Designing your Analysis

By Tallulah Andrews

What is your question?

What data do you need to answer it?

Example Questions:

What progenitor states exist between hematopoietic stem cells and each differentiated blood cell-types?

How do different Huntington gene variants affect neuronal identity/function?

How many cell-types exist in different regions of the liver?

How does the communication between glia and neurons change in Parkinson's disease?

What transcription factor(s) control the differentiation of pancreatic cell-types?

Algorithms/Tools will always give you an answer!

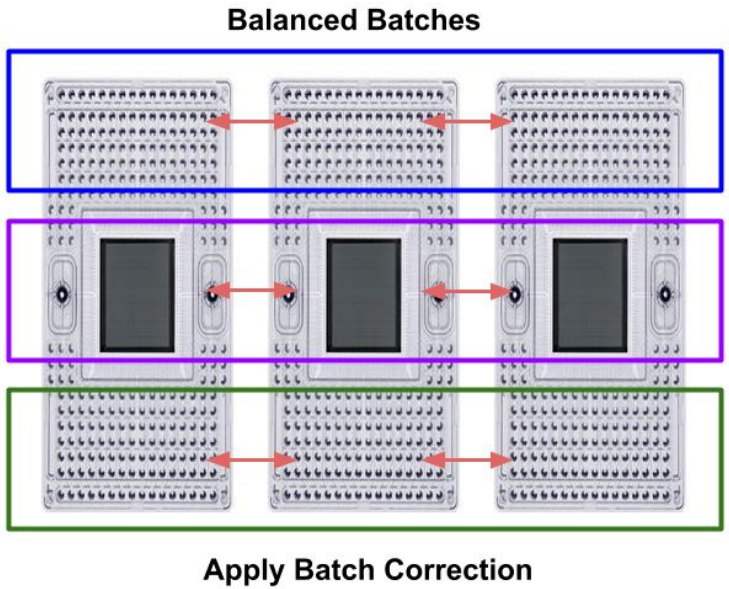
That doesn’t mean that answer is “true”.

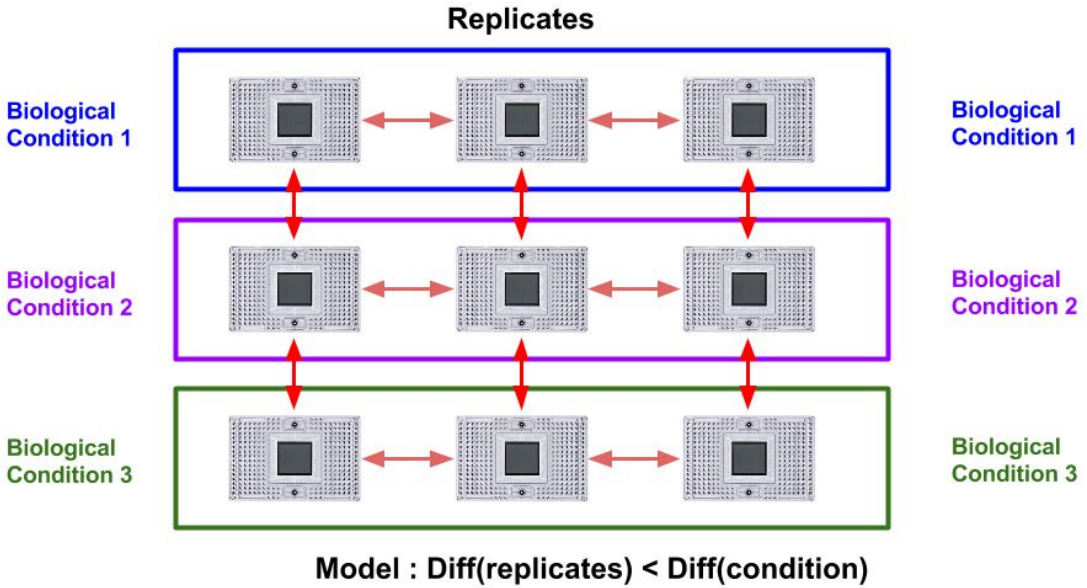
Part 1: When and what to normalize/correct?

Batch Effects / Sample Integration

- Technical Noise (unknown origin)
- Can create false clusters/DE genes in our analysis
- If modelled correctly, helps assess confidence of our conclusions
- Difficult to model because may not be linear / uniform across cell-types

Unconfounded Experiments vs Replicates





Biological “Noise” aka Confounders

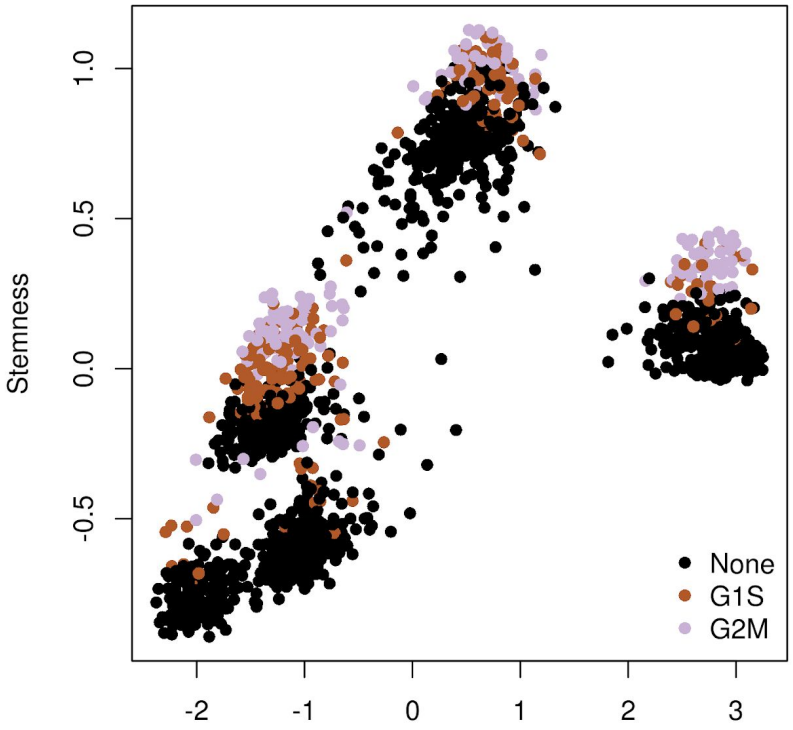
Examples:

- Cell cycle
- Genetic background / individual
- Age
- Circadian rhythm
- Cell stress

Solutions:

- Regress / correct / normalize it
- Exclude cells/genes most affected by it
- Include as a covariate in models
- Ignore it and hope the biology we are interested in still comes through.

The Cell Cycle



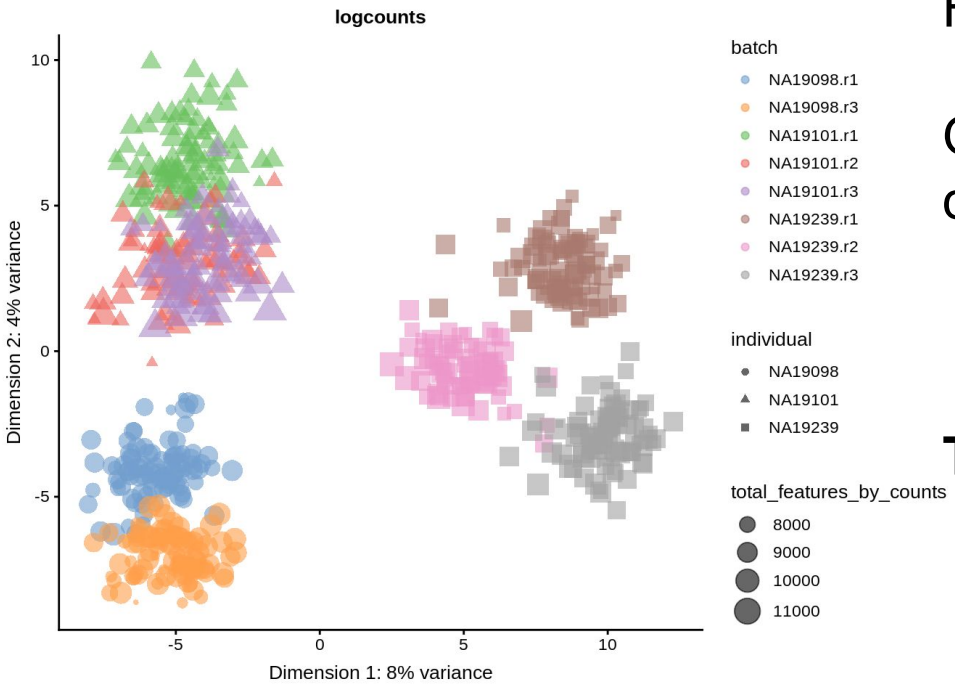
Regressing out the cell-cycle will also remove all variability that is confounded with it.

Development/Differentiation
- Cell cycle may be a confounder

Mature tissue
- Usually cells are not cycling, non-issue

Cancer
- Cell cycle is often biologically interesting

Individual Variation



Human / Patient samples

Often must be confounded with disease or treatment conditions

- 3 patients with diabetes vs 3 patients without

Treated similar to batch effects.

