VEuPathDB
Eukaryotic Pathogen, Vector & Host Informatics Resources

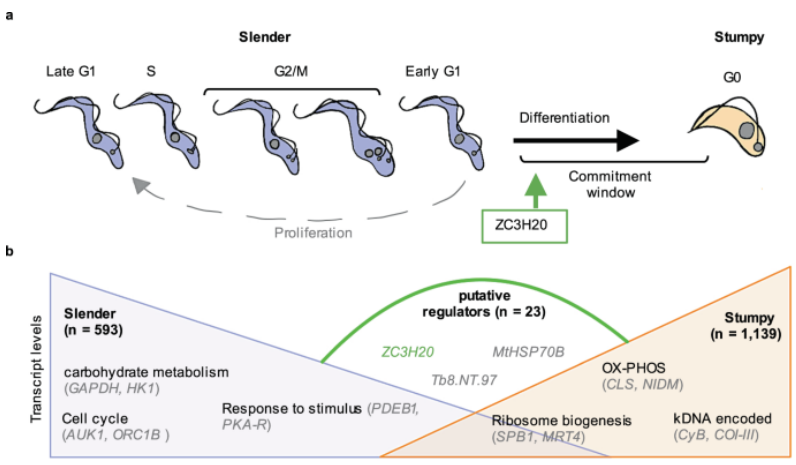
Mining Single Cell RNA-Sequencing (scRNA-seq) Data

Learning objectives

- Find all genes with data from scRNA-seq experiments
- Explore scRNA-seq data on specific gene pages
- Explore scRNA-seq data using the cellxgene application

Introduction

Single-cell RNA sequencing (**scRNA-seq**) offers major advantages over traditional bulk RNA-seq because it measures gene expression in individual cells rather than providing an averaged profile across an entire population. This cell-level resolution makes it possible to identify heterogeneous cell types, rare subpopulations, and continuous developmental transitions that bulk methods obscure.



These benefits are particularly important for studying *Trypanosoma brucei*, which exists in the mammalian bloodstream as a mixture of developmental stages, including **proliferative slender forms** and **transmission-competent stumpy forms**. Because these stages differ in metabolism, cell-cycle state, and gene expression, but coexist within the same sample, bulk RNA-seq cannot distinguish them or reconstruct how slender parasites differentiate into stumpy forms. In contrast, scRNA-seq can resolve each stage as a distinct transcriptional state, identify intermediate cells along the differentiation trajectory, and reveal marker genes that define each step, enabling an understanding of how parasite populations transition from growth to transmission.

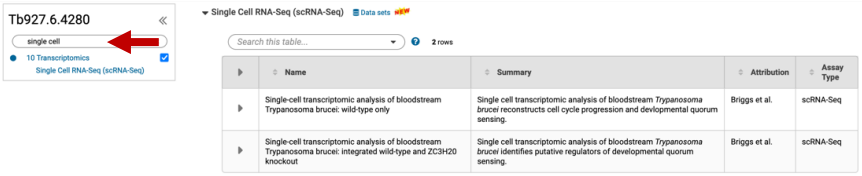
Data used in this exercise is from the paper *Briggs, E.M., Rojas, F., McCulloch, R. et al. Single-cell transcriptomic analysis of bloodstream Trypanosoma brucei reconstructs cell cycle progression and*

developmental quorum sensing. Nat Commun 12, 5268 (2021). The gene GAPDH is a glycolytic enzyme that is a marker for the **proliferative slender form** characterized by active metabolism and cell division, and the gene PAD2 (protein associated with differentiation) is a marker of the **transmission-competent stumpy form**.

Part 1: Exploring scRNA data on the gene page

1. Go to TriTrypDB.org[1], the kinetoplastid informatics resources within VEuPathDB.

2. Find the gene page for glyceraldehyde 3-phosphate dehydrogenase (GAPDH: **Tb927.6.4280**) and go to the single-cell RNA-Seq section of the page. You can quickly do this by filtering the categories on the left side of the gene page.

