Single-cell RNA-seq Analysis in Python

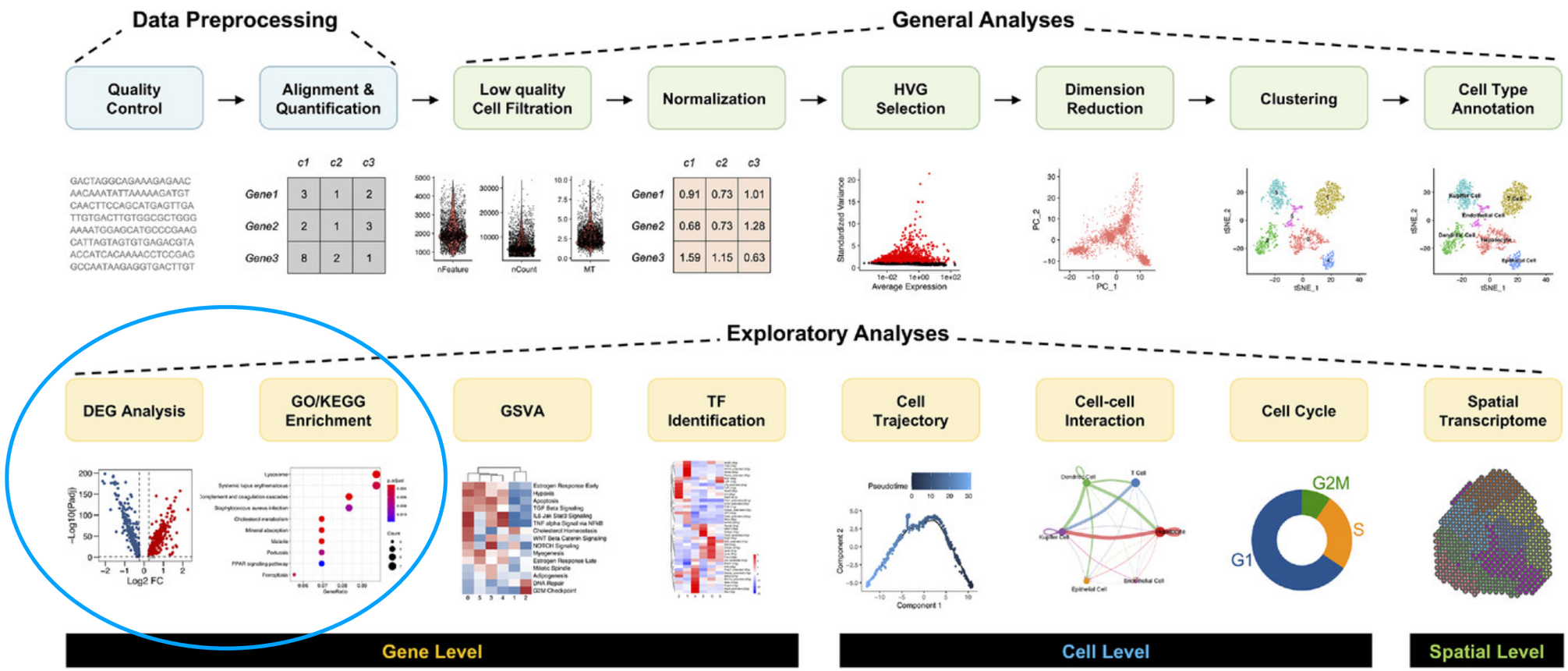
June 2024

Overview of Day #3

Day 3: **Biological Data Interpretation**

- Batch correction and data integration
- Reference mapping
- Differential gene expression
- Gene set enrichment analyses

Roadmap



Jovic et al., 2022, Clinical and Translation Medicine

Batch Correction and Data Integration

What is batch effect?

Changes in measured expression levels that are the result of handling cells in distinct groups or “batches”.

What are the sources of batch effect?

- Technical
  - Sample handling
  - Experimental protocols
  - Sequencing platforms
  - Sequencing depth
- Biological
  - Donor variation
  - Tissue
  - Sampling location

Batch Correction and Data Integration

Why do we need to remove/minimize batch effects?

- Removing batch effects is crucial to enable joint analysis that can focus on finding common structure in the data across batches and enable us to perform queries across datasets.

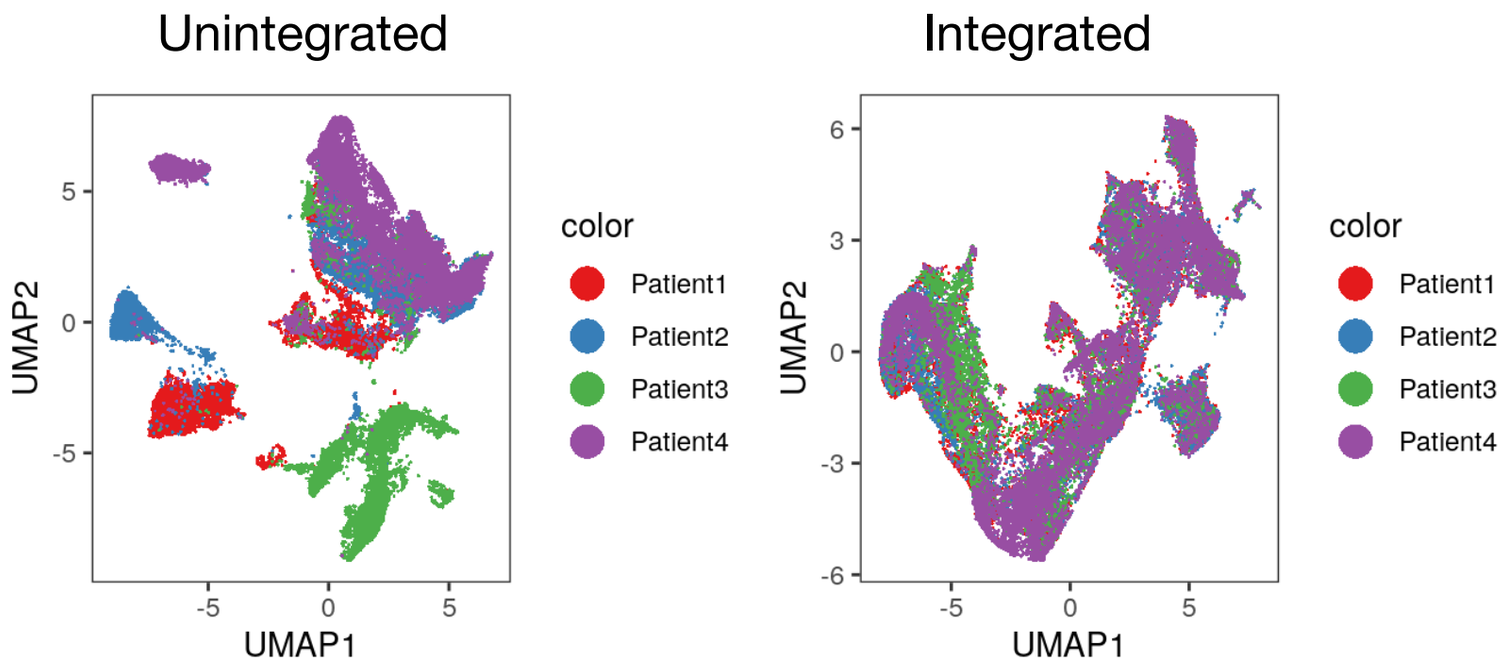
- Often it is only after removing these effects that rare cell populations can be identified that were previously obscured by differences between batches.

- Enabling queries across datasets allows us to ask questions that could not be answered by analysing individual datasets,

