PCOM seminar

Slice, Dice, and Analyze

Spatial Transcriptomics

04/29/2025

Andrew Kossenkov

Regular expression data

Inception

Everything starts from observations:
something that you study is DIFFERENT - phenotype

phenotype

sample	**s1**	**s2**	**s3**	**s4**	**s5**	**s6**
group	g1	g1	g1	g2	g2	g2

Sample: cell line, tissue, patient, collected blood, picture, etc
Groups/phenotype of interest you observed and split into groups:

- Patients with different survival
- Cell lines that grow at different rates
- Tissues with different morphology

Or you did something that resulted in different behavior:

- Treatment
- Gene knockout

Metadata

You start to collect metadata:
something that you can measure/get for the samples

additional metadata

sample	**s1**	**s2**	**s3**	**s4**	**s5**	**s6**
group	g1	g1	g1	g2	g2	g2
age	25	28	35	42	66	55
gender	M	F	F	M	F	F

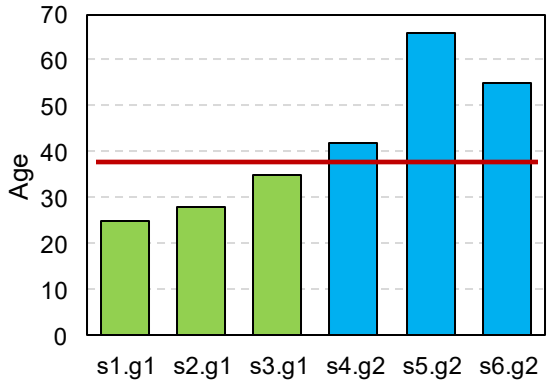
Types of metadata :

- Any clinical characteristics for patients
- Cell line grown rate, invasion, treatment response
- Tissue purity, morphological charachteristics
- Treatment dose, length, response

Metadata can on its own be a subject of analysis

to associate groups (phenotype) with this data (variables)

Answer in the example above can be: groups are associated with age! →

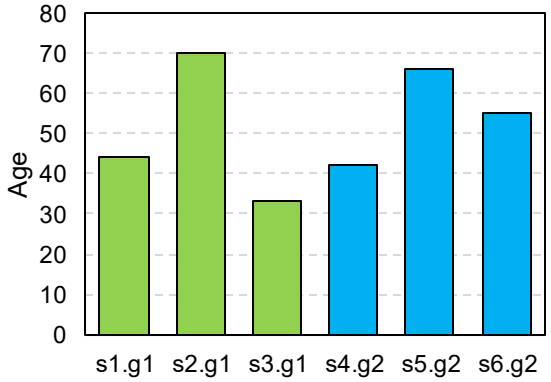


Transcriptomics: single gene

Trascriptomics (or any –omics: proteomics, metabolomics, lipidomics, etc): You can actively start measuring expression of a gene of interest across samples

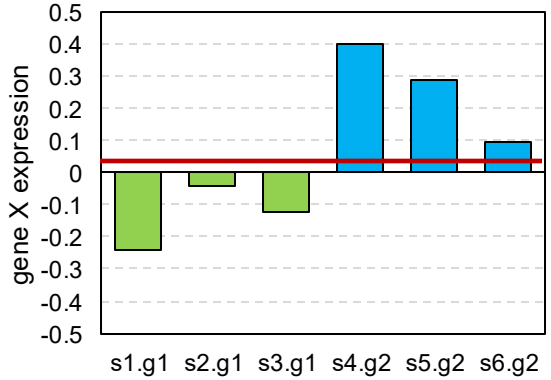
metadata

sample	s1	s2	s3	s4	s5	s6
group	g1	g1	g1	g2	g2	g2
age	44	70	33	42	66	55
gender	M	F	F	M	F	F



expression data

	s1	s2	s3	s4	s5	s6
gene X	-0.24	-0.04	-0.12	0.40	0.29	0.09



p-value (significance)

fold change (magnitude of effect)

Just one gene at a time. Need a hunch/hypothesis from something that it might be important for what you observe. Frequently you don't have it, or it does not explain the groups you observe – you need to expand your search

Transcriptomics: high-throughput

High-throughput – measure a lot of genes at the same time
Total RNA-seq (all genes in a cell) or using pre-defined libraries (only genes or types of genes that you are interested in) will measure expression of bunch of genes

metadata

sample	s1	s2	s3	s4	s5	s6
group	g1	g1	g1	g2	g2	g2
age	44	70	33	42	66	55
gender	M	F	F	M	F	F

expression matrix

measurement of features across sample

	s1	s2	s3	s4	s5	s6
gene1	-0.41	-0.24	0.36	-0.09	0.38	-0.01
.	0.54	-0.30	0.25	0.04	-0.19	-0.33
.	-0.03	0.48	0.44	-0.33	-0.23	-0.34
.	-0.08	-0.14	-0.26	0.58	0.20	-0.30
.	-0.35	-0.32	0.18	0.23	-0.27	0.53
.	0.30	-0.03	-0.35	-0.31	0.06	0.32
.	0.14	0.37	-0.16	-0.51	0.34	-0.19
.	-0.26	-0.22	0.15	-0.15	0.35	0.13
.	-0.09	-0.12	-0.08	0.48	0.23	-0.41
.	0.28	-0.07	-0.31	-0.12	0.21	0.01
.	-0.25	0.11	0.29	0.20	-0.13	-0.22
.	-0.12	0.19	-0.02	-0.12	-0.09	0.16
.	-0.10	-0.35	0.24	0.03	0.09	0.09
.	-0.25	0.13	0.33	0.40	-0.30	-0.30
.	0.21	0.33	-0.46	-0.40	0.04	0.28
.	-0.12	-0.06	-0.28	0.42	-0.31	0.36
.	-0.41	-0.57	0.31	0.40	-0.10	0.36
.	0.25	0.21	-0.36	-0.20	-0.19	0.28
.	0.34	-0.49	-0.10	-0.23	0.15	0.33
.	0.14	0.27	0.33	-0.24	-0.18	-0.33
geneN	-0.09	-0.03	0.21	0.56	-0.35	-0.29

testing

pv	fold
0.533	1.1
0.293	-1.3
0.023	-1.5
0.284	1.2
0.319	1.3
0.858	1
0.467	-1.2
0.322	1.2
0.51	1.1
0.756	1
0.656	-1.1
0.821	-1
0.449	1.1
0.657	-1.1
0.873	-1
0.275	1.2
0.232	1.4
0.798	-1
0.601	1.1
0.002	-1.4
0.876	-1

feature info

	id	symbol	other id
gene1	ENSGxxx	SEP5	na
.			
.			
.			
.			
.			HGNC:5
.			
.			
.			
.			
.			
.			
.			
.			
.			
.			na
.			
.			
.			na
.			
geneN	ENSGxxx	TP53	na

Biological answers:

- Individual significant genes
- Common functions
- Enriched pathways
- Regulation networks
- Hypotheses to test

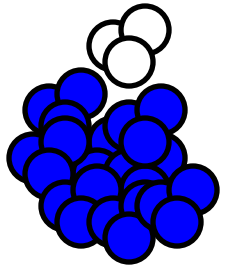
Feasible, all genes, so no hypotheses needed – can be a fishing expedition. Get a snapshot of your sample
Buuuuuut……

Bulk sample, bulk RNA-seq

If not a pure cell line, sample usually is a mix of different cells

Cells on interest (gene X expressed at level 100)
Impurities: some other cells (gene X expressed at level 10)

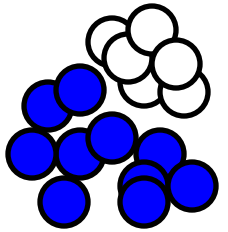
Sample A, group 1



23 our cells
3 “bad” cells

Average expression:
gene A = (23*100+3*10) / 26 = **89.6 per cell**

Sample B, group 2



10 our cells
6 “bad” cells

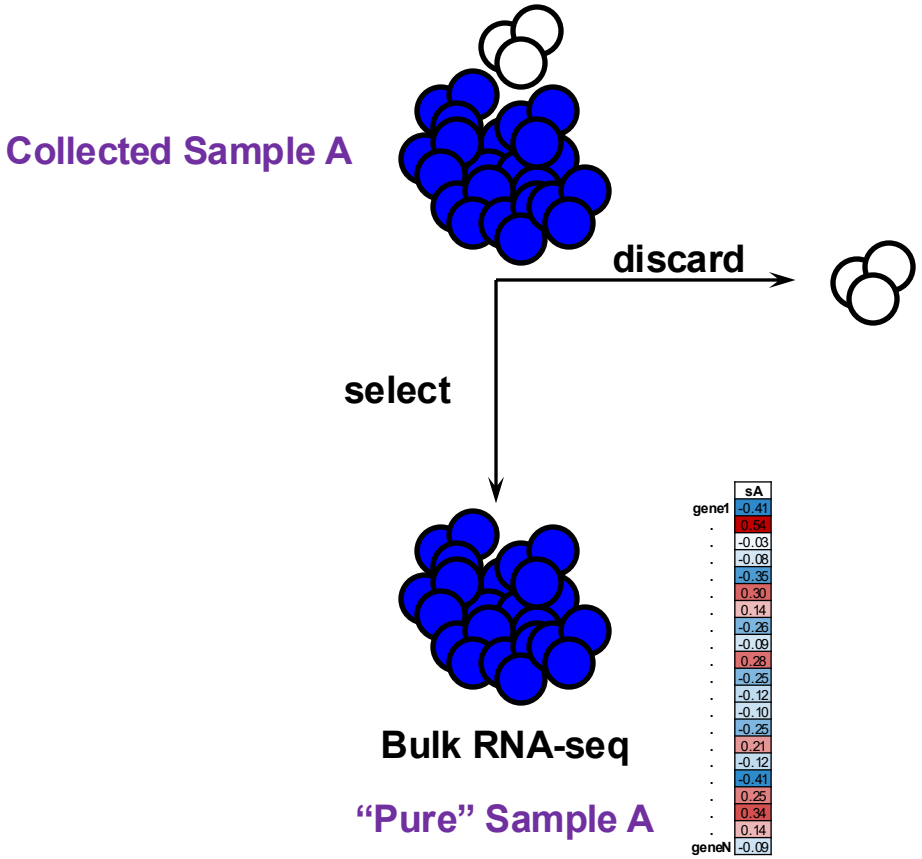
Average expression:
gene A = (10*100+6*10) / 16 = **66.2 per cell**

Difference of **1.4 fold** between measured expression (Sample B / Sample A)
Has nothing to do with actual expression differences (still same 100 or 10 in corresponding cells)
But differences between measured mix/bulk of cell composition resulted in the difference

To the rescue!

