Batch effects in scRNA-seq data

Committee meeting

Almut Lütge
DMLS - University of Zürich
2021-02-26

Batch effects

Batch effects

Systematic differences between same type of cells [..][caused by technical sources] (Oslokov, 2019)

Differences between data sets [..][, that] occur due to **uncontrolled variability** in **experimental factors**, e.g., reagent quality, [.. and] can interfere with downstream analyses if not explicitly modelled. (Lun, 2019)

→ **Systematic unwanted variation**

New Results

Comment on this paper

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

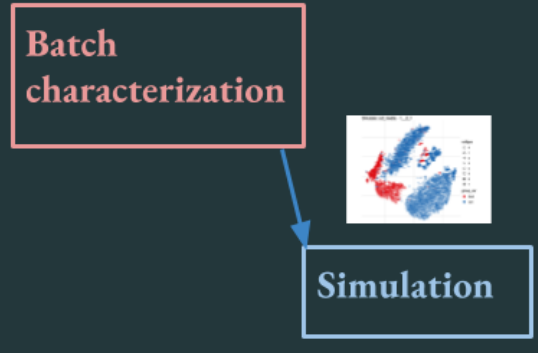
Almut Lütge, Joanna Zyprych-Walczak, Urszula Brykczynska Kunzmann, HelenaL Crowell, Daniela Calini, Dheeraj Malhotra, Charlotte Soneson, Mark D Robinson

