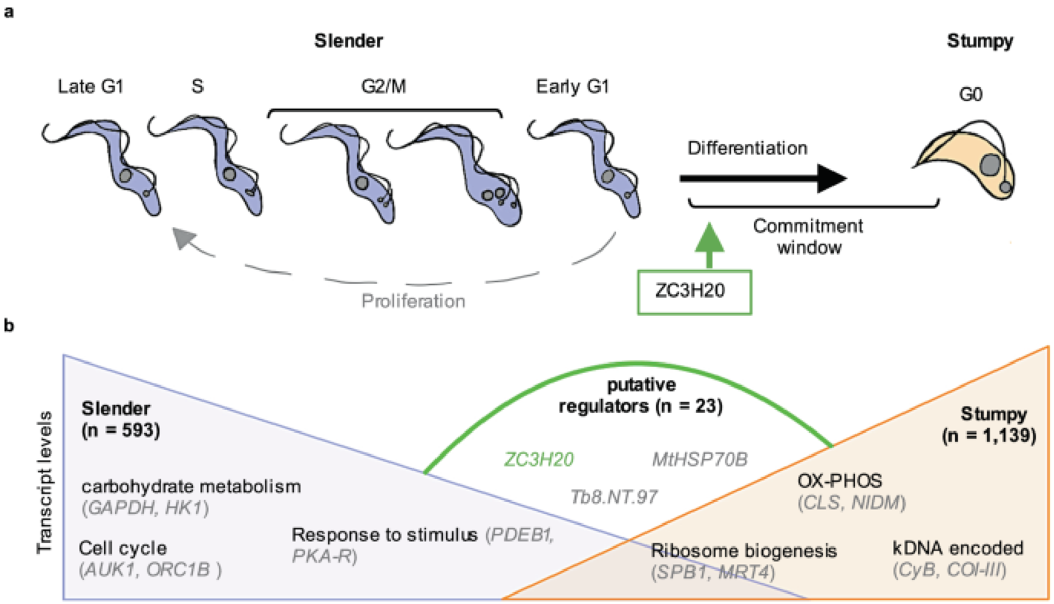
Single Cell RNA-Sequencing (scRNA-seq)

***Note:** 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.*

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.


a
Slender
Stumpy
Late G1
S
G2/M
Early G1
G0
Differentiation
Commitment window
ZC3H20
Proliferation
b
Transcript levels
Slender (n = 593)
carbohydrate metabolism (GAPDH, HK1)
Cell cycle (AUK1, ORC1B)
Response to stimulus (PDEB1, PKA-R)
putative regulators (n = 23)
ZC3H20
MtHSP70B
Tb8.NT.97
Stumpy (n = 1,139)
OX-PHOS (CLS, NIDM)
Ribosome biogenesis (SPB1, MRT4)
kDNA encoded (CyB, COI-III)

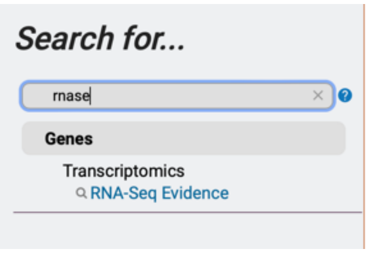
Data used in this exercise is from Briggs, E.M., Rojas, F., McCulloch, R. *et al.* Single-cell transcriptomic analysis of bloodstream *Trypanosoma brucei* reconstructs cell cycle progression and developmental quorum sensing. *Nat Commun* **12**, 5268 (2021). https://doi.org/10.1038/s41467-021-25607-2

Slender markers:
GAPDH: Tb927.6.4280
PYK1: Tb927.10.14140

Stumpy markers:
PAD1: Tb927.7.5930
PAD2: Tb927.7.5940
EP1: Tb927.10.10260

Development regulator:
ZC3H20: Tb927.7.2660

1. Identify genes that are upregulated in the stumpy form compared to the slender form in different experiments.
   a. Start by running a search based on RNA-Seq (hint: to find RNA-Seq searches start typing the word RNAseq in the search filter on the left of the home page). Click on RNA-Seq Evidence to view a list of all the available experiments and search types.



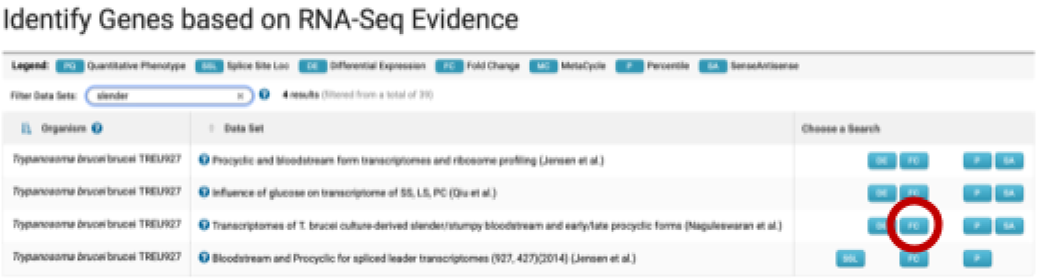
Search for...

rnase

Genes

Transcriptomics
RNA-Seq Evidence

   b. Use the dataset filter to find all experiments that included slender parasites.
   c. Select the fold-change search associated with the experiment by Naguleswaran et al.



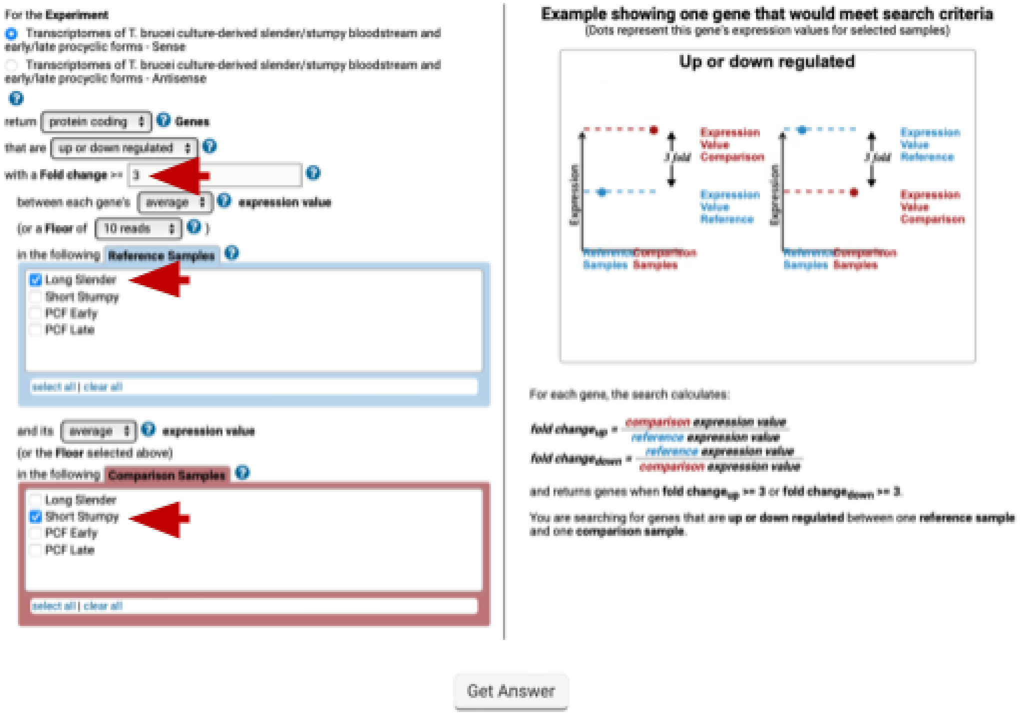
Identify Genes based on RNA-Seq Evidence

