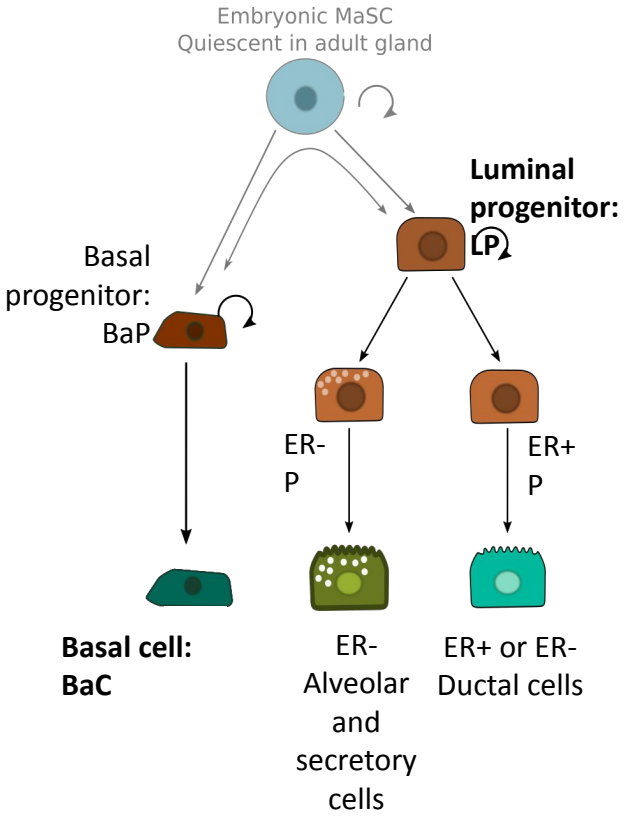
Integration methods in scRNAseq: which one ? why ?

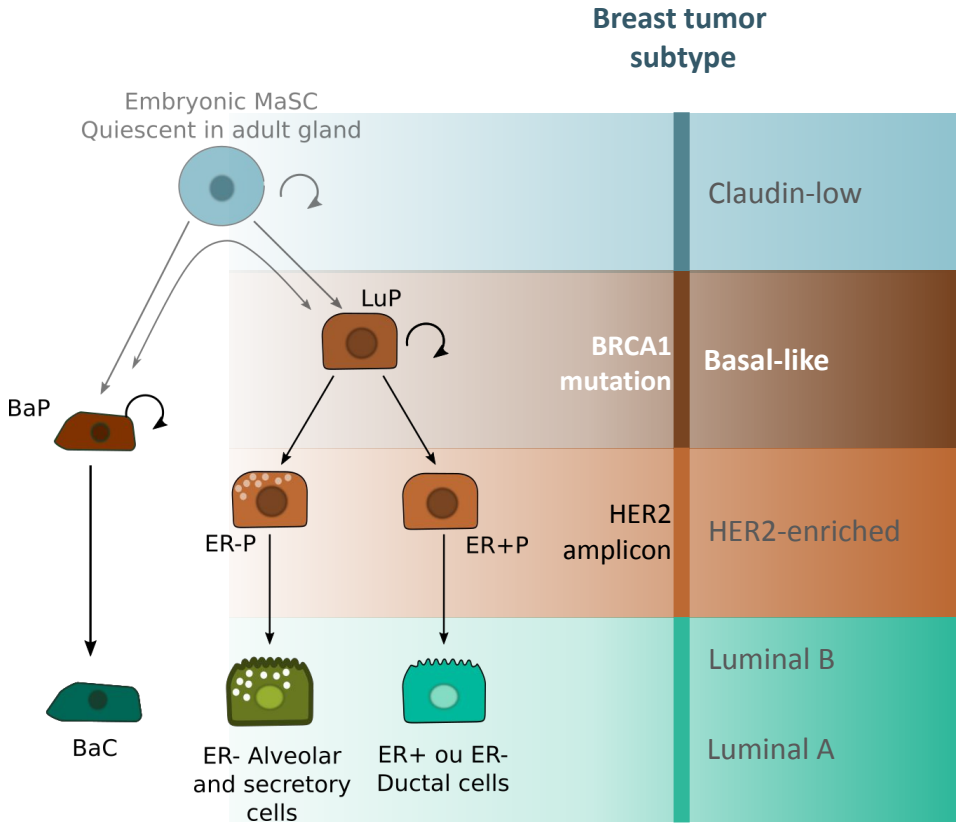
Use case in breast cancer tumorigenesis

Melissa Saichi
1st year PhD student
Dynamics of Epigenetic plasticity in Cancer, Institut Curie

The mammary gland hierarchy is well established

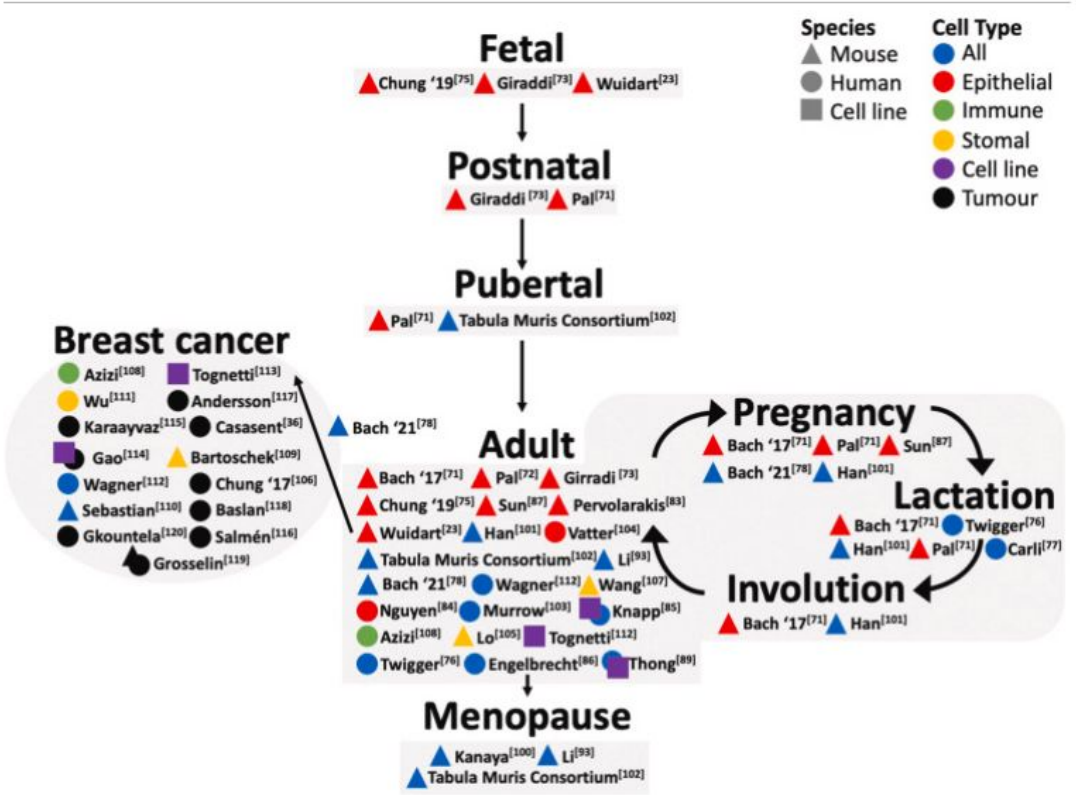


Initial cell identity is determinant for breast tumor subtype



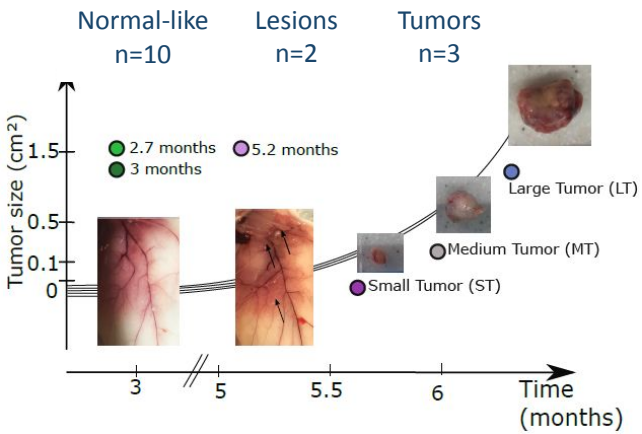
Adapted from Prat et Perou, Nat med 2009

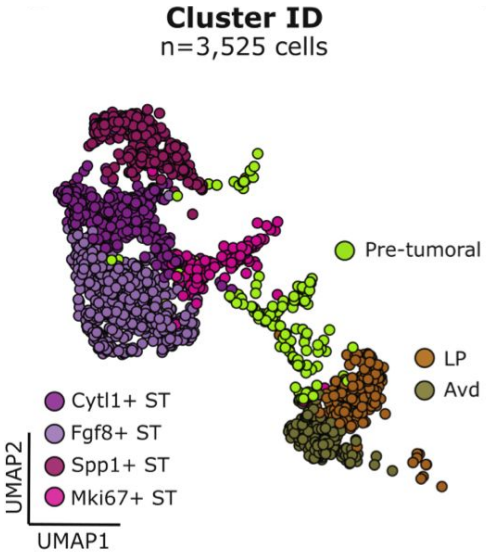
single cell omics revolutionized the characterization of mammary gland cell populations/states in both physiology & disease

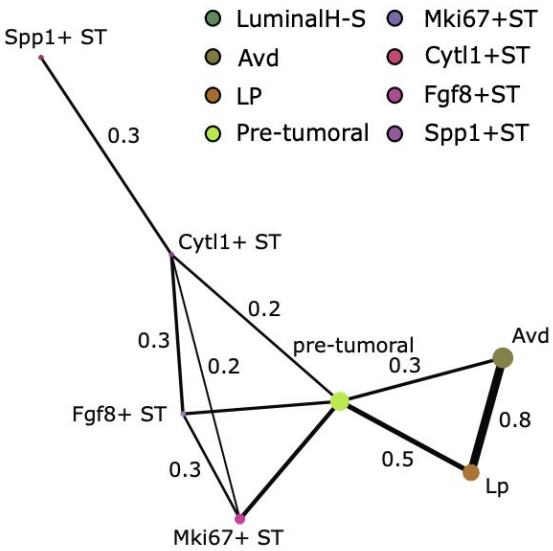


Aim1: Identification of state switches in a mouse model of basal-like tumorigenesis

Tp53 & *Brca1* deletion
In luminal progenitors





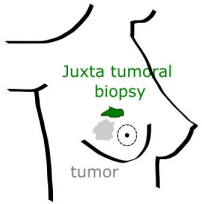


w PAGA algorithm

Abbreviations:
LP: Luminal Progenitor
Avd: Alveolar differentiated cells

Aim2: Searching for pre-tumoral states in humans…

snRNAseq profiling of frozen juxta-tumoral samples from 5 BRCA1 carriers



BRCA1 carriers
n=5 patients

sample ID	Status	Techno
sample1	Frozen	snRNAseq
sample2	Frozen	snRNAseq
sample3	Frozen	snRNAseq
sample4	Frozen	snRNAseq
sample5	Frozen	snRNAseq

Merge the 5 count matrices*

Preprocessing

Step1: Sanity check on doublet/nucleus detection

Step2: QC on count matrices

Step3: Classical workflow on annotation/ clustering ..