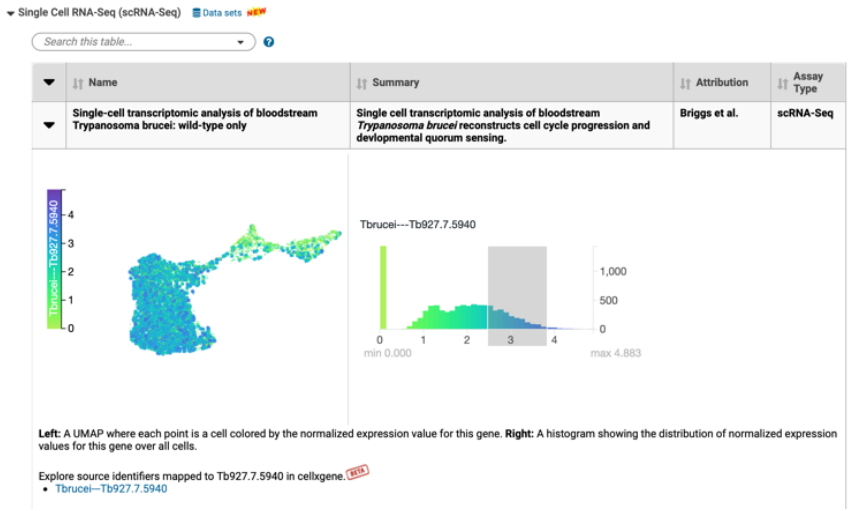


3. Expand the first experiment, which shows wild-type cells only. You see a **UMAP** (Uniform Manifold Approximation and Projection) plot that visualizes high-dimension data to show patterns and clusters-
    a. The left panel is a UMAP with cells colored by expression (green = low, blue = high)
    b. The right panel is a distribution plot showing overall expression levels for Tb927.6.4280 across all cells
    c. You can click and drag in the histogram panel on the right to highlight cells in the left panel. Choose the area between 3 and 4 on the histogram to highlight high-expressing cells on the graph.

4. What does the UMAP plot show? Where are the cells with the highest expression of this gene?

[1] This exercise uses TriTrypDB.org as an example database, but the same functionality is available on all VEuPathDB resources where this type of data is present.

5. Try the same thing by opening a different gene page for “Protein Associated with Differentiation” (*PAD2*: **Tb927.7.5940**).



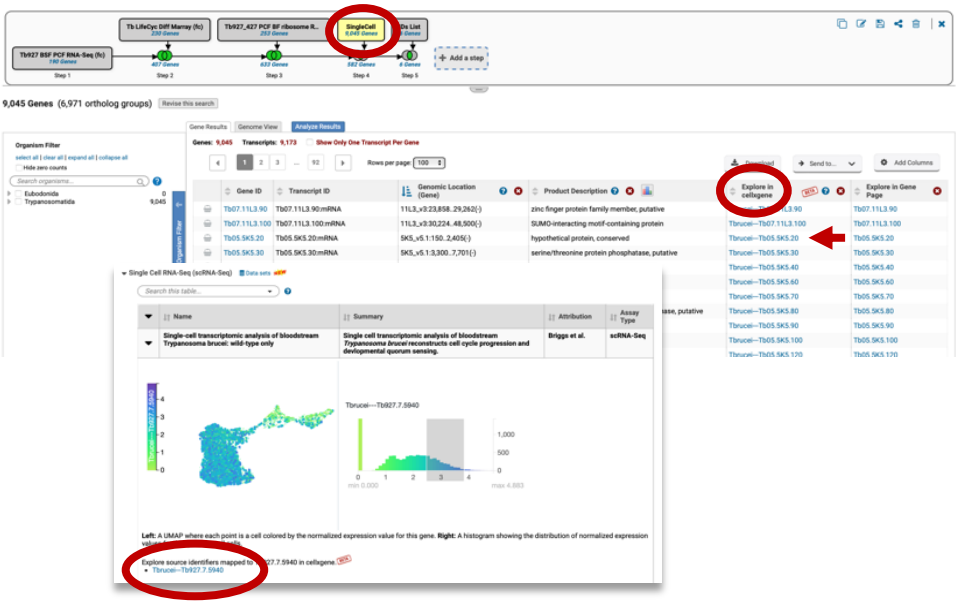
Do cells expressing elevated levels of *PAD2* and *GAPDH* coincide on the UMAP or are they in different regions of the plot?