Legend: PQ Quantitative Phenotype SSL Splice Site Loc DE Differential Expression FC Fold Change MC MetaCycle P Percentile SA SenseAntisense

Filter Data Sets: slender 4 results (filtered from a total of 39)

Organism	Data Set	Choose a Search
Trypanosoma brucei brucei TREU927	Procyclic and bloodstream form transcriptomes and ribosome profiling (Jensen et al.)	DE FC P SA
Trypanosoma brucei brucei TREU927	Influence of glucose on transcriptome of SS, LS, PC (Qiu et al.)	DE FC P SA
Trypanosoma brucei brucei TREU927	Transcriptomes of T. brucei culture-derived slender/stumpy bloodstream and early/late procyclic forms (Naguleswaran et al.)	DE FC P SA
Trypanosoma brucei brucei TREU927	Bloodstream and Procyclic for spliced leader transcriptomes (927, 427)(2014) (Jensen et al.)	SSL FC P

   d. Set up the search parameters to identify genes that are differentially regulated (up or down) by at least 3-fold between the slender and stumpy forms (see



For the **Experiment**
● Transcriptomes of T. brucei culture-derived slender/stumpy bloodstream and early/late procyclic forms - Sense
○ Transcriptomes of T. brucei culture-derived slender/stumpy bloodstream and early/late procyclic forms - Antisense

return protein coding **Genes**
that are up or down regulated
with a **Fold change** >= 3
between each gene's average **expression value**
(or a **Floor** of 10 reads)
in the following **Reference Samples**
☑ Long Slender
□ Short Stumpy
□ PCF Early
□ PCF Late
select all | clear all

and its average **expression value**
(or the **Floor** selected above)
in the following **Comparison Samples**
□ Long Slender
☑ Short Stumpy
□ PCF Early
□ PCF Late
select all | clear all

Example showing one gene that would meet search criteria
(Dots represent this gene's expression values for selected samples)

For each gene, the search calculates:

$$\textit{fold change}_{up} = \frac{\text{comparison expression value}}{\text{reference expression value}}$$

$$\textit{fold change}_{down} = \frac{\text{reference expression value}}{\text{comparison expression value}}$$

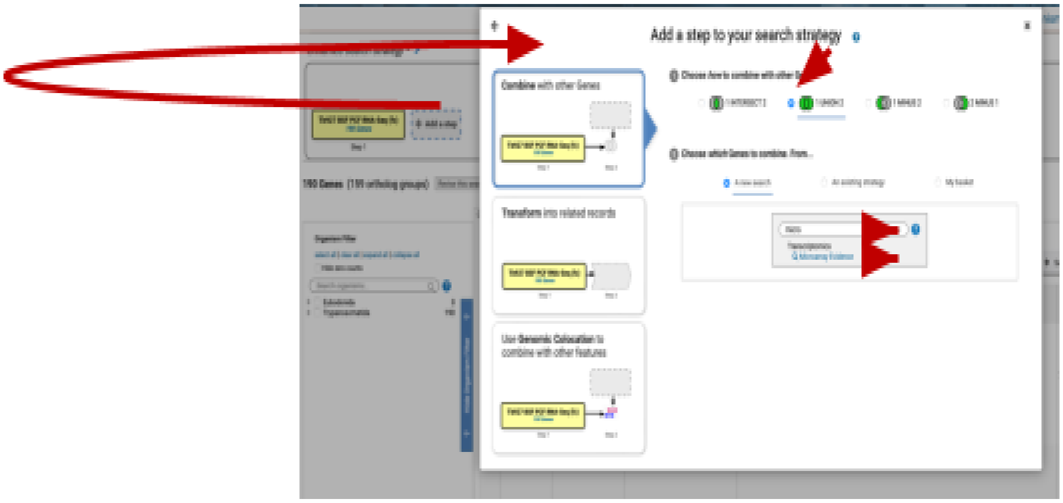
and returns genes when **fold change$_{up}$** >= 3 or **fold change$_{down}$** >= 3.

You are searching for genes that are **up or down regulated** between one **reference sample** and one **comparison sample**.

Get Answer

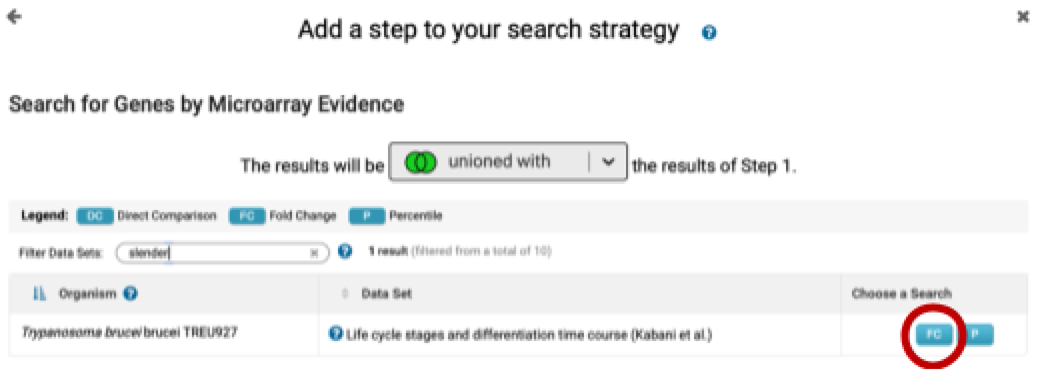
screenshot if you need help). Once you are happy with the parameters click on the “Get Answer” button.

e. Let’s expand the number of genes that might be interesting by combining the results from the above search with results from another experiment that assayed the same stages. To do this follow these steps:
    - Click on the add step button in your search strategy panel.
    - Choose the union operator (why not intersect?).


Add a step to your search strategy
Combine with other Genes
Transform into related records
Use Genomic Colocation to combine with other features
Add a step
Step 1

f. Find the microarray searches by filtering the searches using the key word microarray, then click on the “microarray evidence” link.

g. Filter the experiments and find one that includes slender stages. Select the fold change search associated with the Kabani et al. experiment. Configure the microarray search to find all genes that are differentially regulated by 2-fold between the slender and 0hr, and hours 1-48. When satisfied with your configuration click on “Run step”.