Stress Response

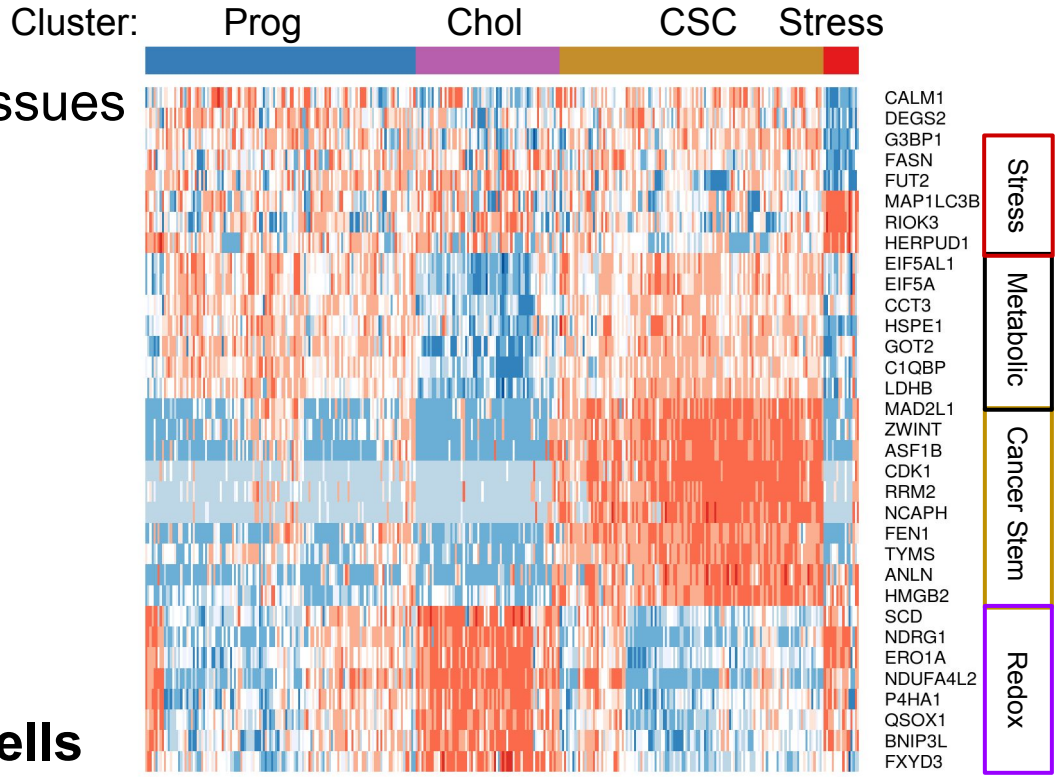
Common when dissociating tissues

Highly cell-type dependent

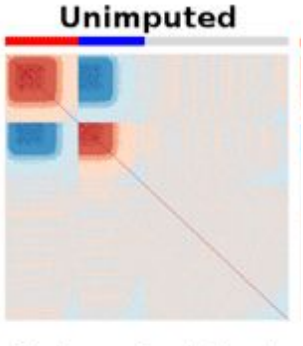
May depend on condition / replicate

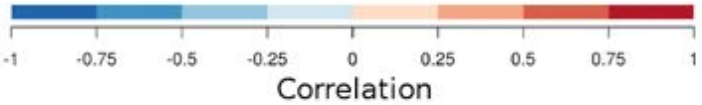
Often non-linear -> Difficult to regress

Typically exclude affected cells



When to impute/smooth data? - ***Visualization only***





Part 2: Clustering vs Pseudotime

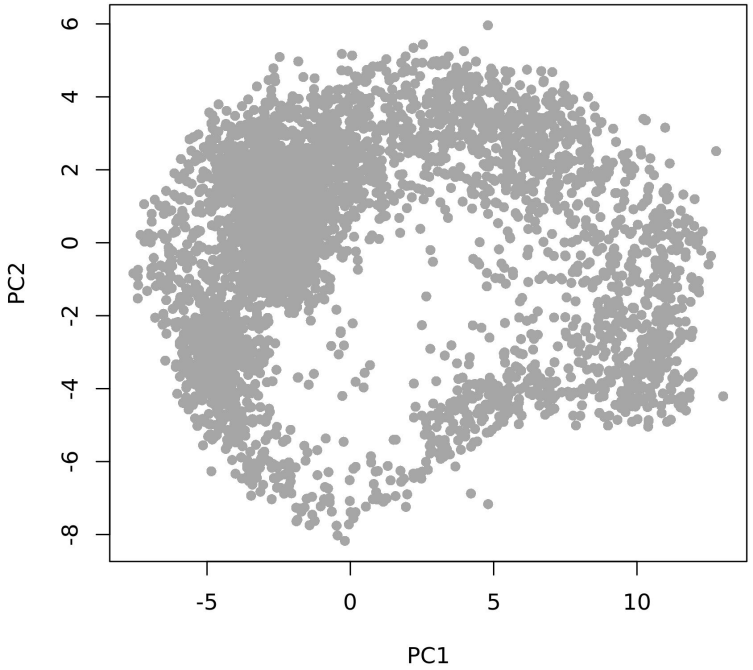
You shouldn’t be doing both.

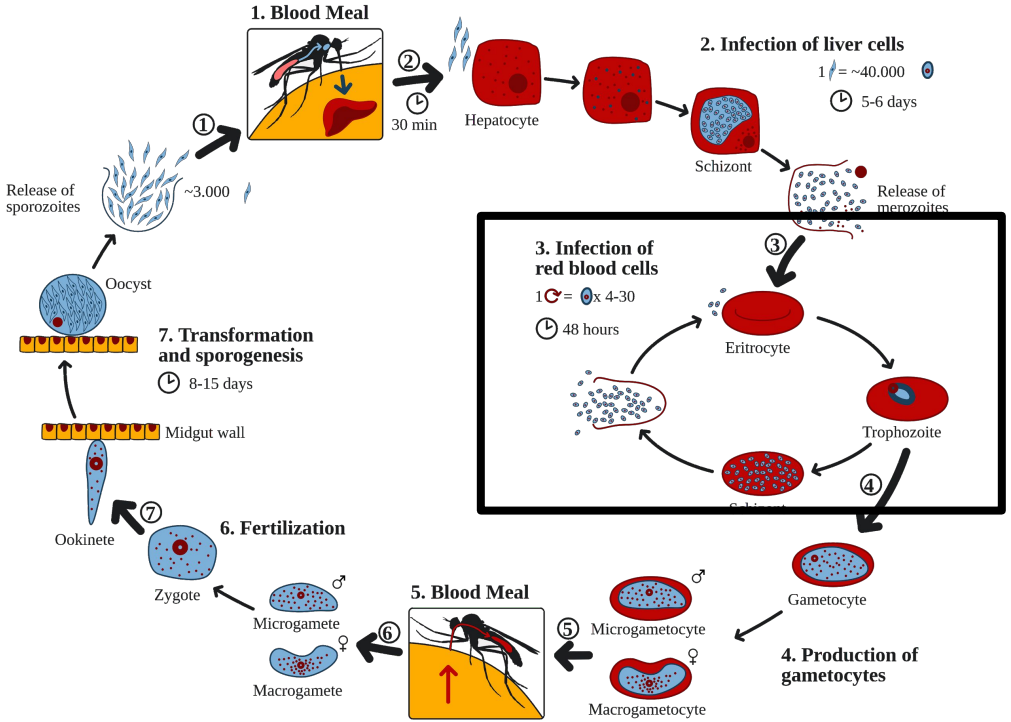
Part 2: Clustering vs Pseudotime

You shouldn’t be ~~doing~~ publishing both.

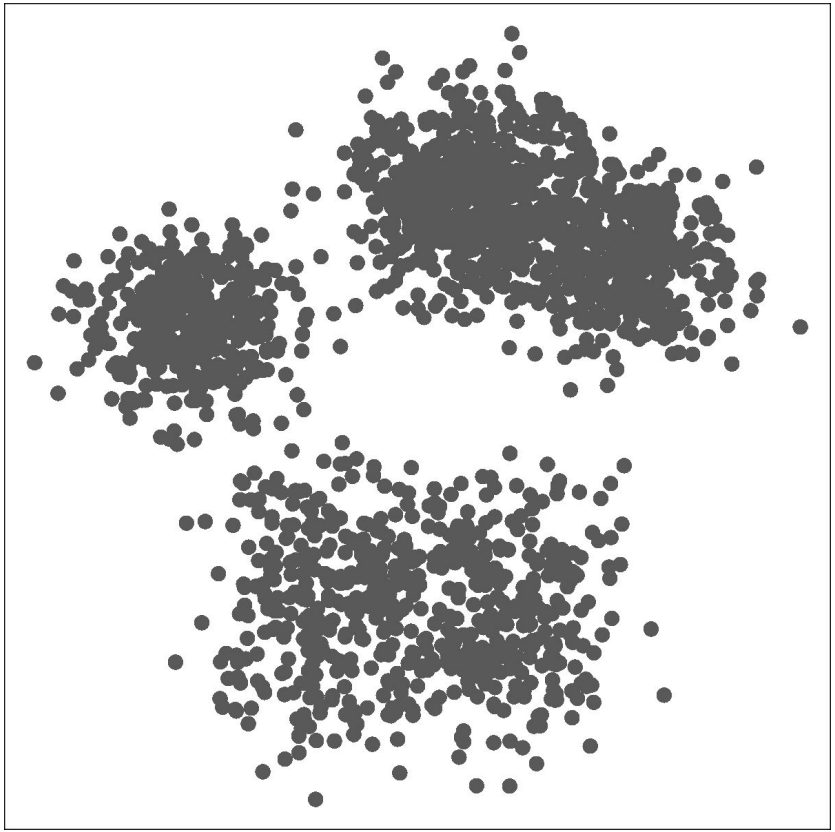
(usually)

Case Study: Malaria Cell Atlas





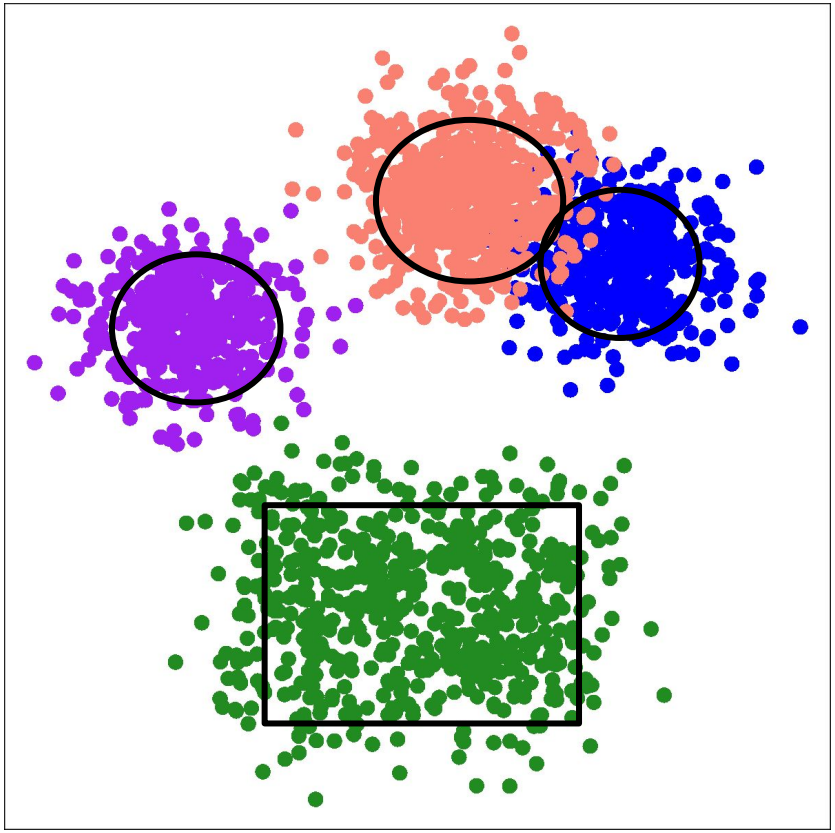
Problem: What is a “good” cluster?