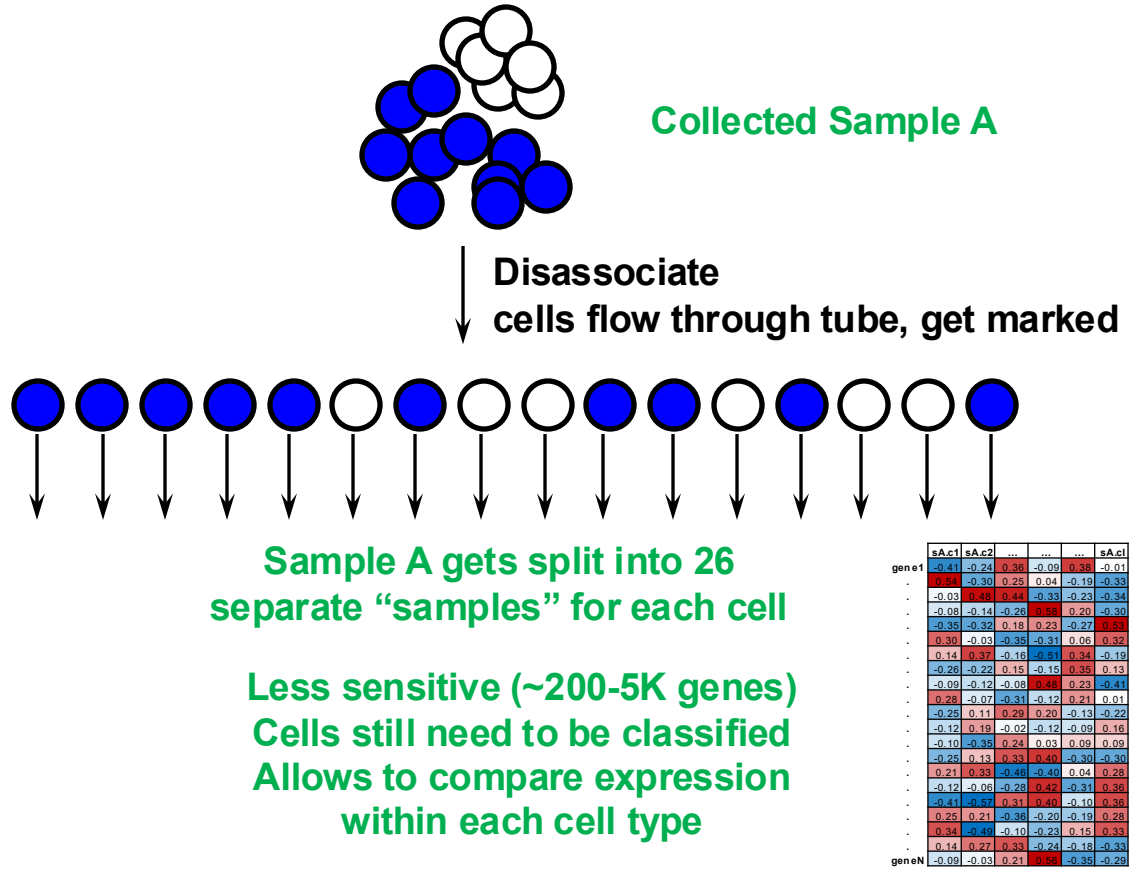
Ways to address: cell sorting or single cell RNA-serq

Sorting (FACS), enriching (pulldown)
Need markers of cells of interest



Need to know what cell you want
Cells should be in a state to be able to be selected
Still not 100% pure, the process creates artifacts

Single cell RNA-seq
Cells are separated, measured on its own



Sample A gets split into 26 separate "samples" for each cell

Less sensitive (~200-5K genes)
Cells still need to be classified
Allows to compare expression within each cell type

In the end – we know what cells are in the sample, their proportion, expression within each but not which ones were close to another (important to tissues)

Single cell

We have measurements of features across samples
Some phenotype to link the expression to

[phenotype for comparisons]

metadata

sample	s1	.	.	.	.	sK
group	g1	g1	g2	g2	g2	g2
quality	good	good	good	bad	good	good
metric	1	2	3	4	5	6

expression matrix

measurement of features across sample

	s1	.	.	.	.	sK
gene1	0.02	0.00	-0.33	0.37	0.27	-0.33
.	-0.30	0.23	-0.21	0.13	0.06	0.09
.	0.32	-0.18	-0.53	0.28	0.34	-0.23
.	-0.07	0.23	0.06	0.33	-0.40	-0.16
.	-0.02	-0.30	0.27	0.39	-0.13	-0.21
.	-0.25	0.22	0.06	0.19	0.11	-0.33
.	-0.02	-0.19	-0.18	0.36	-0.18	0.21
.	0.19	-0.45	-0.05	0.06	-0.15	0.40
.	0.28	-0.07	0.13	-0.20	-0.25	0.11
.	0.30	0.50	-0.13	-0.05	-0.25	-0.37
.	-0.53	0.20	-0.34	0.34	0.29	0.04
.	-0.22	0.07	-0.09	0.06	0.21	-0.03
.	0.56	-0.29	-0.41	-0.26	0.57	-0.17
.	0.43	-0.40	-0.01	-0.34	0.34	-0.03
.	0.08	0.16	-0.26	0.27	0.05	-0.29
.	-0.04	-0.35	-0.22	-0.01	0.09	0.54
.	0.10	-0.18	-0.01	0.11	-0.19	0.18
.	0.03	-0.13	0.12	0.21	0.07	-0.28
.	-0.37	0.13	0.33	-0.28	-0.28	0.46
.	0.29	0.27	0.32	-0.44	-0.27	-0.18
geneN	0.36	-0.03	-0.03	-0.17	-0.41	0.28

[QC/statistics/patterns]

[Enrichments, biology]

feature info

	id	symbol	other id
gene1	ENSGxxx	SEP15	na
.			
.			
.			
.			
.			na
.			
.			
.			
.			
.			
.			
.			
.			
.			
.			
.			
.			na
.			
.			
.			na
.			
geneN	ENSGxxx	TP53	na

QC/Normalization/artifacts/outlier
Stat comparisons/bio context

Advantage (generally):
- **Robust measurements**
- **Selection of assay type**

Missing (generally):
- **unknown mix of cells**
- **sample purity**
- **proportion of even known types**
- **morphology – everything is diced**

Tissues: Spatial transcriptomics

Starts with tissue: normal with different morphology, or tumor with different characteristics, like position within the tissue or origin, surrounding blood cells, etc

Really-really want to know besides expression:

- **Region/morphology where measured expression comes from**
- **what cells are close to each other to see if one type affects expression of another**
- **Best if it is on a single cell level, but can be "small bulk"**
- **Great if can select regions of interest from the tissue**

Need to look at tissue beforehand
Guide the regions of interest selection
Quantify expression of genes
In the end we still get gene expression matrix and all associated analyses
But metadata has now for "samples":

- coordinates (X/Y)
- possibly region it comes from (morphology)
- surrounding information (infiltration)

metadata

sample	s1	s2	s3	s4	s5	s6
group	g1	g1	g1	g2	g2	g2
X	1	2	3	4	5	6
Y	6	5	5	4	3	1
morph	ear	nose	eye	ear	nose	nose
blood	lots	abit	none	lots	none	none

expression matrix

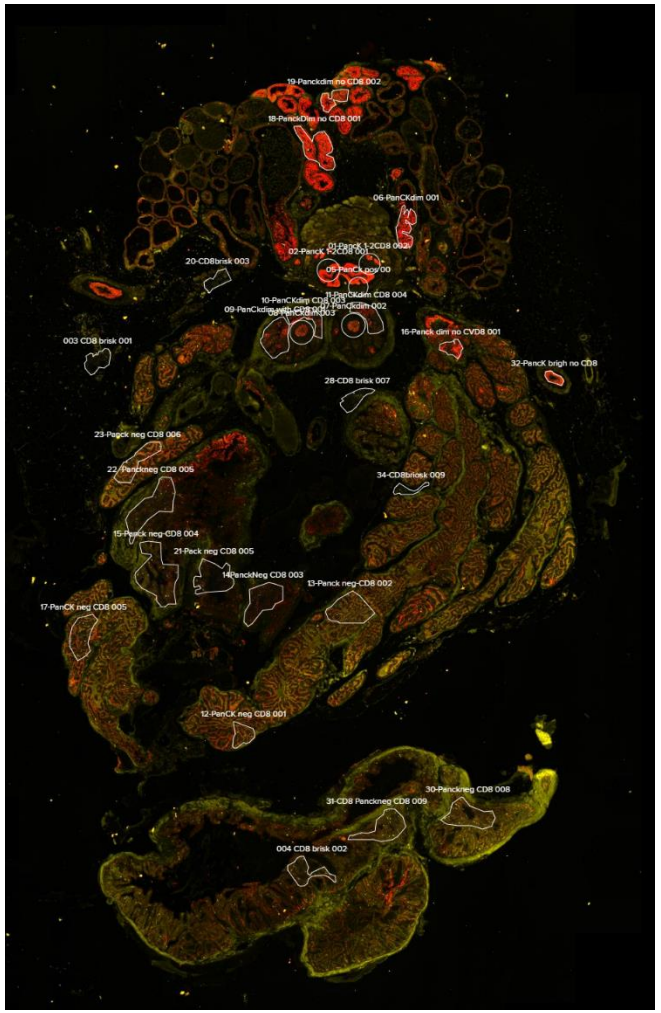
measurement of features across sample

	s1	s2	s3	s4	s5	s6
gene1	-0.41	-0.24	0.36	-0.09	0.38	-0.01
.	0.54	-0.30	0.25	0.04	-0.19	-0.33
.	-0.03	0.48	0.44	-0.33	-0.23	-0.34
.	-0.08	-0.14	-0.26	0.58	0.20	-0.30
.	-0.35	-0.32	0.18	0.23	-0.27	0.53
geneN	-0.09	-0.03	0.21	0.56	-0.35	-0.29

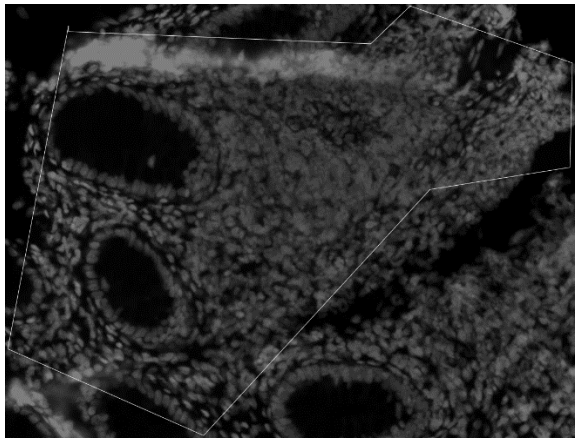
Spatial
(tissue)

Tissue slice required
Colored by something

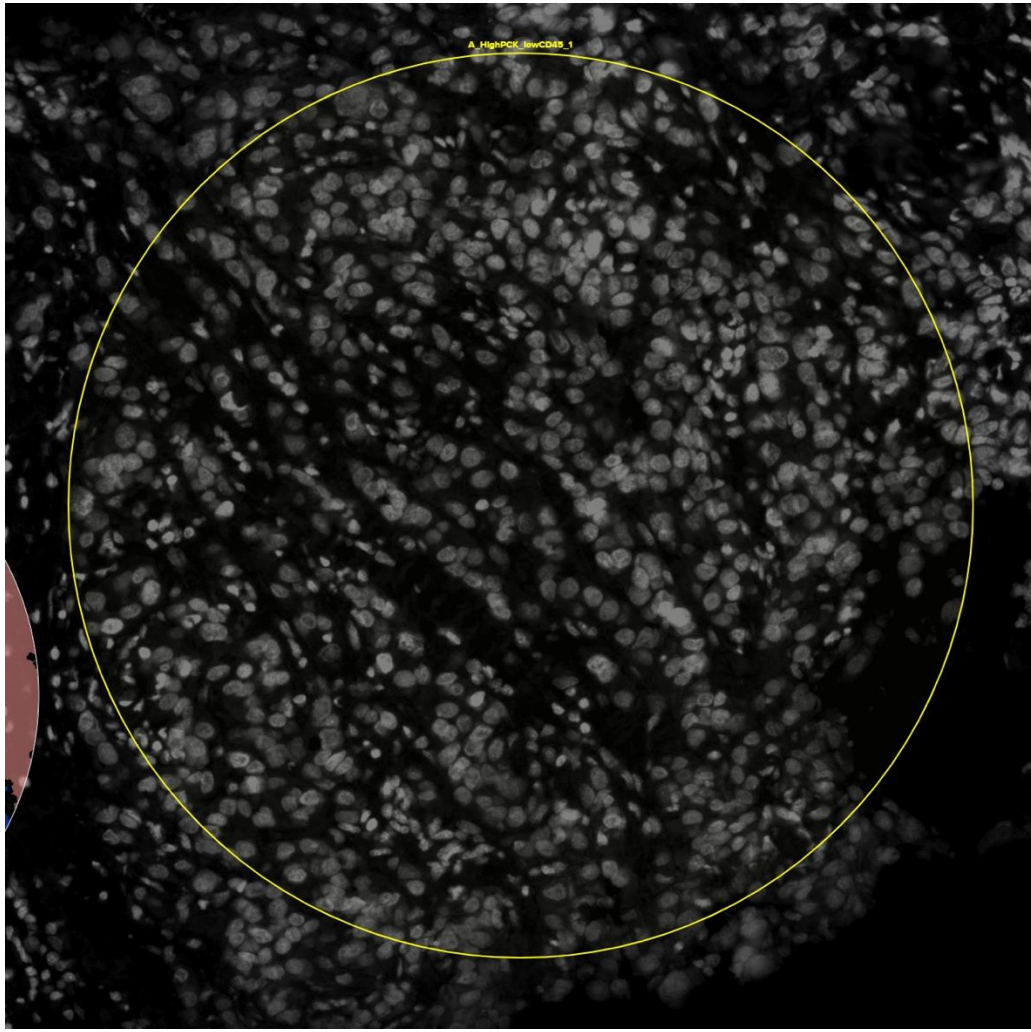
Slice of tissue
Stained for
Cancer Cd8 T cells