Add a step to your search strategy

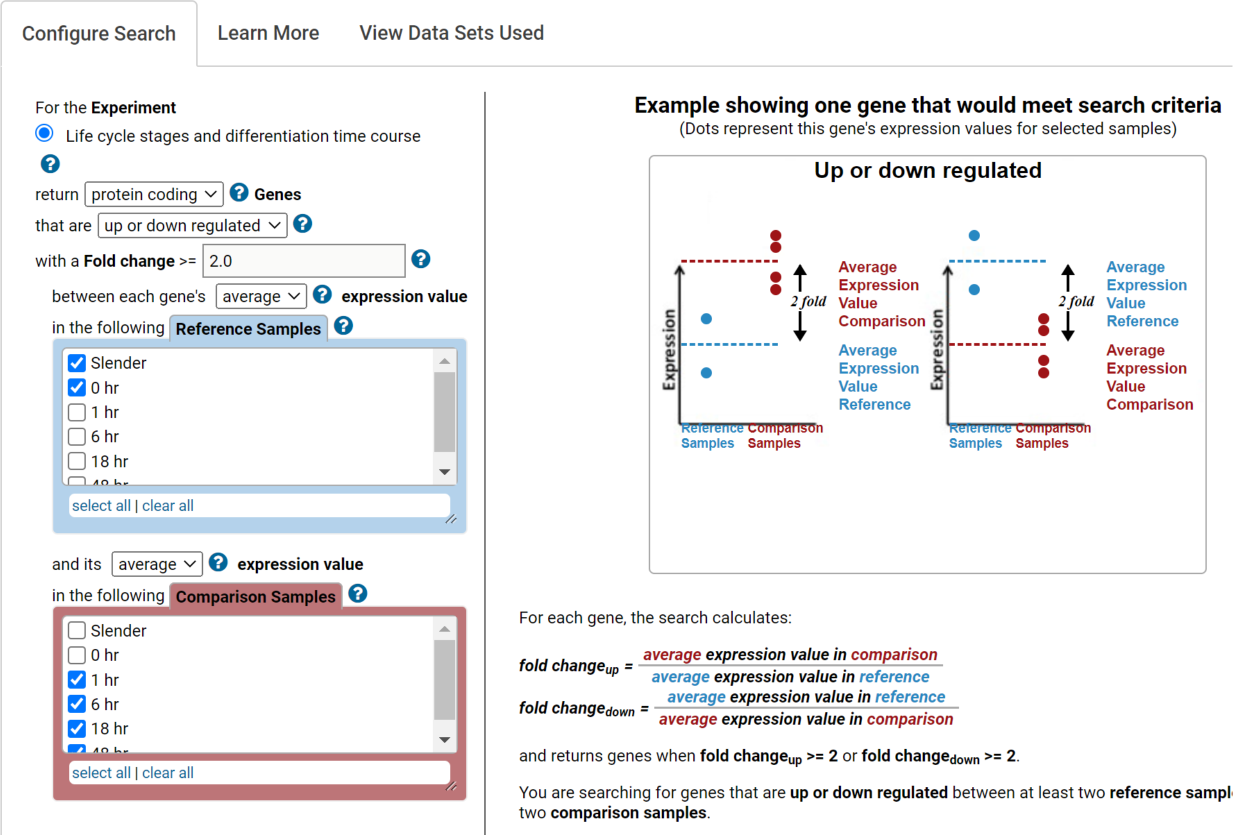
Search for Genes by Microarray Evidence

The results will be unioned with the results of Step 1.

Legend: DC Direct Comparison FC Fold Change P Percentile

Filter Data Sets: slender 1 result (filtered from a total of 10)

Organism	Data Set	Choose a Search
Trypanosoma brucei brucei TREU927	Life cycle stages and differentiation time course (Kabani et al.)	FC P



Configure Search | Learn More | View Data Sets Used

For the **Experiment**

● Life cycle stages and differentiation time course

return [protein coding ⌄] **Genes**

that are [up or down regulated ⌄]

with a **Fold change** >= [2.0]

between each gene's [average ⌄] **expression value**

in the following **Reference Samples**

☑ Slender
☑ 0 hr
□ 1 hr
□ 6 hr
□ 18 hr

select all | clear all

and its [average ⌄] **expression value**

in the following **Comparison Samples**

□ Slender
□ 0 hr
☑ 1 hr
☑ 6 hr
☑ 18 hr

select all | clear all

Example showing one gene that would meet search criteria

(Dots represent this gene's expression values for selected samples)

Up or down regulated

Expression
2 fold
Average Expression Value Comparison
Average Expression Value Reference
Reference Samples
Comparison Samples
Expression
2 fold
Average Expression Value Reference
Average Expression Value Comparison
Reference Samples
Comparison Samples

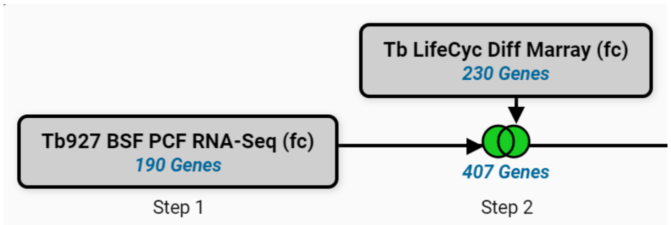
For each gene, the search calculates:

$$\textit{fold change}_{up} = \frac{\textit{average}\ \textbf{expression value in}\ \textit{comparison}}{\textit{average}\ \textbf{expression value in}\ \textit{reference}}$$

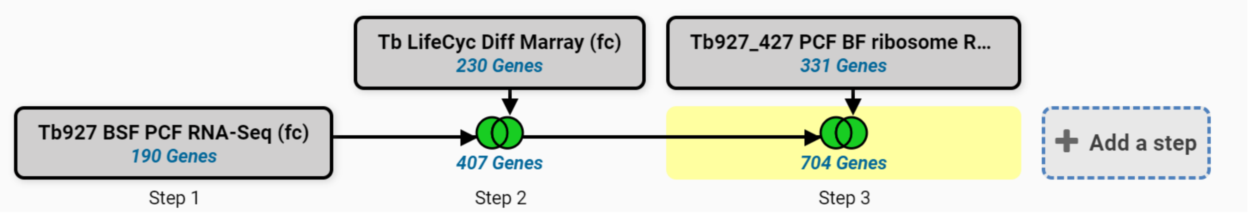
$$\textit{fold change}_{down} = \frac{\textit{average}\ \textbf{expression value in}\ \textit{reference}}{\textit{average}\ \textbf{expression value in}\ \textit{comparison}}$$

and returns genes when **fold change**$_{up}$ **>= 2** or **fold change**$_{down}$ **>= 2**.

You are searching for genes that are **up or down regulated** between at least two **reference sampl** two **comparison samples**.

