AI in Science

Getting closer to the TME

Robert Gentleman

2026-05-28

Outline

- AI and how it can impact your work
- I will use two practical examples
- studying the tumor micro-environment
- using Isolation Forests to do label transfer

Introduction

- Modern molecular biology has progressed rapidly on based primarily on the development of high throughput assays that allow us to measure tens of thousands to millions of features across large populations (e.g. cells, individuals).
- AI in its various forms has also been developing very rapidly and now can aid us in many ways

AI Examples

- alphaFold - protein structure prediction
- alphaFold-multimer - use a similar approach to understand protein complexes
- alphaMisense - start to understand the impact of mutations on protein structure and function
- digital pathology
- find tumor cells in patient samples
- identify different cell types and anatomical structures
- cell segmentation
- foundation models for post-translational modifications

Foundation Models

- as we saw yesterday there are many different ways that foundation models can be brought to bear
- prediction of RNA structure
- we are just beginning to be able to identify candidate anti-bodies from known antigens
- being able to do that is going to change how well and how rapidly we can develop new drugs

Drugs - and the future of drug development

- there are only about 30K genes in the human body
- we know most of them very well now
- we know chemistry and can make small molecule drugs (small molecules transfuse directly in to cells)
- making anti-body therapeutics (these are large molecules and cannot get into cells easily) is much harder
- we have long made protein therapeutics - insulin and human growth hormone

Ozempic Contains Non-Canonical Amino Acids

- Semaglutide (Ozempic) contains one non-canonical amino acid.
- Position 8:
  - Alanine is replaced by 2-aminoisobutyric acid (Aib).
  - Aib is a non-proteinogenic (non-canonical) amino acid.
  - This substitution makes semaglutide resistant to cleavage by the DPP-4 enzyme.
  - Increased DPP-4 resistance contributes to the drug's long duration of action.

Other Modifications to Semaglutide

- Position 26:
  - Lysine is chemically modified with a C18 fatty diacid attached through a linker.
  - This promotes albumin binding and slows clearance from the bloodstream.
- Position 34:
  - Lysine is replaced by arginine.
  - This ensures that only Lys26 is available for fatty acid attachment during synthesis.

Summary

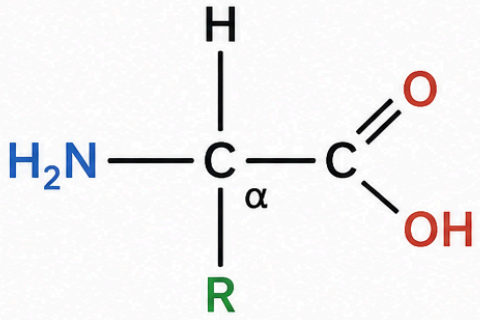
- One non-canonical amino acid: Aib at position 8.
- One chemically modified canonical amino acid: acylated lysine at position 26.
- One canonical amino acid substitution: lysine to arginine at position 34.
- Together, these modifications enable once-weekly dosing.

New classes of Drugs

- there are essentially an infinite number of non-canonical amino acid

Example of a Non-Canonical Amino Acid

Non-canonical amino acids are not among the 20 standard proteinogenic amino acids used in natural proteins.



General structure of an amino acid

H_2N– : amino group

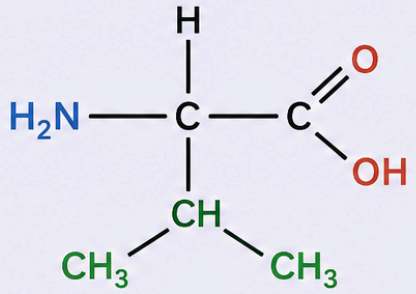
–COOH : carboxyl group

R : side chain (varies)

αC : alpha carbon

Example:

2-Aminoisobutyric acid (Aib)
(a non-canonical amino acid)



- Not encoded by the genetic code
- Often used in peptide and drug design
- Can improve stability, resistance to proteolysis, and bioavailability