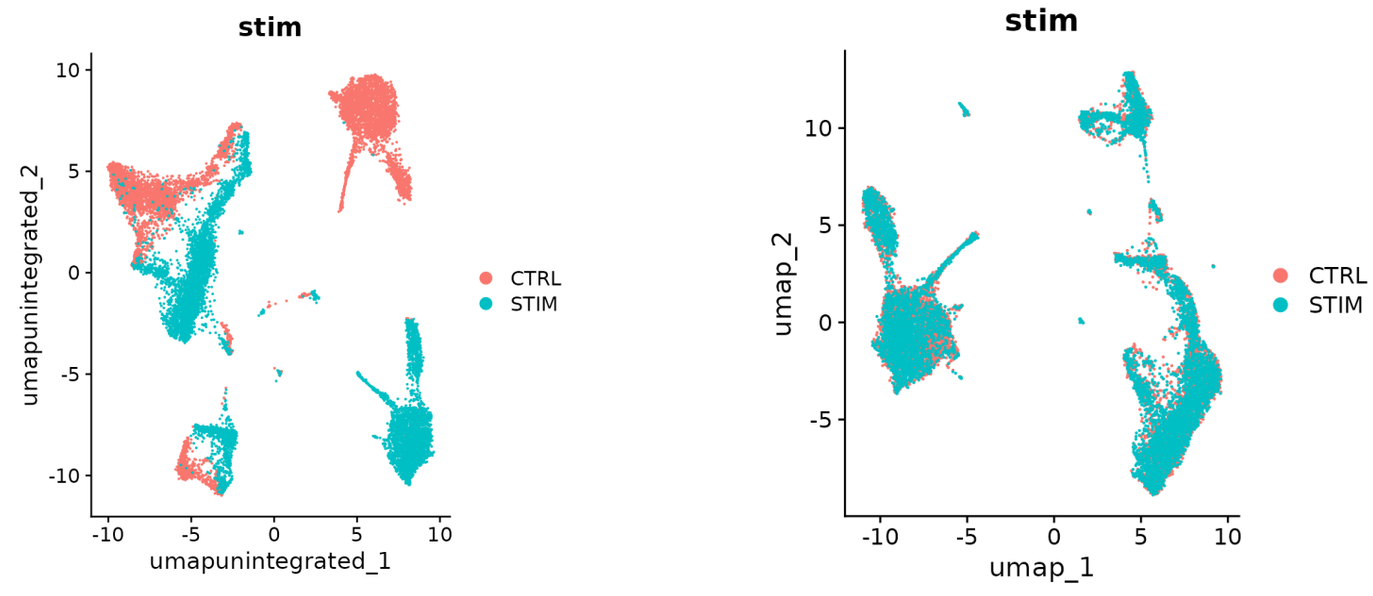
Batch Correction and Data Integration

Examples of batch effect and data integration

- Same tissue, different donors



- Same tissue, different conditions

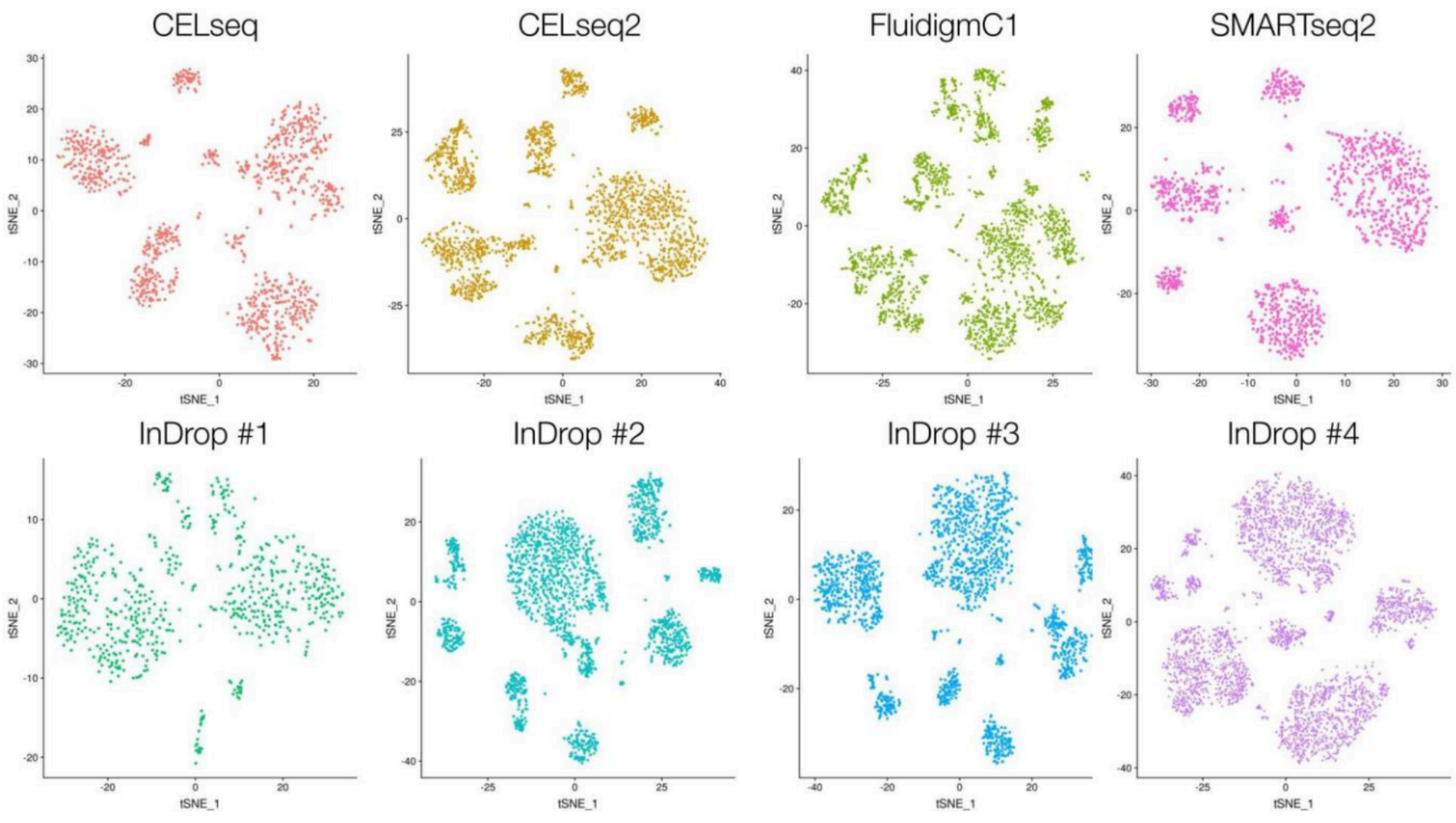


Seurat tutorial

Batch Correction and Data Integration

Examples of batch effect and data integration

- Same tissue, different technologies



Baron et al. 2016, *Cell Syst.*

Lawlor et al. 2017, *Genome Res.*

Grun et al. 2016, *Cell Stem Cell*

Muraro et al. 2016, *Cell Syst.*

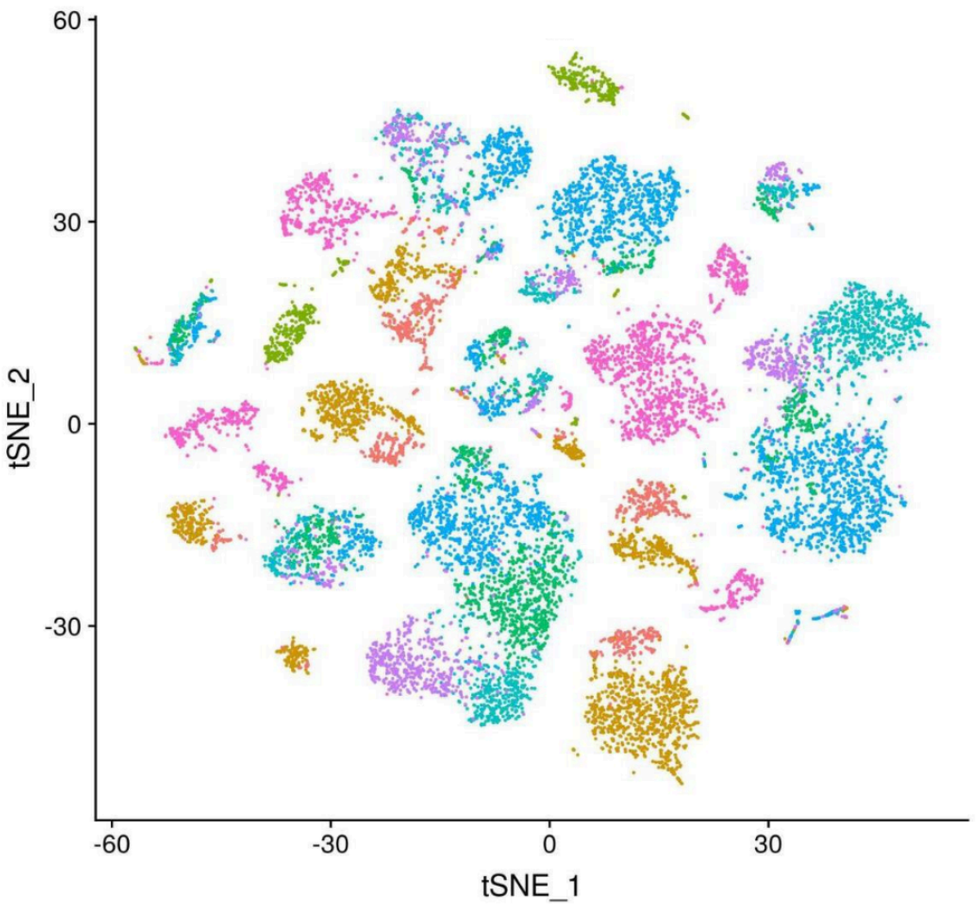
Seurat tutorial

Batch Correction and Data Integration

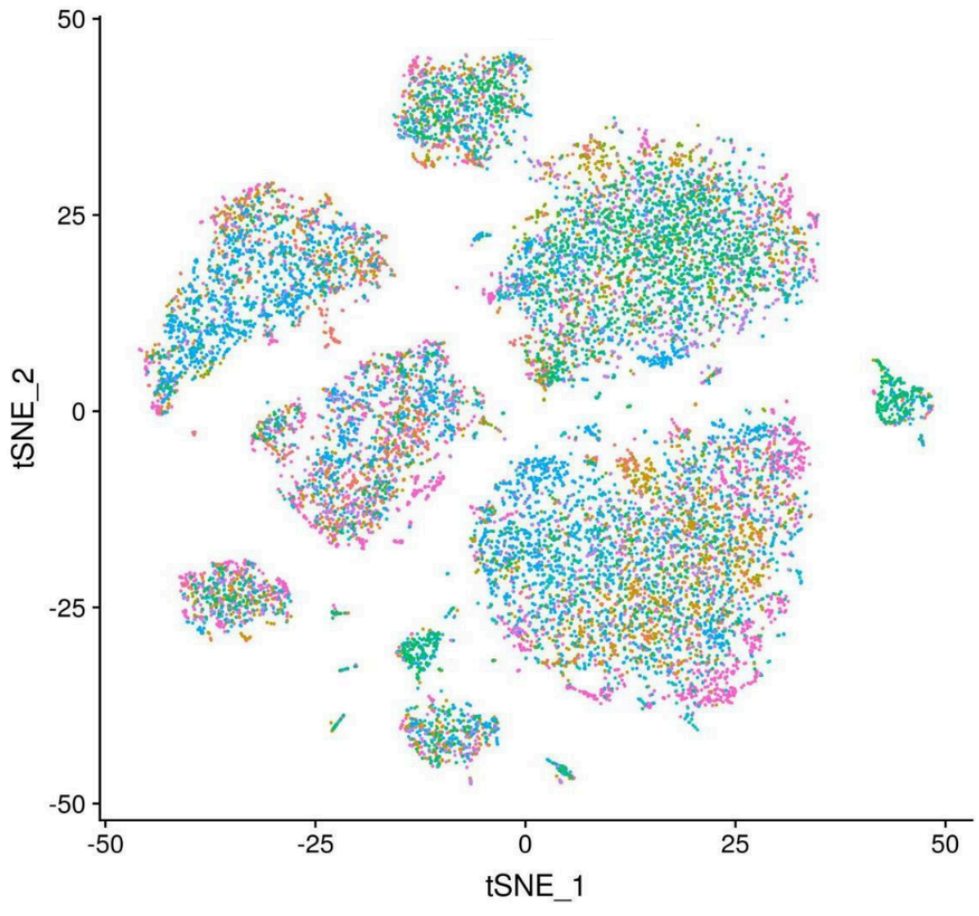
Examples of batch effect and data integration