Step4: Validate annotations

Downstream

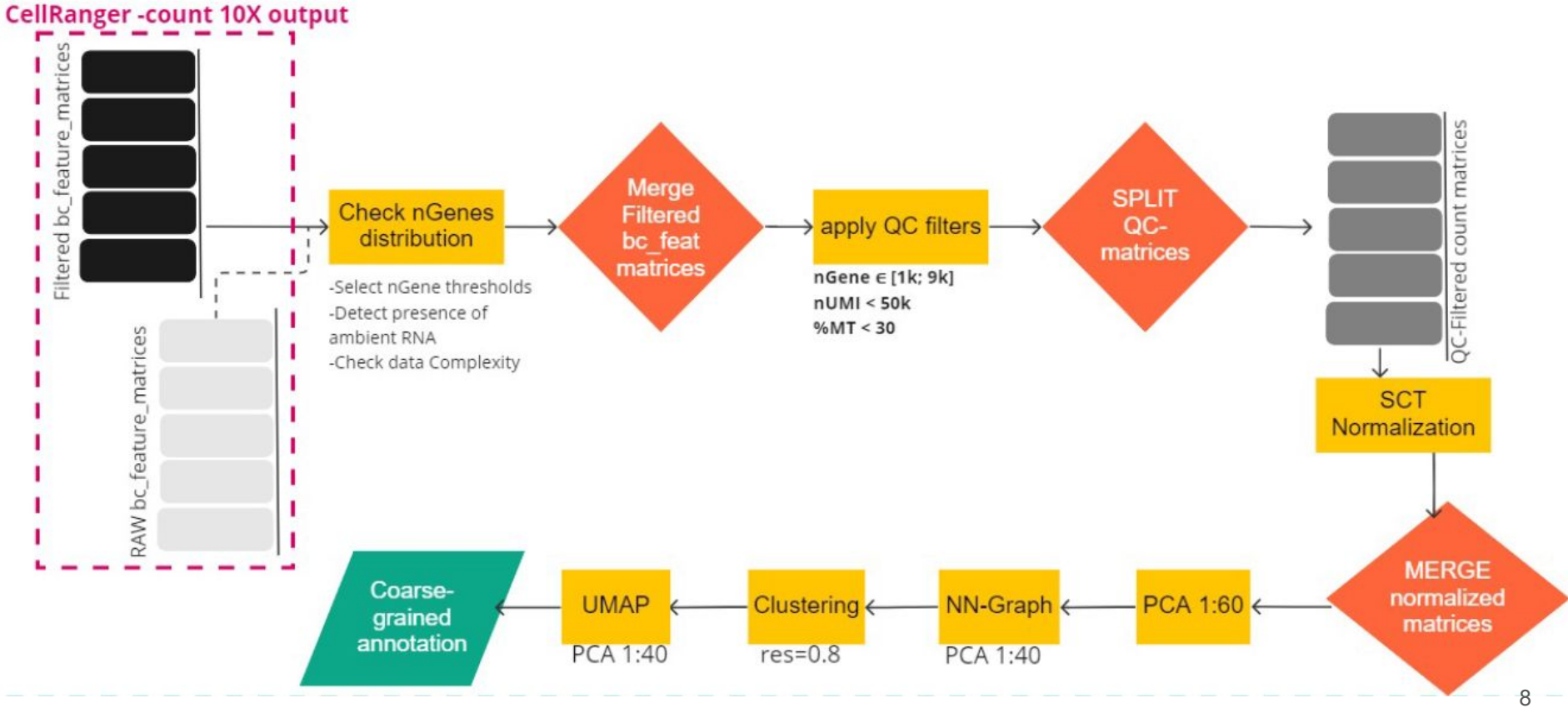
Step1: DEG & PEA

Step2: CNV inference

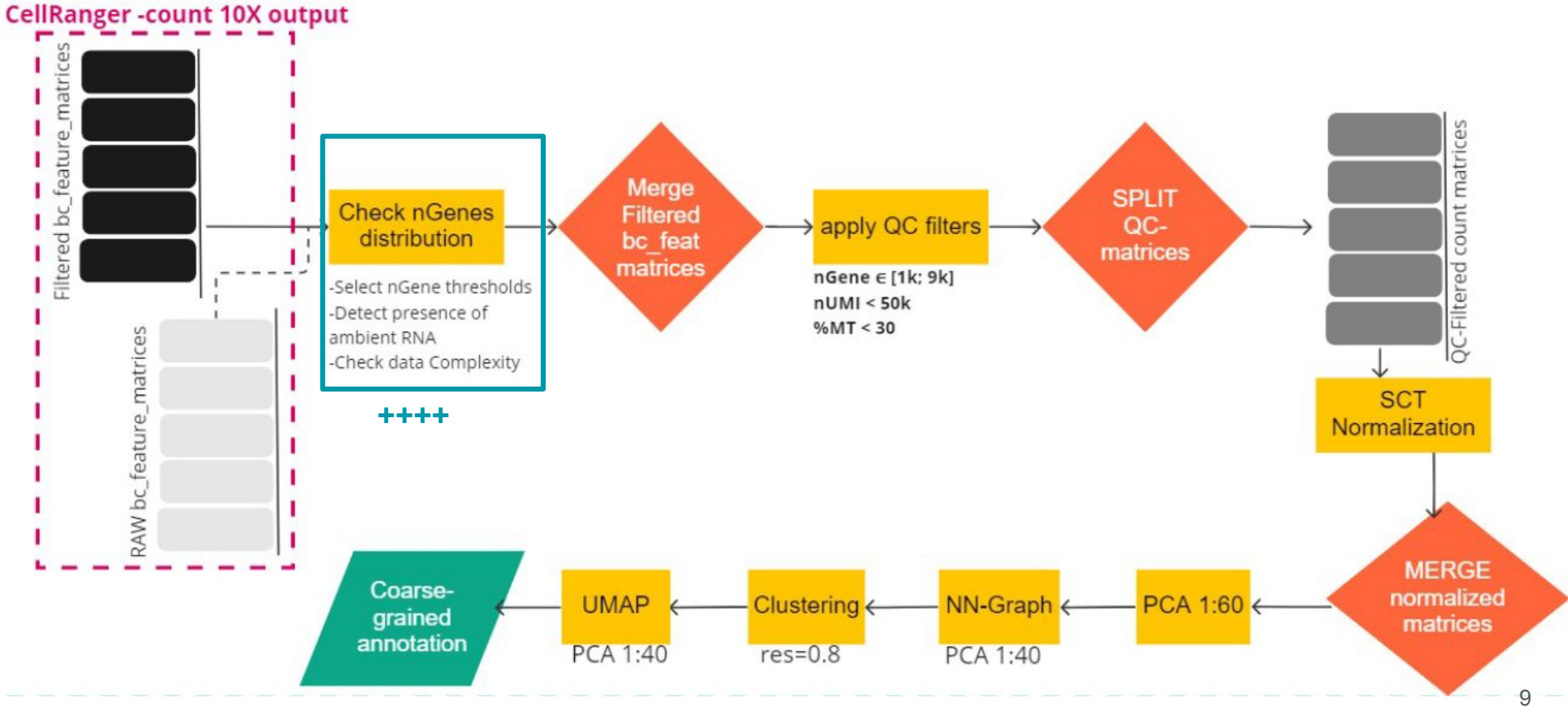
Step3: Comparison with previous findings

With AV Salomon @Pathology Dpt

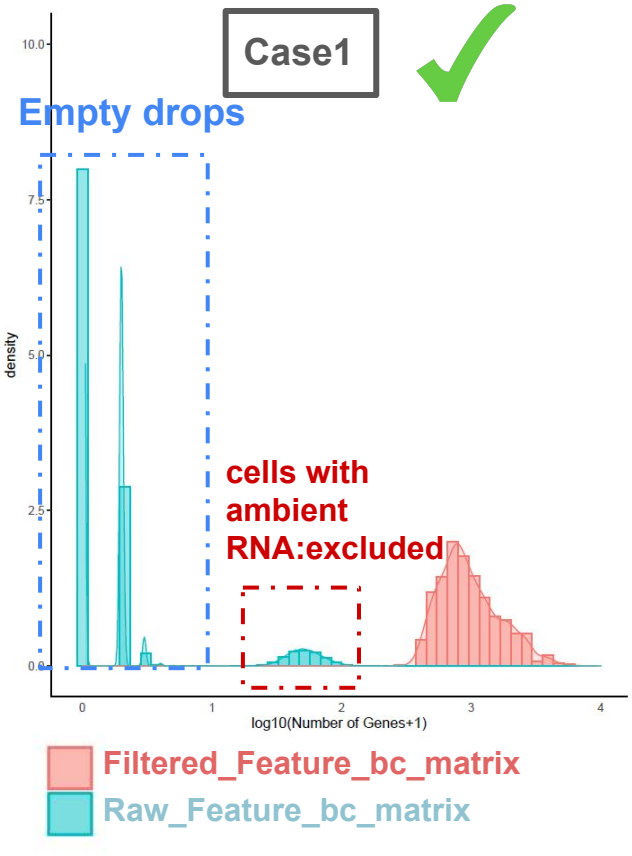
Overview of the pre-processing steps of the snRNAseq datasets



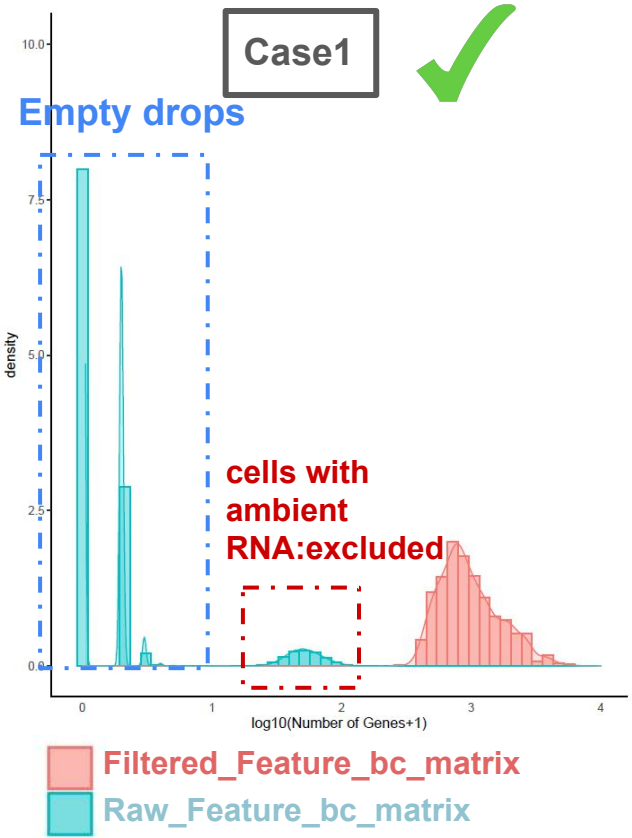
Overview of the pre-processing steps of the snRNAseq datasets