doi: https://doi.org/10.1101/2020.12.11.420885

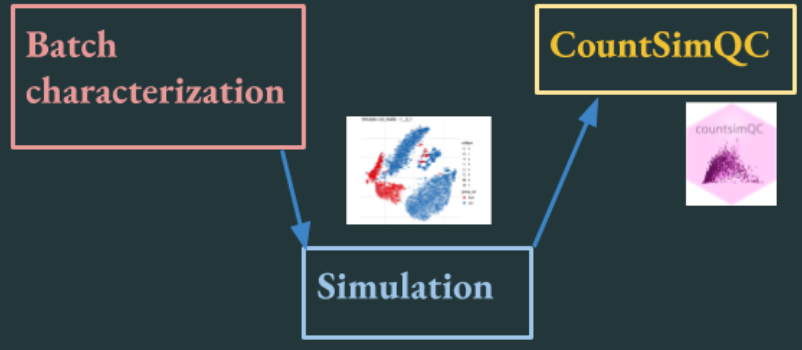
Project overview

Batch characterization

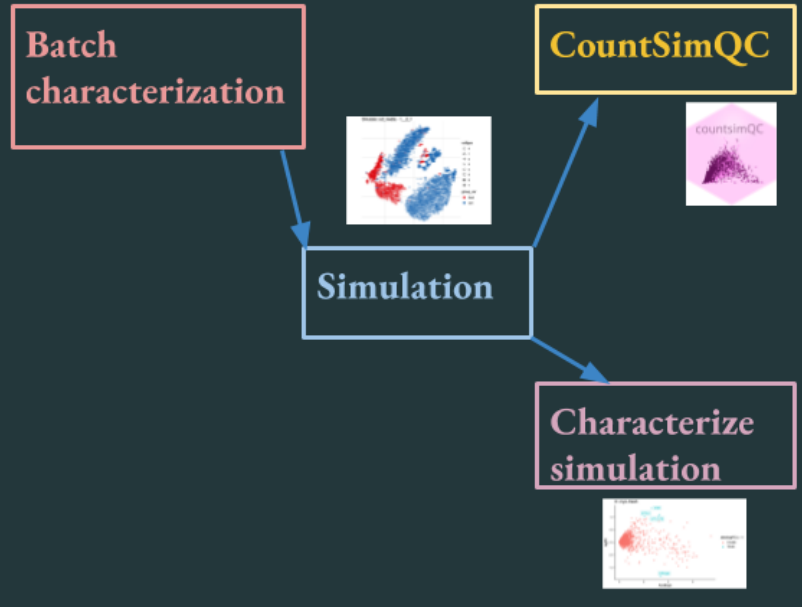
Project overview



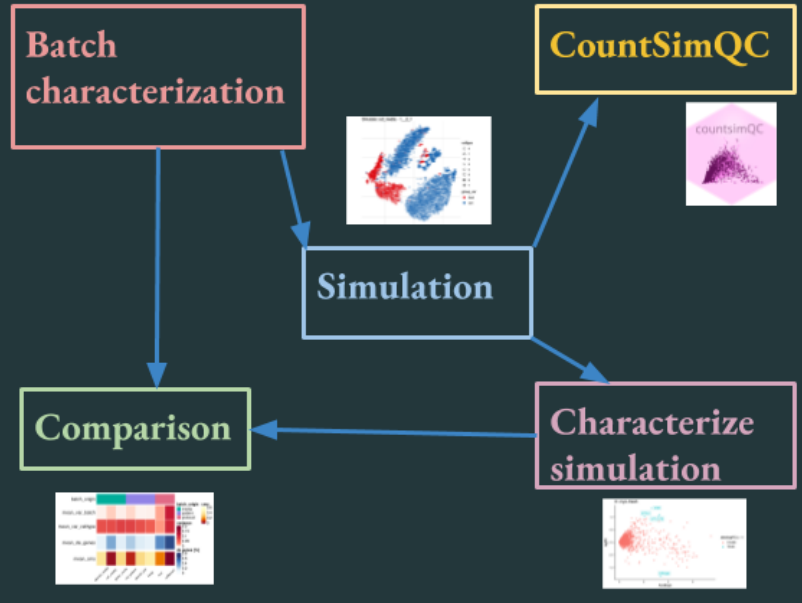
Project overview



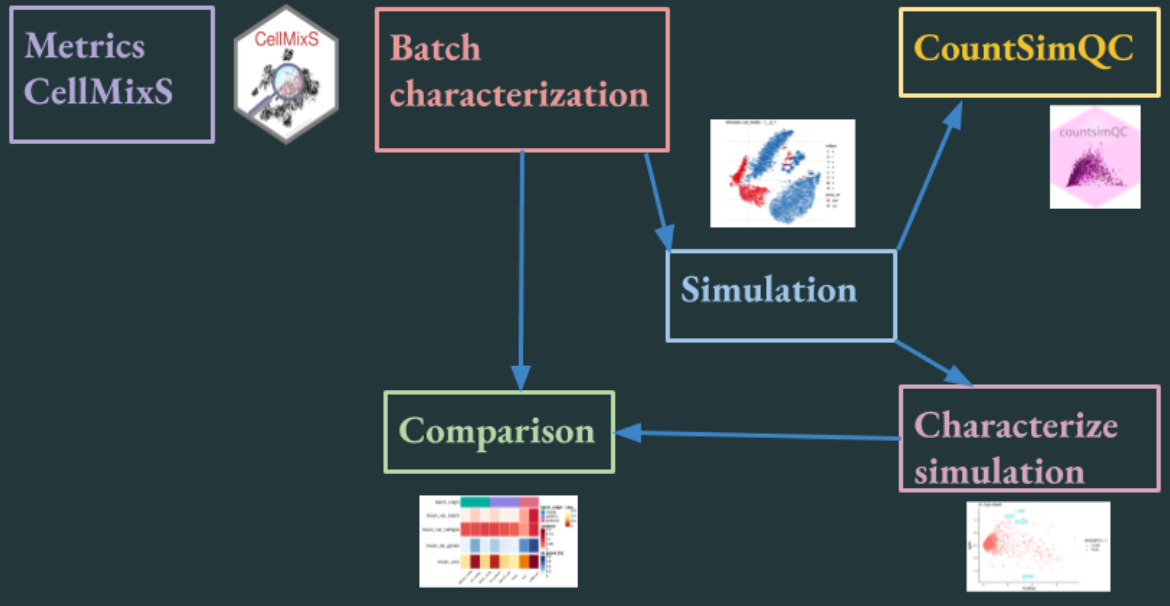
Project overview



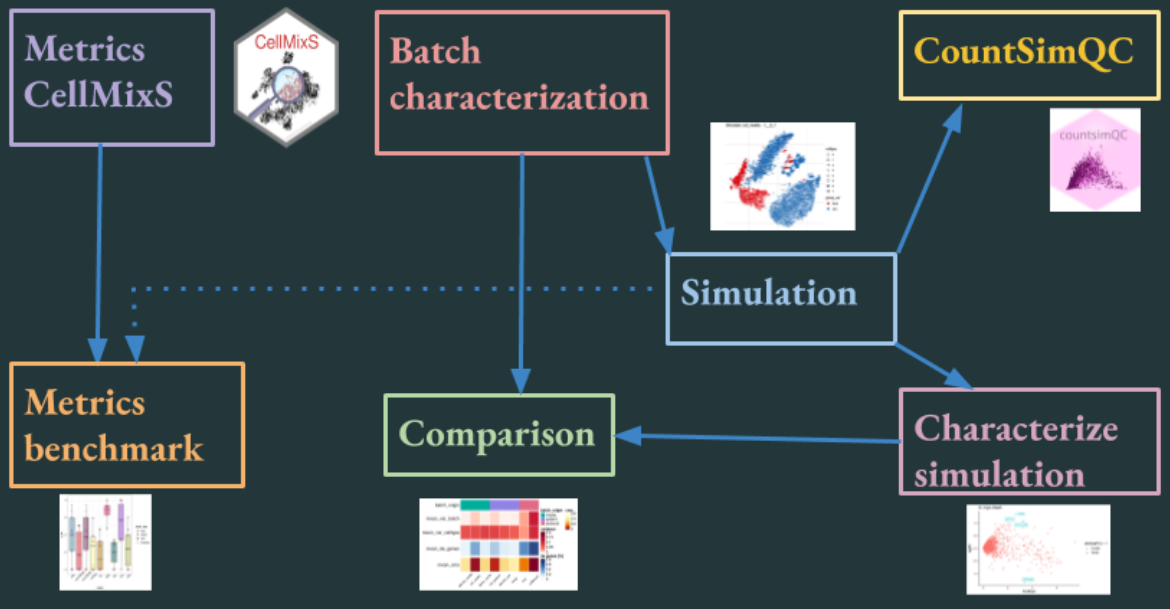
Project overview



Project overview



Project overview



Batch characterization

Batch characterization

- 7 datasets
- 9 batch effects
- *patient, protocol, storage*

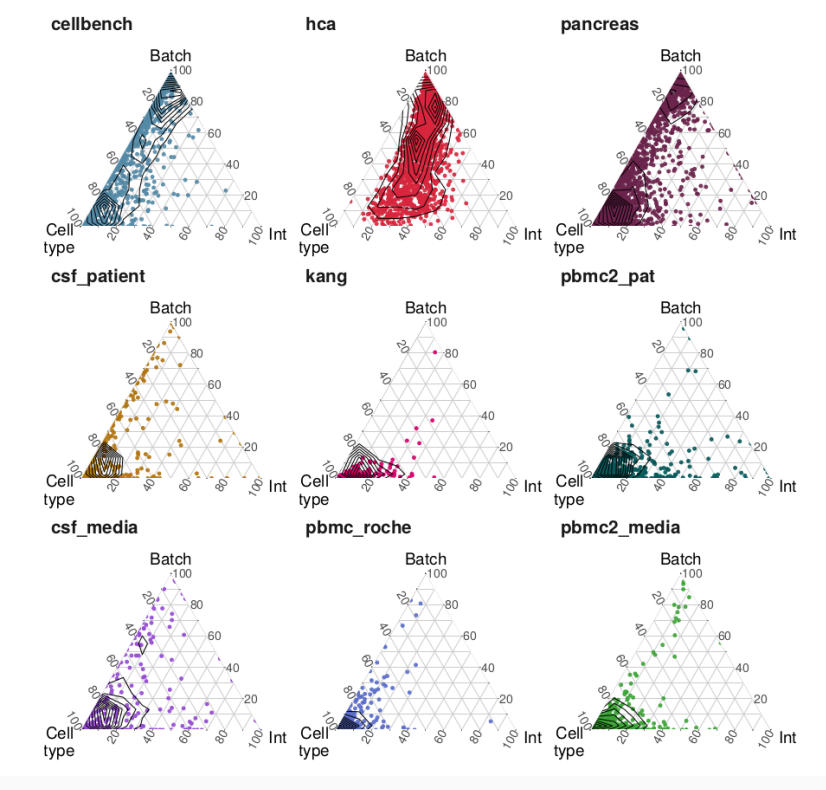
- **Variance partitioning**
- **logFC** distribution and correlation
- DE genes/overlap

Variance partition

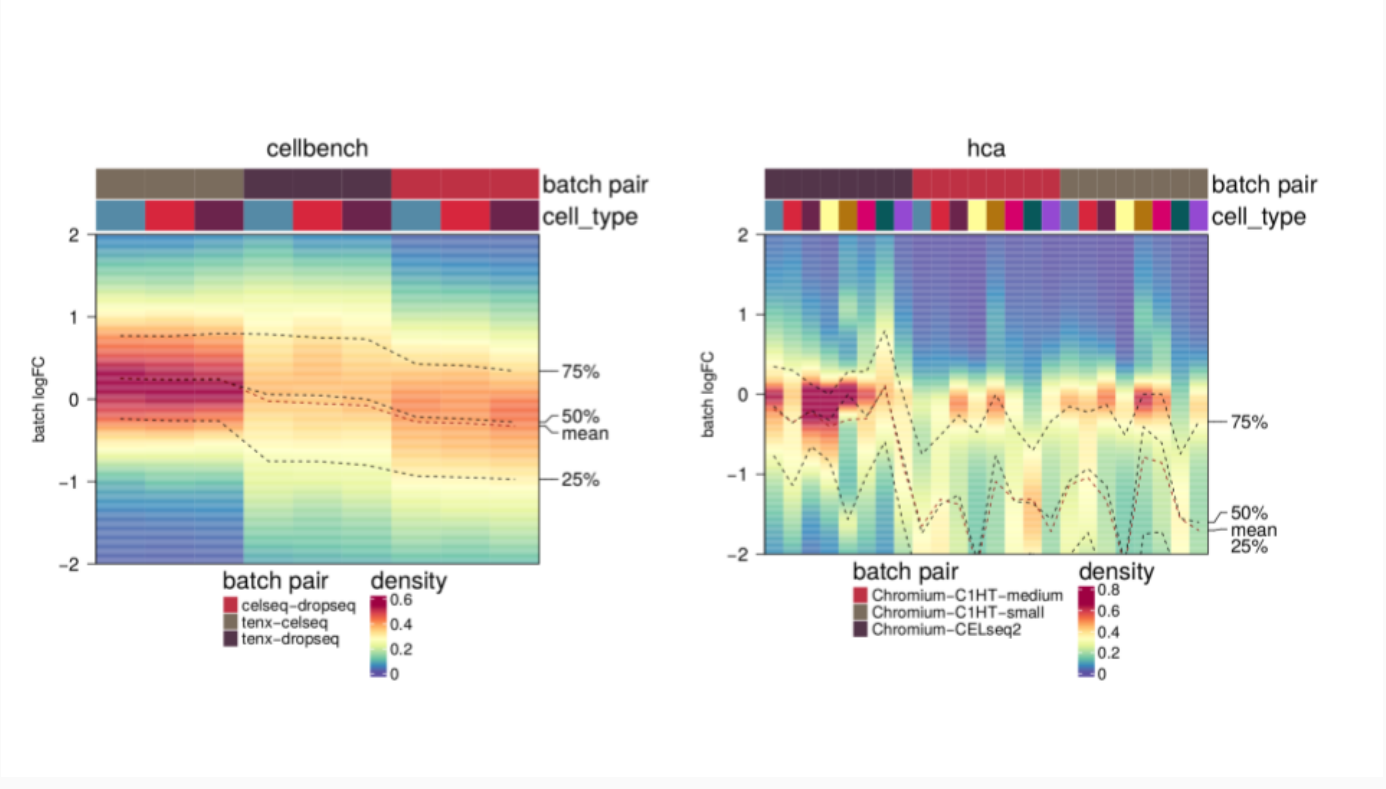
$$Y_g = \mu + X_p\alpha_{pg} + X_b\beta_{bg} + X_{p:b}\gamma_{(p:b)g} + \epsilon_g$$

Y_g: normalized expression gene g
μ: baseline expression
X_p, X_b, $X_{p:b}$: design matrices for the (random) cell types, batches and interactions effect
$\alpha_{pg} \sim N(0, \sigma^2_{pg})$,..: corresponding random effects
$\epsilon_i \sim N(0, \sigma^2_g)$ corresponding errors

Percent variance explained by ..:



Log fold change distributions



Metrics

Metrics

Cell level:

- Cellspecific Mixing Score (**cms**)
- Local Inverse Simpson Index (**lisi**)
- Mixing Metric (**mm**)
- Shannon's **entropy**

Metrics

Cell level:

- Cellspecific Mixing Score (**cms**)
- Local Inverse Simpson Index (**lisi**)
- Mixing Metric (**mm**)
- Shannon's **entropy**

Celltype level:

- k-nearest neighbour Batch Effect Test (**kBet**)
- Average silhouette width (**asw**)
- Graph connectivity (**graph**)

Metrics

Cell level:

- Cellspecific Mixing Score (**cms**)
- Local Inverse Simpson Index (**lisi**)
- Mixing Metric (**mm**)
- Shannon's **entropy**

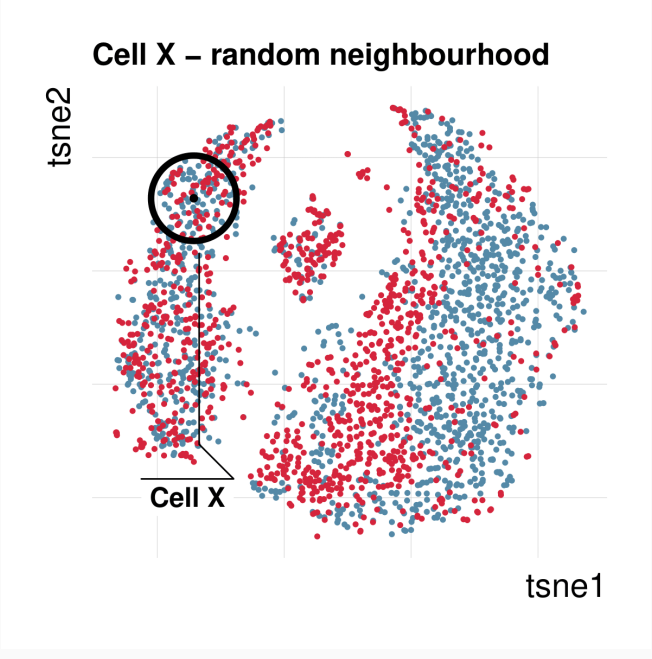
Global level:

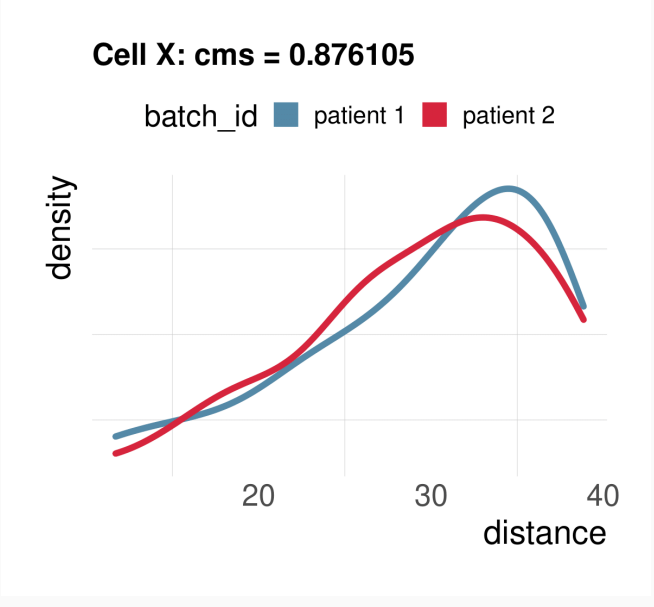
- Principal component regression (**pcr**)

Celltype level:

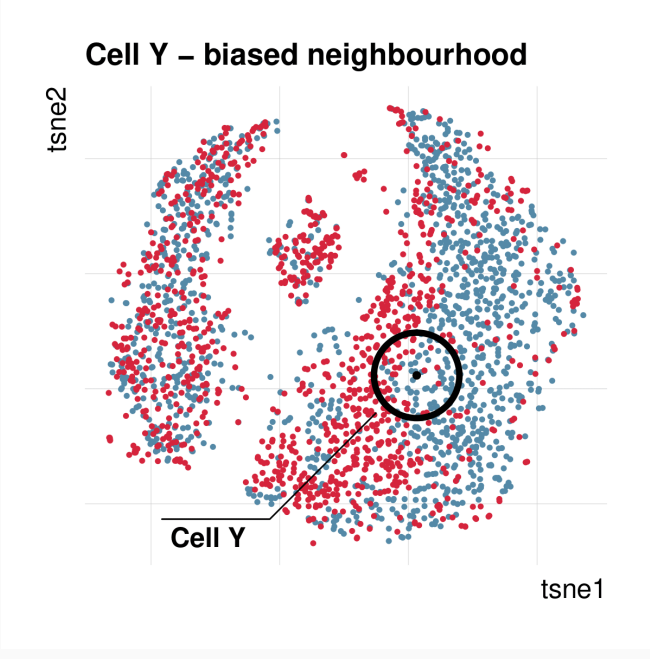
- k-nearest neighbour Batch Effect Test (**kBet**)
- Average silhouette width (**asw**)
- Graph connectivity (**graph**)

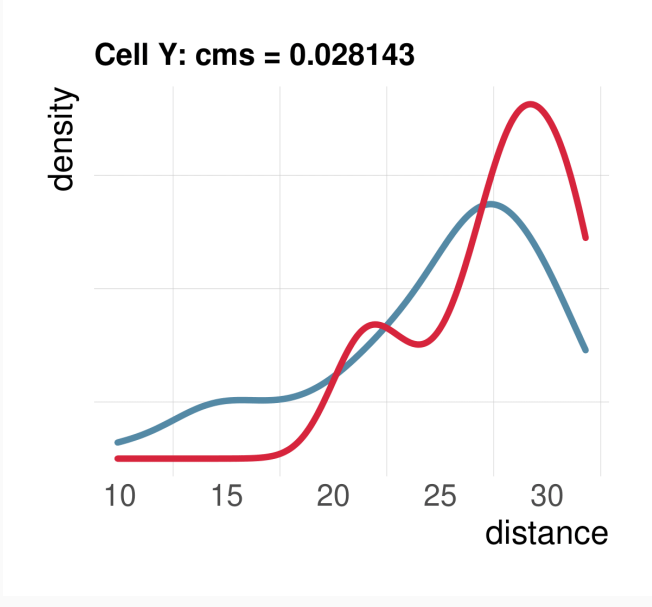
Cell-specific Mixing Score (cms)





Cell-specific Mixing Score (cms)





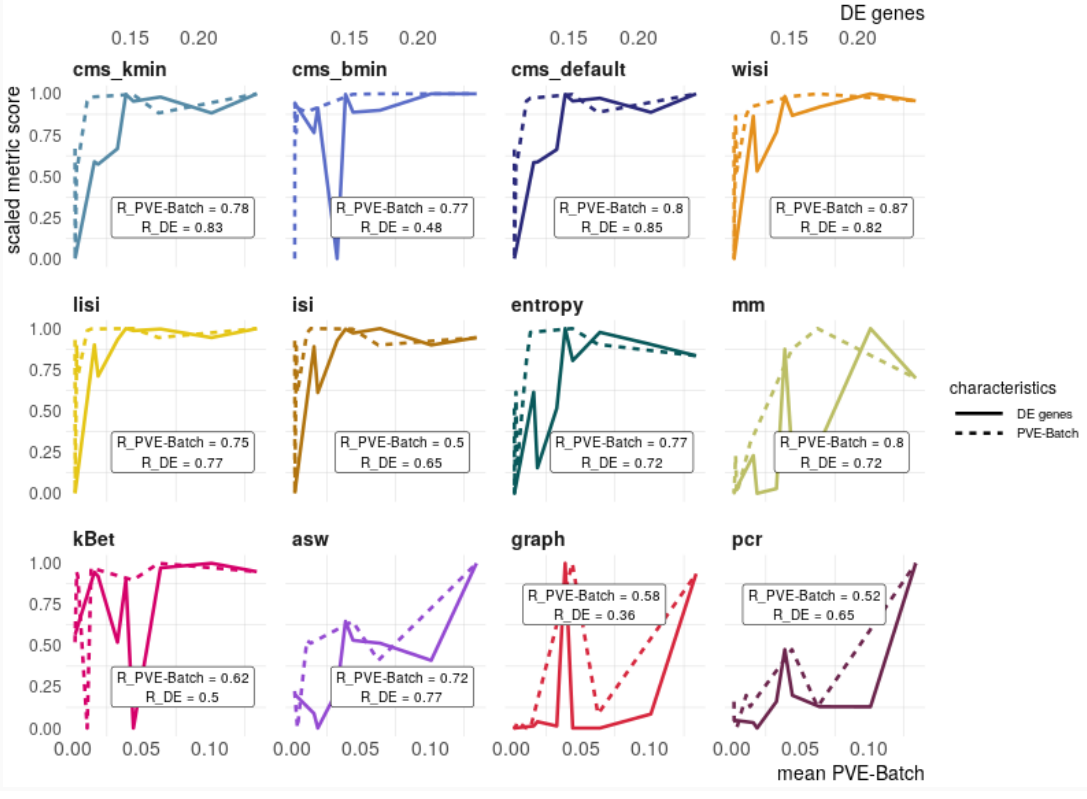
Benchmark

Task 1: Batch characteristics

Aim: Test whether metrics reflect batch strength across datasets

Spearman correlation of metrics with surrogates of batch strength (e.g., PVE-Batch and proportion of DE genes between batches) across datasets