##

Opportunities

- so this opens up a huge area of new therapeutics

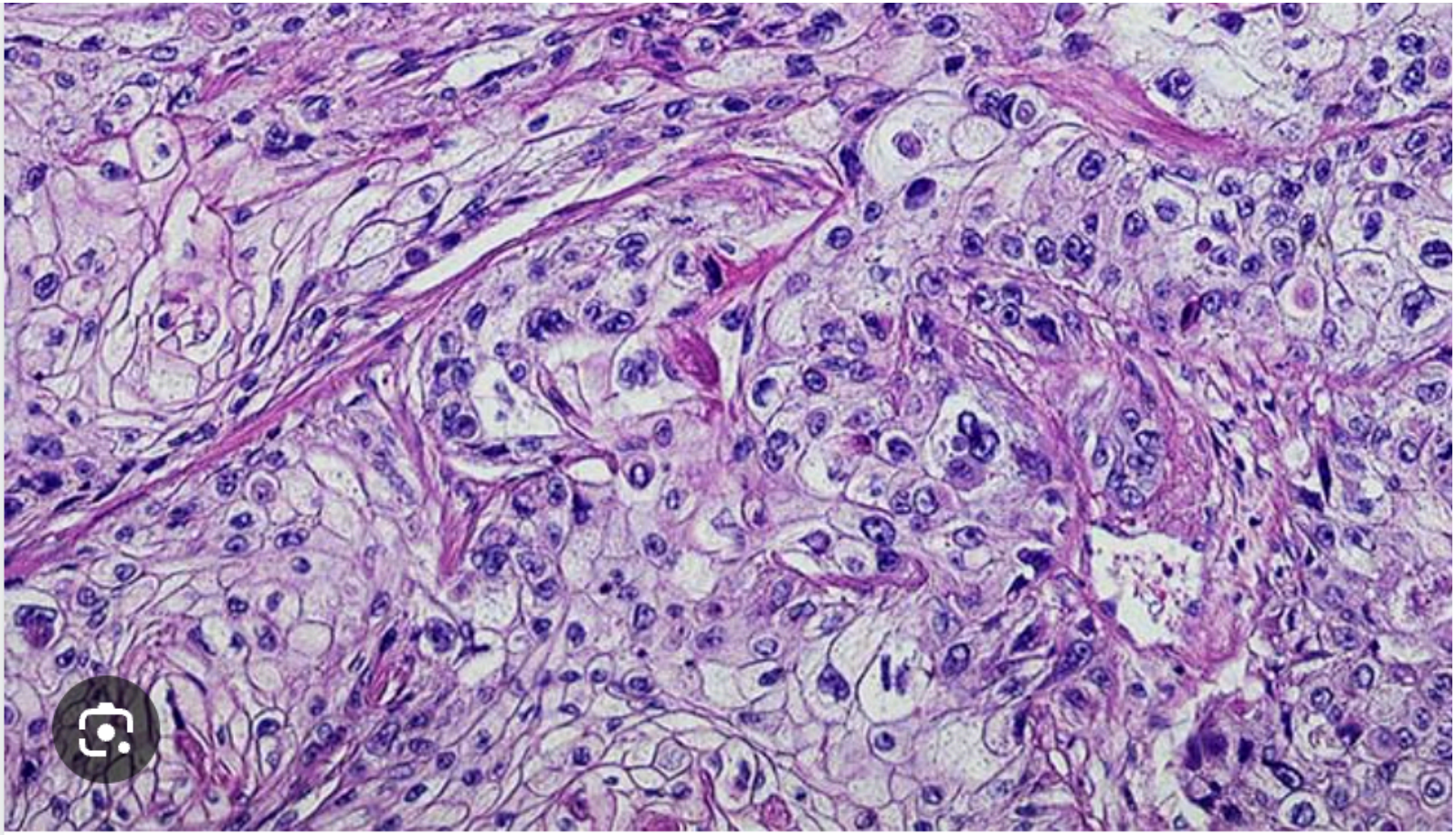
Immuno-therapies

- how do we get the right drug to the right place at the right time
- alot of research now is going in to trying to understand what factors matter the most
- personalized medicine: how much does the patient, their genome and their exposures matter
- we have developed great immuno-therapies but they largely work well on liquid tumors
- why don’t they work on solid tumors and what can we do about it

Assessing the tumor microenvironment

- that was what I hoped to do - but as it turns out…
- without a lab and set of post-docs I made a little less progress
- I am going to show you the tools I am working on to get to a better place
- cell labeling - especially for immune cells is important
- we work primarily with mIF and so have a number of conjugated antibodies (30-ish)
- one key consideration is the difference between markers that influence cell type and those that influence cell activation (we will see one example in the PCA Space)
- second key consideration: if I cannot integrate my predictions and my data so I can visualize both together I will have a very low success rate
- third key consideration: in imaging the tissue slice is confounded with the assay

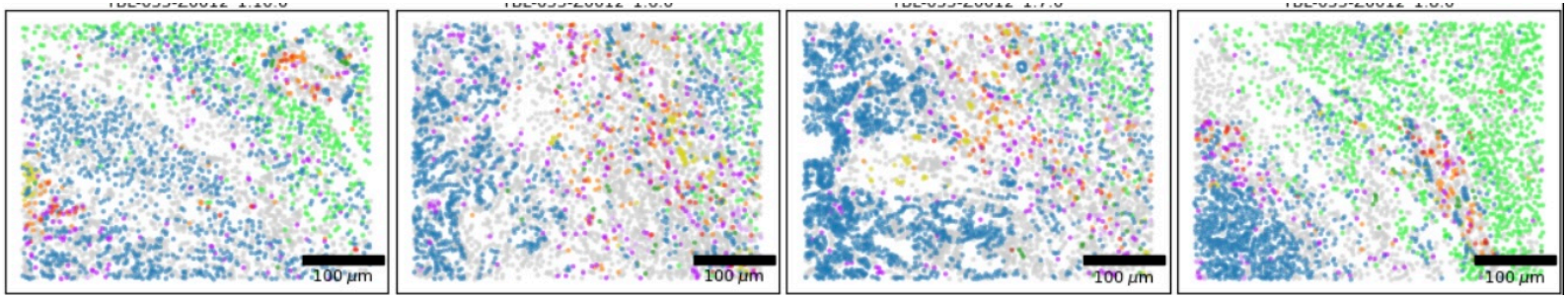
An image



Opinions Stated with very little proof

- we should care about gradients they are things we can measure
- do immune cells change their activation markers depending on distance from tumor/blood vessels/other immune cells
- it is unlikely we will have the accuracy needed to accurately attach cell surface markers to the correct cell (but imaging could get better)
- cell density seems to be interesting in the tumor microenvironment
- anything I can estimate is going to be very spatially local because I only have information in the x and y directions and am blind in z
- there are no good null models that I am aware of

Tumor Microenvironment



- here are 4 ROIs from a single tissue
- we plot cell type predictions in false color
- notice how different my neighborhoods will be for each of these slices

Hypothesis Testing and Inference

- there really is not any way I can see to put this kind of data into an inferential system
- we have one slice of some tissue that was then imaged and regions selected usually by a pathologist
- there is no model that let's me say anything about the tumor in this person (I have no sampling framework) let alone about these kinds of tumors in general
- even if I get data from lots of people, I still do not have a sampling frame work
- you really need to focus on the descriptions of the data in front of you and what can be determined from them