Morphology
Color = DNA



Single cells visible



Information source: Non-spatial vs spatial

Spatial context:

- Guided by microscopy selection of samples/regions/cells
- And/or post-measurement proximity/morphology metrics for each region/cell

Consider:

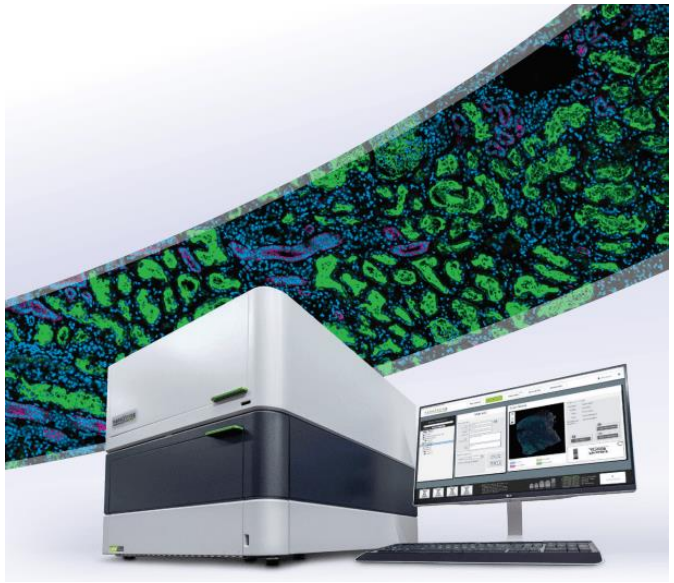
- Bulk: good old matrix
  - Pros: signal-to-noise, sensitivity, de-novo, total RNA pool
  - Cons: mix of cells, therefore signal
- Flow: cells of pure known type
  - Pros: signal-to-noise, sensitivity, de-novo, total RNA pool
  - Cons: hard to isolate, artifacts of isolation, purity, only known types
- Single Cell: de-novo sub-population of cells, cell states
  - Pros: unbiased, single cell resolution, denovo
  - Cons: Sensitivity, don't know only one tiny thing - what cells are nearby

In the end we have same matrix of feature expression across regions/cells, but:

- Better selection, purity estimation or single cell level
- Spatial information about "samples" (basically X/Y coordinates for each region/cell)
- Most of the time some tissue image to go back to for some additional context

Wistar-special spatial

GeoMx



Expression assay: barcode sequencing

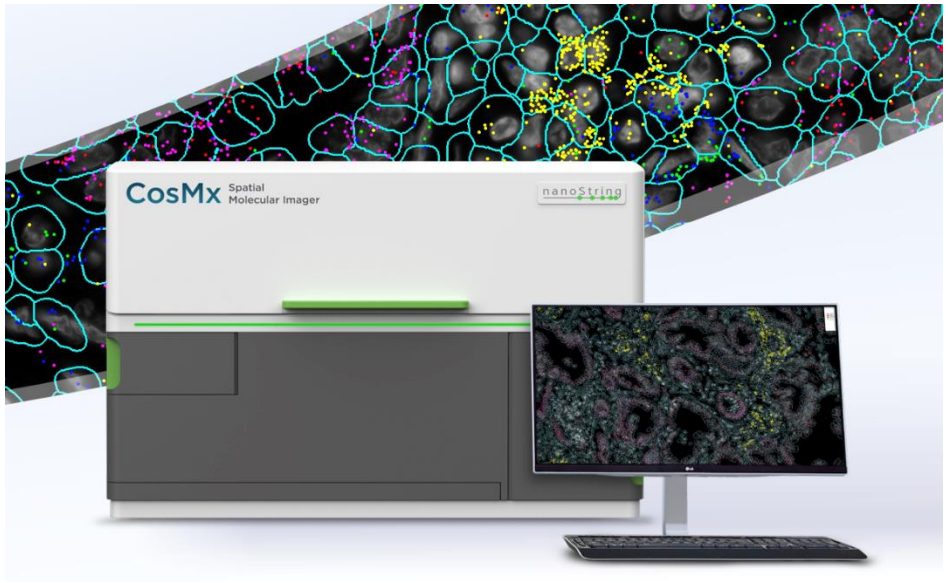
bulk-ish (50-100 cells). 20k genes

Both

Measure:
mRNAs
proteins

Feature scope:
predefined library

CosMx



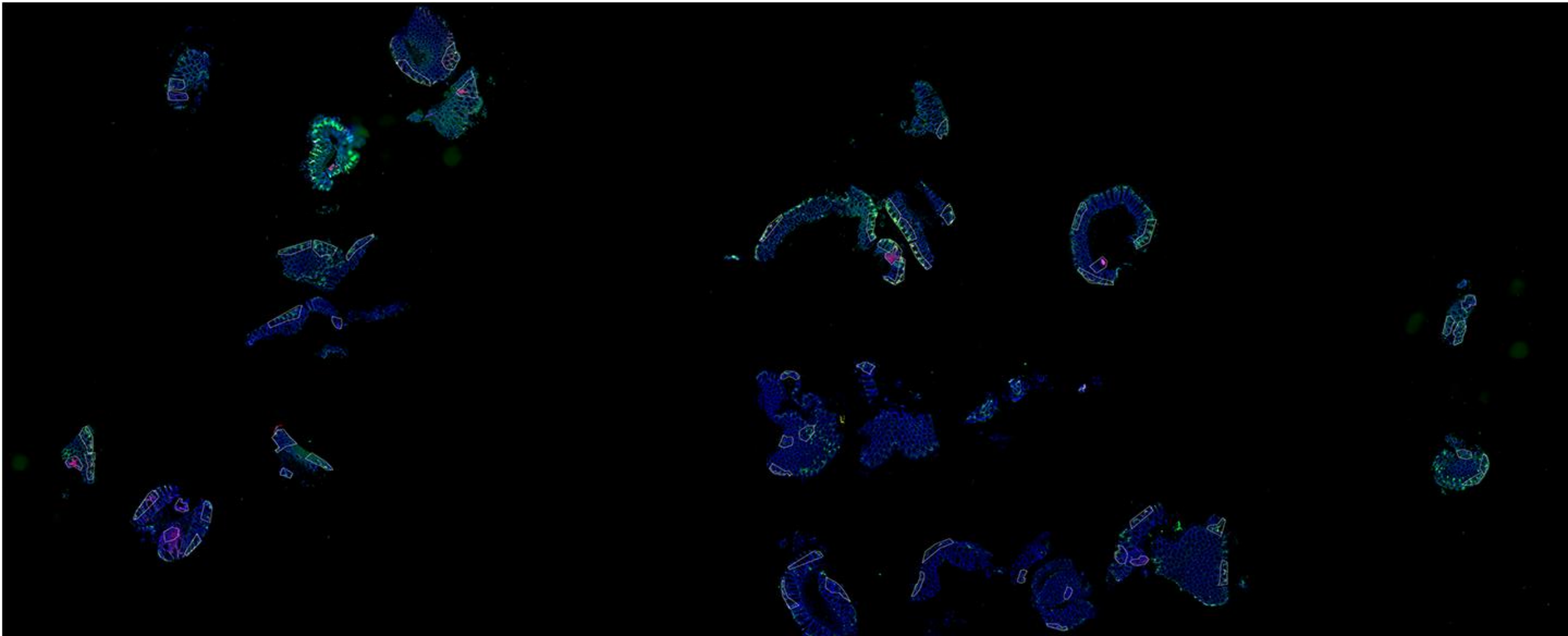
Expression assay: pure microscopy

single cell. 1k/4k/6k/[announced 18k] genes

GeoMx
(spatial, microscopy guided mini-bulk)

GeoMx ROI

Slide you can see with your eyes before selection
59 ROI (regions of interest) you define: here, 20 samples from 1 to 5 ROIs per sample
Size on a computer: ~10GB slide
DNA/CD68/CD45 staining

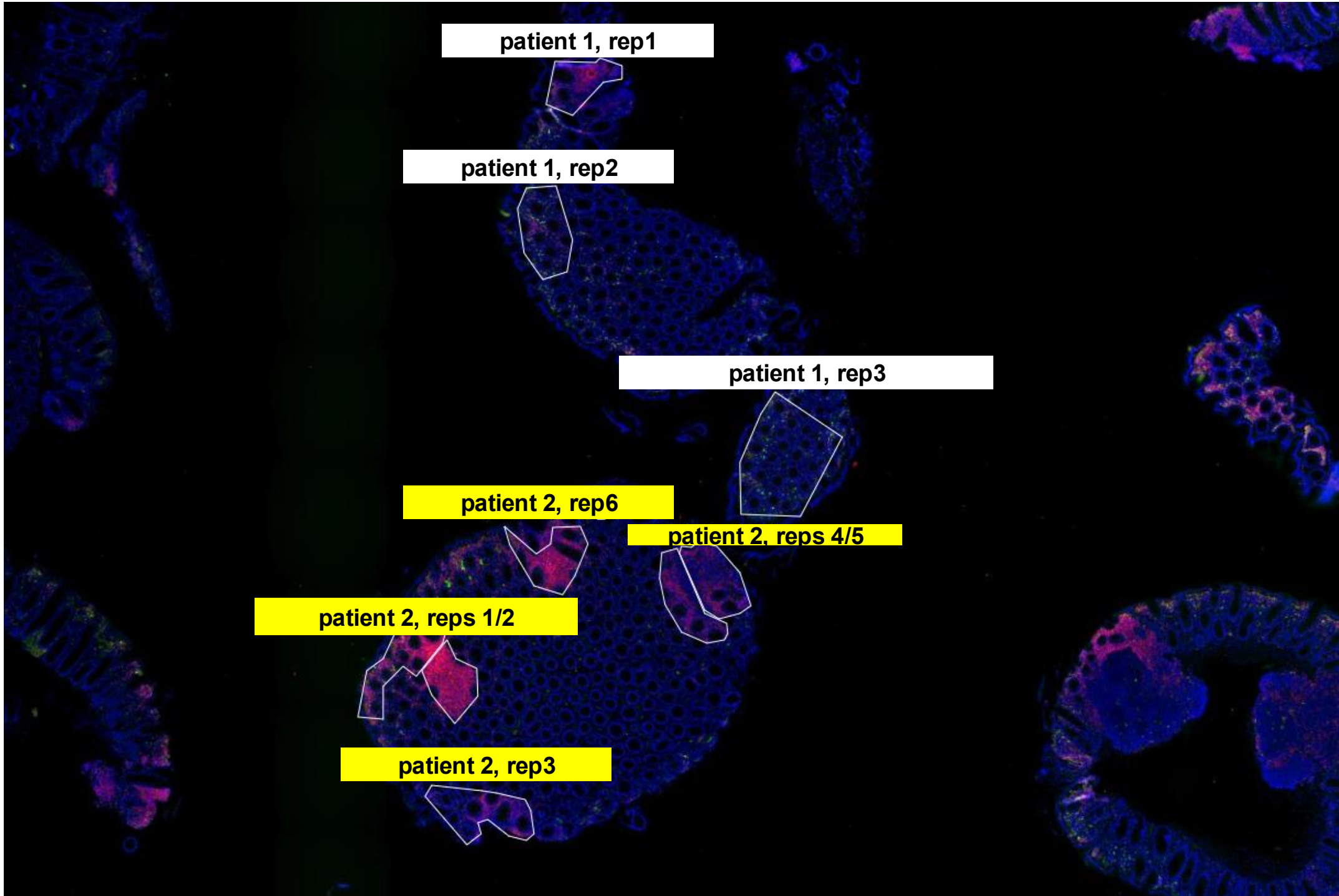


GeoMx ROI

Most important step (and time consuming) Defining ROIs

This fragment
2 stains + DNA
2 samples/patients
9 ROIs

Every ROI (95 total for the project is done manually)



