Genomic Data Visualization

Jan 28, 2022

Quick review:
- lab website: https://jef.works/genomic-data-visualization/
  - previous notes from class, homework assignments
  - class slack
    - previous recordings from class
    - catch up
  - first homework assignment (due Monday)
  - any questions, feel free to reach out to me or Lyla

Spatially Resolved Transcriptomics Data

Learning objectives:
- motivation
- important to understand where your data comes from
-> graduate level
-> please raise your hand if you have questions, chat Lyla

Motivation
-> why do we care about profiling transcriptomics spatially

When we consider a complex mamalian organ like the brain, what are the cell-types of the brain?
- glial cells
    - oligodendrocytes
    - astrocytes
- neurons
    - excitatory neurons
    - inhibitory neurons
        - subtypes

- cell-types are arranged in very specific manners
    - organization is related to function
    - characterizing how these cell-types (cell-states) are organized in the brain

- how can we distinguish one cell-type from another?
    -> all cells have DNA
    -> all cells have the same DNA
    -> how is one cell different from another?
        - different cells express different proteins
        -> we can measure what proteins are expressed in different cells, tell us about what cell-type it is

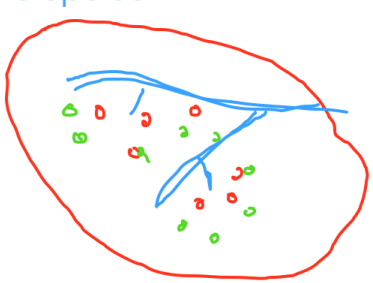
Measuring proteins is really hard
-> we don't really have a way to measure every protein
-> we're limited to measuring proteins where we have good antibodies

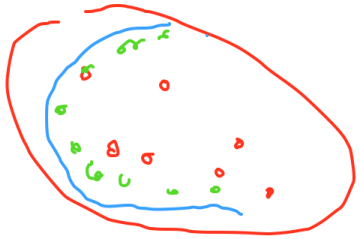
So what's an alternative?
-> we can look at morphology
    -> neurons vs. glia
    -> excitatory vs. inhibitory neurons
    -> subtypes neurons it becomes harder
    -> breast cancer, morphology can allow us to distinguish between cancer and non-cancer
        -> it gets hard to distinguish between invasive ductal carcinoma vs. BRCA+ vs. HER2+

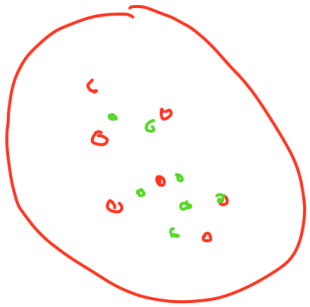
So what's another alternative?
- look at what makes protein -> genes/mRNAs
- mRNAs are much easier to measure

-> want to be able to measure gene expressions in single cells (cellular transcriptome)
-> preserve spatial organization of these cells in tissues
    -> if we can do this:
        -> see how cell-types spatially organized?
        -> how cell-types may be expressing different receptors or ligands or chemokines (cell-cell communications)

tumor biopsies







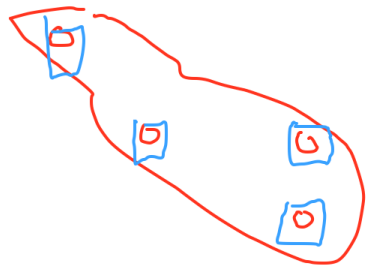
cancer cells exhist in a complex millieu of other cell-types in the tumormicroenvironment

spatial organization of blood vessels?

spatial organization of immune cells (T cells)?
- infiltrated vs. excluded spatial organization
-> potential impact prognosis, response to treatment (immunotherapies)

Historic techniques
- manual microdissection + sequencing
- sequential single molecule imaging

Manual microdissection

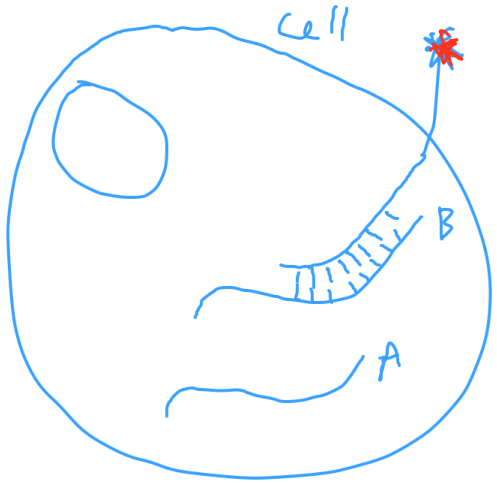


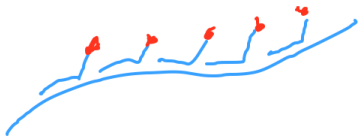
RNA sequence manually sectioned out regions
-> what this will give me is what genes are being expressed in each of my manually sectioned chunks