Live demo here

- these phenotype maps are nice, but it is really hard to check and see if the labels I am predicting are accurate so I want to be able to view the predictions with the actual images
- work with Ian Dryg, Satyakam Mishra and Allison Nau
- we are exploring the use of vizarr (https://github.com/hms-dbmi/vizarr)
- which is very lightweight but has some nice functionality that we are working to extend

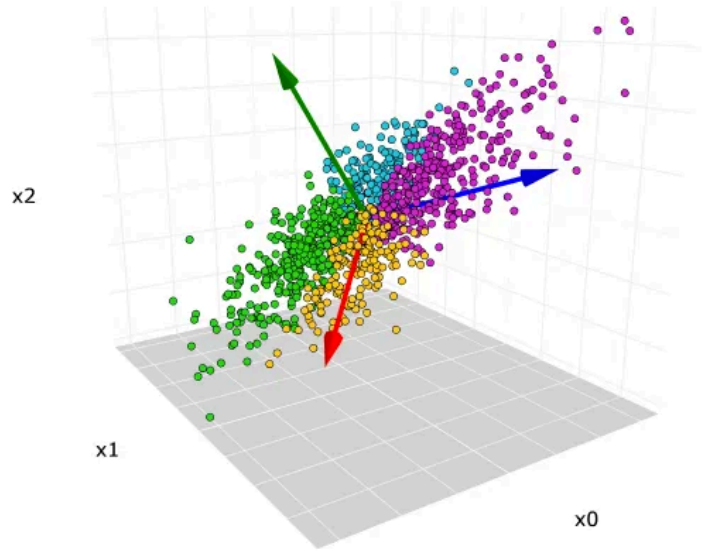
In the end…

- we want to get something that looks like a cells x genes matrix
- we want to retain important spatial and structural information
- we want to be able to model and understand the important features
- those might be genes but they could be secreted factors, chemokines and cytokines

PCA refresher

- We will refer to the data matrix as $\mathbf{X}$, and each row as $\mathbf{x_i} = (x_{i,1}, x_{i,2}, \ldots, x_{i,p})$.
- We can see in the picture that we could rotate the coordinate axis to align differently with the data
- There are many possible rotations, but we want to consider one where the axis are oriented towards directions of greatest variation in the point cloud.

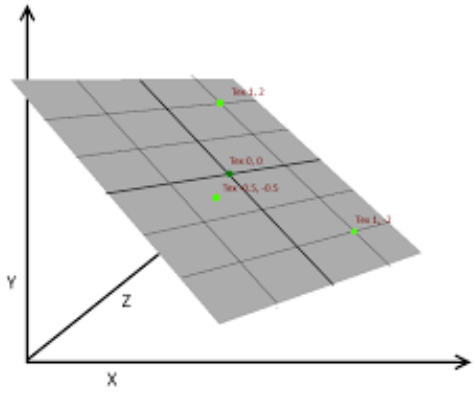
Three important points



- If we rotate the coordinate axis, as in the plot, the relationship between the data points has not changed at all.
- The data matrix ($\mathbf{X}'$) has N rows and p columns.
- The PCs are chosen to be orthogonal

The Null Space

- Sometimes we have p columns, but the point cloud really only occupies some lower dimensional space.



- in the rotated coordinate system we only need 2 dimensions, so one of the rotated variables will be zero
- if we restrict ourselves to the first k PCs then the null space becomes $PC_{k+1} \ldots PC_n$

PCA in Molecular Biology

- in many different molecular biology assays we collect data on individuals or cells
- that data could be genotypes, gene expression values, methylation etc
- the data structures for these assays are usually transposed from those used above for the exam scores on students
- in biological examples there tend to be many more quantities that are measured than there are individuals
- so for most of these examples we will be working with the transpose of the data matrix
- we are often interested in finding a low dimensional set of features that explain most of the between individual variation

Case Study with Single cell RNA-seq experiments

- In a single cell experiment we typically have hundreds to thousands of cells.
- We have 10’s of thousands of genes….
- So each cell is a point in some 10-40K space…but we think that the there are useful summaries
- We want to use PCA analysis as a way to do some form of dimension reduction - to those directions where there is a lot of variation in the data.
- outliers can greatly skew the results and we need to be careful to ensure that our analysis is not too reliant on a few observations.

The basics of the process

- get your single cell data and do QA/QC to remove genes (rows) and cells (columns) that seem to have technical or biological reasons to be be suspect
- log transform and size normalize the data
- do some sort of filtering for the top K most variable genes (K and how you measure variable are parameters)
- do PCA
- use the first M PCs to do clustering
- revert back to the full count matrix and use UMAP and tSNE to visualize

Set up the data

- We will examine the single cell data from workflow #3 in the OSCA book
- http://bioconductor.org/books/3.17/OSCA.workflows/
- This is a peripheral blood mononuclear cell (PBMC) dataset from 10X Genomics (Zheng et al. 2017). The data are publicly available from the 10X Genomics website, from which we download the raw gene/barcode count matrices, i.e., before cell calling from the CellRanger pipeline. The data were processed as described in that workflow.