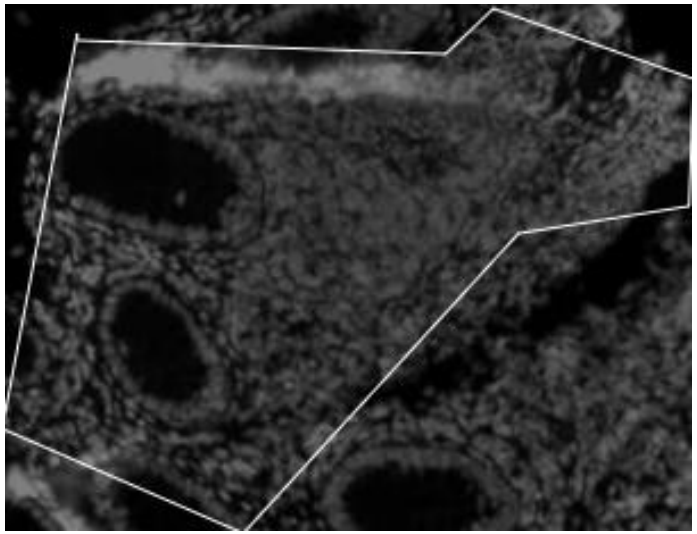
GeoMx ROI: focus on one ROI

Rectal sample (no tumor)

DNA/CD68/CD45 staining

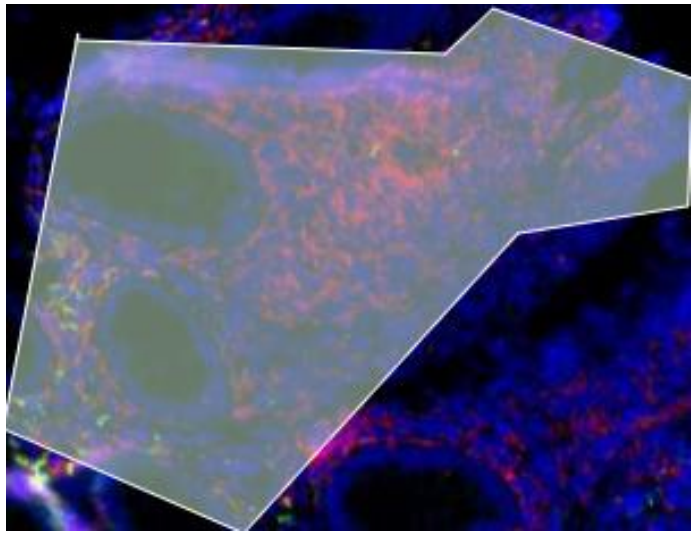
You can see rectal tissue ducts (affects cell numbers and perception of intensity by eye during selectin)

DNA

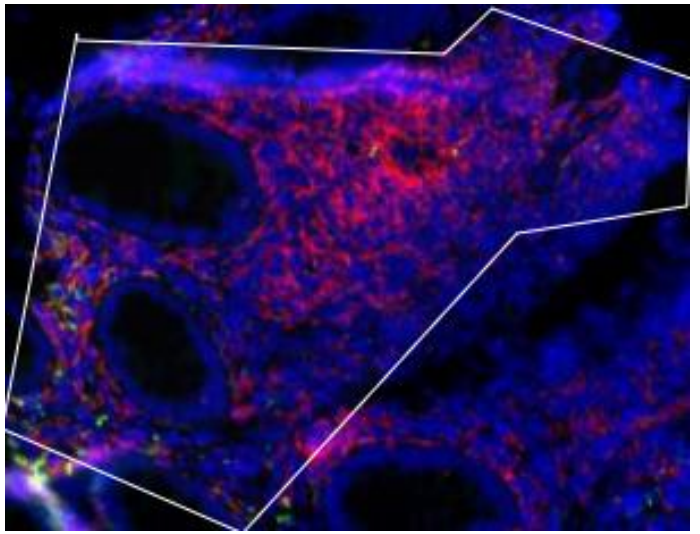


DNA + CD68 + CD45

with overlay



without overlay

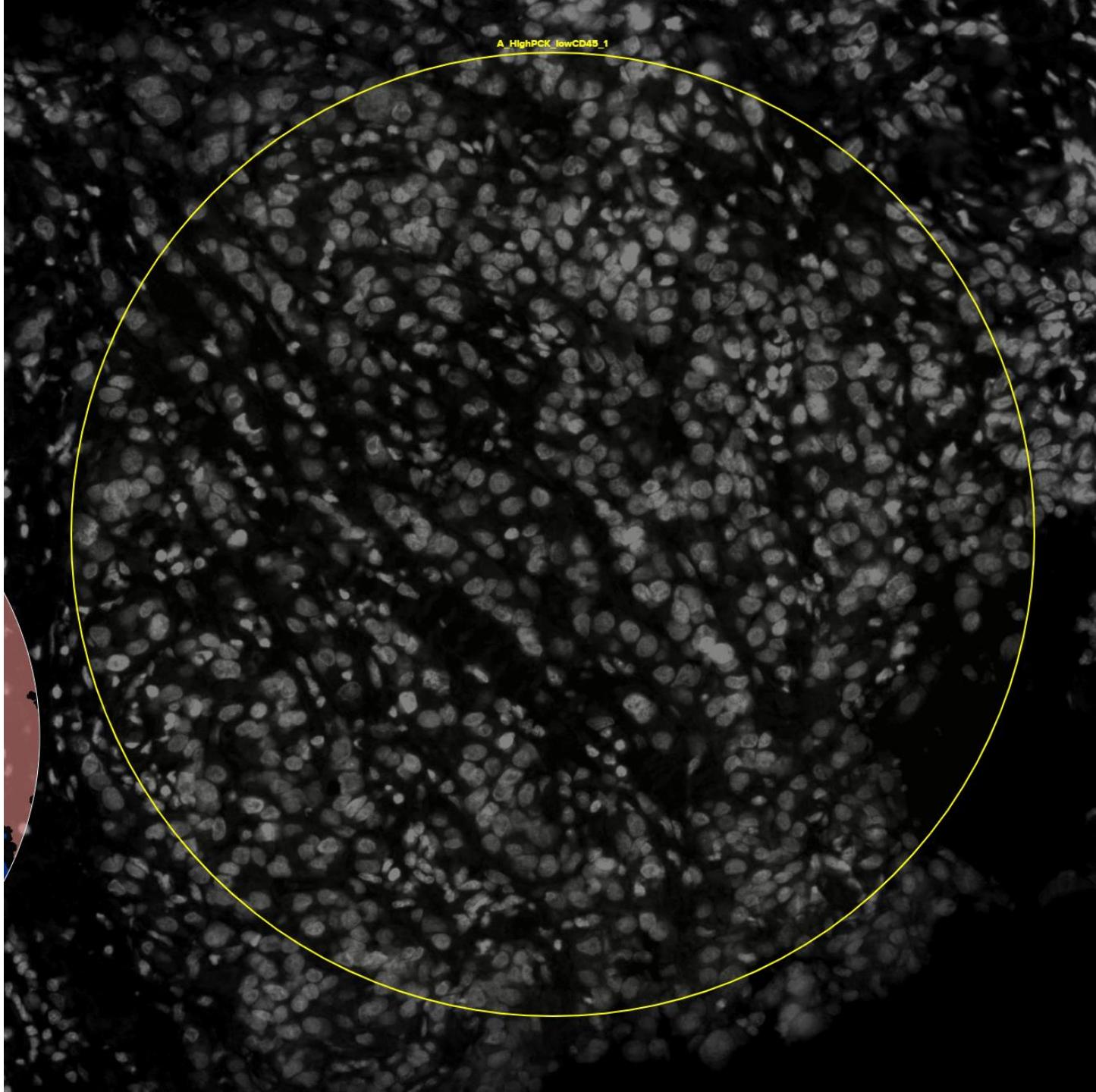


GeoMx: segmentation

Simple circle for ROI
DNA channel – cells visible
1267 cells

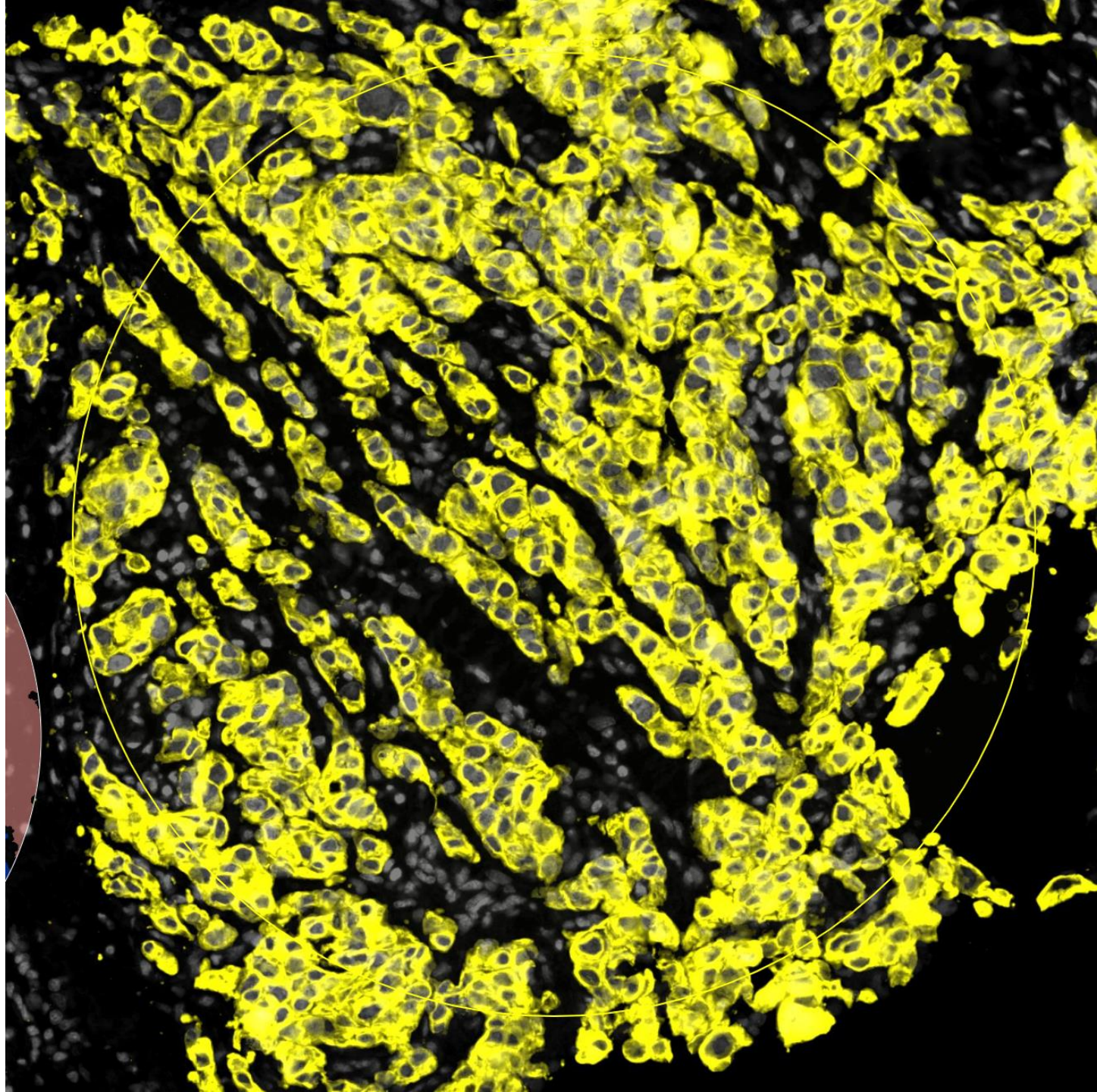
PanCK and CD45 channels
Defining segments on PanCK