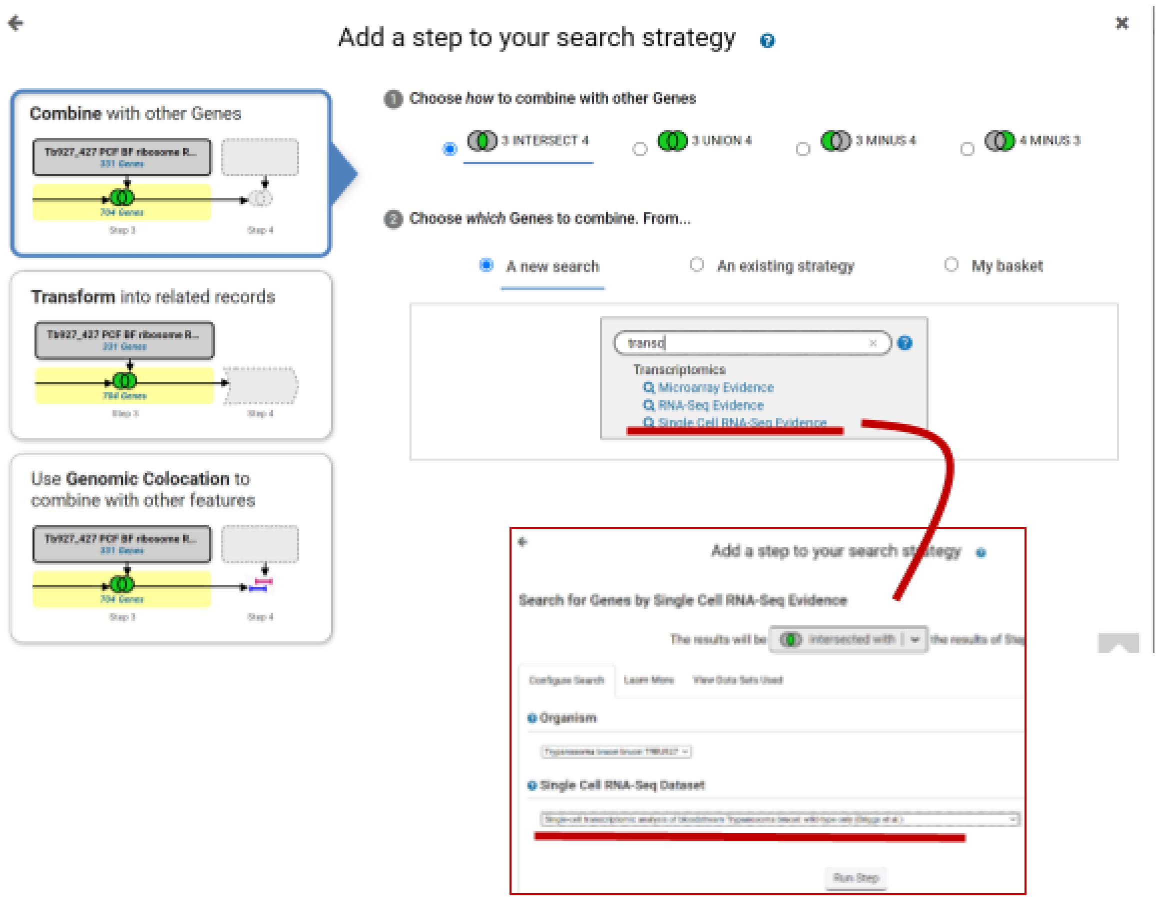


h. Using the same logic as above, add another step and find the RNA-Seq experiment from Jensen et al. and configure the fold change search to find all differentially expressed genes by 2-fold comparing the blood form (cBF mRNA) to the slender form (slBF mRNA). (don't forget to use the union operator).



Tb LifeCyc Diff Marray (fc)
230 Genes
Tb927_427 PCF BF ribosome R...
331 Genes
Tb927 BSF PCF RNA-Seq (fc)
190 Genes
407 Genes
704 Genes
+ Add a step
Step 1
Step 2
Step 3

2. **Identify genes with expression in single cell RNA-Seq experiments.**
    a. How many of the genes from #1 above were also identified in the single cell RNA-Seq experiment: “Single-cell transcriptomic analysis of bloodstream Trypanosoma brucei: wild-type only (Briggs et al.)”. Using the same logic as above for adding searches, find the single cell RNA-seq experiments and use the intersect operation to find genes from your previous search that have results in the scRNA-Seq experiment. Why are some genes not represented in the single cell experiment?



Add a step to your search strategy

Combine with other Genes
Tb927_427 PCF BF ribosome R...
331 Genes
704 Genes
Step 3
Step 4
Transform into related records
Tb927_427 PCF BF ribosome R...
331 Genes
704 Genes
Step 3
Step 4
Use Genomic Colocation to combine with other features
Tb927_427 PCF BF ribosome R...
331 Genes
704 Genes
Step 3
Step 4

1 Choose *how* to combine with other Genes

● 3 INTERSECT 4 ○ 3 UNION 4 ○ 3 MINUS 4 ○ 4 MINUS 3

2 Choose *which* Genes to combine. From...

● A new search ○ An existing strategy ○ My basket

transc

Transcriptomics
Microarray Evidence
RNA-Seq Evidence
Single Cell RNA-Seq Evidence

Add a step to your search strategy

Search for Genes by Single Cell RNA-Seq Evidence

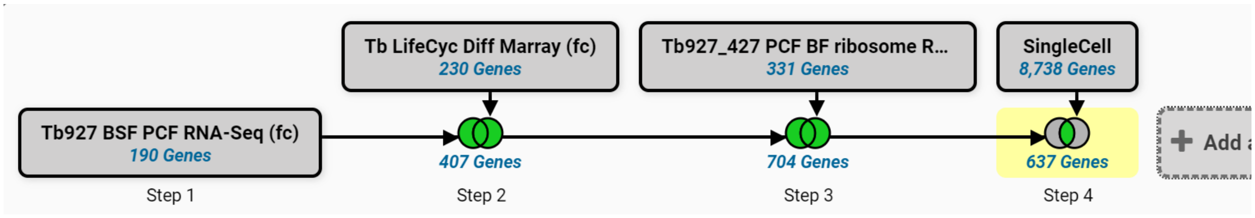
The results will be intersected with the results of Step

Organism

Trypanosoma brucei brucei TREU927

Single Cell RNA-Seq Dataset

Single-cell transcriptomic analysis of bloodstream Trypanosoma brucei: wild-type only (Briggs et al.)



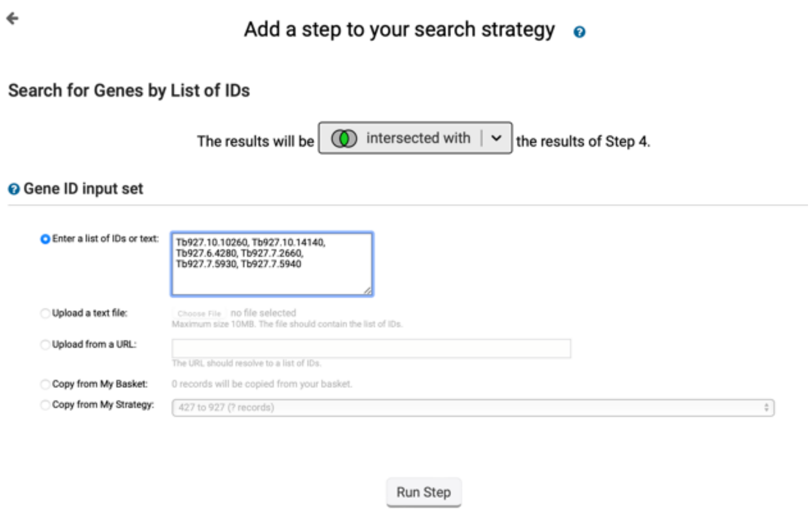
Tb927 BSF PCF RNA-Seq (fc)
190 Genes
Step 1

Tb LifeCyc Diff Marray (fc)
230 Genes
407 Genes
Step 2

Tb927_427 PCF BF ribosome R...
331 Genes
704 Genes
Step 3

SingleCell
8,738 Genes
637 Genes
Step 4

b. Does this list of genes include any of the markers described in the paper? You can add another step and search using a list of IDs. Copy and paste the following IDs into the search window: Tb927.10.10260, Tb927.10.14140, Tb927.6.4280, Tb927.7.2660, Tb927.7.5930, Tb927.7.5940



Search for Genes by List of IDs

The results will be intersected with the results of Step 4.

Gene ID input set

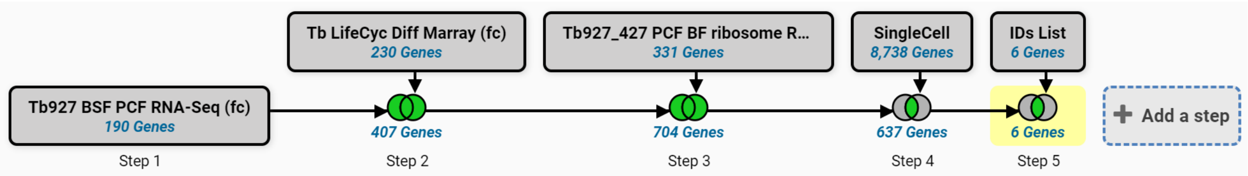
● Enter a list of IDs or text: Tb927.10.10260, Tb927.10.14140, Tb927.6.4280, Tb927.7.2660, Tb927.7.5930, Tb927.7.5940

○ Upload a text file: Choose File no file selected
Maximum size 10MB. The file should contain the list of IDs.

○ Upload from a URL:
The URL should resolve to a list of IDs.

○ Copy from My Basket: 0 records will be copied from your basket.