-> 10s, 100s of regions
-> regions you dissect out will be pretty big (more than 30 cells)
-> somewhat low throughput approach

Sequential imaging

- smFISH = single molecule flourescence in situ hybridization

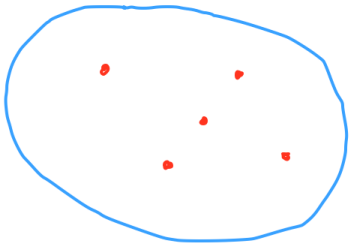




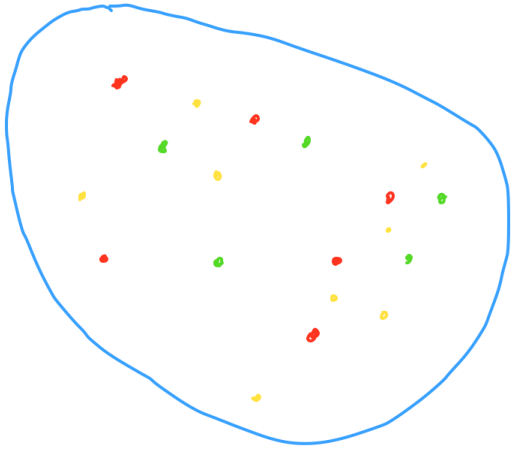
AT
TA
CG
GC

ACGGGAA
TGCCCTT

image:



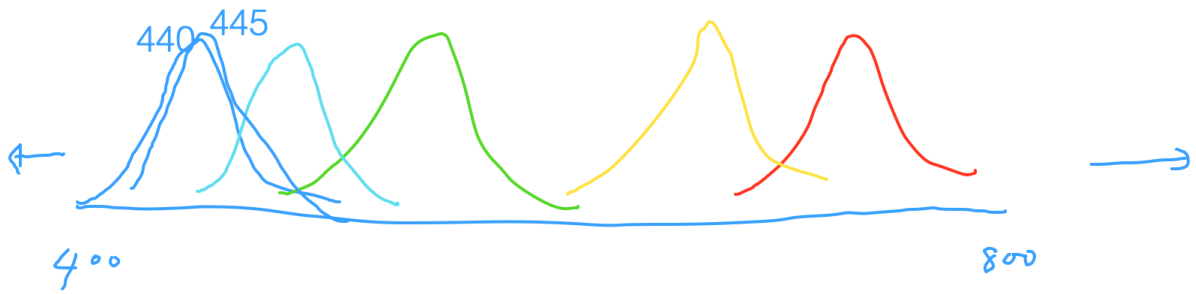
5 copies of gene B
in this particular cell



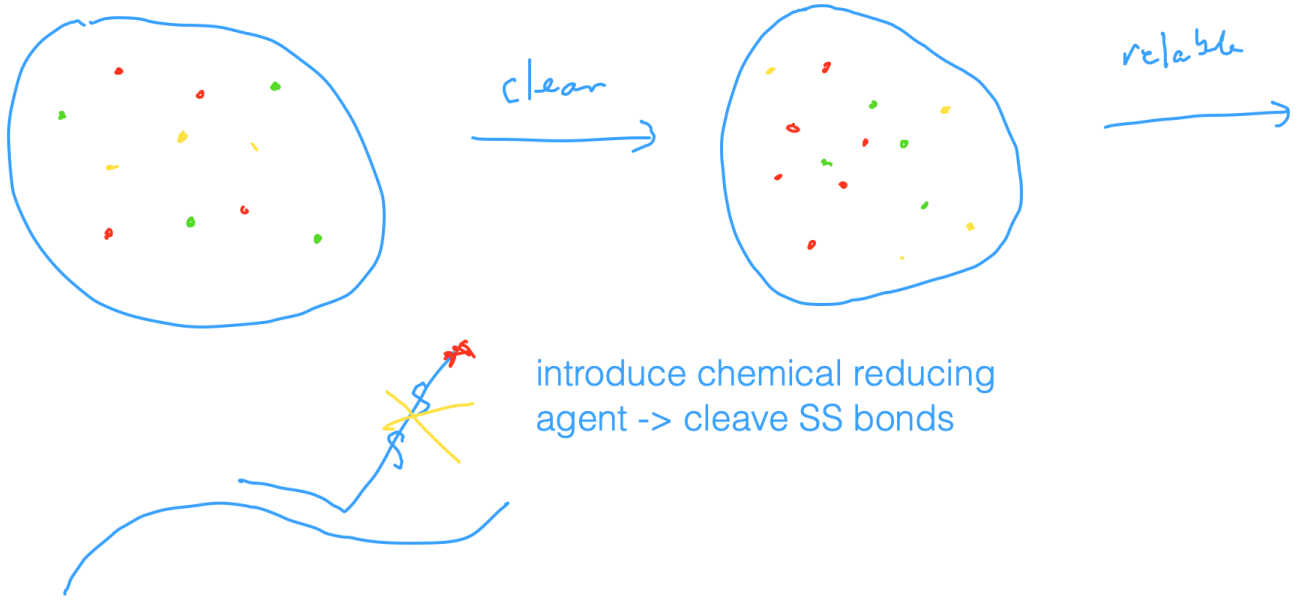
How many genes are there in mamallian cell?
- 20K
-> not every gene is expressed in a cell at the same time
-> 4K genes being expressed

