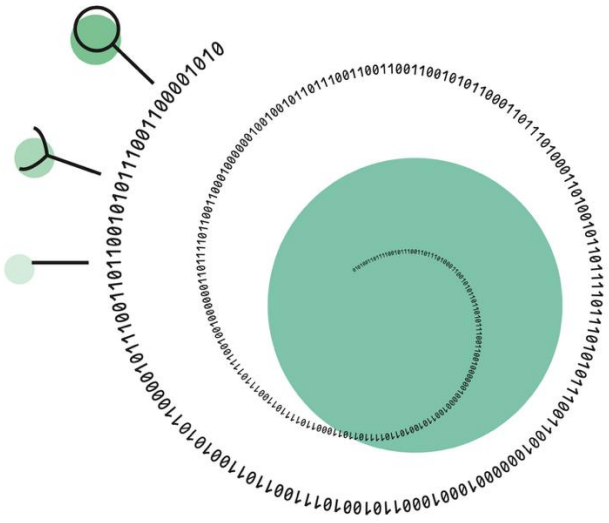
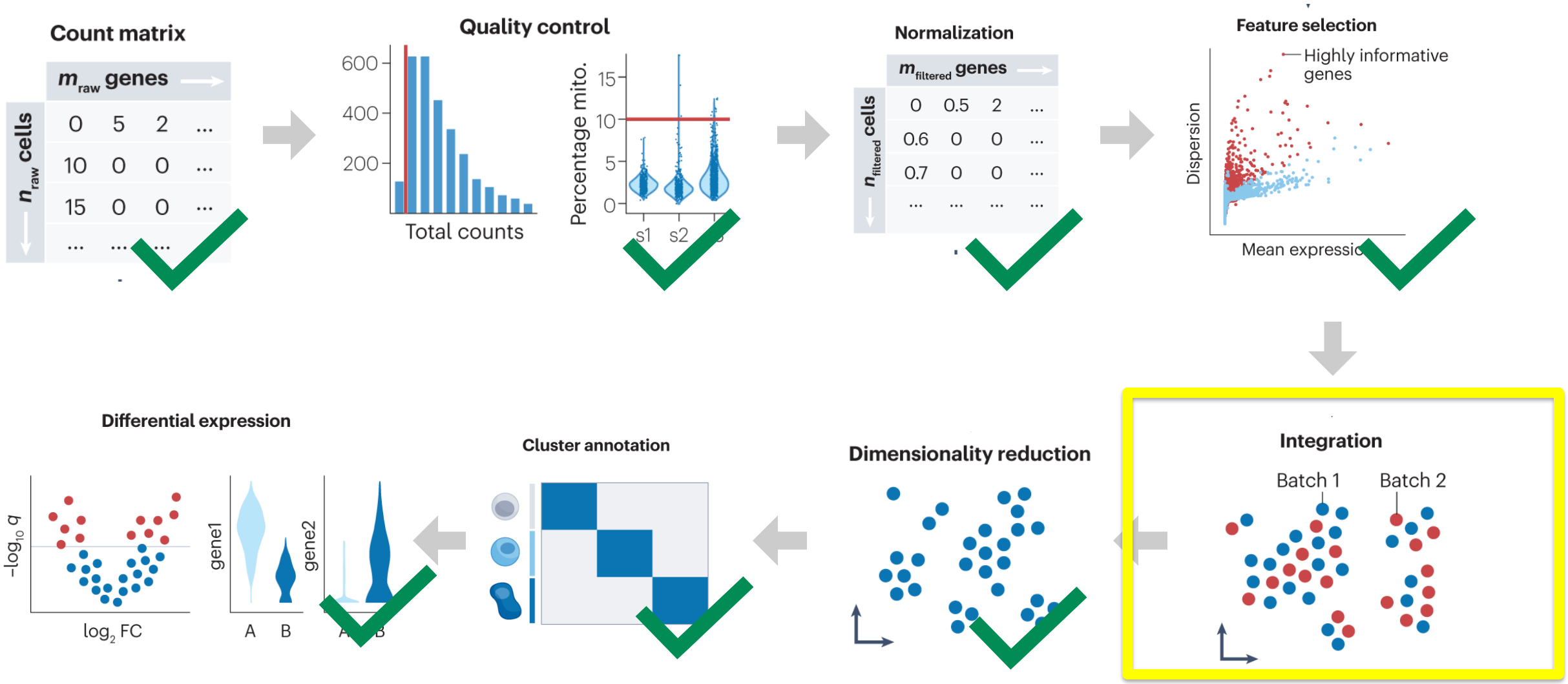


Single Dell Data Analysis Course

Batch correction algorithms for dataset combination (integration)

Lisa Buchauer
Professor of Systems Biology of Infectious Diseases
Department of Infectious Diseases and Intensive Care
Charité - Universitätsmedizin Berlin

Today



Heumos, L., Schaar, A.C., Lance, C. et al. Best practices for single-cell analysis across modalities. Nat Rev Genet 24, 550–572 (2023). https://doi.org/10.1038/s41576-023-00586-w

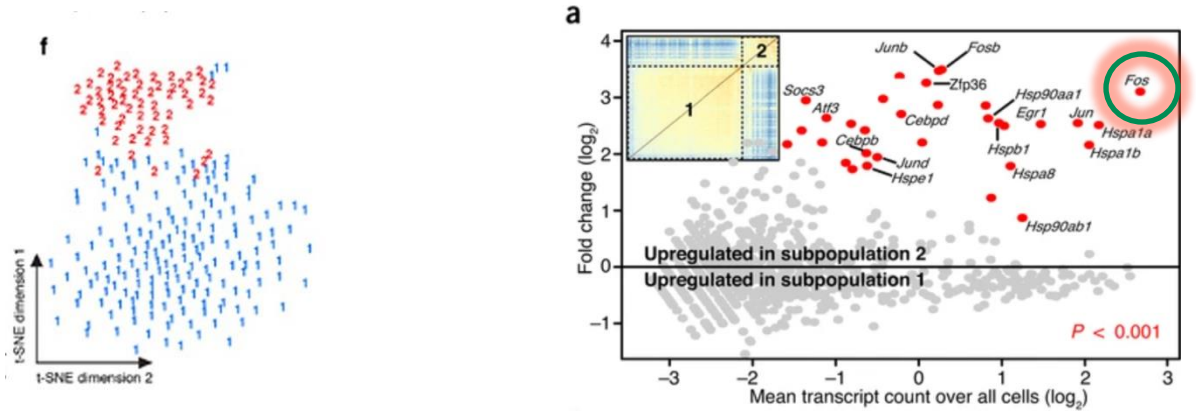
To the Editor:

In many gene expression studies, cells are extracted by tissue dissociation and fluorescence-activated cell sorting (FACS), but the effect of these protocols on cellular transcriptomes is not well characterized and is often ignored. Here, we applied single-cell mRNA sequencing (scRNA-seq) to muscle stem cells, and we found a subpopulation that is strongly affected by the widely used dissociation protocol that we employed. One implication of this finding is that several published transcriptomics studies may need to be reinterpreted. Importantly, we detected similar subpopulations in other single-cell data sets, suggesting that cells from other tissues may be affected by this artifact as well.

van den Brink, S., Sage, F., Vértesy, Á. et al. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. Nat Methods 14, 935–936 (2017). https://doi.org/10.1038/nmeth.4437

van den Brink example: dissociation protocol induces heterogeneity into single cell data sets.

scRNA seq of dissociated, FACS'ed skeletal muscle cells

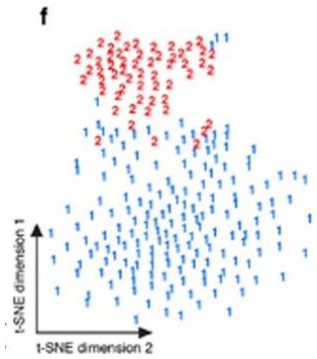


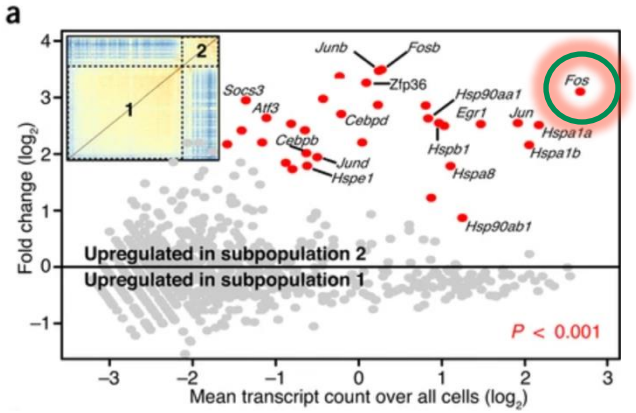
Two populations of satellite cells!!

van den Brink, S., Sage, F., Vértesy, Á. et al. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. Nat Methods 14, 935–936 (2017). https://doi.org/10.1038/nmeth.4437

van den Brink example: dissociation protocol induces heterogeneity into single cell data sets.

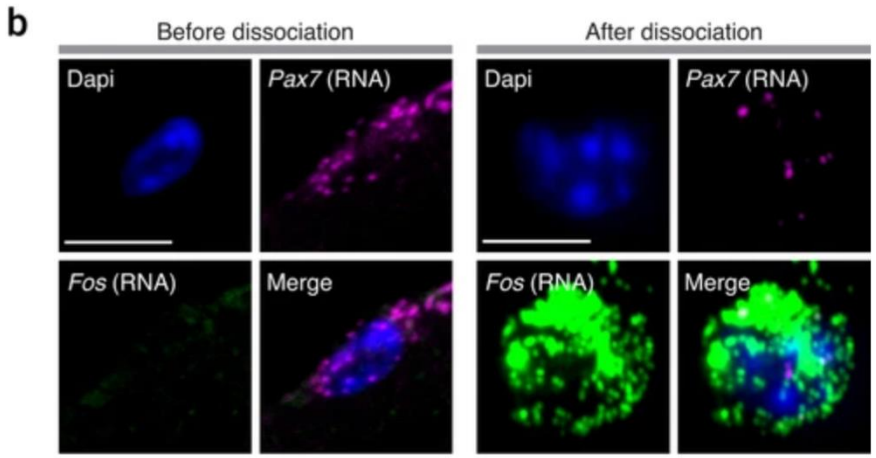
scRNA seq of dissociated, FACS'ed skeletal muscle cells





Two populations of satellite cells!!

(de)validation via smFISH: heterogeneous *Fos* expression is caused by isolation protocol



van den Brink, S., Sage, F., Vértesy, Á. et al. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. Nat Methods 14, 935–936 (2017). https://doi.org/10.1038/nmeth.4437

What are “batch effects” in the context of single-cell transcriptomics data?

Differences in handling and processing for different groups of cells can introduce differences in measured gene expression levels.

Storage differences

- fresh vs frozen tissue
- Freezer temperatures
- Storage media
- …

What are “batch effects” in the context of single-cell transcriptomics data?

Differences in handling and processing for different groups of cells can introduce differences in measured gene expression levels.

Storage differences

- fresh vs frozen tissue
- Freezer temperatures
- Storage media
- …

Experimental processing differences

- Preprocessing
- Technology platforms
- Reagent lots
- Timing
- Personnel
- …

What are “batch effects” in the context of single-cell transcriptomics data?

Differences in handling and processing for different groups of cells can introduce differences in measured gene expression levels.

Storage differences

- fresh vs frozen tissue
- Freezer temperatures
- Storage media
- …

Experimental processing differences

- Preprocessing
- Technology platforms
- Reagent lots
- Timing
- Personnel
- …

Bioinformatical processing differences

- Transcriptome versions
- Alignment software and parameters
- Post-alignment QC filtering
- …

What are “batch effects” in the context of single-cell transcriptomics data?

Differences in handling and processing for different groups of cells can introduce differences in measured gene expression levels.

Storage differences

Experimental processing differences

Bioinformatical processing differences

But: combining datasets increasingly allows to ask new questions and uncover rare events!

Need for computational mitigation strategies ~ batch effect removal methods

Goal: After computational batch correction, biologically equivalent cells should cluster together.

Pre

Cell types

- CD10⁺ B cells
- CD14⁺ monocytes
- CD16⁺ monocytes
- CD20⁺ B cells
- CD4⁺ T cells
- CD8⁺ T cells
- Erythrocytes
- Erythroid progenitors
- HSPCs
- Megakaryocyte progenitors
- Monocyte progenitors
- Monocyte-derived dendritic cells
- NK cells
- NKT cells
- Plasma cells
- Plasmacytoid dendritic cells

Batches

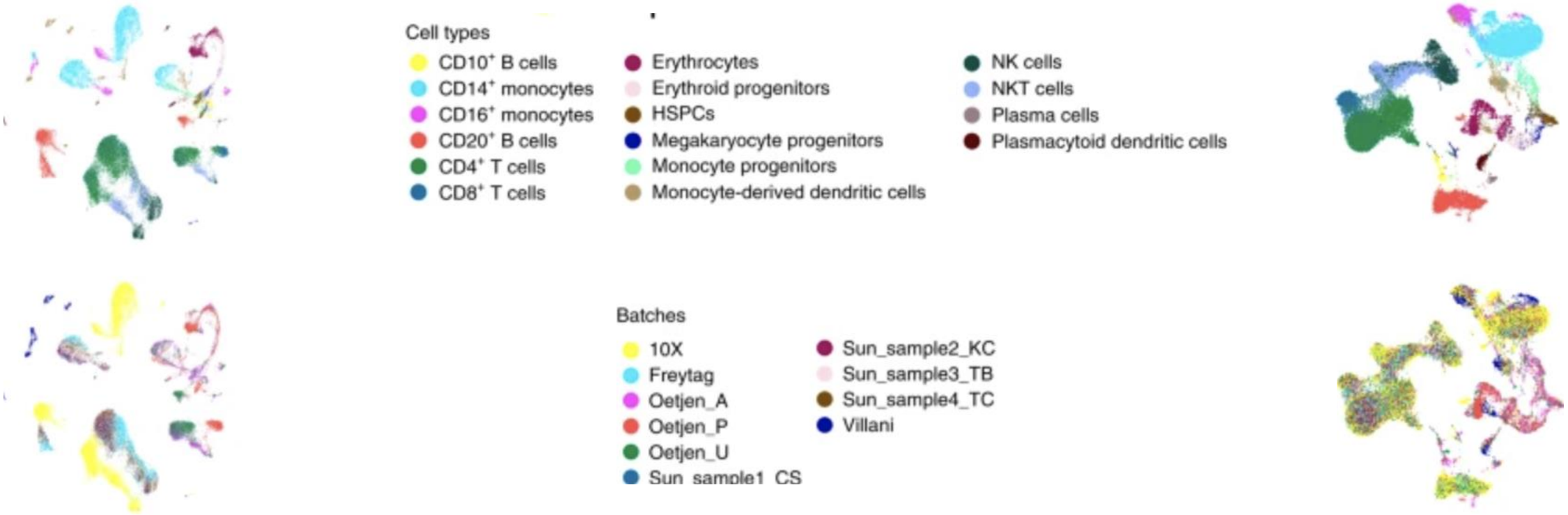
- 10X
- Freytag
- Oetjen_A
- Oetjen_P
- Oetjen_U
- Sun_sample1_CS
- Sun_sample2_KC
- Sun_sample3_TB
- Sun_sample4_TC
- Villani

Luecken, M.D., Büttner, M., Chaichoompu, K. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods* **19**, 41–50 (2022). https://doi.org/10.1038/s41592-021-01336-8

Goal: After computational batch correction, biologically equivalent cells should cluster together.