○ Copy from My Strategy: 427 to 927 (? records)


Step 1: Tb927 BSF PCF RNA-Seq (fc) — 190 Genes

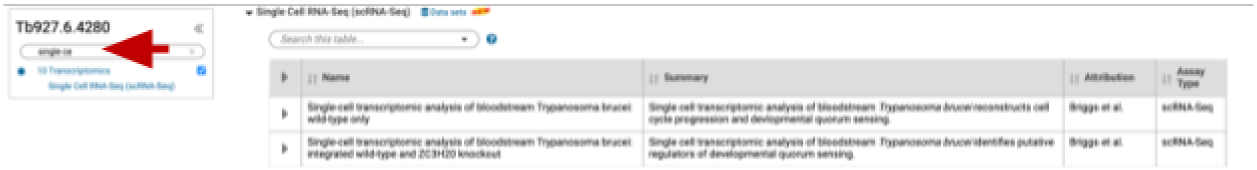
Step 2: Tb LifeCyc Diff Marray (fc) — 230 Genes; 407 Genes

Step 3: Tb927_427 PCF BF ribosome R... — 331 Genes; 704 Genes

Step 4: SingleCell — 8,738 Genes; 637 Genes

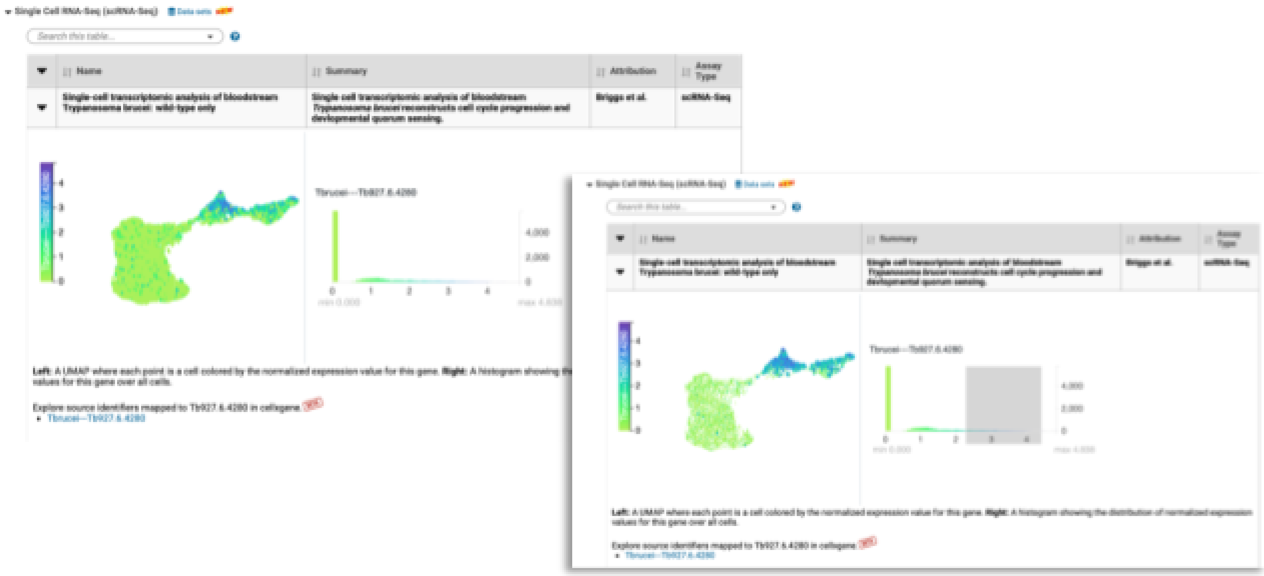
Step 5: IDs List — 6 Genes; 6 Genes

c. Visit 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.



Name	Summary	Attribution	Assay Type
Single-cell transcriptomic analysis of bloodstream Trypanosoma brucei wild-type only	Single cell transcriptomic analysis of bloodstream Trypanosoma brucei reconstructs cell cycle progression and developmental quorum sensing	Briggs et al.	scRNA-Seq
Single-cell transcriptomic analysis of bloodstream Trypanosoma brucei integrated wild-type and ZC3H20 knockout	Single cell transcriptomic analysis of bloodstream Trypanosoma brucei identifies putative regulators of developmental quorum sensing	Briggs et al.	scRNA-Seq

d. Expand the first experiment showing wild-type only cells. What does the UMAP plot show? Where are the cells with the highest expression of this gene? 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.



Single Cell RNA-Seq (scRNA-Seq) Data sets NEW

	Name	Summary	Attribution	Assay Type
	Single-cell transcriptomic analysis of bloodstream Trypanosoma brucei: wild-type only	Single cell transcriptomic analysis of bloodstream *Trypanosoma brucei* reconstructs cell cycle progression and devlopmental quorum sensing.	Briggs et al.	scRNA-Seq

Left: A UMAP where each point is a cell colored by the normalized expression value for this gene. **Right:** A histogram showing the distribution of normalized expression values for this gene over all cells.

Explore source identifiers mapped to Tb927.6.4280 in cellxgene.

- Tbrucei---Tb927.6.4280

Single Cell RNA-Seq (scRNA-Seq) Data sets NEW

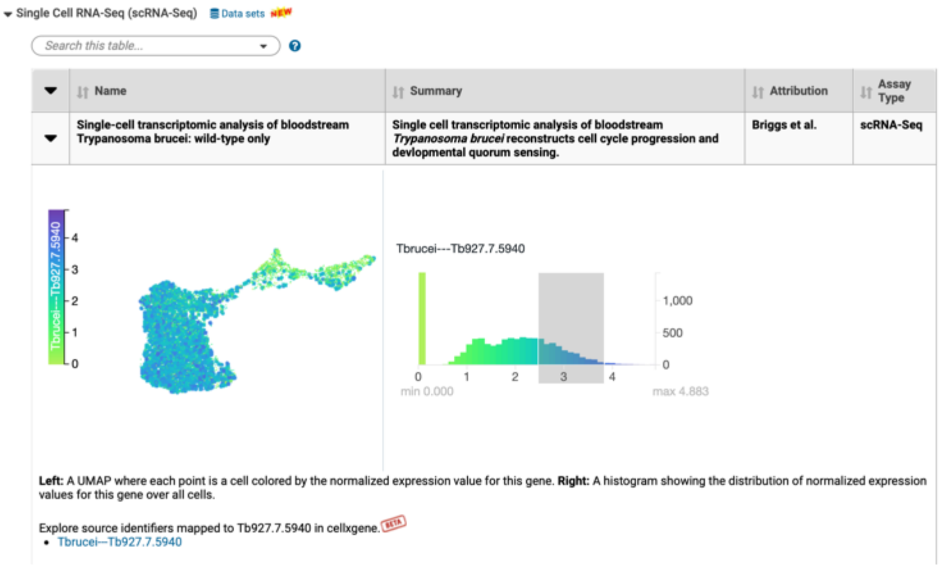
	Name	Summary	Attribution	Assay Type
	Single-cell transcriptomic analysis of bloodstream Trypanosoma brucei: wild-type only	Single cell transcriptomic analysis of bloodstream Trypanosoma brucei reconstructs cell cycle progression and devlopmental quorum sensing.	Briggs et al.	scRNA-Seq

Left: A UMAP where each point is a cell colored by the normalized expression value for this gene. **Right:** A histogram showing the distribution of normalized expression values for this gene over all cells.

Explore source identifiers mapped to Tb927.6.4280 in cellxgene.

- Tbrucei---Tb927.6.4280

e. Try the same thing with “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?



