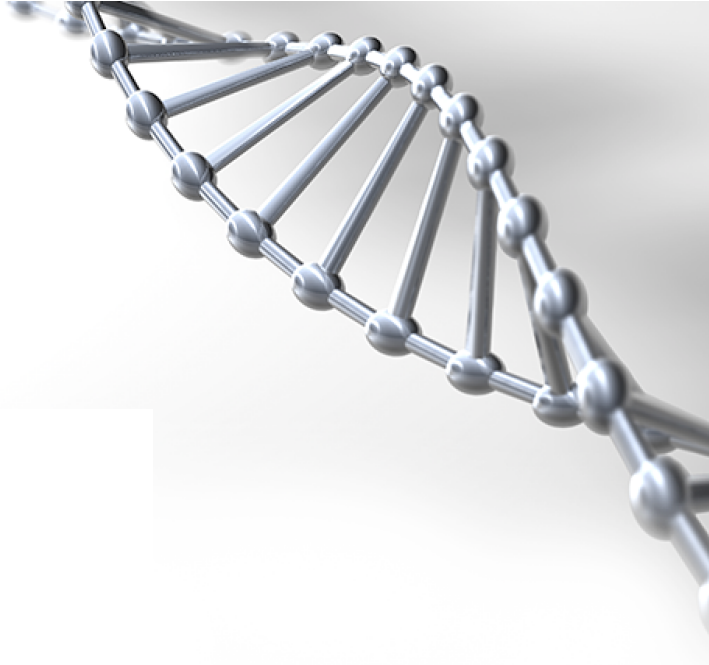


Visium data analysis using Chipster

Autumn 2022

Eija Korpelainen, Maria Lehtivaara Iida Hakulinen

CSC – Suomalainen tutkimuksen, koulutuksen, kulttuurin ja julkishallinnon ICT-osaamiskeskus



Introduction

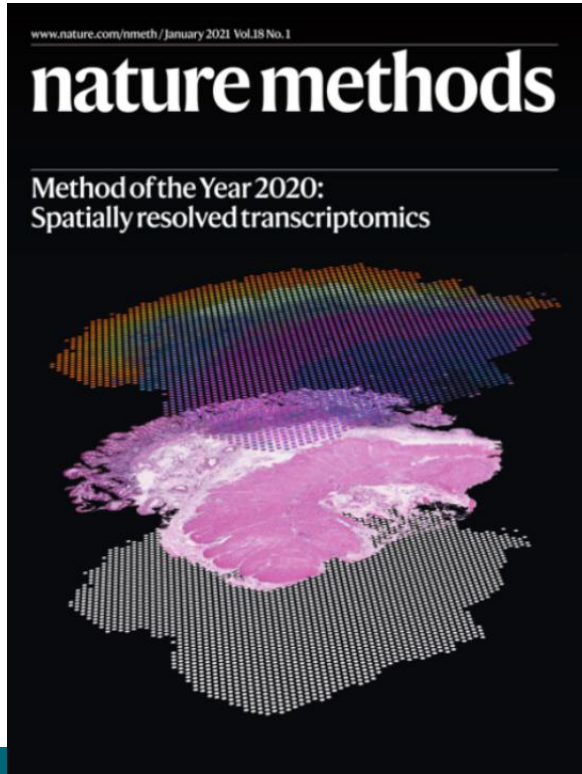
Spatially resolved transcriptomics

Overview

- Intro to spatially resolved transcriptomics
- How does the Visium system work?
- Things to keep in mind when working with Visium data
- What will you learn during this course

Spatially resolved transcriptomics

- Spatial context: Gene expression data overlaid with a tissue image
  - retains organization of tissue and cellular microenvironment
  - cell type identification in the context of heterogeneous tissue
- Several technologies available
  - this course focuses on 10X Genomics Visium data



Visium – how does it work?

- Place tissue slice (frozen or FFPE) on a capture area on a slide
- Capture area contains about 5000 barcoded spots
  - Spot diameter 55 um, center to center distance 100 um
  - NOTE: about 1-10 cells per spot
- Each spot has millions of capture probes
  - 16 nt spatial barcode to track back the location
  - 12 nt UMI for counting
  - 30 nt poly(dT) to capture polyA
- Stain, image, permeabilize cells
- cDNA synthesis
- Library construction

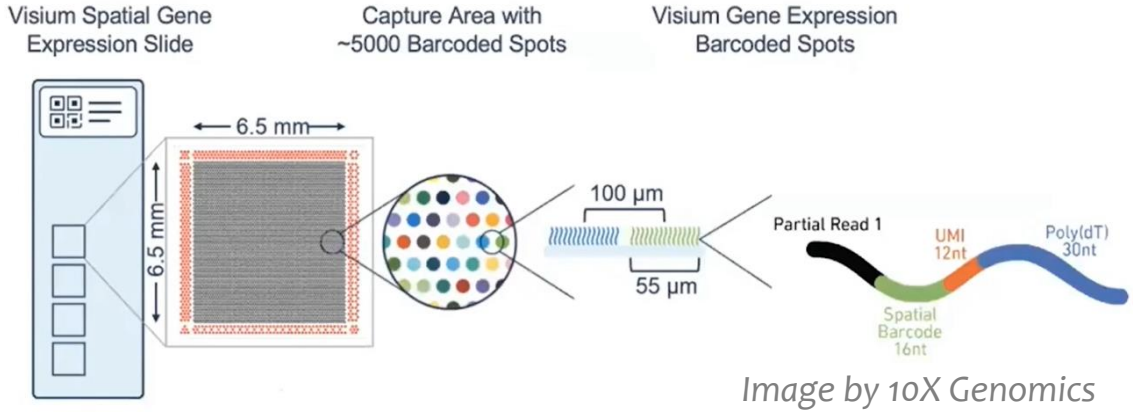


Image by 10X Genomics

Visium data – things to remember

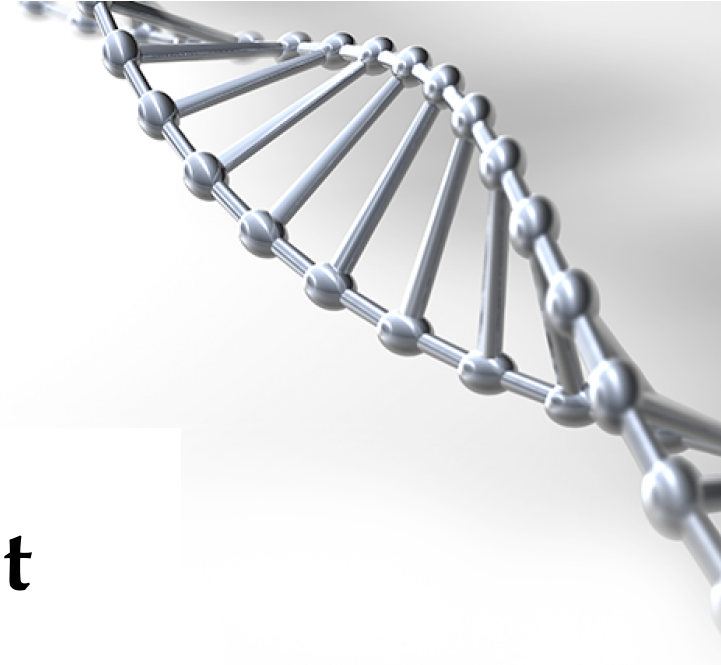
- Each spot typically includes several cells, not just one
- There can be different types of cells in a spot
- The gene expression values measured are an average from the cells in a spot
- Clusters represent a group of spots with **similar composition of cell types**