LM on Clusters

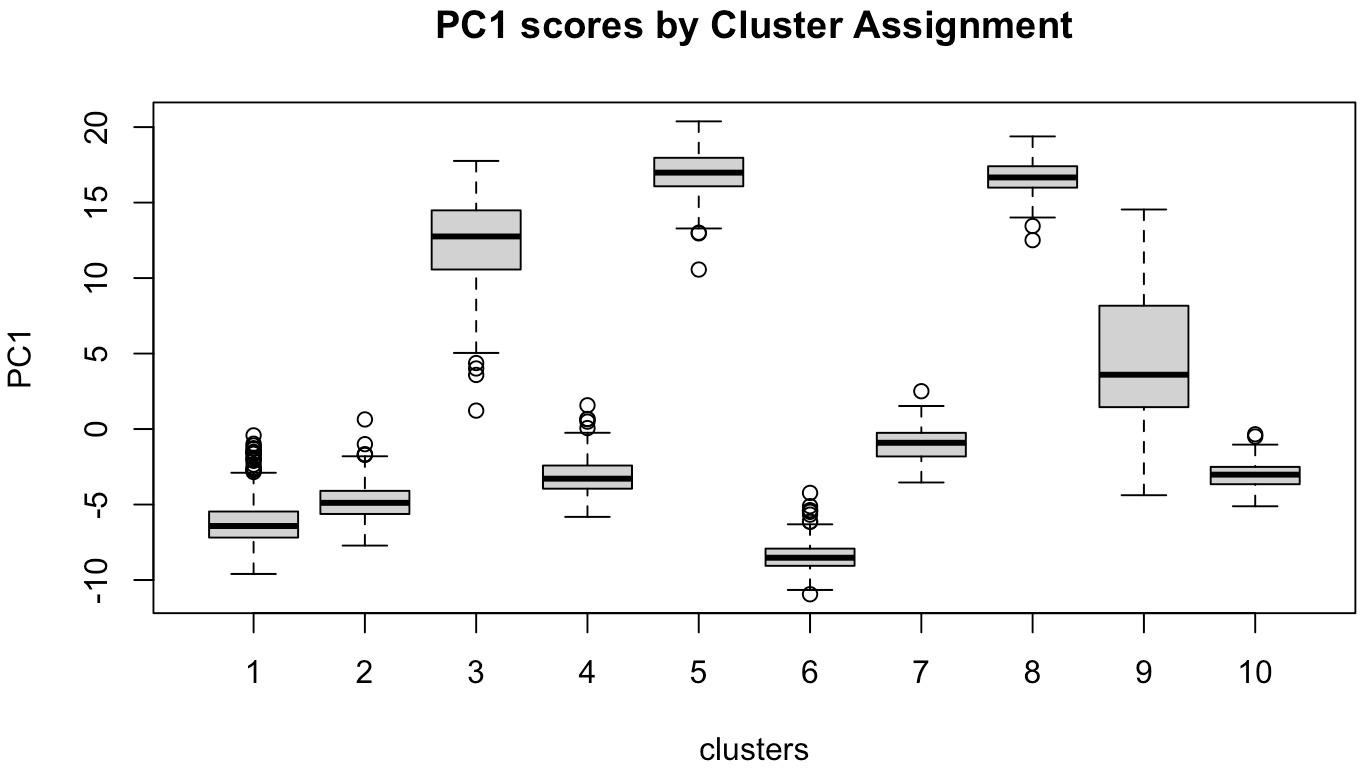
Here we consider a set of models, one for each PC, where we ask how much of the variation in the PC is explained by the clusters

The variables 1_{Cj} are indicator variables, the i^{th} element is 1 if the i^{th} observation is in the j^{th} cluster and 0 otherwise.

$$PC_l = \beta_1 \cdot 1_{C1} + \cdots + \beta_k \cdot 1_{Ck} + \epsilon$$

We can fit this model for each of some selected number of PCs.

Boxplot on Clusters



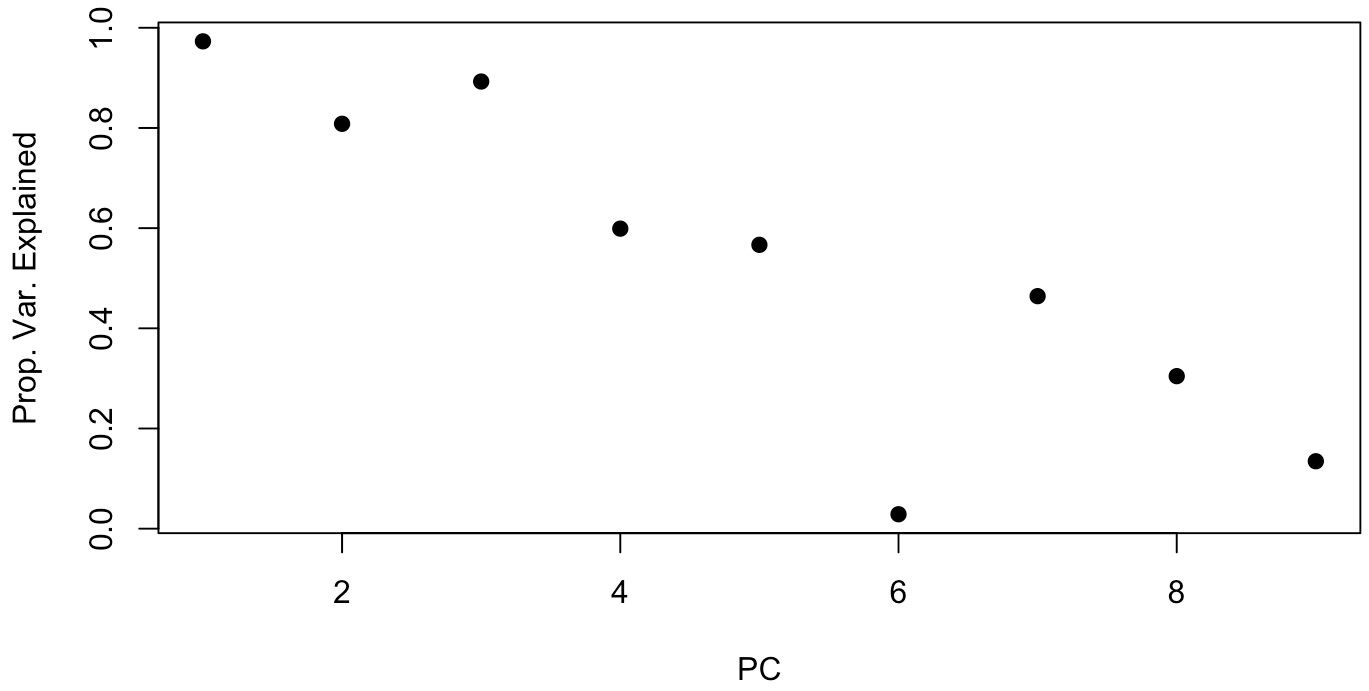
- note that this shows PC1 is essentially a contrast between a subset of the clusters

Is there info in the PCs we have not used?