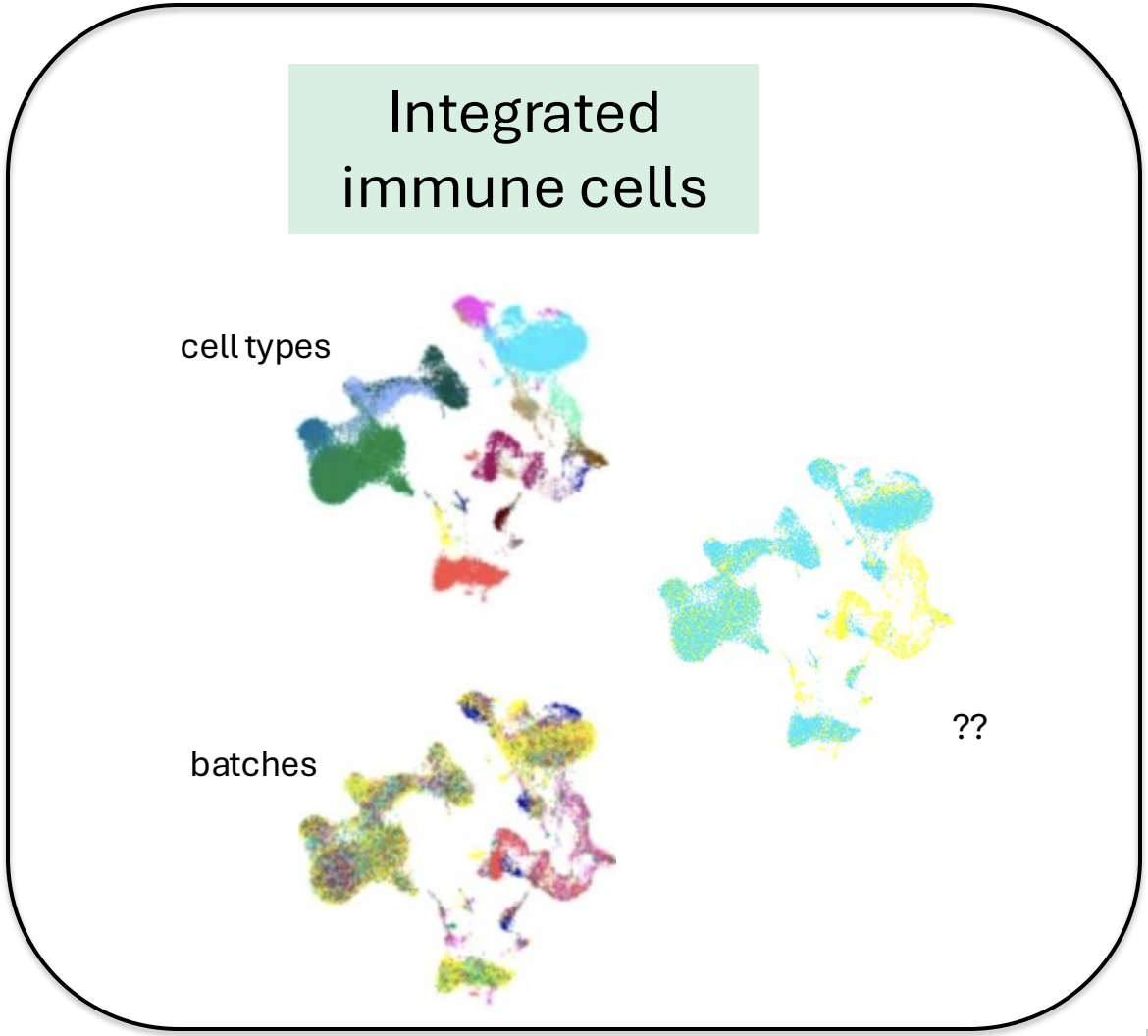
Pre

Post



Luecken, M.D., Büttner, M., Chaichoompu, K. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods* **19**, 41–50 (2022). https://doi.org/10.1038/s41592-021-01336-8

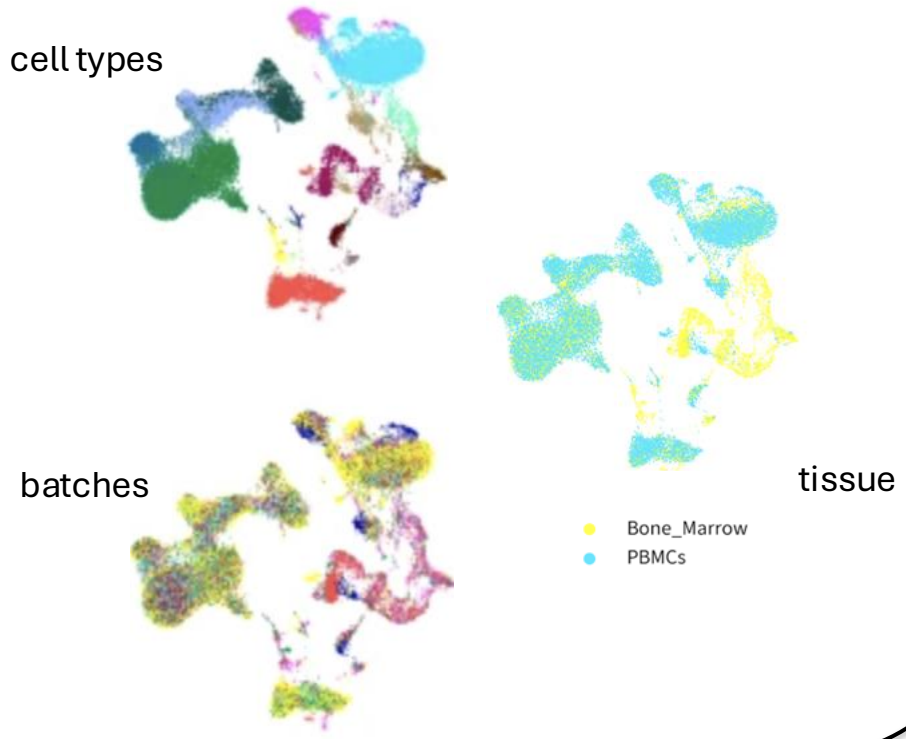
Core problem 1: Which differences between batches are relevant for our question?



Luecken, M.D., Büttner, M., Chaichoompu, K. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods* **19**, 41–50 (2022). https://doi.org/10.1038/s41592-021-01336-8

Core problem 1: Which differences between batches are relevant for our question?

Integrated immune cells

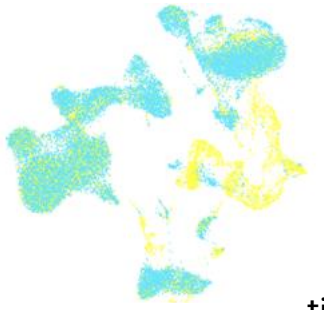


Luecken, M.D., Büttner, M., Chaichoompu, K. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods* **19**, 41–50 (2022). https://doi.org/10.1038/s41592-021-01336-8

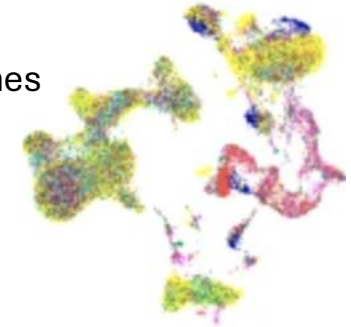
Core problem 1: Which differences between batches are relevant for our question?

Integrated immune cells

cell types



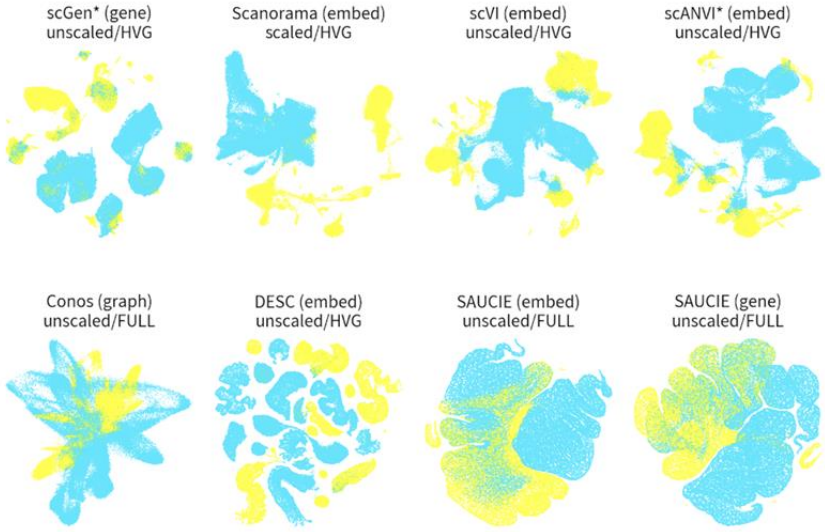
batches



tissue

Bone_Marrow
PBMCs

Integrated immune cells

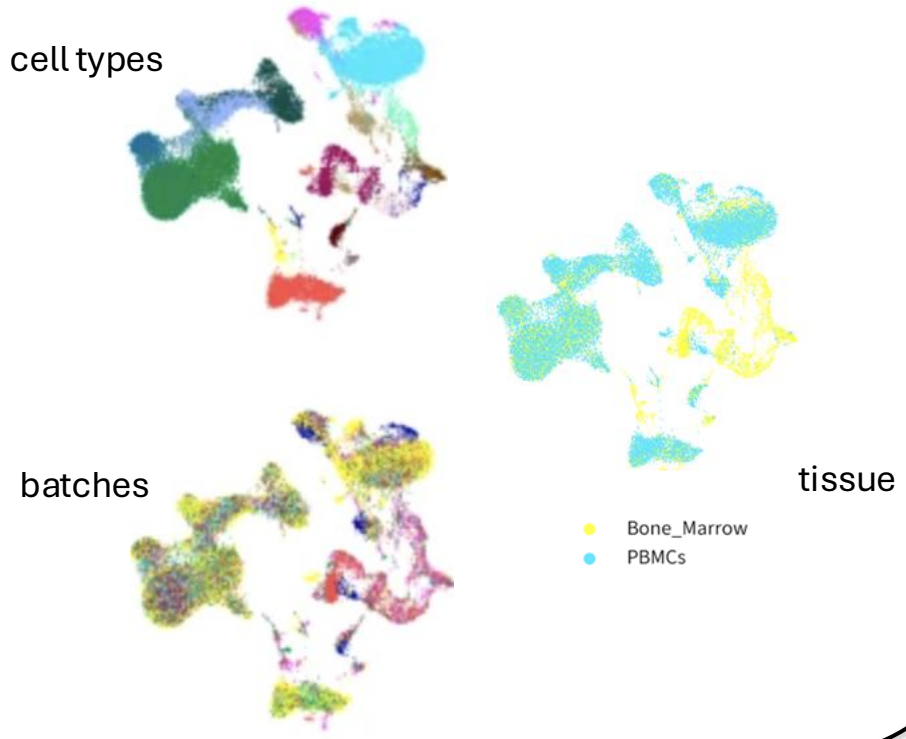


??

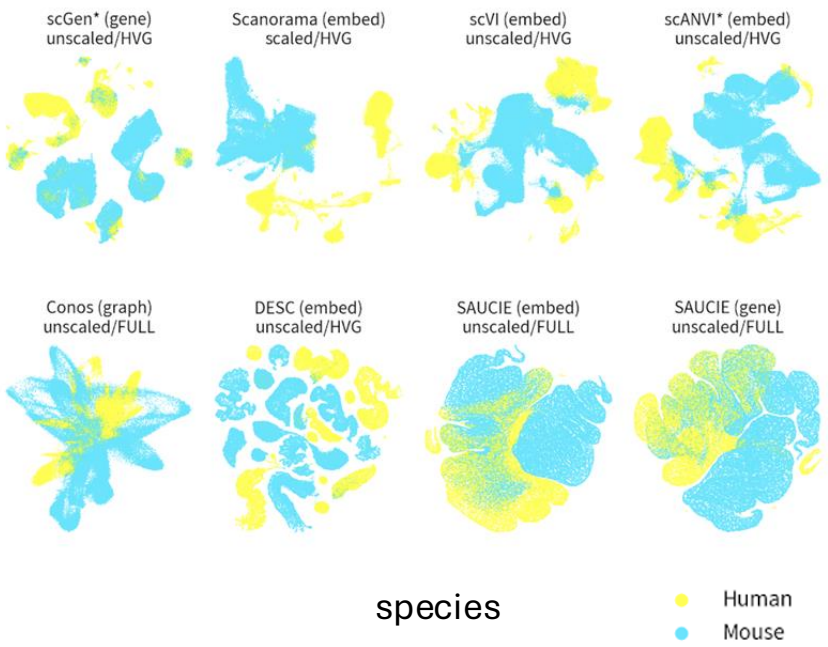
Luecken, M.D., Büttner, M., Chaichoompu, K. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods* **19**, 41–50 (2022). https://doi.org/10.1038/s41592-021-01336-8

Core problem 1: Which differences between batches are relevant for our question?

Integrated immune cells



Integrated immune cells



Luecken, M.D., Büttner, M., Chaichoompu, K. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods* **19**, 41–50 (2022). https://doi.org/10.1038/s41592-021-01336-8

Core problem 1: Which differences between batches are relevant for our question?

Whether an effect is an undesirable batch effect depends on the question being asked!

- Processing
- Disease state
- Stimulation
- Tissues
- Species
- …

Core problem 1: Which differences between batches are relevant for our question?