Batch characteristics

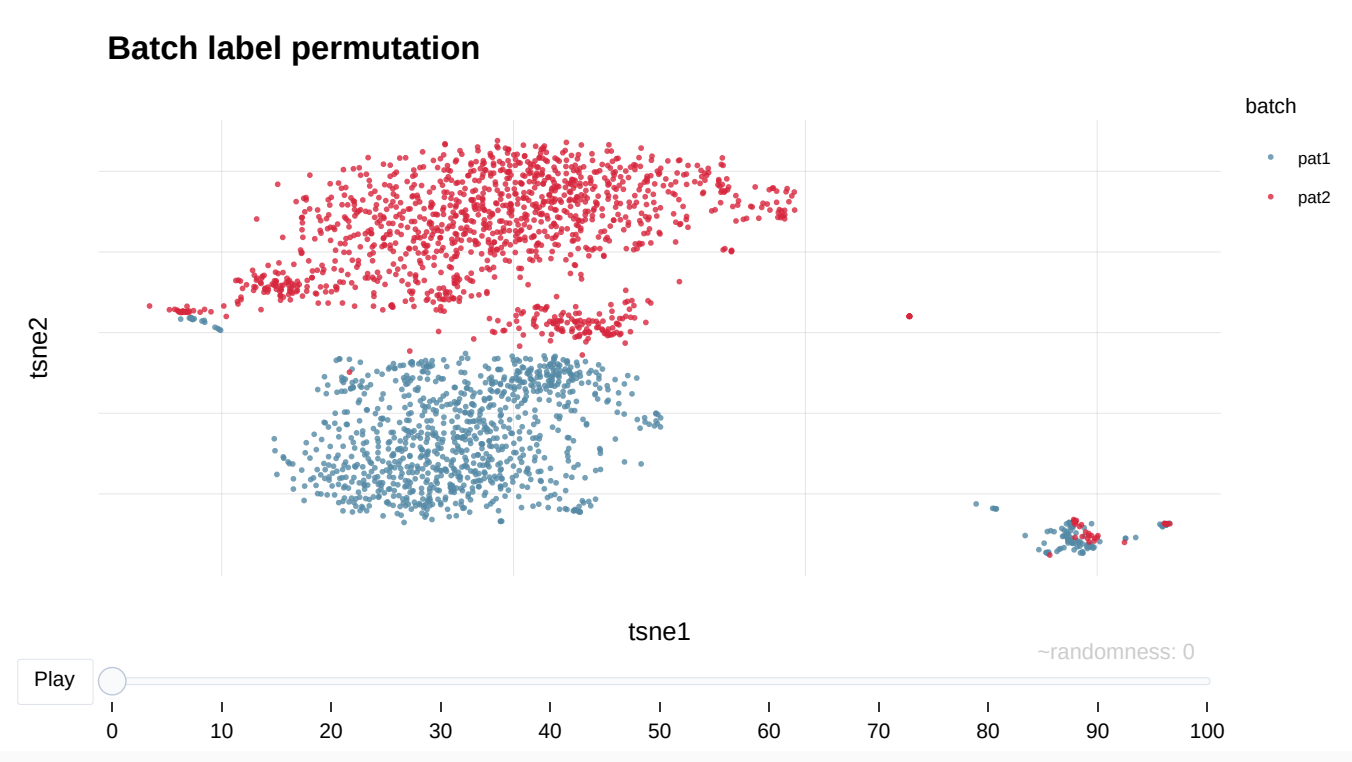


Task 2: Batch label permutation

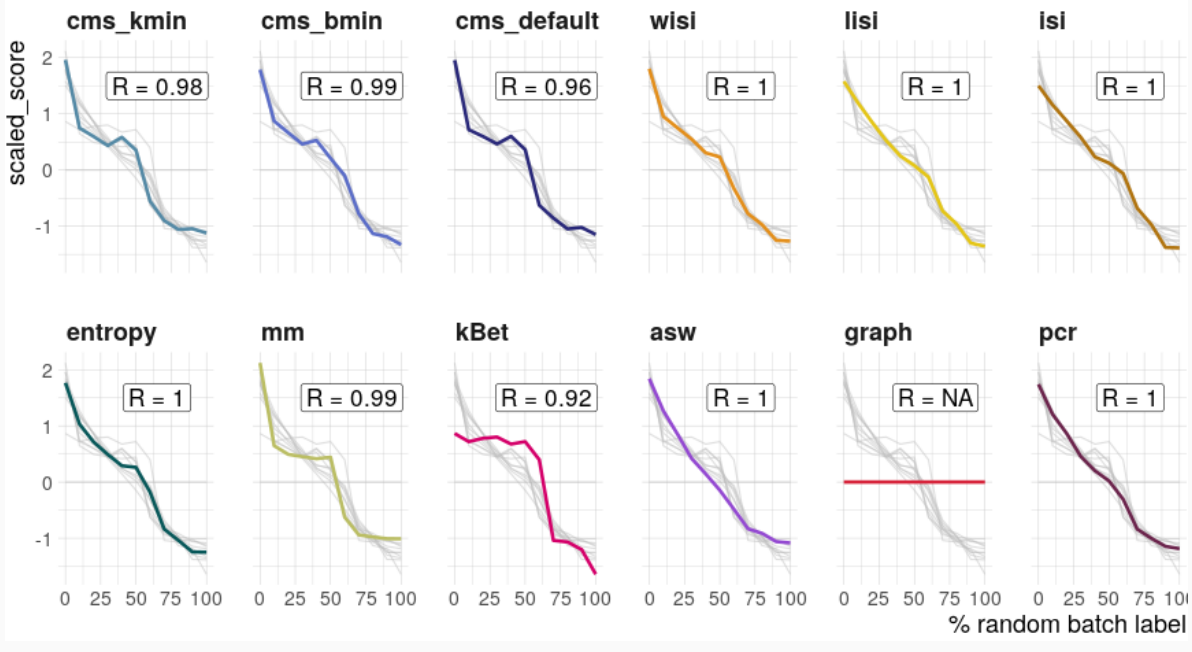
Aim: Negative control and test whether metrics scale with randomness

Spearman correlation of metrics with the percentage of randomly permuted batch label

Batch label permutation



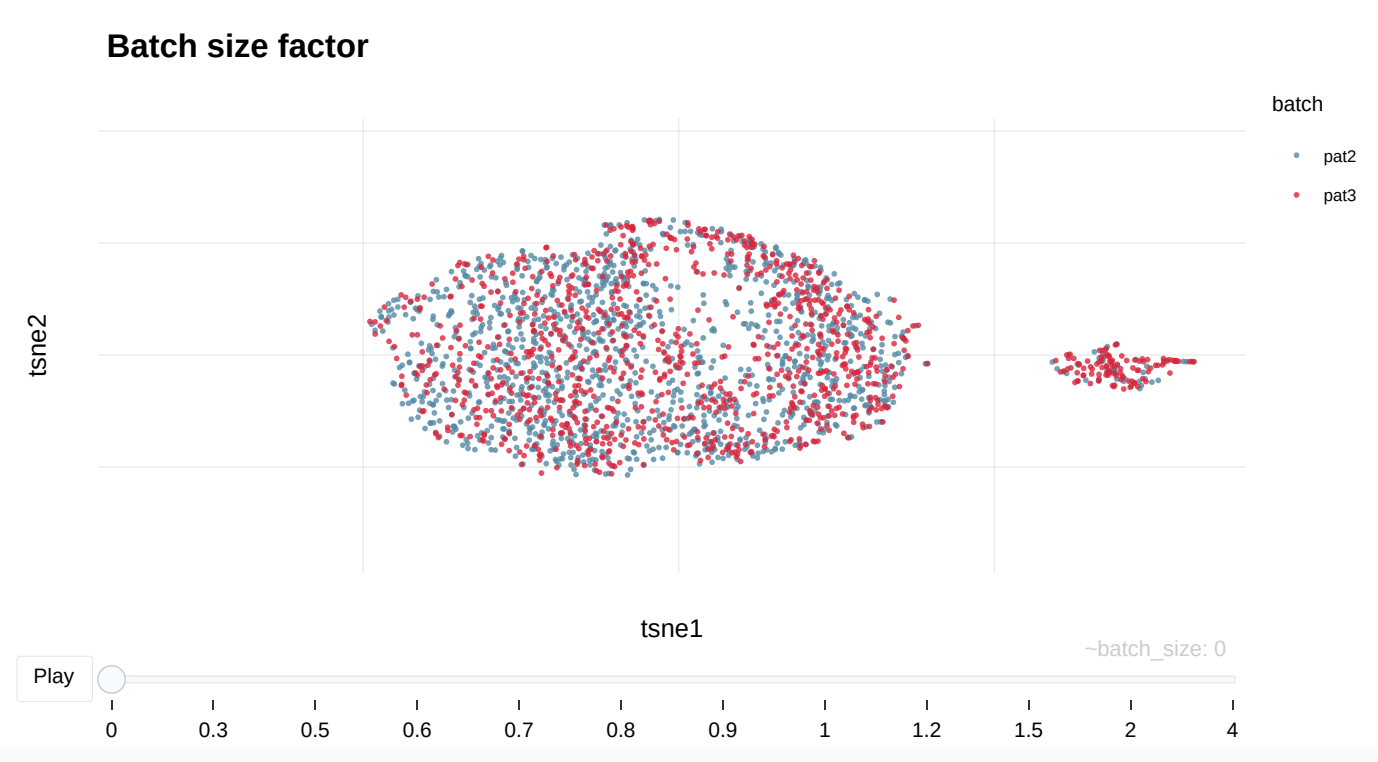
Batch label permutation



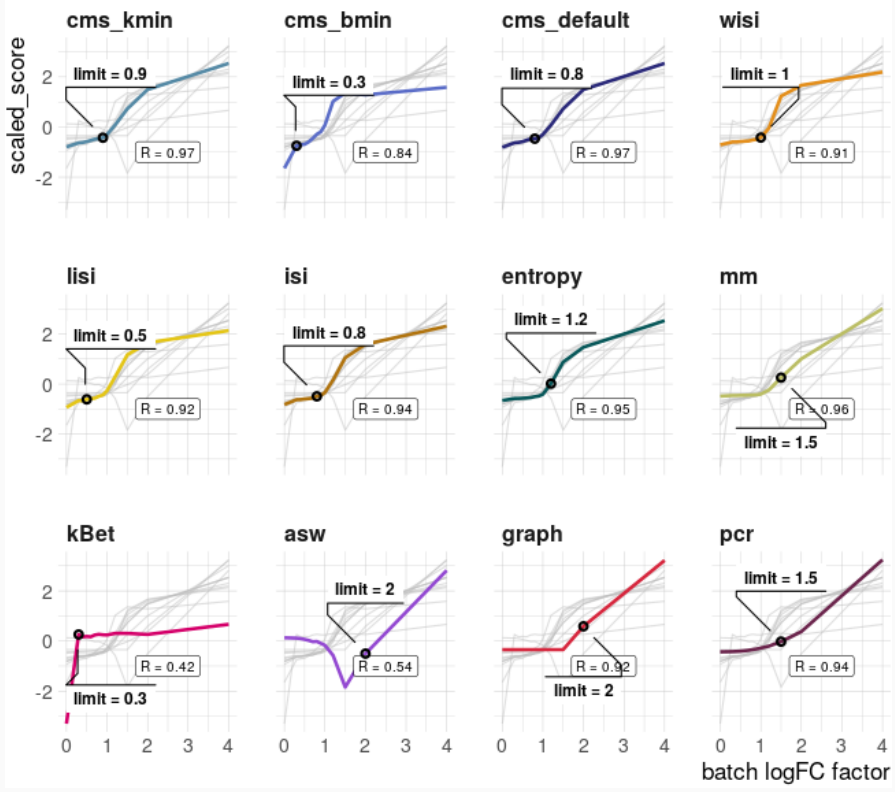
Task 3: Scaling and detection limits

Aim: Test whether metrics scale with (synthetic) batch strength; Estimate lower limit of batch detection

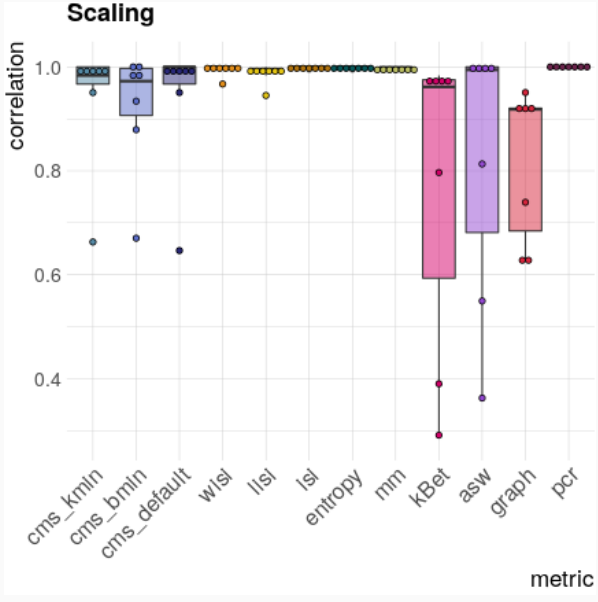
Spearman correlation of metrics with the batch logFC in simulation series on the same dataset; Minimal batch logFC that is recognized from the metrics as batch effect

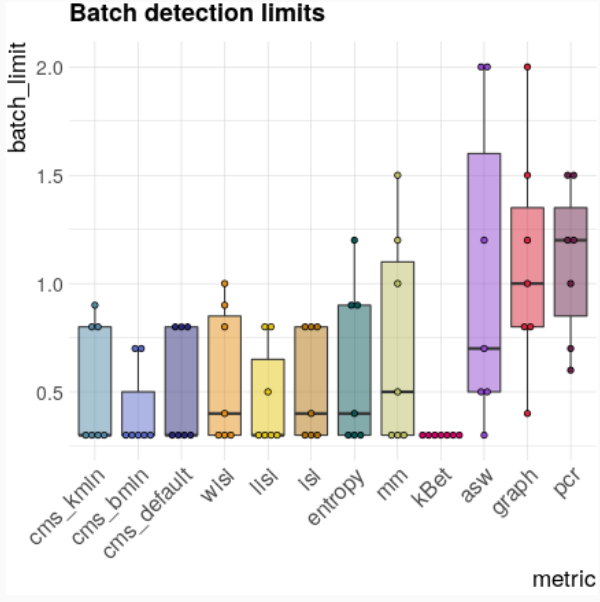

Batch size factor
batch
pat2
pat3
tsne2
tsne1
~batch_size: 0
Play
0
0.3
0.5
0.6
0.7
0.8
0.9
1
1.2
1.5
2
4