Deemed by eye
as high PCK/low CD45



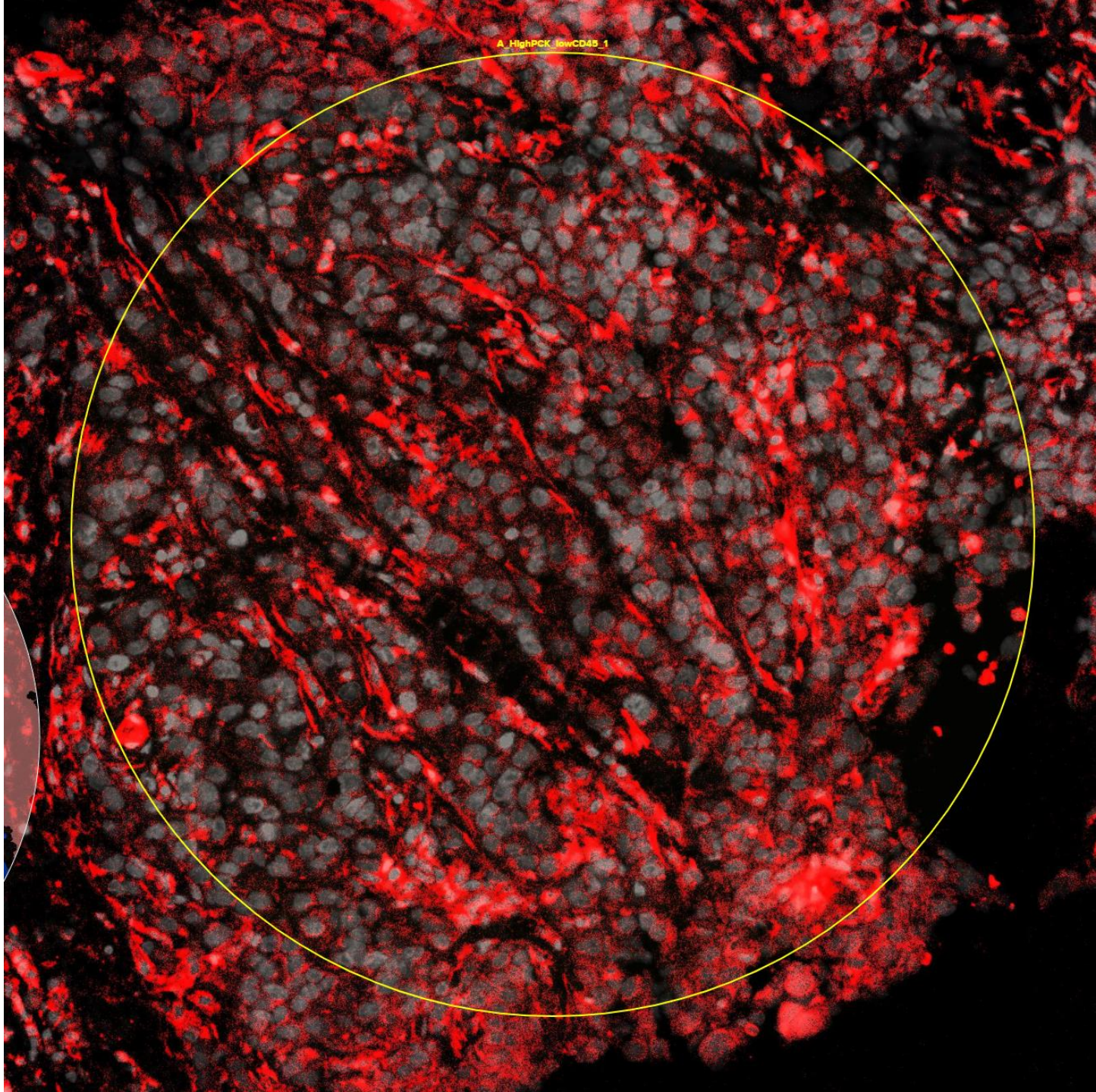
GeoMx: segmentation

DNA + PanCK



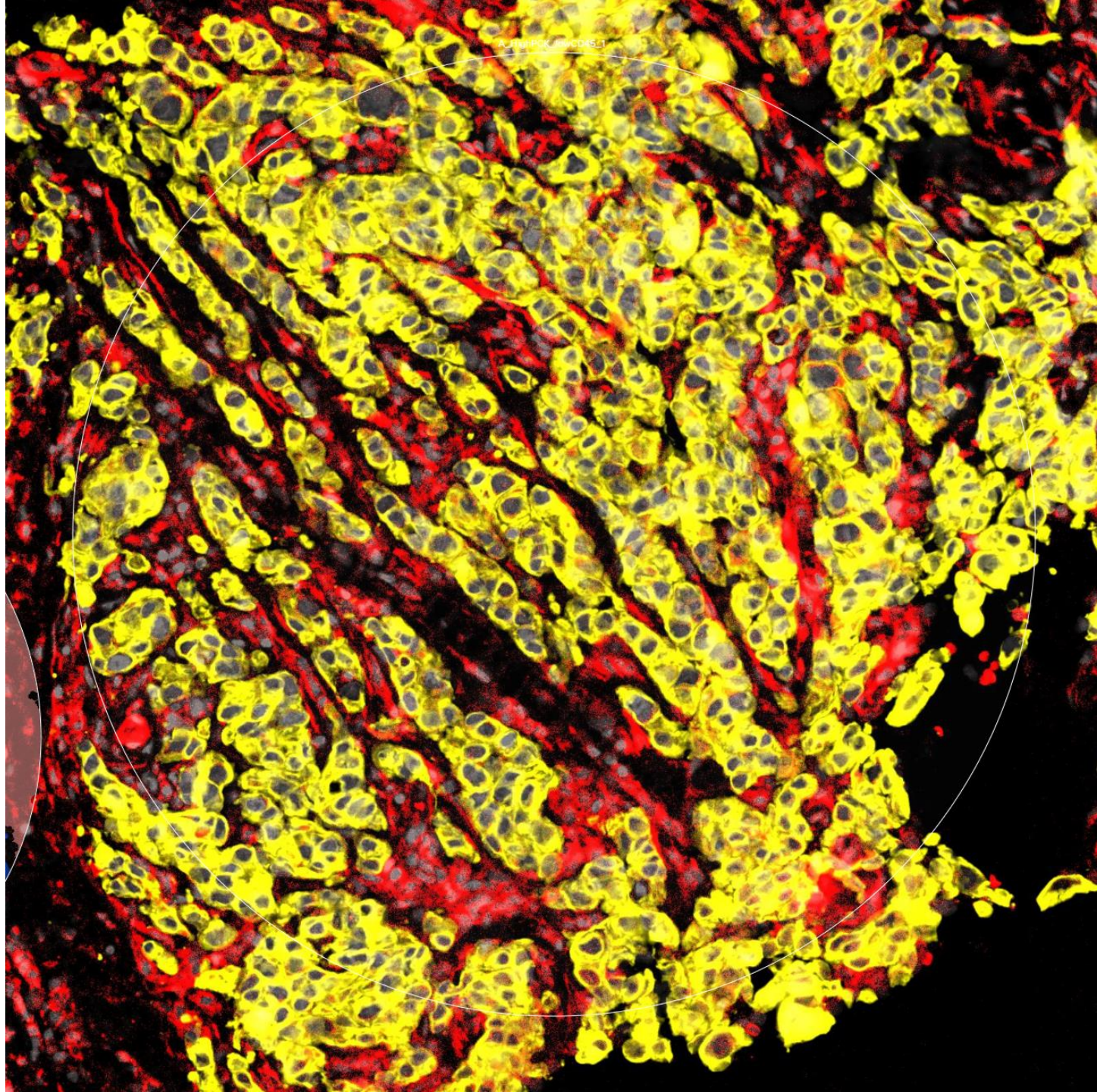
GeoMx: segmentation

DNA + CD45



GeoMx: segmentation

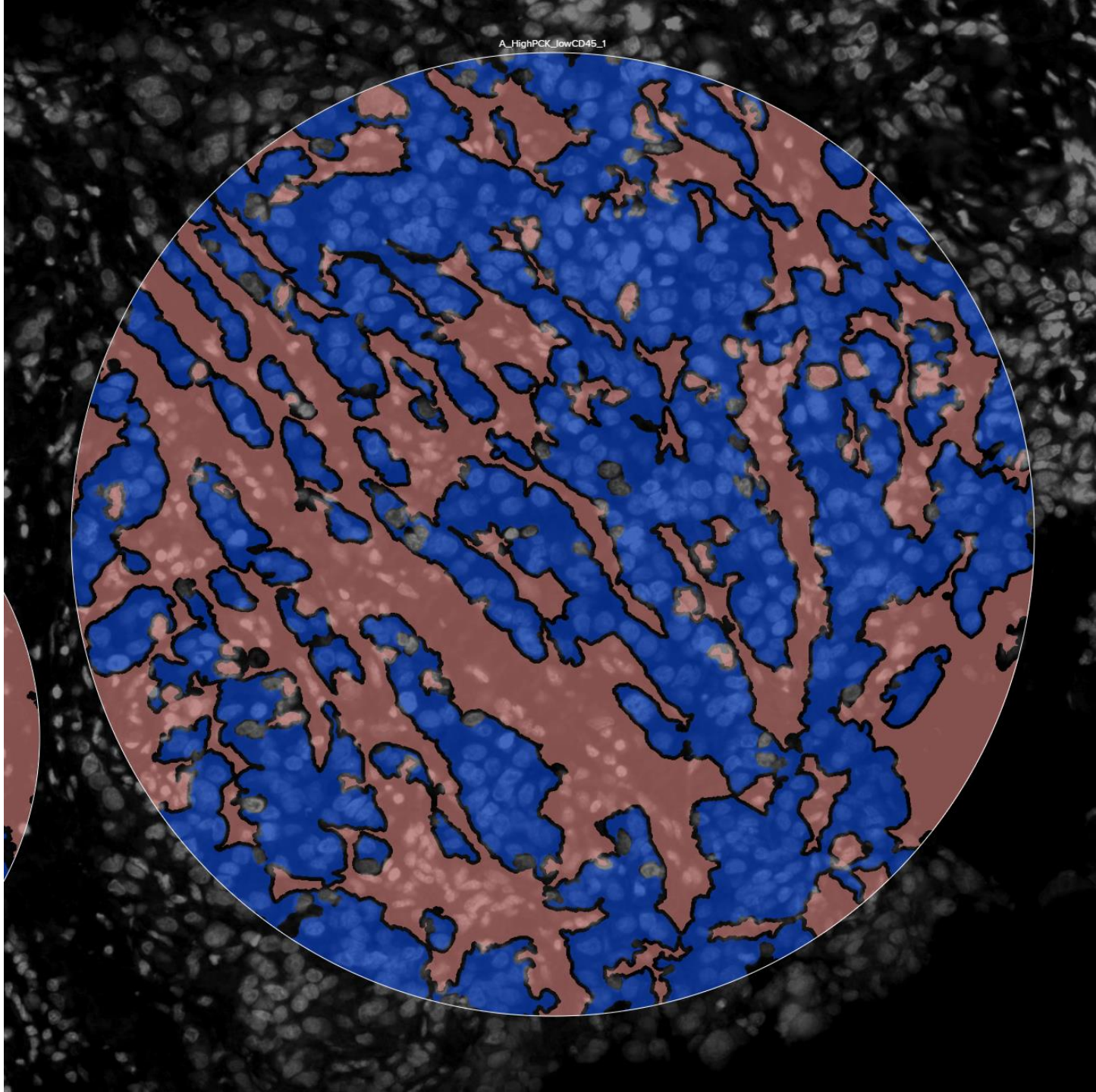
DNA + PanCK + CD45



GeoMx: segmentation

Segments defined in the ROI
Based on PanCK color

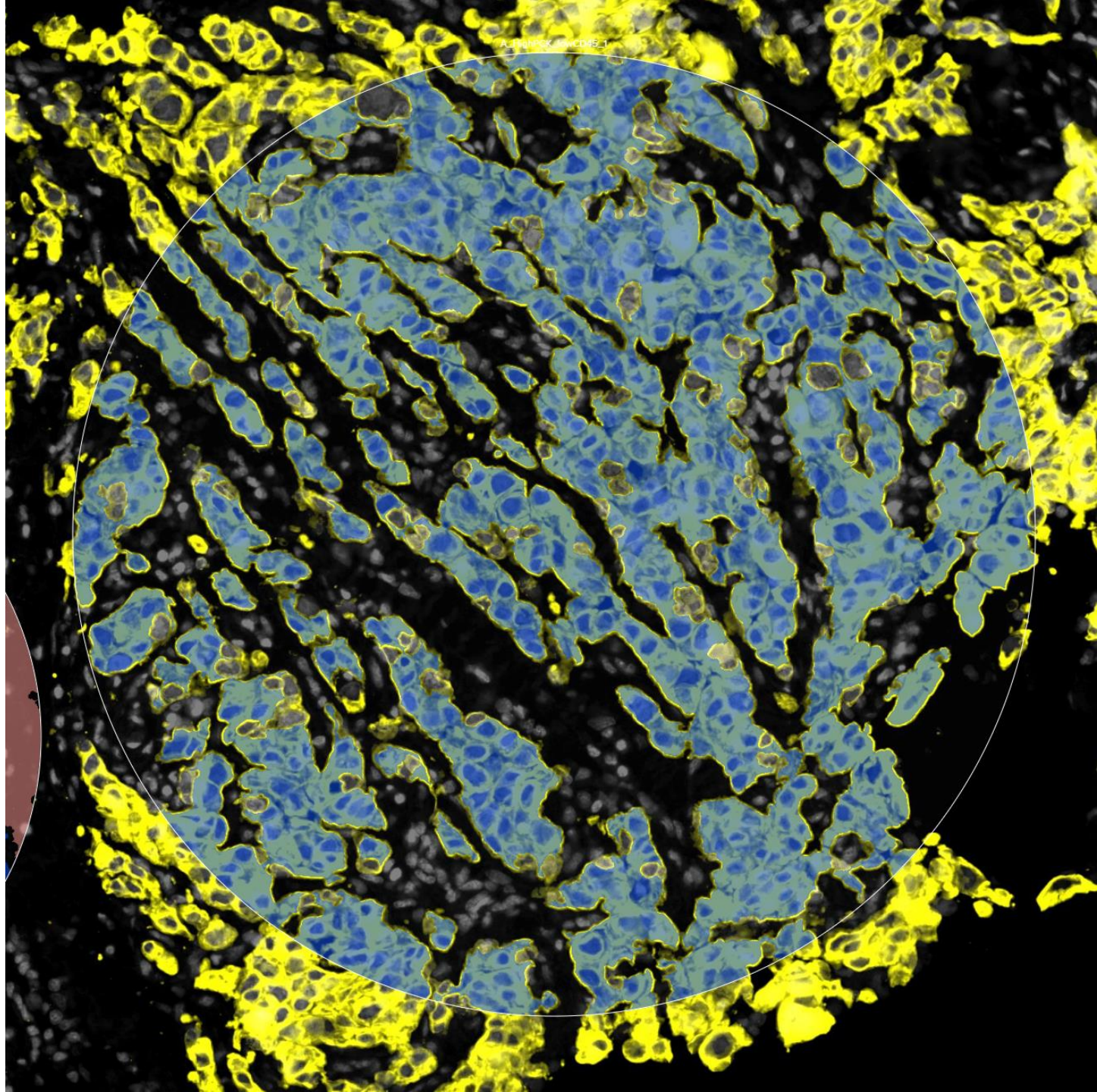
Mask (pink-blue)
Blue = PanCK high
Pink = PanCK low