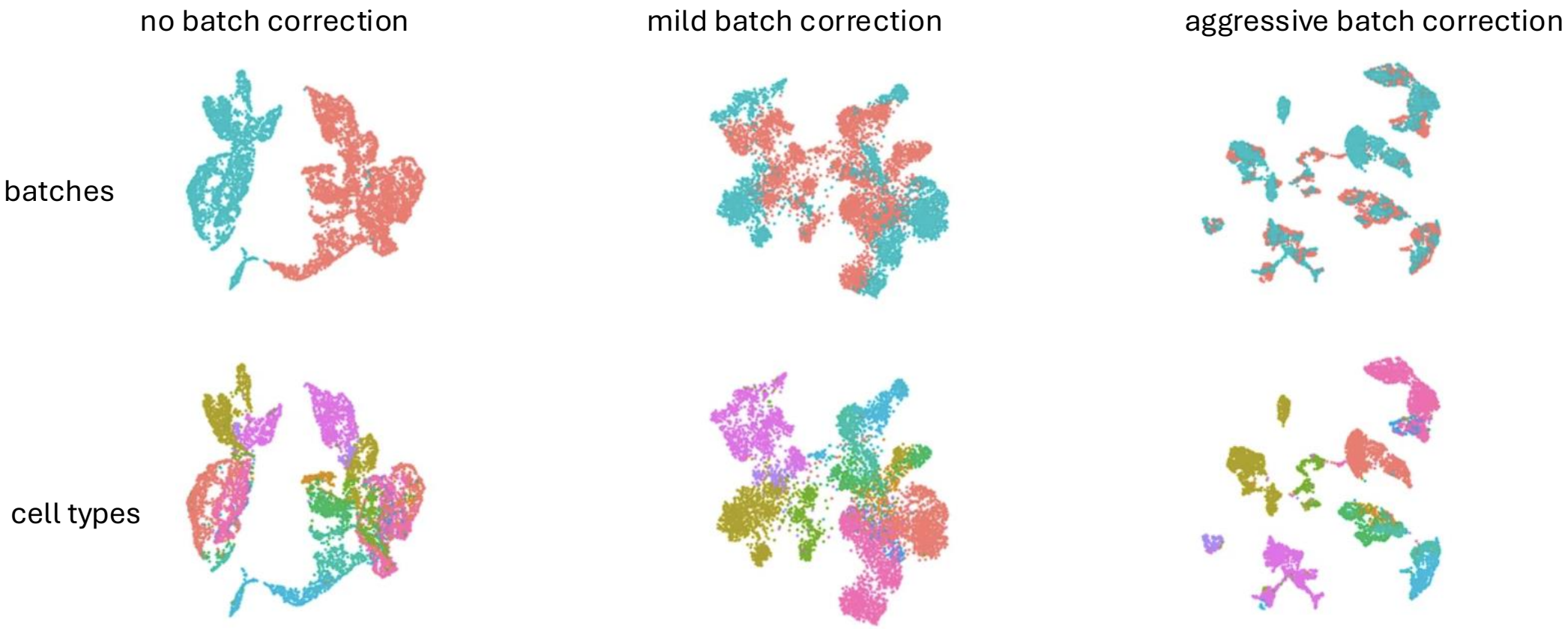
Whether an effect is an undesirable batch effect depends on the question being asked!

- Processing
- Disease state
- Stimulation
- Tissues
- Species
- ...

Which is (are) the right batch covariate(s)?

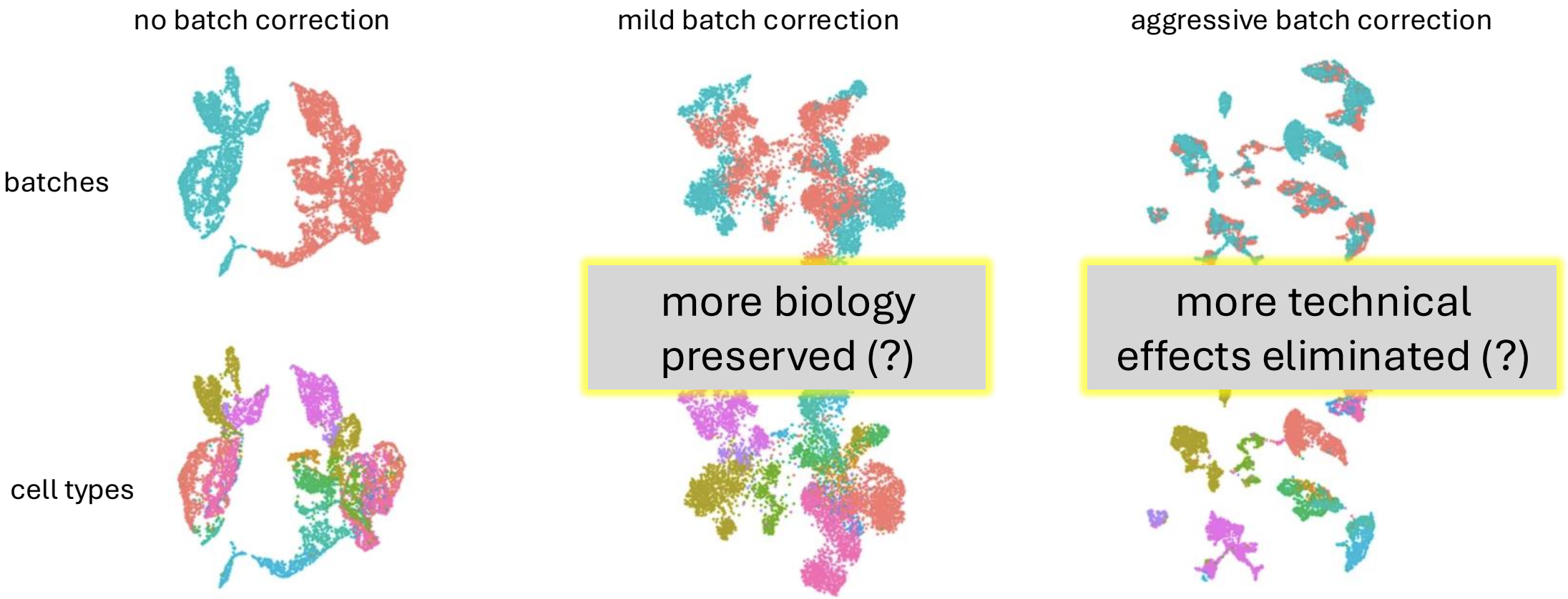
Which genes do we consider likely to carry effects associated with these ?

Core problem 2: how can we tell “technical” and “biological” batch effects apart?



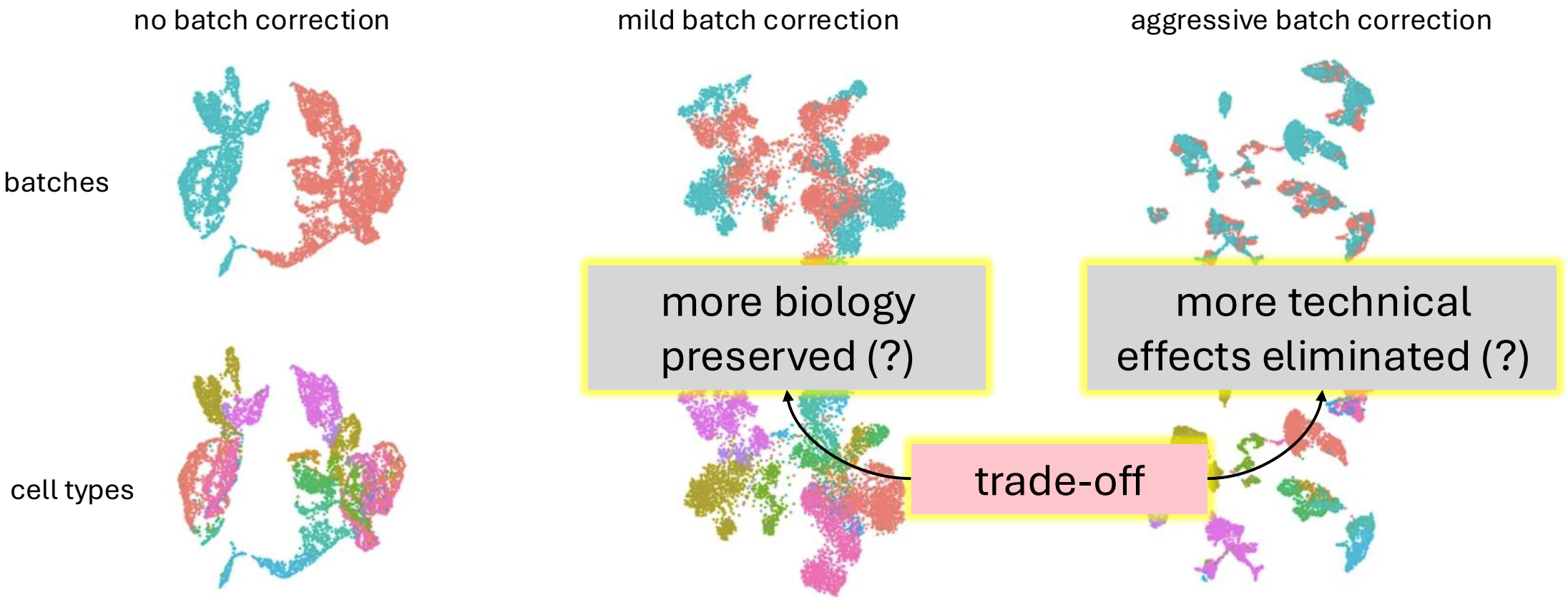
Tran, H.T.N., Ang, K.S., Chevrier, M. et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biol 21, 12 (2020). https://doi.org/10.1186/s13059-019-1850-9

Core problem 2: how can we tell “technical” and “biological” batch effects apart?



Tran, H.T.N., Ang, K.S., Chevrier, M. et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biol 21, 12 (2020). https://doi.org/10.1186/s13059-019-1850-9

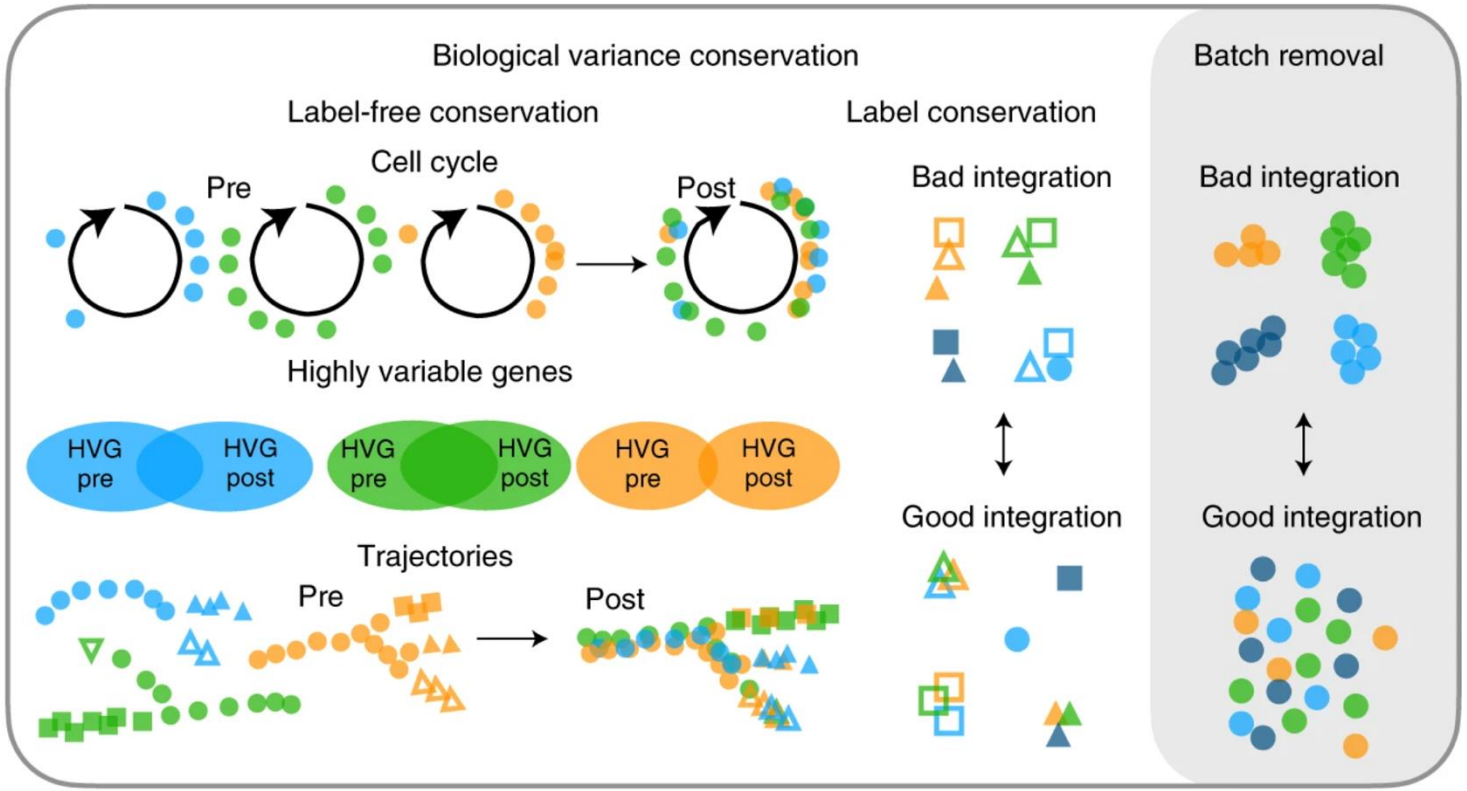
Core problem 2: how can we tell “technical” and “biological” batch effects apart?



Tran, H.T.N., Ang, K.S., Chevrier, M. et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biol 21, 12 (2020). https://doi.org/10.1186/s13059-019-1850-9

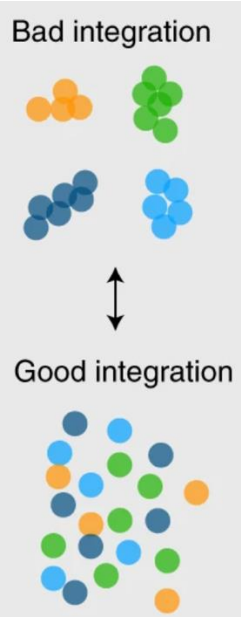
Core problem 2: how can we tell “technical” and “biological” batch effects apart?

What do we mean by “technical” and “biological” effects?