Scaling and Sensitivity



Scaling and Sensitivity

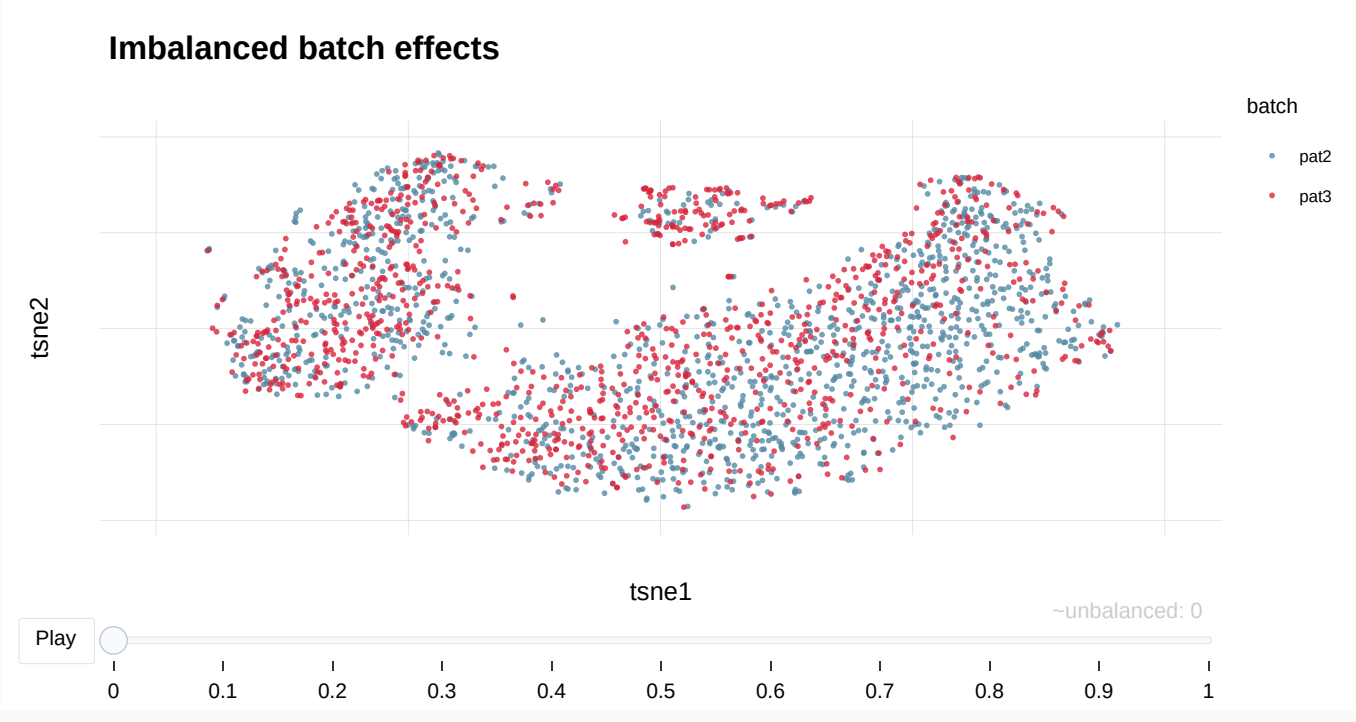




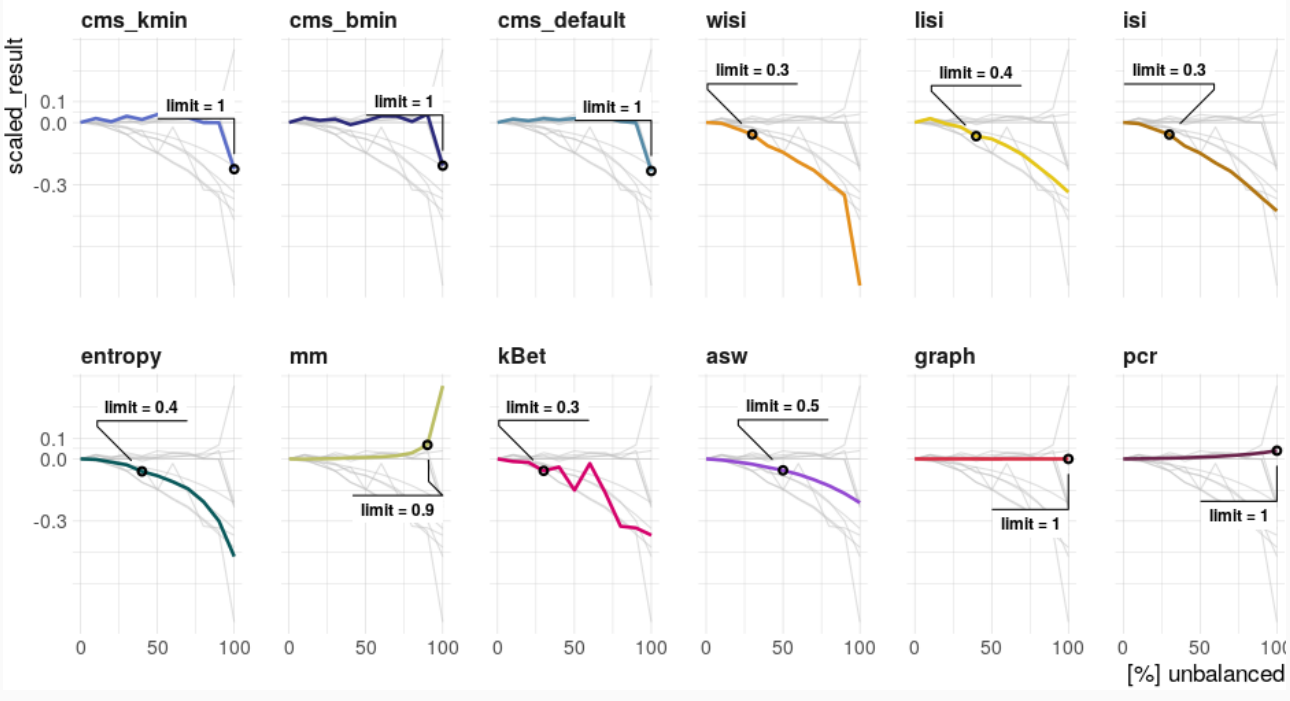
Task 4: Imbalanced batches

Aim: Reaction of metrics to imbalance cell type abundance within the same dataset

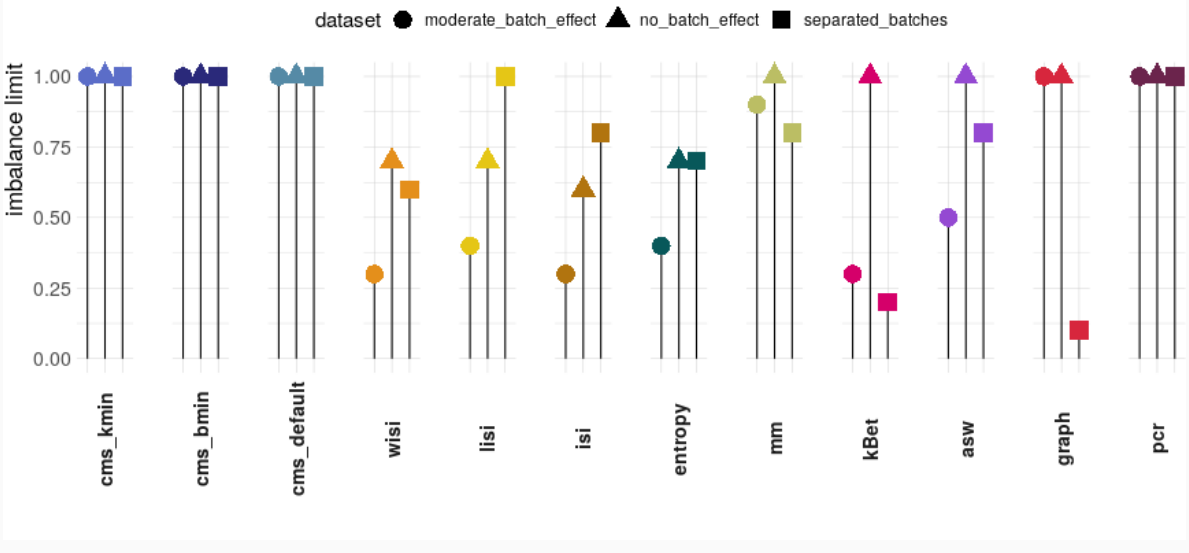
Test sensitivity towards imbalance of cell type abundance


Imbalanced batch effects
batch
pat2
pat3
tsne2
tsne1
~unbalanced: 0
Play
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Imbalanced batch effects