During this course you will learn how to

- Create a Seurat object and check the quality of spots
  - Filter out low quality spots (damaged tissue)
- Normalise gene expression values and identify highly variable genes
- Reduce dimensions with PCA using the highly variable genes
- Use the PCs to cluster spots with graph based clustering
  - Visualise clusters (UMAP and overlay with the tissue image)
- Detect spatially variable genes
  - Visualise gene expression on the tissue image
- Subset anatomical regions
- Predict cell type composition in spots: Integrate with scRNA-seq data
- Integrate several samples

During this course you will learn how to

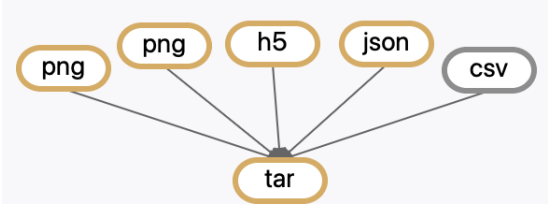
- Create a Seurat object and check the quality of spots
  - Filter out low quality spots (damaged tissue)
- Normalise gene expression values and identify highly variable genes
- Reduce dimensions with PCA using the highly variable genes
- Use the PCs to cluster spots with graph based clustering
  - Visualise clusters (UMAP and overlay with the tissue image)
- Detect spatially variable genes
  - Visualise gene expression on the tissue image
- Subset anatomical regions
- Predict cell type composition in spots: Integrate with scRNA-seq data
- Integrate several samples



Setting up Seurat object

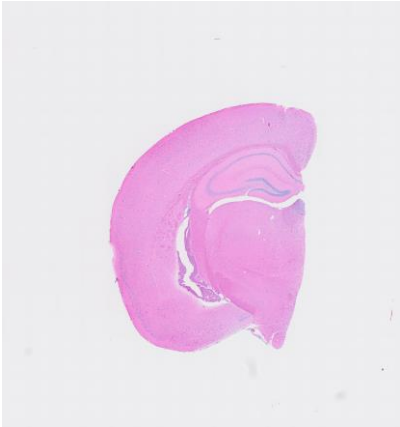
Input for the tool "Seurat – setup and QC"

- 10X Genomics output files (from Visium Space Ranger software):
  - **Filtered**_feature_bc_matrix.h5 (= spot by gene expression matrix)
  - Tissue_hires_image.png (= image of the tissue slice)
  - Tissue_lowres_image.png
  - Scalefactors_json.json (= relate the high res image to low res)
  - Tissue_positions_list.csv (make sure that doesn't contain the column names) (= spot positions over the image)

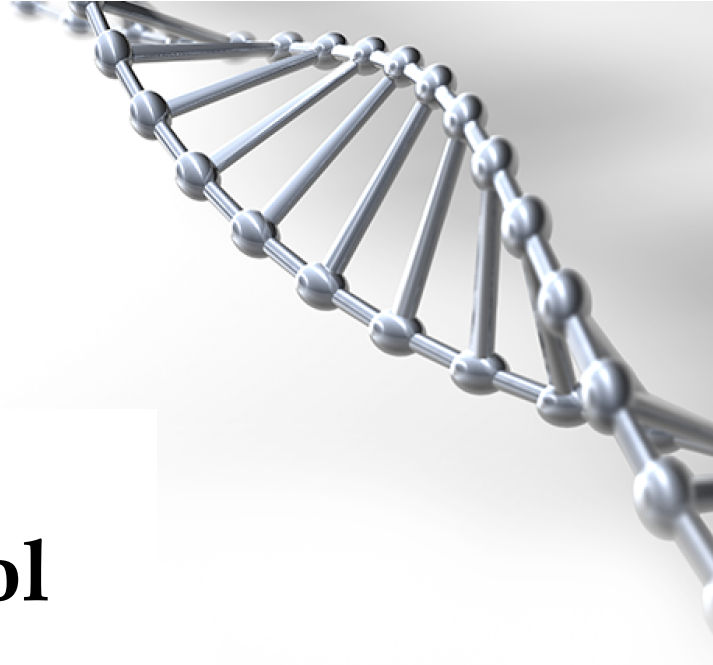


- Make a tar package of these files
  - You can use the Chipster Utilities tool "Make a tar package" for this

→ Seurat object containing spot-level expression data & the associated image of the tissue slice



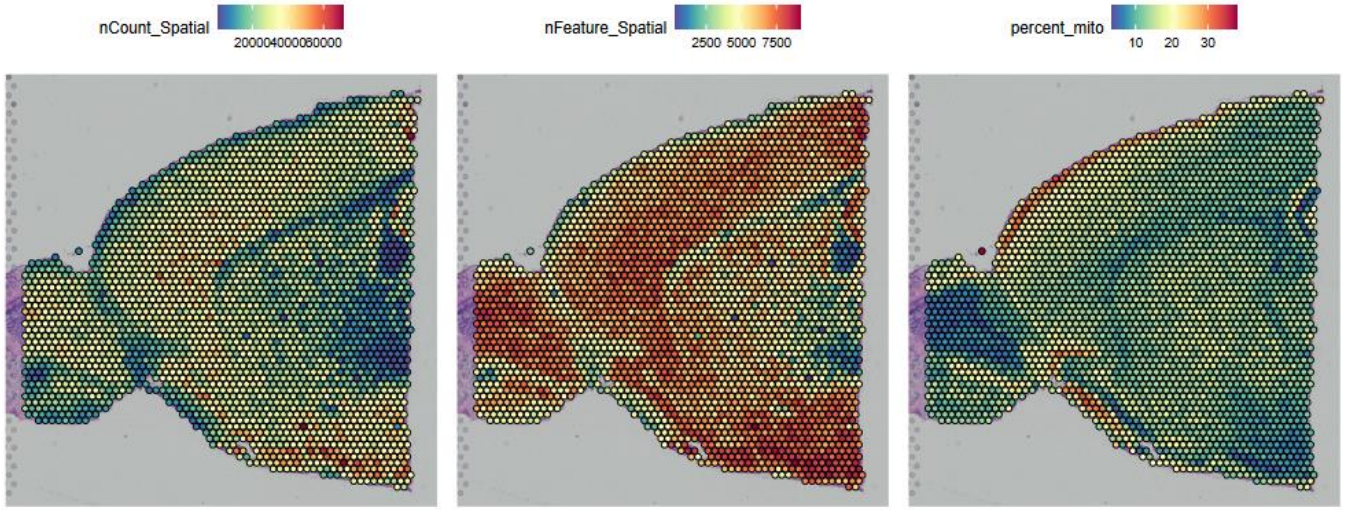
- If you have multiple samples, do this for every sample



