Integration of scRNAseq datasets for studying cellular heterogeneity in solid tumors

Stefano Volinia

ICSC
Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

Università degli Studi di Ferrara

Integration

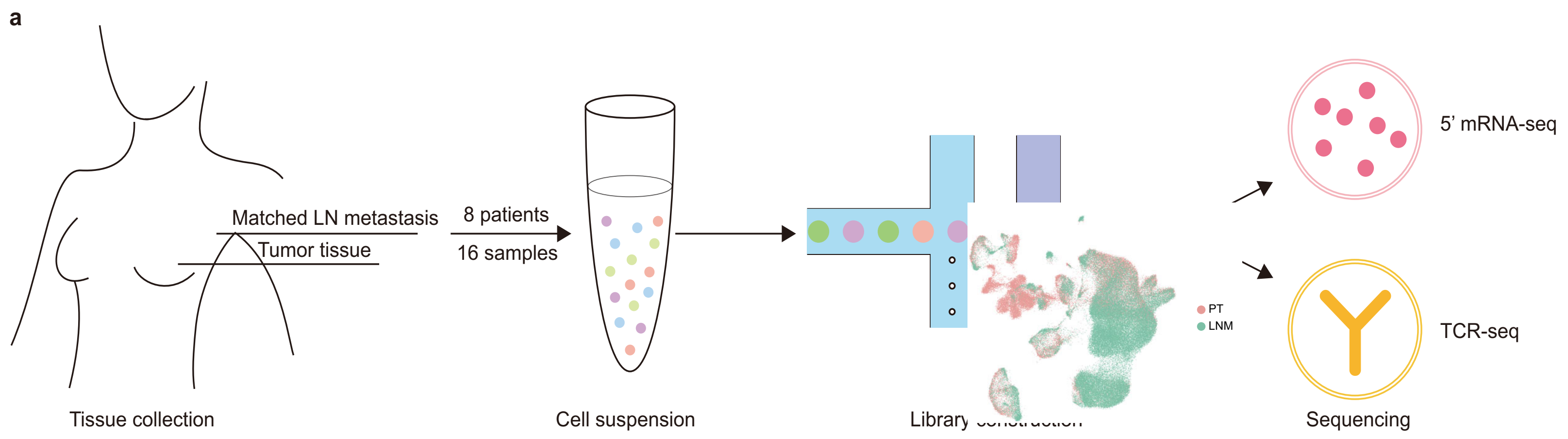
Integration of single cell profiles from normal breast, breast cancer primary tumors and metastatic lymph nodes.

Prior to final integration of cancer and normal single cell datasets, we first separately integrated and **down-sampled** the cancer and normal datasets separately. ER+, HER2+, TNBC, and tumors derived from BRCA1 patients were obtained from GSE161529. Tumor involved lymph nodes and normal breast tissues were also obtained from the same dataset. Additional primary tumors and paired metastatic lymph nodes were obtained from the GSE167036 dataset. The human breast cell atlas, assembled from 55 donors, who underwent reduction mammoplasties or risk reduction mastectomies, was used to integrate single cell normal breast data. To reduce the size, without affecting complexity, normal breast (from HBCA and from GSE161529) datasets were integrated with Scanorama and each cluster was randomly down-sampled to 1000 cells. Identical procedure was also performed on the primary tumor datasets.

The **aneuploidy** was tested using scevan and copykat, using the stromal (CAFs and TECs) as baseline.

All datasets were integrated using Scanorama **integration** and down-sampled to 1000 cells per cluster.

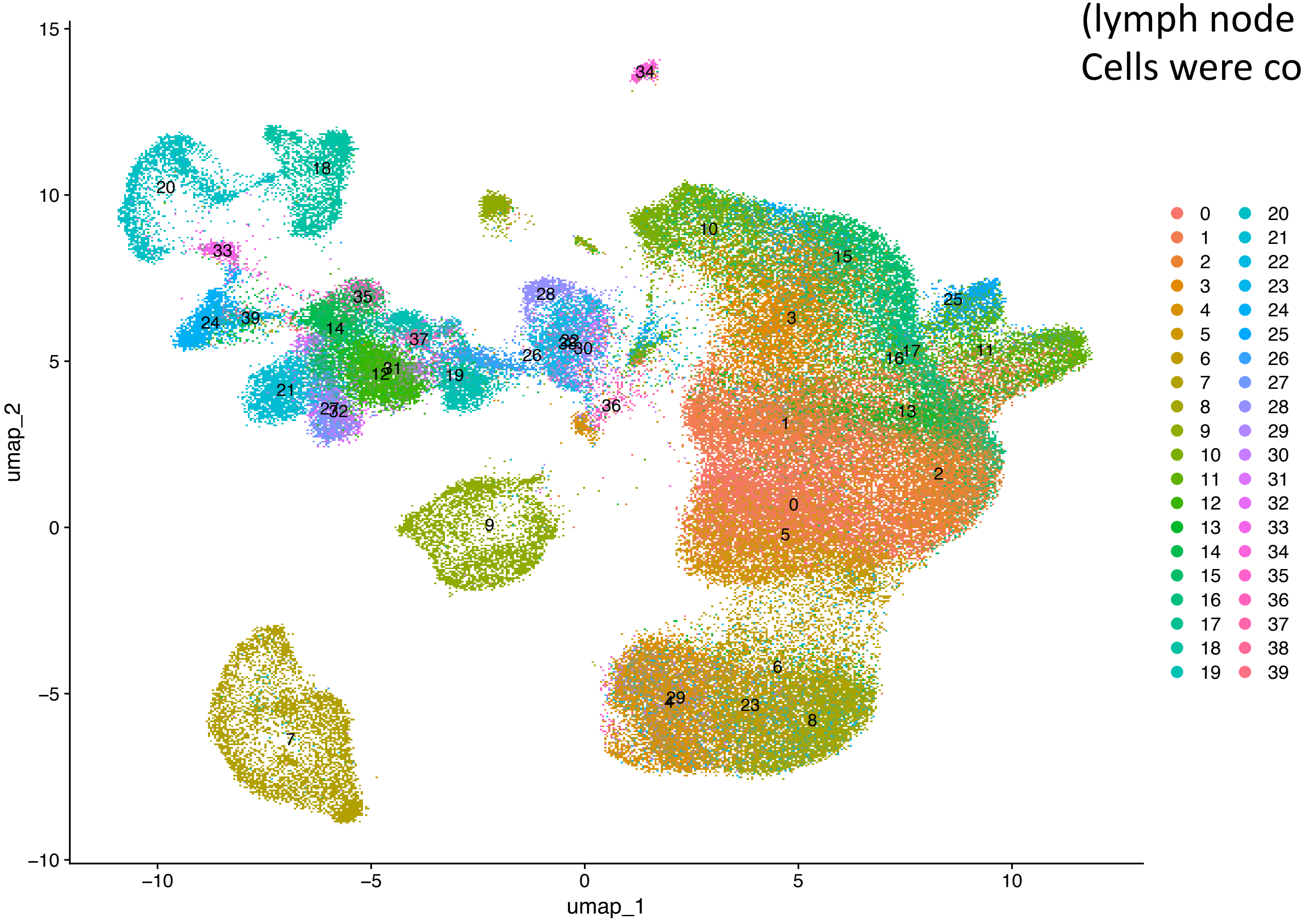
Single cell profiling of primary and paired metastatic lymph node tumors in breast cancer patients. GSE167036