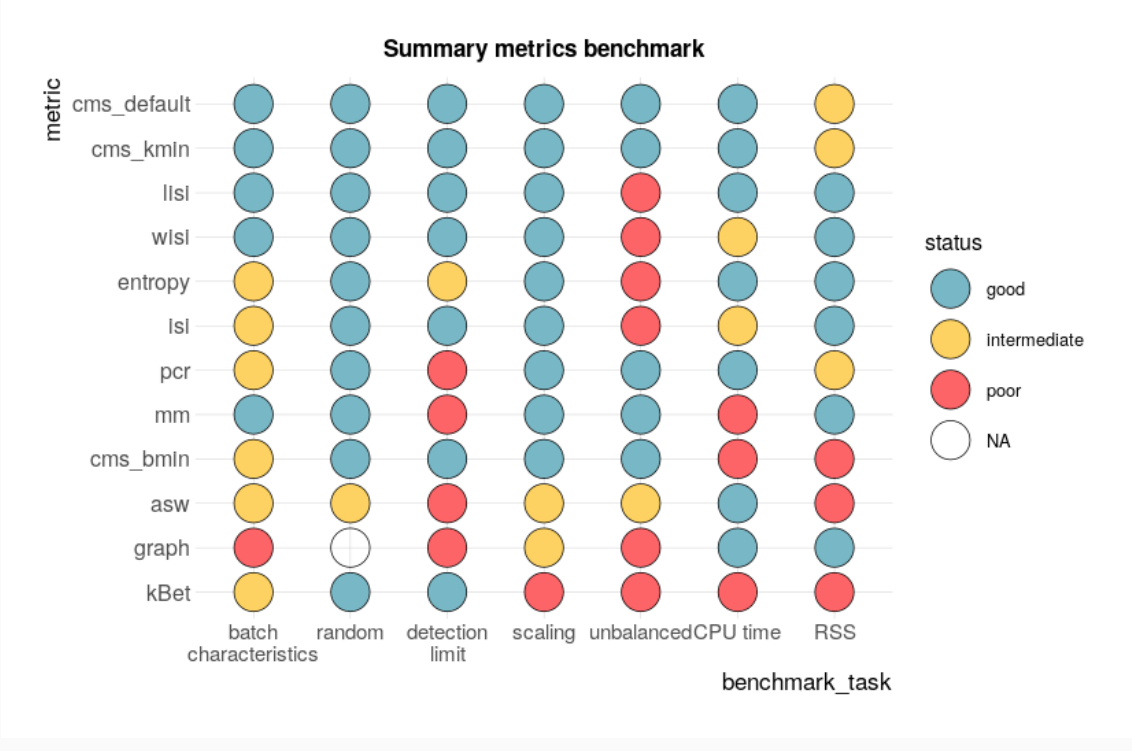


Imbalanced batch effects



Summary

Summary


Summary metrics benchmark
metric
cms_default
cms_kmln
lisi
wisi
entropy
lsi
pcr
mm
cms_bmln
asw
graph
kBet
batch characteristics
random
detection limit
scaling
unbalanced
CPU time
RSS
benchmark_task
status
good
intermediate
poor
NA

2. *Omni_batch*: open and continuous benchmarking of single cell batch correction methods

State of the art

Research | Open Access | Published: 16 January 2020

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

Hoa Thi Nhu Tran, Kok Siong Ang, Marion Chevrier, Xiaomeng Zhang, Nicole Yee Shin Lee, Michelle Goh & Jinmiao Chen

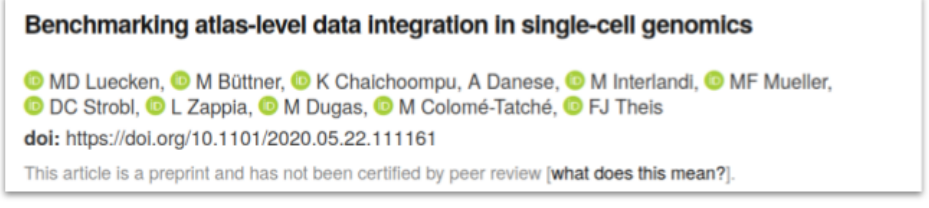
Genome Biology **21**, Article number: 12 (2020) | Cite this article

Flexible comparison of batch correction methods for single-cell RNA-seq using BatchBench

Ruben Chazarra-Gil, Stijn van Dongen, Vladimir Yu Kiselev, Martin Hemberg

Nucleic Acids Research, gkab004, https://doi.org/10.1093/nar/gkab004

Published: 01 February 2021 **Article history**



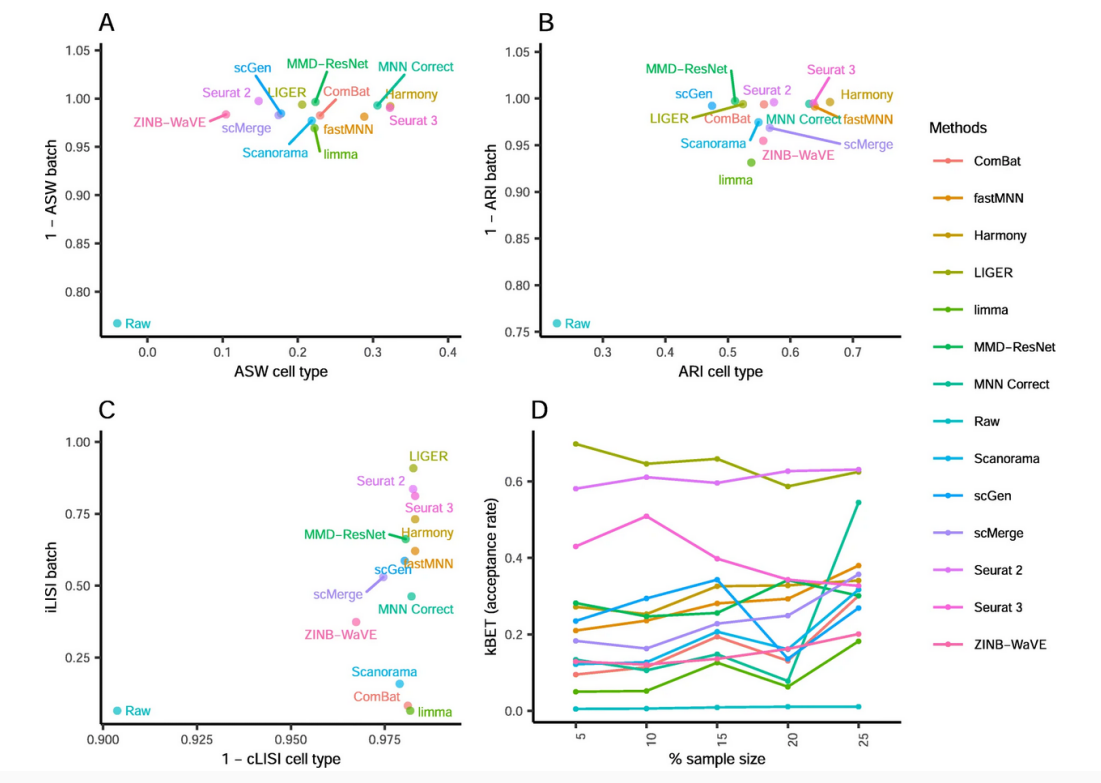
Benchmarking atlas-level data integration in single-cell genomics

MD Luecken, M Büttner, K Chaichoompu, A Danese, M Interlandi, MF Mueller, DC Strobl, L Zappia, M Dugas, M Colomé-Tatché, FJ Theis

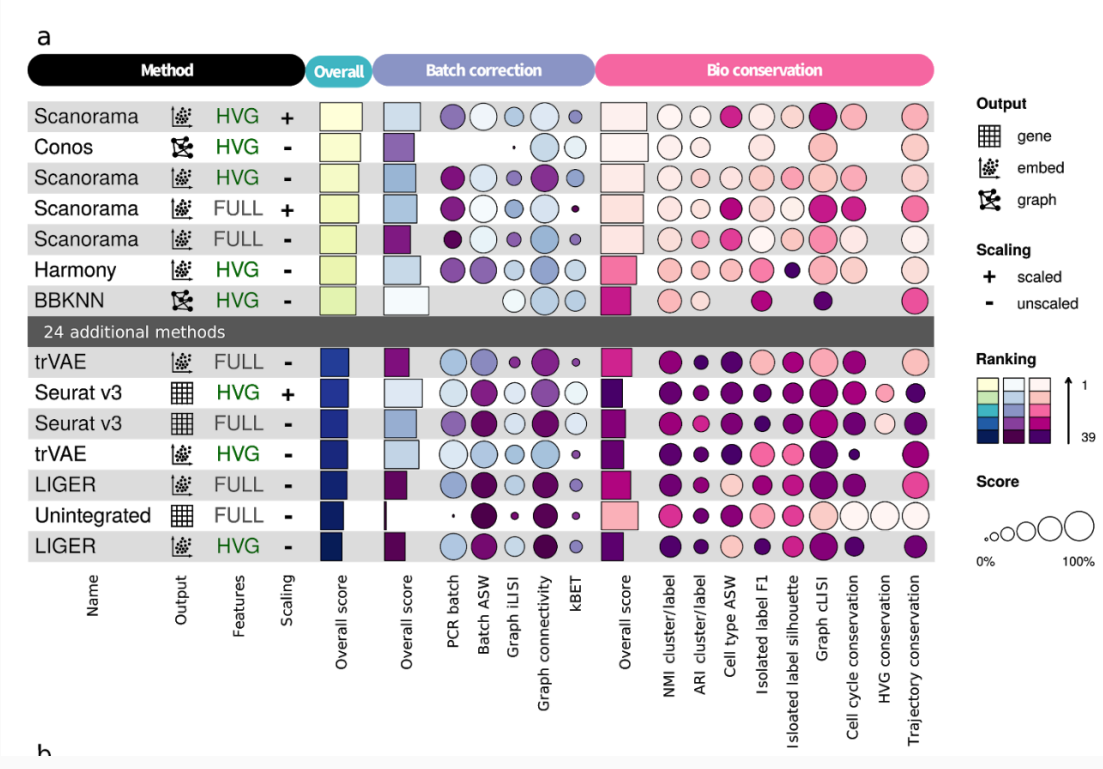
doi: https://doi.org/10.1101/2020.05.22.111161

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

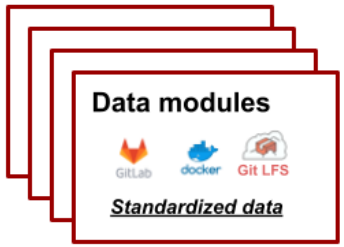
Benchmark A


A
1 – ASW batch
ASW cell type
scGen
MMD–ResNet
MNN Correct
Seurat 2
LIGER
ComBat
Harmony
ZINB–WaVE
scMerge
fastMNN
Seurat 3
Scanorama
limma
Raw
B
1 – ARI batch
ARI cell type
C
iLISI batch
1 – cLISI cell type
D
kBET (acceptance rate)
% sample size
Methods
ComBat
fastMNN
Harmony
LIGER
limma
MMD–ResNet
MNN Correct
Raw
Scanorama
scGen
scMerge
Seurat 2
Seurat 3
ZINB–WaVE