- Same tissue, different technologies

Unintegrated



Integrated

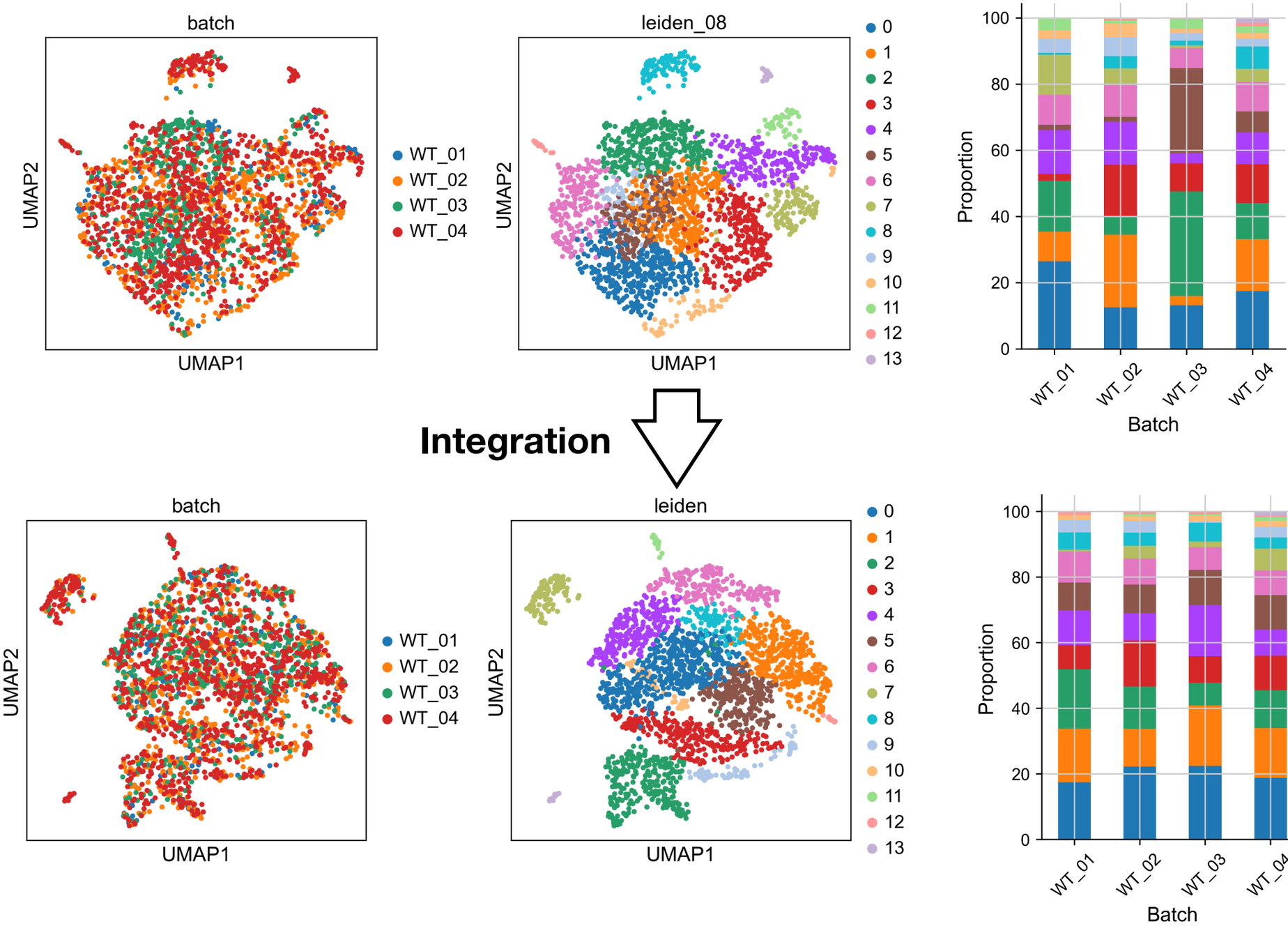


Seurat tutorial

Batch Correction and Data Integration

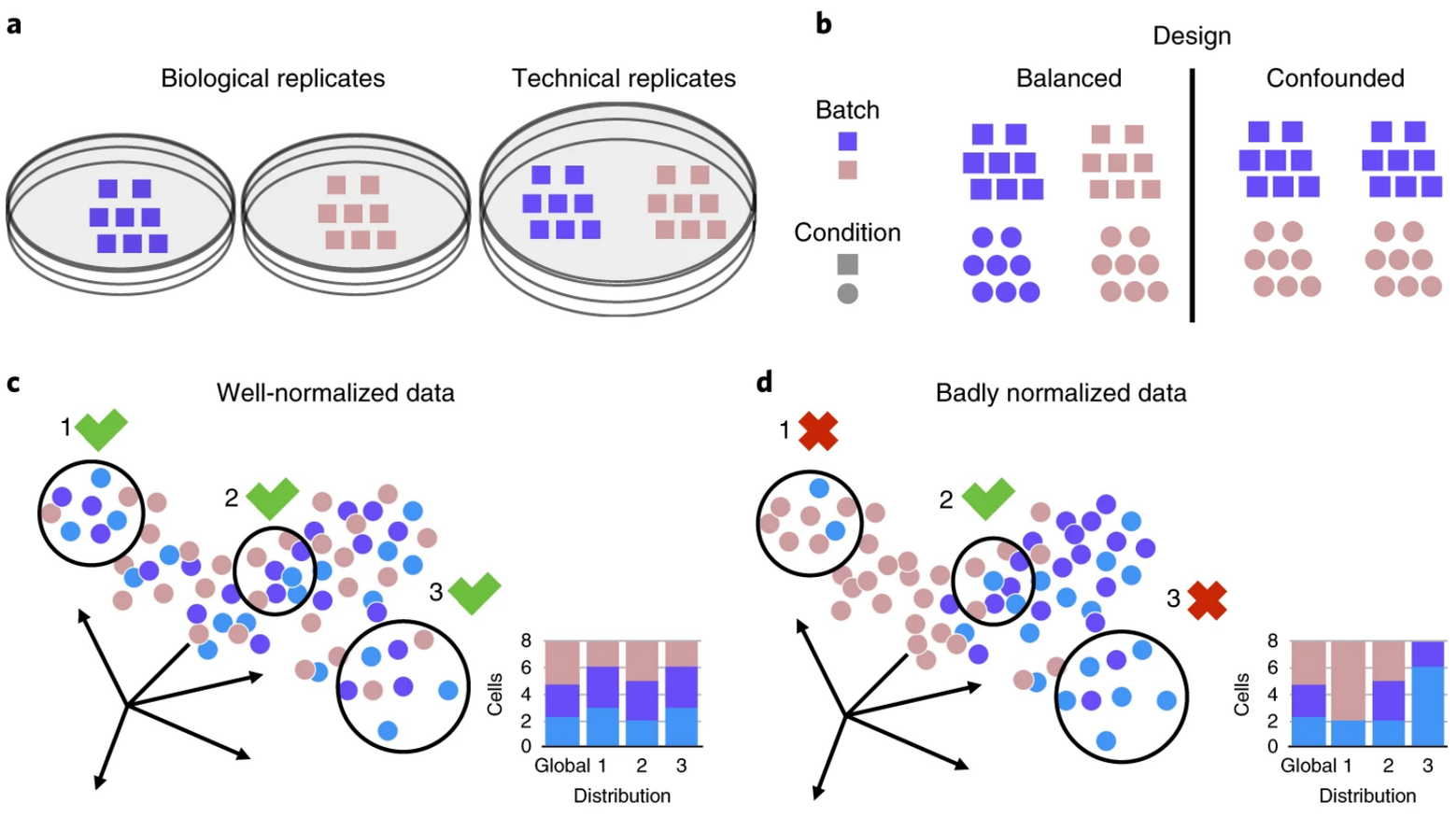
Examples of batch effect and data integration

- Same tissue, same technology, different days



Batch Correction and Data Integration

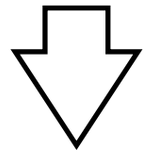
Batch-balanced experiment design



Buettner et al., 2019, Nature Methods

Batch Correction and Data Integration

It is always recommended that you check the raw data before performing any batch correction and data integration



- Whether any batch correction is required
- Determine the batch covariate

Batch Correction and Data Integration

Batch-aware feature selection

Due to the changes in gene expression between batches, we need to modify the highly variable gene selection process to account for these differences.

This is because genes that are variable across the whole dataset could be capturing batch effects rather than the biological signals we are interested in.

```
sc.pp.highly_variable_genes(adata, layer="log1p_norm", batch_key="batch",
n_top_genes=4000)
```

Batch Correction and Data Integration

Batch correction vs. Data integration

Batch correction Maximmaly correct differences between technical replicates

Unsualy corrects the raw expression matrix

A. Linear correction: Limma
B. Bayesian correction: COMBAT

Data integration Correct only inferred technical effects but keep differences that could be biologically meaningful. The aim is to combine data from multiple sources or experiments to create a unified dataset.

Usually operates on embeddings or graphs

A. Linear embedding integration e.g. Harmony, Liger, **Seurat CCA**, Seurat RPCA
B. Graph-based integration e.g. BBKNN
C. Variational autoencoder (VAE)-based integration e.g. scVI, scANVI, trVAE

Batch Correction -> Combat

```
sc.pp.combat(adata, key=batch_key)
```

Negative binomial regression models

Gene-wise model: for a certain gene *g*, count in sample *j* from batch *i* $y_{gij} \sim NB(\mu_{gij}, \phi_{gi})$

$$\log \mu_{gij} = \alpha_g + X_j \beta_g + \gamma_{gi} + \log N_j$$

$$Var(y_{gij}) = \mu_{gij} + \phi_{gi} \mu_{gij}^2$$

Decompose scaled counts into 3 components

- α_g Average level for gene g (in “negative” samples)
- $X_j \beta_g$ Biological condition of sample j
- γ_{gi} Mean batch effect

ϕ_{gi} Dispersion batch effect

Estimate batch effect parameters Estimate parameters using established methods in edgeR

Calculate “batch-free” distributions

We assume the adjusted data also follow a negative binomial distribution: $y^*_{gj} \sim NB(\mu^*_{gj}, \phi^*_g)$

$$\log \mu^*_{gj} = \log \hat{\mu}_{gij} - \hat{\gamma}_{gi}$$

$$\phi^*_g = \frac{1}{N_{batch}} \sum_i \hat{\phi}_{gi}$$

Adjust the data

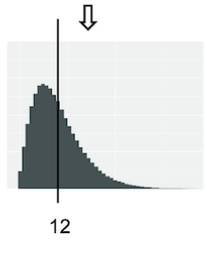
Mapping the data from the estimated empirical distributions to ”batch-free” distributions

Original count matrix

	S 1	S 2
G 1	14	12
G 2	0	0
G 3	112	11
…		

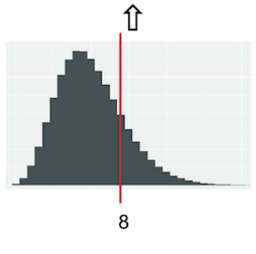
Empirical distribution of original counts:

$y_{gij} \sim NB(\hat{\mu}_{gij}, \hat{\phi}_{gi})$



Adjusted count matrix

	S 1	S 2
G 1	9	8
G 2	0	0
G 3	60	47
…		



Batch-free distribution for adjusted counts:

$y^*_{gj} \sim NB(\mu^*_{gj}, \phi^*_g)$

Zhang et al., 2020, NAR genom. bioinform.

Data Integration

MNNcorrect(https://doi.org/10.1038/nbt.4091)

RPCA + anchors (Seurat v3)(https://doi.org/10.1101/460147)

CCA+ anchors (Seurat v3) (https://doi.org/10.1101/460147)

CCA+ dynamic time warping (Seurat v2) (https://doi.org/10.1038/nbt.4096)

LIGER (https://doi.org/10.1101/459891)

Harmony(https://doi.org/10.1101/461954)

Conos(https://doi.org/10.1101/460246)

Scanorama(https://doi.org/10.1101/371179)

scMerge(https://doi.org/10.1073/pnas.1820006116)

STACAS (https://doi.org/10.1093/bioinformatics/btaa755)

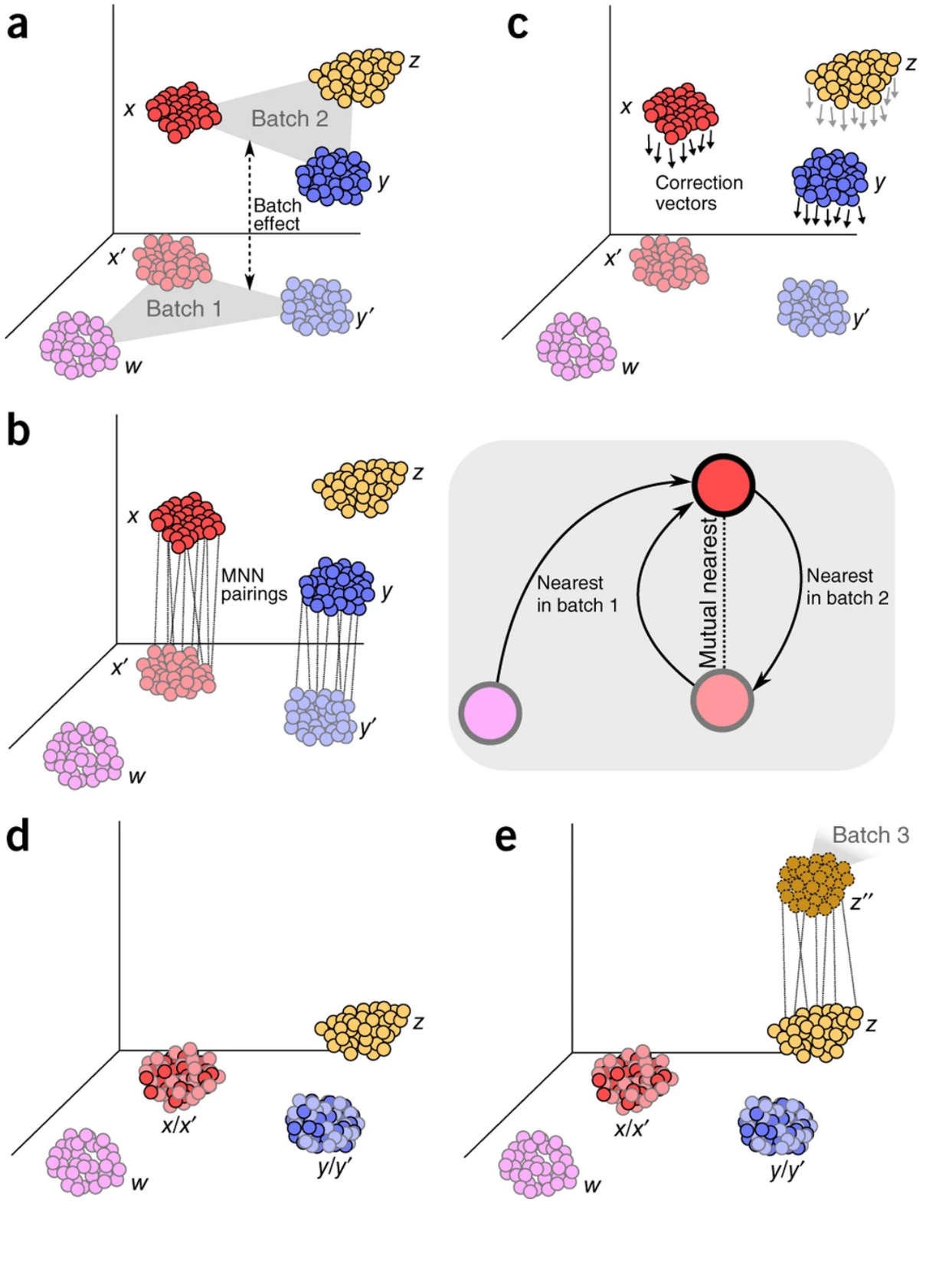
Butler et al., 2018, Nature Biotech.