- we regress each PC in turn against the cluster labels and get the multiple R^2
- for each regression, the multiple R^2 tells us about how much of the variation in that PC is explained by the clusters
- if that value is especially low, as it is for PC6 - then that direction is not really reflecting the clustering - something else is driving the variability in that direction

Multiple R^2 per PC

- note that for PC6 the multiple R^2 is very close to zero, suggesting that the variation in that direction is not described by the clusters



PC4

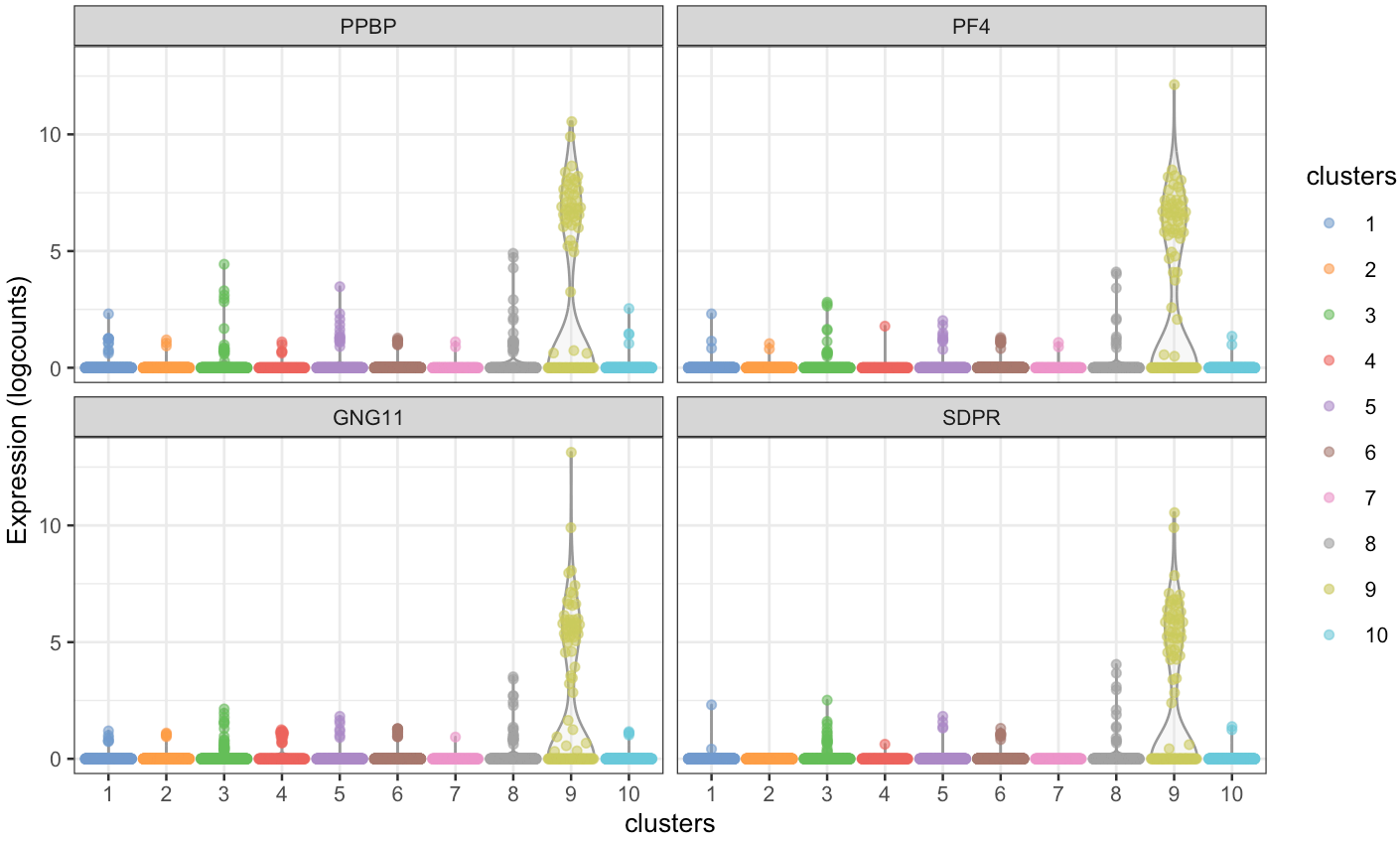
- we identify cluster 9 as being the unusual cluster in our plot
- we now find the highly variable genes within cluster 9
- so we only use the cluster 9 cells

```
exprs9 = assays(sc_exp_norm_trans)$logcounts[, clusters==9]
sdbyg9 = rowSds(exprs9)
names(sdbyg9) = row.names(exprs9)
topsdbyg9 = sort(sdbyg9,dec=TRUE)[1:50]
topsdbyg9[1:10]
```

```
##      PPBP       PF4     GNG11      SDPR HIST1H2AC      CCL5    MALAT1     TUBB1
##  3.456763  3.334462  3.000800  2.929638  2.928735  2.841697  2.821872  2.664090
##      NRGN    TMSB4X
##  2.660178  2.634199
```