Quality control

QC violin plots produced by the Seurat setup tool

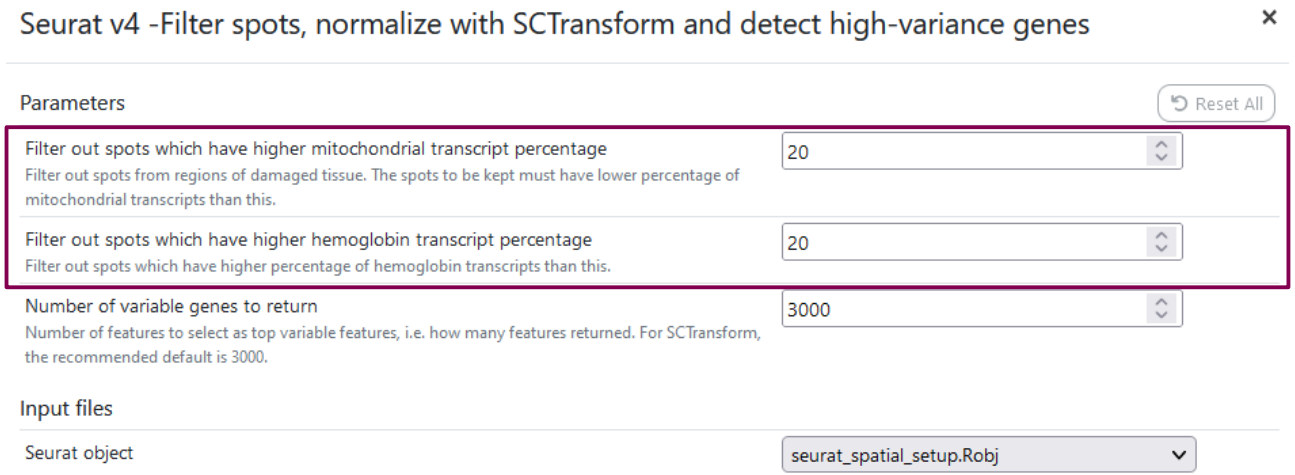
QC data plotted on tissue image

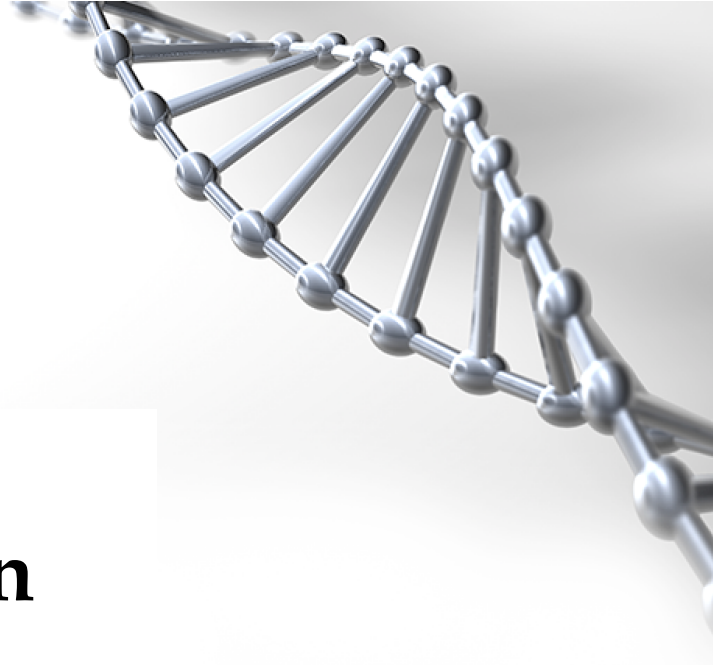


- nCount_Spatial: big differences in UMI counts per spot
  - technical and biological reasons (tissue anatomy) → normalization required
- High percentage of mitochondrial reads near the edges, tears and folds of tissue
  - damaged tissue

Filtering bad quality spots

- You can filter spots prior to normalization
  - Mitochondrial transcript percentage
  - Hemoglobin transcript percentage

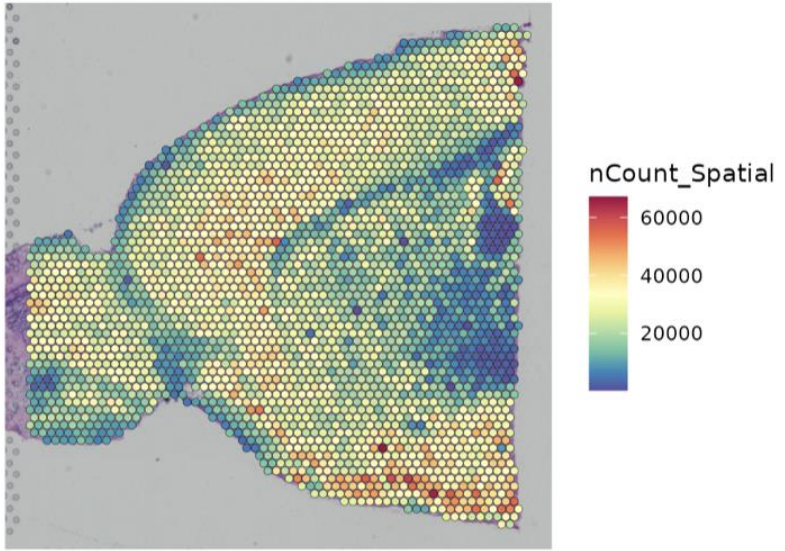




Normalisation

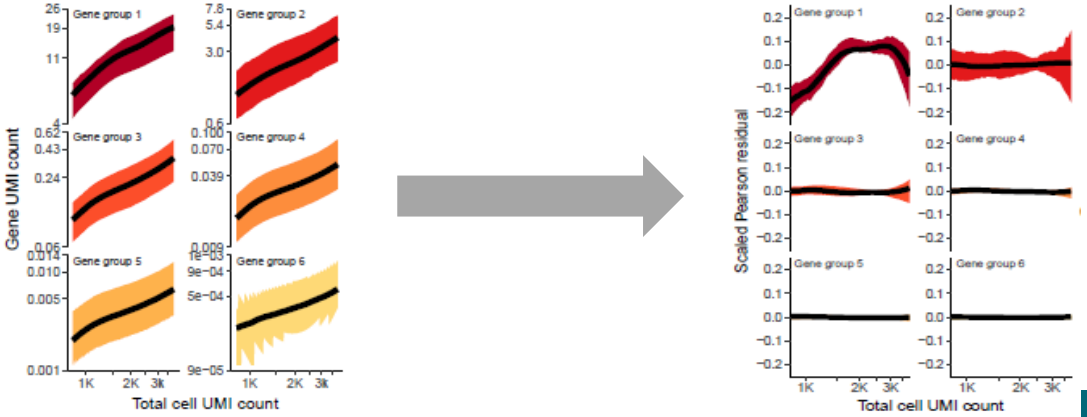
Normalising expression values across spots