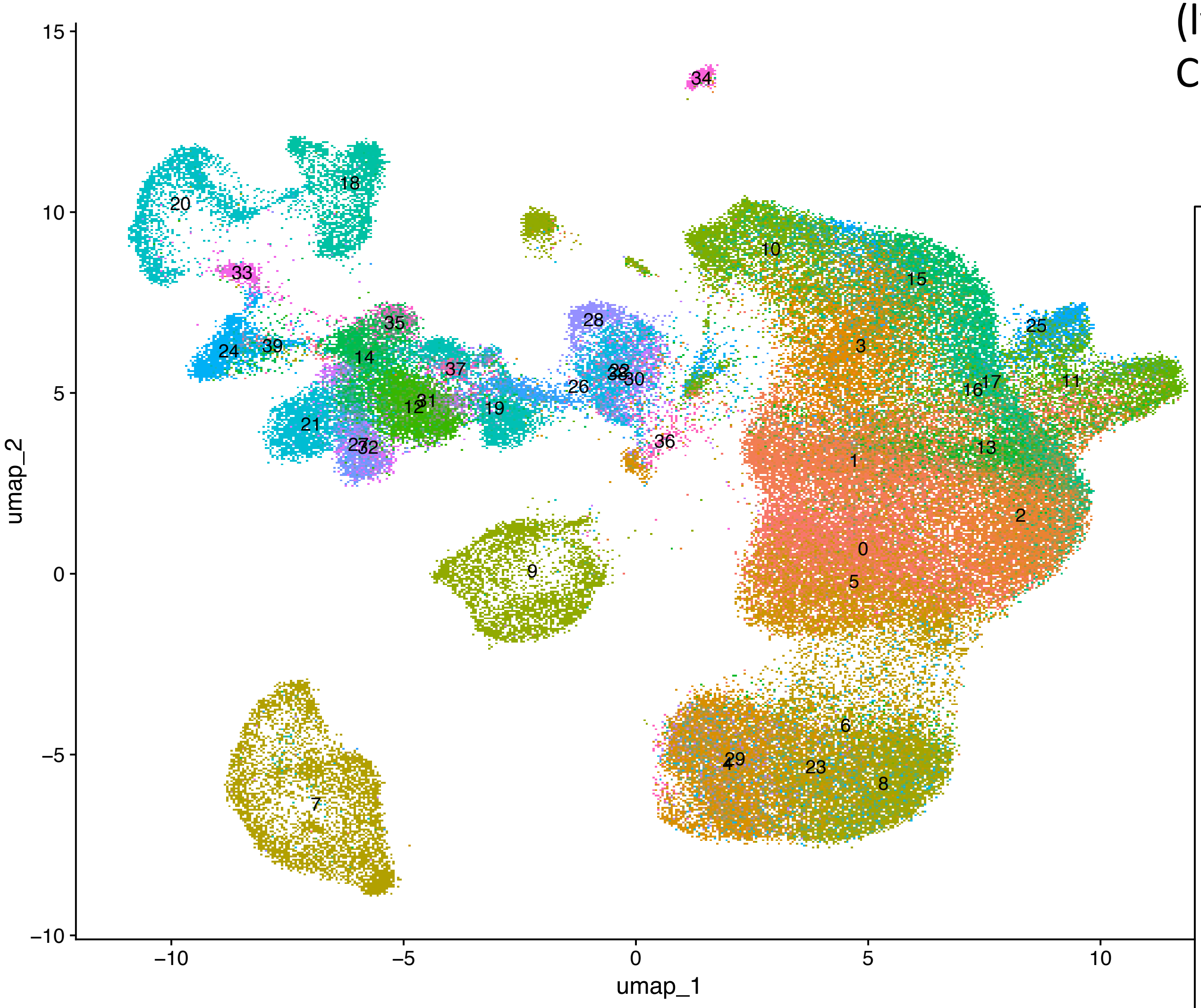
- Characterize the microenvironment of LNMT and PT, which may shed light on the individualized therapeutic strategies for breast cancer patients with lymph node metastasis.

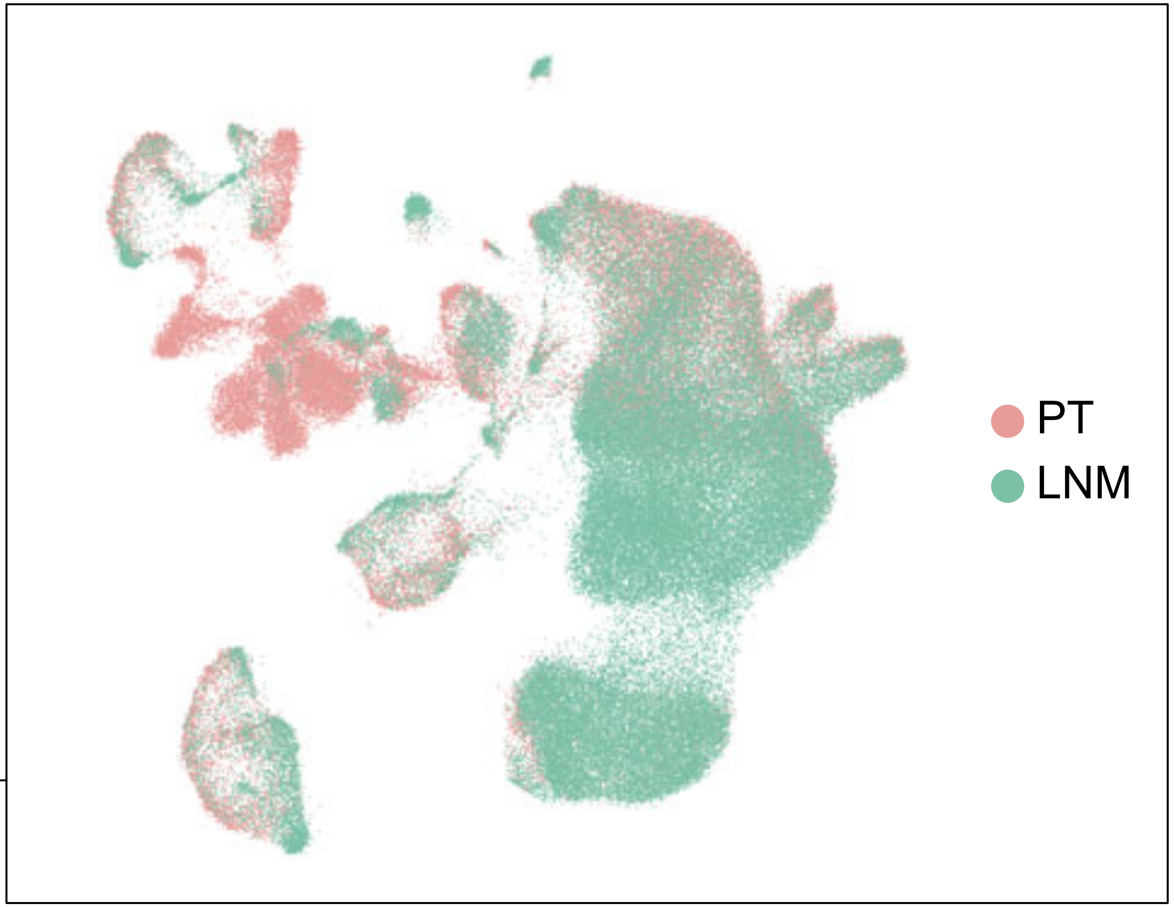
- Single-cell maps of primary tumors (PTs) and paired LNMTs in 8 breast cancer patients
- They demonstrate that the activation, cytotoxicity, and proliferation of T cells are suppressed in LNMT compared with PT

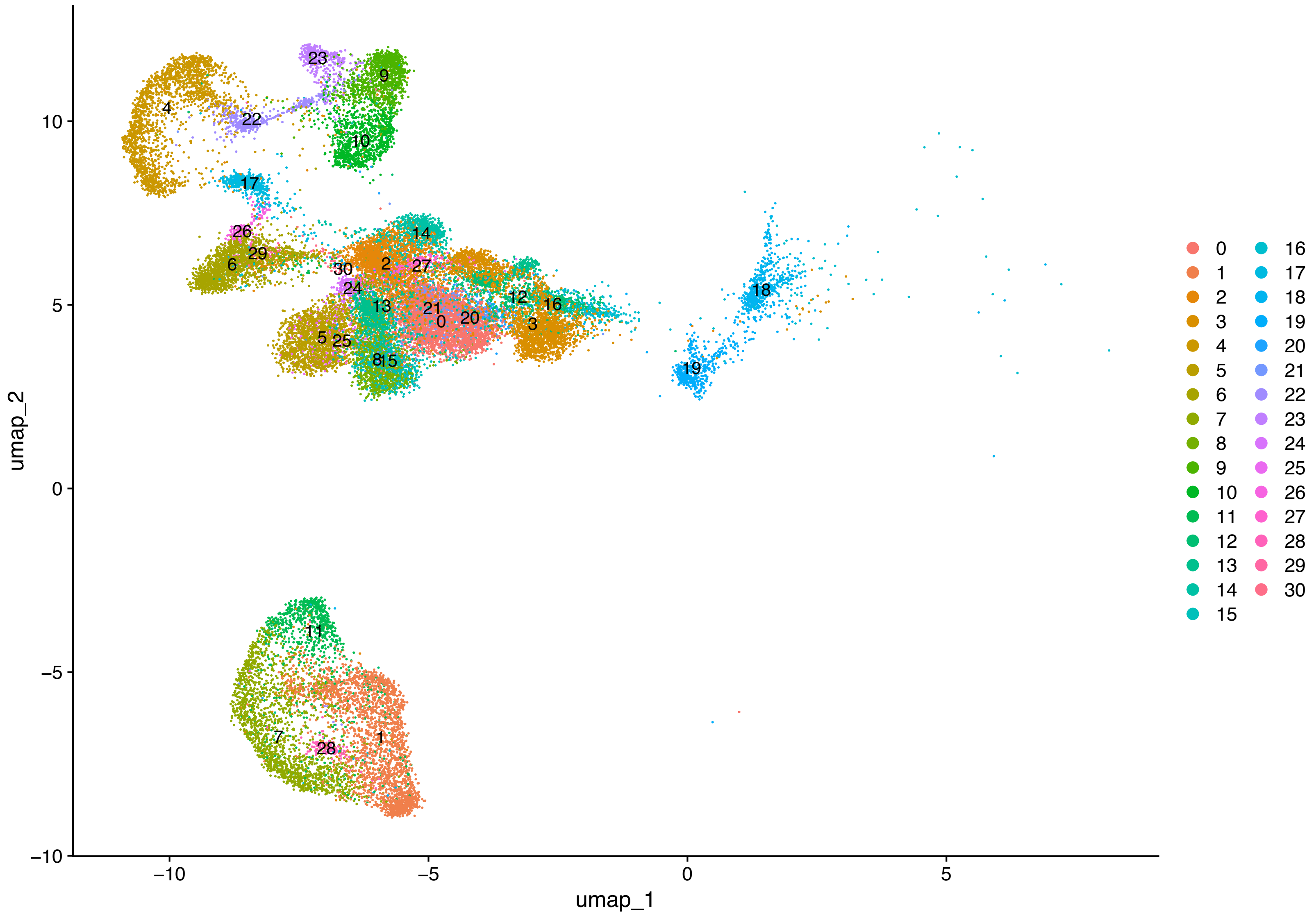
UMAP embedding plot showing identified clusters of all 118K cells from paired PT (primary tumor) and LNMT (lymph node metastasized tumors) of 8 LNMT patients. Cells were colored according to their clusters