How many clusters are there in this data?

Pause the video and decide on your answer

The Truth



Clustering: Common Assumptions

- Clusters are roughly the same size

May fail to detect a rare cell-types

Clustering: Common Assumptions

- Clusters are roughly the same size
  - May fail to detect a rare cell-type
  - **May arbitrarily split up large clusters**

Clustering: Common Assumptions

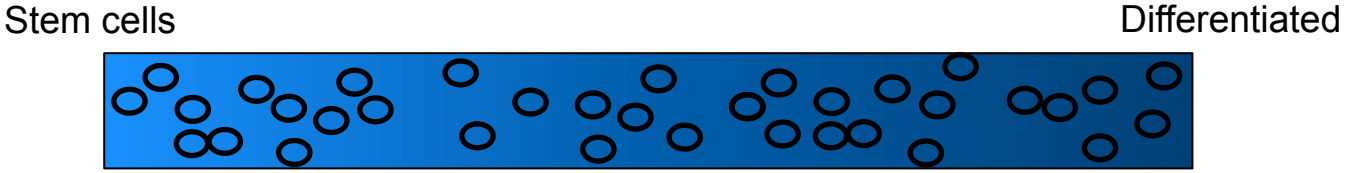
- Clusters are roughly the same size
  - May fail to detect a rare cell-type
  - May arbitrarily split up large clusters

- Clusters are roughly the same density

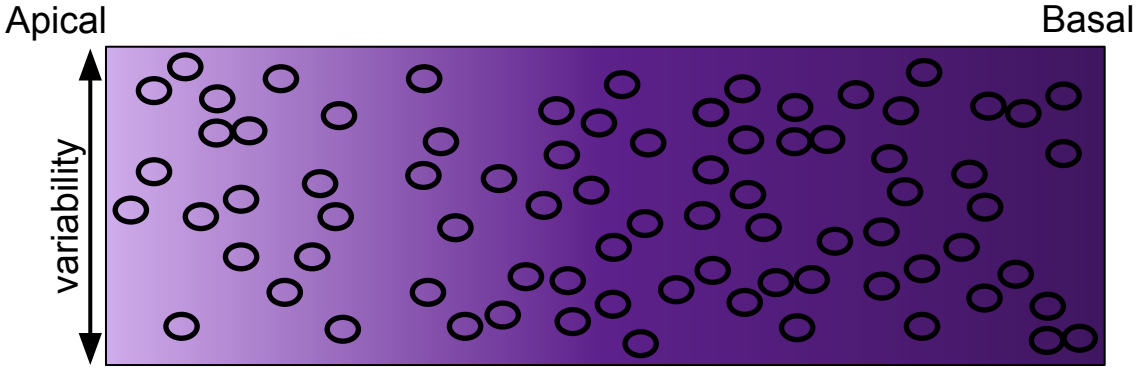
Noisy / sparse clusters often split up into many smaller clusters

What happens if you use a clustering method on a gradient?

Gradient 1:



Gradient 2:

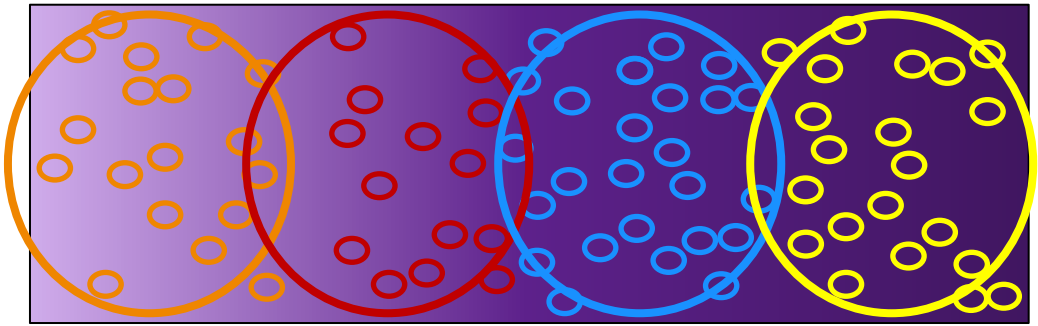


Pause the video and decide on your answer

What happens if you use a clustering method on a gradient?

Gradient 1:

Gradient 2:



Trajectories are split up along their length in a way to optimize uniformity of clusters.

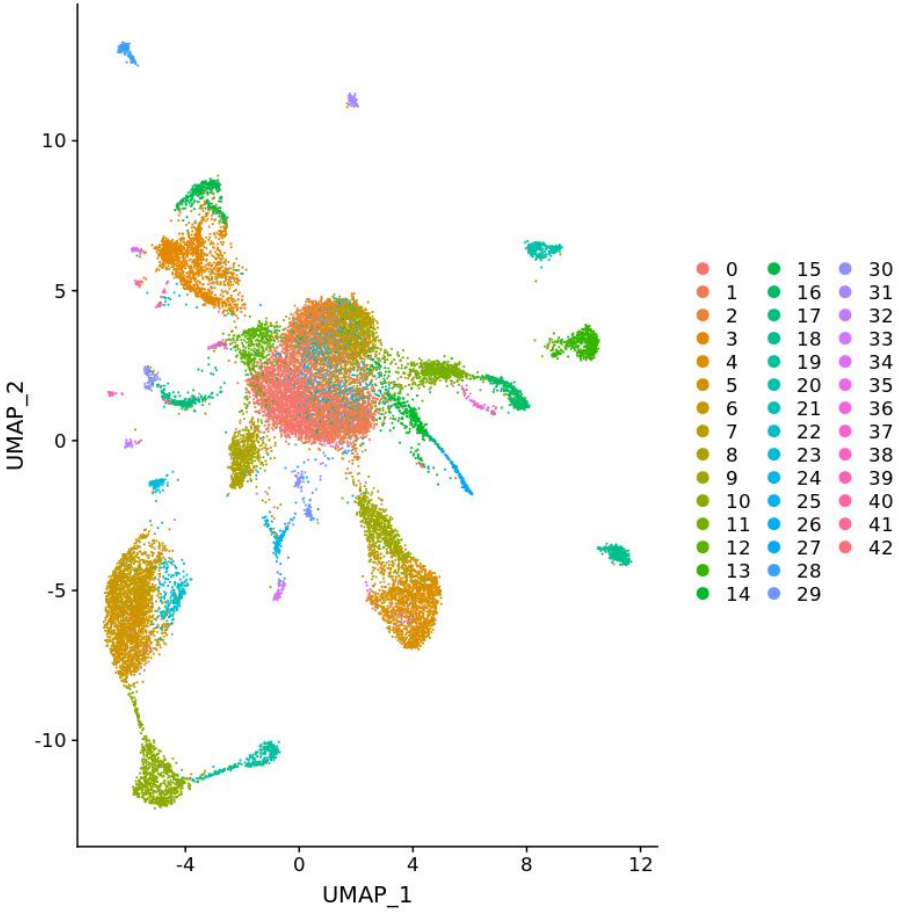
Clustering: Common Assumptions

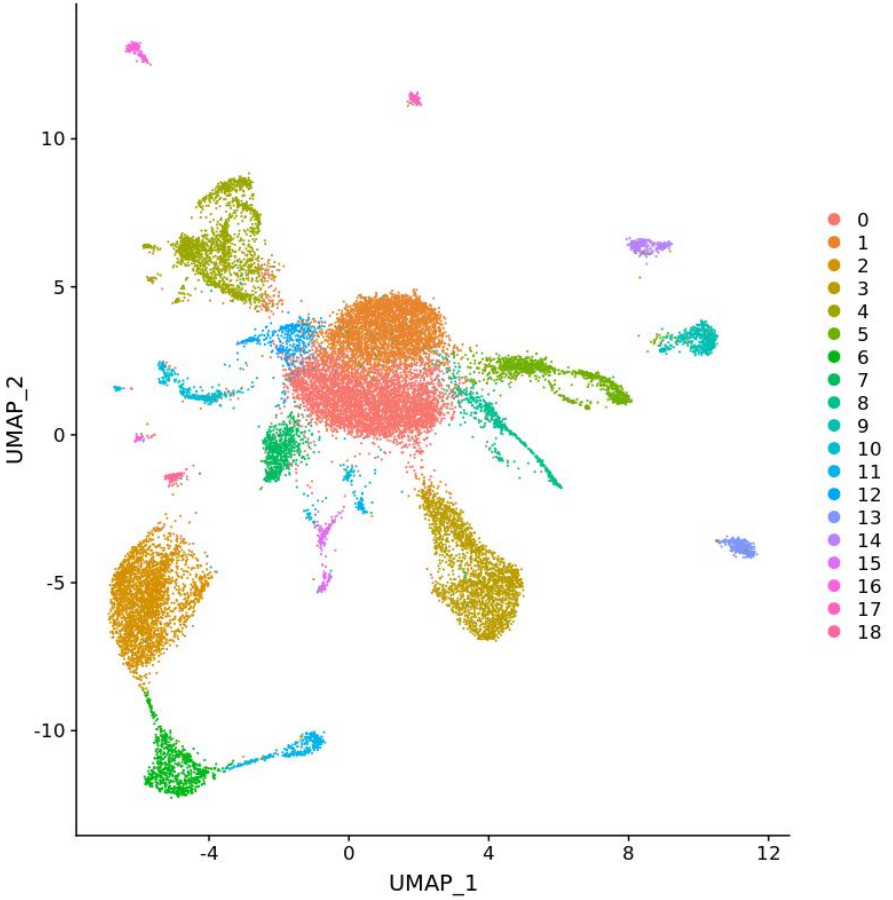
- Clusters are roughly the same size
  - May have difficulty detecting a rare cell-type
  - May arbitrarily split up large clusters

