \cdot & \cdot \\ nh^1 & 4 & 3 & \cdot & \cdot & \cdot \\ nh^2 & 1 & 3 & \cdot & \cdot & \cdot \\ nh^3 & 1 & 3 & \cdot & \cdot & \cdot \\ nh^4 & 4 & 2 & \cdot & \cdot & \cdot \\ nh^5 & 2 & 2 & \cdot & \cdot & \cdot \\ nh^6 & 1 & 3 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ nh^i & C_i & \cdot & \cdot & \cdot & \cdot \end{array}$$

$$C_i \sim NB(\mu_i, \phi_i)$$

Test neighbourhoods for differential abundance

DE vs DA? Two sides of the same coin

The distinction between DA and DGE is inherently artificial for scRNA-seq because the labels used in DA are defined based on the genes that are also tested for DGE.



Summary III

- DA analysis can tests for differences in cell proportions across conditions but it is limited by the quality of annotation.
- Clustering-free methods like Milo can be used to circumvent this, particularly for development processes of changes that might appear in transitional states between cell types or in a specific subset of cells of a given cell type. Milo will be limited by the quality of integration

Topics

1. Dimensionality reduction and batch correction (integration)
2. Differential gene expression analysis (DGE)
    a. Gene set enrichment analysis
3. Differential abundance analysis (DA)
4. **Cell-cell communication**

Cell-cell communication (CCC)

- CCC consists of using repositories of ligands, receptors and their interactions to predict interactions between annotated clusters. They use the gene expression information as a proxy of protein abundance. These CCC tools infer intercellular crosstalk between pairs of cell groups, one group being the source and the other the receiver of a CCC event.

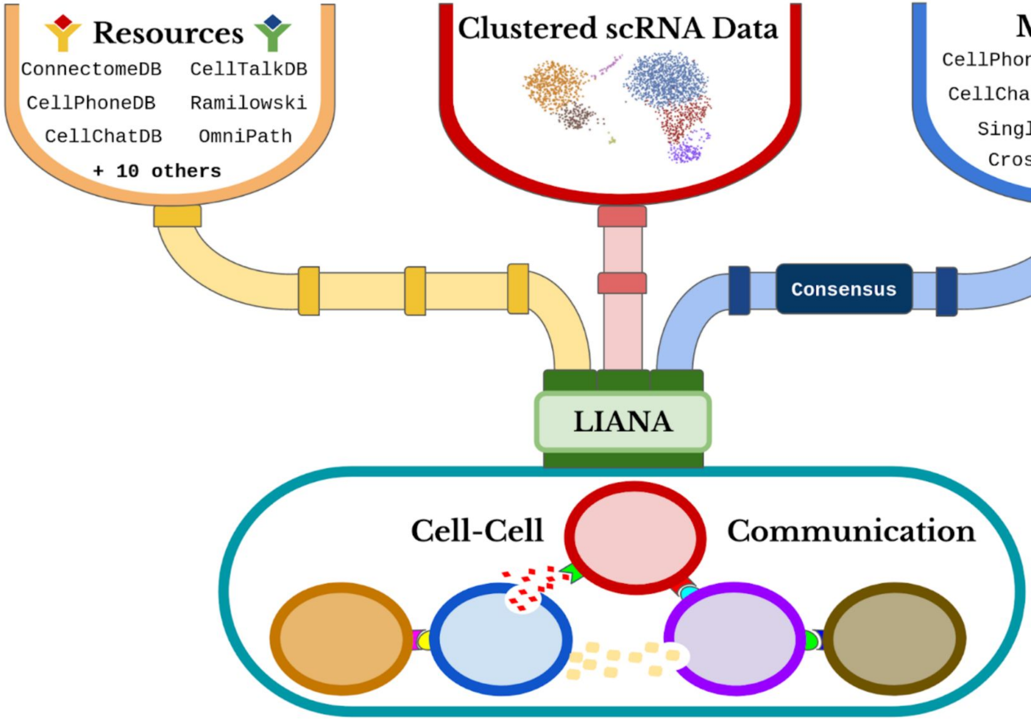
Cell-cell communication (CCC)

- CCC consists of using repositories of ligands, receptors and their interactions to predict interactions between annotated clusters. They use the gene expression information as a proxy of protein abundance. These CCC tools infer intercellular crosstalk between pairs of cell groups, one group being the source and the other the receiver of a CCC event.

- Cons:
  - These databases tend to be biased towards specific pathways, functional categories and tissue-enriched proteins (1)
  - The choice of method and interaction database has a strong effect on the predicted interactions (1)
  - Assumes that protein co-expression leads to cellular interactions.

1. *https://doi.org/10.1038/s41467-022-30755-0*

LIANA



https://saezlab.github.io/liana/

Summary IV

- CCC analysis aims to predict interactions between different cell types using single-cell transcriptomics data.
- CCC analysis is an emergent field with methods being constantly developed. One should proceed with caution when running these methods as results can change depending on the database and method

Other applications

- Network modelling
- Splicing analysis
- Correlating gene expression with genotype data (eQTL)

Additional resources

- Papers about best single-cell practices:
  - https://www.nature.com/articles/s41576-023-00586-w
  - https://www.embopress.org/doi/full/10.15252/msb.20188746

- Tutorials on single-cell analysis and best practices:
  - https://bioconductor.org/books/3.19/OSCA/
    - Tutorial to perform DA with edgeR: https://bioconductor.org/books/3.14/OSCA.multisample/differential-abundance.html
  - https://www.sc-best-practices.org/preamble.html
- Tutorial about DGE: https://lcolladotor.github.io/cshl_rstats_genome_scale_2024/differential-gene-expression-analysis-overview.html

- Design matrices and contrasts for DGE: https://f1000research.com/articles/9-1444

PRACTICE TIME !