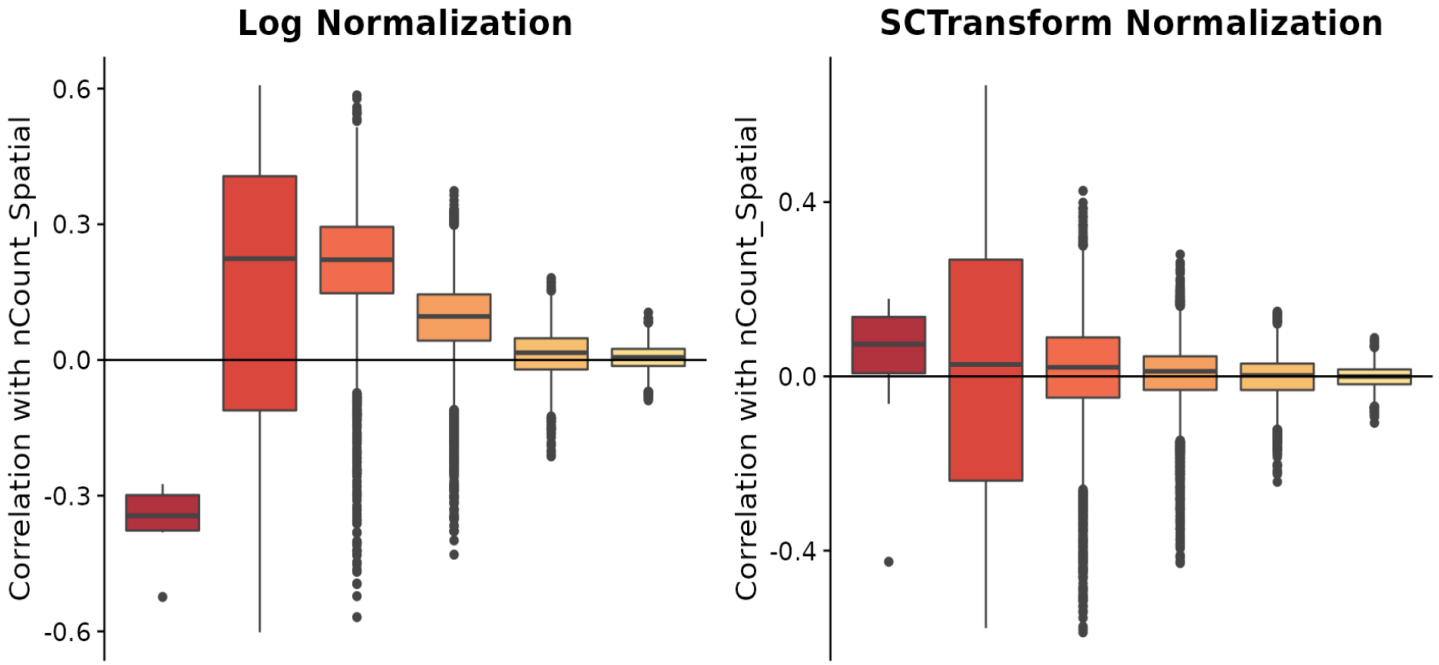
- Sequencing depth (number of UMIs per cell) varies significantly between spots
- Normalized expression values of a gene should be independent of sequencing depth
- Variance can be substantial for spatial datasets
- Cell density varies across the tissue → global scaling normalization doesn't work
- Use SCTransform
  - Doesn't force the same "size"



SCTransform

- Models gene expression as a function of sequencing depth using GLM
  - Constrains the model parameters through regularization, by pooling information across genes which are expressed at similar levels
  - Normalized expression values = Pearson residuals from regularized negative binomial regression
- Works well also for high expressing genes
- In addition to normalization, identifies highly variable genes and scales data
- SCTransform v2 even better





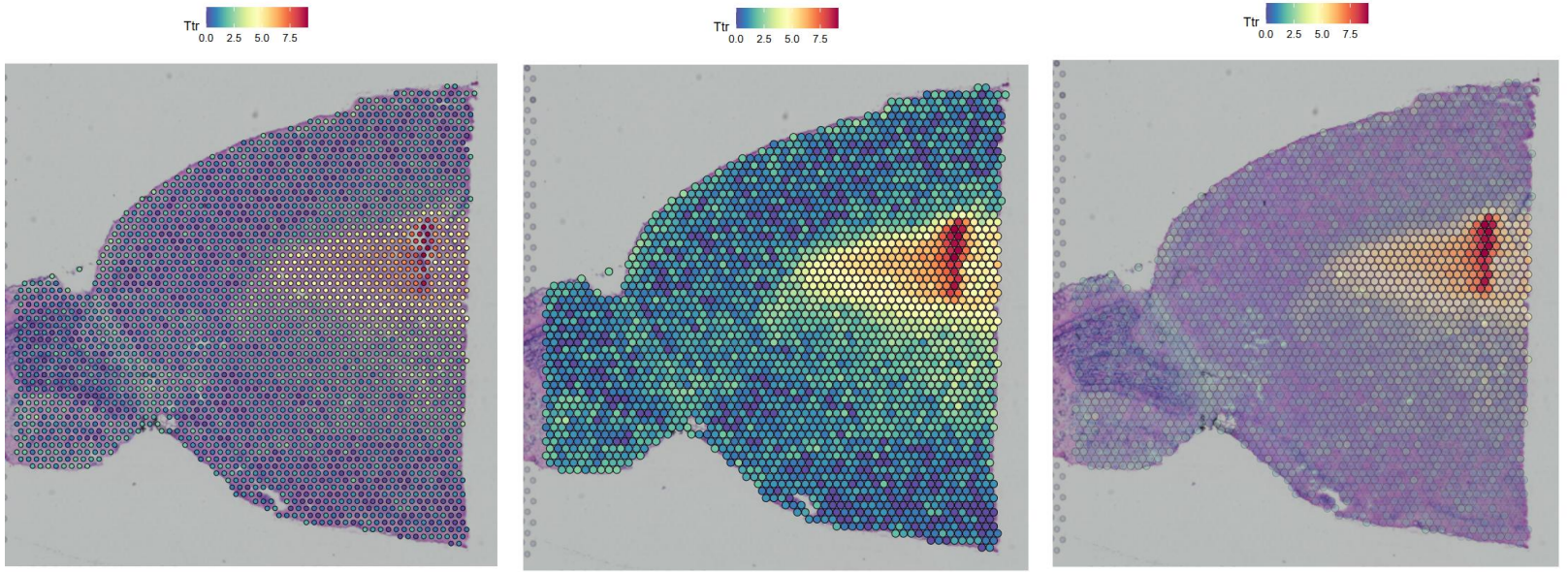
Picture 1: https://satijalab.org/seurat/articles/spatial_vignette.html#data-preprocessing-2

Gene expression visualization

Spatially resolved transcriptomics

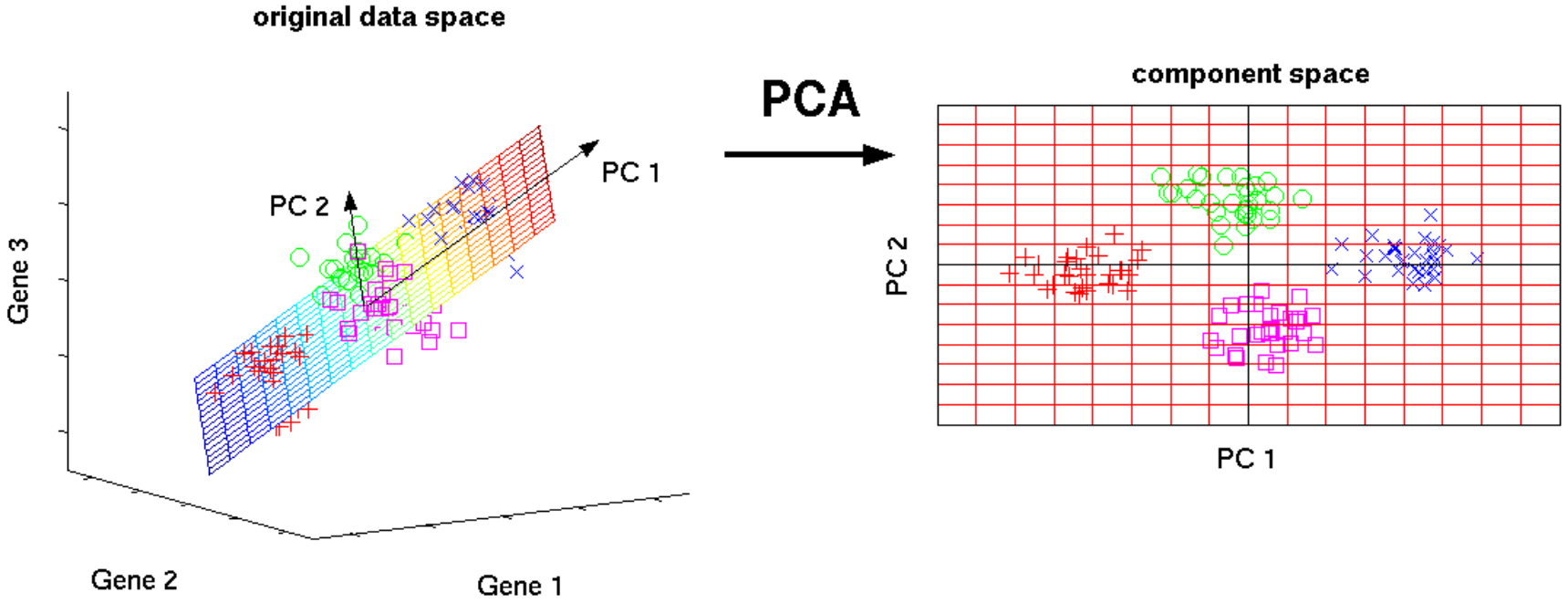
Overlay gene expression values on top of histology image