- Clusters are roughly the same density
  - Noisy / sparse clusters often split up into many smaller clusters

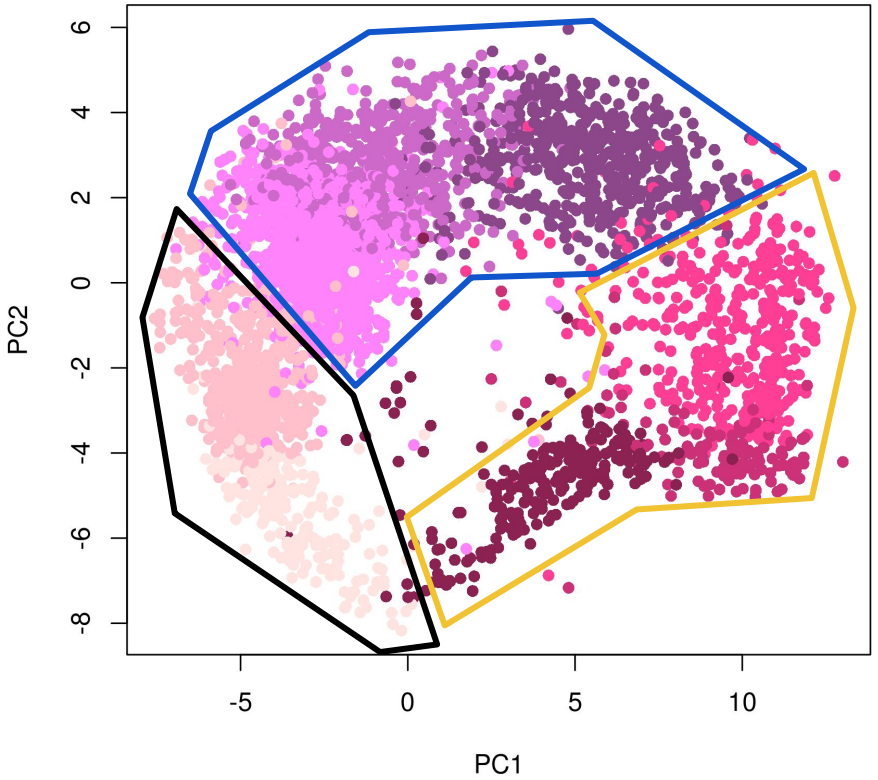
- Clusters exist at a fixed & predetermined resolution

Clustering the same data with different parameters:





Seurat Applied to MCA data:



Morphologically there are 3 distinct stages of parasites, using marker genes we could link these clusters to them.

Are the subdivisions actually meaningful?

Or is this a smooth continuum that has been arbitrarily divided up by a clustering algorithm?

Clustering: How do you know a cluster is “real”?

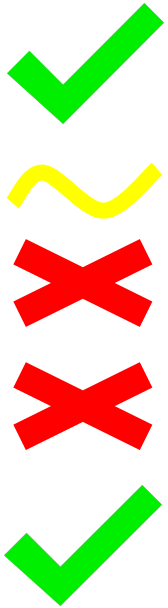
1. Robustness to clustering method/parameters
2. Significant marker genes / differentially expressed genes
3. Known marker genes
4. Q
5. C

   e.g. scmap
6. Spatial structure
7. Experimental validation

Seeing that they “look good” in a visualization is not a good way to validate clustering

Malaria Cell Atlas Data:

1. Robustness to clustering method/parameters
2. Significant marker genes / differentially expressed genes
3. Known marker genes
4. Quality statistics:
    a. e.g. Silhouette Index
5. Consistency across experimental replicates or with reference data
    a. e.g. scmap
6. Spatial structure
7. Experimental validation

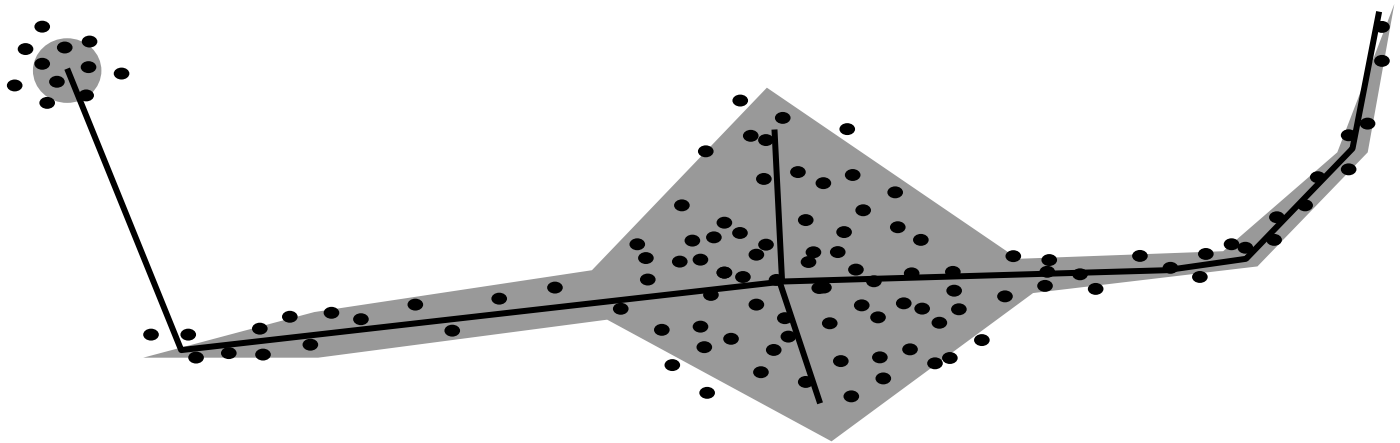


NA

NA

Pseudotime: Common Assumptions

- **All cells exist on a smooth continuum**
  - Cells belonging to a distinct cluster create a false branch
  - Varying thickness (cell density) may be mistaken for a branch

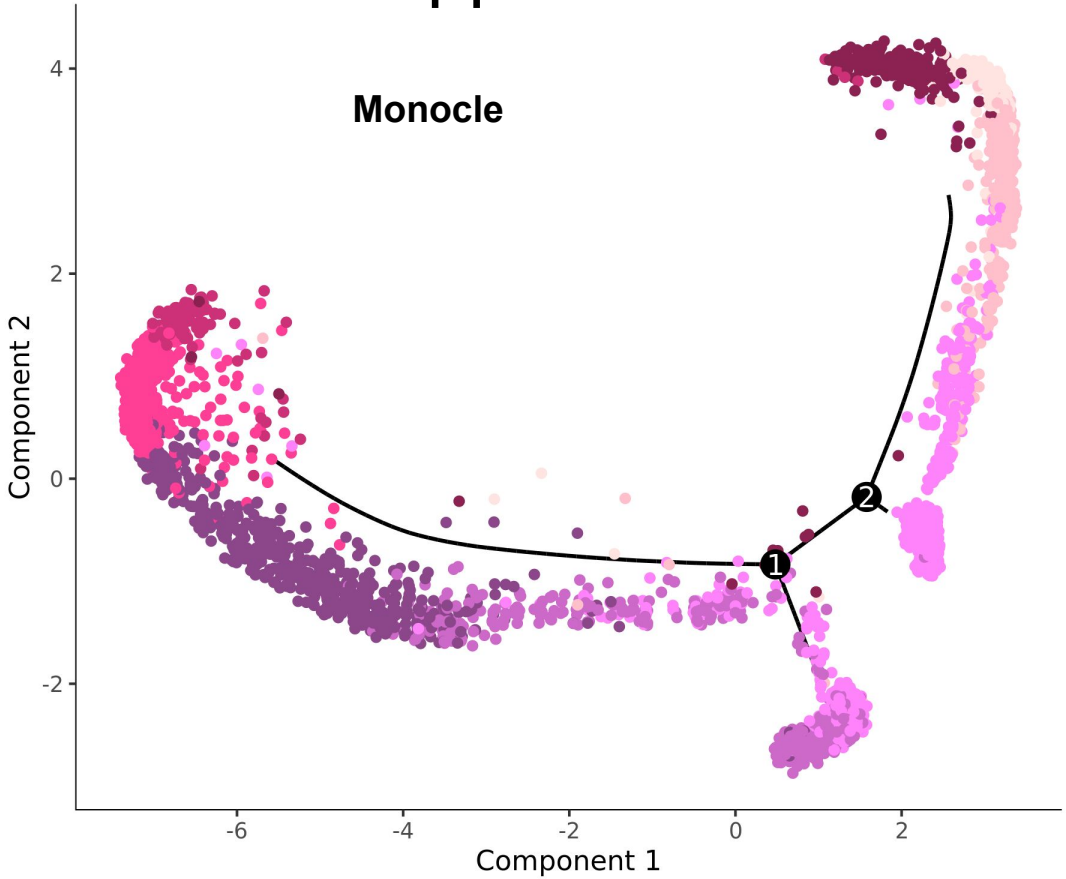


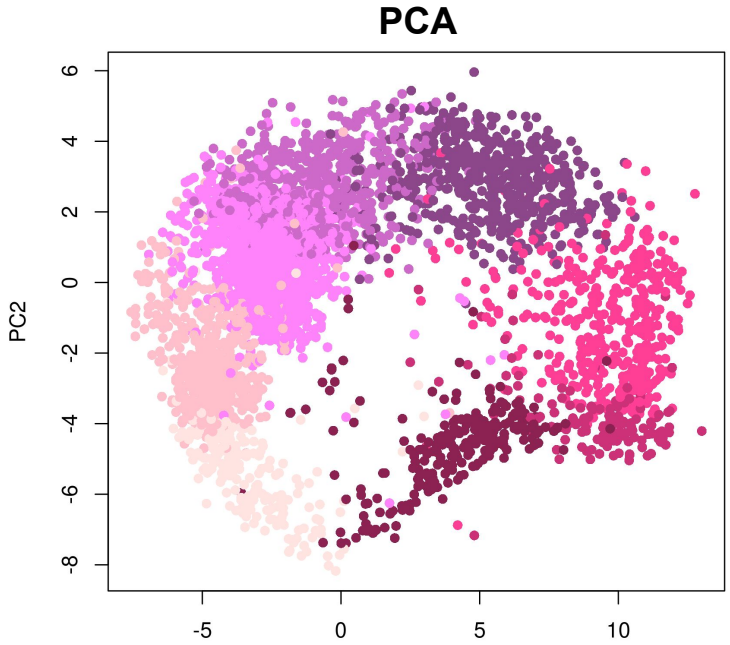
Pseudotime: Common Assumptions

- All cells exist on a smooth continuum
  - Cells belonging to a distinct cluster create a false branch
  - Varying thickness (cell density) may be mistaken for a branch

- **The continuum is a line/curve/tree and has a start and an end**
  - Cycles fail to be properly modelled
  - Multiple overlapping gradiants create “shapes” (e.g. triangle) which may be forced into a “Y” shaped topology
  - If you specified the number of branches/clusters the tool will find that number even if that is not correct.