How many colors do you think I could distinguish?
(a machine could distinguish?)
- 100?
- 3?
- 10?
-> turns out it's somewhere around 8

Reason is something called: spectral overlap



Solution: Ok, I can only measure 8 genes at a time, but I can clear my tissue and do sequentially another 8 genes
-> required chemistry innovations for clearing a tissue



introduce chemical reducing agent -> cleave SS bonds

Through many rounds of sequential smFISH -> measure more genes

-> 4k (ideally 20k) all genes

How many rounds of imaging is it going to take to measure 20k genes if every round I can measure 8 genes?
- 2500 rounds of imaging

If it takes 1 hour to do a round of imaging, how many PhDs is it going to take…
-> take a long time

- good for profiling a limited set of gene targets
-> low throughput, not amenable for full transcriptome characterization

High throughput spatially resolved transcriptome profiling technologies
- MERFISH = multiplexed error robust flourescence in situ hybridization (smFISH analogue)
- Visium = spatially resolved capture + sequencing (microdissection analogue)

MERFISH
- builds on smFISH

represent genes as binary codes

	imaging round 1	2	3	4
gene A				
gene B				
gene C				
gene D				

in 4 imaging rounds, I measured 4 genes by smFISH

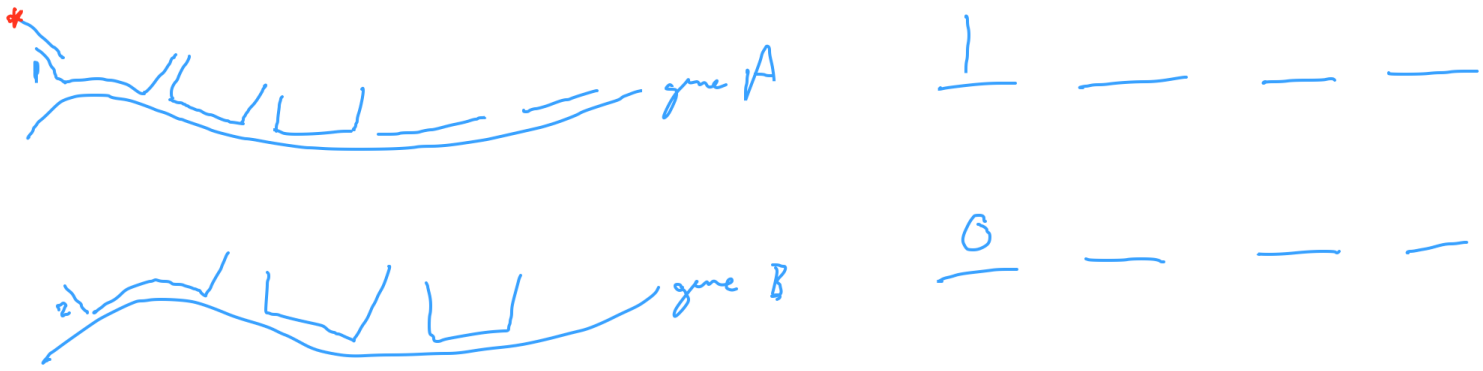
-> scales linearly

	imaging round 1	2	3	4
gene A	1	1	0	0
gene B	1	0	1	0
gene C	0	1	1	0
	0	0	0	1
	1	0	0	0
	1	1	1	1
	1	0	0	1
	⋮			

binary barcode of length 4
~~with 2 "on bits" (2 1s)~~

0 1
2 × 2 × 2 × 2

$2^4 = 16$



we can design probes to imprint binary barcodes onto genes that we can then read out with sequential smFISH