- some of these are platelet factors, the first is a sign of platelet activation

Plot all four genes by groups



Summary of our Findings about PC4

- our regression of the PCs against the cluster identities suggested that in the direction of PC4 there was some variability that was not being explained by the clusters
- we found that most of that variation was contained in one cluster, cluster 9
- when we looked at the highly variable genes in that cluster we found that the top four seemed to be coordinately expressed
- these genes were involved in platelet activation
- but we have not proven anything, and we would want to follow up with a biologist who is knowledgeable about platelets to see if there are further experiments we should carry out

Isolation Forests and Label Transfer

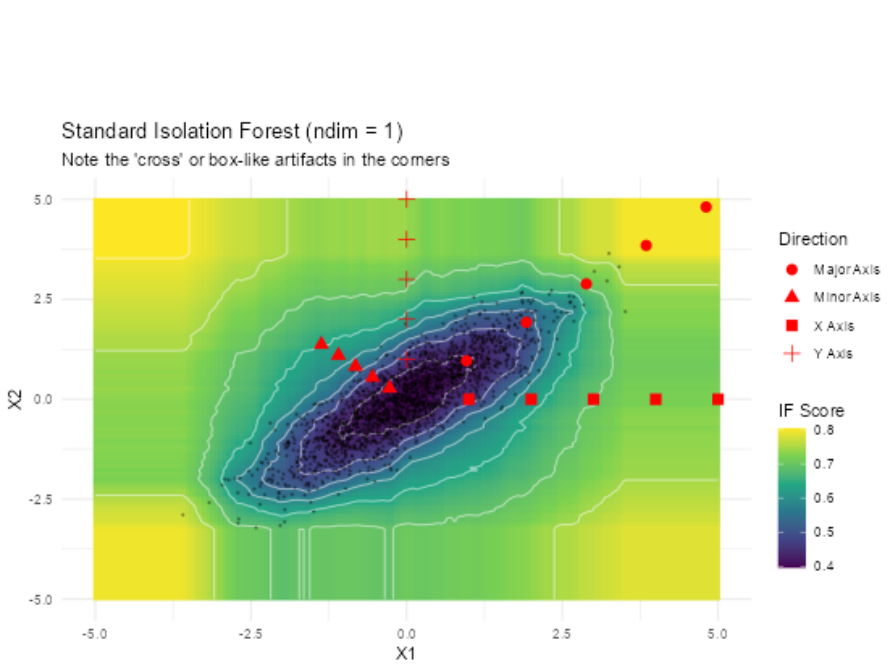
- finding anomalous points in high dimension can be very challenging
- in some cases small number of similar but unusual points might be the most interesting findings
- the isotree package (10.32614/CRAN.package.isotree) provides a good R implementation
- Liu, Fei Tony, Kai Ming Ting, and Zhi-Hua Zhou. “Isolation forest.” 2008 Eighth IEEE International Conference on Data Mining. IEEE, 2008.
- the key observation was: unusual observations are often easier to isolate by choosing a small number of random split points for the data

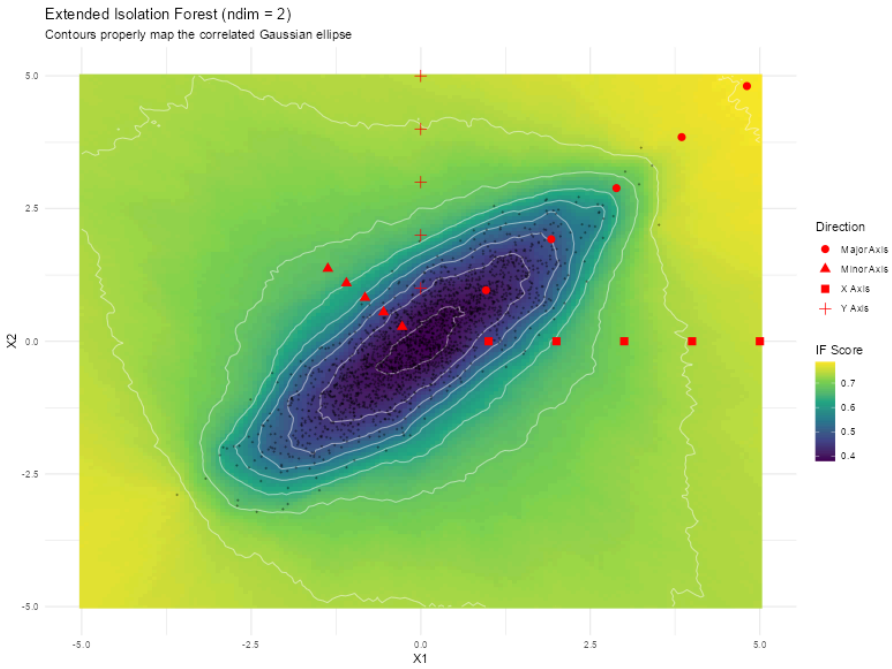
What is an Isolation Forest?