Single Cell RNA-Seq (scRNA-Seq) Data sets NEW

	Name	Summary	Attribution	Assay Type
	Single-cell transcriptomic analysis of bloodstream Trypanosoma brucei: wild-type only	Single cell transcriptomic analysis of bloodstream *Trypanosoma brucei* reconstructs cell cycle progression and devlopmental quorum sensing.	Briggs et al.	scRNA-Seq

Left: A UMAP where each point is a cell colored by the normalized expression value for this gene. **Right:** A histogram showing the distribution of normalized expression values for this gene over all cells.

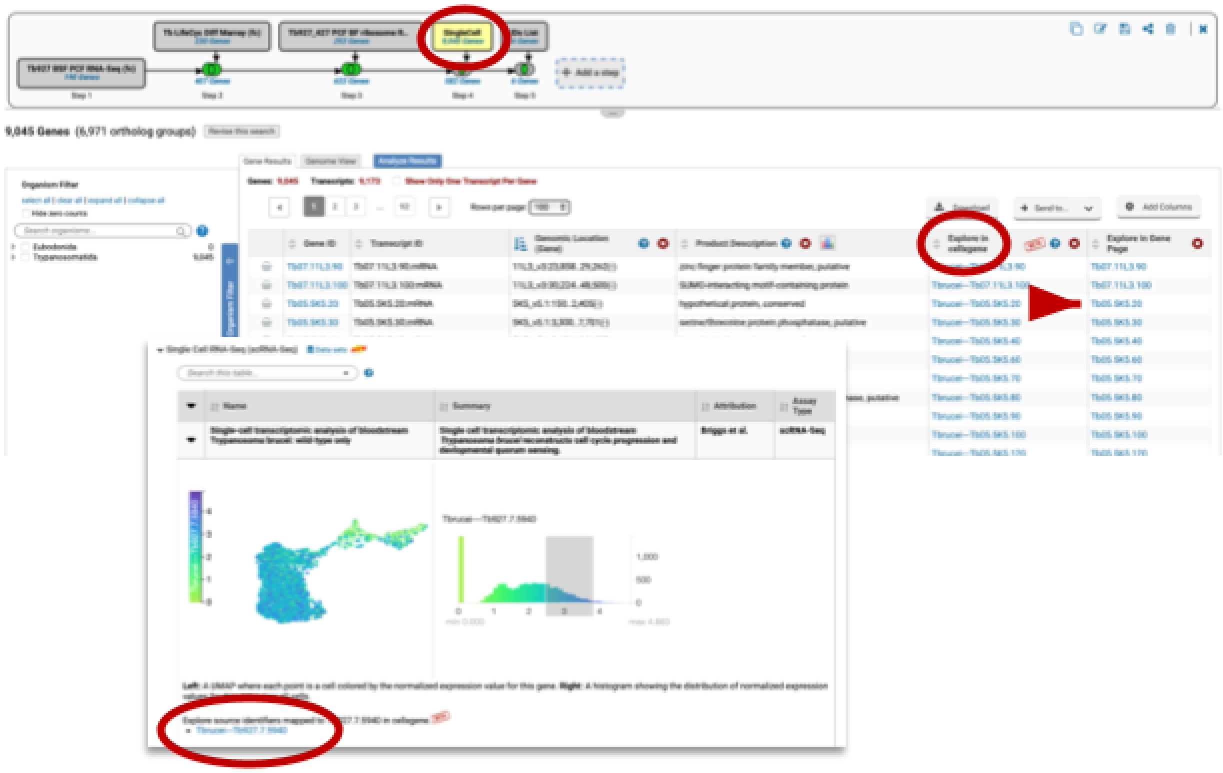
Explore source identifiers mapped to Tb927.7.5940 in cellxgene. BETA

- Tbrucei---Tb927.7.5940

3. Explore scRNA-Seq 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.

a. Start with the Briggs et al. wild-type experiment. There are two ways to open cellbygene From the gene page there is a link below the graphs for each experiment. You can also add a column to the strategy result for the Single Cell search.



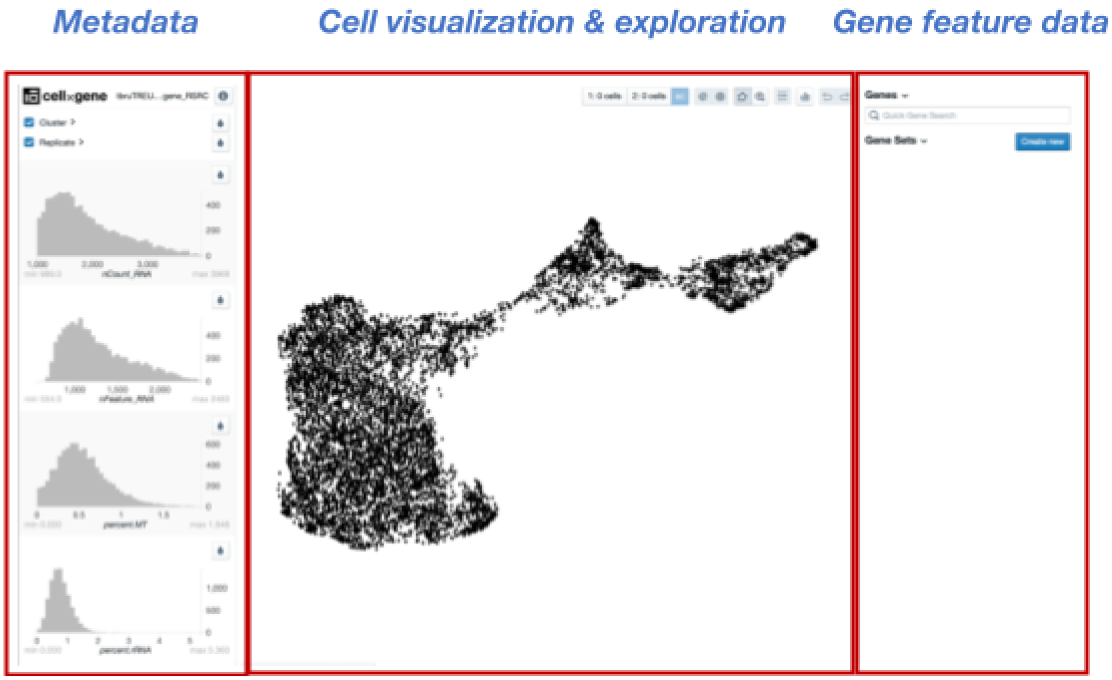
9,045 Genes (6,971 ortholog groups)

	Name	Summary	Attribution	Assay Type
	Single-cell transcriptomic analysis of bloodstream Trypanosoma brucei: wild-type only	Single cell transcriptomic analysis of bloodstream Trypanosoma brucei reconstructs cell cycle progression and developmental quorum sensing.	Briggs et al.	scRNA-Seq

Left: A UMAP where each point is a cell colored by the normalized expression value for this gene. **Right:** A histogram showing the distribution of normalized expression values