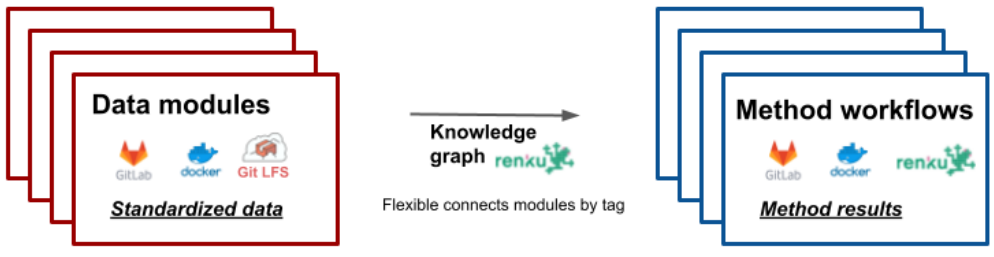
Benchmark B


a
Method
Overall
Batch correction
Bio conservation
Scanorama HVG +
Conos HVG -
Scanorama HVG -
Scanorama FULL +
Scanorama FULL -
Harmony HVG -
BBKNN HVG -
24 additional methods
trVAE FULL -
Seurat v3 HVG +
Seurat v3 FULL -
trVAE HVG -
LIGER FULL -
Unintegrated FULL -
LIGER HVG -
Name
Output
Features
Scaling
Overall score
Overall score
PCR batch
Batch ASW
Graph iLISI
Graph connectivity
kBET
Overall score
NMI cluster/label
ARI cluster/label
Cell type ASW
Isolated label F1
Isloated label silhouette
Graph cLISI
Cell cycle conservation
HVG conservation
Trajectory conservation
Output
gene
embed
graph
Scaling
+ scaled
- unscaled
Ranking
1
39
Score
0%
100%
b