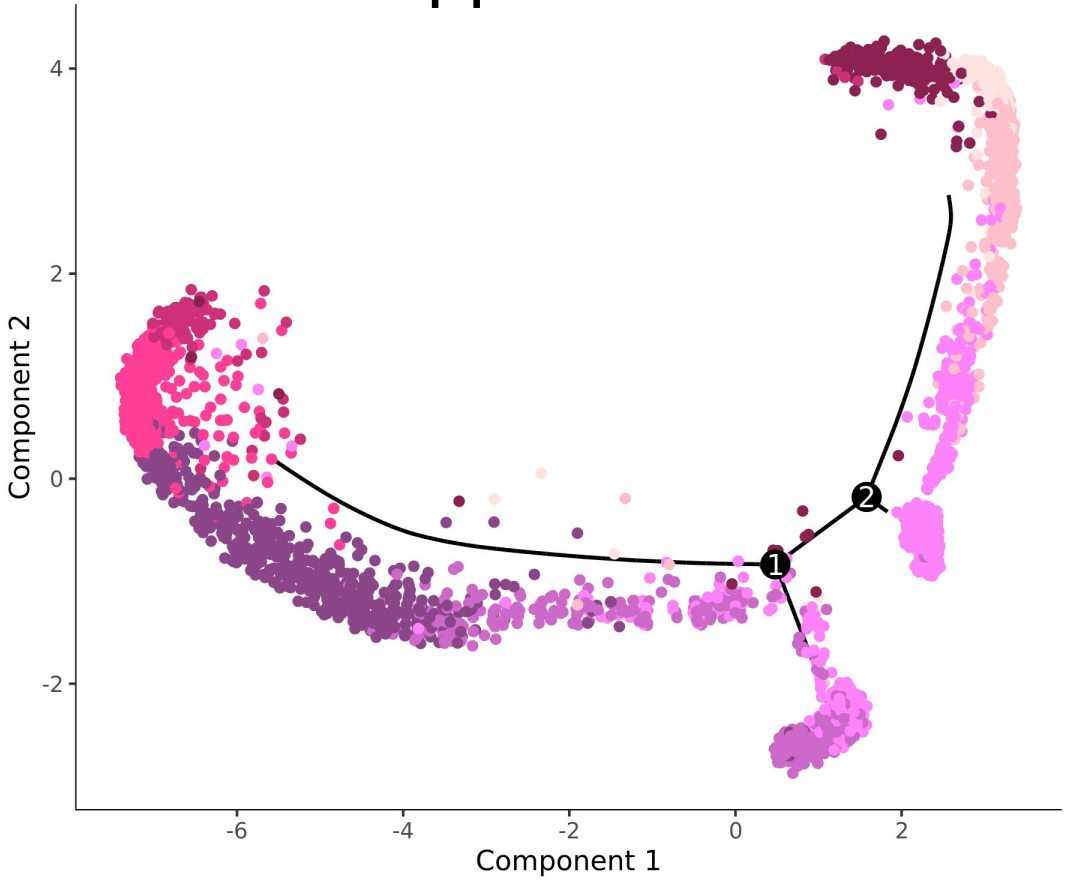
Monocle Applied to MCA data: What happened?





Pause the video and decide on your answer

Monocle Applied to MCA data:



Errors due to model assumptions:

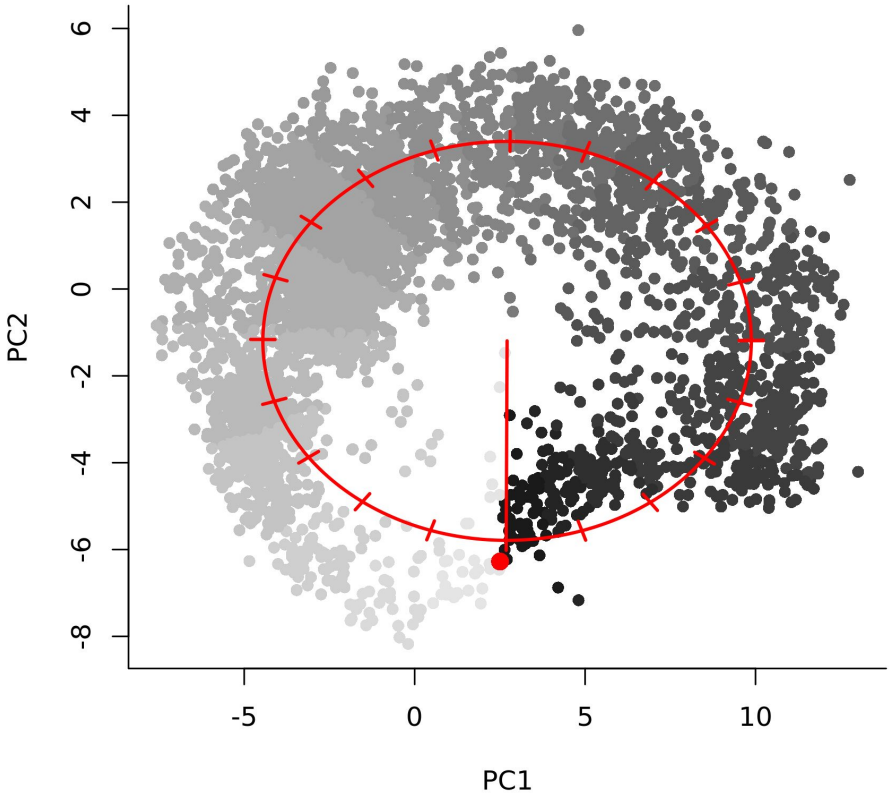
1. Arbitrarily cut the circle open

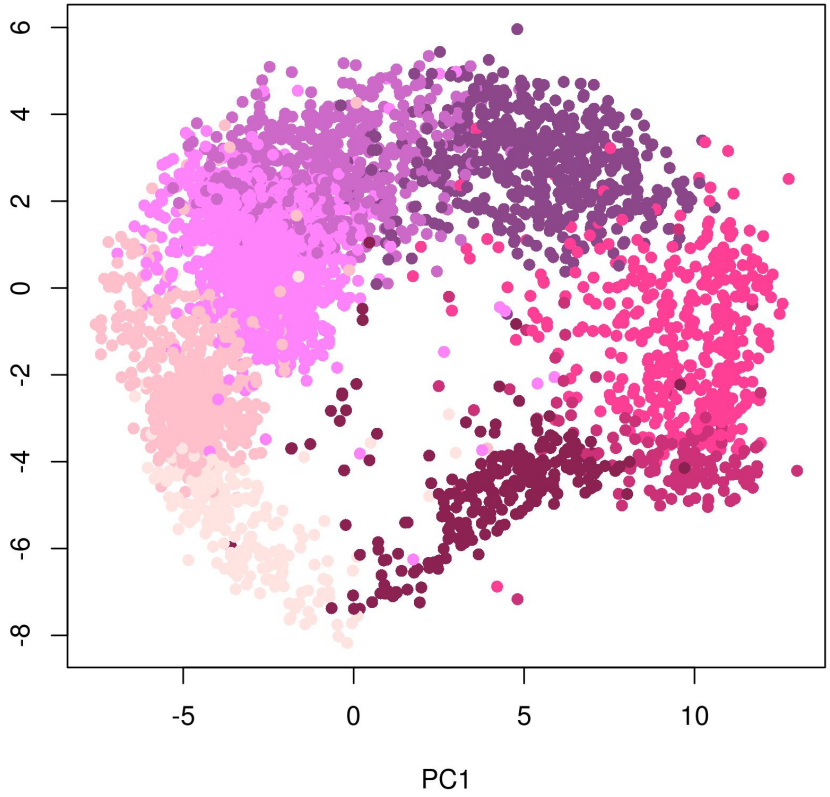
2. False branches at most heavily sampled portion of the cycle

Danger!