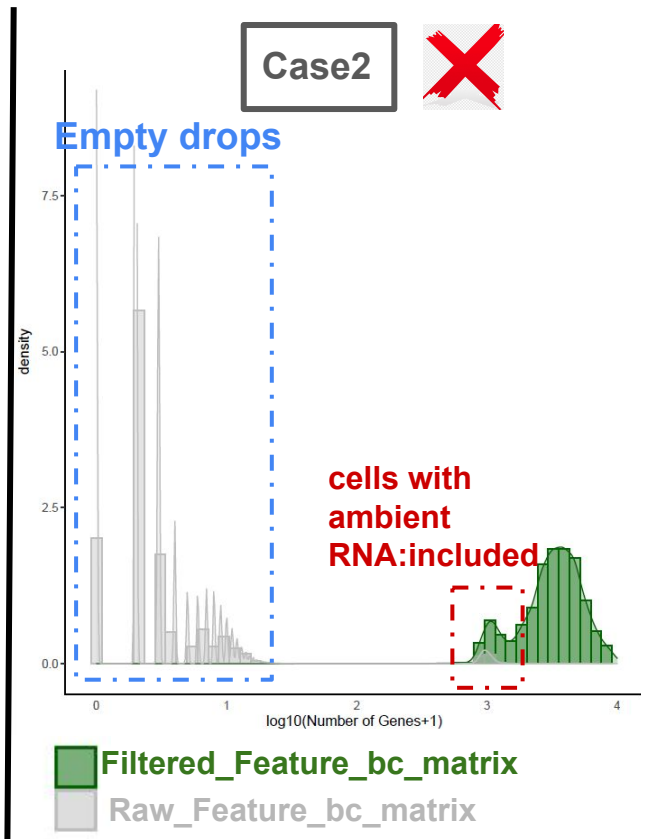


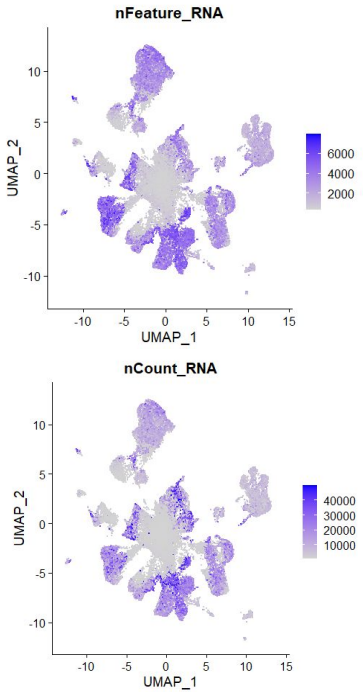
Comparing cellranger filtered & raw matrices per sample are crucial for downstream analysis



Comparing cellranger filtered & raw matrices per sample are crucial for downstream analysis

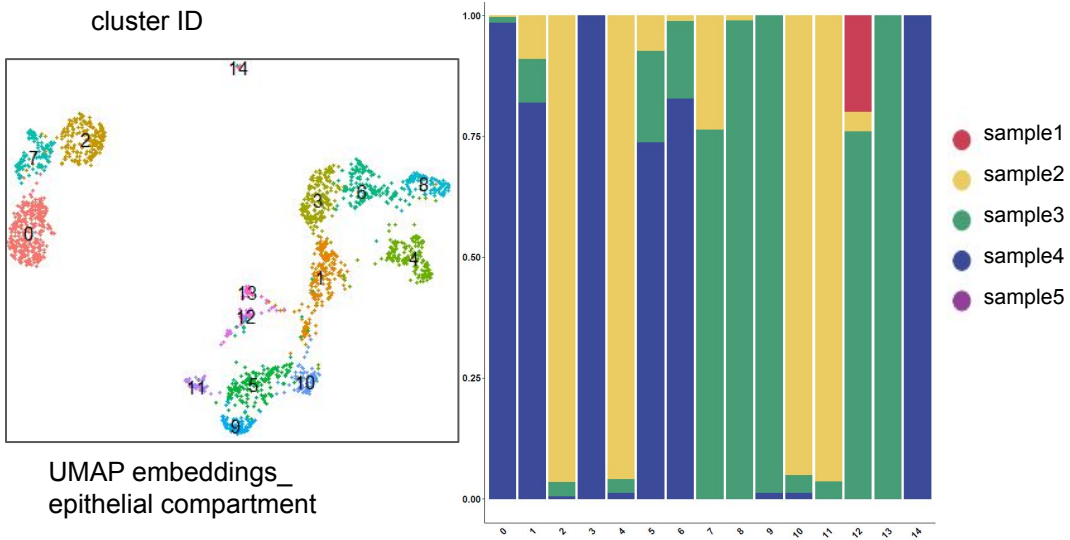






Simple merging of the 5 expression matrices generated heterogeneous cluster composition by sample

Focus on the cluster composition of the epithelial compartment



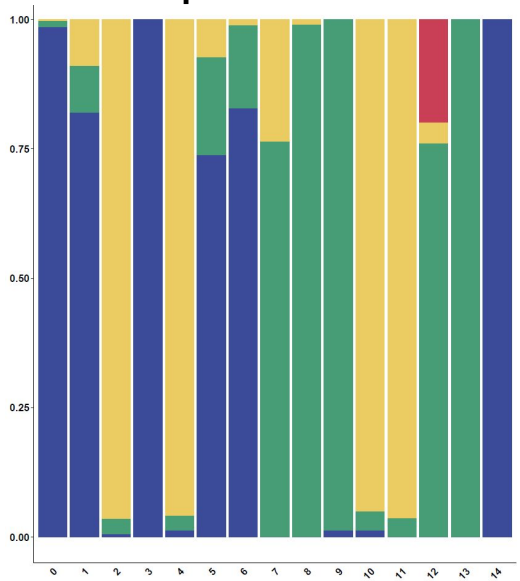
UMAP embeddings_
epithelial compartment

-Presence of "**sample-effect**" rather than a "batch effect" since all samples were processed at the same time using same conditions

-probably due to "**sample composition**"

How to quantify the “sample effect” on cluster composition?

a.Have a look at the representation of samples inside clusters



(+) visual representation

(-) - not quantitative

- assumption about similarity in cell identity composition between the samples

How to quantify the “sample effect” on cluster composition?

b.Compute “Shannon index” diversity metric

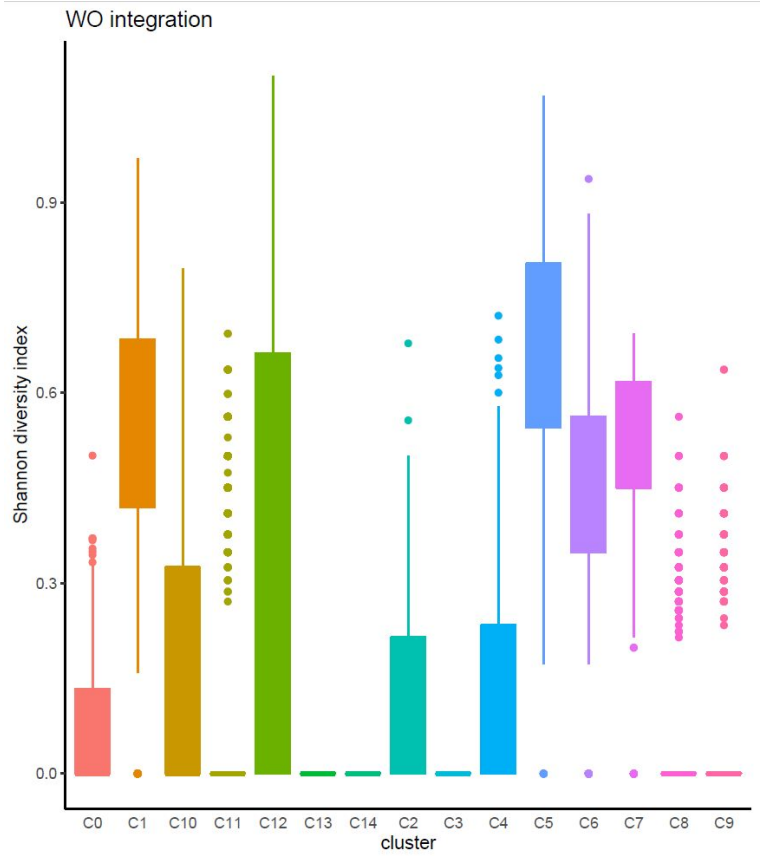
confusion matrix

	S0	S1	S2	…	SN	sum
C1	35	0	26	0	8	N1
C2	21	10	0	17	0	N2
…	18	9	23	26	0	N3
CN	9	2	7	9	1	Nn

n

p=n/N1
H = -sum(p*log(p))

run 1000 simulations



(+) quantitative
(-) - assumption about similarity in cell identity composition between the samples

Simple merging of the 5 expression matrices generated heterogeneous cluster composition by patients

Limitations:

-cells were **grouped by sample** rather than **identity**

-challenge to **define a sample of reference** to seek for CNV alterations

Solution:

-Minimizing the sample of origin covariate would help deciphering true biological signal

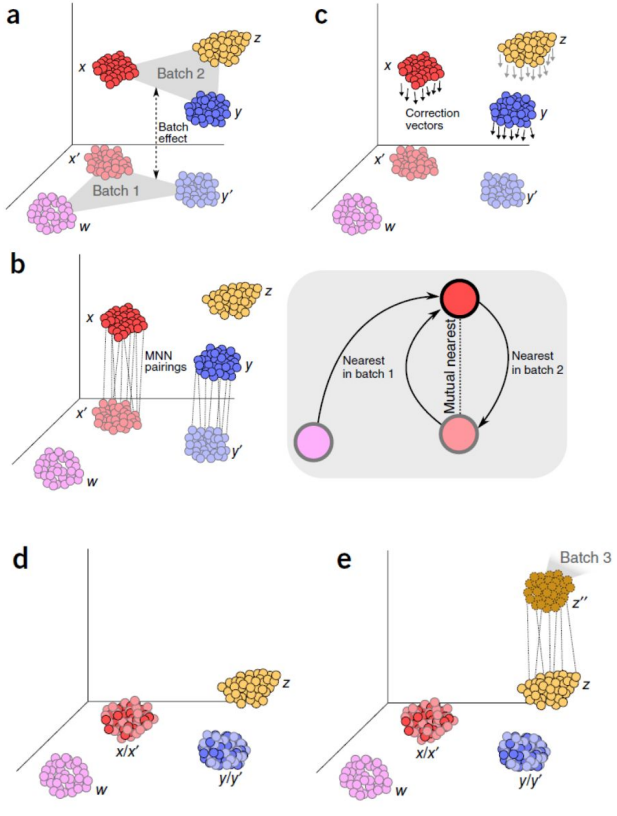
-Common solution to reduce covariates effect in scRNAseq is **“Data Integration”**

Computational challenges

- What’s the best suitable integration method?
- How to choose it?
- What is the **optimal number** of biologically relevant clusters?