Data Integration

1. Find corresponding cells across datasets (by computing a distance between cells in a certain space)
2. Compute a data adjustment based on correspondences between cells
3. Apply the adjustment

Data Integration

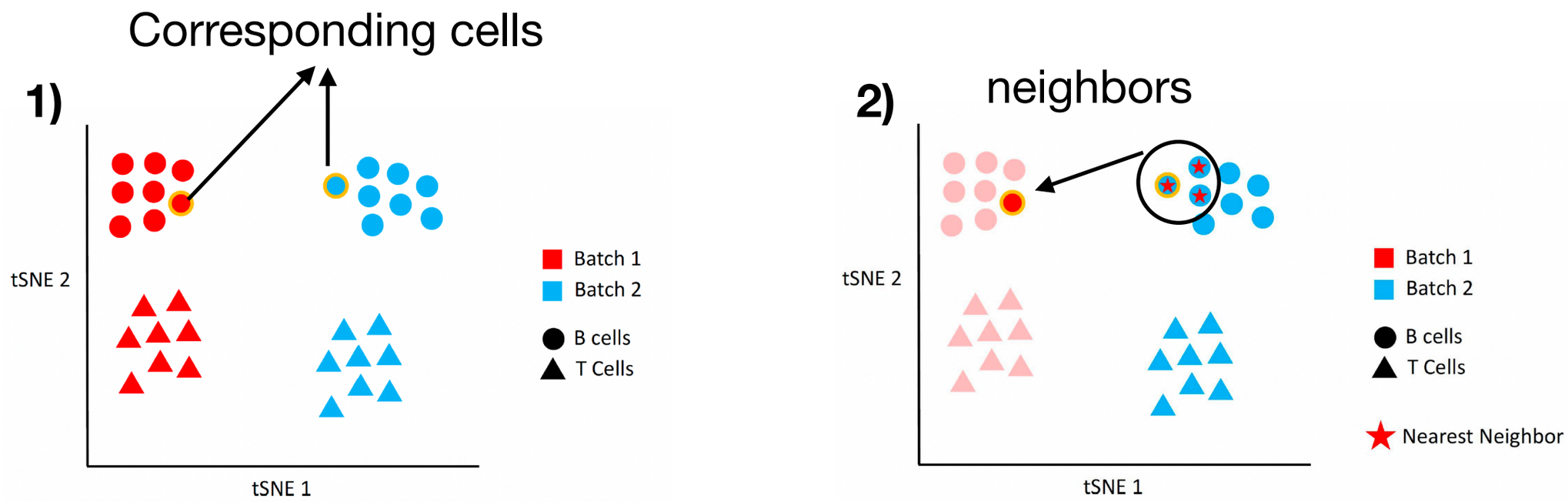
Mutual Nearest Neighbors (MNN)



Haghverdi et al., 2018, Nature Biotech.

Data Integration

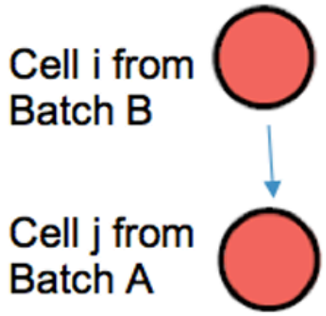
Mutual Nearest Neighbors (MNN)



Haghverdi et al., 2018, Nature Biotech.

Data Integration

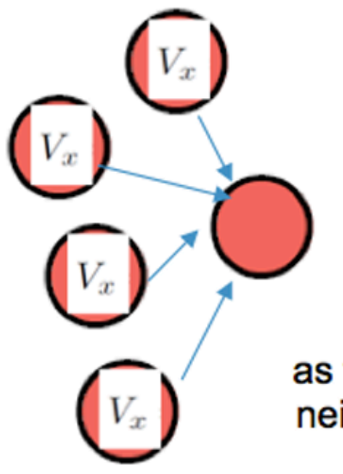
Mutual Nearest Neighbors (MNN)



1) For each MNN pair, a pair-specific batch-correction vector is computed as the vector difference between the expression profiles of the paired cells.

$$V_x = \begin{pmatrix} gene1_a - gene1_b \\ gene2_a - gene2_b \\ gene3_a - gene3_b \\ \dots \\ geneN_a - geneN_b \end{pmatrix}$$

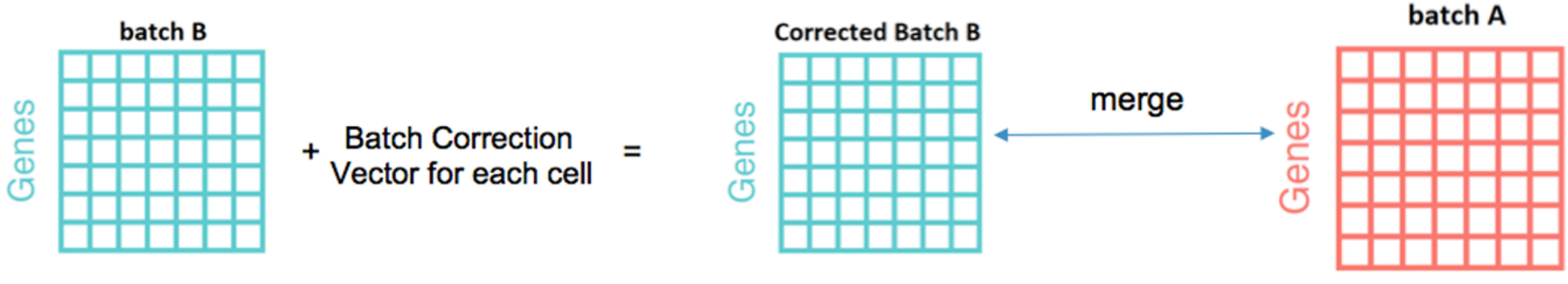
2) A cell-specific batch-correction vector is then calculated as a weighted average of these pair-specific vectors, as computed with a Gaussian kernel.



Gaussian Kernel Smoothing

Real valued function
$f : R^p \rightarrow R$
as the weighted average of neighboring observed data

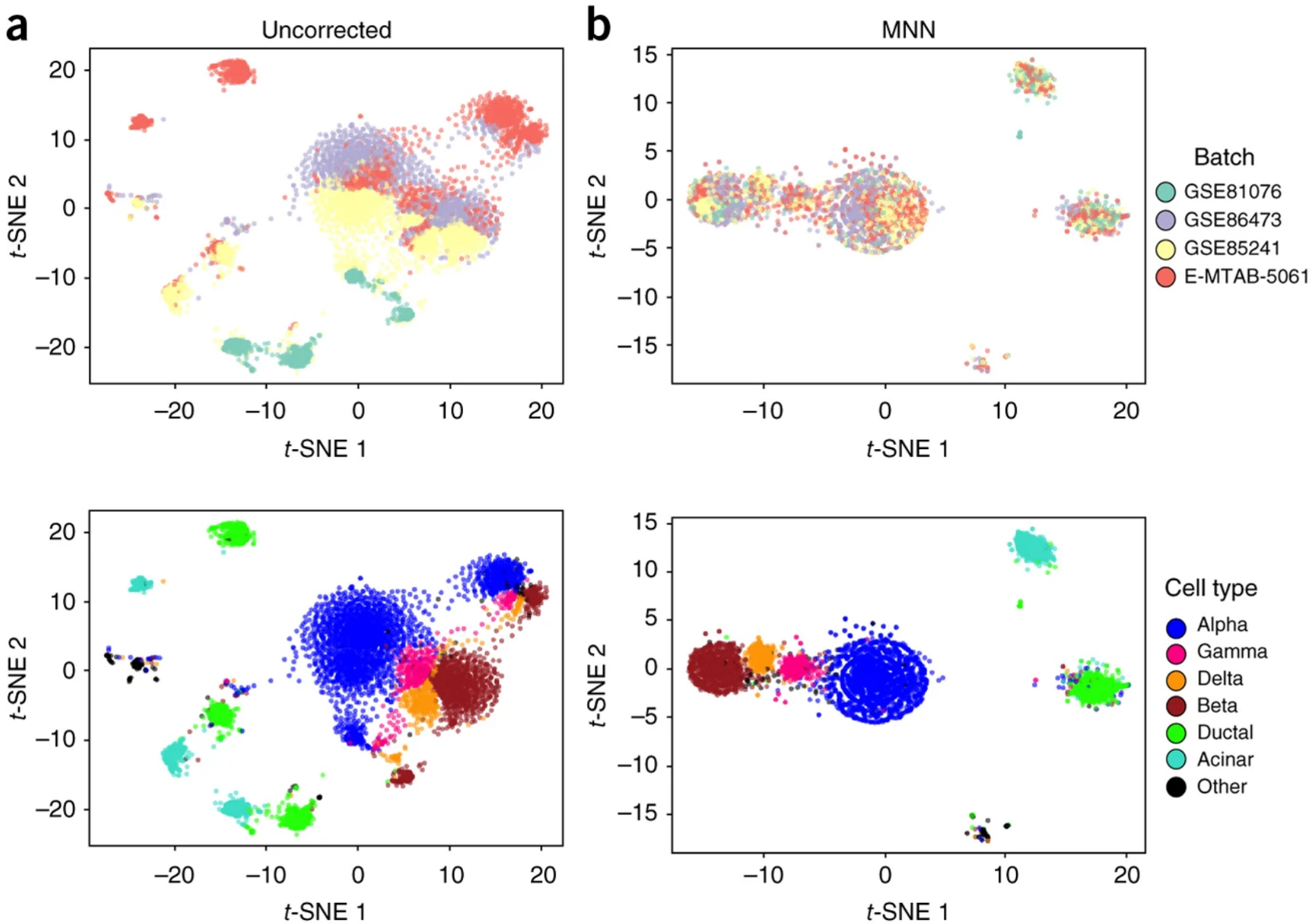
Batch Correction vector for each cell



Haghverdi et al., 2018, Nature Biotech.