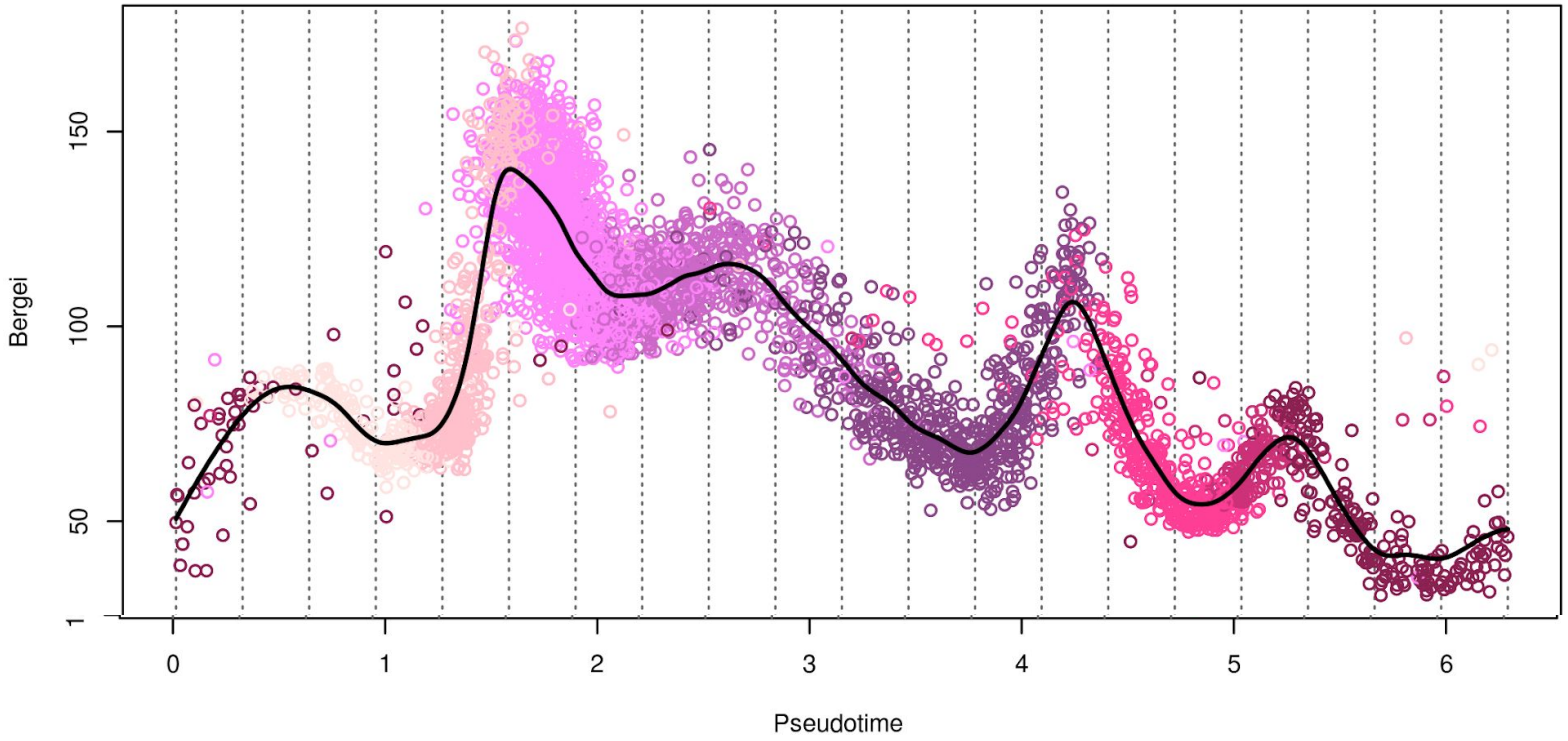
We could have interpreted one/both of the branches as cells moving in/out of the IDC

Malaria Cell Atlas - what did we do? (both)





Why did we end up doing both?

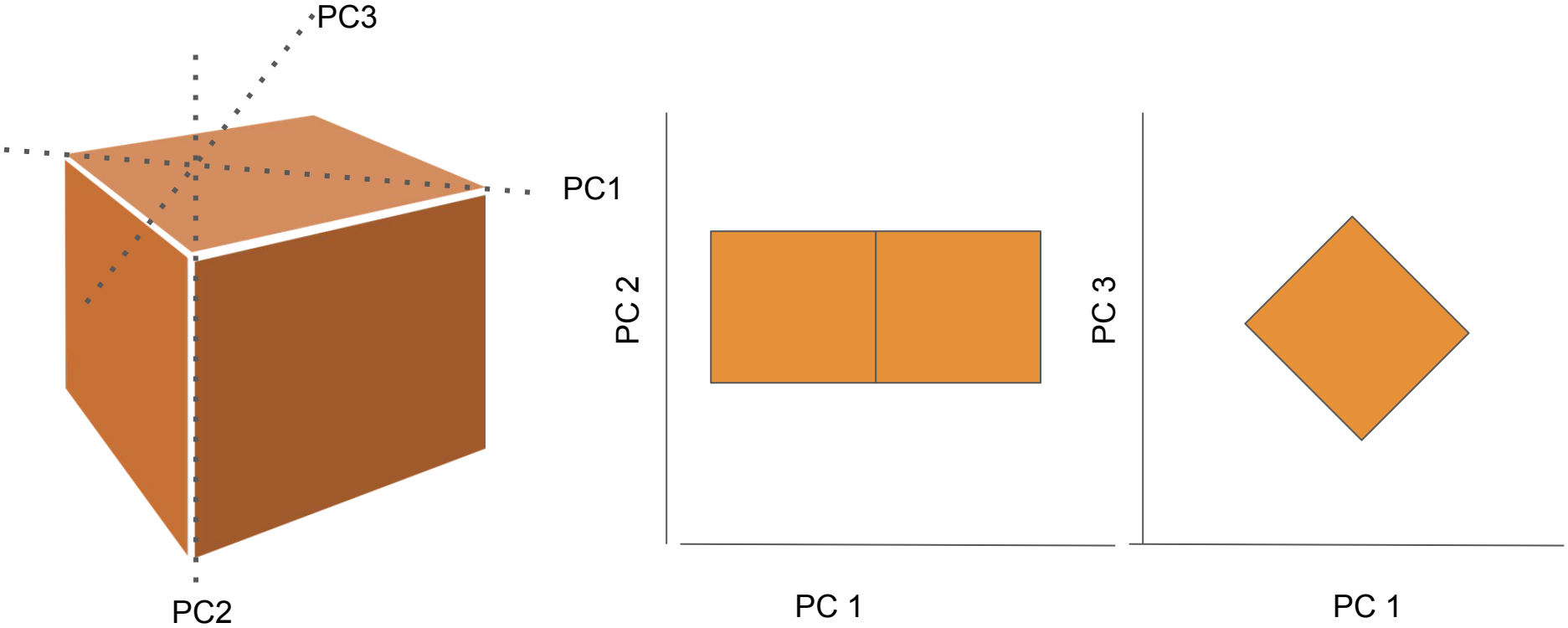


Part 3: Visualization

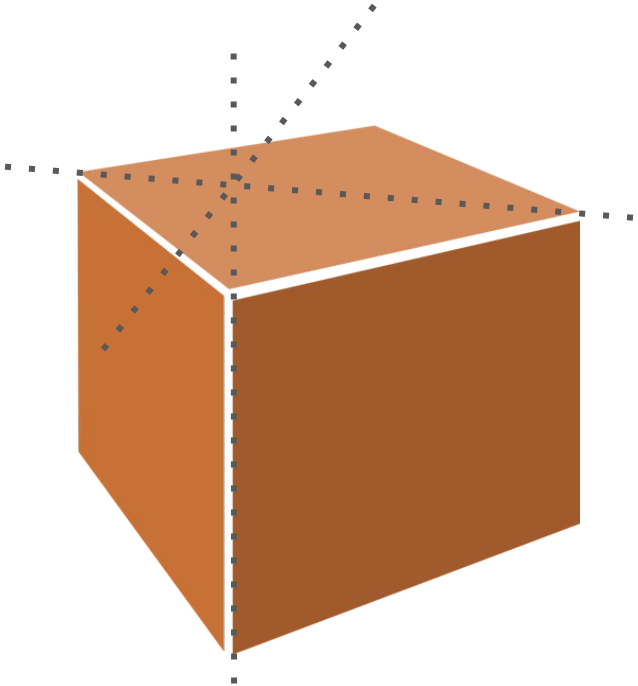
All models are wrong, some are useful *-George Box*

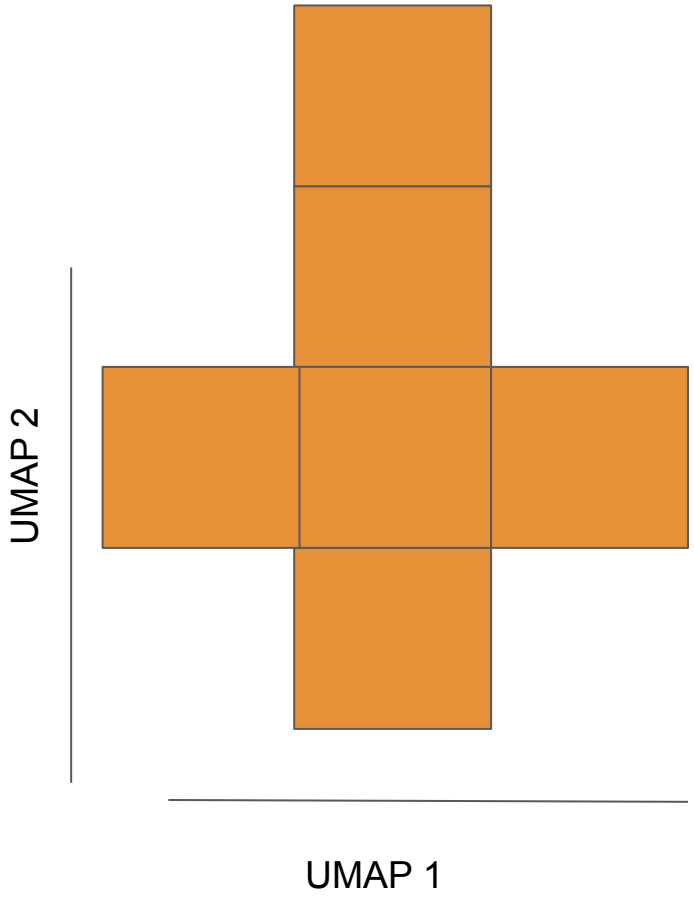
Visualizations

Visualization: Box Analogy



Visualization: Box Analogy





Visualization: What do you choose to show?