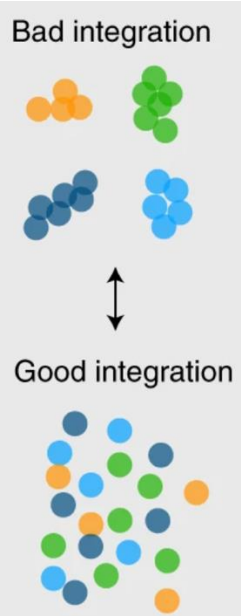
Luecken, M.D., Büttner, M., Chaichoompu, K. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods* **19**, 41–50 (2022). https://doi.org/10.1038/s41592-021-01336-8

Excursion: integration quality metrics



Luecken, M.D., Büttner, M., Chaichoompu, K. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods* **19**, 41–50 (2022). https://doi.org/10.1038/s41592-021-01336-8

Excursion: integration quality metrics

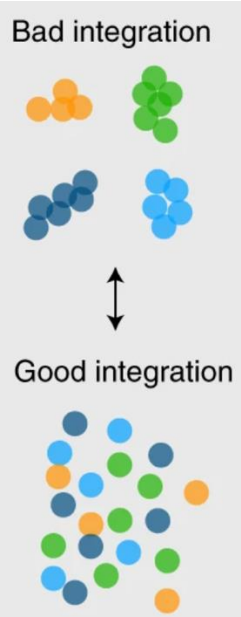


The four batches are integrated properly, because they are well mixed in the UMAP.

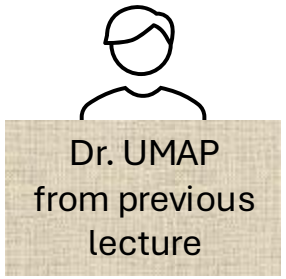
Dr. UMAP from previous lecture

Luecken, M.D., Büttner, M., Chaichoompu, K. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods* **19**, 41–50 (2022). https://doi.org/10.1038/s41592-021-01336-8

Excursion: integration quality metrics



The four batches are integrated properly, because they are well mixed in the UMAP.



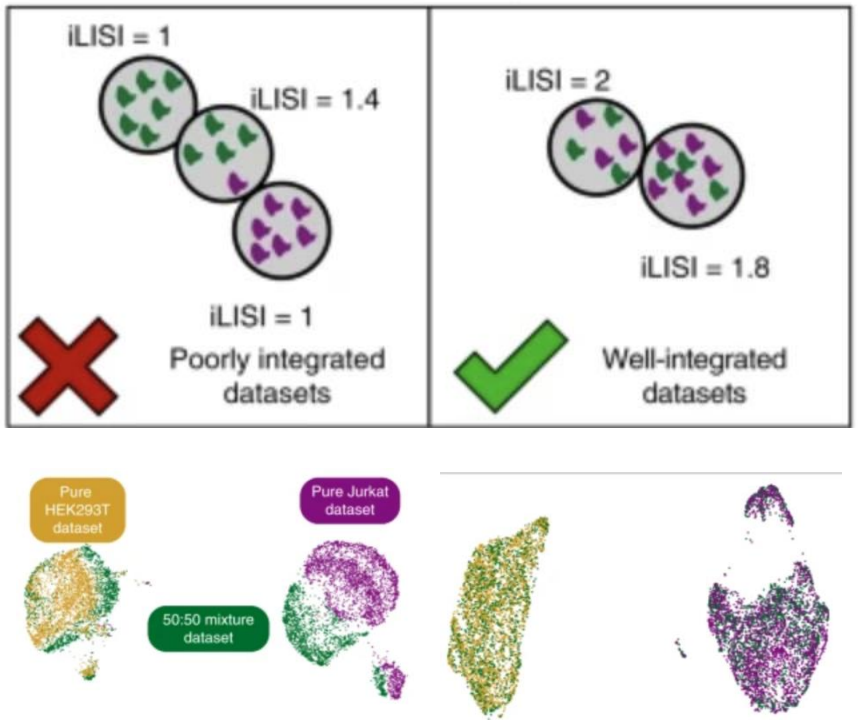
Good first indicator, but follow up with a quantitative metric.

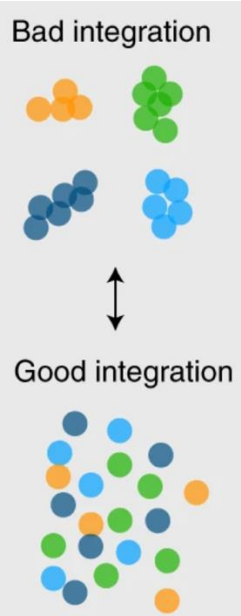
Luecken, M.D., Büttner, M., Chaichoompu, K. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods* **19**, 41–50 (2022). https://doi.org/10.1038/s41592-021-01336-8

Excursion: integration quality metrics

iLISI

integration Local
Inverse Simpson's Index



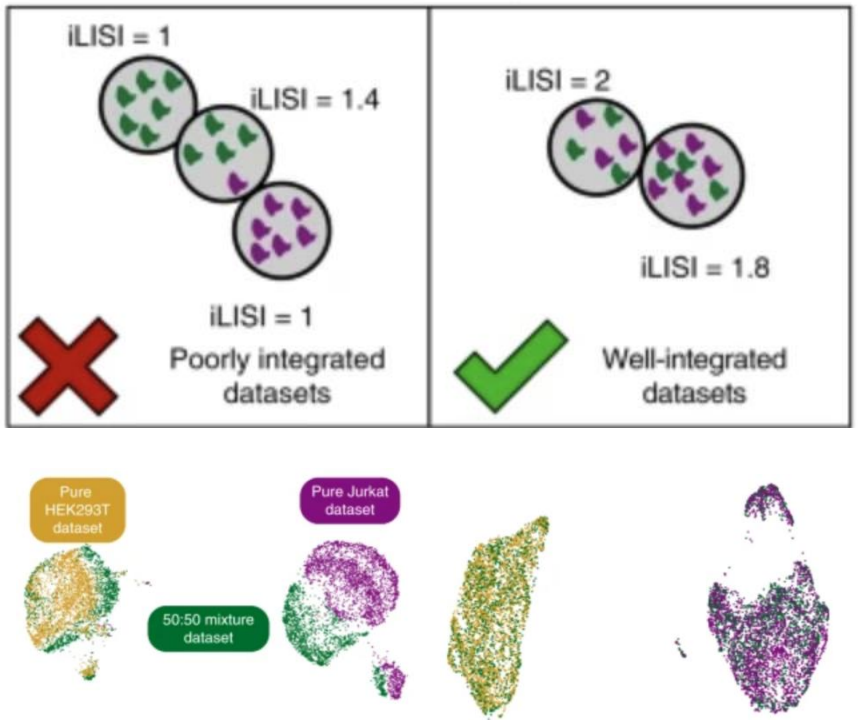


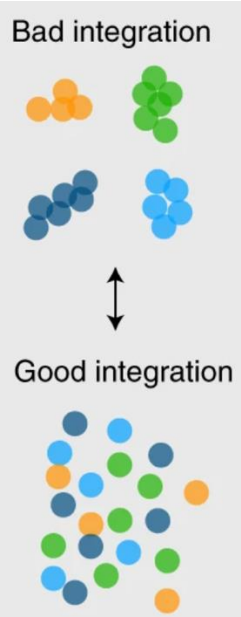
Korsunsky, I., Millard, N., Fan, J. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods 16, 1289–1296 (2019). https://doi.org/10.1038/s41592-019-0619-0

Excursion: integration quality metrics

iLISI

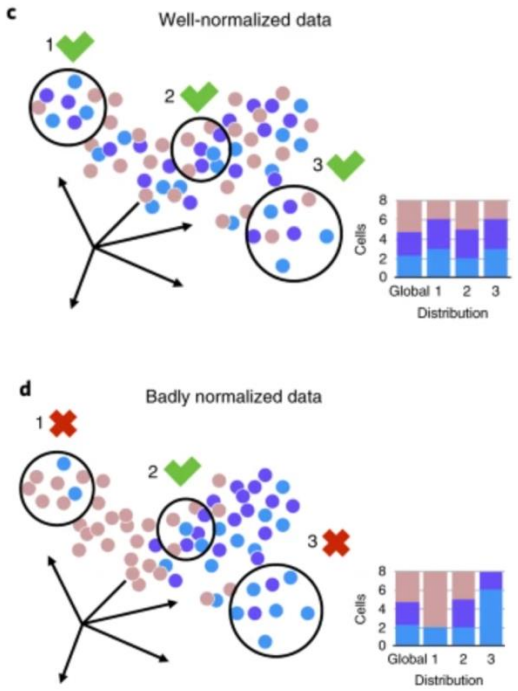
integration Local
Inverse Simpson's Index





kBET

k-nearest-neighbor
batch-effect test



Korsunsky, I., Millard, N., Fan, J. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods 16, 1289–1296 (2019). https://doi.org/10.1038/s41592-019-0619-0

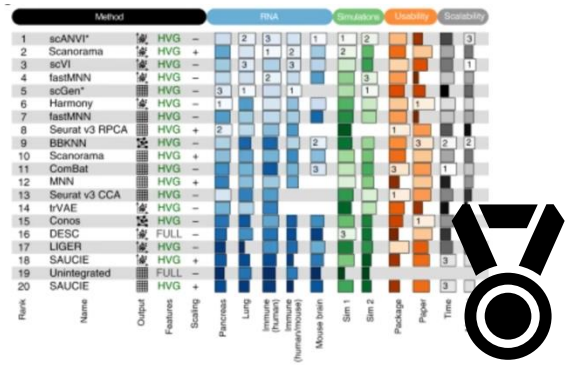
Büttner, M., Miao, Z., Wolf, F.A. *et al.* A test metric for assessing single-cell RNA-seq batch correction. *Nat Methods* **16**, 43–49 (2019). https://doi.org/10.1038/s41592-018-0254-1

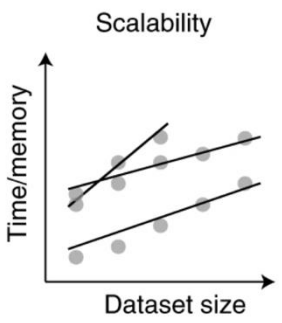
Core problem 2: how can we tell “technical” and “biological” batch effects apart?

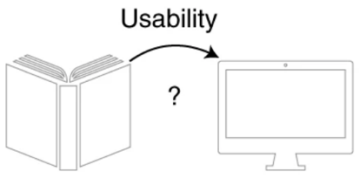
Which method and parameters to use?

Performance in benchmarking studies
(ideally, performance in similar problems)

resource constraints and convenience
(e.g. availability in your ecosystem)







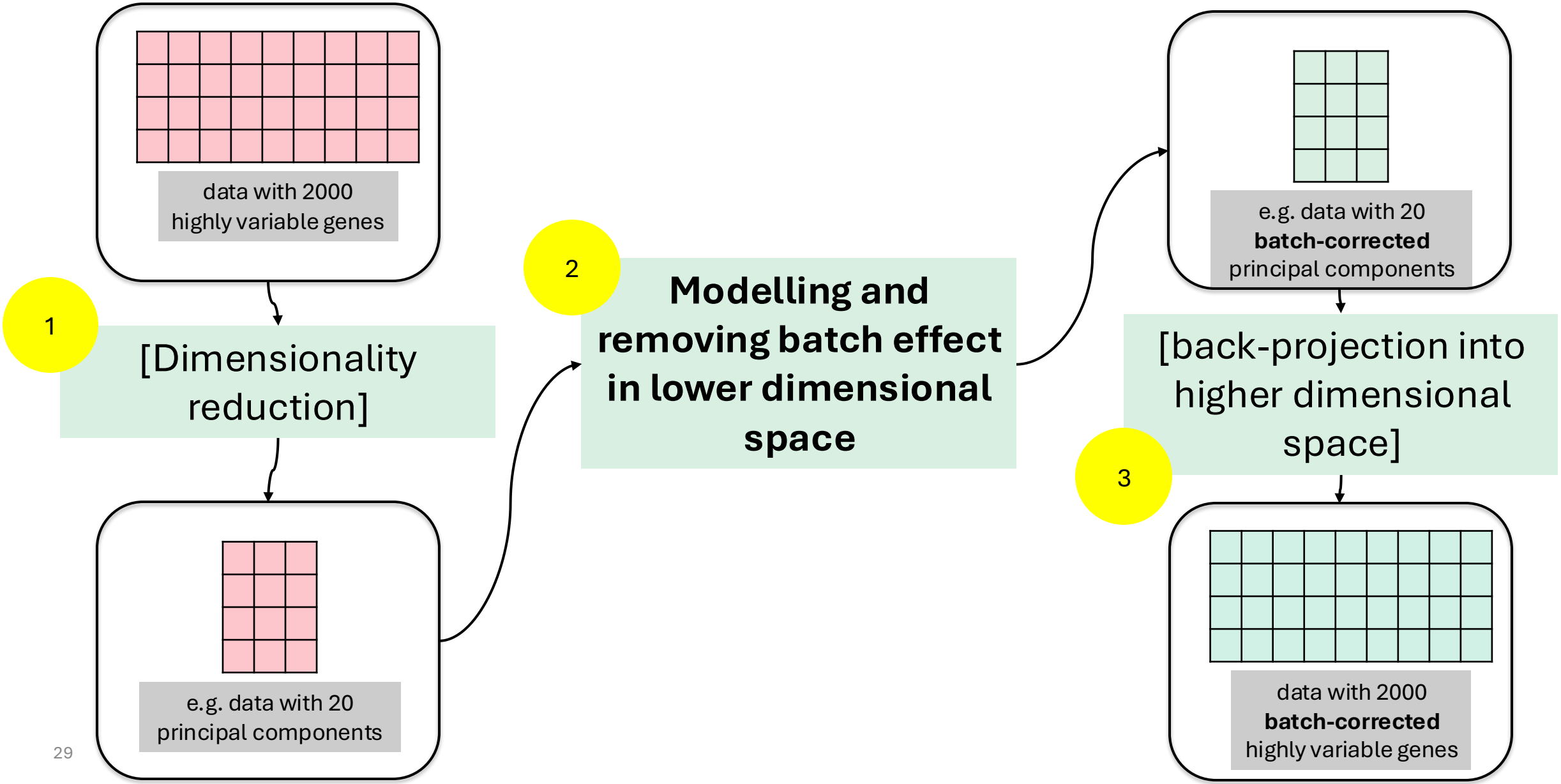
If possible, especially for large and/or complex integrations (many data sets, many cell types) evaluate several methods and parameter combinations

Luecken, M.D., Büttner, M., Chaichoompu, K. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods* **19**, 41–50 (2022). https://doi.org/10.1038/s41592-021-01336-8

Basic steps of (most) batch integration protocols

Modelling and removing batch effect in lower dimensional space

Basic steps of (most) batch integration protocols



Categories of batch removal 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