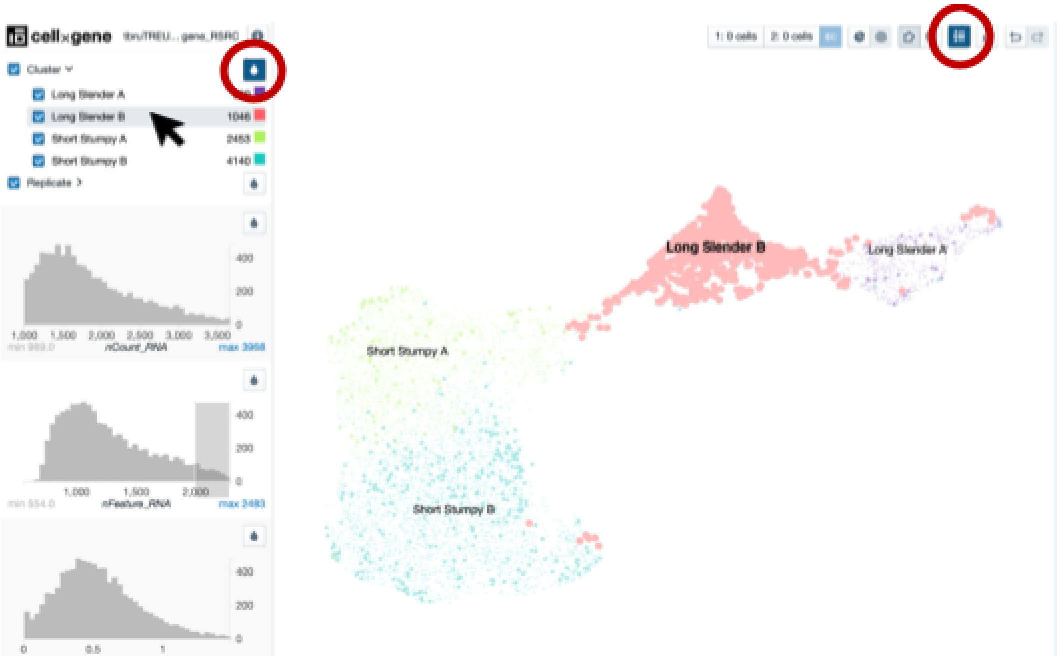
b. Your initial view will be a UMAP plot of all cells from this experiment. This may be black and white, or may be colored to show expression of a specific gene depending on how you got there.

c. The left-hand panel includes ***metadata*** while the right-hand 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.

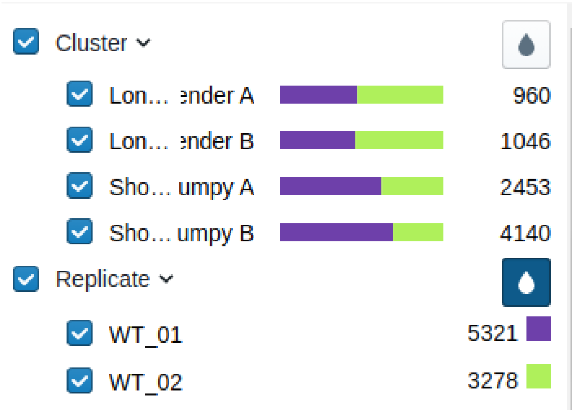


d. Note that the metadata section includes numerical metadata represented as interactive histograms and categorical metadata such as the cluster assignments or replicates. The exact data shown here will vary by experiment.

e. 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. Try this:
    - Expand the “Cluster” metadata category to see the cluster names. Note that these have been annotated by the author of the dataset
    - Use the droplet icon to color the cells by cluster. Do the annotations fit with what you saw when you looked at *GAPDH* and *PAD2* on the gene pages earlier?
    - Hover over the cluster names to bring them into focus in the UMAP.
    - Label the UMAP with the cluster names by clicking on the labels button in the central panel menu.



- Turn the labeling off again. Expand the “Replicate” metadata category. Use the droplet icon to color based on replicate. Mouseover 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?



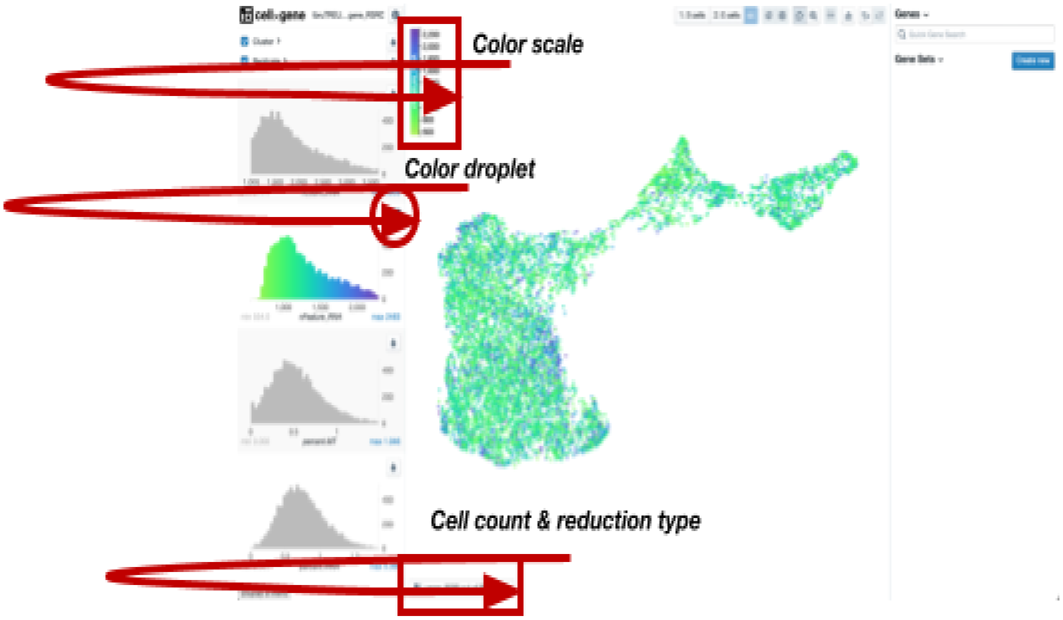
☑ Cluster

☑ Lon… ender A	960	
☑ Lon… ender B	1046	
☑ Sho… umpy A	2453	
☑ Sho… umpy B	4140	

☑ Replicate

☑ WT_01	5321	🟪
☑ WT_02	3278	🟩

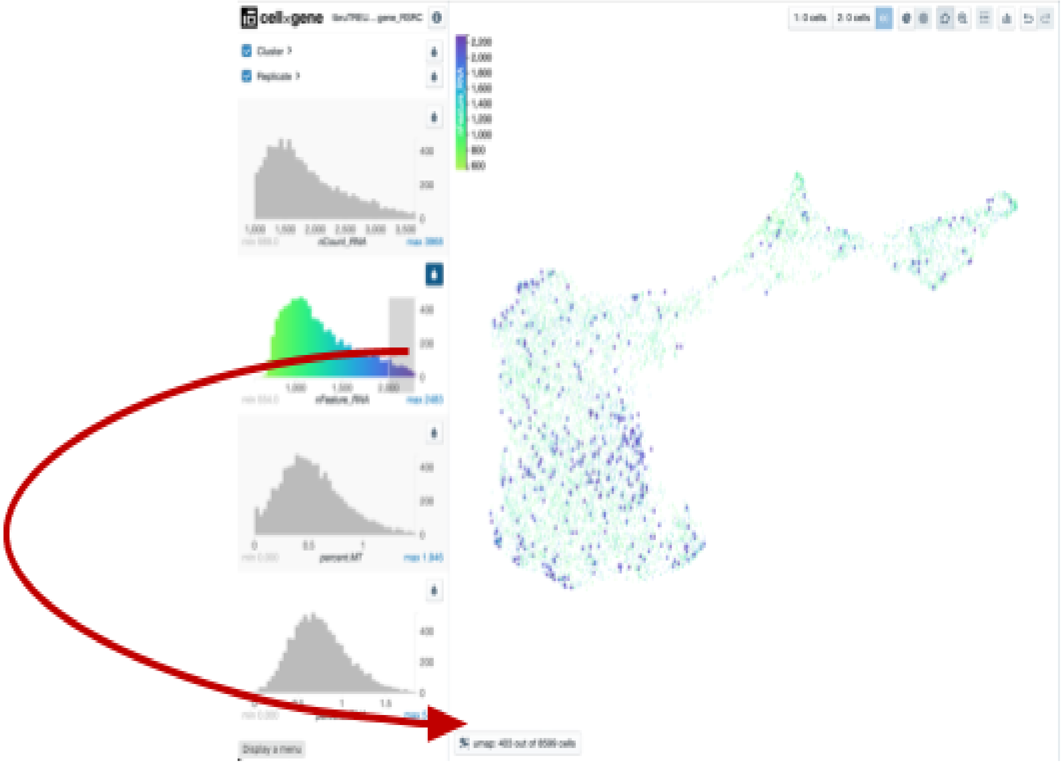
f. The droplet icon can also be used to color cells based on continuous metadata. Generally, the continuous metadata available is provided for QC purposes. Try this: click on the droplet icon for the nFeature_RNA (number of genes detected in each cell). How many cells are displayed?