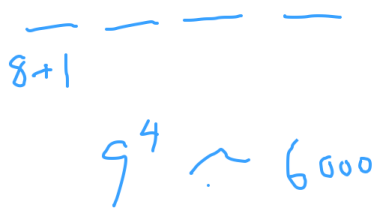
rather 1-1 relationship between genes and phlores, 2^N (N is the number of imaging rounds)

How many rounds of imaging would it take to profile the whole transcriptome?

through N rounds of imaging, we can profile 2^N genes
-> profile 20,000 genes

2^N = 20,000
-> 14 to 15 rounds of imaging

Gene A = 1 0 1 0 (red)
Gene B = 1 0 1 0 (blue)



Ideal world != real world

Real world != perfect
-> poorly exposed images
-> phores that don't actually anneal
-> make it hard to read off the binary barcodes

Gene A 1 1 0 0
Gene B 1 1 1 0

If I make a mistake in imaging round 3, I wouldn't be able to distinguish whether an RNA is gene A or gene B

MERFISH introduces an “error robust” binary barcode
- rather than using all possible binary barcodes
2^14 - 2^15 to capture 20k genes
- Hamming distance
-> rather than using all possible binary barcodes, we use only a subset that are all seprated by some “Hamming distance”

Hamming distance = minimal number of errors that could have transformed one string to another

Jean
Sean
-> what is the Hamming distance between these two strings?
-> Hamming distance of 1

Dear Bean
-> who is Lyla is referring to?
-> we wouldn’t know who Lyla is referring to

- depending on Hamming distance, we may be able to identify errors
- if all our gene binary barcodes are separated by Hamming distnace of 1, we would be able to tell when errors were made