- If you want to see the tissue better you can modify
  - point size
  - transparency of the points (spots with lower expression for gene X are more transparent)



Dimension reduction

Dimensionality reduction

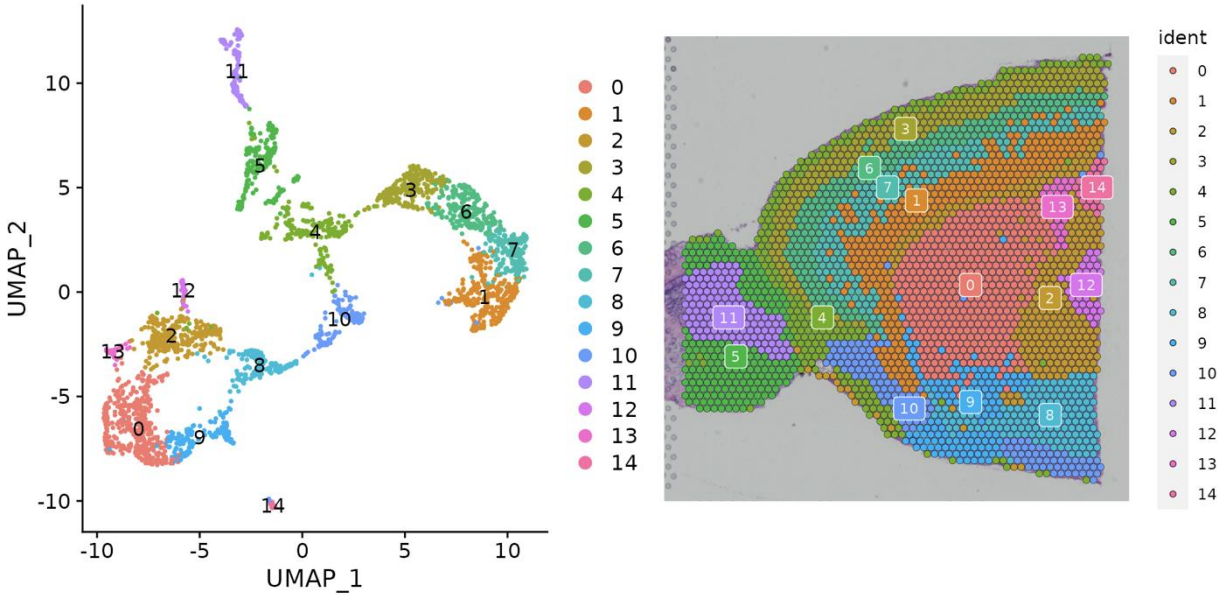


Picture 1. Im, Jonas. "Introduction to PCA" Medium.com, 6th Dec. 2018, https://medium.com/@jamesim2077/introduction-to-pca-principal-component-analysis-c26dffe2a857. Accessed 10 Aug. 2022

Clustering

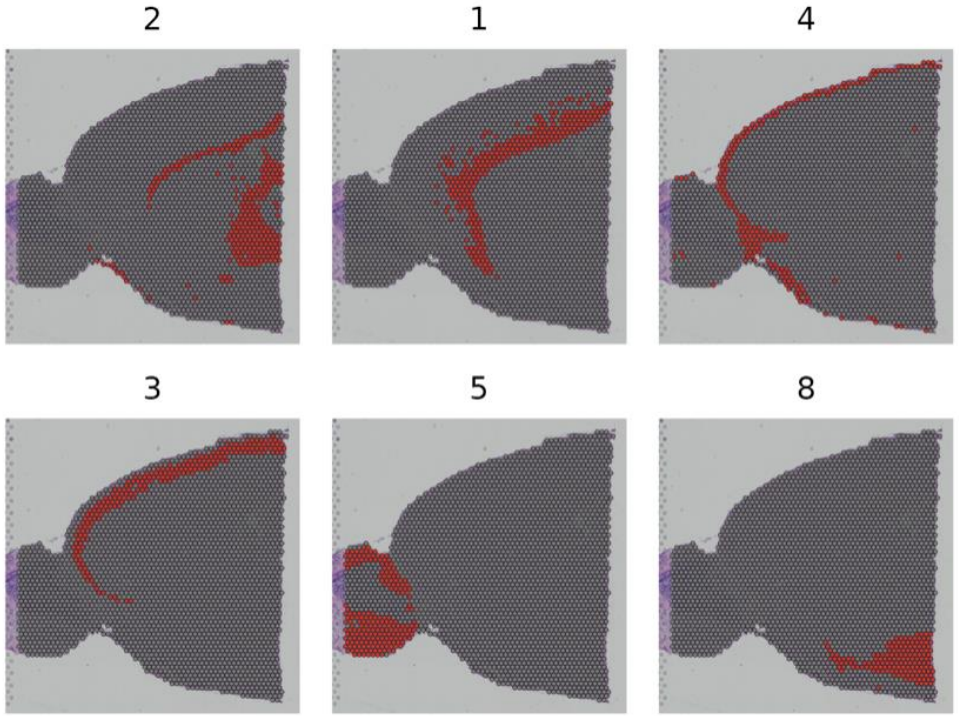
Cluster visualisation

- UMAP plot
- Clusters on top of the slice image



Tricky to see the clusters?

- Using **Visualise clusters** tool and drawing one cluster at a time helps

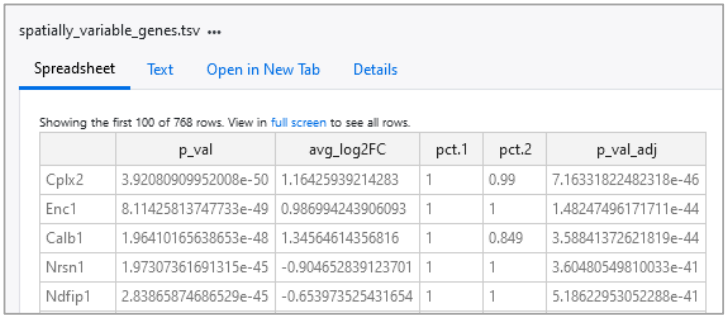


Identifying spatially variable genes

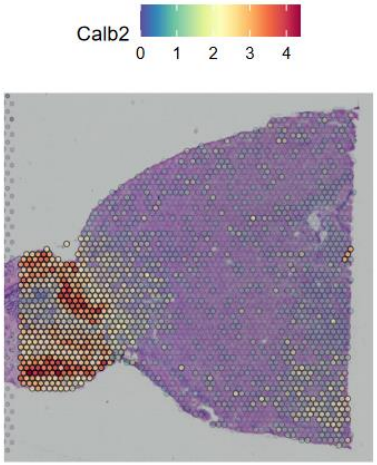
Two approaches for detecting genes whose expression level depends on spatial location