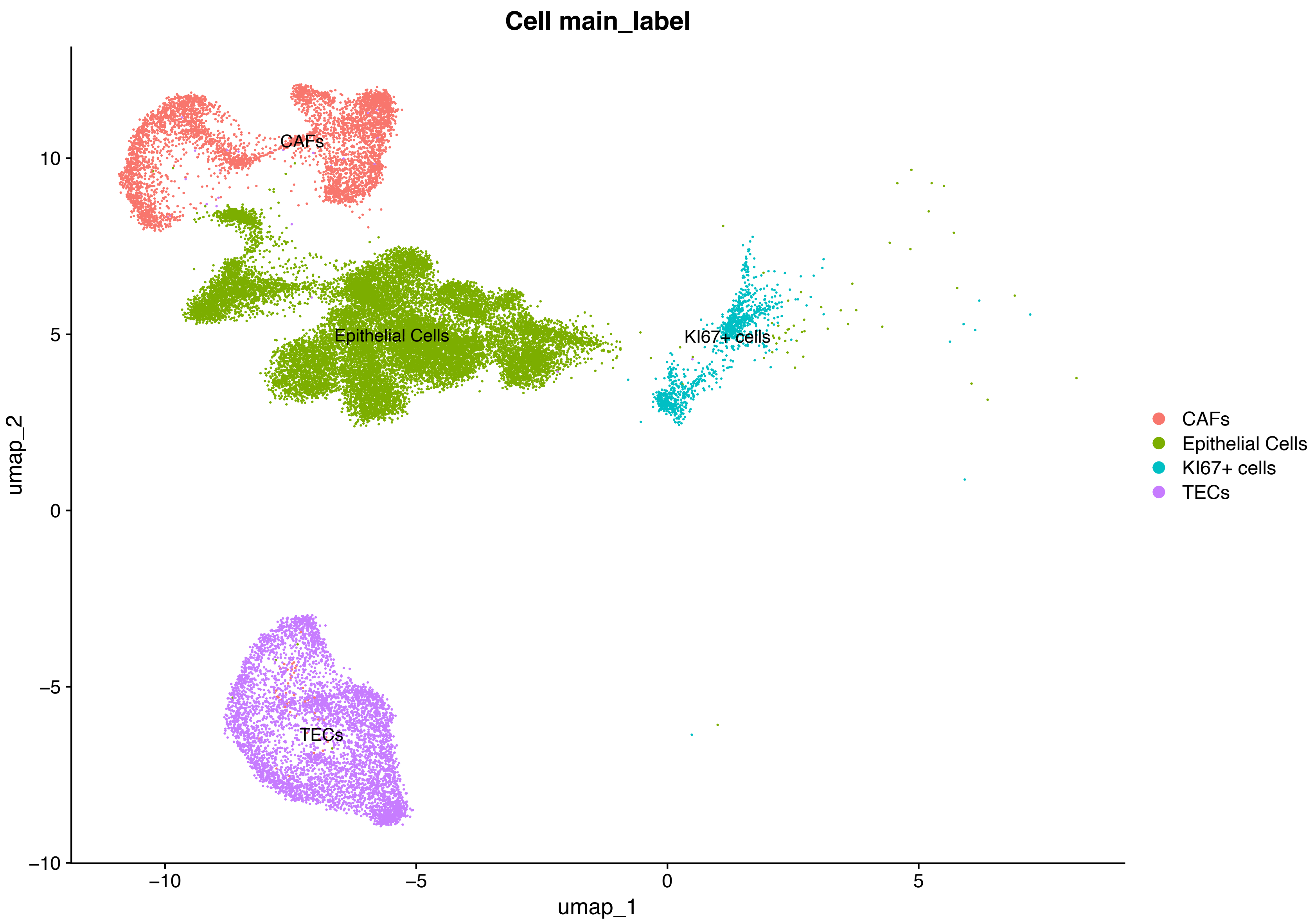
UMAP embedding plot showing identified clusters of all 118K cells from paired PT (primary tumor) and LNMT (lymph node metastasized tumors) of 8 LNMT patients. Cells were colored according to their clusters







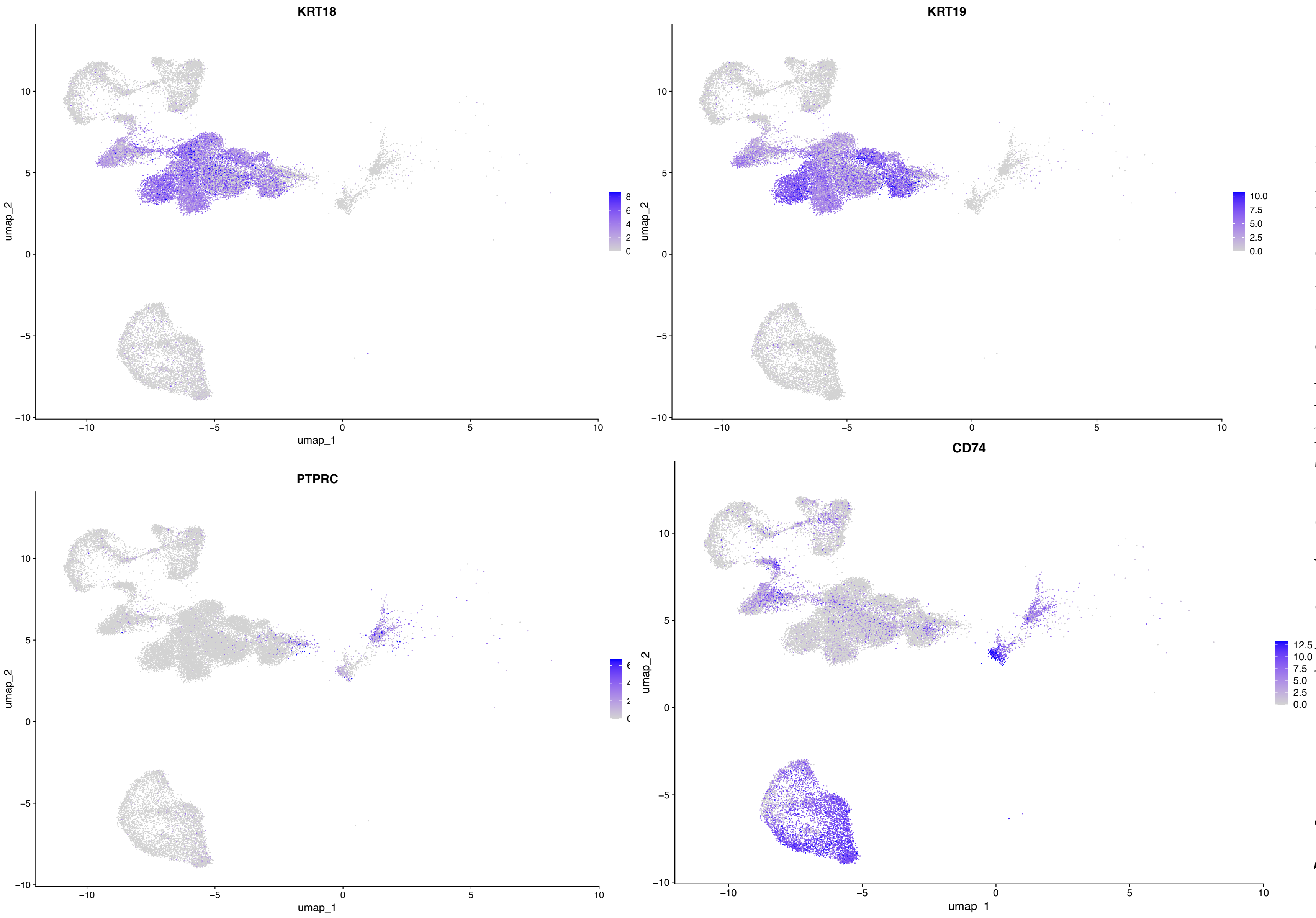
UMAP embedding plot showing remaining clusters cells from paired PT (primary tumor) and LNMT (lymph node metastasized tumors) of 8 LNMT patients, after CD45-positive cells are removed (hematopoietic cells)



What is left in the tumor after hematopoietic cells are removed.