gene A 1 0 1 0
gene B 1 1 0 0

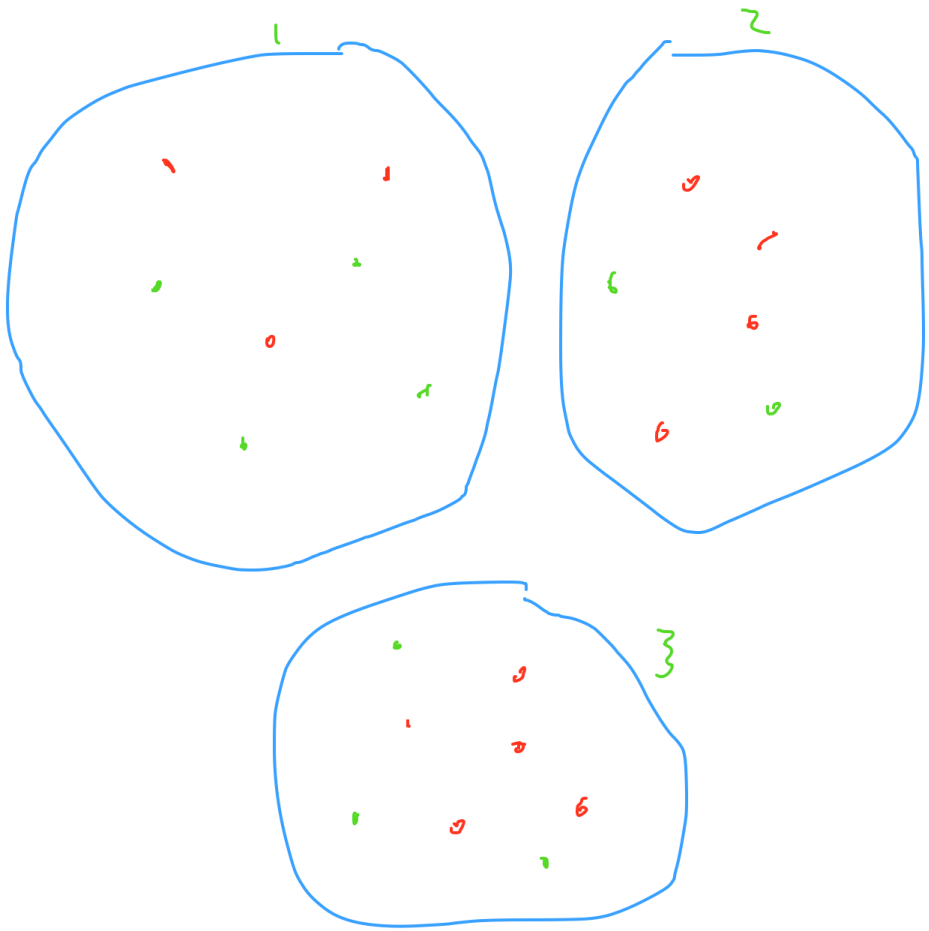
read out 1 0 0 1 -> error

Jean
Jose

Hamming distance between our names?
Hamming distance of 3

Lyla emailed ‘Bean’ -> she probably misspelled Jean
Lyle emailed ‘Josy’ -> she probably misspelled Jose
I’m inherently assuming that the truth has fewer errors

Summarize for MERFISH
- representing genes as error robust binary barcodes
- combinatorial labeling to imprint these barcodes and read them
- using many rounds of smFISH

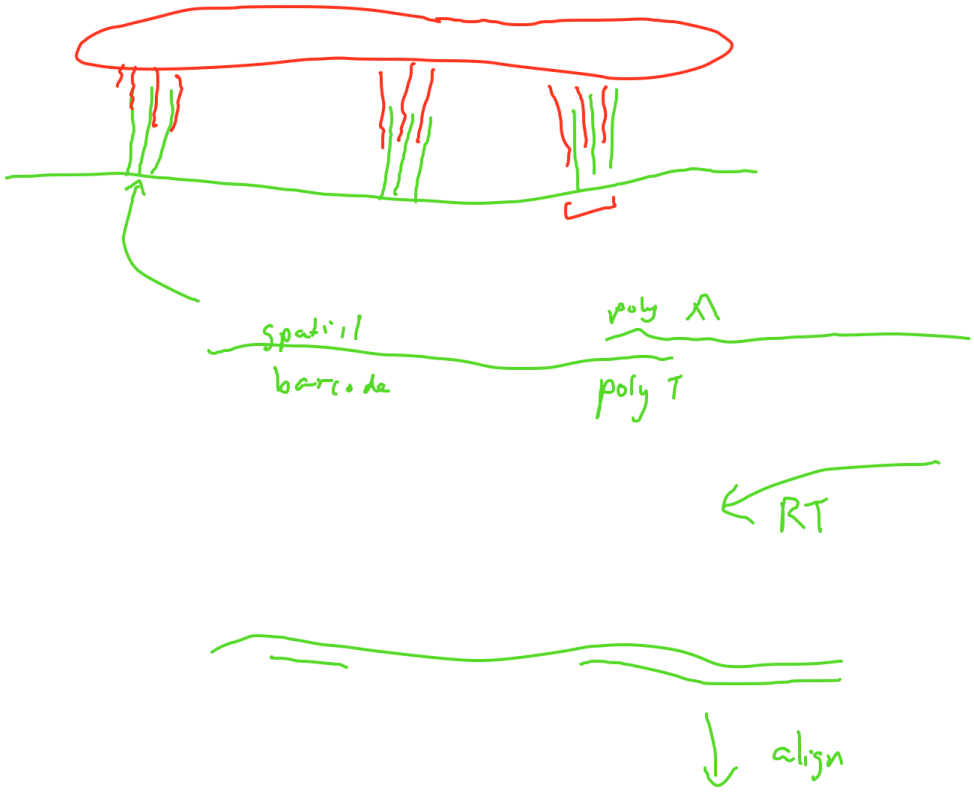


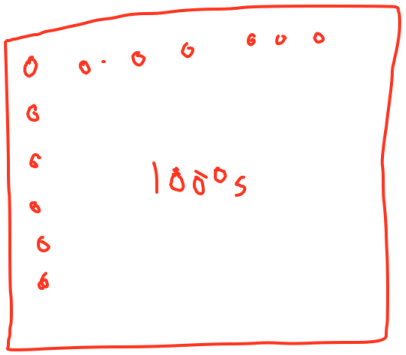
cell	1	2	3	...
gene A	3	4	5	
red	4	2	3	
13				
gr				

Gene counts matrix of cells
- thousands of cells
- hundreds of genes

Spatial positions of the cells
- x y coordinate

Visium - spatially resolved capture + sequencing





End up with
- gene expression matrix
- genes that are expressed in each spatially resolved spot

Spatial positions for the spots

MERFISH vs. Visium

single molecule, single cell resolution

design probes to target specific genes

spot resolution (multiple cells)

polyA capture -> measure all polyA transcripts -> unbiased way to capture the transcriptome -> end up with more genes