Since *GAPDH* is a slender marker and *PAD2* is a stumpy marker, what can you conclude about the cells that coincide with those markers?

Part 2: Exploring scRNA data in the cellxgene application

Cellxgene ("cell-by-gene") is an open-source data visualization and exploration tool designed to help interrogate high-dimensional data. We use cellxgene in VEuPathDB as a supplement to allow investigators to explore scRNA-Seq data.

1. Start with the Briggs et al. wild-type experiment. There are two ways to open cellxgene. The gene page has a link below the graphs for each experiment. You can also run a search that identifies genes that have scRNAseq data available. The results of this search include a column with links to explore the results in cellxgene.

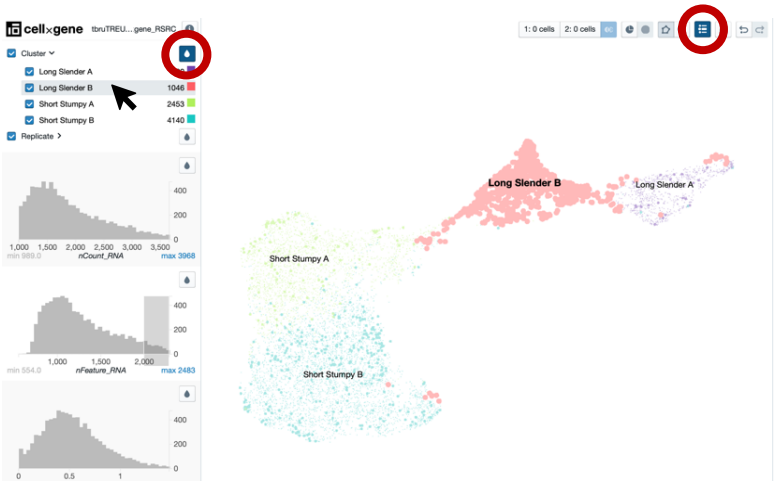


2. Your initial view will be a UMAP plot of all cells from this experiment. This may be black and white or colored to show the expression of a specific gene, depending on how you got there.

3. The left panel includes metadata, while the right panel includes gene feature data, where data for any gene measured in the dataset can be explored. The central area is the cell visualization and exploration panel.

4. Note that the metadata section includes numerical metadata represented as interactive histograms and categorical metadata such as the cluster assignments or replicates. The exact data shown here will vary by experiment.

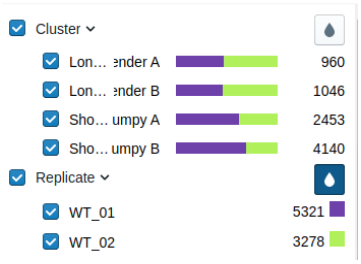
5. The droplet icon can be used to color the cells in the central panel with metadata from the left panel or gene expression data from the right panel. Do the annotations fit with what you saw when you looked at GAPDH and PAD2 on the gene pages earlier?



6. Expand the “Cluster” metadata category to see the cluster names. Note that these have been annotated by the author of the dataset.

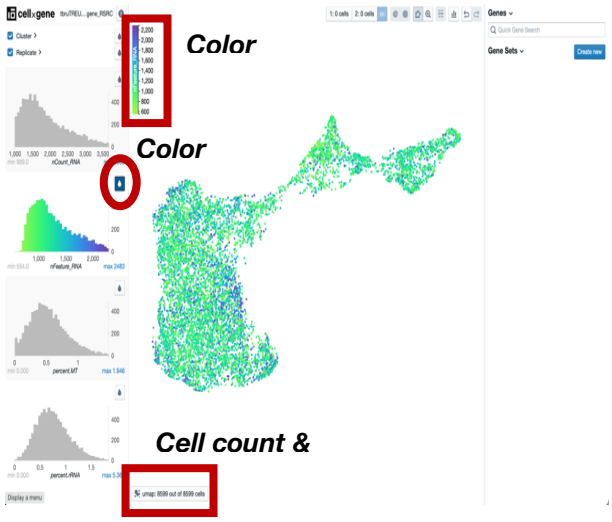
7. Label the UMAP with the cluster names by clicking the labels button in the central panel menu. Hover over the cluster names to bring them into focus in the UMAP.