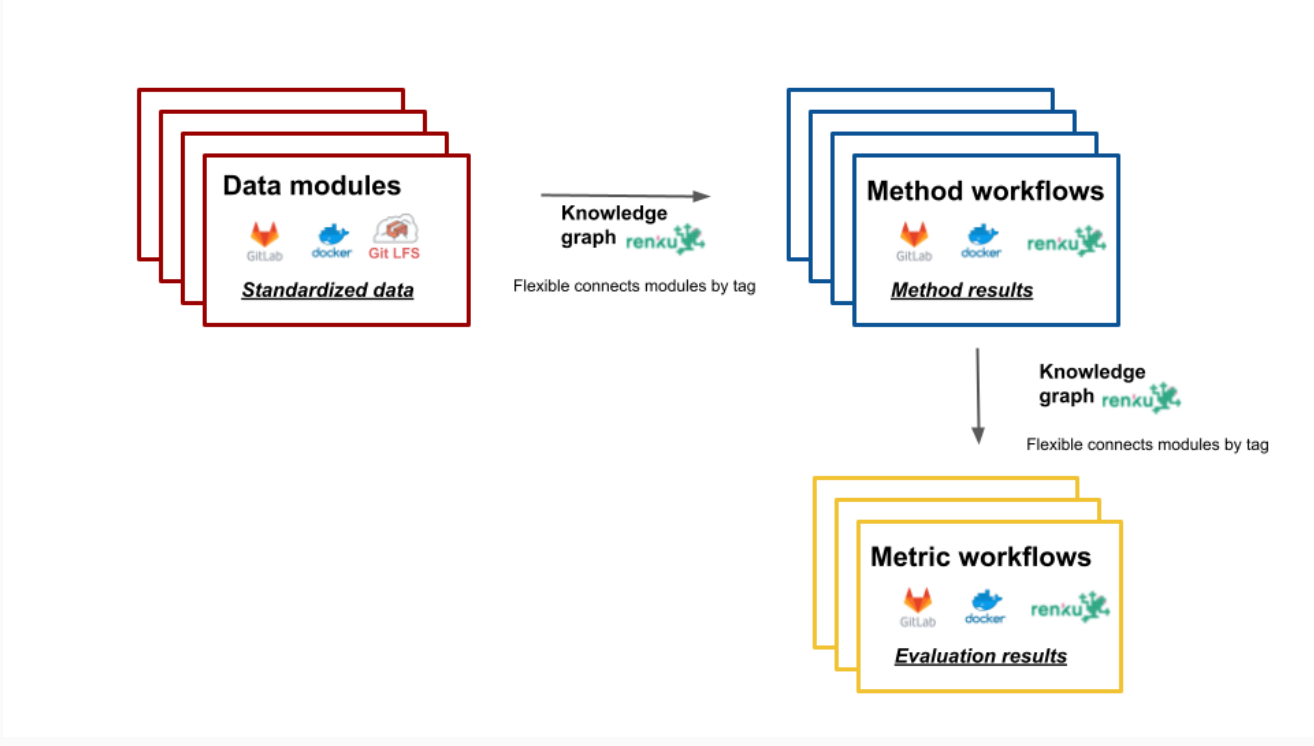
Omni-benchmark



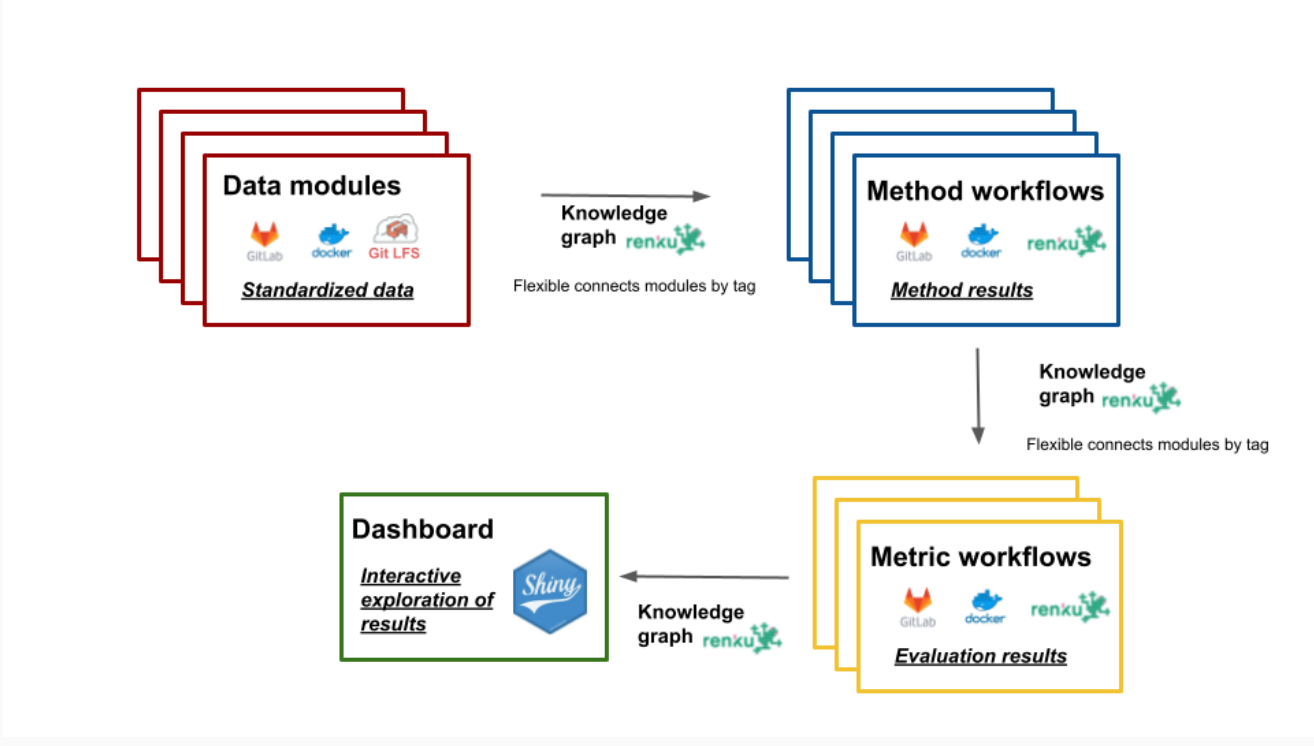
Omni-benchmark



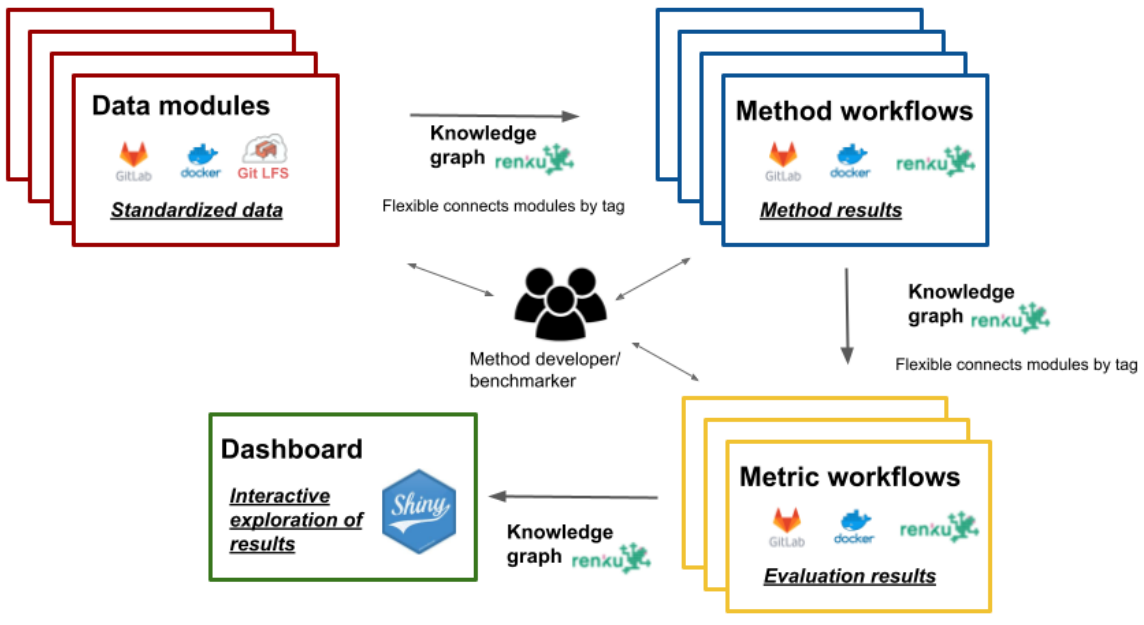
Omni-benchmark



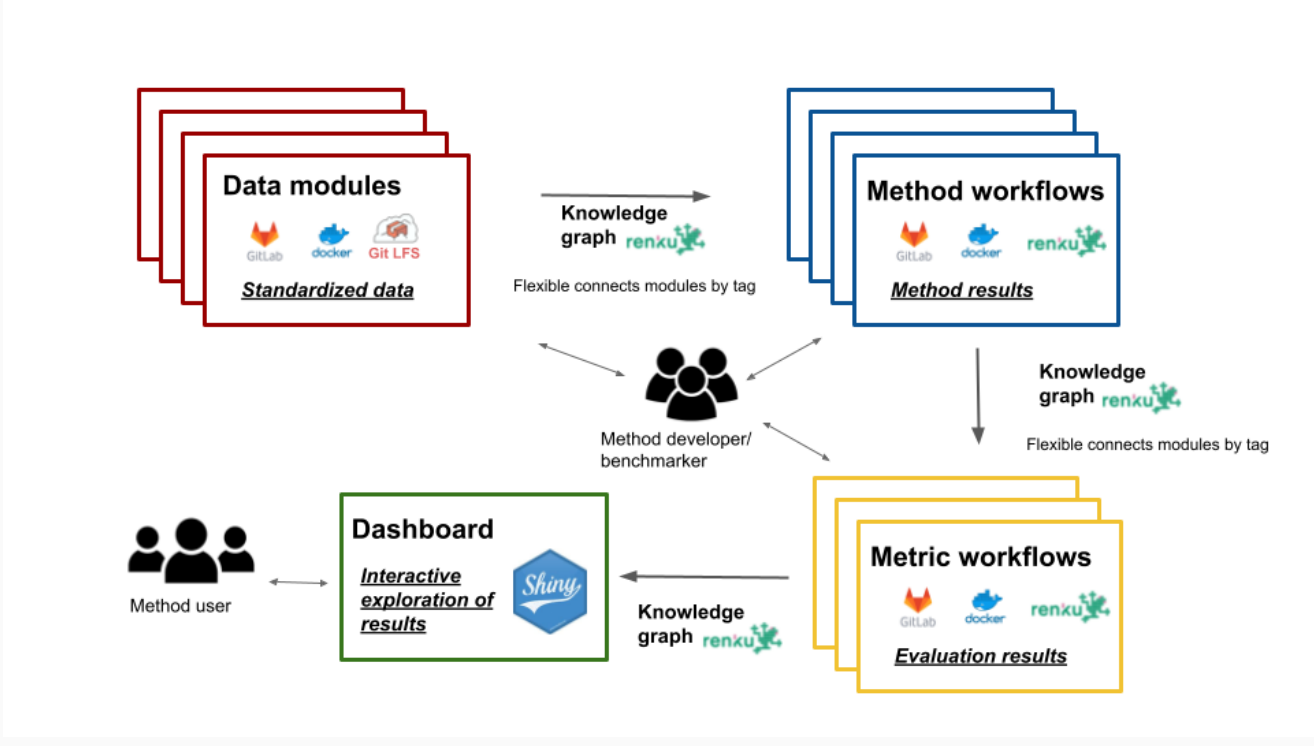
Omni-benchmark



Omni-benchmark



Omni-benchmark



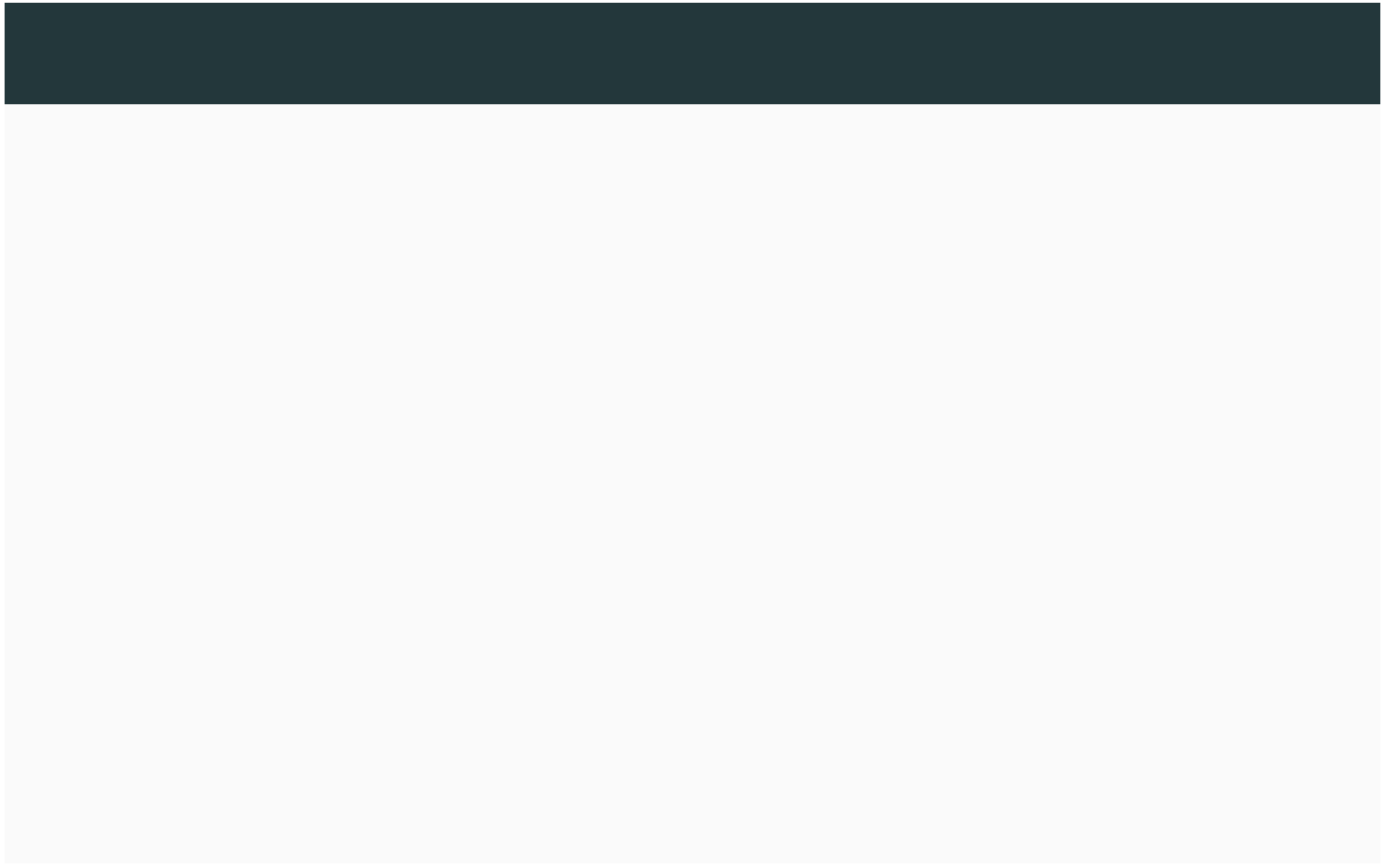
3. Teaching and course work

Teaching and course work

Teaching: BIO 134 (170 hours), R course ✔

Course work: 8/12 credits

Missing: Method seminar (1 credits)
Transferable skills courses (2 credits)
Pre-doctoral exam (1 credit)



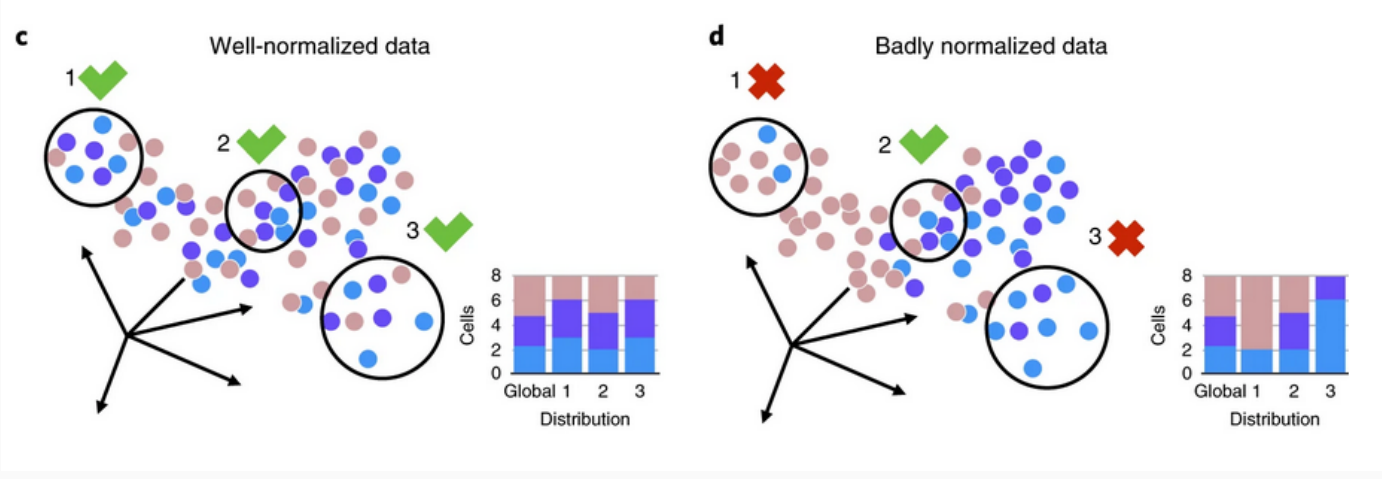
Extra slides

Local Inverse Simpson Index (lisi)

- neighborhood diversity
- effective number of batches
- neighbor weighting:
  - euclidean distance --> wisi
  - no weighting --> isi
  - Gaussian kernel based weighting --> lisi

$$\frac{1}{\sum_{b=1}^{B} p(b)}$$

k-nearest neighbour batch effect test (kBet)



(Buttner et. al., 2019)

Mixing metric (mm)