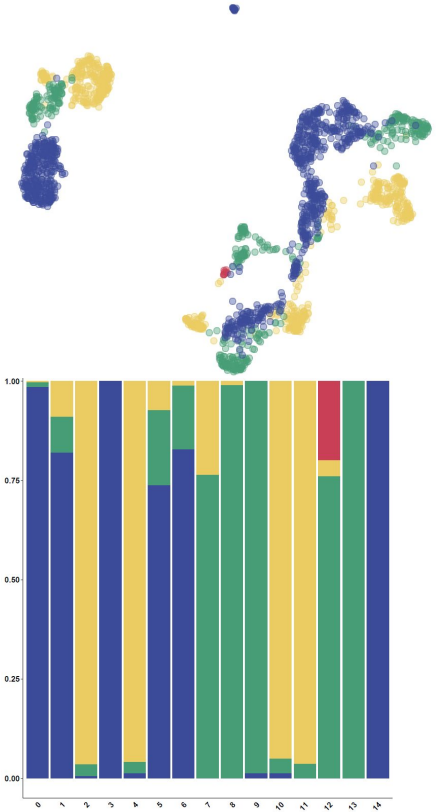
Briefly, what's an integration method in sc data?

Example of how integration works:

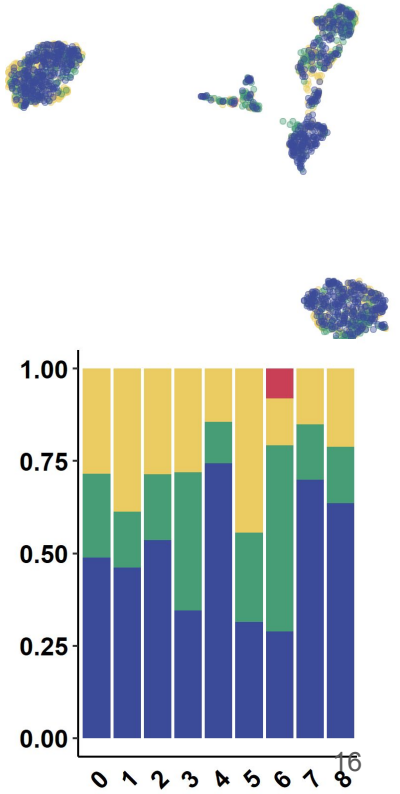


Application on our dataset:

Before integration



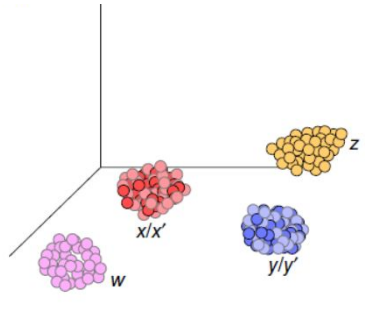
After integration



5 integration methods were used on the 5 merged samples(1)

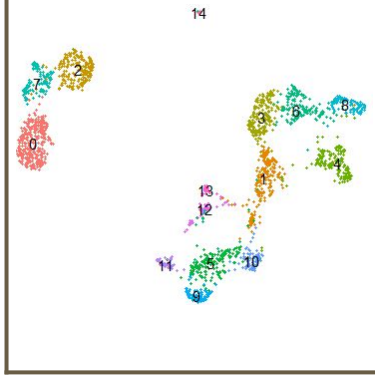
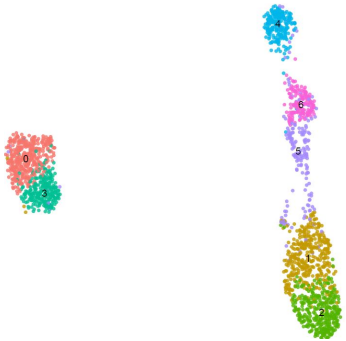
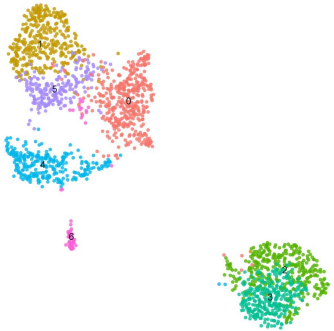
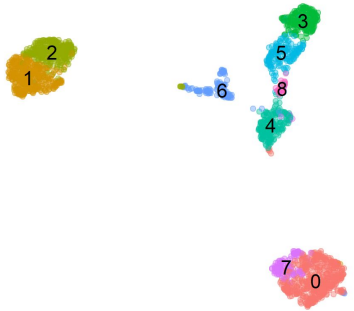
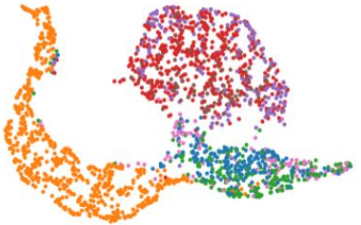
sample ID	After QC
sample1	521
sample2	737
sample3	788
sample4	366
sample5	1313

Method				
1	scANVI*		HVG	–
2	Scanorama		HVG	+
3	scVI		HVG	–
4	fastMNN		HVG	–
5	scGen*		HVG	–
6	Harmony		HVG	–
7	fastMNN		HVG	–
8	Seurat v3 RPCA		HVG	+
9	BBKNN		HVG	–
10	Scanorama		HVG	+
11	ComBat		HVG	–
12	MNN		HVG	+
13	Seurat v3 CCA		HVG	–
14	trVAE		HVG	–
15	Conos		HVG	–
16	DESC		FULL	–
17	LIGER		HVG	–
18	SAUCIE		HVG	+
19	Unintegrated		FULL	–
20	SAUCIE		HVG	+
Rank	Name	Output	Features	Scaling



scRNAseq classical workflow

4 integration methods were used on the 5 merged samples(2)

Not integrated(merged)	Seurat V3	Fast MNN	Harmony	scGen
				

NB: the colors and cluster IDs do not match between the integration methods

Strategies to define the most suitable integration method

Strategy 1: Compare the **mean diversity index** between the integration methods

Strategy 2: Find a well-defined mammary cell population (example: Basal cells, LPs..), and check the “integration efficiency” through the **number of Differentially Expressed Genes (DEG) between samples within each population**

Strategy 3: Look for **a population of interest** & compare the specificity of the enriched pathways across the integration methods

=> Choose the integration method which matches the criteria above

Strategy 1: Cluster composition ~ Diversity index

1)For **each integration method**, compute clustering at a high resolution (=1.0)

2)-Replicate 1000 times:

-Subsample 200 cells from the entire integrated matrix

-Create a contingency table of: the cluster IDs & the sample of origin:

	Cluster1	Cluster2
sample1	65	23
sample2	87	9
sample3	176	0
…	…	…
sampleN	x	y

3) Calculate the Shannon diversity index for each cluster

Strategy 1: Cluster composition ~ Diversity index

1)For **each integration method**, compute clustering at a high resolution (=1.0)

2)-Replicate 1000 times:

-Subsample 200 cells from the entire integrated matrix

-Create a contingency table of: the cluster IDs & the sample of origin:

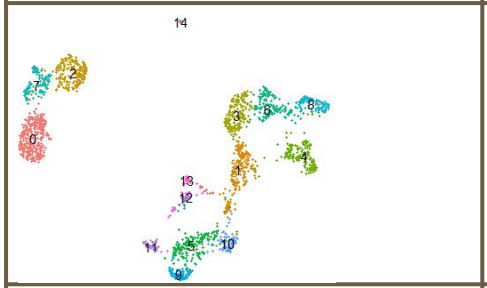
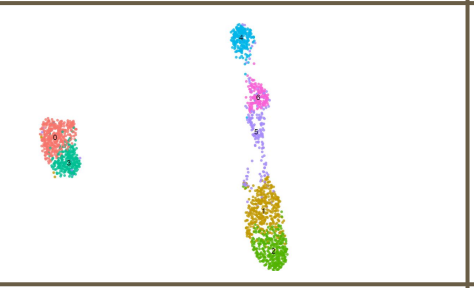
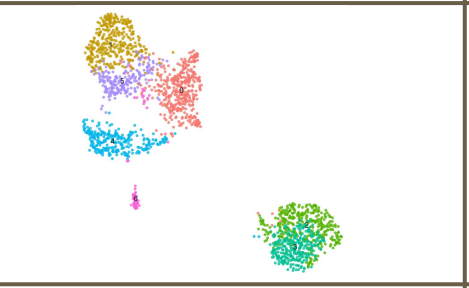
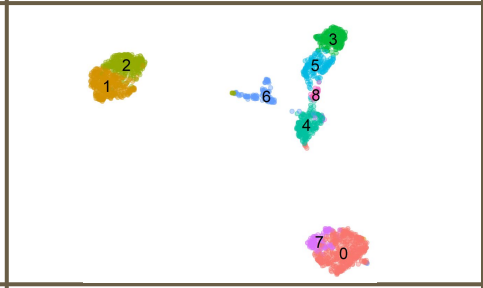
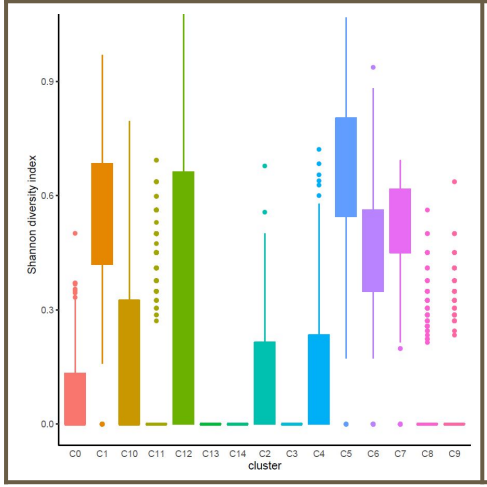
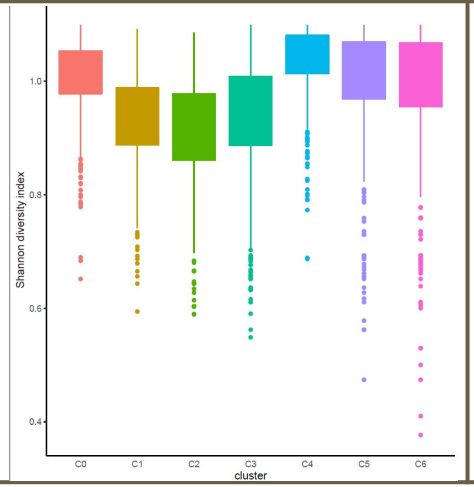
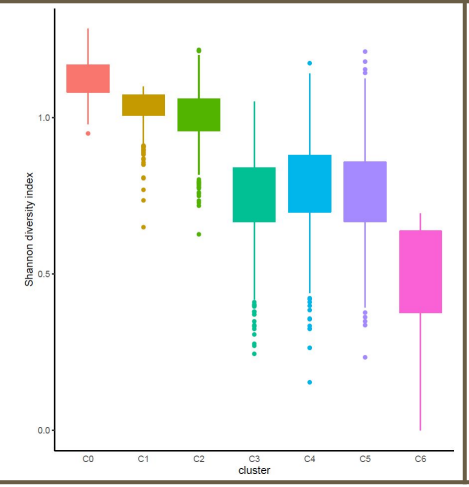
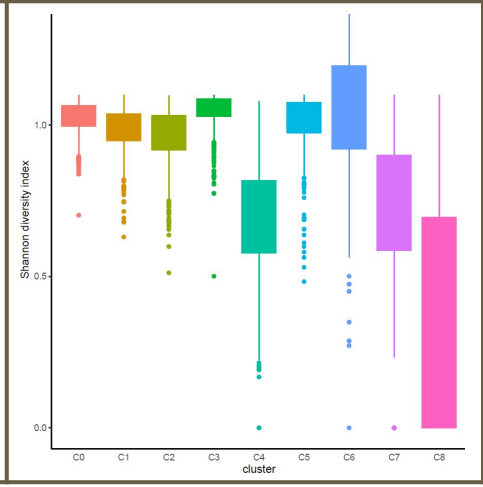
	Cluster1	Cluster2
sample1	65	23
sample2	87	9
sample3	176	0
…	…	…
sampleN	x	y