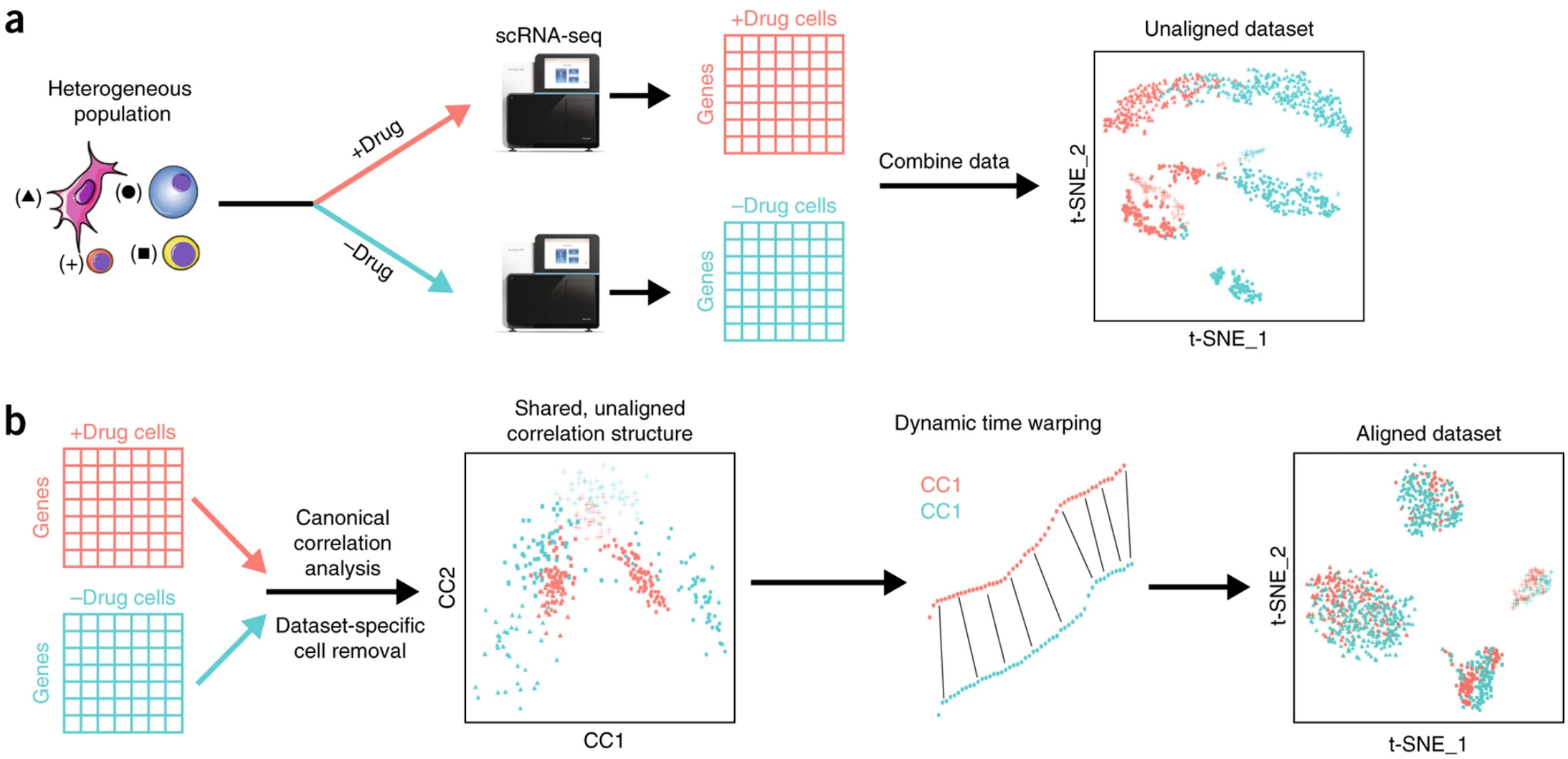
Data Integration

Mutual Nearest Neighbors (MNN)



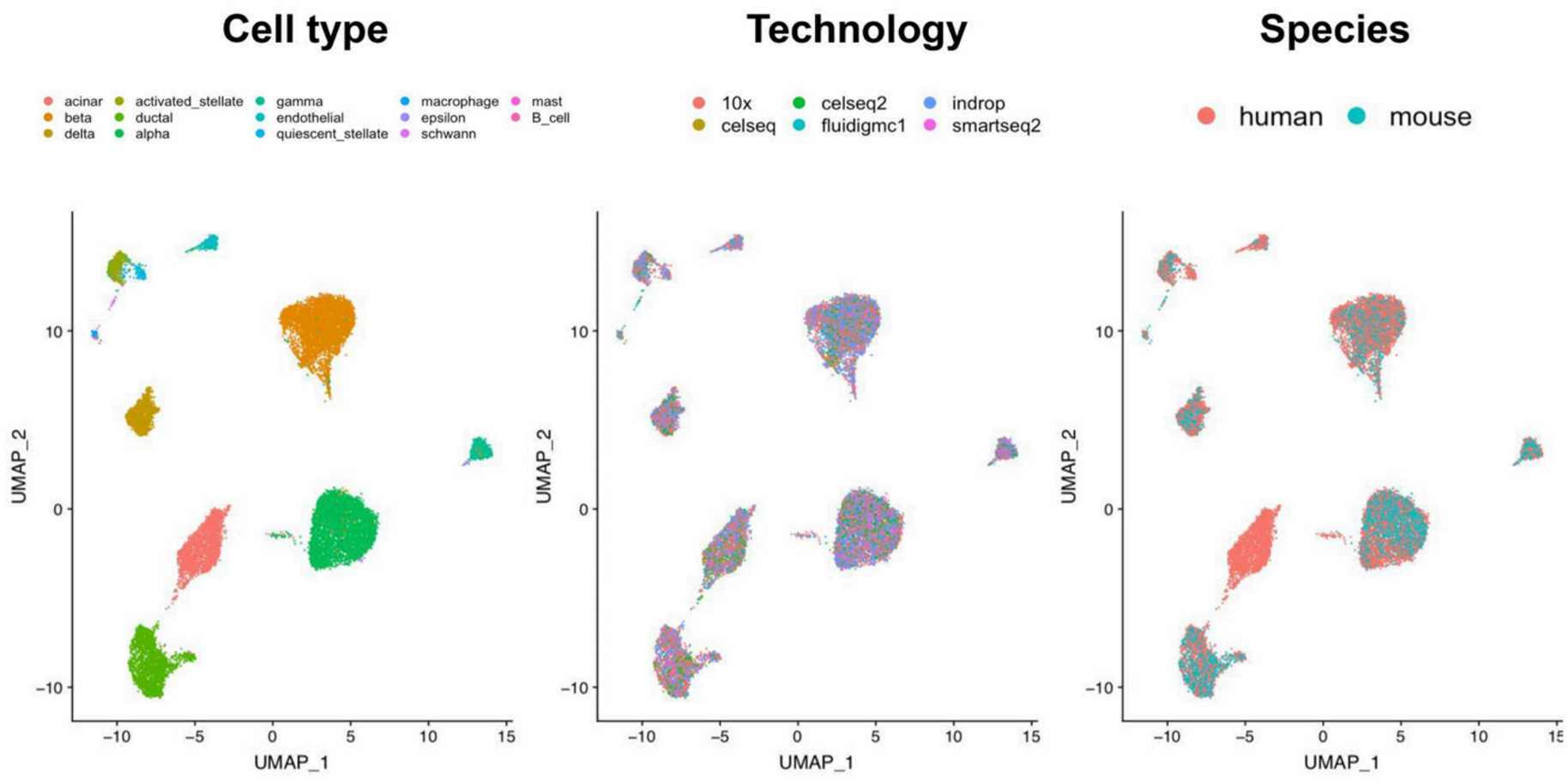
Haghverdi et al., 2018, Nature Biotech.

Data Integration -> CCA + anchors (Seurat V3)

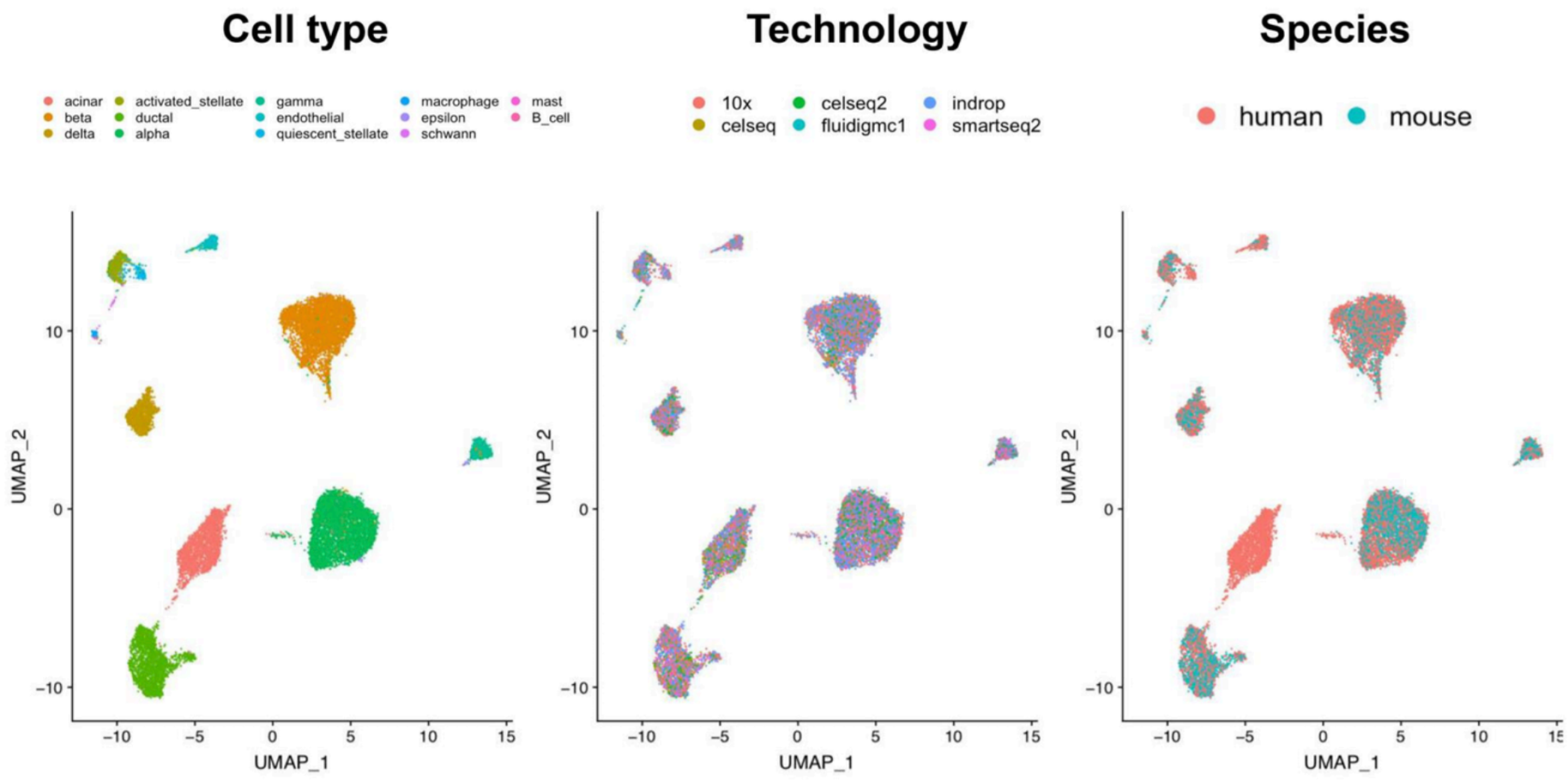


Butler et al., 2018, Nature Biotech.