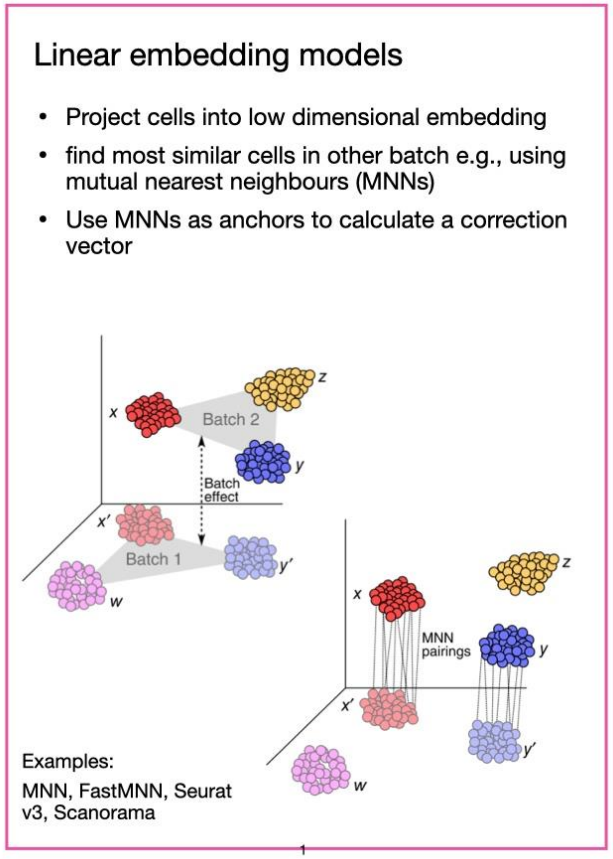
multiplicative batch effect

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

- from the “bulk ages”
- batch effect assumed consistent across all cells

https://www.sc-best-practices.org/cellular_structure/integration.html#types-of-integration-models

Categories of batch removal methods

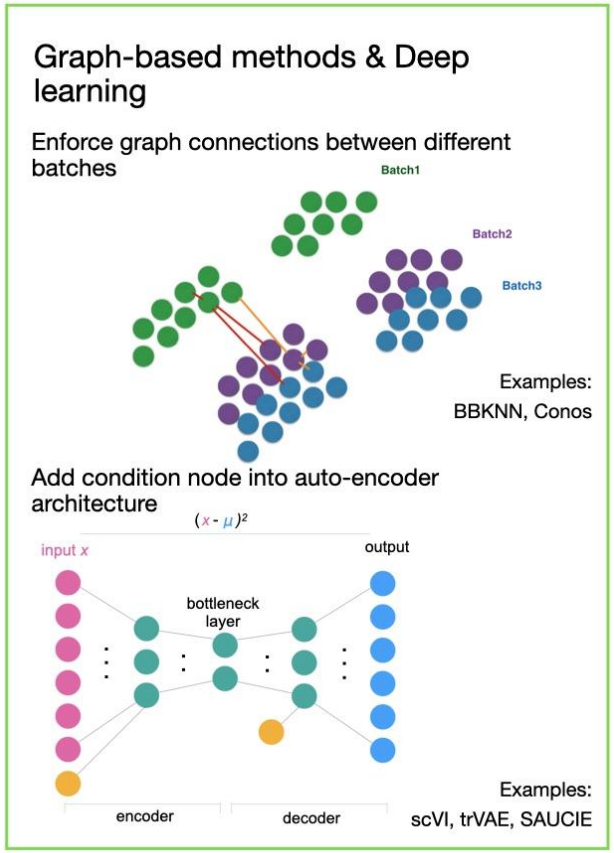


- Developed specifically for single cell data
- Consider local neighbourhoods

https://www.sc-best-practices.org/cellular_structure/integration.html#types-of-integration-models

Categories of batch removal methods

- Developed specifically for single cell data
- Typically need more data (and resources) to run well
- Best performance in recent benchmarking studies for large integrations



https://www.sc-best-practices.org/cellular_structure/integration.html#types-of-integration-models

Categories of batch removal 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

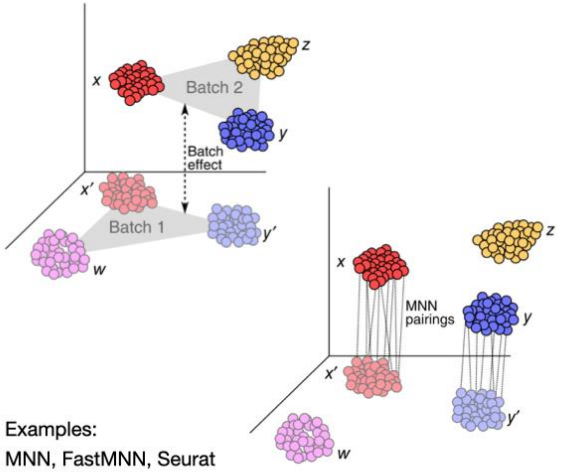
additive batch effect

multiplicative batch effect

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

Linear embedding models

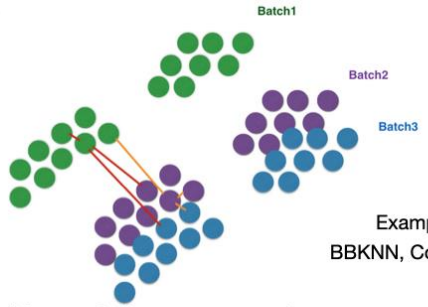
- Project cells into low dimensional embedding
- find most similar cells in other batch e.g., using mutual nearest neighbours (MNNs)
- Use MNNs as anchors to calculate a correction vector



Examples:
MNN, FastMNN, Seurat v3, Scanorama

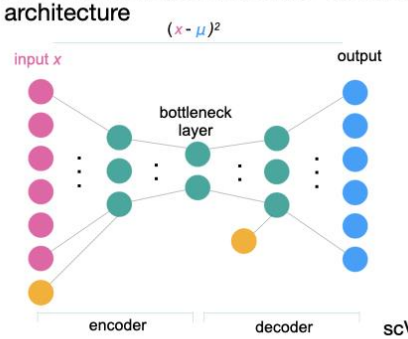
Graph-based methods & Deep learning

Enforce graph connections between different batches



Examples:
BBKNN, Conos

Add condition node into auto-encoder architecture



Examples:
scVI, trVAE, SAUCIE

Typically used for dataset integration

https://www.sc-best-practices.org/cellular_structure/integration.html#types-of-integration-models

Categories of batch removal 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

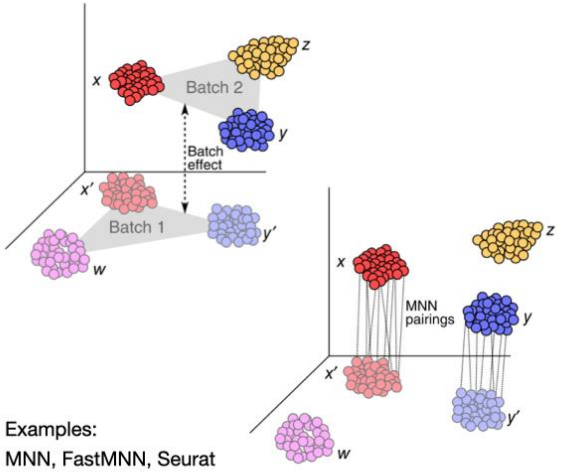
additive batch effect

multiplicative batch effect

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

Linear embedding models

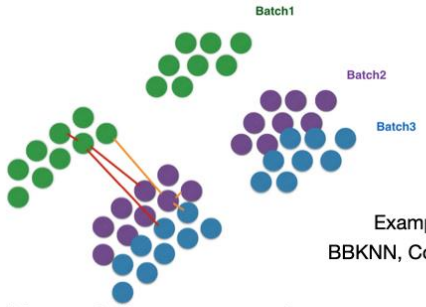
- Project cells into low dimensional embedding
- find most similar cells in other batch e.g., using mutual nearest neighbours (MNNs)
- Use MNNs as anchors to calculate a correction vector



Examples:
MNN, FastMNN, Seurat v3, Scanorama

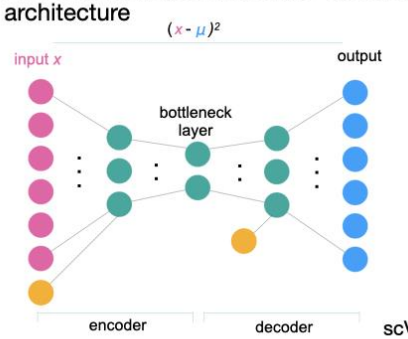
Graph-based methods & Deep learning

Enforce graph connections between different batches



Examples:
BBKNN, Conos

Add condition node into auto-encoder architecture



Examples:
scVI, trVAE, SAUCIE

Best for small datasets / simple tasks

https://www.sc-best-practices.org/cellular_structure/integration.html#types-of-integration-models

Categories of batch removal 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

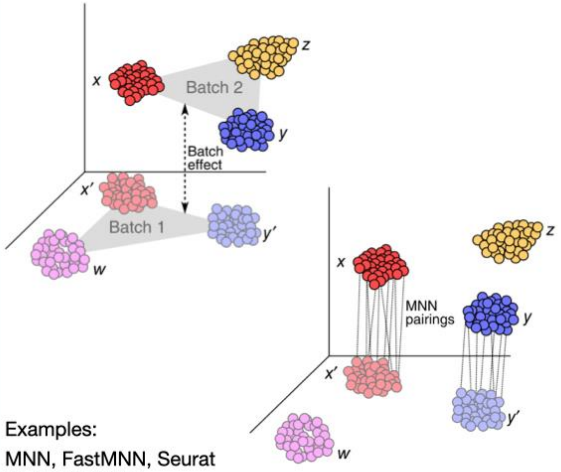
additive batch effect

multiplicative batch effect

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

Linear embedding models

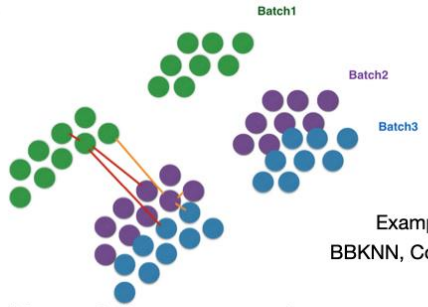
- Project cells into low dimensional embedding
- find most similar cells in other batch e.g., using mutual nearest neighbours (MNNs)
- Use MNNs as anchors to calculate a correction vector



Examples:
MNN, FastMNN, Seurat v3, Scanorama

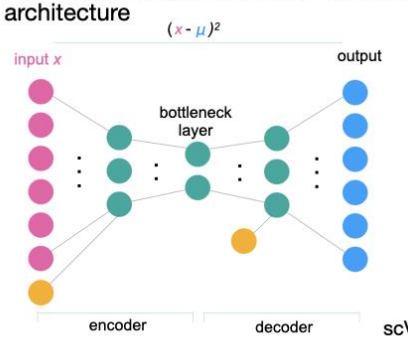
Graph-based methods & Deep learning

Enforce graph connections between different batches



Examples:
BBKNN, Conos

Add condition node into auto-encoder architecture



Examples:
scVI, trVAE, SAUCIE

Best for large datasets / complex tasks

https://www.sc-best-practices.org/cellular_structure/integration.html#types-of-integration-models

Categories of batch removal 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

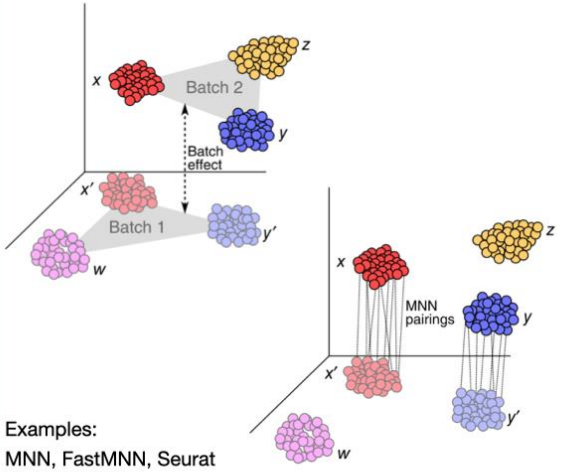
additive batch effect

multiplicative batch effect

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

Linear embedding models

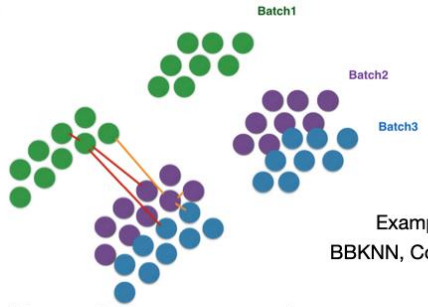
- Project cells into low dimensional embedding
- find most similar cells in other batch e.g., using mutual nearest neighbours (MNNs)
- Use MNNs as anchors to calculate a correction vector



Examples:
MNN, FastMNN, Seurat v3, Scanorama

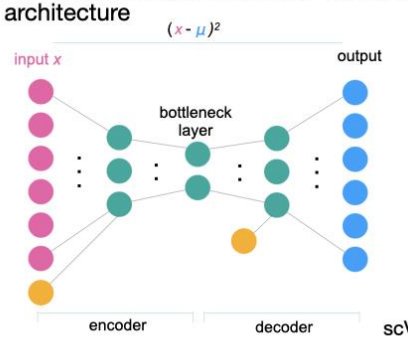
Graph-based methods & Deep learning

Enforce graph connections between different batches



Examples:
BBKNN, Conos

Add condition node into auto-encoder architecture

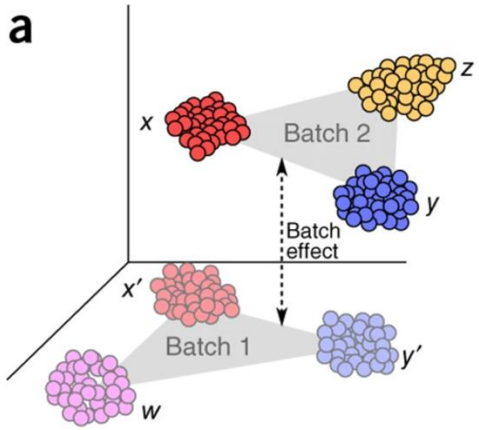


Examples:
scVI, trVAE, SAUCIE

Typically used for removing unwanted effects within datasets (e.g. cell cycle regression, nUMI regression)