3) Calculate the Shannon diversity index for each cluster

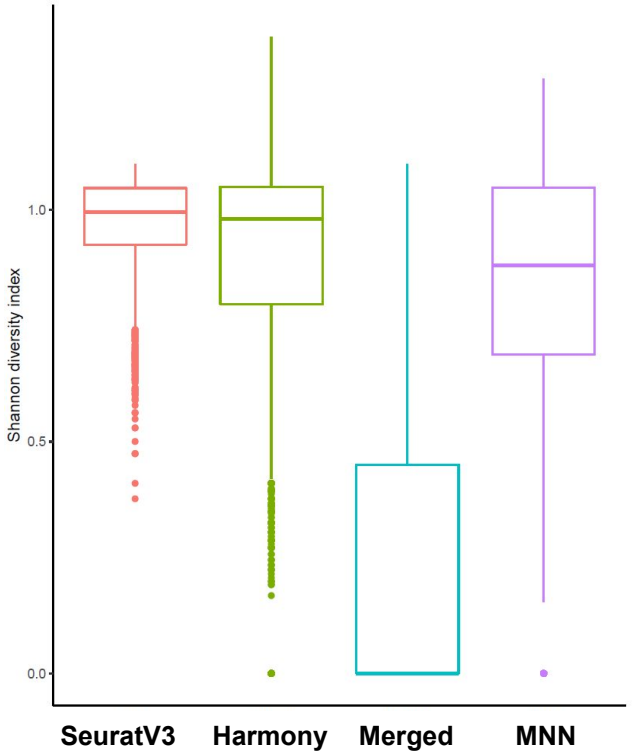
1000 shannon diversity values for each cluster in each integration method

	Cluster1	Cluster2	Cluster3
iteration1	0.76	0.96	0.65
iteration2	0.56	0.65	0.06
iteration3	0.2	0.12	0.32

Strategy 1: Cluster composition ~ Diversity index

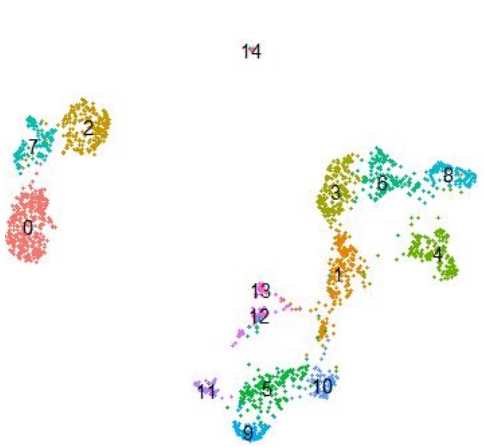
Not integrated	Seurat V3	Fast MNN	Harmony
			
			

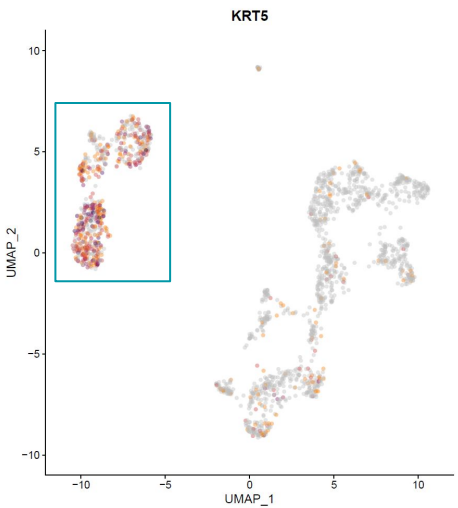
Strategy 1: Harmony & Seurat V3 integration methods give the highest Shannon values

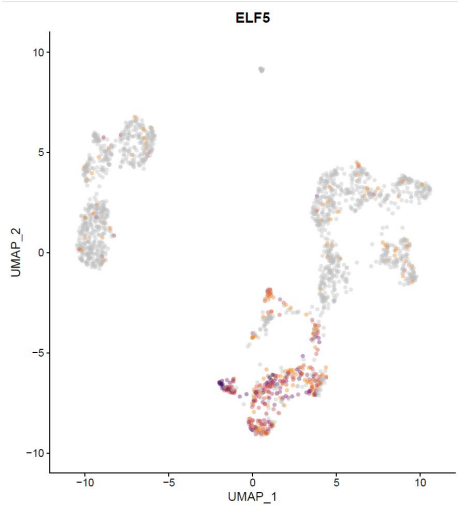


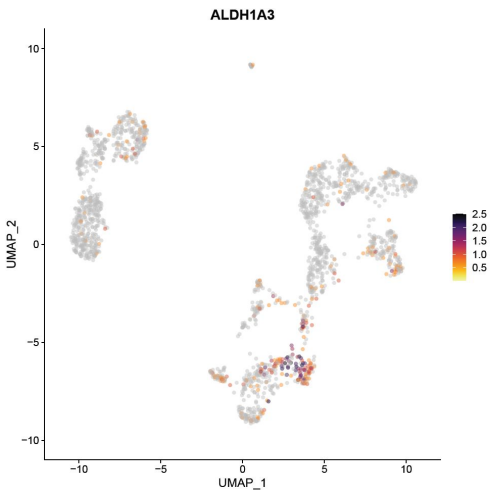
High Shannon index values => High diversity=> Homogeneous mixing of samples within the cluster

Strategy 2: a) Identify a known cell type cluster







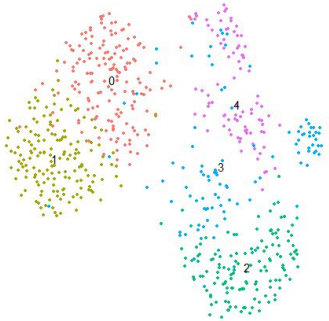
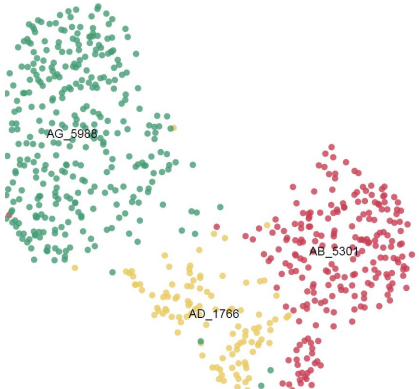
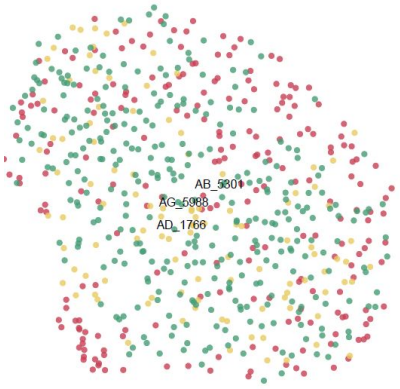
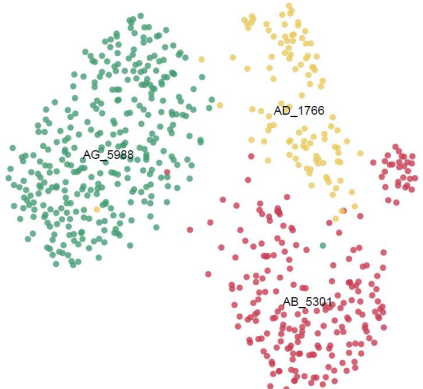


Basal cluster

Luminal Progenitors

Strategy 2: b) represent the sample distribution ~cell population

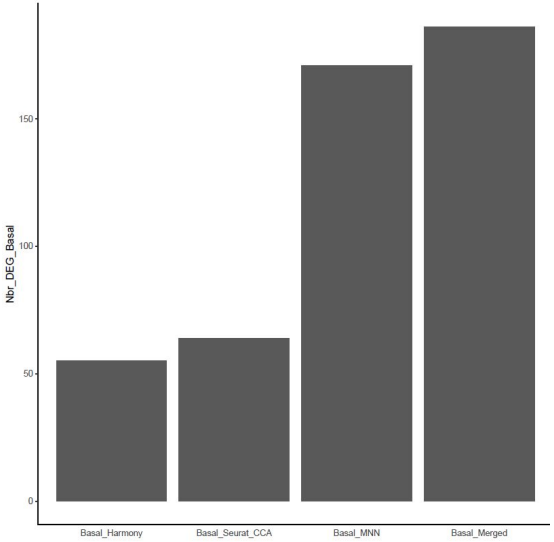
Case of Basal cells (Krt5 positive)

Not integrated	Seurat V3	Harmony
		
		

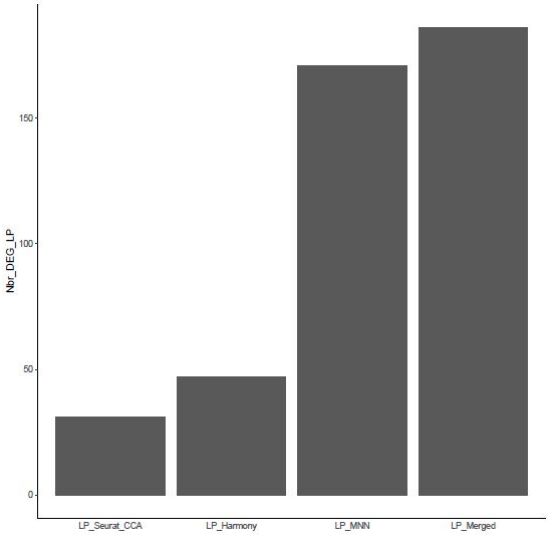
c) In each known cell population (Basal, LPs), compare the number of DEG between cells from each sample

Assumption: Low number of DEG between cells from each sample ⇔good integration method