Data Integration -> CCA + anchors (Seurat V3)

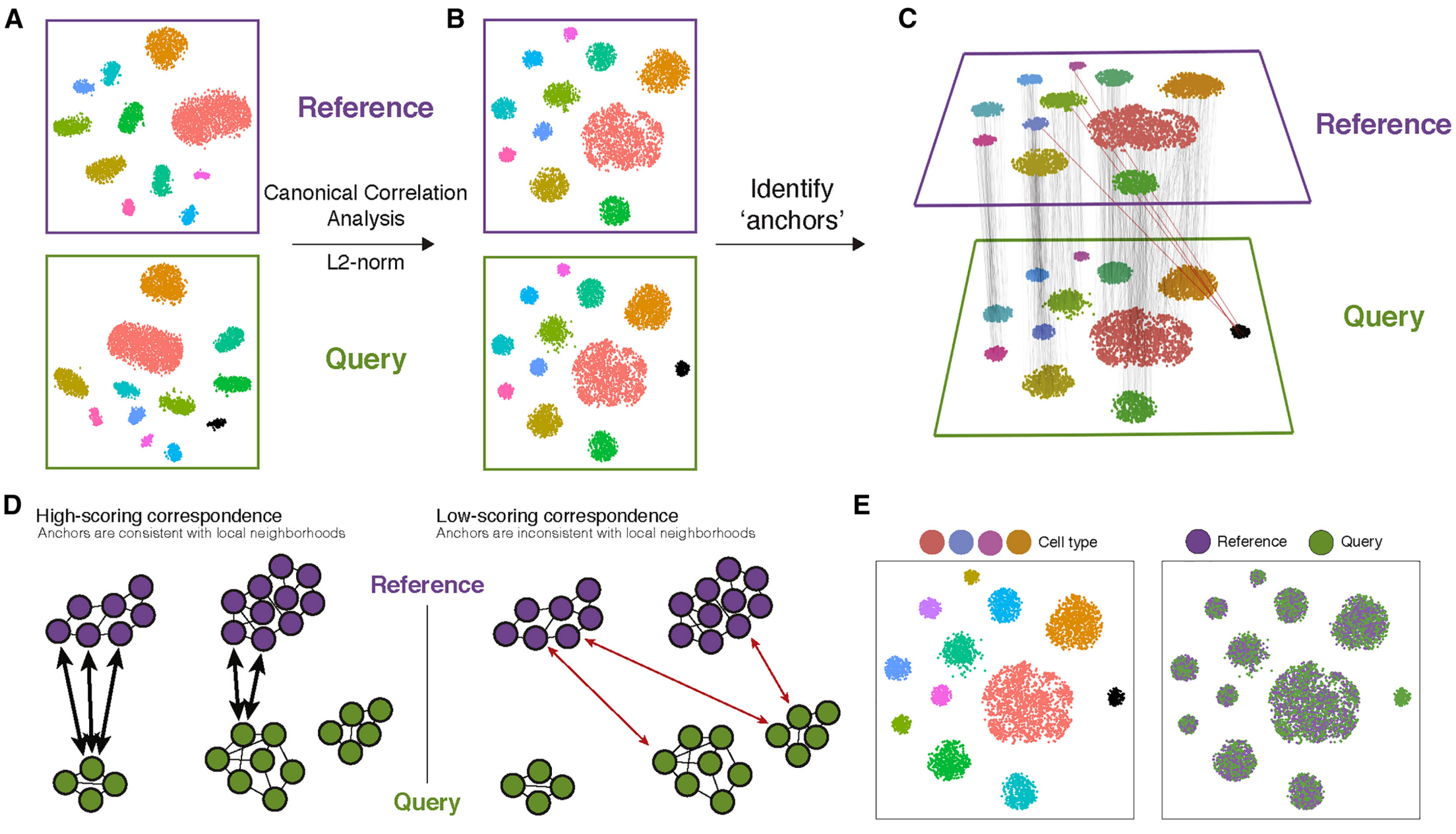


Data Integration -> CCA + anchors (Seurat V3)



It is not recommended to use the corrected gene expression vectors for performing differentially expressed genes analysis

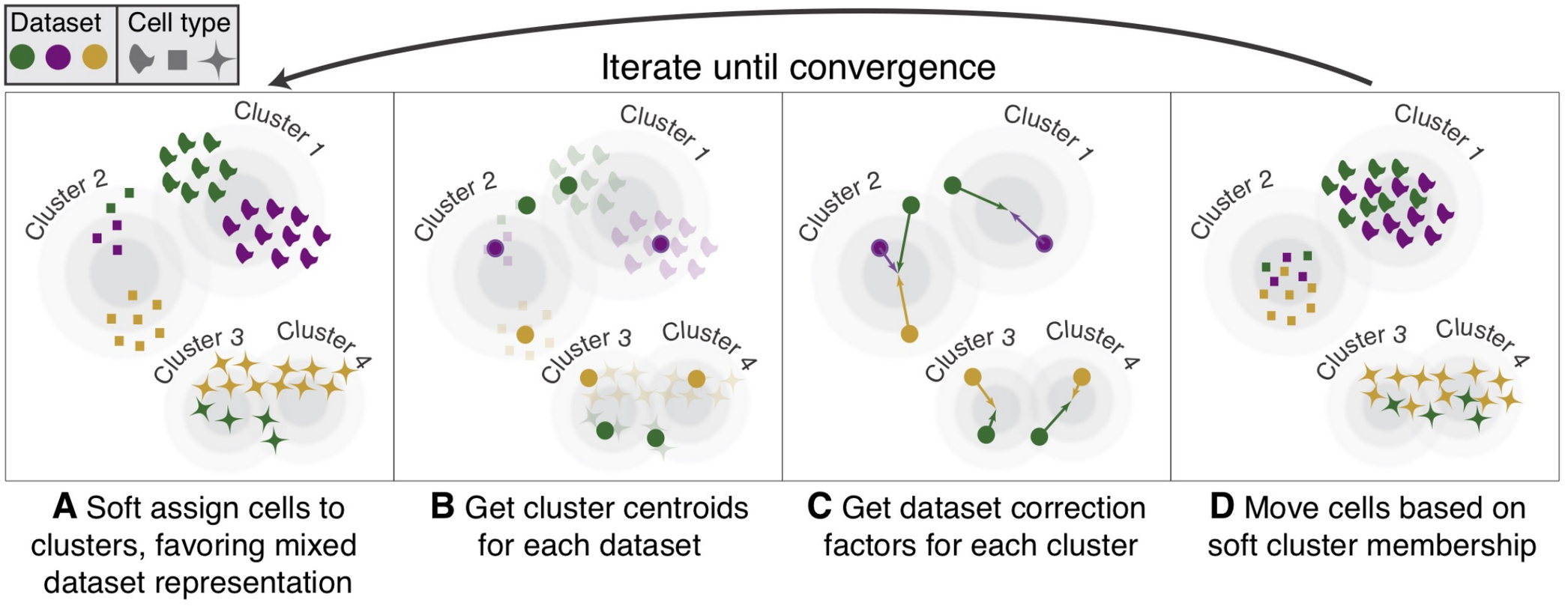
Label Transfer -> CCA + anchors (Seurat V3)



Data Integration -> Harmony

Instead of using CCA, Harmony applies a transformation to the principal component (PCs) values, using all available PCs, e.g. as pre-computed within the Seurat workflow

Iterative clustering



Korsunsky et al., 2019, Nature Methods

Question

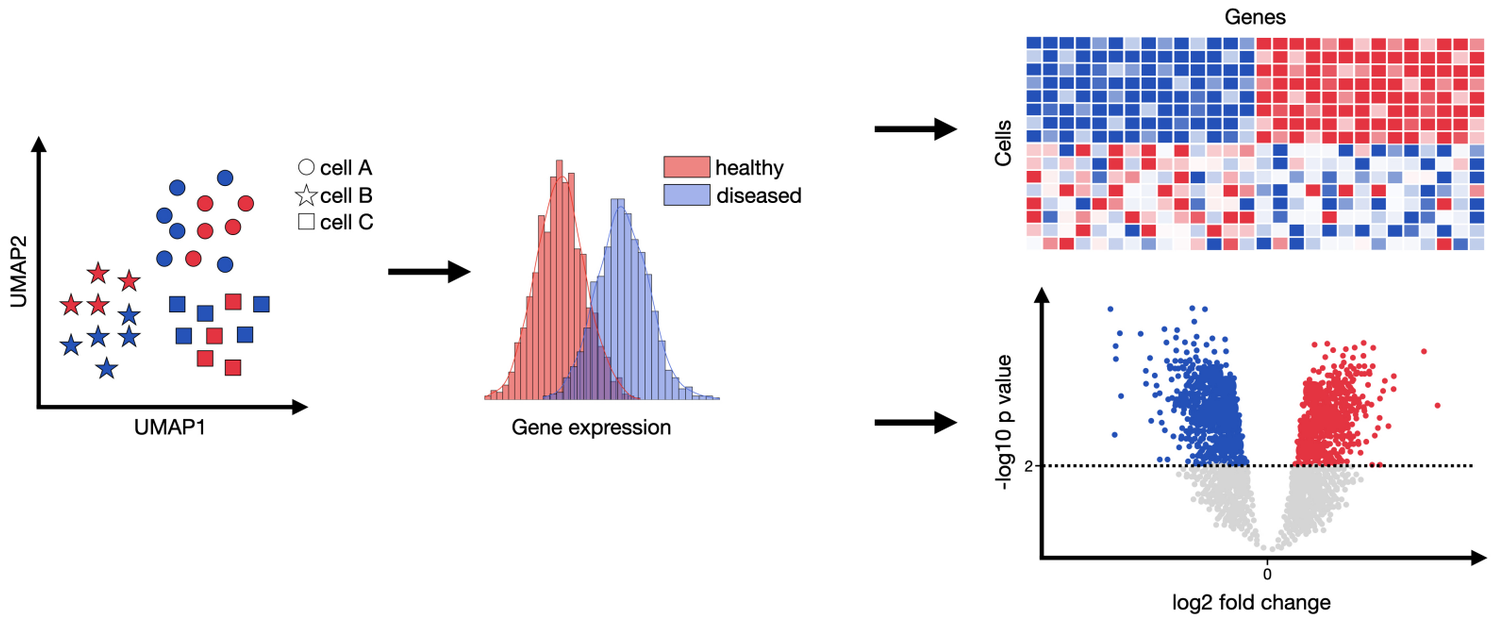
Integration is useful:

In order to have a corrected table of expression used for further differential gene expression of data coming from two or more sources

When I want to obtain a visual representation of data coming from two or more sources

Both of the above

Differential Expression Gene Analysis



- Marker gene identification: genes overexpressed by each cell type, cell cluster, ..., within the dataset => can help in cell type annotation

- Differential gene expression analysis: genes impacted by experimental conditions within a cell type, cell cluster, ..., etc

https://www.sc-best-practices.org/

Differential Expression Gene Analysis

Approaches:

- **Sample-level:** The expression is aggregated to create “pseudobulks”, then, analysed by methods originally designed for bulk RNA-seq such as DESeq2