- An isolation forest is built out of isolation trees
- the algorithm takes a bootstrap (or random) sample and chooses a random split point for one variable
- all observations are divided into two nodes based on that variable and split point
- this algorithm is applied recursively to each node until (conceptually) all observations are in a node by themselves
- very many such trees are built, creating a forest
- and one can compute an anomaly score for each element of the data set
- there are very many different suggestions on how to choose the splits
- one can also use an isolation forest to see if other data points are in-liers or out-liers

Isolation Trees and Outliers

- here we show how the isolation forest scores align well with known Gaussian behaviors
- work with Anthony Christidis





Isolation forests and distances

- you can calculate distances between pairs of points based on how many splits it takes to separate them
- or compute a proximity score by looking at how many trees two points share the same terminal node
- these distances can be used for clustering or other purposes

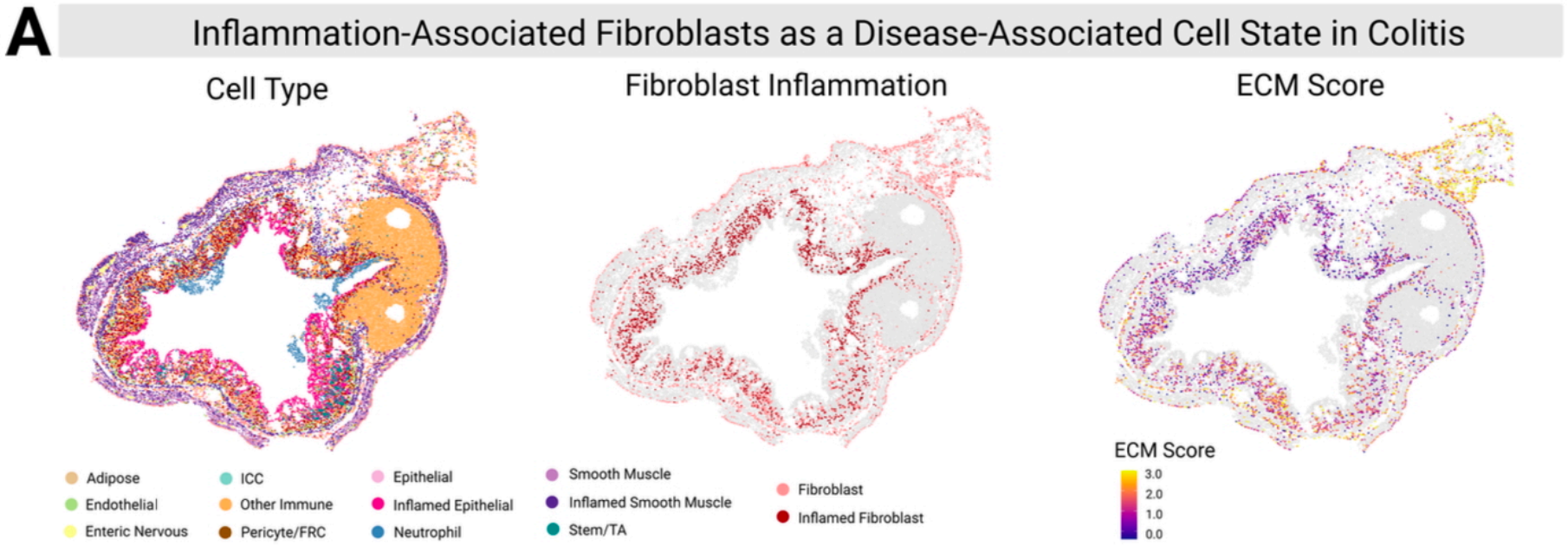
Label Transfer using isolation forests

- if one has a reference data set for cell types, say
- then for each different cell type you can create an isolation forest
- and from the query data set you can test whether a cell is an in-lier or an out-lier for each of the different reference cell types
- this would allow you to identify cells that look different from everything in the reference
- the method is not overly sensitive to the sizes of the reference cells
- the method does not seem to be too sensitive to the number of features used
- work primarily of Anthony Christidis

An example

- we show an example based on the `scDiagnostics` package applied to the MERFISH data from Cadinu et al. Cell 187, 2010–2028 (2024)
- the data profile a mouse model of colitis over a timecourse of 21 days, with peak induction of colitis at Day 9.
- We compared the mouse colon at two timepoints: Day 0 (healthy baseline) and Day 9 (Dextran Sodium Sulfate [DSS]-induced colitis).
- We used Day 0 as *normal* and then compared fibroblasts at day 9 to an isolation forest built on fibroblasts at day 0