- Compare selected clusters
  - Uses cluster information, similar to what we did with scRNAseq data
  - Works when clusters show clear spatial restriction
- Look for spatially variable genes in the absence of pre-annotation (e.g. cluster info)
  - Markvariogram
    - Models spatial data as a mark point process and computes a variogram.
    - Takes a lot of time (will be parallelised)
  - Moransi
    - Will be integrated in Chipster later
  - SpatialDE and Splotch are not in Seurat yet

Identify spatially variable genes based on clusters



	p_val	avg_log2FC	pct.1	pct.2	p_val_adj
Cplx2	3.92080909952008e-50	1.16425939214283	1	0.99	7.16331822482318e-46
Enc1	8.11425813747733e-49	0.986994243906093	1	1	1.48247496171711e-44
Calb1	1.96410165638653e-48	1.34564614356816	1	0.849	3.58841372621819e-44
Nrsn1	1.97307361691315e-45	-0.904652839123701	1	1	3.60480549810033e-41
Ndfip1	2.83865874686529e-45	-0.653973525431654	1	1	5.18622953052288e-41

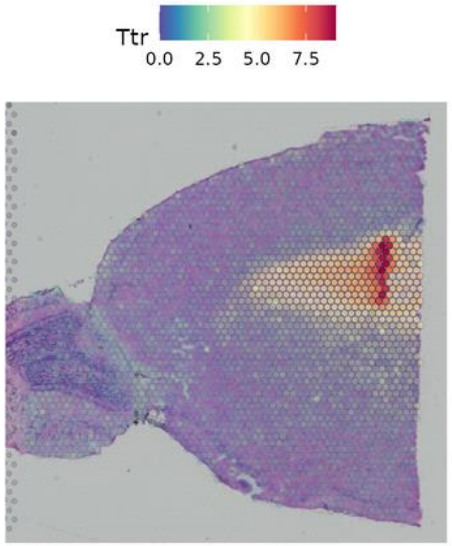


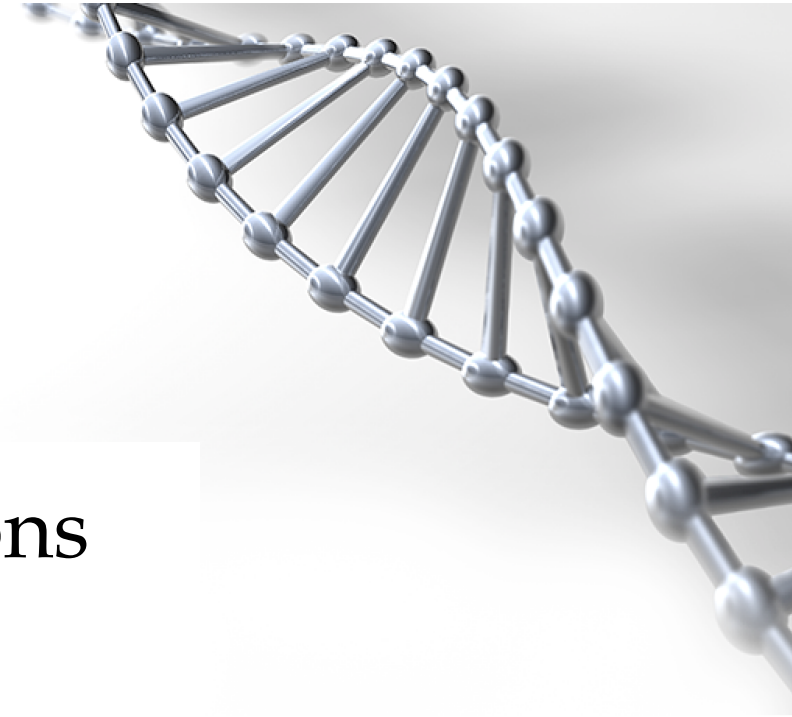
Parameters will be added to limit testing to genes which

- are expressed in at least X fraction of cells
  - Seurat's default 10%
- show at least Y fold difference
  - Seurat's default 0.25

Identify spatially variable genes using markvariogram

- Models spatial data as a mark point process and computes a variogram
  - This process calculates gamma(r) values measuring the dependence between two spots a certain “r” distance apart. By default, Seurat uses an r-value of 5, and only computes these values for variable genes to save time.

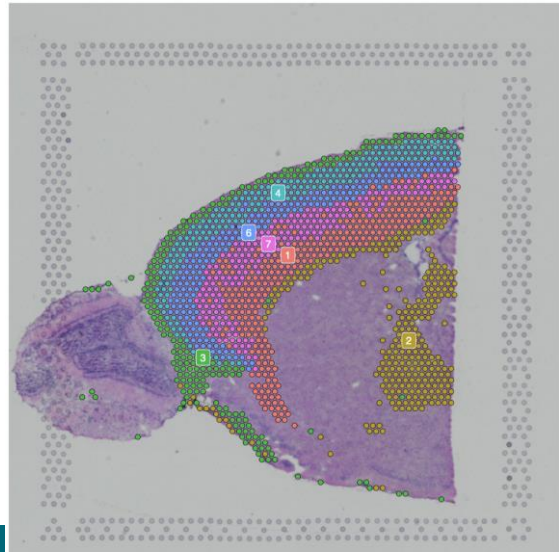




Subset anatomical regions

Subset out clusters

- Select clusters you want to subset to study them further
  - You can use the "Visualise clusters" tool to more easily see which clusters correspond to specific regions in the slide
- If you are going to integrate with scRNA-seq data, select the clusters that correspond to that data
  - Other clusters can give rise to false positive cell type assignments

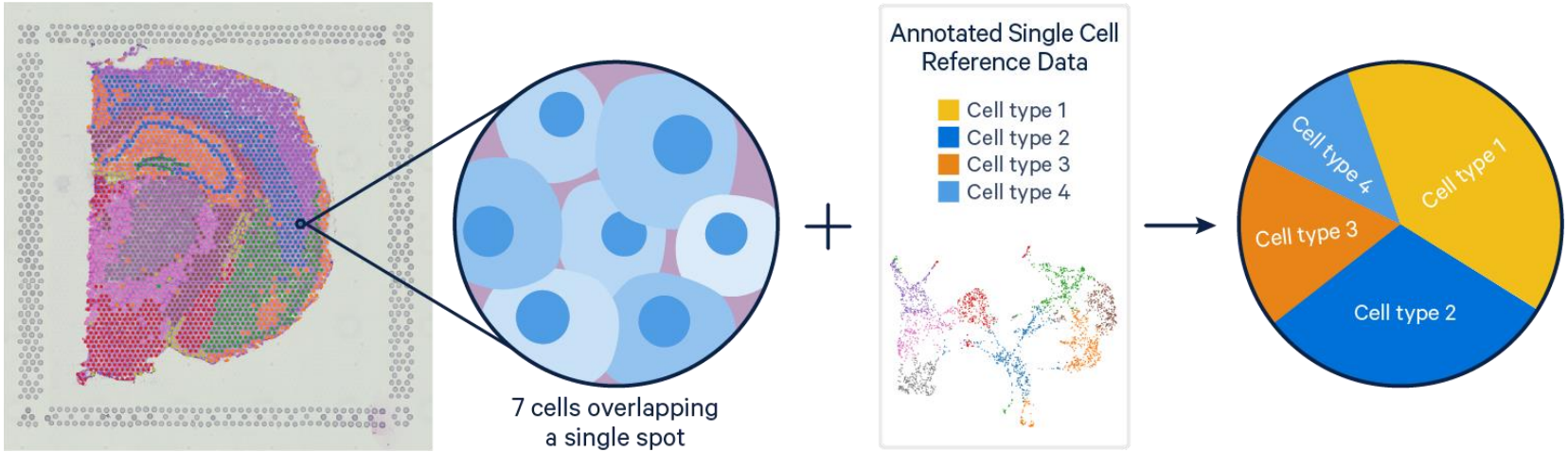


Integration with scRNA-seq data

Integration with scRNA-seq data

- What is the cell type composition in the spots?
- We use scRNA-seq data as a reference data set
- Several methods available, Seurat uses anchor-based method