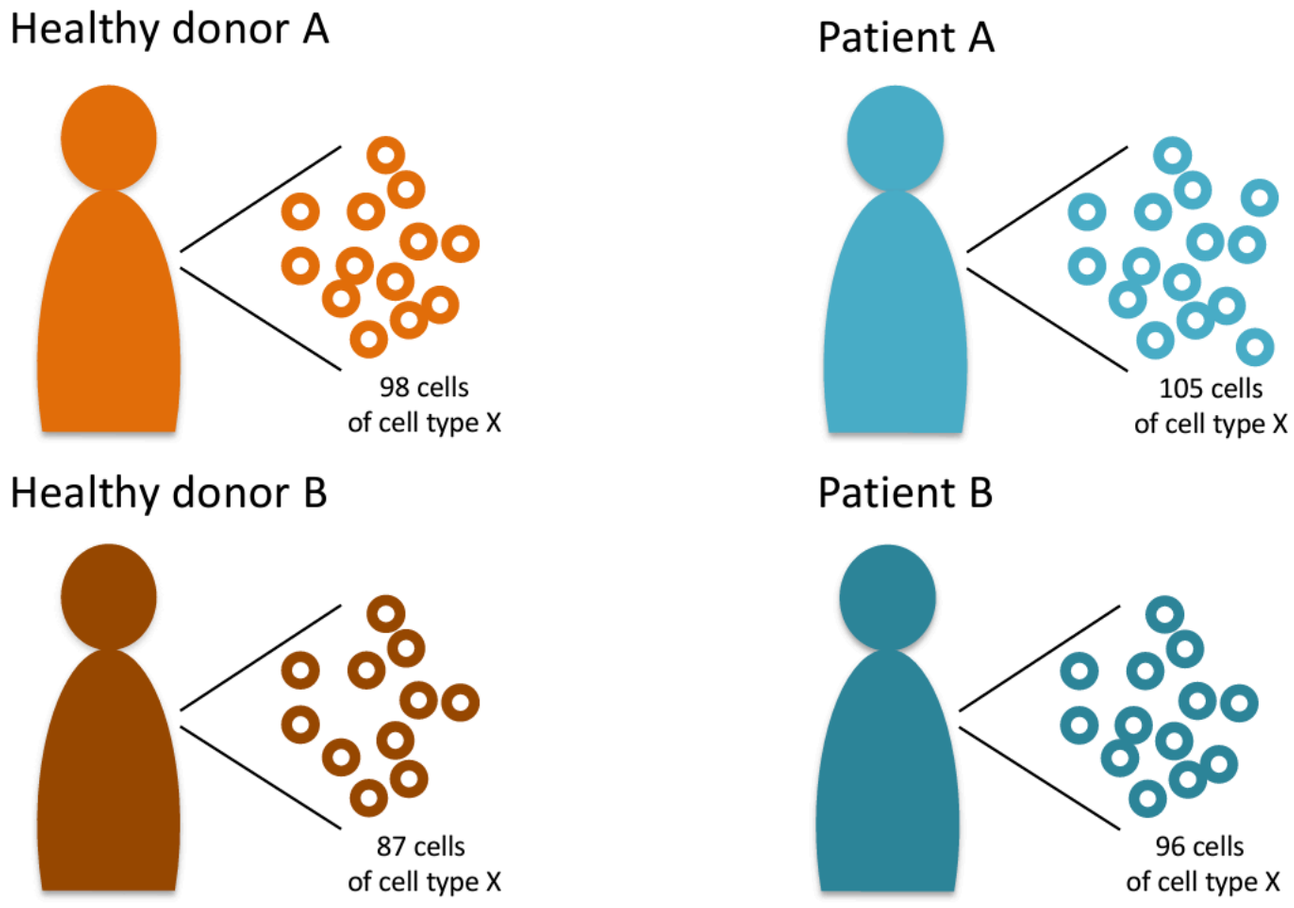
- **Cell-level:** Cells are modeled individually using generalized mixed effect models such as MAST

Love et al., 2014, Genome Biol.
Das et al., 2021, Genes (Basel)

DESeq2

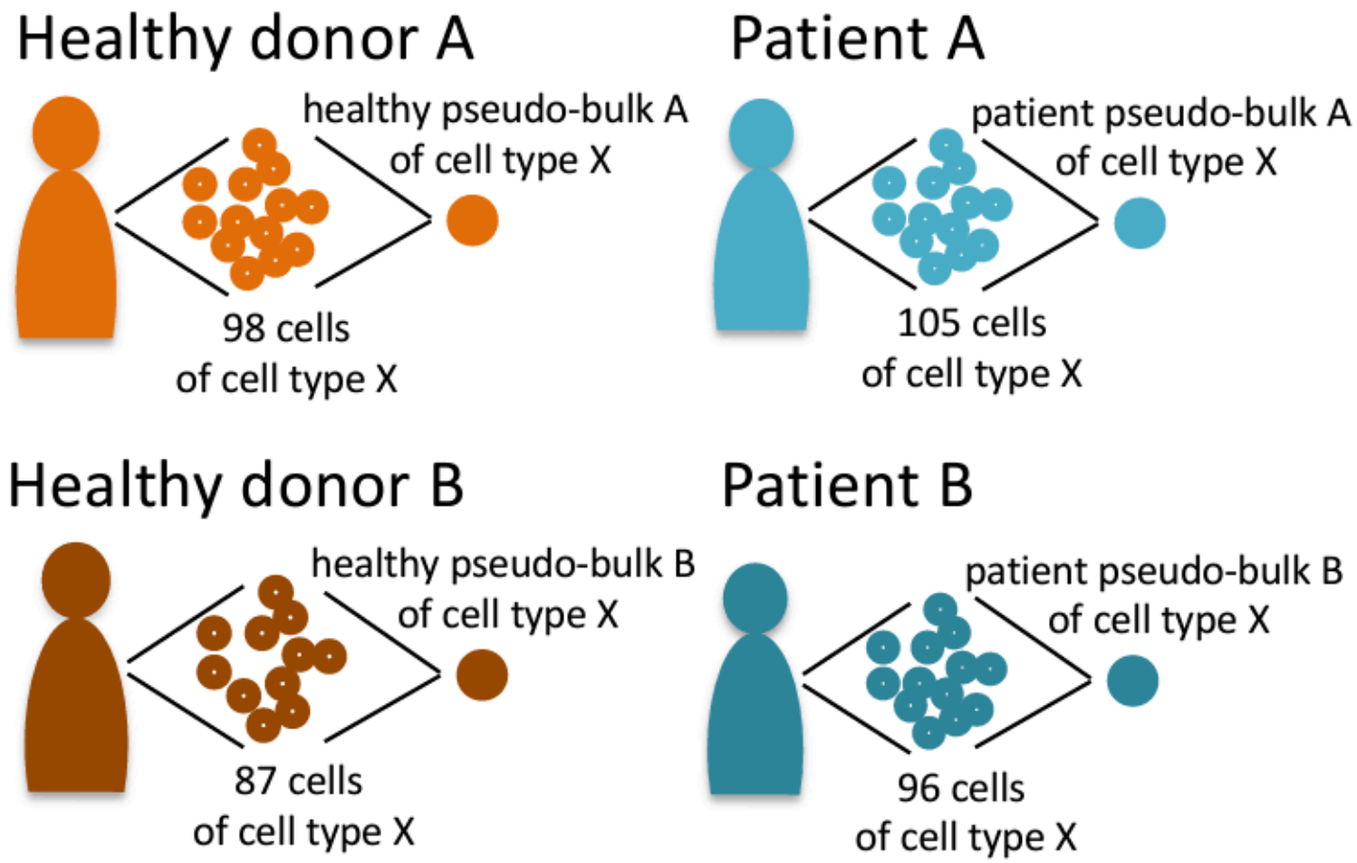
We can include complex designs i.e. if we have batch effects we can include that in our model as covariate

How many independent replicates do we have,
~200 or 2 replicates per condition?



DESeq2

We need to create pseudobulk per cell type per condition/sample



Gene set enrichment and pathway analysis

Goal: to gain biologically meaningful insights from long gene lists – test if differentially expressed genes are enriched in genes associated with a particular function

Depending on the type of the enrichment test chosen, gene expression measurements may or may not be used for the computation of the test statistic.

Gene set enrichment and pathway analysis

- test if differentially expressed genes are enriched in genes associated with a particular function
- test a small number of gene sets, or a large collection of gene sets

What is gene set

- Genes working together in a pathway (e.g. energy release through Krebs cycle)
- Genes located in the same compartment in a cell (e.g. all proteins located in the cell nucleus)
- Proteins that are all regulated by a same transcription factor
- Custom gene list that comes from a publication and that are down-regulated in a mutant List of genes that contain SNPs associated with a disease

MSigDB (https://www.gsea-msigdb.org/gsea/msigdb/index.jsp) is the most comprehensive database

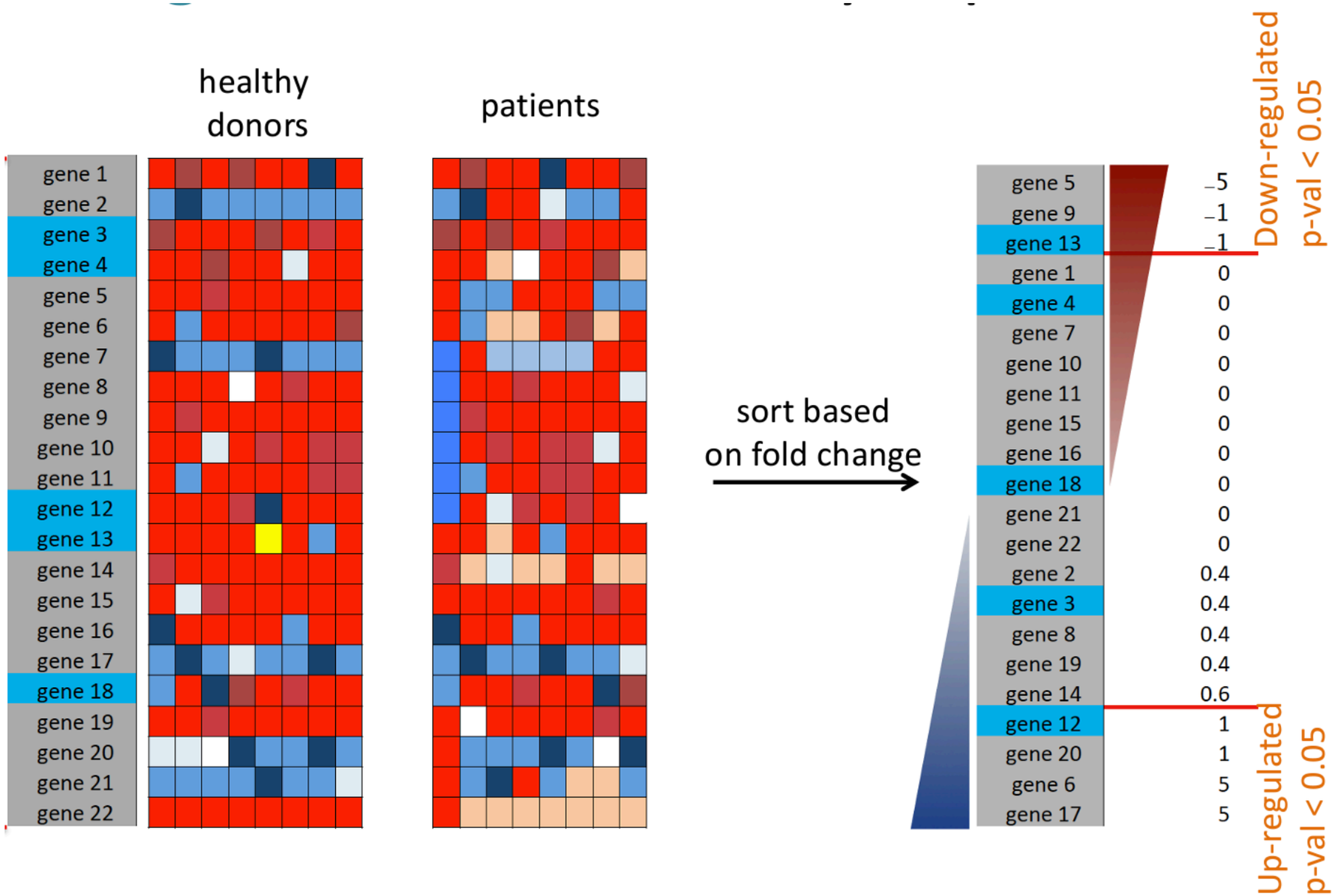
Gene set enrichment and pathway analysis

Example: Gene Ontology gene sets -> Biological Process

	Gene Set	Genes
0	'De Novo' AMP Biosynthetic Process (GO:0044208)	[ATIC, PAICS, PFAS, ADSS1, ADSS2, GART]
1	'De Novo' Post-Translational Protein Folding (...	[SDF2L1, HSPA9, CCT2, HSPA6, ST13, ENTPD5, HSP...
2	2-Oxoglutarate Metabolic Process (GO:0006103)	[IDH1, PHYH, GOT2, MRPS36, GOT1, IDH2, ADHFE1,...
5	3'-Phosphoadenosine 5'-Phosphosulfate Metaboli...	[PAPSS1, PAPSS2, SULT1A2, SULT1C4, SULT2A1, SU...
3	3'-UTR-mediated mRNA Destabilization (GO:0061158)	[UPF1, TRIM71, RC3H1, ZFP36L1, ZFP36L2, MOV10,...
4	3'-UTR-mediated mRNA Stabilization (GO:0070935)	[DAZ4, TIRAP, DAZ3, YBX3, DAZ2, DAZ1, ELAVL1, ...
6	5S Class rRNA Transcription By RNA Polymerase ...	[GTF3C2, GTF3C3, GTF3C4, GTF3C5, GTF3C6]
8	7-Methylguanosine Cap Hypermethylation (GO:003...	[SNRPD3, SNRPD2, SNRPD1, TGS1, SNRPF, SNRPG, S...
7	7-Methylguanosine RNA Capping (GO:0009452)	[RNGTT, RNMT, NCBP1, CMTR2, RAMAC]
9	7-Methylguanosine mRNA Capping (GO:0006370)	[RNGTT, RNMT, NCBP1, CMTR2, RAMAC, RAMACL]