Basal cluster



LP cluster



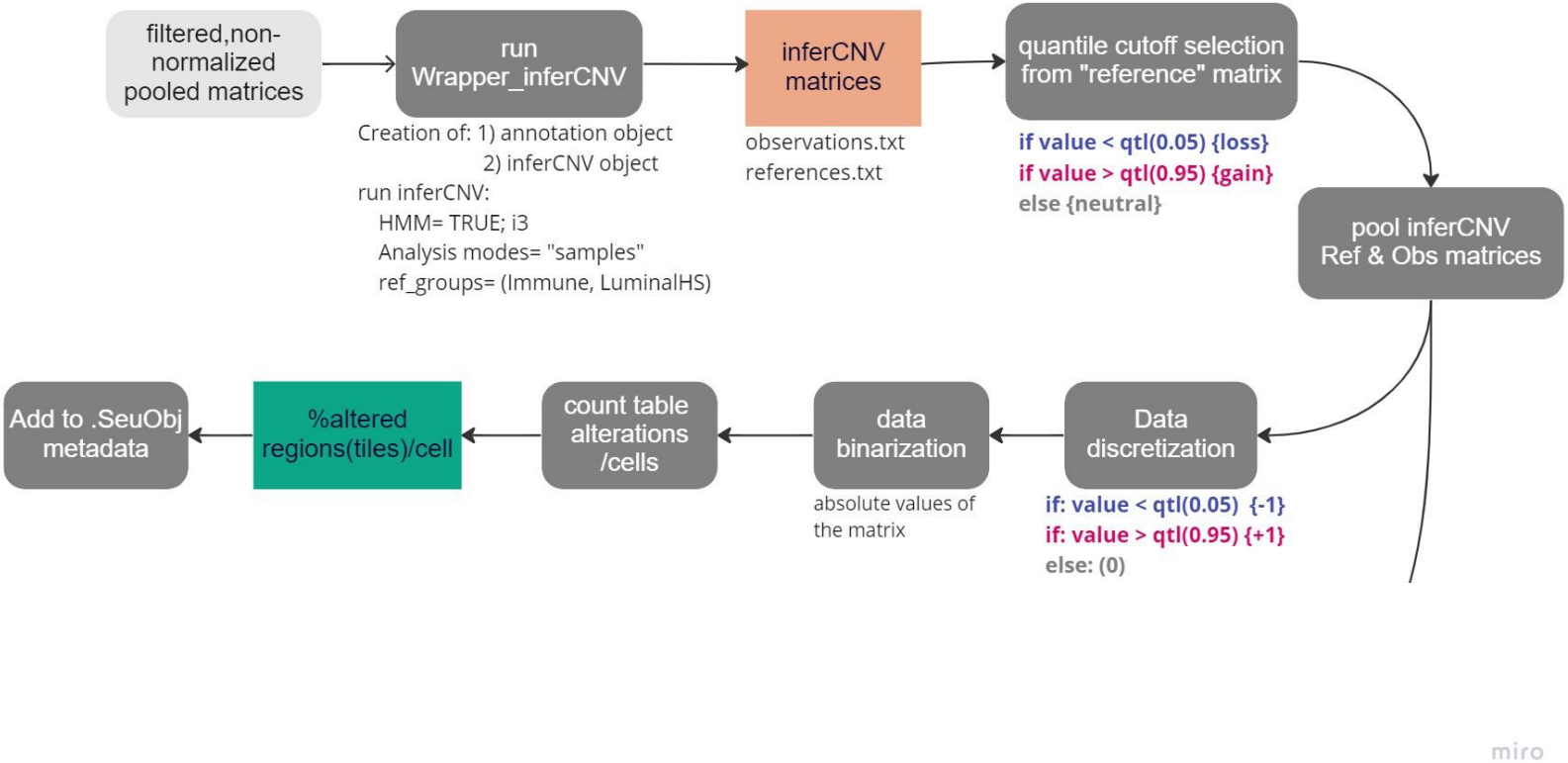
In both cell populations, Seurat V3 & Harmony integration methods gave the lowest numbers of differentially expressed genes between the cells from each sample

Strategy3: Identify a population of interest & check the specifically enriched pathways from its top genes

A population of interest in Juxta-tumor samples:

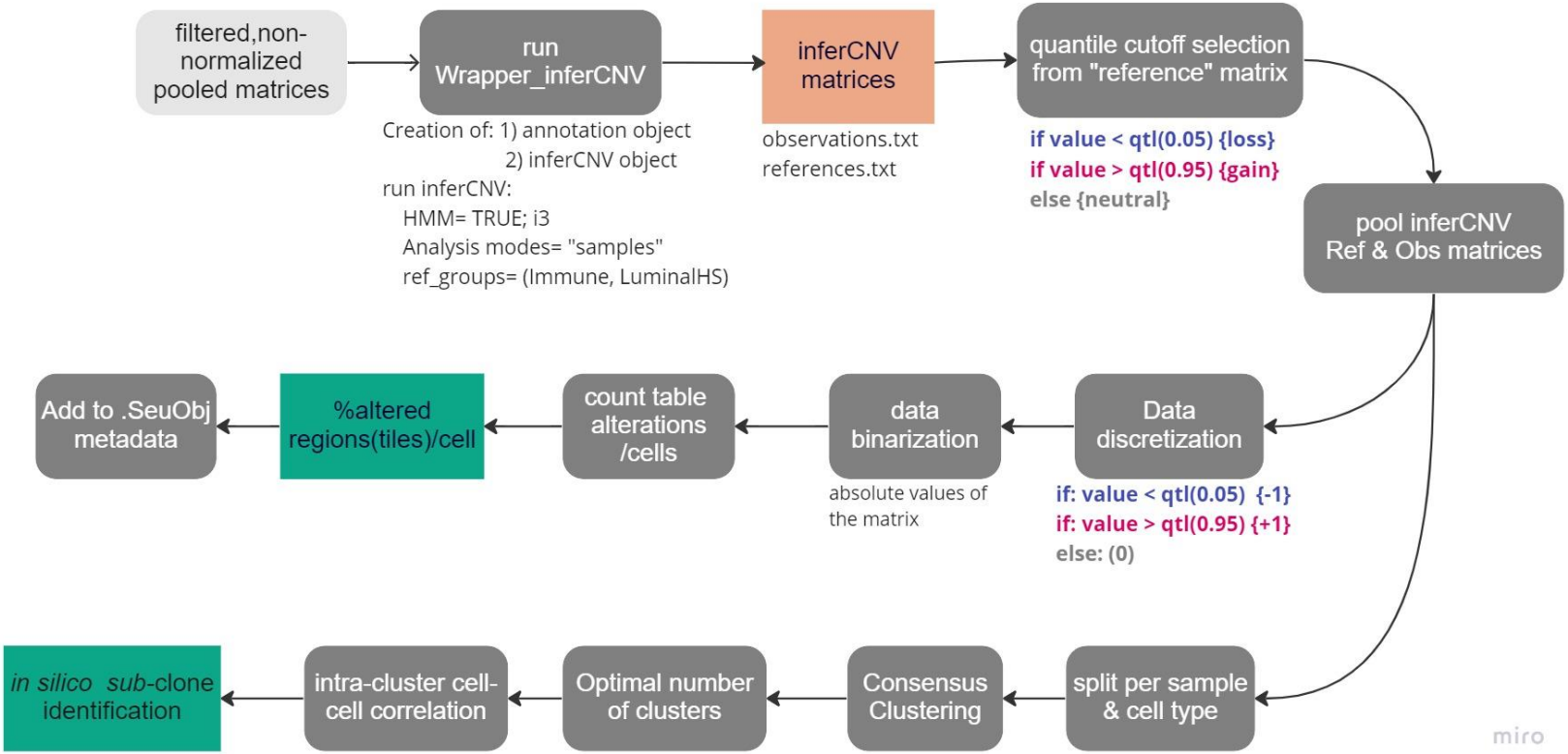
-harbors **abnormal CNV profiles** as compared to the remaining cells
-displays differentially enriched pathways

custom approach to estimate the fraction of altered genome -1-



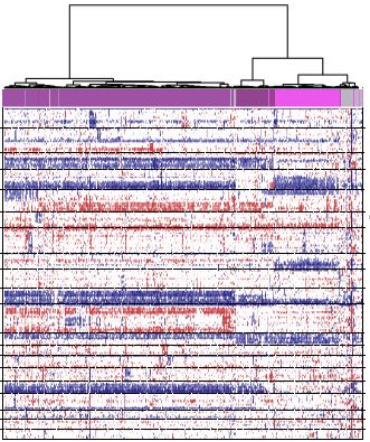
inferCNV: https://github.com/broadinstitute/inferCNV/wiki

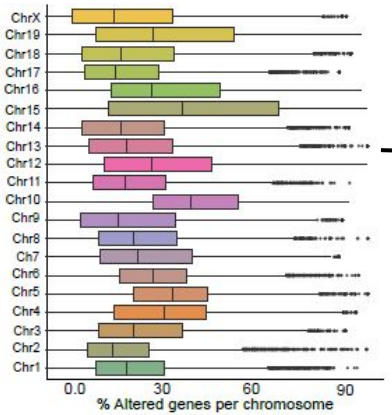
custom approach to estimate the fraction of altered genome -2-

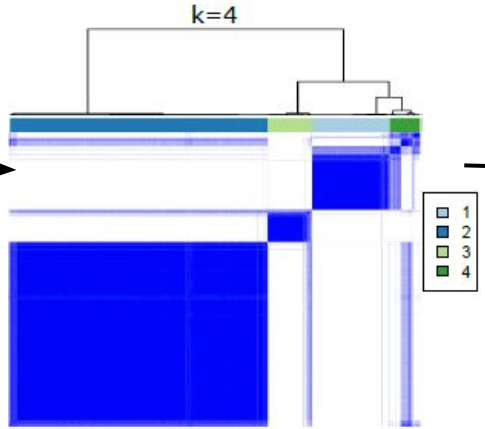


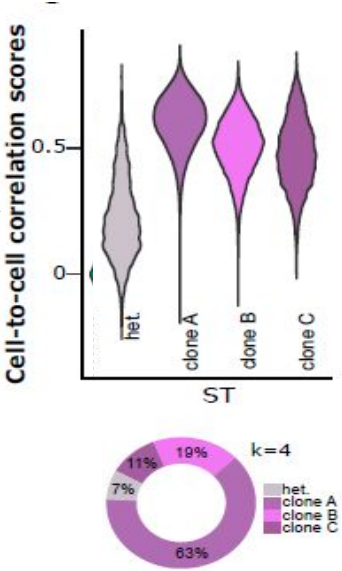
Example of identifying in silico sub-clones from inferCNV output matrix