Major cell types include epithelial cells, cancer associated fibroblasts (CAFs) and enfothelial cells (TECs).

Mitotic cells (KI67+ cells are mostly of hematopoietic origin)



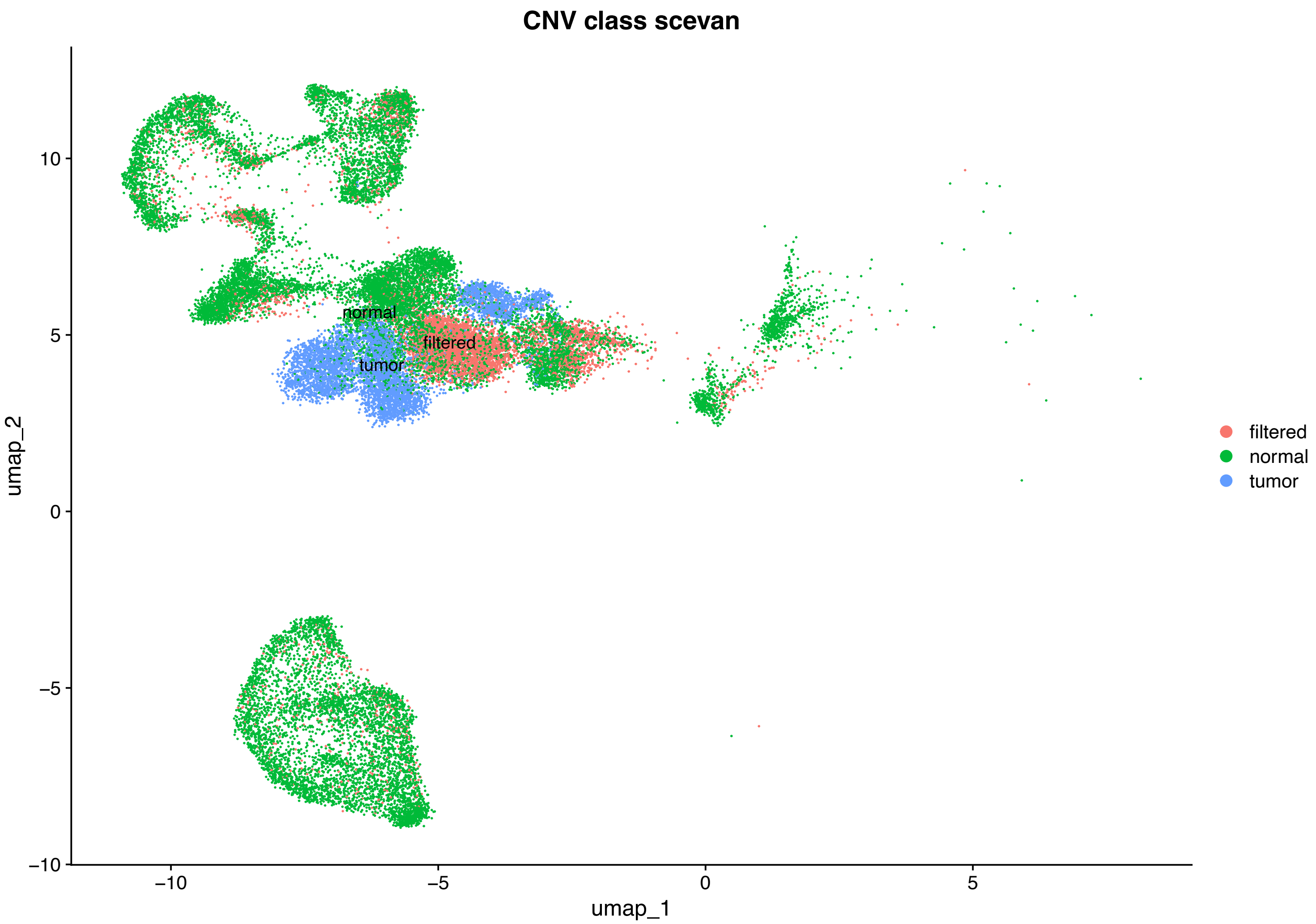
Where are the tumor cells?

Single cell profiles from metastatic breast cancer lymph nodes.

Cells from the hematopoietic lineage (B cells, CD4+ and CD8+ T cells, DCs, macrophages, NK and plasma cells) were previously removed from the lymph nodes. Tumor stromal cells, like endothelial cells (TEC), and CAFs, were also identified and removed. CD74 is involved in antigen presentation (MHC class II).

To select true cancer cells, we can infer ploidy by using copykat ans scevan.

Inferring CNV status



To select true cancer cells, we can infer ploidy by using *copycat* or *scevan (R packages)*.

De Falco, A., Caruso, F., Su, XD. *et al.* A variational algorithm to detect the clonal copy number substructure of tumors from scRNA-seq data. *Nat Commun* **14**, 1074 (2023). https://doi.org/10.1038/s41467-023-36790-9

Why do we need Integration ?