8. Turn the labeling off again. Expand the “Replicate” metadata category. Use the droplet icon to color based on the replicate. Mouse over the replicates to see how they are distributed in the UMAP. Notice the bars that appear for the cluster categories showing the proportion of cells from each replicate in each cluster. Do these look like good replicates?



9. The droplet icon can also be used to color cells based on continuous metadata. Generally, the continuous metadata available is provided for QC purposes. Try this: click on the droplet icon for the **nFeature_RNA** (number of genes detected in each cell). How many cells are displayed?

10. In single-cell data, capturing a variable number of genes from each cell is common. How many cells were captured in which 2000 or more genes were observed? To find this, click and drag the histogram area in the left panel to highlight the area representing 2000 and above. Note: don’t worry about being exact here; you are just trying to understand what the data looks like.

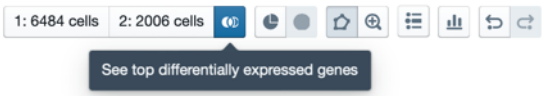


11. Do you think you have a higher percentage of stumpy cells with more genes assayed than the slender forms? Can you get the number of cells in each of the stages that met the >/= 2000 genes representation (See step 10)? To do this, click and draw around the stumpy cells in the central panel or use the checkboxes next to the cluster labels to deselect the slender cells.

12. Do the same thing for the slender population. Do you see a difference in the number of cells? Don’t forget to take account of the overall number of cells in each population. You can see this in the left panel.

13. Now let us identify genes that differentiate between the stumpy and slender populations. Follow these steps to do this:
    a. Select the stumpy population (both A and B). You can do this by clicking and drawing round them, or by using the check boxes in the left pane. Click on population 1 in the menu bar to save the selection for differential expression.
    b. Repeat the same process to select the slender population and save it as population 2.

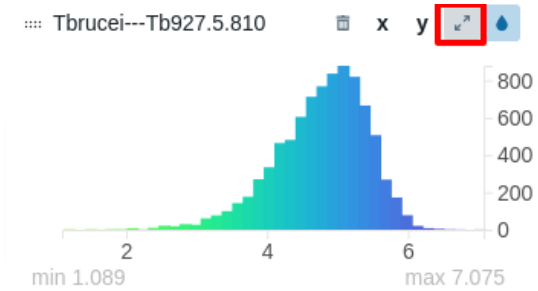
c. When done with your selections and saving populations, click on the differential expression icon.

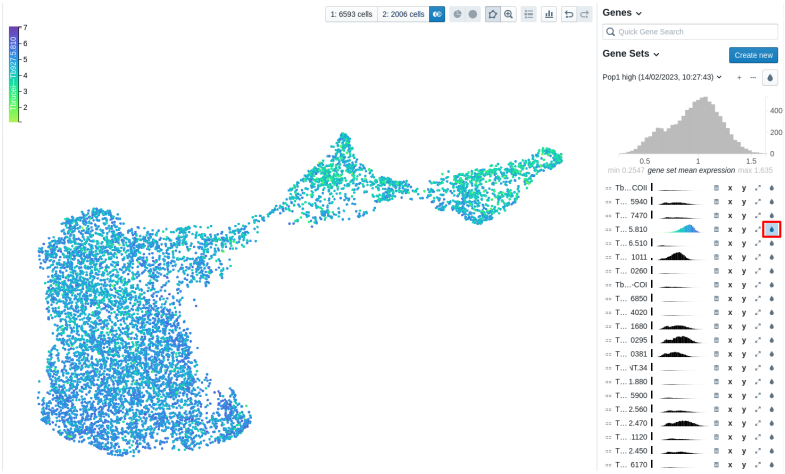


14. Click on population 1 in the right-hand gene feature panel to reveal the **top stumpy genes**.
    a. Click on the expand icon to view a gene more clearly. The histogram in the right panel shows the expression of this gene over all the cells.

    b. You can color the UMAP by clicking on the droplet icon next to each gene.

    c. The expression of this gene in each cluster can be viewed as histograms in the left panel.