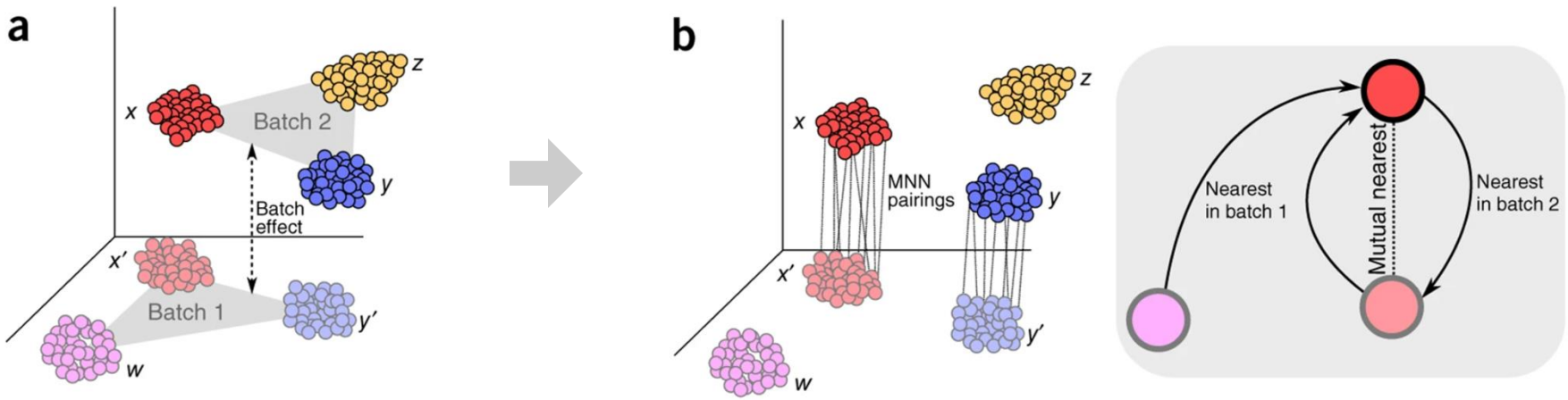
https://www.sc-best-practices.org/cellular_structure/integration.html#types-of-integration-models

Example method: Mutual Nearest Neighbours (MNN)



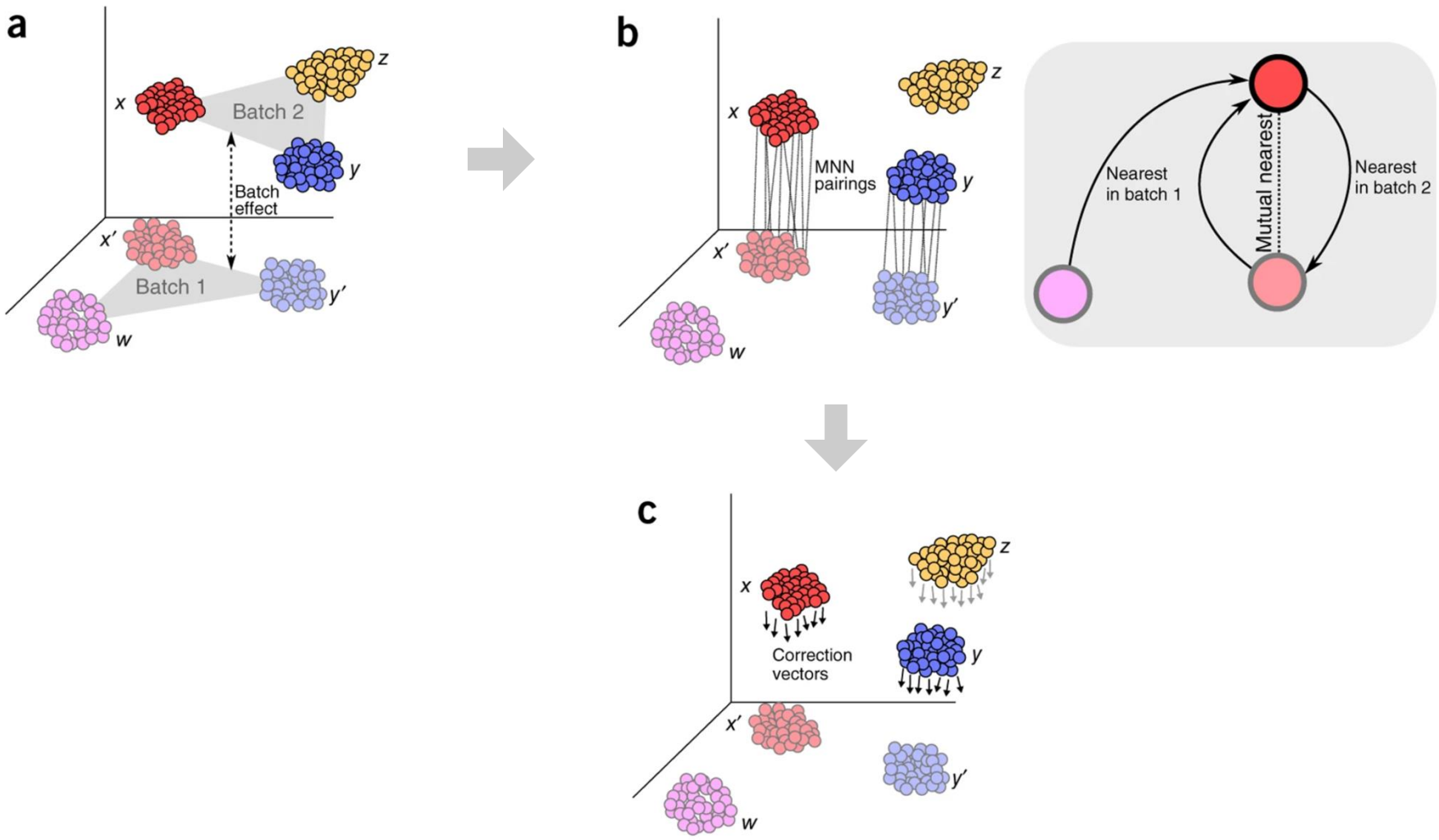
Haghverdi, L., Lun, A., Morgan, M. et al. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol 36, 421–427 (2018). https://doi.org/10.1038/nbt.4091

Example method: Mutual Nearest Neighbours (MNN)



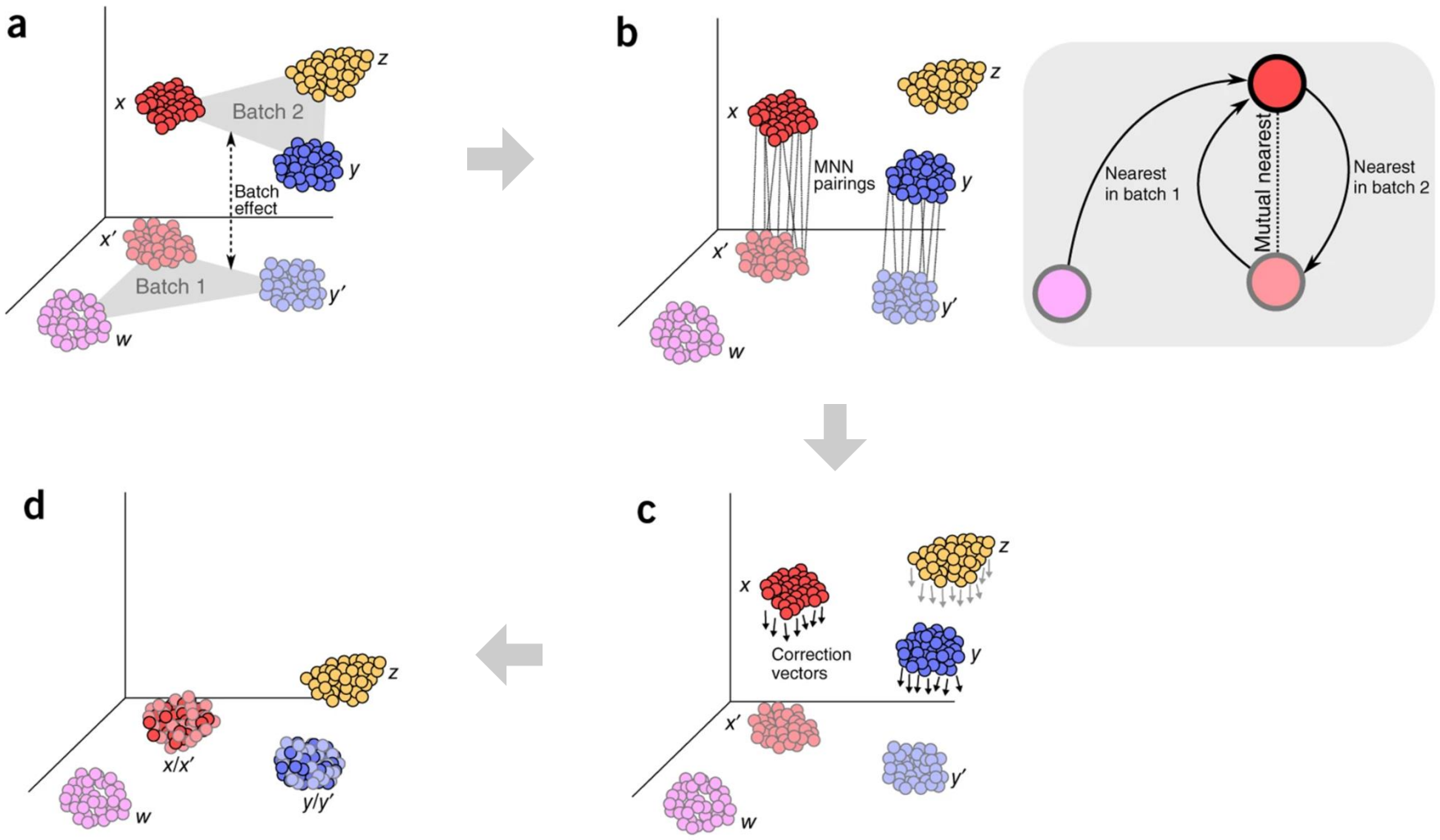
Haghverdi, L., Lun, A., Morgan, M. et al. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol 36, 421–427 (2018). https://doi.org/10.1038/nbt.4091

Example method: Mutual Nearest Neighbours (MNN)



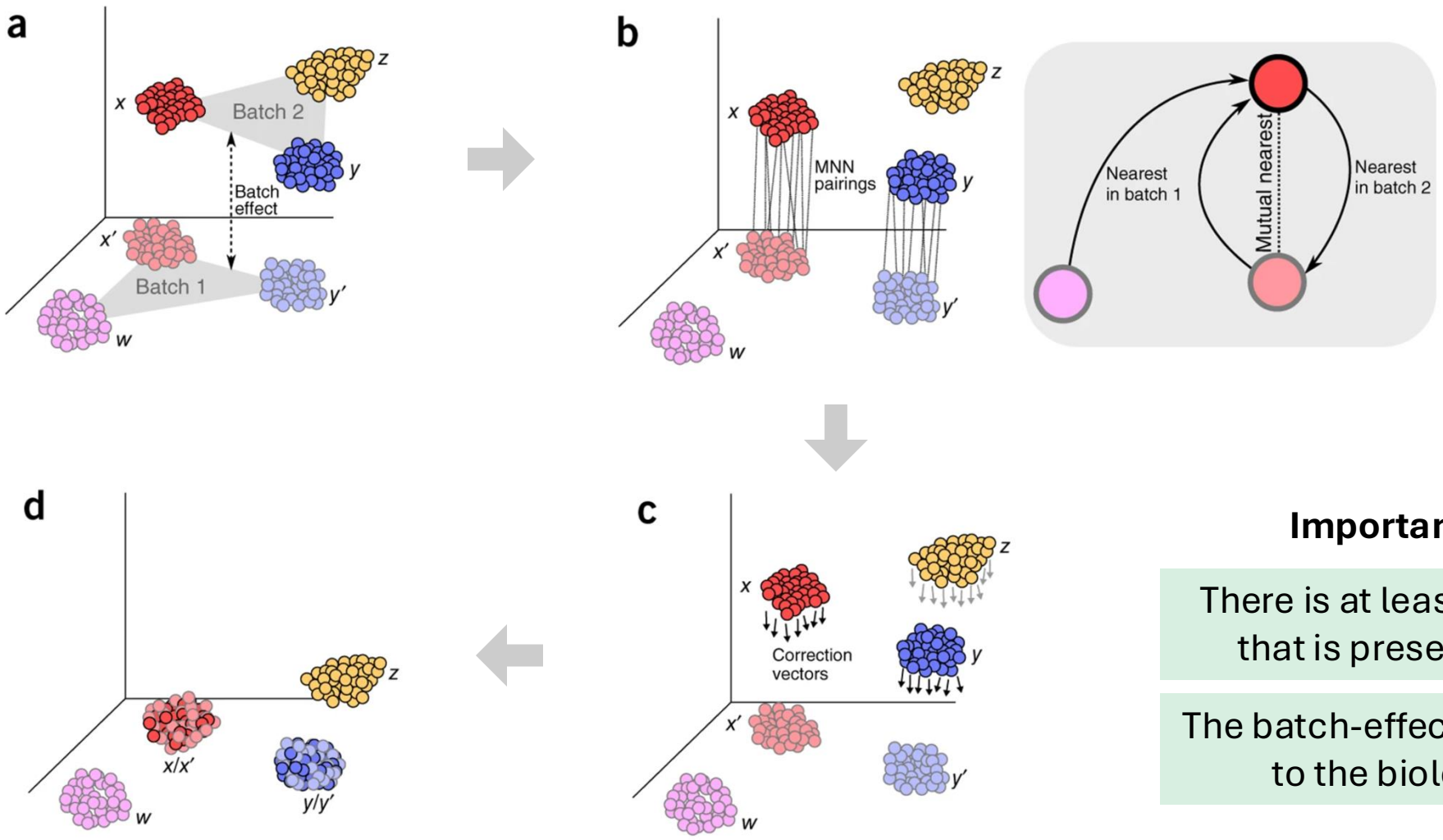
Haghverdi, L., Lun, A., Morgan, M. et al. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol 36, 421–427 (2018). https://doi.org/10.1038/nbt.4091

Example method: Mutual Nearest Neighbours (MNN)



Haghverdi, L., Lun, A., Morgan, M. et al. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol 36, 421–427 (2018). https://doi.org/10.1038/nbt.4091

Example method: Mutual Nearest Neighbours (MNN)



Important assumptions

There is at least one cell population that is present in both batches.