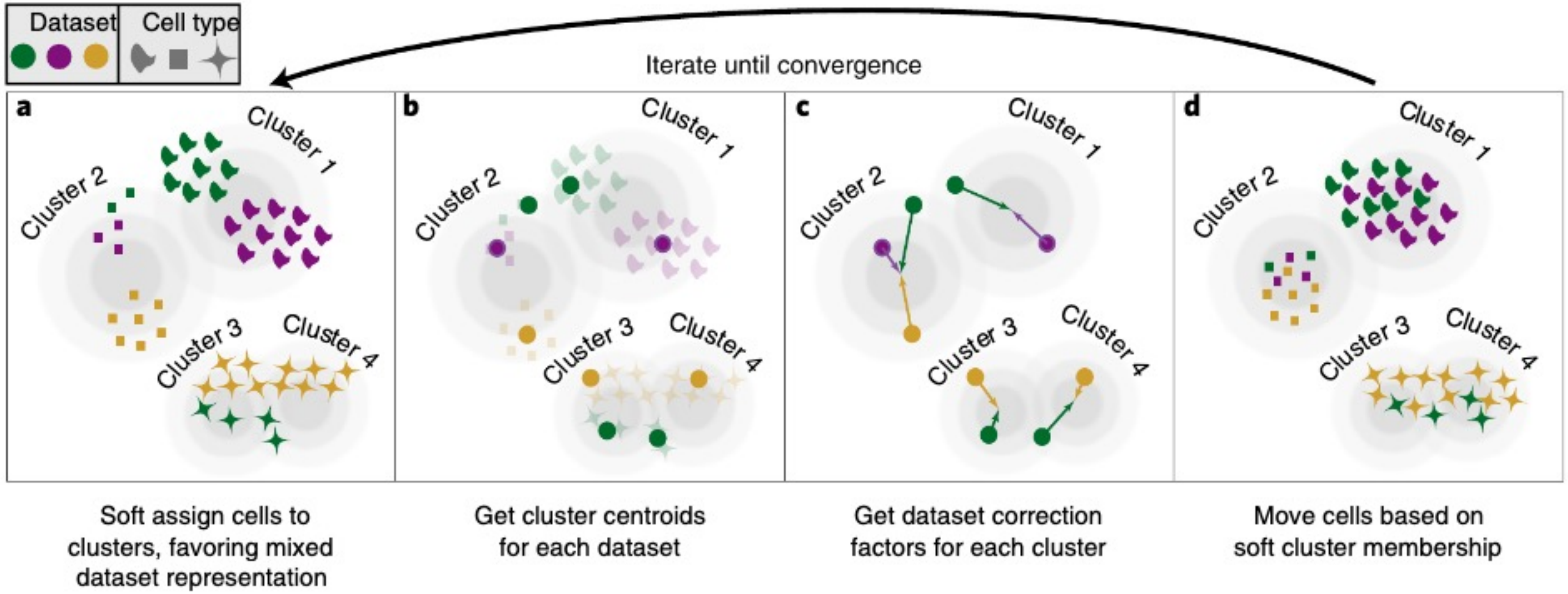


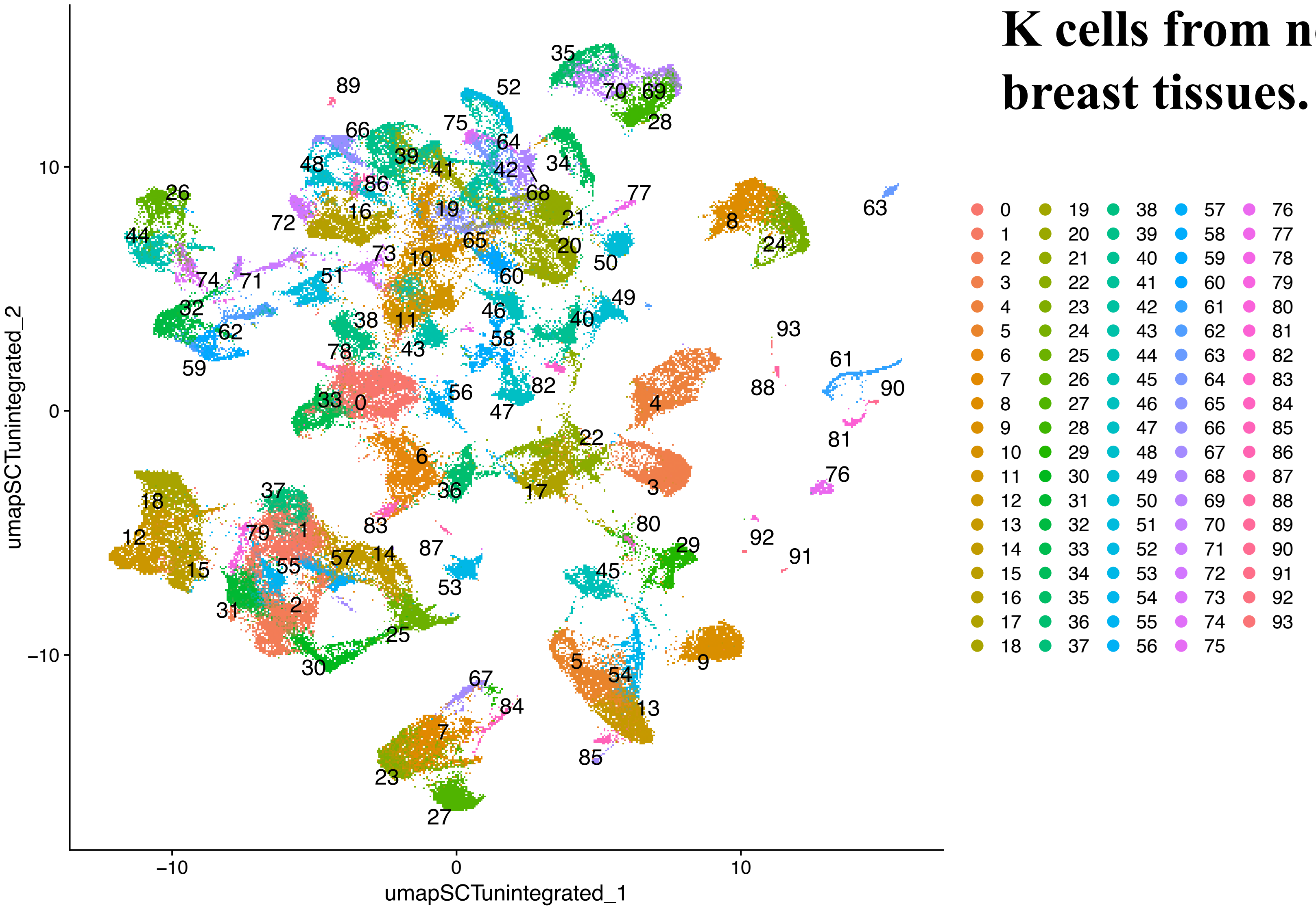
Fig. 1 | Overview of Harmony algorithm. PCA embeds cells into a space with reduced dimensionality. Harmony accepts the cell coordinates in this reduced space and runs an iterative algorithm to adjust for dataset specific effects. **a**, Harmony uses fuzzy clustering to assign each cell to multiple clusters, while a penalty term ensures that the diversity of datasets within each cluster is maximized. **b**, Harmony calculates a global centroid for each cluster, as well as dataset-specific centroids for each cluster. **c**, Within each cluster, Harmony calculates a correction factor for each dataset based on the centroids. **d**, Finally, Harmony corrects each cell with a cell-specific factor: a linear combination of dataset correction factors weighted by the cell's soft cluster assignments made in step **a**. Harmony repeats steps **a** to **d** until convergence. The dependence between cluster assignment and dataset diminishes with each round. Datasets are represented with colors, cell types with different shapes.

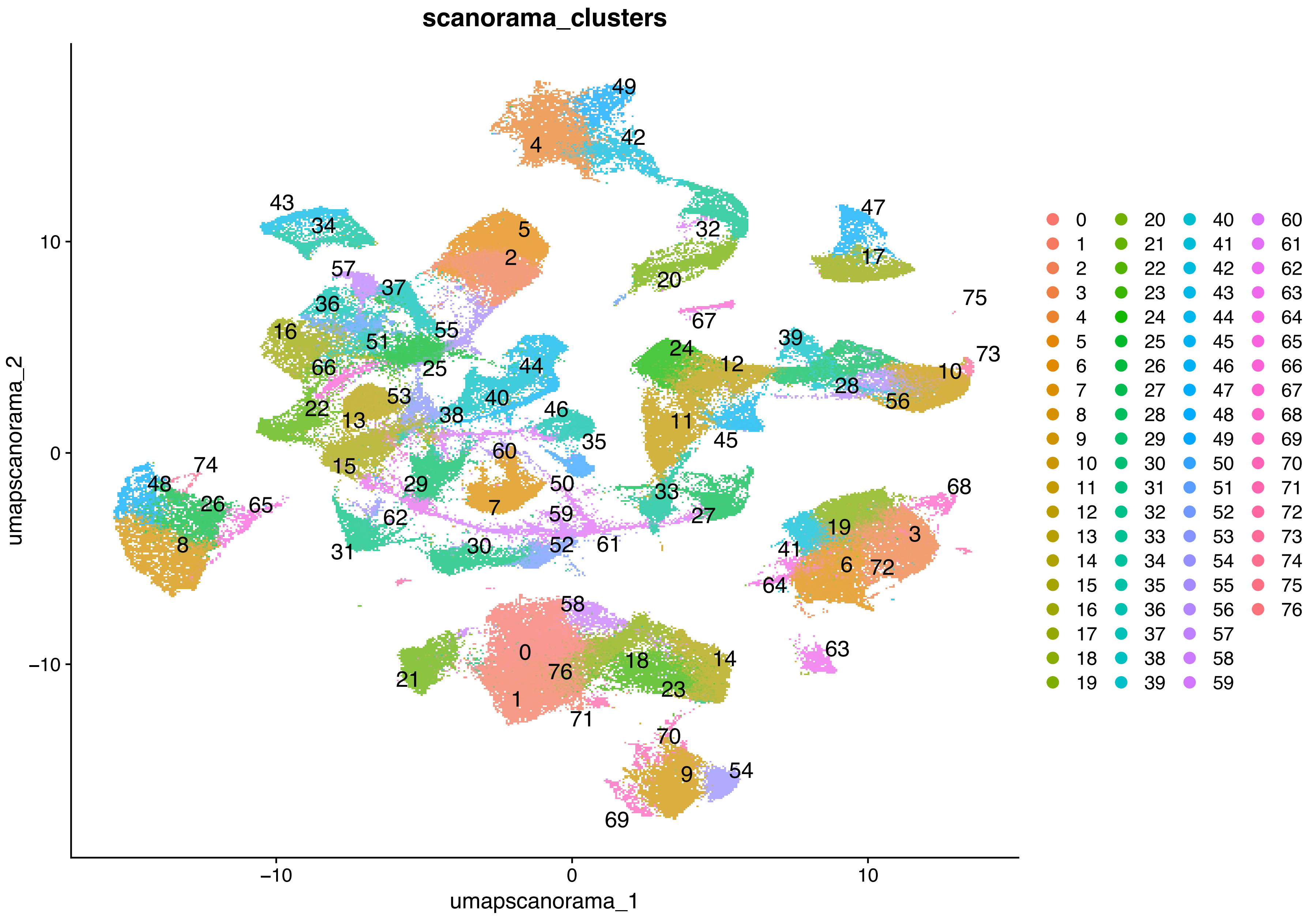
Stitching together many breast cancer scRNAseq datasets

- ER+
- HER2
- TNBC
- BRCA1
- Normal mammary tissue
- Metastatic lymph nodes

  - *Many patients and controls*
  - *Removing non cancer cells (if needed)*

The unintegrated dataset with 229 K cells from normal and cancer breast tissues.