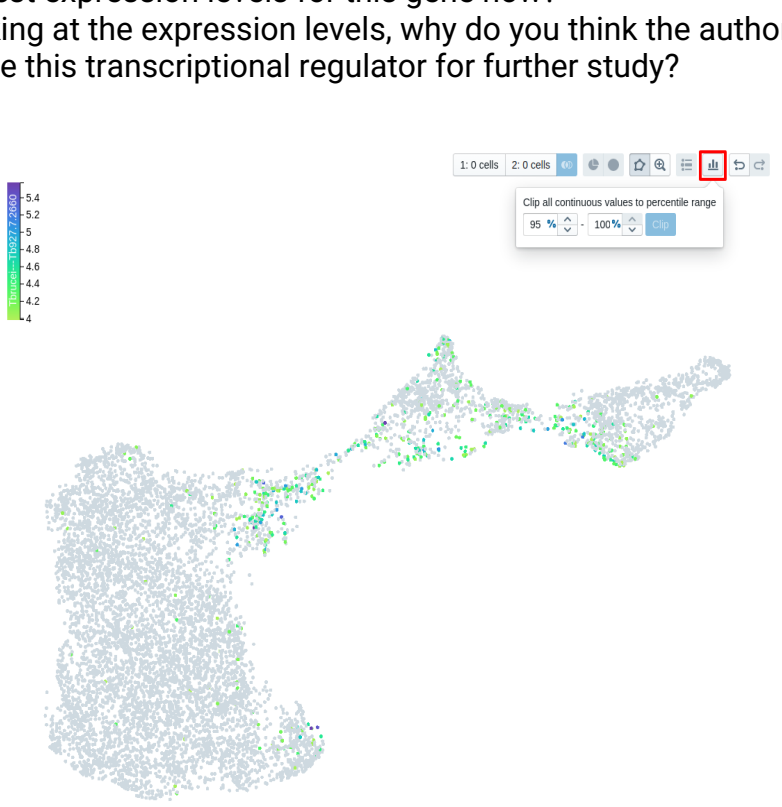


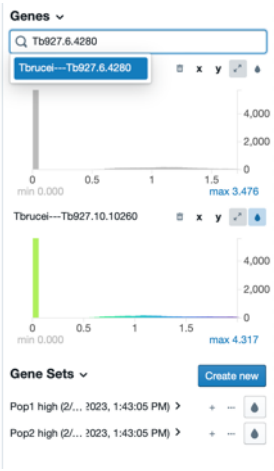
15. Copy one of the gene IDs and explore it in TriTrypDB. Can you come up with a rational reason why your selected gene might be important in stumpy development? *Note that copying gene IDs from cellxgene is frustrating. If you click on the expand icon for the individual gene, it becomes easier to copy the gene ID.*

16. Repeat this for the slender forms.

17. How do the gene sets you identified in your differential expression compare to the marker genes used in the paper? You search for specific genes by pasting the gene ID in the quick gene search window in the right-hand panel. You can select the gene to explore it further if it is found. Here is the list of marker genes: Tb927.10.10260, Tb927.10.14140, Tb927.6.4280, Tb927.7.2660, Tb927.7.5930, Tb927.7.5940