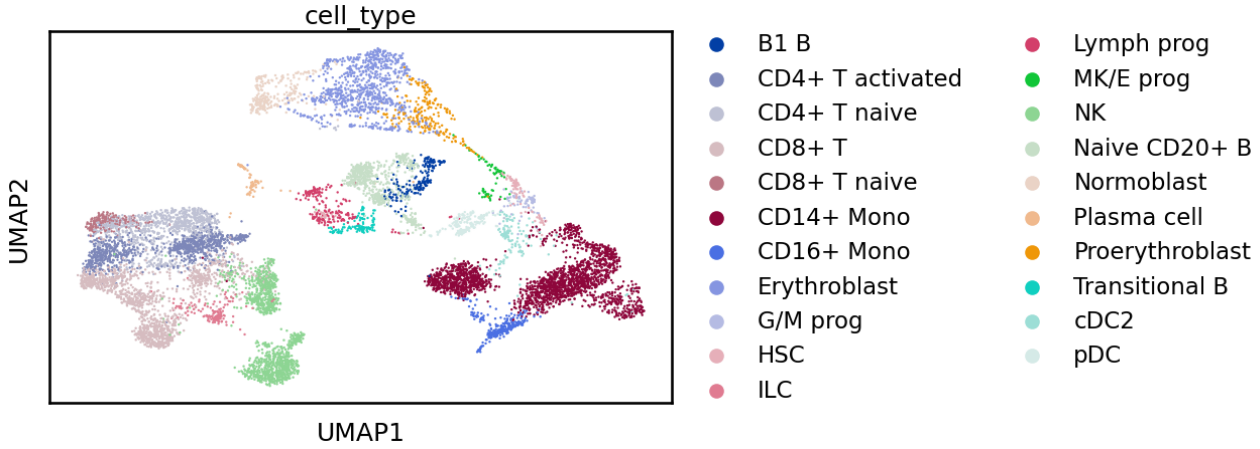
The batch-effect is almost orthogonal to the biological subspace.

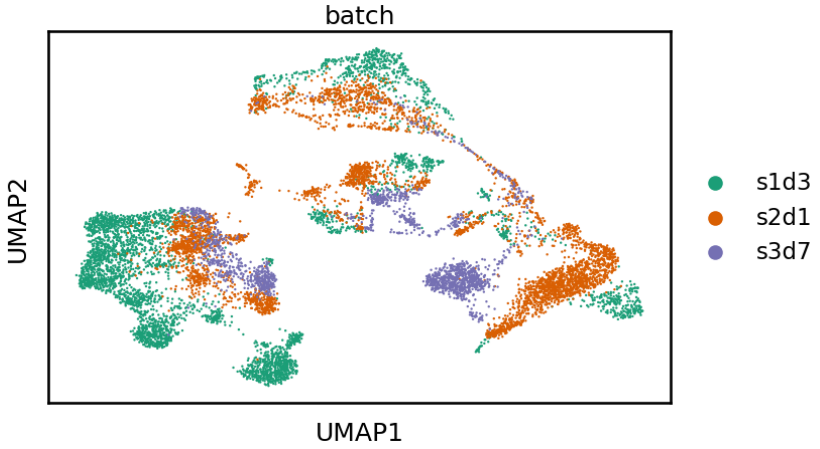
Haghverdi, L., Lun, A., Morgan, M. et al. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol 36, 421–427 (2018). https://doi.org/10.1038/nbt.4091

Integration workflow - outline

1 Inspect unintegrated data

[3 bone marrow datasets]





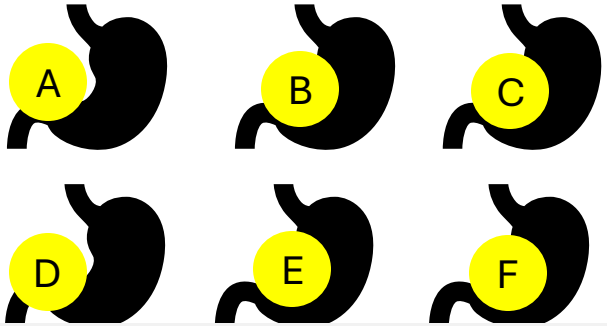
- Do we observe batch effects?
- Which co-variates seem to be driving them?

https://www.sc-best-practices.org/cellular_structure/integration.html#dataset

Integration workflow - outline

1 Inspect unintegrated data

2 Decide on meaningful batch covariates for your research question

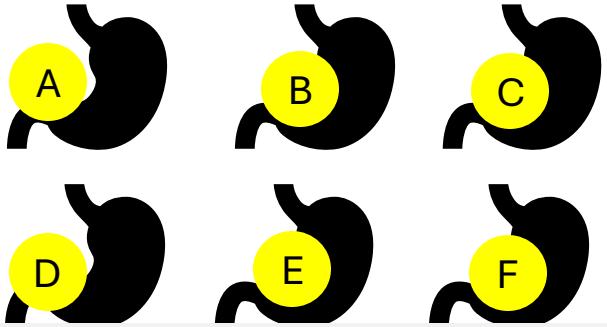


Common choice: sample as batch co-variate

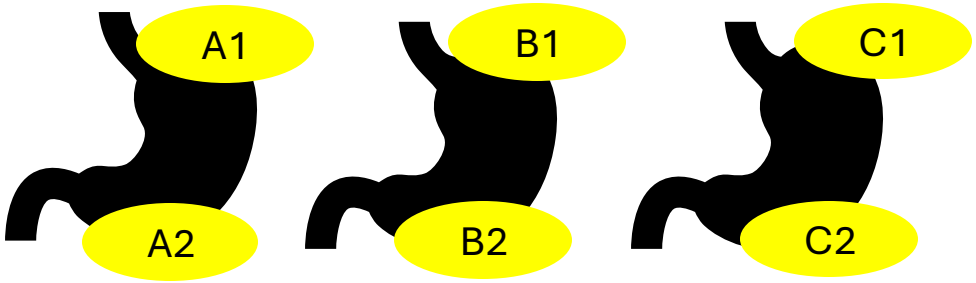
Integration workflow - outline

1 Inspect unintegrated data

2 Decide on meaningful batch covariates for your research question



Common choice: sample as batch co-variate

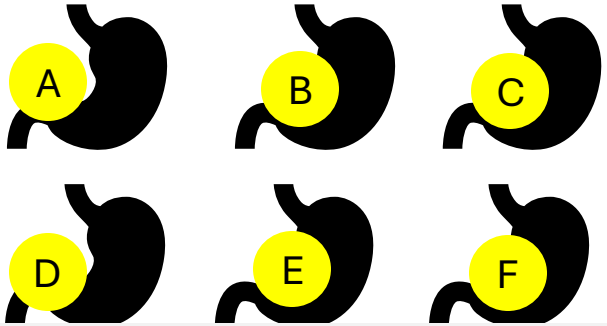


Additional biological factors: Donor as batch co-variate to preserve intra-donor variability

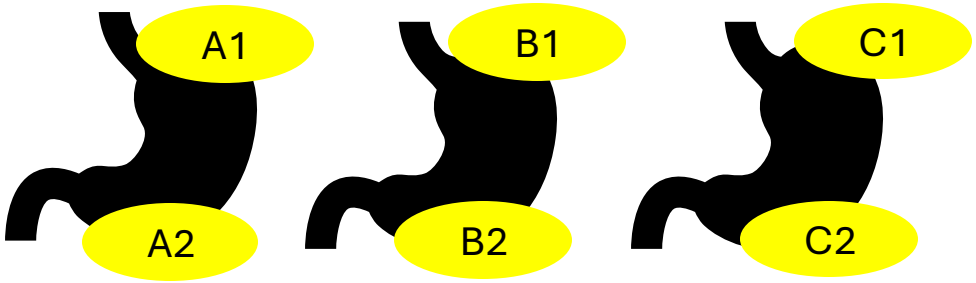
Integration workflow - outline

1 Inspect unintegrated data

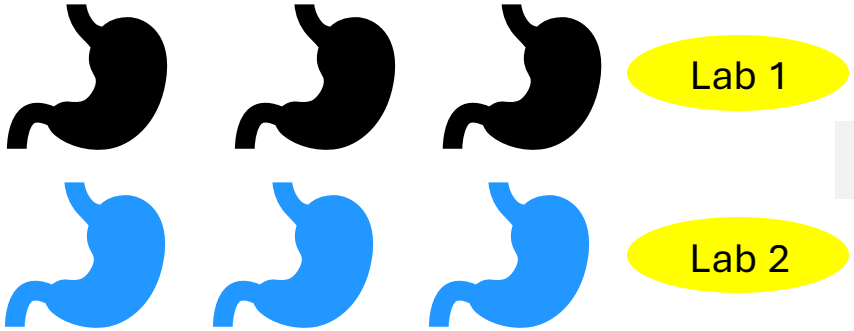
2 Decide on meaningful batch covariates for your research question



Common choice: sample as batch co-variate



Additional biological factors: Donor as batch co-variate to preserve intra-donor variability

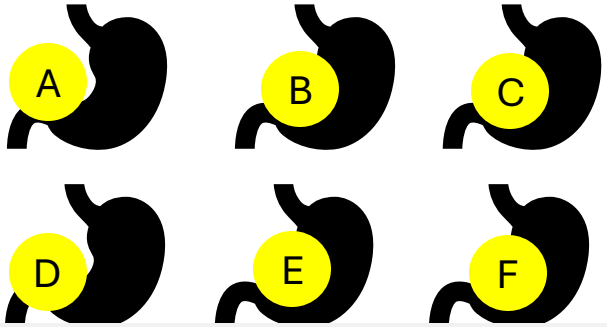


Technical factors as co-variates

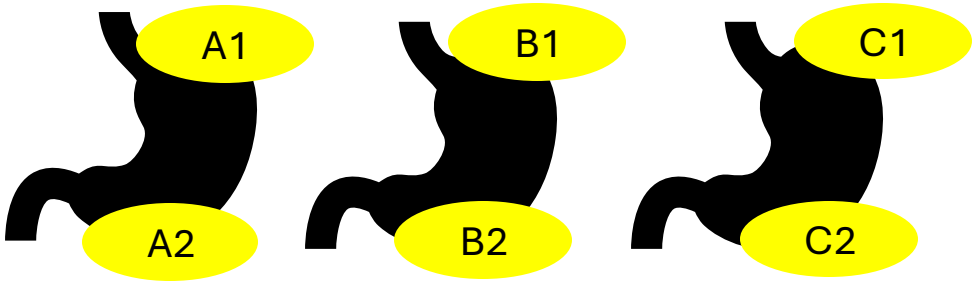
Integration workflow - outline

1 Inspect unintegrated data

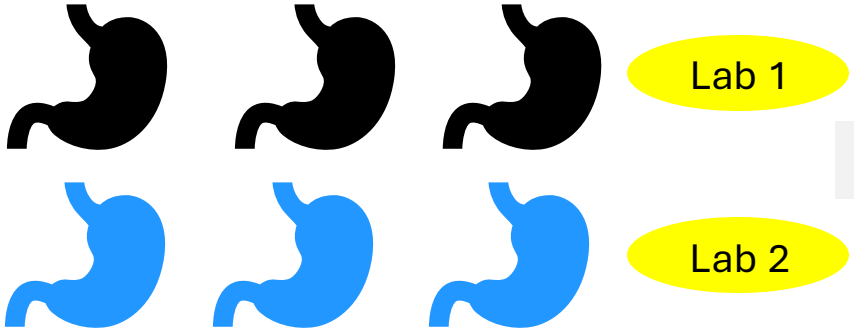
2 Decide on meaningful batch covariates for your research question



Common choice: sample as batch co-variate



Additional biological factors: Donor as batch co-variate to preserve intra-donor variability



Technical factors as co-variates

When needed, several batch co-variates can and should be selected.

Integration workflow - outline

1. Inspect unintegrated data
2. Decide on meaningful batch covariates for your research question
3. Batch-aware feature selection (~ “highly variable genes” step)

https://www.sc-best-practices.org/cellular_structure/integration.html#dataset

Integration workflow - outline

3

Batch-aware feature selection (~ “highly variable genes” step)

Naïve approach

Merge all datasets and perform highly variable gene selection on the combined dataset.

Integration workflow - outline

3

Batch-aware feature selection (~ “highly variable genes” step)

Naïve approach

Merge all datasets and perform highly variable gene selection on the combined dataset.

Genes which are only variable in one of the datasets may be missed in the global approach.

Genes which are only variable variable between datasets, but not within (~ are potentially not biologically informative) get picked up a lot.

Integration workflow - outline

3

Batch-aware feature selection (~ “highly variable genes” step)

Naïve approach

Merge all datasets and perform highly variable gene selection on the combination.

1. Perform highly variable gene selection ***per dataset***.
2. Combine all gene lists into one master gene list.
3. Subset all datasets to this gene list.
4. Merge data sets.

Integration workflow - outline

3

Batch-aware feature selection (~ “highly variable genes” step)

Naïve approach

Merge all datasets and perform highly variable gene selection on the combination.

1. Perform highly variable gene selection per dataset.
2. Combine all gene lists into one master gene list.
3. Subset all datasets to this gene list.
4. Merge data sets.

OR

1. Build biologically informed gene list independently of data (i.e. from the literature), e.g. comprised of marker genes for the anticipated cell types.
2. Subset all datasets to this gene list.
3. Merge data sets.

Integration workflow - outline

3

Batch-aware feature selection (~ “highly variable genes” step)

Naïve approach

Merge all datasets and perform highly variable gene selection on the combination.

1. Perform highly variable gene selection per dataset.
2. Combine all gene lists into one master gene list.
3. Subset all datasets to this gene list.
4. Merge data sets.

Good for exploration of functional clusters.

OR

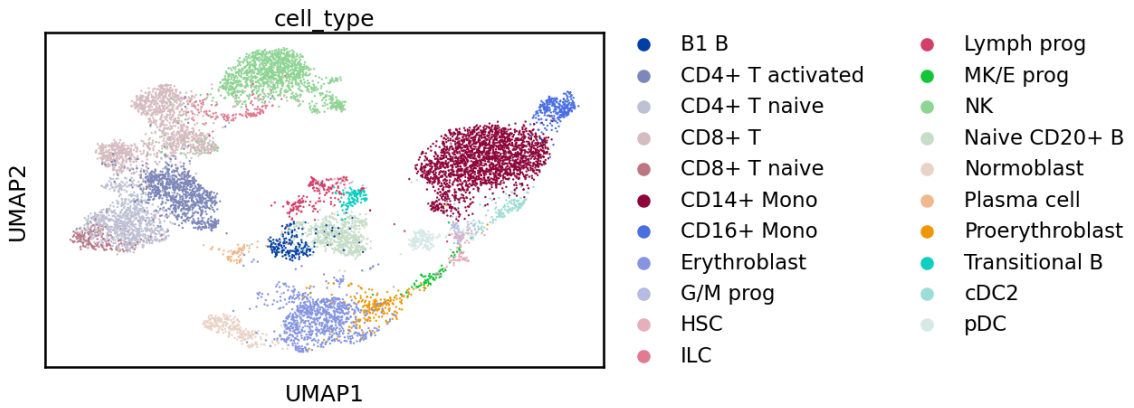
1. Build biologically informed gene list independently of data (i.e. from the literature), e.g. comprised of marker genes for the anticipated cell types.
2. Subset all datasets to this gene list.
3. Merge data sets.

