Why bother integrating?

picture after analysis pipeline: (raw data -> no integration)

1) loading in raw data of 10X Genomics scRNA-seq data from 2-month old human organoids of six iPSC lines and 1 ESC (H9) line ~49k cells from neuronal lineage ⇒ paper: reproducibility of gene expression patterns across PSC-lines

2) QC: threshold for mitochondrial RNA < 5%
threshold for genes detected per cell 500 - 5000
↳ filter out doublets/multiplets or not enough deep sequencing

3) Normalization to make gene expression levels between diff. single cells comparable

4) Feature selection: identification of highly variable genes -> most varied expression levels across cells (top 3000 genes with highest variance)

5) Scaling: to account for diff. base expression level of diff. genes

6) PCA: chose first 50 PC for first look

7) 2D embedding -> UMAP ("uniform manifold approximation and projection)
↳ captured global structure better

⇒ PIC: colored by cell line (six iPSC + one ESC -> H9)
⇒ see already some areas that look like they are overlapping

Top 5 cluster marker genes per cluster (31)

=> already see correlation

Example: -> Same 5 genes for cluster 0, 2, 4, 5
-> Same 5 genes for cluster 9, 10, 14

-> cell lines share quite many cell types (based on marker genes)
↳ should not be in seperate clusters -> ideally same cell types from diff. cell lines should be mixed

ideally: Same cell types of diff. cell lines should be mixed, while diff. cell types seperated

=> batch effect ->need to integrate

MNN, CCA, Harmony

MNN: mutual nearest neighbour integration (= batch correction)
↳ estimates cell-specific correction vector based on MNN between cells of diff. batches in high-dimensionality expression space
=> introduce correction to dimension reduction (PCA) of cells

-> we used default parameters got 18 clusters
↳ much nicer overlapping between cell lines
still some similarities between top marker genes between clusters
-> annotation for cell types later

CCA: Canonical correlation analysis

1) → rotating datasets seperately that covariance is maximised
= maximize similarities

2) anchoring mechanism → anchors = cell pairs from diff. datasets
→ are eachothers MNN in CCA space (NN of one cell in same dataset tend to be NN to anchor cell in other dataset)

3) two anchor cells are seen as corresponding → expression of 1 dataset substracted by other expression via transformation matrix from anchor pairs

⇒ took very long time to find anchors + expression level correction
→ subjectively looks a little worse than MNN (and 1 more cluster → 19)
→ paper used CCA (clusters not well seperated) ⇒ harmony better

Harmony: uses fuzzy clustering → calculates correction factor for each dataset to move centroid of cluster of each dataset closer to global centroid of cluster
⇒ cellspecific correction factor calculated ⇒ iteratively repeated until convergence

⇒ key advantage: large datasets (up to 10^6 cells) with multiple batches + computationally most efficient

→ least amount of time + best overlap + least clusters

⇒ decided to continue with harmony

(+ most highlighted, fastest ⇒ best compromise in 2020 benchmark paper Tran et al.

⇒ best cluster seperation