Integration with sc reference



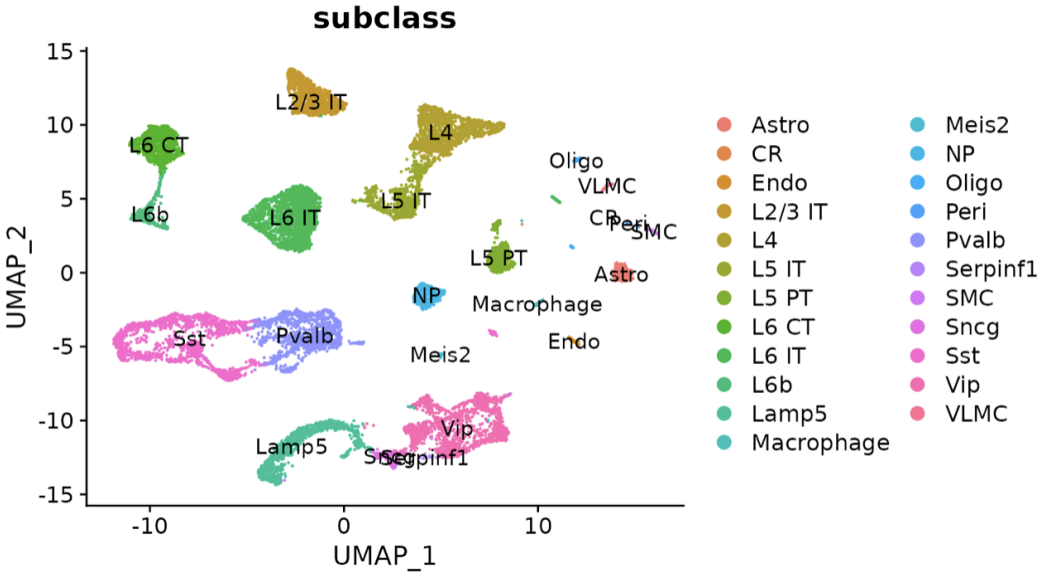
Picture 1.. ” Integrating Single-cell and Visium Spatial Gene Expression Data” 10xgenomics..com, 17th Mar. 2022, https://www.10xgenomics.com/resources/analysis-guides/integrating-single-cell-and-visium-spatial-gene-expression-data. Accessed 19 Aug. 2022

Integration with scRNA-seq data using Seurat

- Find anchors between Visium data and scRNA-seq data (MNN)
- Create correction vector based on differences in expression
- Use correction vectors to remove platform effects
- Integrate data sets
- Transfer cell type information from scRNA-seq data to spatial spots

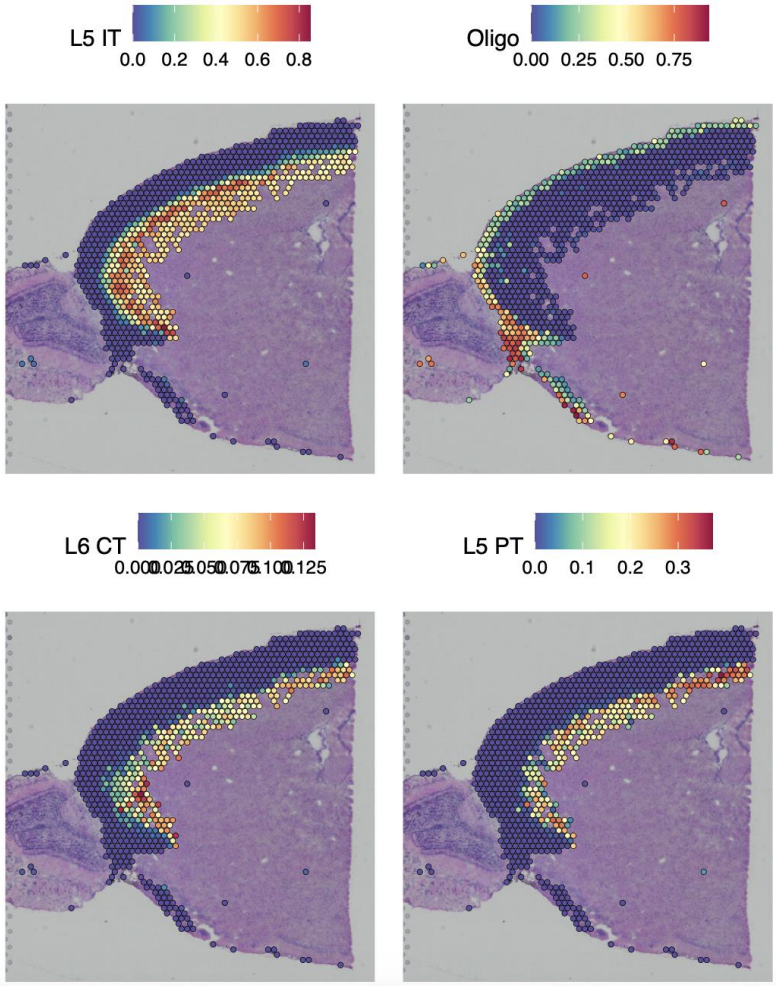
Spatial Transcriptomics / Seurat v4 -Visualise integration

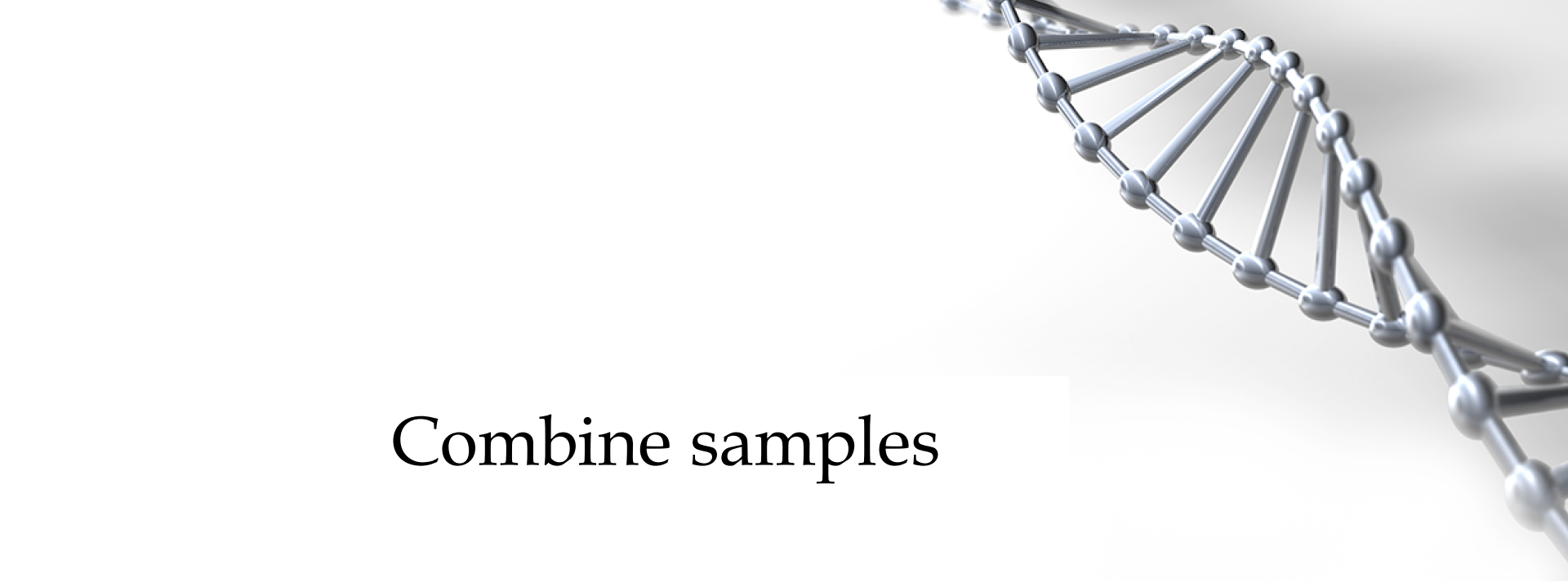
- Currently, this tool takes as input a allen_cortex.rds reference file and the subsetted Seurat object
- As output:
  - Seurat object with predictions for spots cell types
  - A pdf showing the reference data as UMAP plot