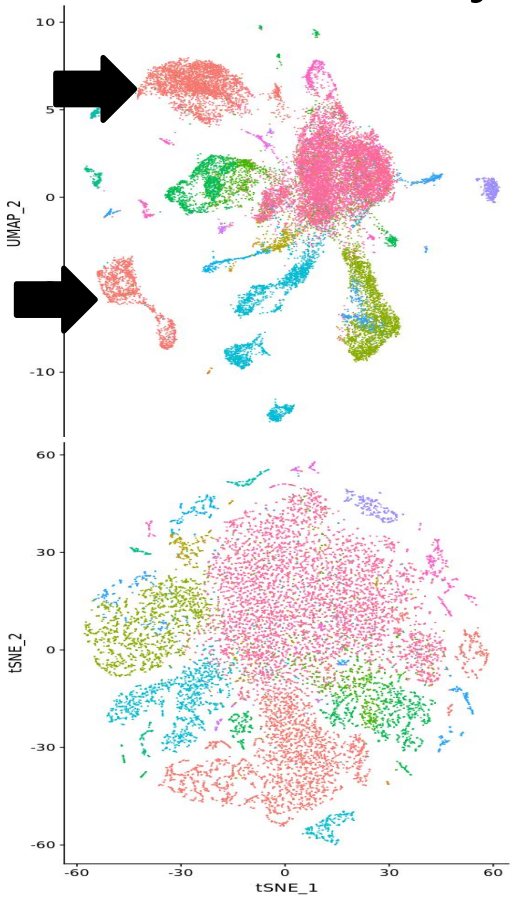
Accurate distances between points. - PCA

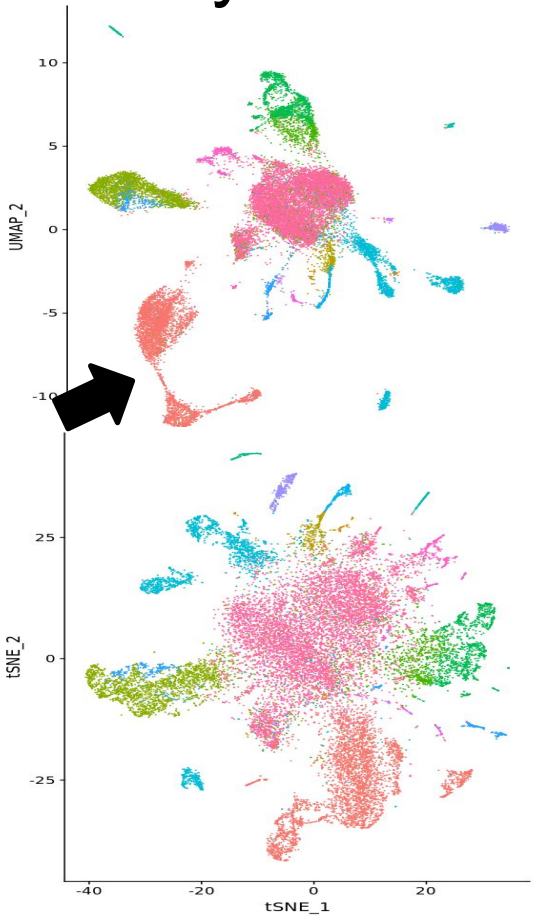
Overall structure - UMAP

Distinct Clusters - t-SNE

Trajectories/gradiants - Diffusion Map

How many clusters can you see?





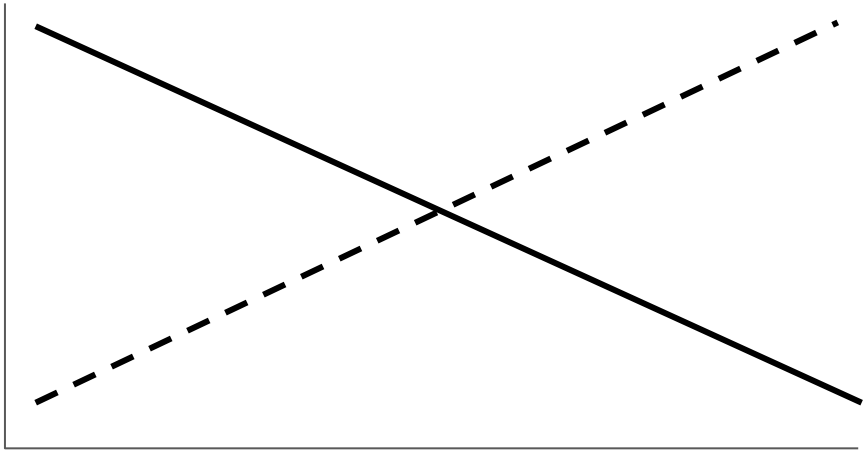
UMAP

tSNE

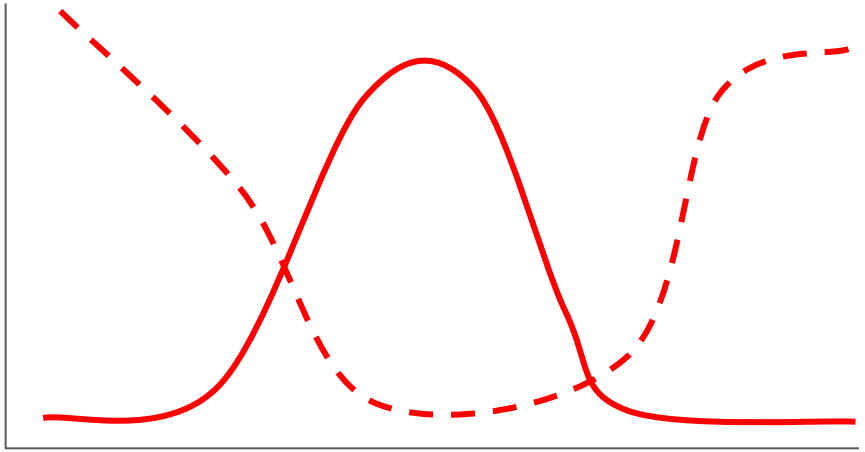
Part 4: Differential Expression

General Linear Models: Common Assumptions

- Changes are Linear
  - Oscillatory patterns will not be detected!



Pseudotime



Pseudotime

General Linear Models: Common Assumptions

- Changes are Linear
  - Oscillatory patterns will not be detected!

- Noise follows a specific distribution
  - Negative Binomial - Suitable for UMI counts
  - Zero Inflated Negative Binomial - Suitable for UMI counts or Read counts

General Linear Models: Common Assumptions

- Changes are Linear
  - Oscillatory patterns will not be detected!

- Noise follows a specific distribution
  - Negative Binomial - Suitable for UMI counts
  - Zero Inflated Negative Binomial - Suitable for UMI counts or Read counts

- Changes in Expression caused by different factors are Additive
  - i.e. batch effects are the same across cell-types

Non-parametric Tests

- Changes are Monotonic
  - Oscillatory patterns will not be detected!

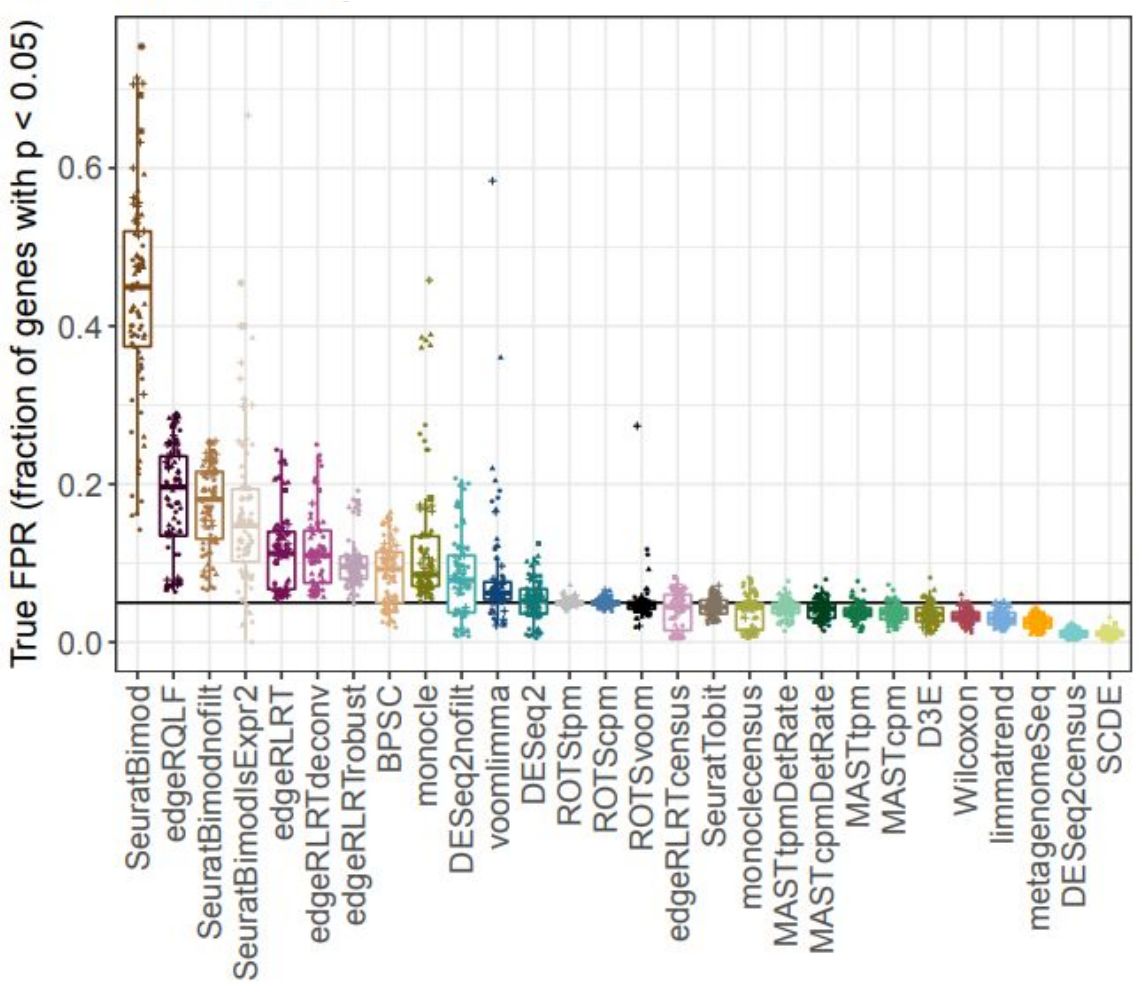
- Independent of distribution of data values
  - Can be applied to batch-corrected, scaled, normalized, regressed data

- Do not account for confounding factors contributing to differences in expression

What is the “best”?

Conclusions:

- Seurat performs poorly
- Both GLM and non-parametric tests work well when used correctly
- **MAST** is best GLM
- **Wilcoxon** is best nonparametric



Soneson and Robinson. 2018. Nature Methods. 15 : 255-261 doi: 10.1038/nmeth.4612

DE across multiple batches and multiple conditions

- Subset your data to a single cell-type, then perform a DE test:
- General Linear Model
  - Include biological condition(s) of interest as a predictor.
  - Include batch ID as a random effect.
  - MAST or NEBULA
- Pseudobulks
  - Sum raw counts of cells in each batch x condition. These are “pseudobulk” samples.
  - Use existing bulk RNAseq DE tools on the pseudobulks (e.g. edgeR).