18. The authors identified one gene as a putative regulator of slender to stump transition. This is a zinc-finger protein which has been described as having a role post-transcriptional regulation. Let's look at the expression of this protein.
    a. The gene id is Tb927.7.2660. Find this gene using the quick gene search in the right pane and color the UMAP with expression values for this gene.
    b. Which cells are expressing this gene at the highest levels? Is it easy to see a pattern just by coloring for this gene?
    c. We can explore this further in two different ways. First, try clicking and dragging on the expression histogram for this gene to highlight cells where the expression value is > 4.5. You have done this already using the nFeature_RNA histogram.
    d. The second method is to use the clipping tool. Select the clipping tool in the top menu. Leave the upper value at 100%. Change the lower value to 95% and click "Clip". You are now coloring only the cells in the 95th percentile of expression for this gene.
    e. What happens to the UMAP and the histogram? Is it easier to find the cells with the highest expression levels for this gene now?
    f. Looking at the expression levels, why do you think the authors chose this transcriptional regulator for further study?

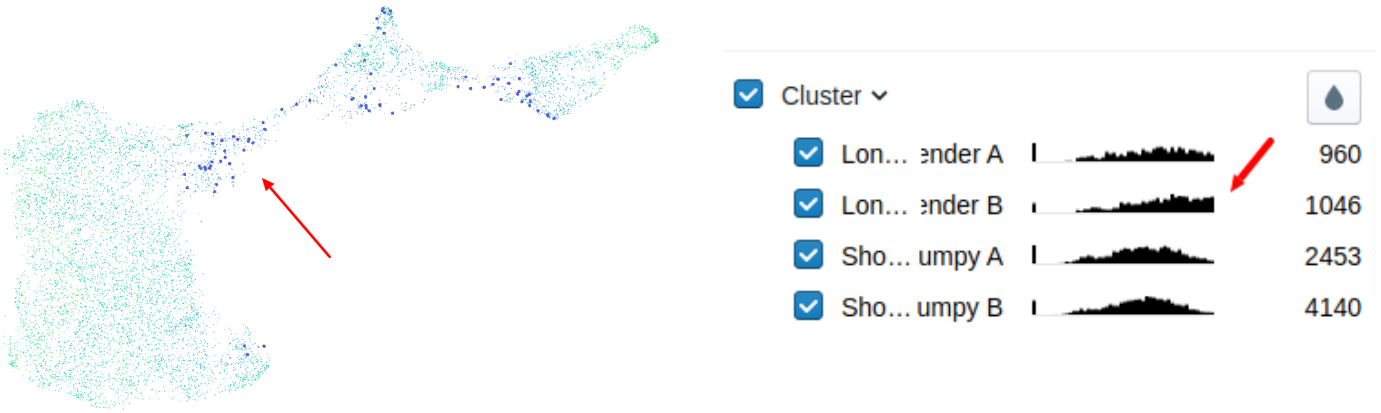




Part 3: Optional exercise

The dataset you have just explored is one of two from the same paper. In the first dataset, the authors explored the transition from the replicative slender form of *T. brucei* to the non-replicative, transmissible stumpy form. During this work, they identified a putative regulator, a zinc finger protein ZC3H20: Tb927.7.2660. We looked at the expression of this gene at the end of part 3.

Note that this gene is most highly expressed in the slender B population, and the cells that are transitioning between slender to stumpy. This can be observed in the histograms in the left pane, or by using the gene expression histogram in the right pane to select the cells with the highest expression levels.

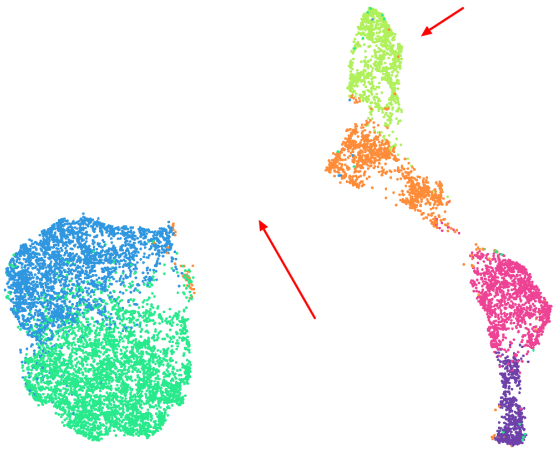


In a subsequent experiment, the group knocked this gene out. The sequenced data from the knockout was integrated with the wild-type cells you've already looked at. This data can be viewed in TriTrypDB using the methods you learned above, or it can be directly accessed in cellxgene here: https://tritrypdb.org/cellxgene/view/tbruTREU927_briggs_ZC3H20_KO_cellxgene_RSRC.h5ad/

This exercise has fewer instructions - explore this dataset using the tools you learnt about above.