Visualise integrated data

- Allows you to visualise the cell type predictions of interest (give as parameter)
- Cell types whose location is spatially restricted
  - The same methods used as to define spatially variable features (markvariogram), but use the cell type prediction scores as the "marks" rather than gene expression
  - Currently, top4 of these types are plotted

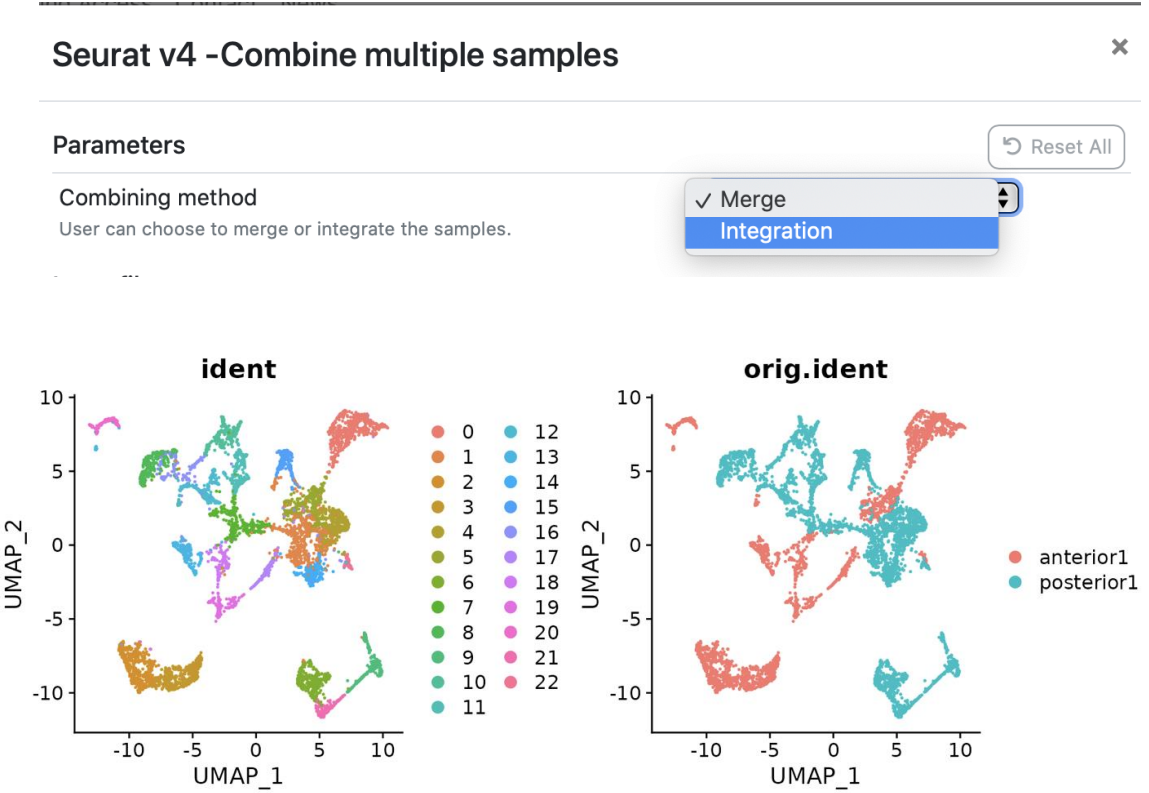




Combine samples

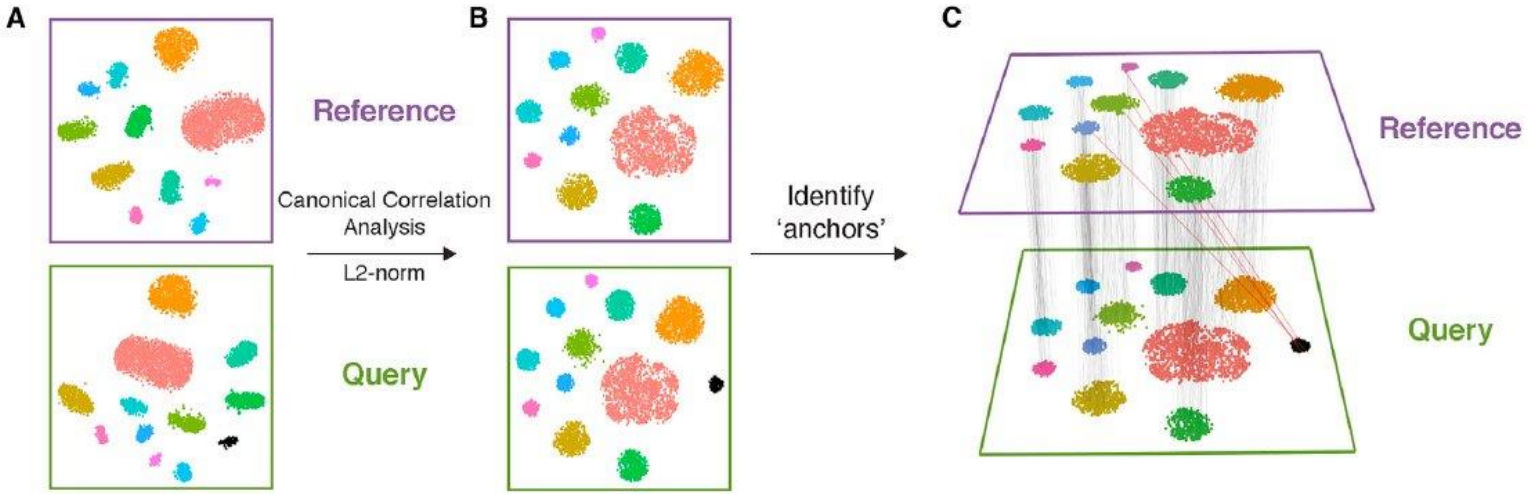
Combine samples

- Currently, two options:
  - Merge
    - Simple 1+1 merging, used in Seurat vignette
    - Ok, when there's no big batch effect
  - Integrate
    - Similar to what is used to combine samples in scRNA-seq data



Batch effects

- Strong batch effects between samples can affect the multiple sample analysis
  - Can be removed with integration



Picture 1: https://www.nature.com/articles/nbt.4096