Analysis of complex experiments

How would you test the following questions? (Hint: there is more than one good answer)

1) Effect of stimulation on gene expression in three batches of T-cells, each batch contains 50% stimulated and 50% unstimulated.

2) Cell-type specific effects of diabetes in the pancreas, three replicates were performed for diabetic and non-diabetic pancreas samples.

Pause the video now to write down your answers
(15 minutes allotted)

Effect of stimulation on gene expression in three batches of T-cells, each batch contains 50% stimulated and 50% unstimulated.

	Stimulated	Unstimulated
Batch 1	150	150
Batch 2	150	150
Batch 3	150	150

Example Analysis:

1. Assign cells from each batch to original sample
2. Batch Correction
3. Clustering & Marker detection (identify sub-types of T-cells)
4. Differential expression with Wilcox-rank-sum test for each cluster before/after stimulation

Effect of stimulation on gene expression in three batches of T-cells, each batch contains 50% stimulated and 50% unstimulated.

	Stimulated	Unstimulated
Batch 1	150	150
Batch 2	150	150
Batch 3	150	150

Analysis Option 2:

1. Assign cells from each batch to original sample
2. Clustering & Marker detection (identify sub-types of T-cells)
3. Differential expression with a GLM including Batch and Stimulation as predictors.

Cell-type specific effects of diabetes in the pancreas, three replicates were performed for diabetic and non-diabetic pancreas samples.

	Diabetic	Healthy
Donor 1	1500	0
Donor 2	1500	0
Donor 3	1500	0
Donor 4	0	1500
Donor 5	0	1500
Donor 6	0	1500

Example Analysis:

1. Cluster each sample separately
2. Match clusters across donors
3. Use a GLM with Donor as a random effect, and Diabetic/Healthy as the predictor

Cell-type specific effects of diabetes in the pancreas, three replicates were performed for diabetic and non-diabetic pancreas samples.

	Diabetic	Healthy
Donor 1	1500	0
Donor 2	1500	0
Donor 3	1500	0
Donor 4	0	1500
Donor 5	0	1500
Donor 6	0	1500

Analysis Option 2:

1. Integrate samples together without correcting batch effects
2. Cluster on Integrated data
3. Create mini-bulks by summing raw counts for cells from each donor in each cluster.
4. Perform traditional bulk-RNAseq DE within each cell-type.

END

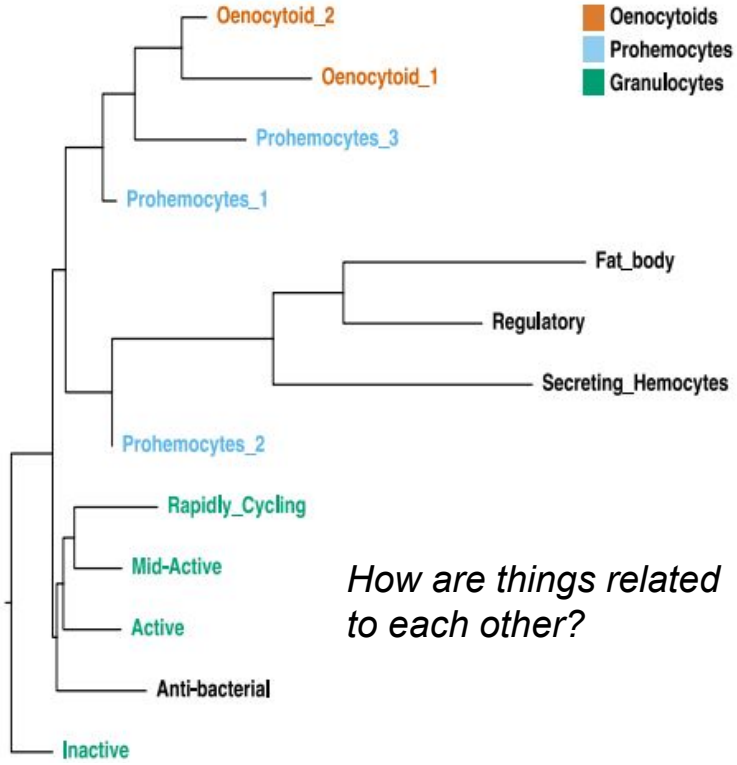
Types of questions

- Descriptive Analysis (one condition):
  - How many cell-types are there? What are they?
  - Is Gene X expressed in a particular cell-type?
  - Are there developmental trajectories/gradients?
  - Where is this cell-type located?

What is this?

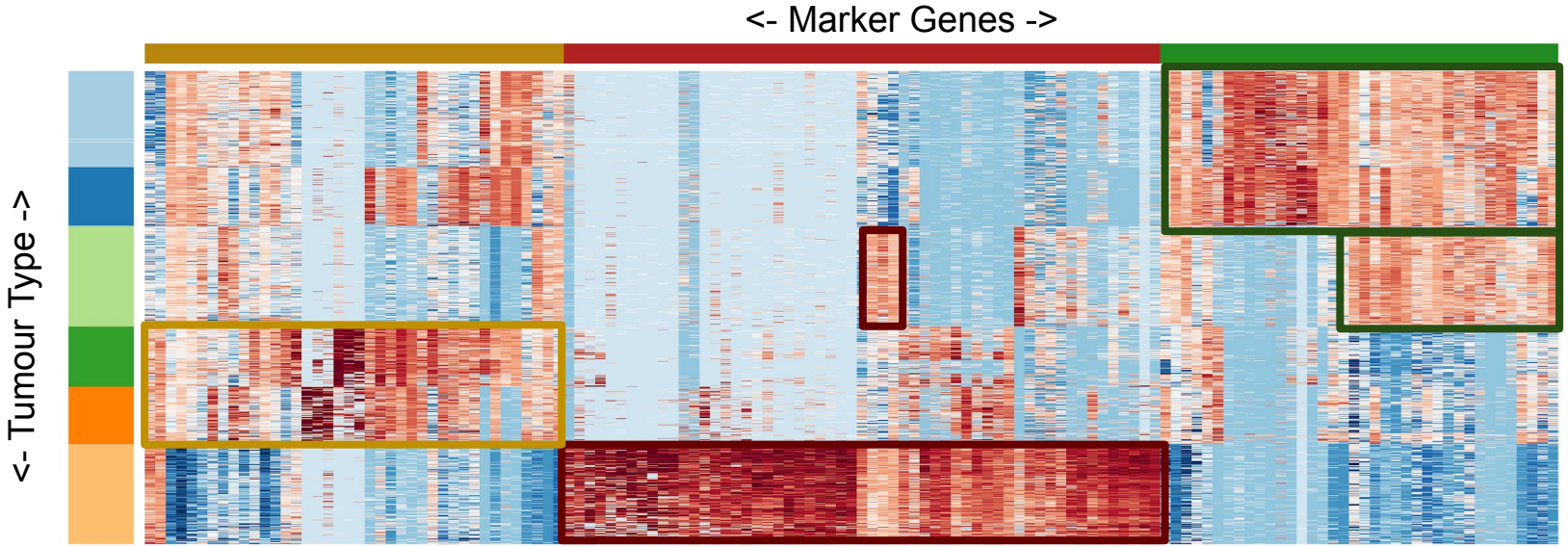
Types of questions

- Relational (one condition, multiple cell-types)
  - Which cell-types are similar to other cell-types?
  - Does cell-type A communicate with cell-type B?
  - Is cell-type A a progenitor of cell-type B?



How are things related to each other?

Types of questions



How do things change when X happens?

- Comparative Analyses (multiple conditions):
  - How does this cell-type or developmental process change during disease?
  - In response to a drug/mutation?
  - When in contact with a different cell-type?

Types of questions

- Descriptive Analysis (one condition):
  - How many cell-types are there? What are they?
  - Is Gene X expressed in a particular cell-type?
  - Are there developmental trajectories/gradients?
  - Where is this cell-type located?
- Relational (one condition, multiple cell-types):
  - Which cell-types are similar to other cell-types?
  - Does cell-type A communicate with cell-type B?
  - Is cell-type A a progenitor of cell-type B?

Most current tools exist in this space.

- Comparative Analyses (multiple conditions):
  - How does this cell-type or developmental process change during disease?
  - In response to a drug/mutation?
  - When in contact with a different cell-type?

Differential Expression

Interactive Exercise: Research Questions

For each situation:

- What type of question is being asked?
- What data do you need to answer the question?

Pause the video to write down your answers before continuing.

Exercise Situations:

(1) What progenitor states exist between hematopoietic stem cells and each differentiated blood cell-types?

(2) How do different Huntington gene variants affect neuronal identity/function?

(3) How many cell-types exist in different regions of the liver?

(4) How does the communication between glia and neurons change in Parkinson's disease?

(5) What transcription factor(s) control the differentiation of pancreatic cell-types?

Pause the video now to write down your answers
(5 minutes allotted)

Exercise Answers: (1)

What progenitor states exist between hematopoietic stem cells and each differentiated blood cell-types?

- What type of question is being asked?

Relational & Descriptive

- What data do you need to answer the question?

Single cell expression for one sample of differentiated blood cells, and from bone marrow and all other sites of blood cell maturation.

Exercise Answers: (2)

How do different Huntington gene variants affect neuronal identity/function?

- What type of question is being asked?

Comparative

- What data do you need to answer the question?

Single cell RNAseq from brain samples or brain organoids from multiple samples carrying different Huntington gene variants.

(Bulk RNAseq of sorted neurons is also valid)

Exercise Answers: (3)

How many cell-types exist in different regions of the liver?

- What type of question is being asked?

Descriptive (perhaps Comparative)

- What data do you need to answer the question?

Single cell RNAseq from one or more different regions a liver.

Exercise Answers: (4)

How does the communication between glia and neurons change in Parkinson's disease?

- What type of question is being asked?

Comparative & Relational

- What data do you need to answer the question?

Single-cell RNAseq from both glia and neurons in normal samples and samples affected by Parkinson's disease.

Exercise Answers: (5)

What transcription factor(s) control the differentiation of pancreatic cell-types?

- What type of question is being asked?

Relational & Descriptive

- What data do you need to answer the question?

Single-cell RNAseq at multiple time points during pancreatic development or *in vitro* differentiation