GeoMx: segmentation

PanCK channel

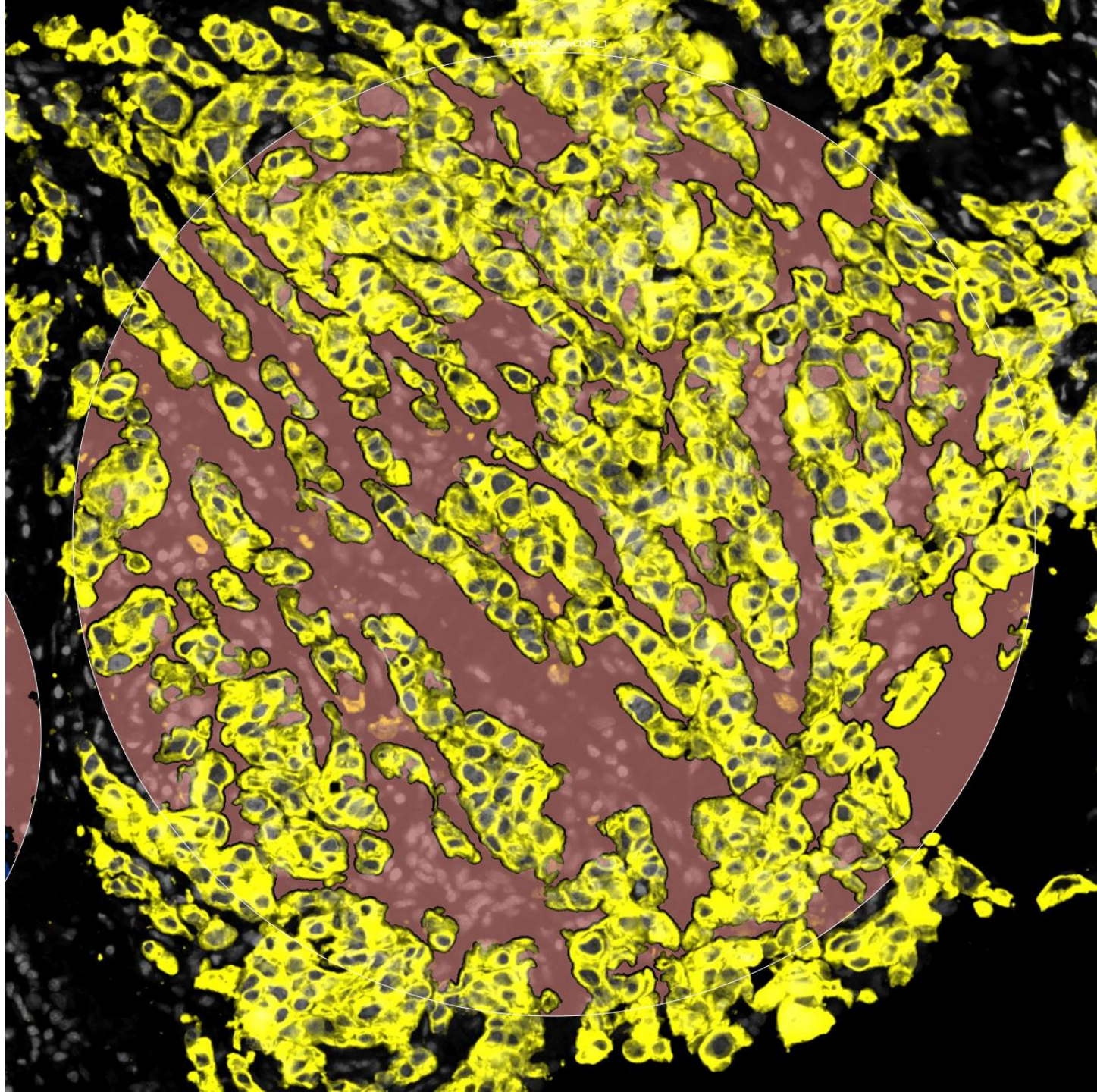
PanCK+ mask



GeoMx: segmentation

PanCK channel

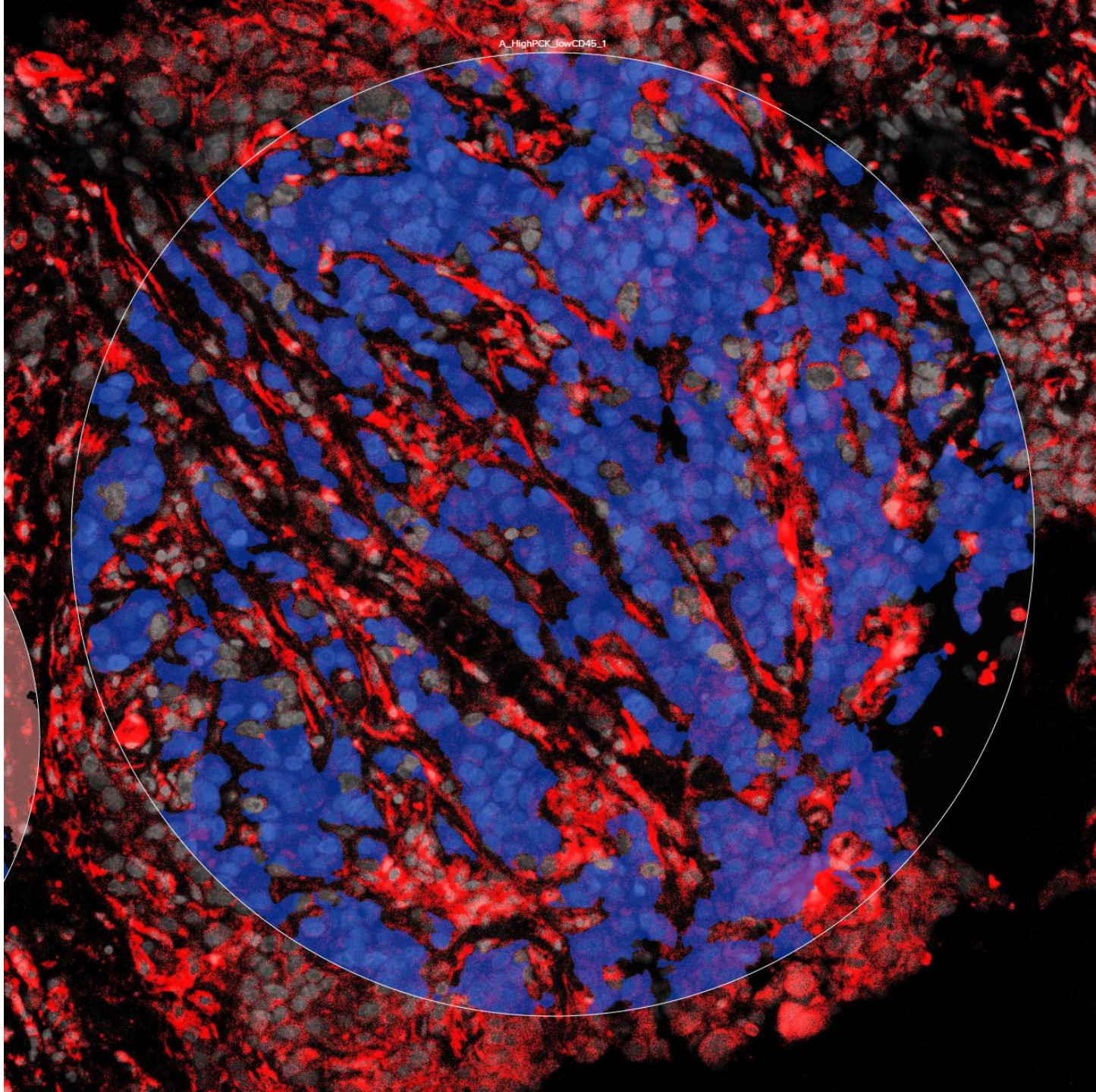
PanCK- mask



GeoMx: segmentation

CD45 channel

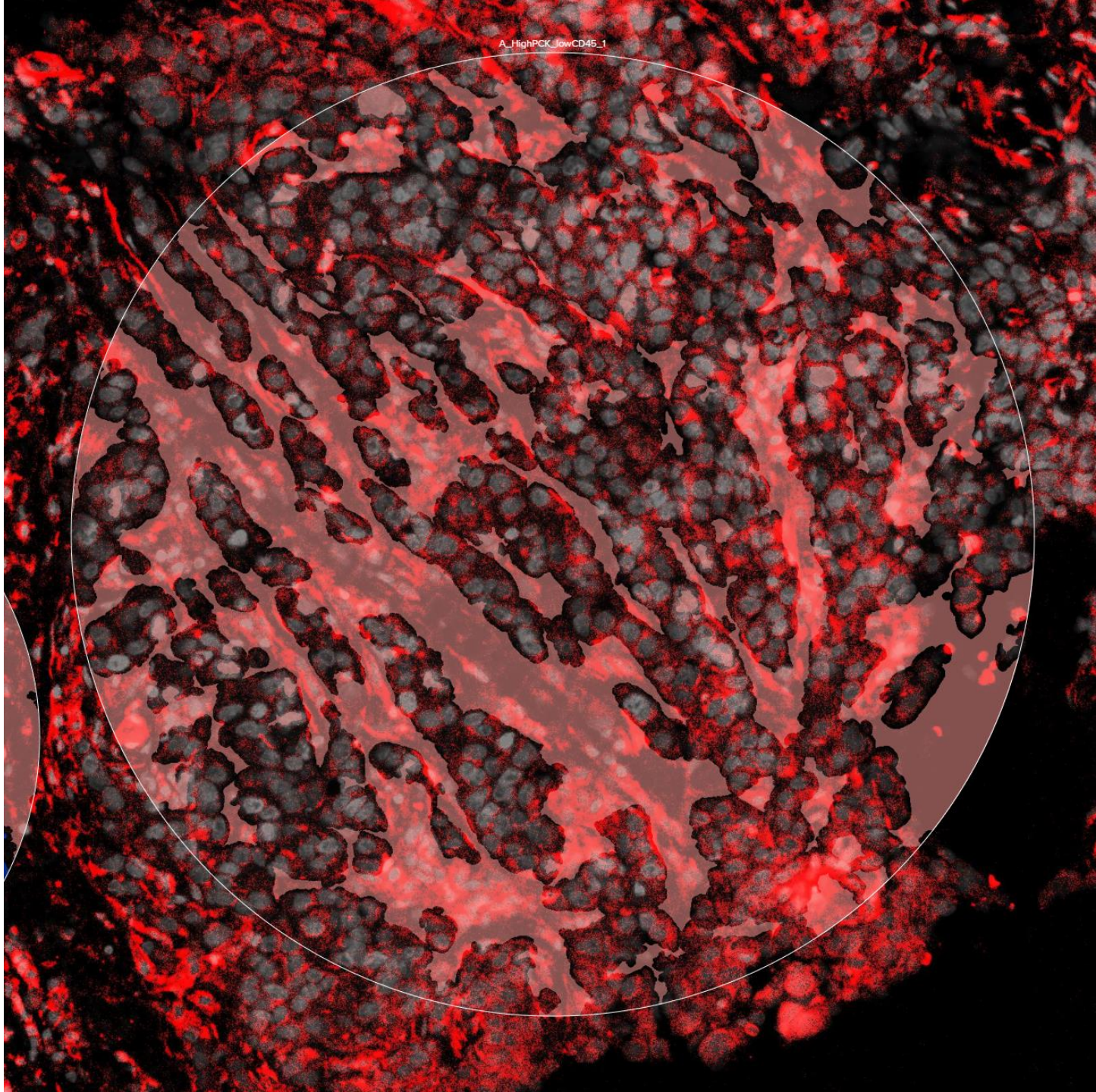
PanCK+ mask



GeoMx: segmentation

CD45 channel

PanCK- mask



GeoMx: segmentation

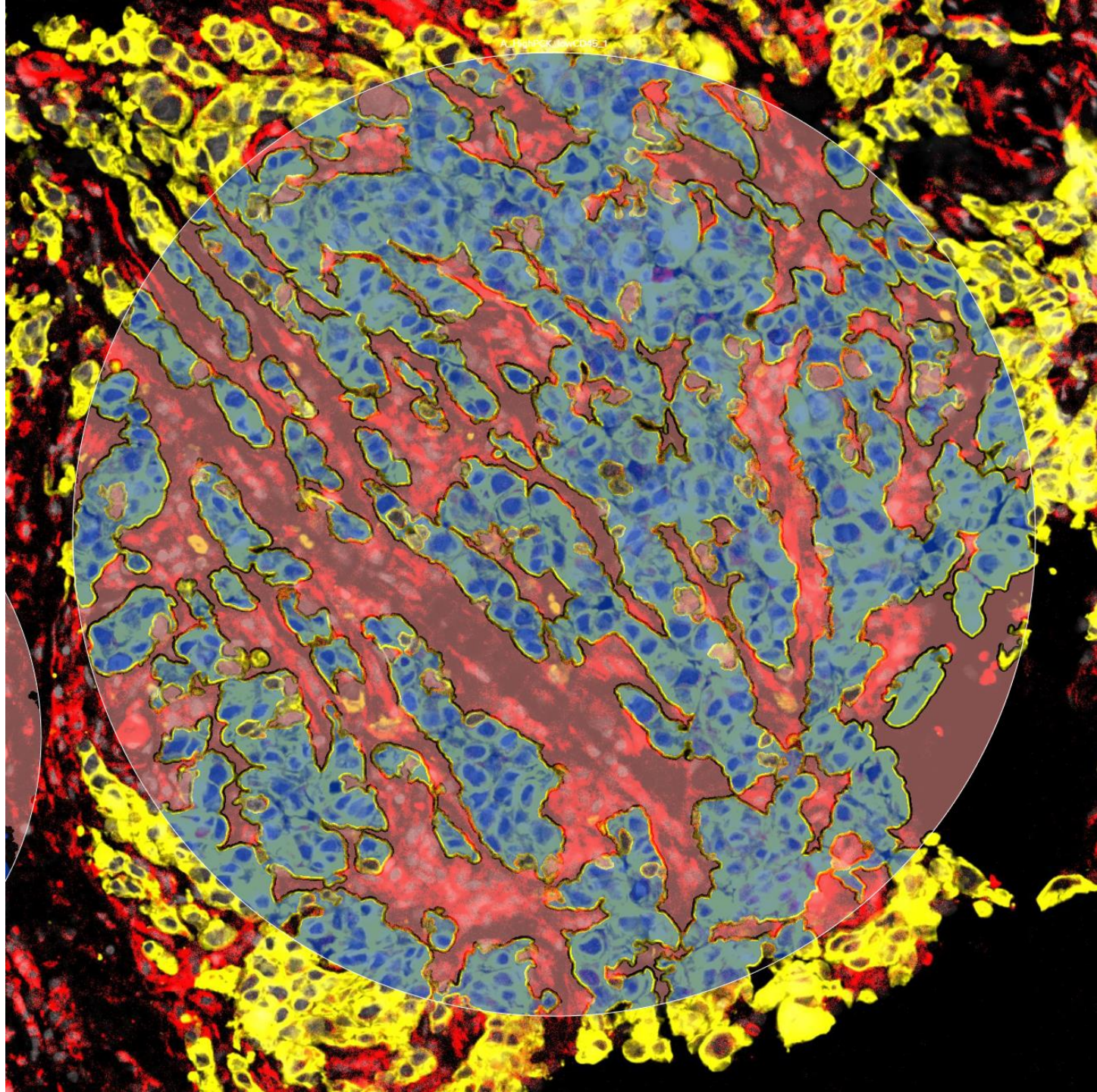
PanCK+CD45 channel

PanCK+ and PanCK- mask

Machine takes two samples from this ROIx:

ROIx PanCK+

ROIx PanCK-



GeoMx: process

Color tissue by up to 3 colors (+ DNA to see cells): most of the time blood and tumor cells are highlighted
Define regions of interest: lots of tumor, lots of blood cells, morphologically interesting, etc
Segment regions to separate into subsets base on color: within a region, take only tumor or blood
Let machine extract RNA from the slide
RNA-seq is done of samples (95 regions/samples from a slide)
Get the matrix with metainformation - analyze