Small tumor from a mouse mammary gland









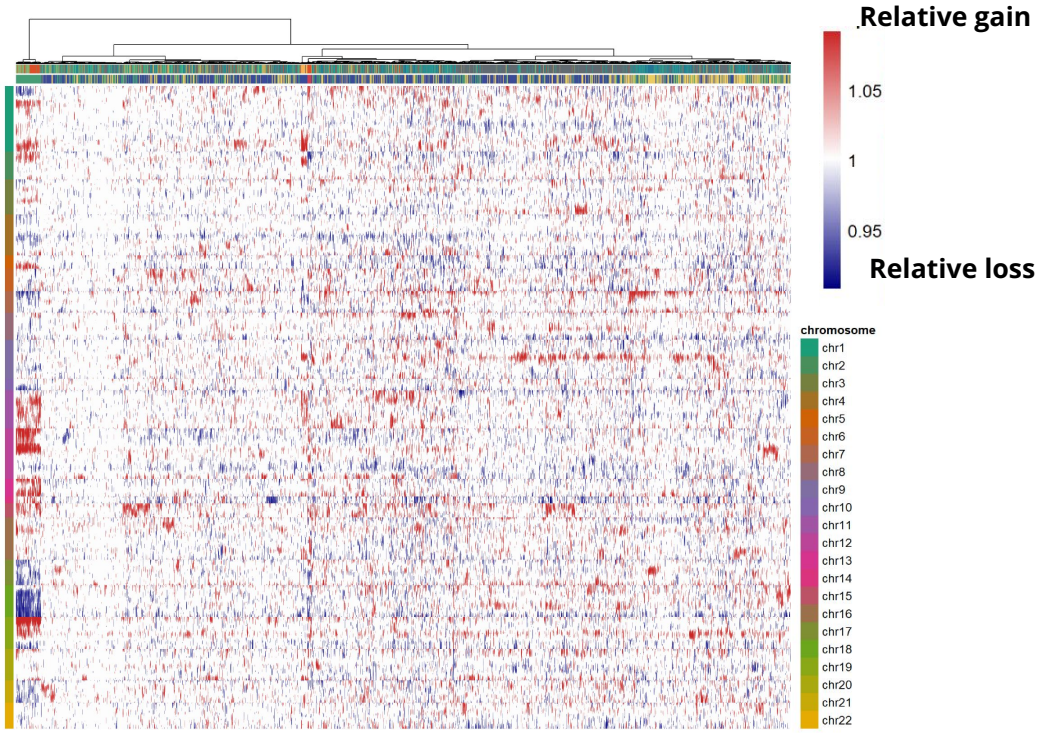
Strategy3: Identify a population of interest & check the specifically enriched pathways from its top genes

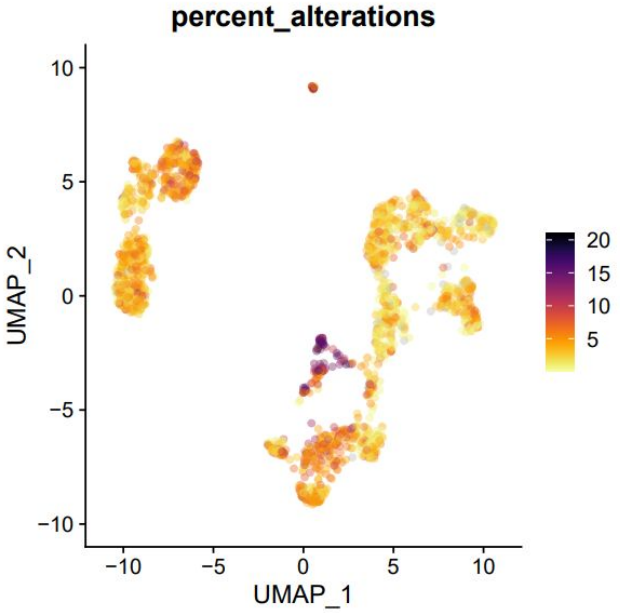
Procedure:

- ➔ infer CNV profiles from the scRNAseq epithelial NON-Integrated data
- ➔ Estimate percentage of alterations
- ➔ Identify the cell clusters corresponding to this population (if it exists)
- ➔ Run differential expression of this cell pop versus others
- ➔ Pathway enrichment analysis from the up-regulated genes of this cell pop in each integration method

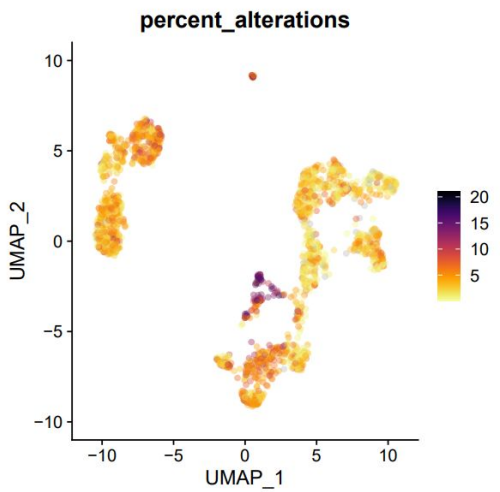
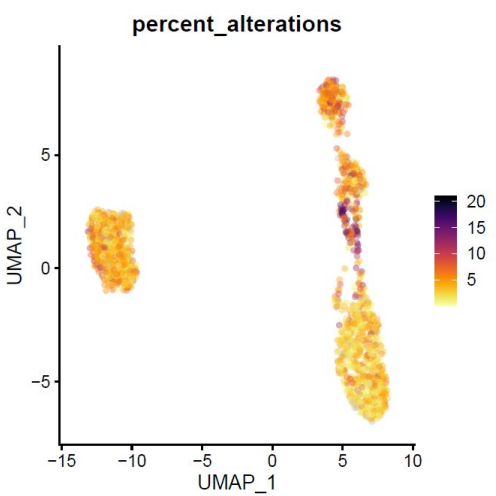
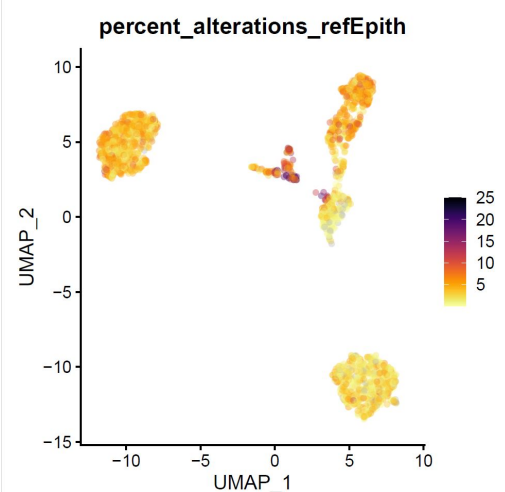
inferCNV highlighted a small subset of cells with relatively high alteration rate

inferCNV matrix of the epithelial cells



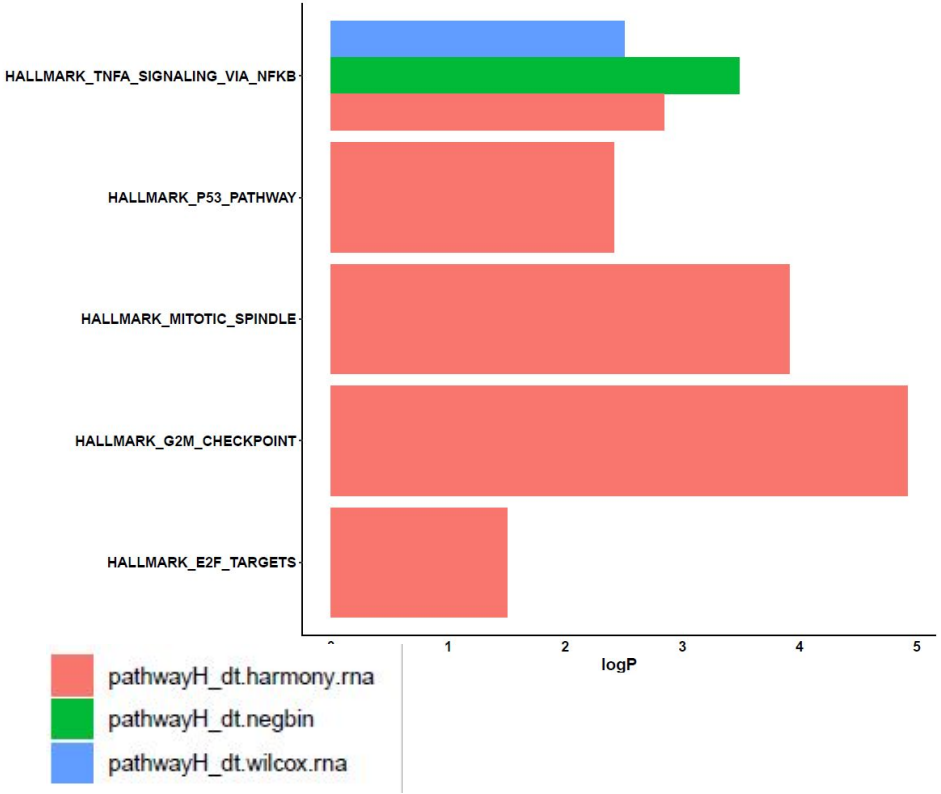


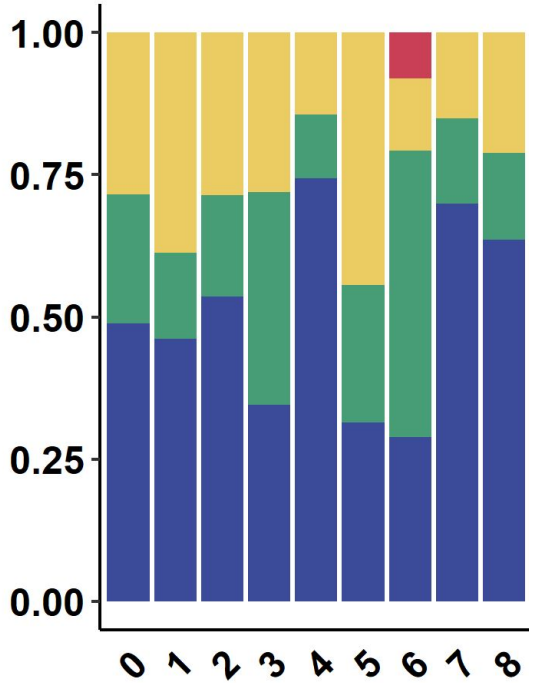
Visualization of the alteration rate on the embeddings from the integration methods

Not integrated	Seurat V3	Harmony
		
Clusters of interest: **12, 13**	Clusters of interest: **5**	Clusters of interest: **6**

=> Perform DEG (**Clusters of interest** vs **All** in each integration method)

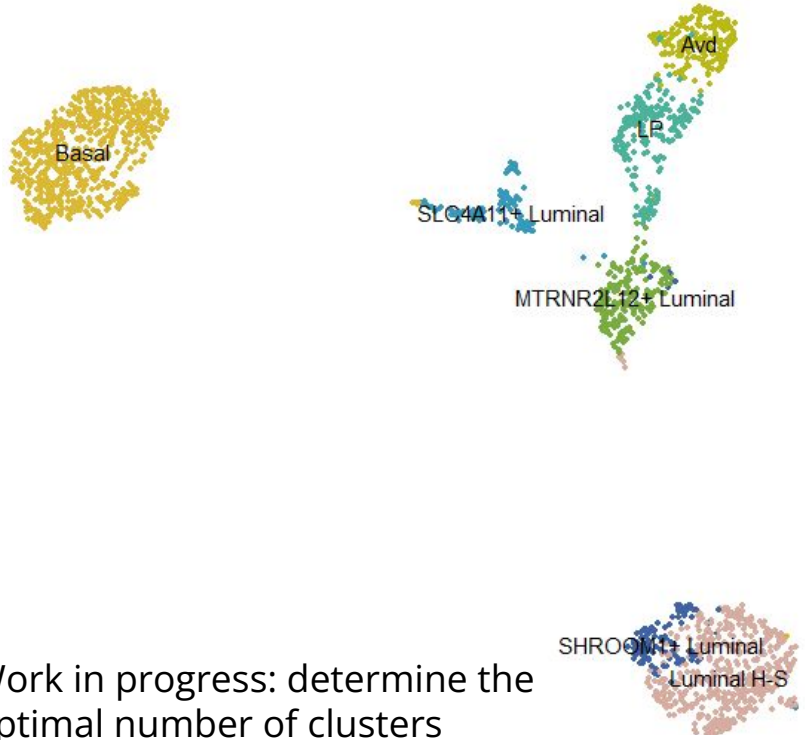
Integration with Harmony displayed the highest number of enriched pathways from overexpressed genes in the “cluster of interest”