The final integrated dataset contained 229 K cells.

The initial Scanorama integration yielded 94 clusters, some with cells engaged in cell cycle, as unveiled by the Wilcoxon tests.

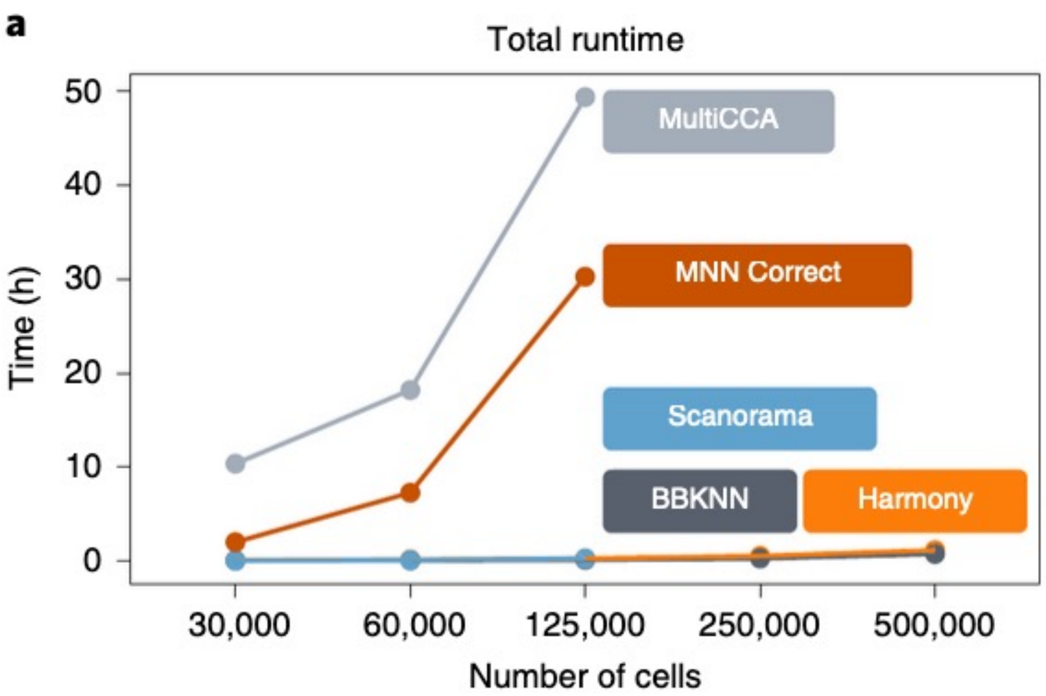
Therefore, we removed the cell cycle genes (n=97, Seurat S and G2 genes) prior to the integration, leading to a smaller number of Scanorama clusters (n=77).

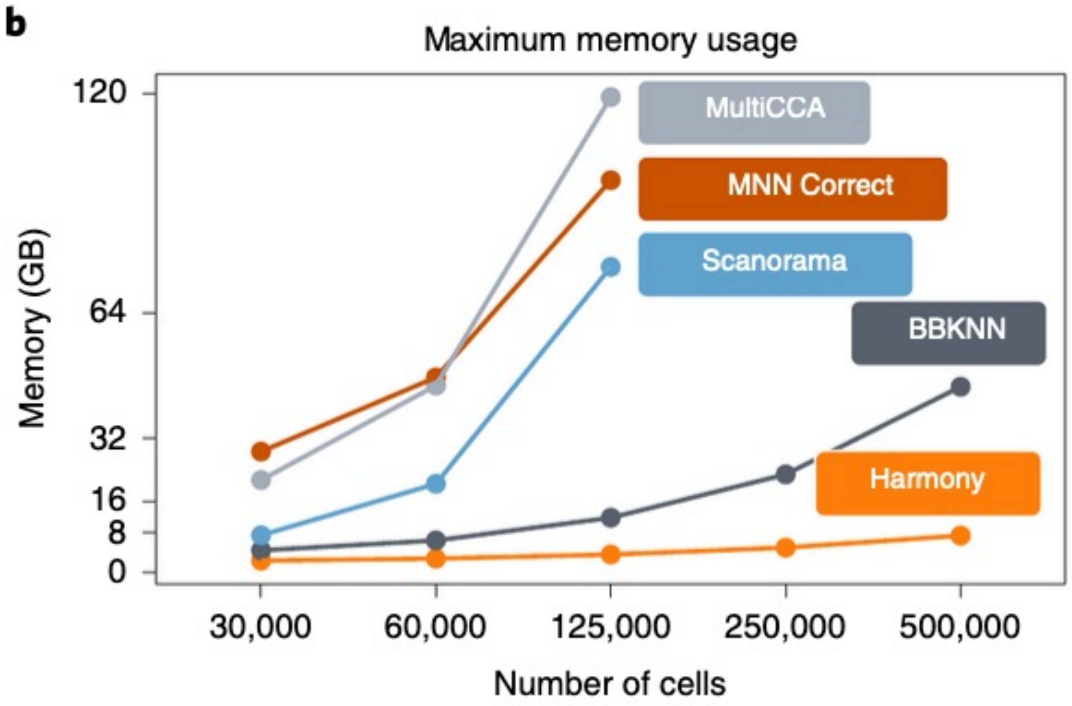
Integration methods

- Harmony and Seurat – R packages
- Scanorama and scVI Python modules run from R using reticulate