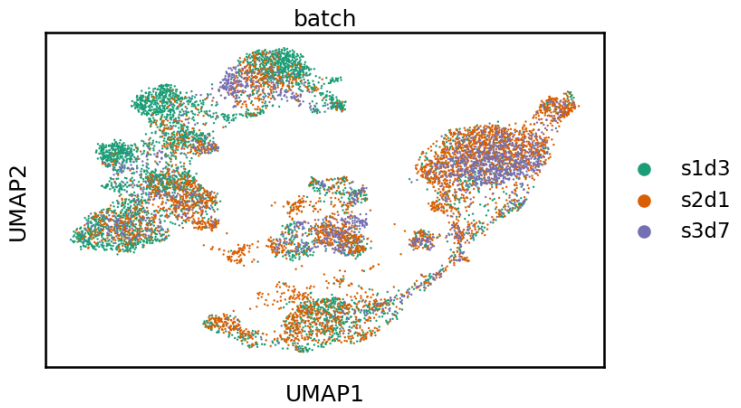
Good for repeatable, consistent cell type annotation.

Integration workflow - outline

1. Inspect unintegrated data
2. Decide on meaningful batch covariates for your research question
3. Batch-aware feature selection (~ “highly variable genes” step)
4. Run batch correction method of your choice

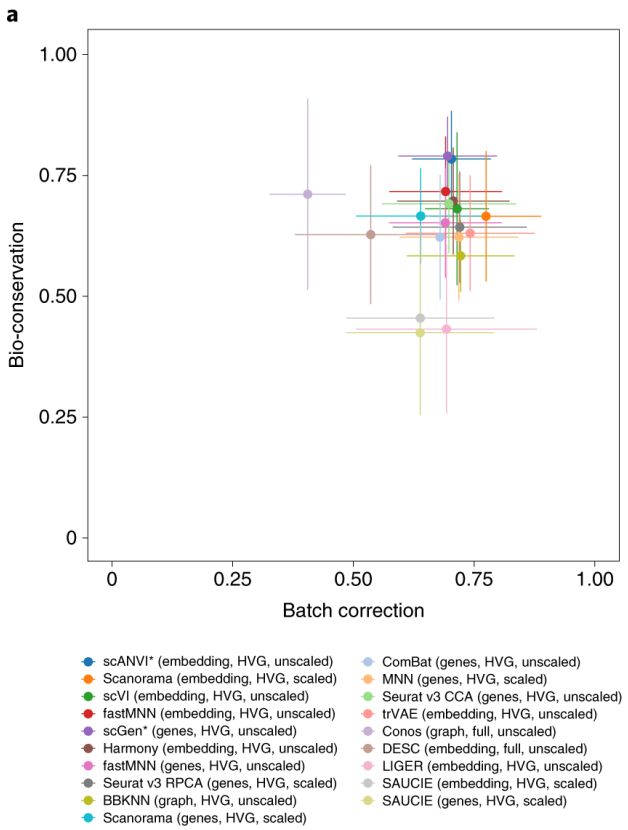
Seurat and scanpy provide this out of the box, but make sure to select the corresponding option!

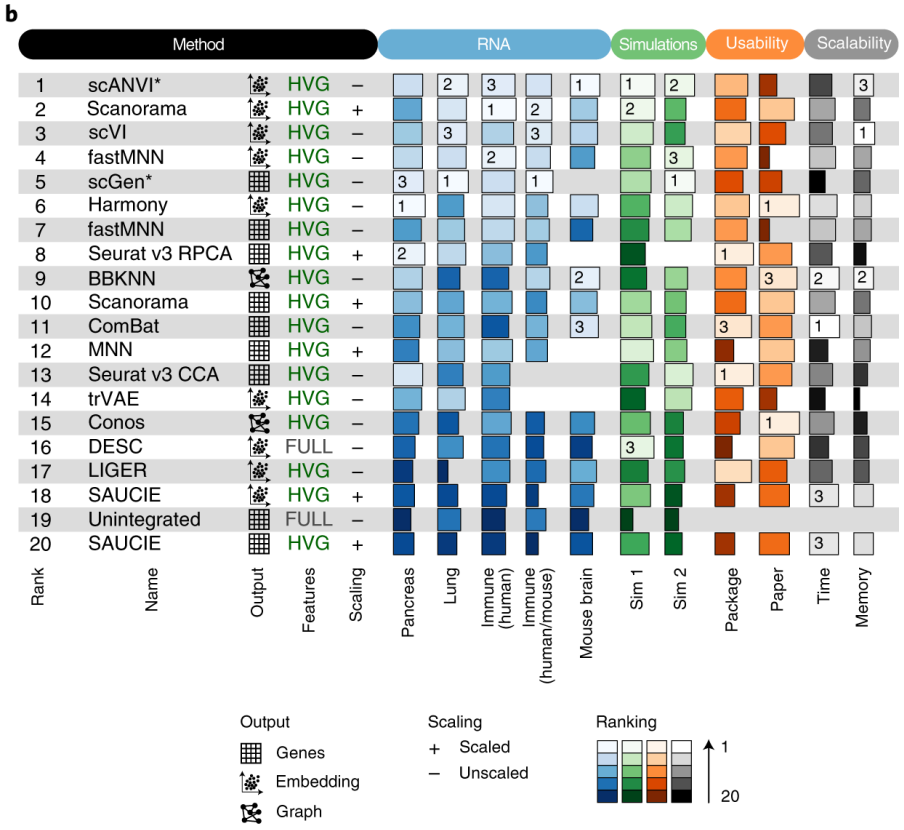




https://www.sc-best-practices.org/cellular_structure/integration.html#dataset

… which method to choose?





Simple tasks → Harmony, Seurat

Harder tasks → scVI, scGen, scANVI, scanorama

https://www.sc-best-practices.org/cellular_structure/integration.html#choosing-an-integration-method

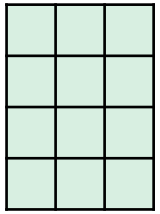
Integration workflow - outline

1. Inspect unintegrated data
2. Decide on meaningful batch covariates for your research question
3. Batch-aware feature selection (~ “highly variable genes” step)
4. Run batch correction method of your choice
5. Work with the integrated dataset

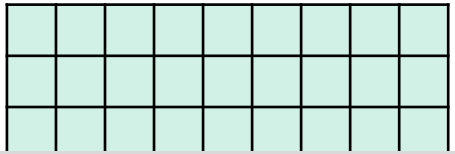
https://www.sc-best-practices.org/cellular_structure/integration.html#dataset

Integration workflow - outline

5 Work with the integrated dataset



(Almost) All methods return a batch-corrected latent space (e.g. 20D PCA space).

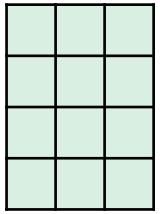


Many methods return a batch-corrected gene expression object (all input genes, e.g. 2k highly variable genes).

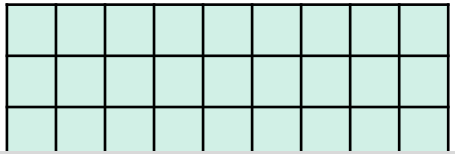
https://www.sc-best-practices.org/cellular_structure/integration.html#dataset

Integration workflow - outline

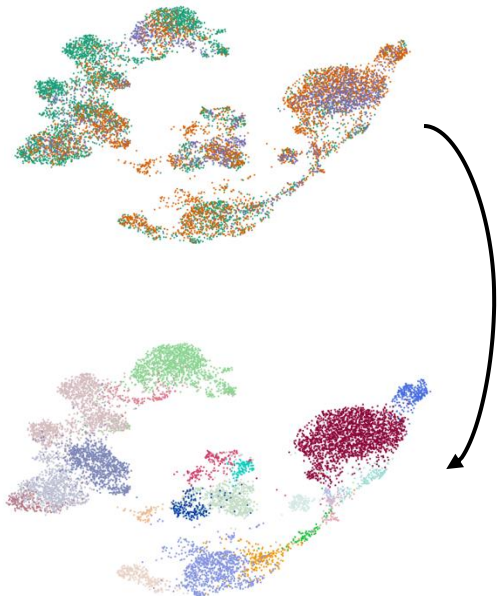
5 Work with the integrated dataset



(Almost) All methods return a batch-corrected latent space (e.g. 20D PCA space).



Many methods return a batch-corrected gene expression object (all input genes, e.g. 2k highly variable genes).



clustering ✓

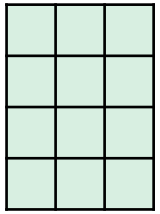
annotation ✓

visualization ✓

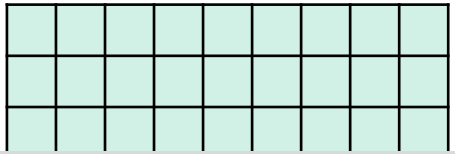
https://www.sc-best-practices.org/cellular_structure/integration.html#dataset

Integration workflow - outline

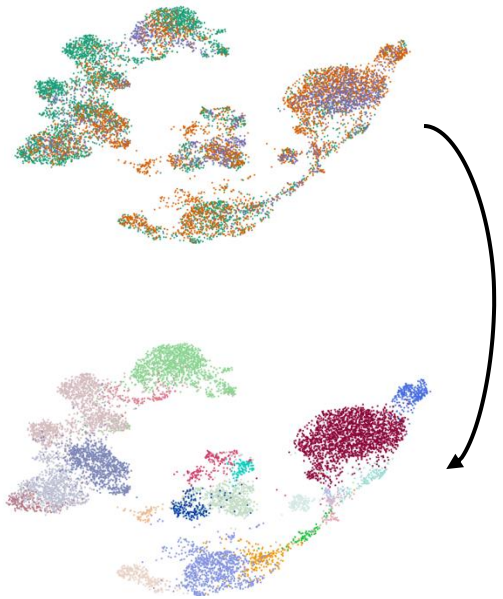
5 Work with the integrated dataset



(Almost) All methods return a batch-corrected latent space (e.g. 20D PCA space).



Many methods return a batch-corrected gene expression object (all input genes, e.g. 2k highly variable genes).



clustering ✓

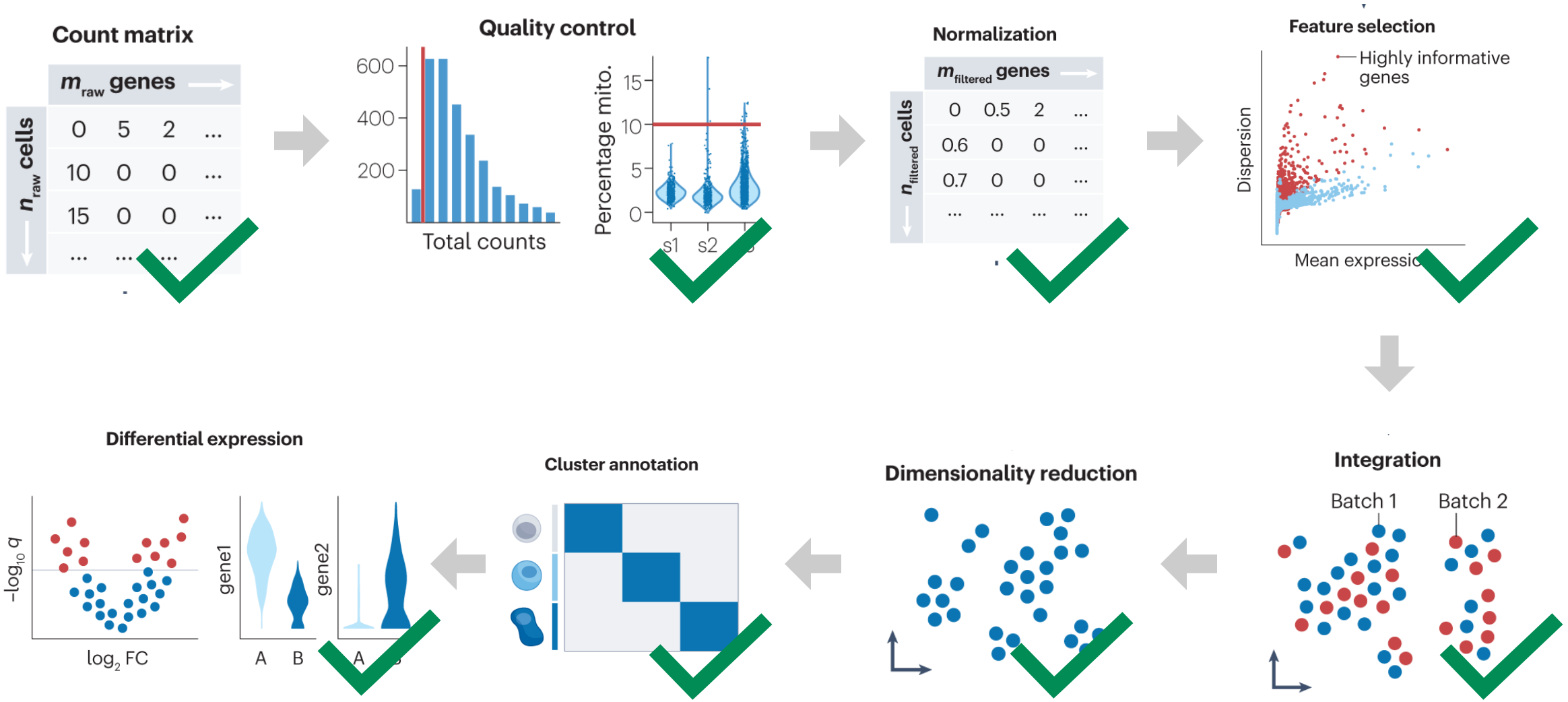
annotation ✓

visualization ✓

Most analysts do not currently use imputed / corrected gene expression matrices for quantitative analysis.

Integrated clusters and **original gene expression values** are typically used for differential gene expression analysis.

https://www.sc-best-practices.org/cellular_structure/integration.html#dataset



Heumos, L., Schaar, A.C., Lance, C. et al. Best practices for single-cell analysis across modalities. Nat Rev Genet 24, 550–572 (2023). https://doi.org/10.1038/s41576-023-00586-w