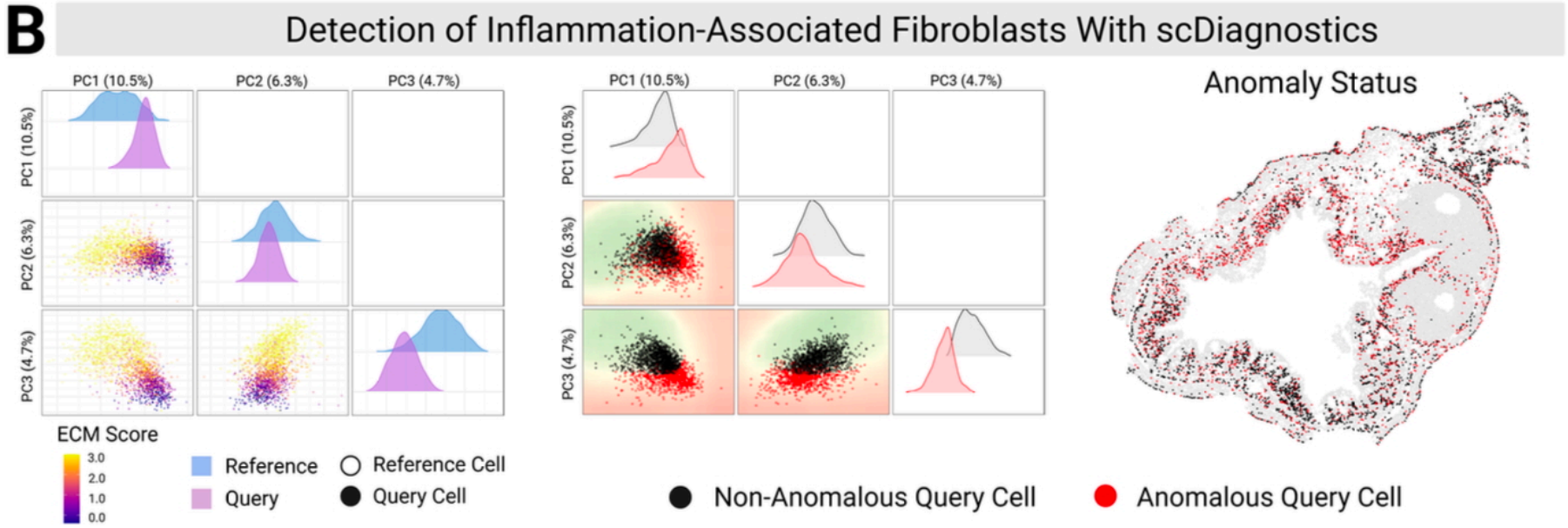
Figure A


A
Inflammation-Associated Fibroblasts as a Disease-Associated Cell State in Colitis
Cell Type
Fibroblast Inflammation
ECM Score
Adipose
Endothelial
Enteric Nervous
ICC
Other Immune
Pericyte/FRC
Epithelial
Inflamed Epithelial
Neutrophil
Smooth Muscle
Inflamed Smooth Muscle
Stem/TA
Fibroblast
Inflamed Fibroblast
ECM Score
3.0
2.0
1.0
0.0

Processing the samples:

- from the Day 0 reference we removed all cells annotated as inflammation-associated (inflammation-associated epithelial cells [IAE], inflammation- associated fibroblasts [IAF], and inflammation-associated smooth muscle cells [IASMC])
- to ensure the reference represented a strictly homeostatic baseline
- we created a we calculated an ECM homeostasis signature score using five canonical matrisome genes: Col1a2, Timp2, Col6a1, Sparc, and Dpt.
- the score was computed as the mean log-normalized expression of these genes, winsorized at 1.5 to reduce the influence of extreme values, and scaled by a factor of 2 for visualization
- we then applied the detectAnomaly function to the fibroblast lineage
- projecting query fibroblasts onto the PCA manifold defined by Day 0 reference fibroblasts and calculating anomaly scores using the isolation forest algorithm with a threshold of 0.5.

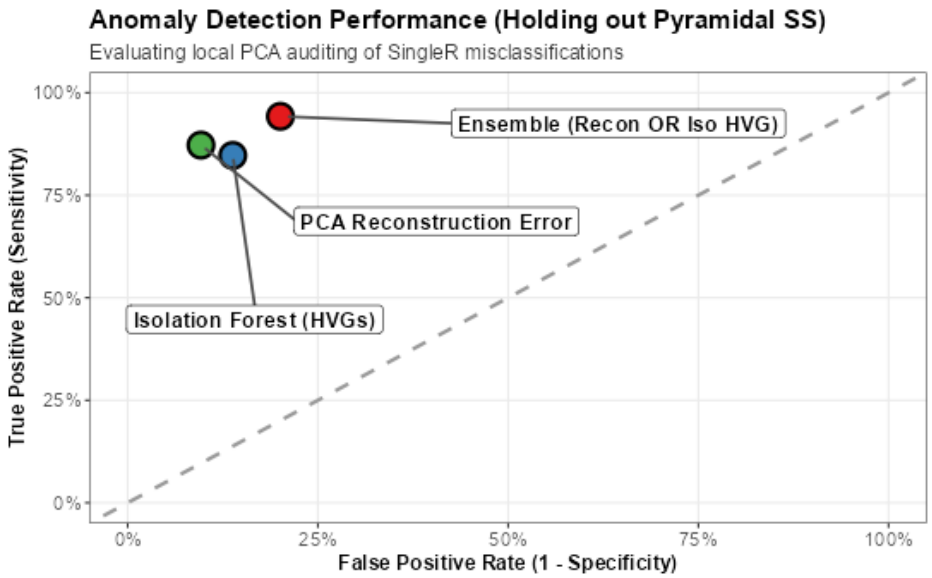
Figure B


B
Detection of Inflammation-Associated Fibroblasts With scDiagnostics
PC1 (10.5%)
PC2 (6.3%)
PC3 (4.7%)
ECM Score
3.0
2.0
1.0
0.0
Reference
Query
Reference Cell
Query Cell
Anomaly Status
Non-Anomalous Query Cell
Anomalous Query Cell

A second example

- Zeisel brain data (https://pubmed.ncbi.nlm.nih.gov/25700174/)
- we removed the pyramidal SS cells from the reference
- then ran SingleR to annotate the full query data,
- almost all of the SS cells in the query were annotated as “pyramidal CA1” cells
- we then took all of those cells and used an isolation forest based assignment
- we also projected the query into both the PC space and the null space for the “pyramidal CA1” cells
- combined these two scores with an “OR”

Anomaly Detection Performance



Thanks

- this was work done while collaborating with the teaching teams at CSAMA (https://csama2025.bioconductor.eu/)
- and Summer School Biological Data Science in Uzhhorod, Ukraine, July 2023
- with lots of input from the students and faculty
- Anthony Christidis, HMS for work on scDiagnostics
- Wolfgang Huber (EMBL)

References

- Modern Statistics for Modern Biology (Holmes and Huber); https://www.huber.embl.de/msmb/
- Elements of Statistical Learning (Hastie, Tibshirani and Friedman) https://hastie.su.domains/Papers/ESLII.pdf
- Modern Applied Statistics with S. Fourth Edition, by W. N. Venables and B. D. Ripley; MASS package in R
- CRAN https://cran.r-project.org/ Task Views
- Wikipedia pages are pretty good for a simple introduction, but often do not provide appropriate guidance/interpretation