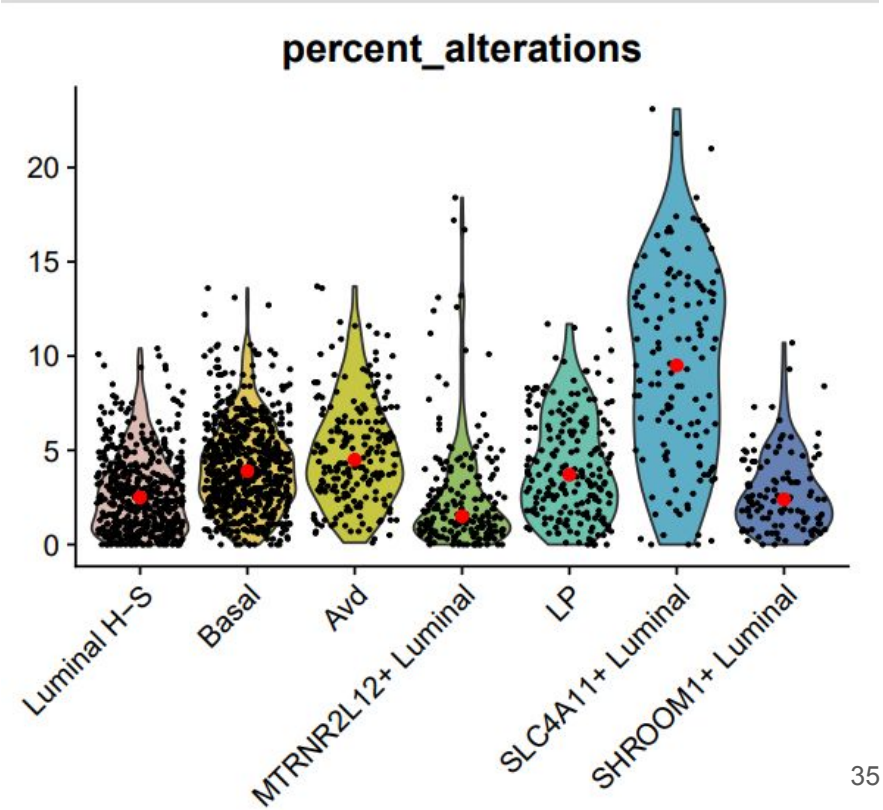




0 enriched pathways in Seurat V3

Cell cluster annotation on the Harmony embeddings





Work in progress: determine the optimal number of clusters

Conclusion & Perspectives

- ➔ We could justify the use of Harmony to process our dataset
- ➔ pre-tumoral cells could be identified using a customized analysis from the inferCNV object

- ➔ /!\ data integration is not a “systematic” approach : sometimes, a simple merge works!!
- ➔ Choice of performing or not data integration depends on:

  => the initial question: inter-patient heterogeneity vs atlas construction

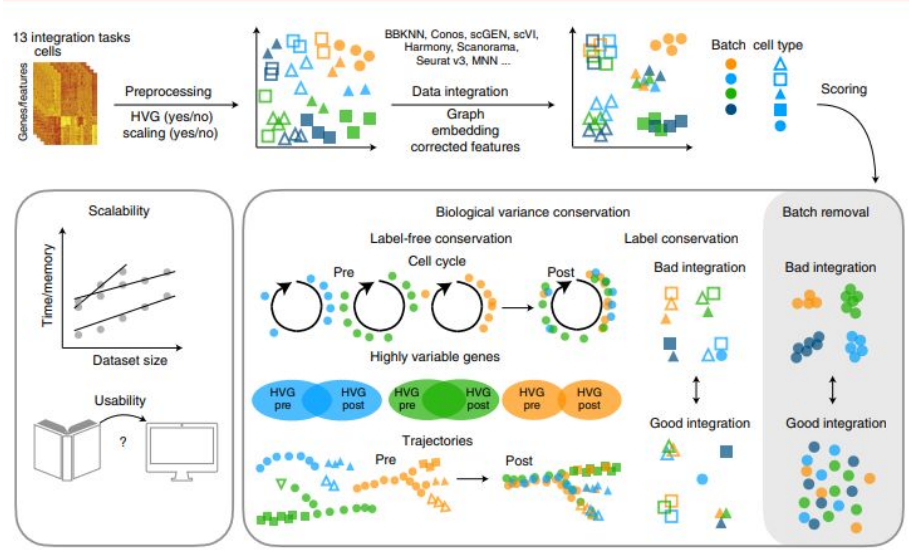
  => the dataset quality

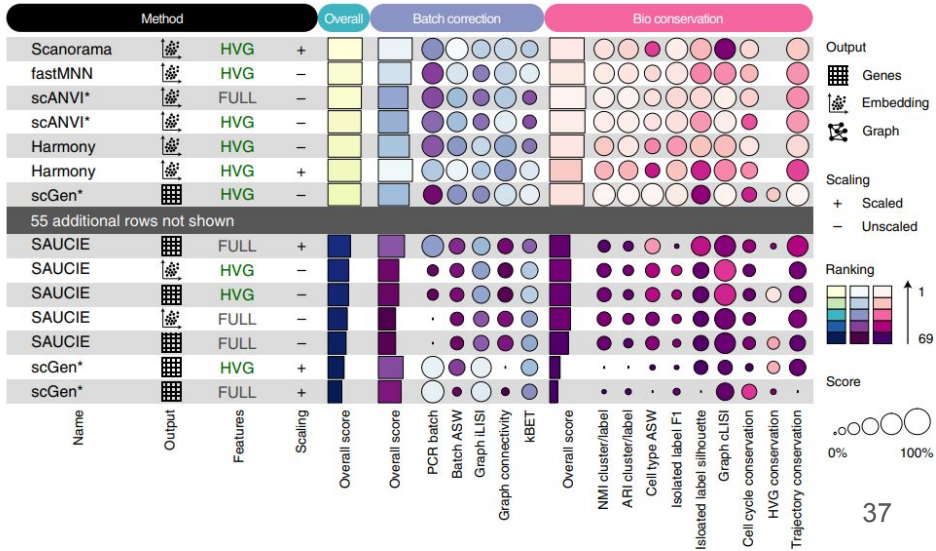
- ➔ Caution on integration methods which correct the count matrices /!\

A more comprehensive tool to evaluate data integration methods for atlas-level data

Benchmarking atlas-level data integration in single-cell genomics

Malte D. Luecken[1], M. Büttner[1], K. Chaichoompu[1], A. Danese[1], M. Interlandi[2], M. F. Mueller[1], D. C. Strobl[1], L. Zappia[1,3], M. Dugas[4], M. Colomé-Tatché[1,5,6] ✉ and Fabian J. Theis[1,3,5] ✉





Thank you !

Vallot Lab

Léa Baudre
Juliette Bertorello
Adeline Durand
Grégoire Jouault
Marthe Laisne
Yuna Landais
Justine Marsolier
Félix Raimundo
Melissa Saichi
Mathias Schwartz

Pacôme Prompsy
Camille Landragin

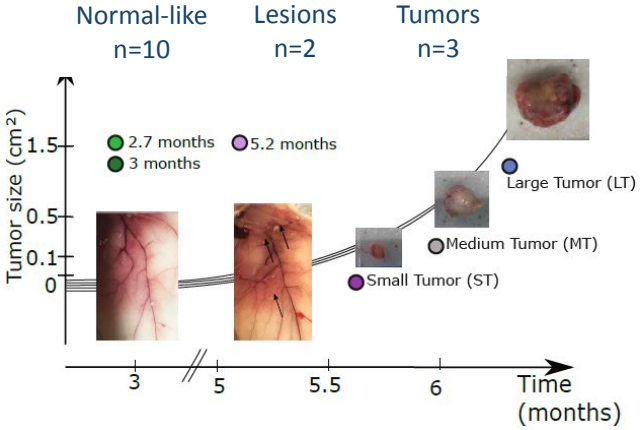
institut Curie

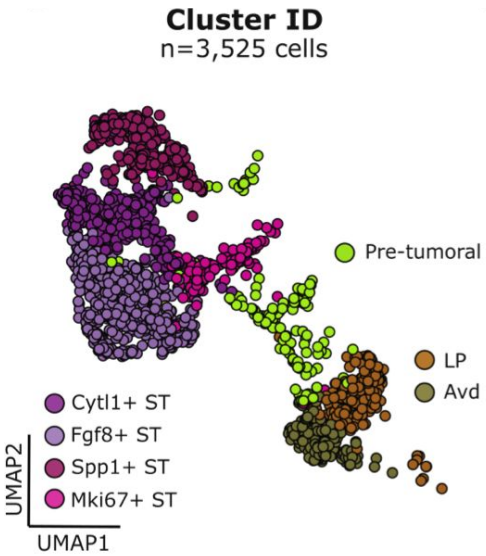
Fondation pour la Recherche Médicale

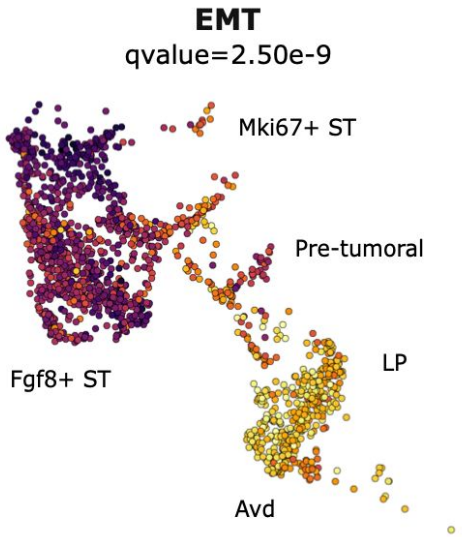
SORBONNE UNIVERSITÉ

Partial EMT occurs at the onset of basal-like breast tumorigenesis

Tp53 & *Brca1* deletion
In luminal progenitors







++ Vimentin, fibronectin,
- Cdh1, Cldn4/3/1
+ Twist1

Abbreviations:
LP: Luminal Progenitor
Avd: Alveolar differentiated cells
EMT: Epithelial to Mesenchymal transition