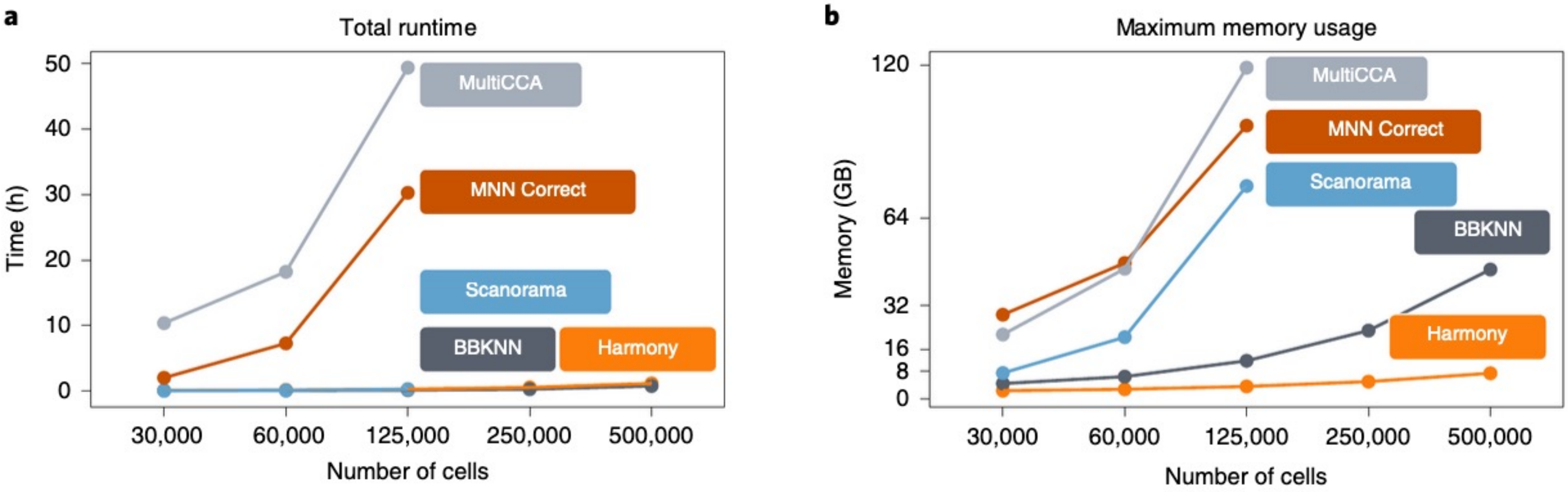
- GPU : scVI

Resources needed for Integration of scRNAseq datasets



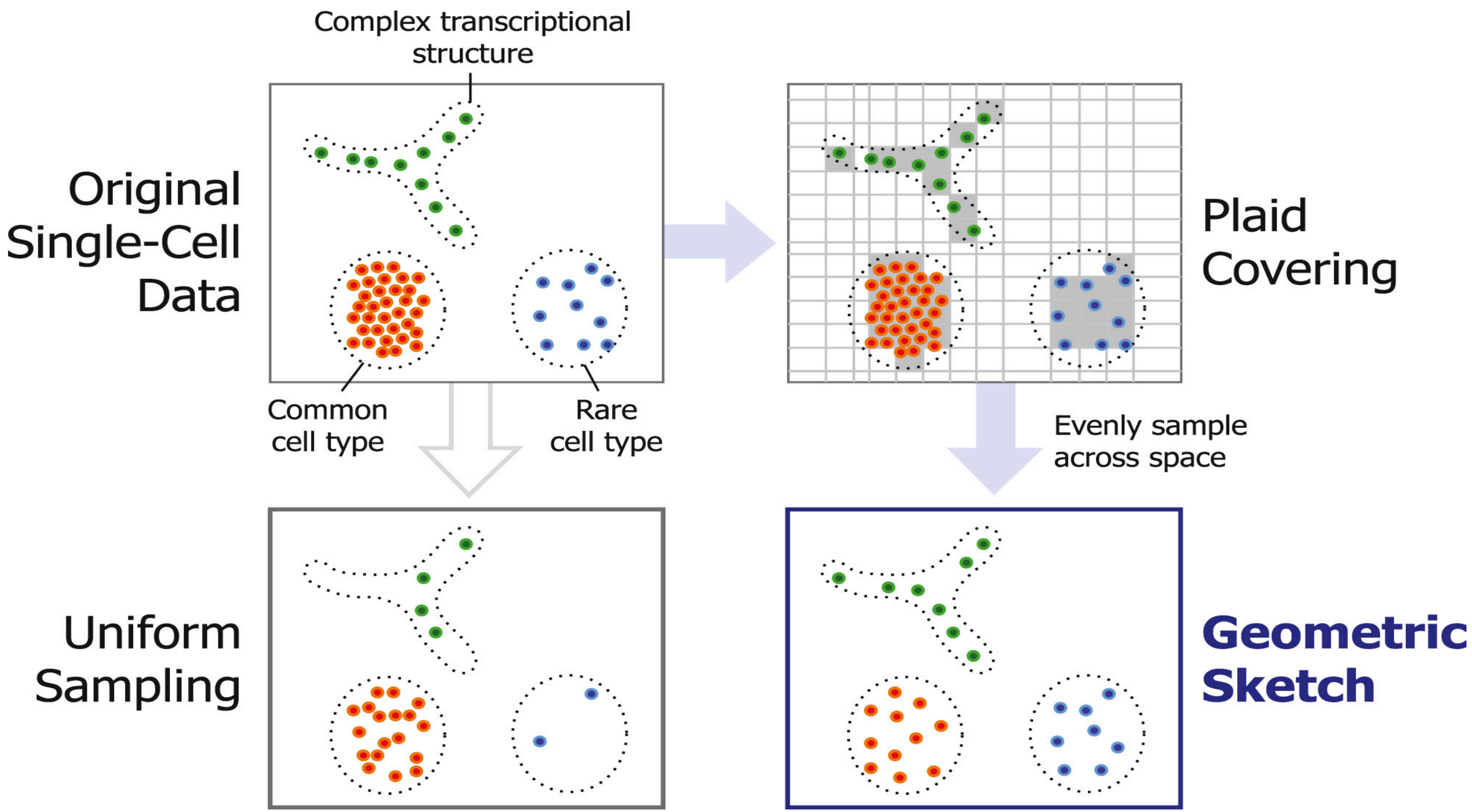


Resources needed for Integration of scRNAseq datasets



Small cohort : 100 patients (50K per patient) = 5 million cells

Remedy for limited resources: *Sketching*



Benchmarking Integration of scRNAseq

a

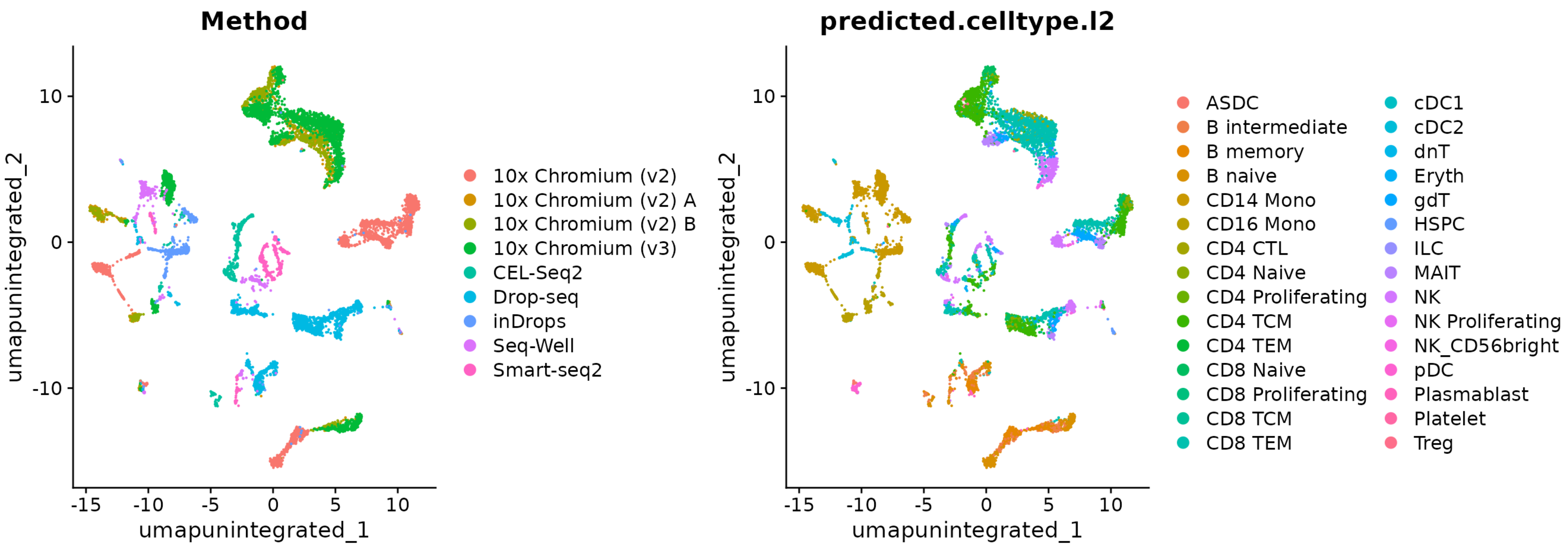
	Considerations	scANVI	Scanorama embed	scVI	FastMNN embed	scGen	Harmony	FastMNN gene	Seurat v3 RPCA	BBKNN	Scanorama gene	ComBat	MNN	Seurat v3 CCA	trVAE	Conos	DESC	LIGER	SAUCIE embed	SAUCIE gene
Input	Programming language	Python	Python	Python	R	Python	R	R	R	Python	Python	Python / R	Python / R	R	Python	R	Python	R	Python	Python
Input	Method runs without additional information	✗				✗														
Scib results	Consistent top performer	✓	✓	✓		✓														
Scib results	Top method on small/ simple tasks		✓		✓	✓	✓													
Scib results	Top method on large/ complex tasks	✓	✓	✓		✓														
Scib results	Top method on ATAC data	–		–			✓											✓		
Task details	Integrates strong batch effects	✓	–	–		✓			–	–				–						
Task details	Top method for recovery cell states or modules	✓	✓								✓	✓	✓							
Task details	Confounding of bio and batch variance	✓	–			✓														
Task details	Top method for trajectories	–	✓	–	✓	✓														
Task details	Method deals with varying compositions											✗								
Speed	Fast method for quick results									✓		✓								
Speed	Scales well to large datasets on CPU	✓	–	✓						✓	–								✓	✓
Speed	Method has GPU support	✓		✓		✓									✓		✓		✓	✓
Speed	Scales well to feature spaces beyond genes														✓	✓				
Output	Method shows corrected expression					✓		✓	✓		✓	✓	✓	✓						✓
Output	Method gives relative cell embeddings									✗						✗				

✓ Fulfills the criterion
– Partial fulfillment of criterion
✗ Does not fulfill criterion
Python
R