1. Look at the UMAP for the integrated experiment. How does this differ compared to the wild-type cells?
    - Are there any cell populations you didn't see before?
    - Is there still a clear transition from slender to stumpy?

- What do you think is the effect of knocking out this gene?

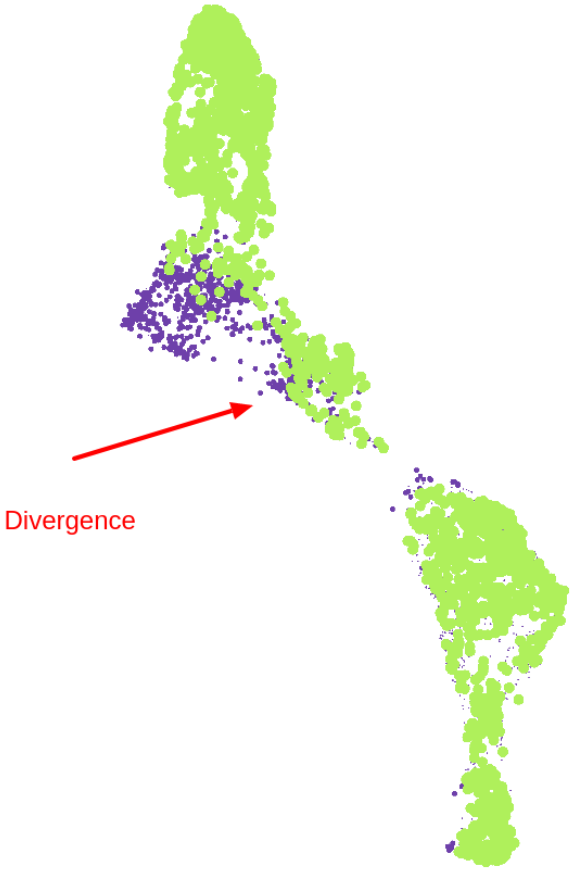


2. Color the UMAP for the expression of the gene that was knocked out (Tb927.7.2660). Use the tools you've already learnt about to explore this.
    - Does this look as you expect?
    - Are there any cell populations where this gene is not expressed?
    - Do you think the knockout was successful?

3. Expand the "Line" metadata category in the left menu. Color the UMAP based on this. Mouseover to highlight the different populations.
    - How are the knockout cells different from the wild-type cells?
    - Expand the "Cluster" metadata category to see the proportions of cells.

4. It appears that knock-out cells cannot differentiate from slender to stumpy. Instead, they form a novel cluster labeled "Long Slender B.2". Let's see what genes are highly expressed in this cluster:
    - Uncheck all the clusters except "Long Slender B.2". Add this to Population 1
    - Inverse this operation. Add the other clusters to Population 2.
    - Do the differential expression. Can you find any genes in the Pop 1 high list (upregulated in B.2) that look interesting to explore further?

5. Color the UMAP by the "Line" metadata category again
    - Mouseover the categories and watch what happens in the Long Slender B.1 population
    - It looks like the cells diverge in their differentiation program here



- The best way to explore this is through trajectory inference. Unfortunately, cellxgene does not offer this. However, we can begin to explore this with differential expression
- Select only the Long Slender B.1 population. Use the check box, then draw around it to remove outlier cells.
- Use the subset button to remove the other cells.

- Now add the ZC3H20_KO cells in this subset to population 1, add the WT cells in this subset to population 2, and do a differential expression.
- What genes in this gene set do you think might be worth exploring further?