Gene Set Enrichment Analysis

Over-representation analysis - Enrichr



Gene Set Enrichment Analysis

Over-representation analysis - Enrichr

```
        Fisher's Exact Test for Count Data

data:  cont.table
p-value = 1
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  0.1012333 18.7696686
sample estimates:
odds ratio
   1.56456
```

2X2 count table	Differentially expressed	Not Differentially expressed	total
blue	2	3	**5**
Not blue	5	12	**17**
total	**7**	**15**	**22**

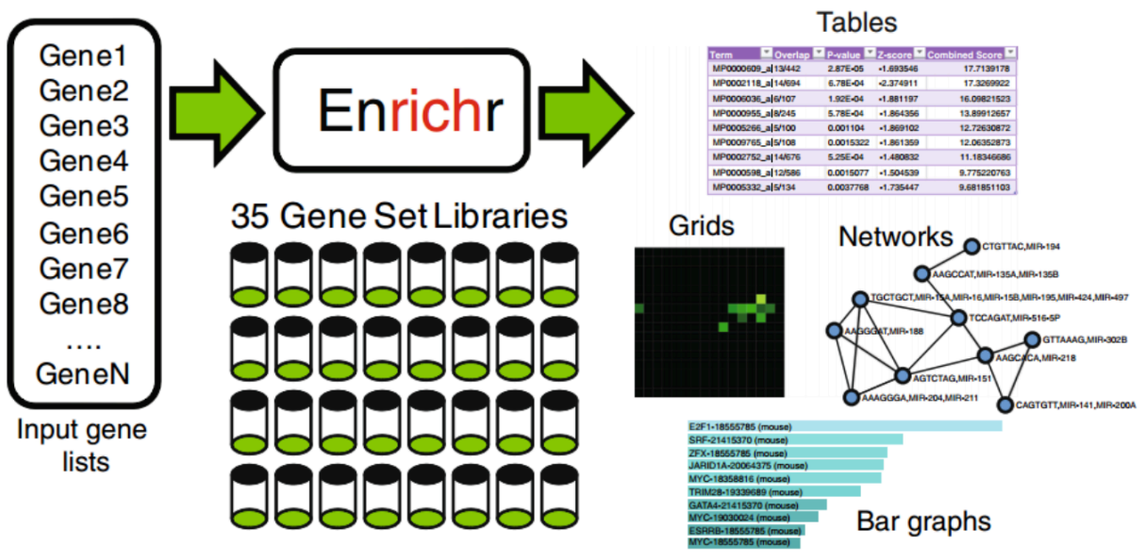
2/7 = 0.29

3/15 = 0.20

Gene set enrichment and pathway analysis

Over-representation analysis - Enrichr

1. Subset the specific cell type
2. Run DE analysis
3. Subset the genes that are significantly different between conditions (with adjusted p-values) and are either upregulated or downregulated (with log2 fold change values)
4. Select a relevant collection of gene sets e.g. 'GO_Biological_Process_2021', 'GO_Cellular_Component_2021', 'GO_Molecular_Function_2021', 'KEGG_2021_Human', 'WikiPathways_2021_Human',



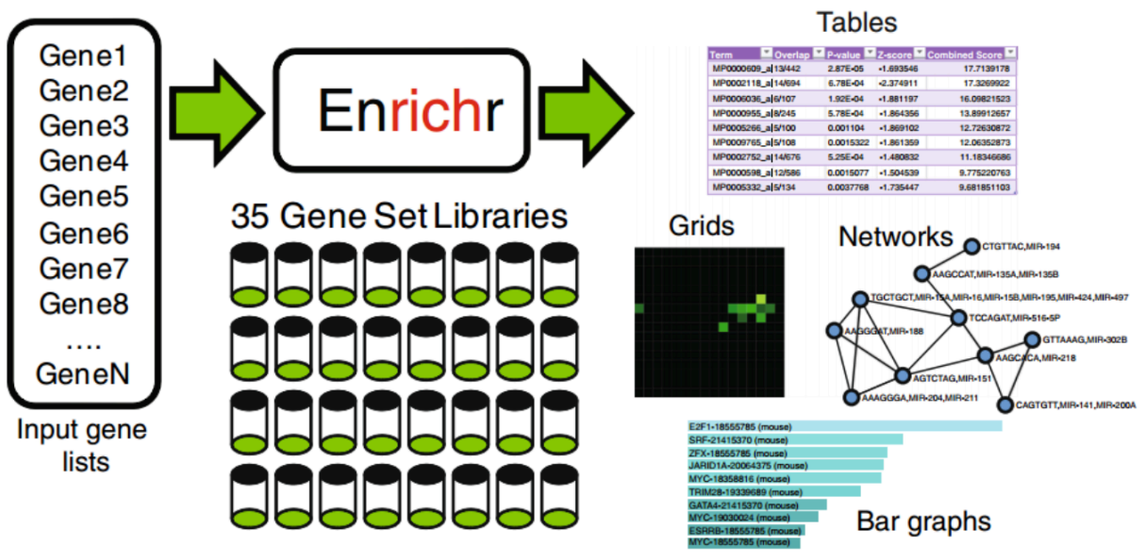
You just need to provide the list of the genes for the enrichr function

https://gseapy.readthedocs.io/en/latest/singlecell_example.html

Gene set enrichment and pathway analysis

Over-representation analysis - Enrichr

1. Subset the specific cell type
2. Run DE analysis
3. Subset the genes that are significantly different between conditions (with adjusted p-values) and are either upregulated or downregulated (with log2 fold change values)
4. Select a relevant collection of gene sets e.g. 'GO_Biological_Process_2021', 'GO_Cellular_Component_2021', 'GO_Molecular_Function_2021', 'KEGG_2021_Human', 'WikiPathways_2021_Human',



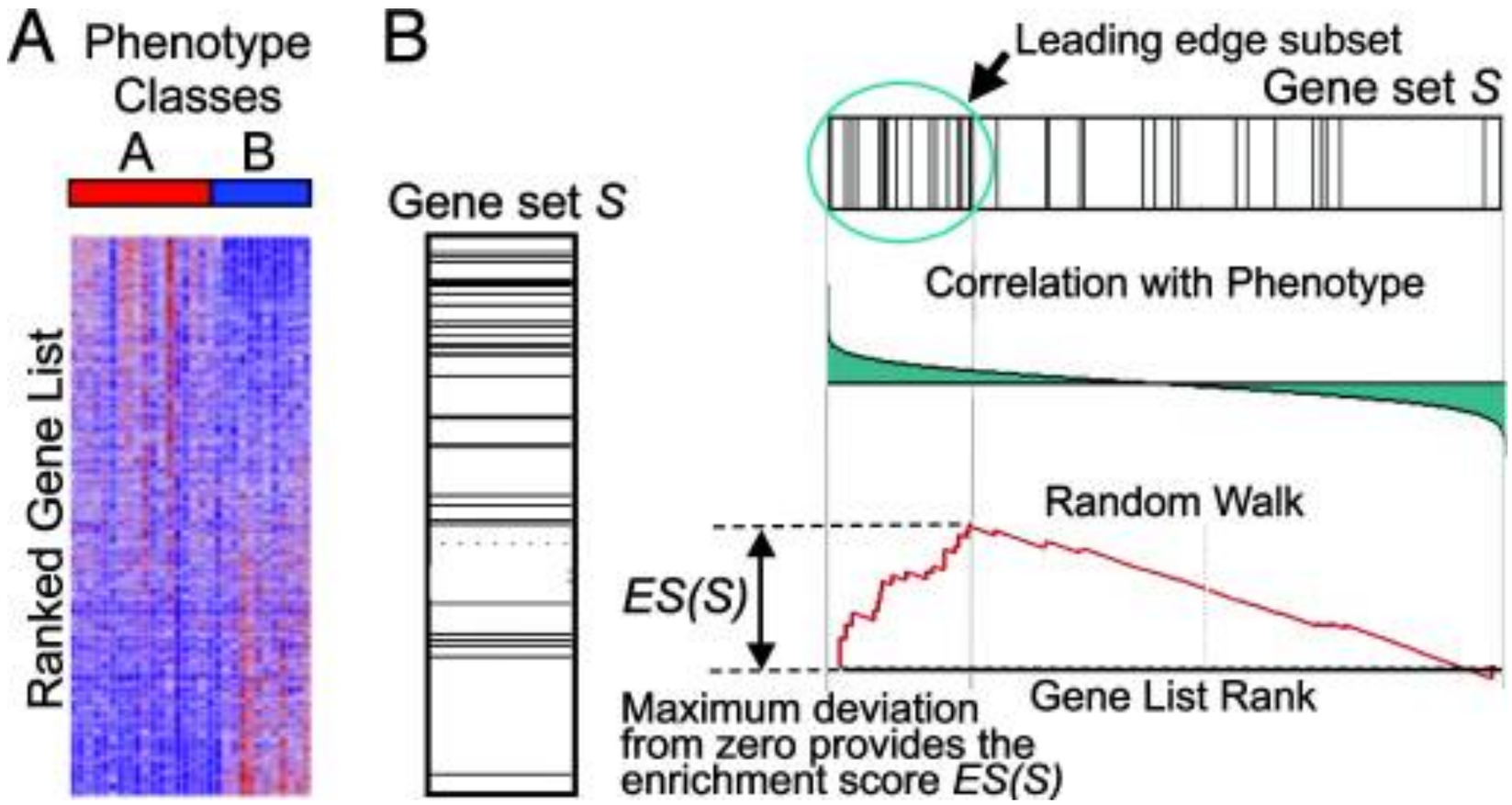
You just need to provide the list of the genes for the enrichr function

https://gseapy.readthedocs.io/en/latest/singlecell_example.html

Gene Set Enrichment Analysis

Gene Set Enrichment Analysis (GSEA)

Can be performed if you have statistics for all genes detected in the scRNAseq dataset, when using DESeq2, etc.



Gene Set Enrichment Analysis

Gene Set Enrichment Analysis (GSEA)

1. Normalize your adata (.X should contain the log normalized values)
2. Order your adata observations based on the category your are interested in having the comaprison for (e.g. stim vs. ctrl)
3. Select a relevant collection of gene sets

Q: Which of the following statements correctly describes a key difference between Gene Set Enrichment Analysis (GSEA) and Enrichr?

- a) GSEA performs enrichment analysis by ranking genes and evaluating gene sets over the entire ranked list, whereas Enrichr uses predefined gene sets and evaluates their enrichment based on statistical tests like the Fisher exact test.
- b) Enrichr is used exclusively for RNA sequencing data, while GSEA can be used for any type of gene expression data.
- c) GSEA requires permutation testing to estimate the significance of enrichment scores, whereas Enrichr calculates p-values directly from the observed data without permutations.
- d) Enrichr allows for real-time gene expression analysis, while GSEA requires batch processing of data.

Q: How is the Enrichment Score (ES) in GSEA calculated?

- a) By taking the average expression level of all genes in the gene set.
- b) By calculating the maximum deviation from zero of a running-sum statistic as you move down a ranked list of genes.
- c) By summing the absolute expression levels of all genes in the gene set.
- d) By counting the number of genes in the gene set that are differentially expressed.