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

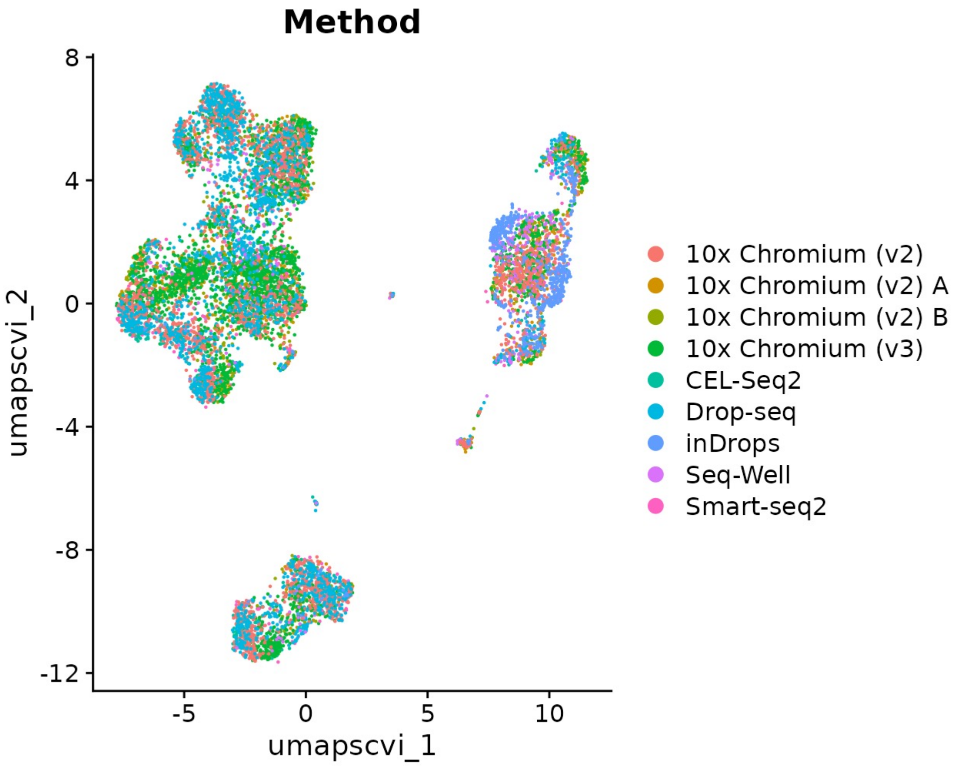
Automated cell type annotation

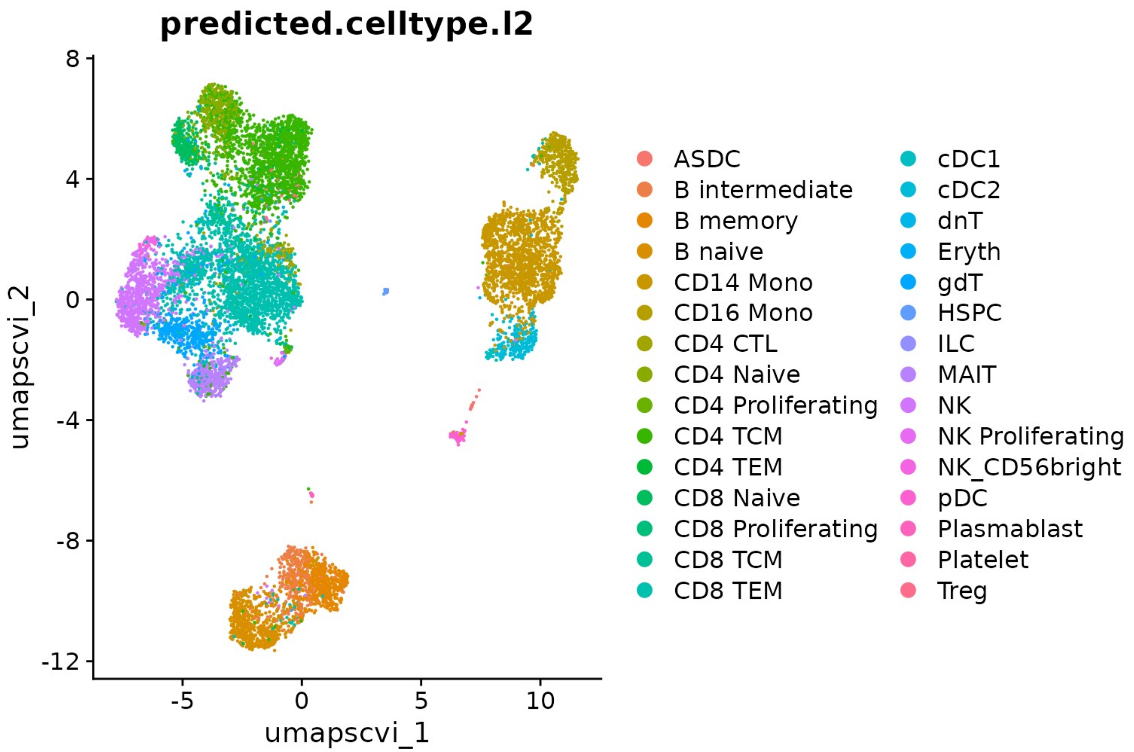
Automated annotation of scRNAseq : *Azimuth*



Unintegrated PBMC (blood) datasets

Automated annotation of scRNAseq : *Azimuth*



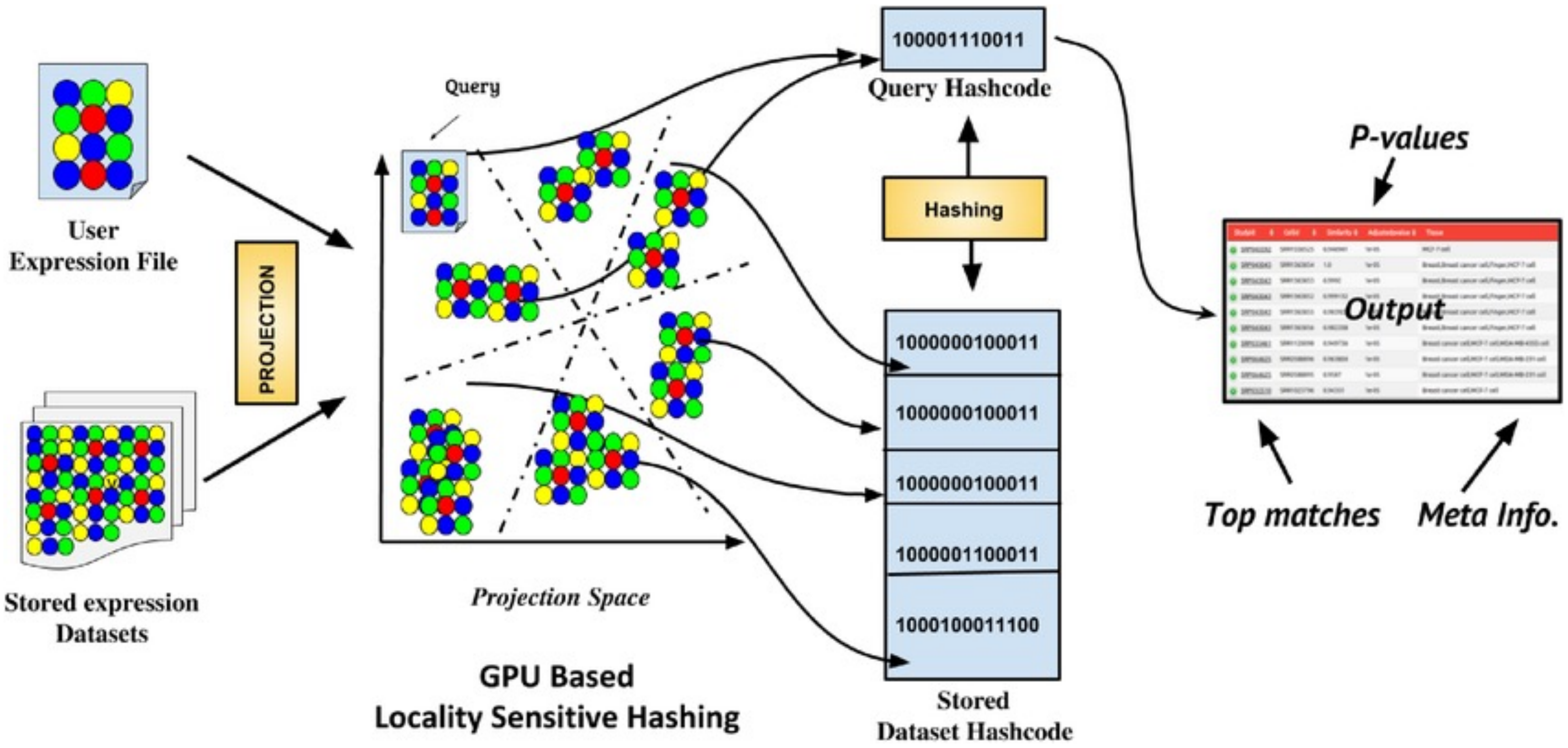


Integrated PBMC (blood) datasets

Automated cell annotation (using a reference)

- Azimuth
- singleR
- cellBlast
- scLearn
- cellAtlasSearch
- Many others

Automated annotation tools can be GPU-enabled



Conclusions

- The number and size of single cell cancer datasets is steadily increasing
- Integration of scRNAseq datasets is required to study cancer at a single cell level
- With larger number of cells projects become more informative and more demanding
- The current tools for single cell RNAseq need to be adapted to enable efficient management and analysis