Samples + regions

metadata

sample	s1	s2	s3	s4	s5	s6
group	g1	g1	g1	g2	g2	g2
X	1	2	3	4	5	6
Y	6	5	5	4	3	1
morph	ear	nose	eye	ear	nose	nose
blood	lots	abit	none	lots	none	none

expression matrix

measurement of features across sample

	s1	s2	s3	s4	s5	s6
gene1	-0.41	-0.24	0.36	-0.09	0.38	-0.01
.	0.54	-0.30	0.25	0.04	-0.19	-0.33
.	-0.03	0.48	0.44	-0.33	-0.23	-0.34
.	-0.08	-0.14	-0.26	0.58	0.20	-0.30
.	-0.35	-0.32	0.18	0.23	-0.27	0.53
geneN	-0.09	-0.03	0.21	0.56	-0.35	-0.29

Biological answers:

Tests with spatial Information in the model

- **Individual significant genes**
- **Common functions**
- **Enriched pathways**
- **Regulation networks**
- **Hypotheses to test**

Still bulk (50-500 cells per) Library for mRNA, but almost all genes Not a single cell lvl

CosMx
(spatial, single cell level)

CosMx – tissue [FOV focus]

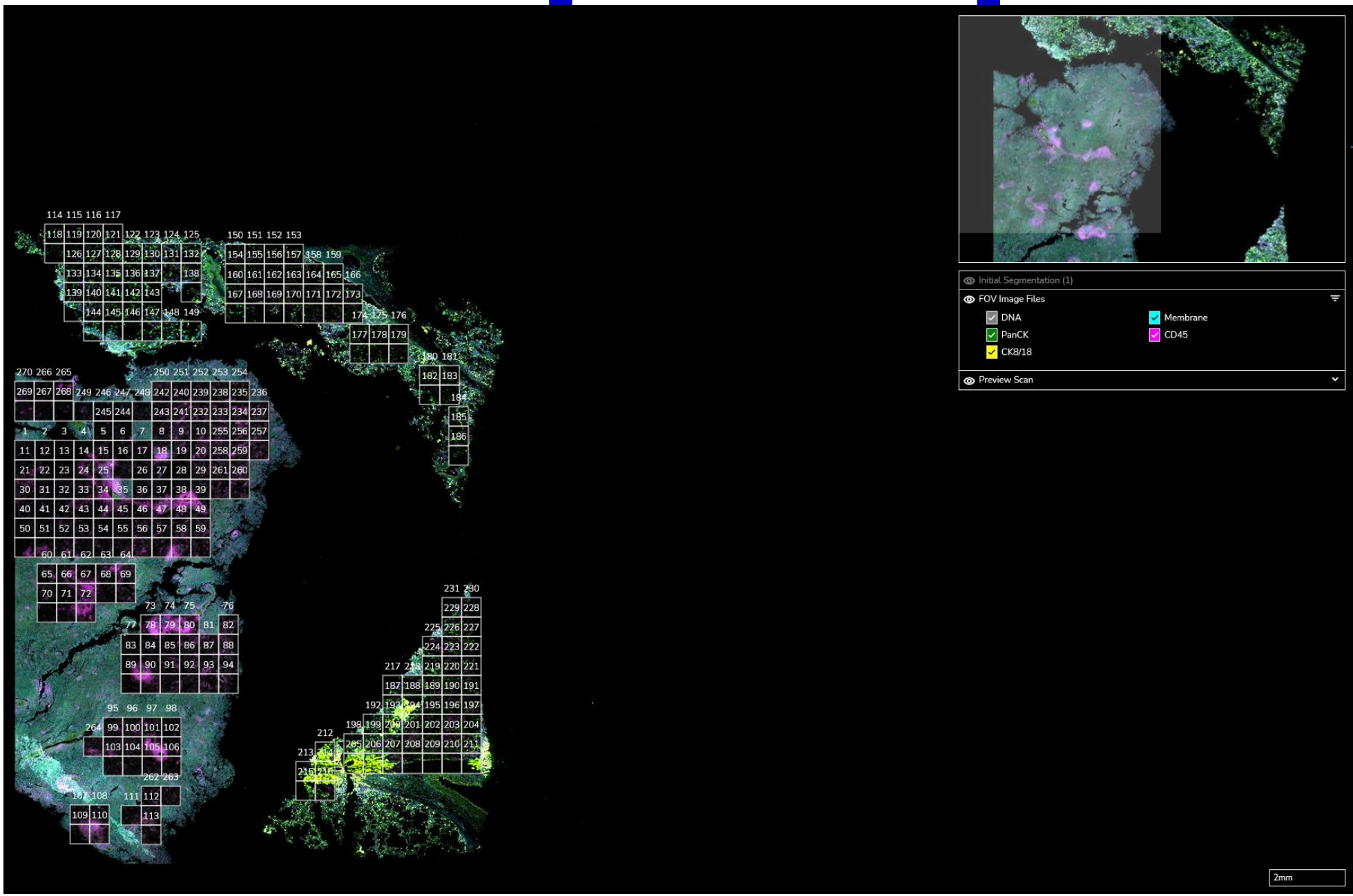
5 stain colors

Tissue overview

FOVs –
Field of view regions defined (~100-300)

Mostly performance reasons – FOVs allows to generate data only for global tissue regions of interest, image quantification/recalculation can be done only for individual FOVs

2mm resolution



CosMx – cells [segmentation]

5 stain colors

Cells segmented based on staining (DNA/membrane)

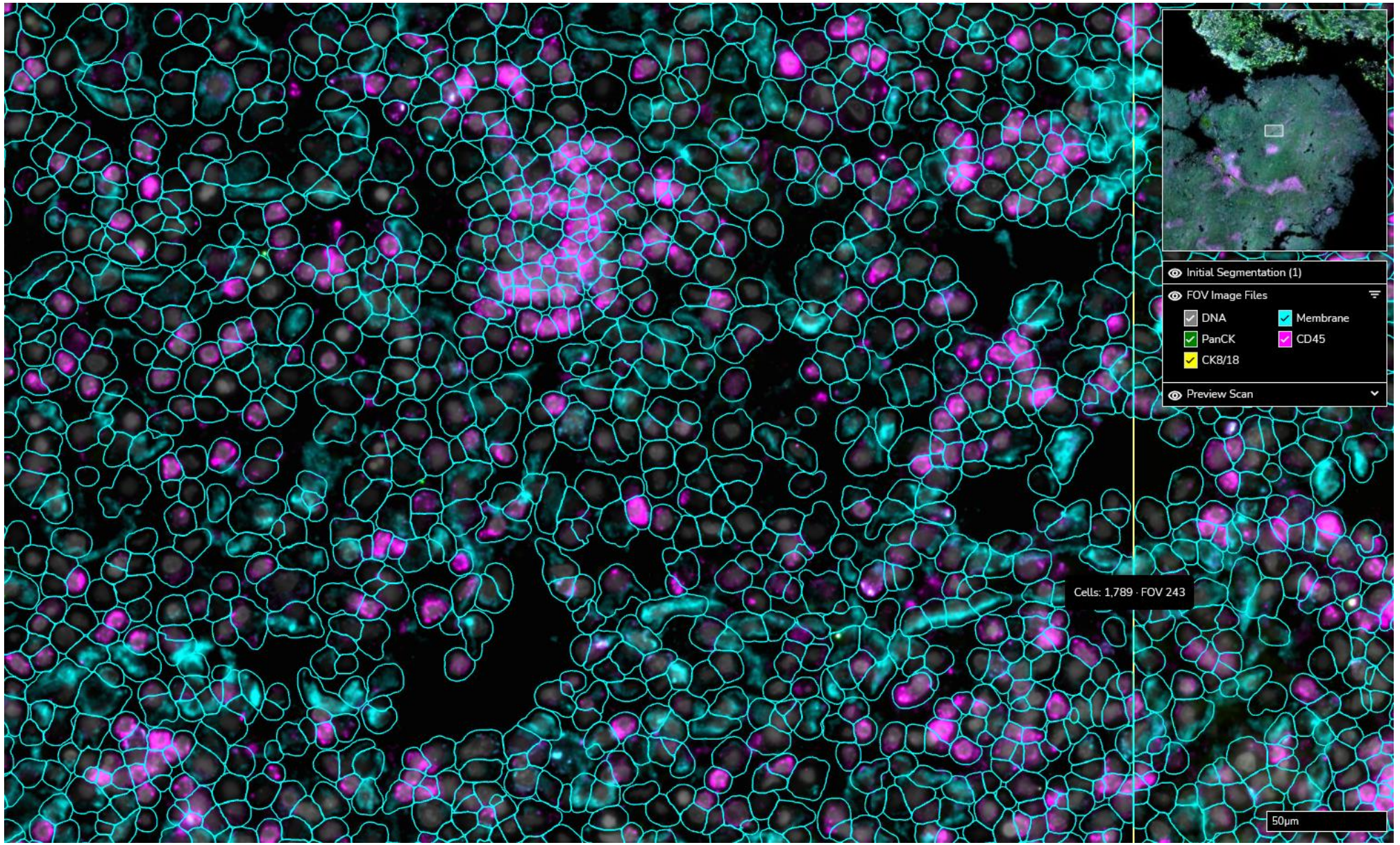
Each teal border region is algorithmically imputed cell

1789 cells in one FOV

50um resolution

Each cell now has:
- **coordinate**
- **defined shape**
- **Neighbors**
- **Intensity of 5 colors**

Segmentation can be redone – takes hours, changes downstream analysis results



Subjective – segmentation has many parameters and "goodness" is basically evaluated by user based on this image

CosMx - genes

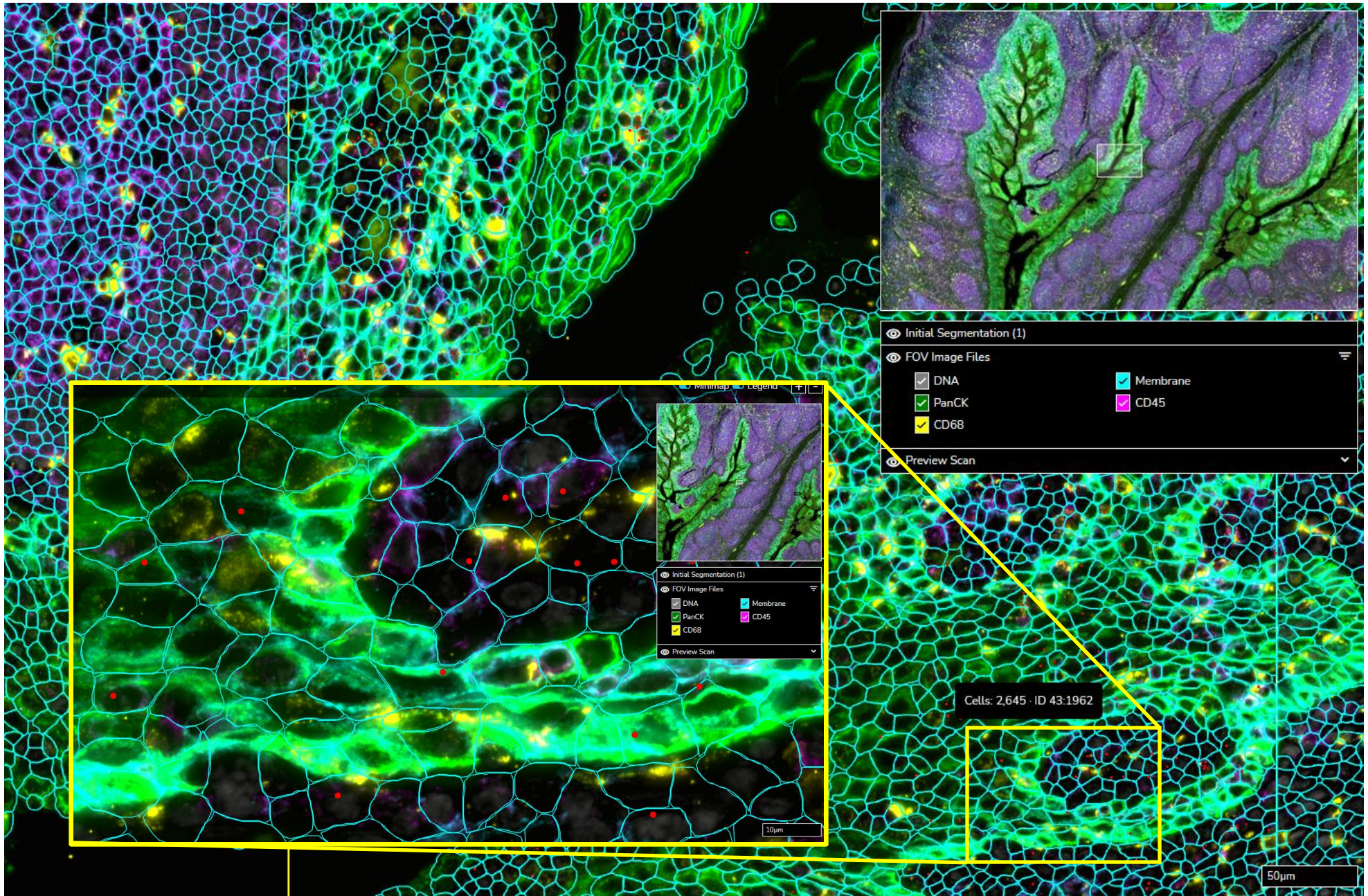
Gene expression counting

Red = CD8A gene selected to be highlighted

One dot is a barcode:
Geometrical point with fluorescence [ATCG] code that matches CD8A

So we have for each gene:
- coordinate
- cell it is in
- N copies

Cell re-segmentation can drastically change outcome



CosMx

No sequencing – all microscopy:
- Single cell resolution
- Segmentation of images to identify cells borders based on several markers
- Based on hybridization of oligos to mRNAs
- Decoding of fluorescence to get ATCG barcodes
- Counting genes in cells

Final data and possible analyses:
- **Regular single cell:**
- gene expression in cells:
  - ~100-300 FOVs, 1-3K cells in each
  - 100-600 (for 1K panel) genes in each
- QC/clustering/cell typing/DEG

- **Spatial:**
- Proximal Ligand-Receptor
- Other neighborhood analysis
- Micro-environment conditional differential expression

Pipelines for most things
Export options with customized analysis:
images (takes forever and huge): derive intensities, segmentation DIY
segmentations results: cells, coordinates, intensities
quantification results: gene counts in cells

Samples + single cells

metadata

sample	**s1**	**s2**	**s3**	**s4**	**s5**	**s6**
group	g1	g1	g1	g2	g2	g2
X	1	2	3	4	5	6
Y	6	5	5	4	3	1
morph	ear	nose	eye	ear	nose	nose
blood	lots	abit	none	lots	none	none

expression matrix

measurement of features across sample

	s1	**s2**	**s3**	**s4**	**s5**	**s6**
gene1	-0.41	-0.24	0.36	-0.09	0.38	-0.01
.	0.54	-0.30	0.25	0.04	-0.19	-0.33
.	-0.03	0.48	0.44	-0.33	-0.23	-0.34
.	-0.08	-0.14	-0.26	0.58	0.20	-0.30
.	-0.35	-0.32	0.18	0.23	-0.27	0.53
geneN	-0.09	-0.03	0.21	0.56	-0.35	-0.29

Total summary: go spatial?

- **Without strong spatial component in experimental design, it can easily become a slightly more refined bulk RNA-seq**
- **Tissues with unique morphology that may have an effect on expression are good candidate**
- **Tumors with/without proximity of certain immune cells, etc**
- **Hopefully GeoMx could improve ROI defining process with more sophisticated and user-friendly interface. Why they do not provide computationally-aided selection, not quite clear.**
- **Analysis is still a straightforward pipeline like for a regular bulk RNA-seq: QC, identifying ROI outliers, DEG, DEG corrected by estimated computationally cell proportions, enrichment IPA/GSEA, Heatmap/Volcano/etc for a paper/grant**