Color scale

Color droplet

Cell count & reduction type

g. In single cell data, it is common to capture a variable number of genes from each cell. 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 get an idea of what the data looks like.

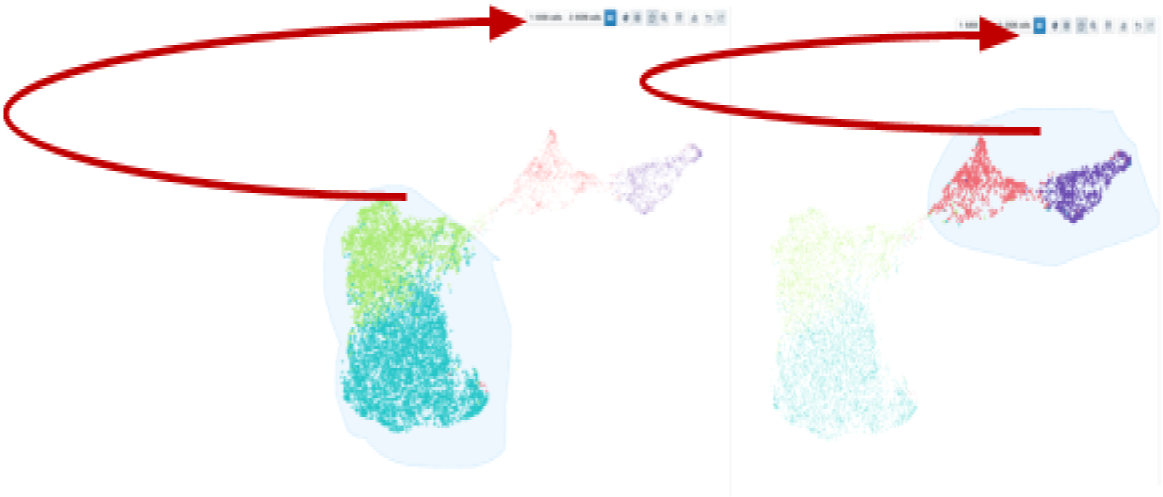


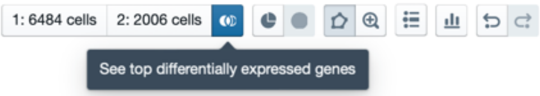
h. Do you think you have a higher percentage of stumpy cells with more genes assayed than the slender forms? Can you get the number of cells in each of the stages that met the >/= 2000 genes representation (See step f)? To do this, click and draw around the stumpy cells in the central panel or use the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

i. 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.

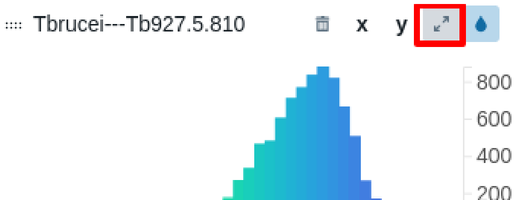
j. Now let us identify genes that differentiate between the stumpy and slender populations. Follow these steps to do this:

    - 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.
    - Repeat the same process to select the slender population and save it as population 2.

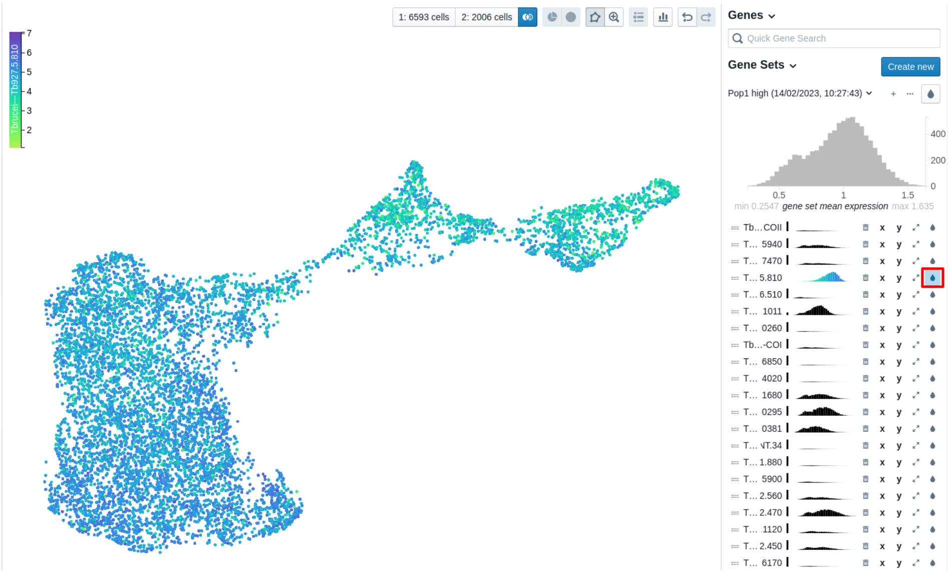


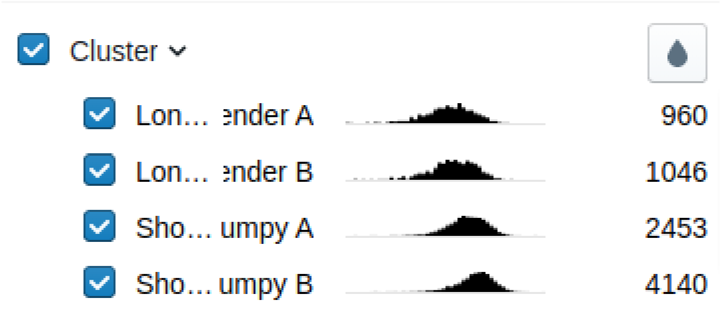


    - When done with your selections and saving populations, click on the differential expression icon.
    - Click on population 1 in the right-hand gene feature panel to reveal the top stumpy genes. Click on the expand icon to view a gene more clearly.



- The histogram in the right panel shows the expression of this gene over all the cells. You can color the UMAP by clicking on the droplet icon next to each gene. The expression of this gene in each cluster can be viewed as histograms in the left panel.



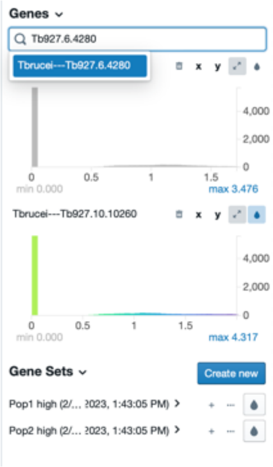


☑ Cluster

☑ Lon… ender A	960
☑ Lon… ender B	1046
☑ Sho… umpy A	2453
☑ Sho… umpy B	4140

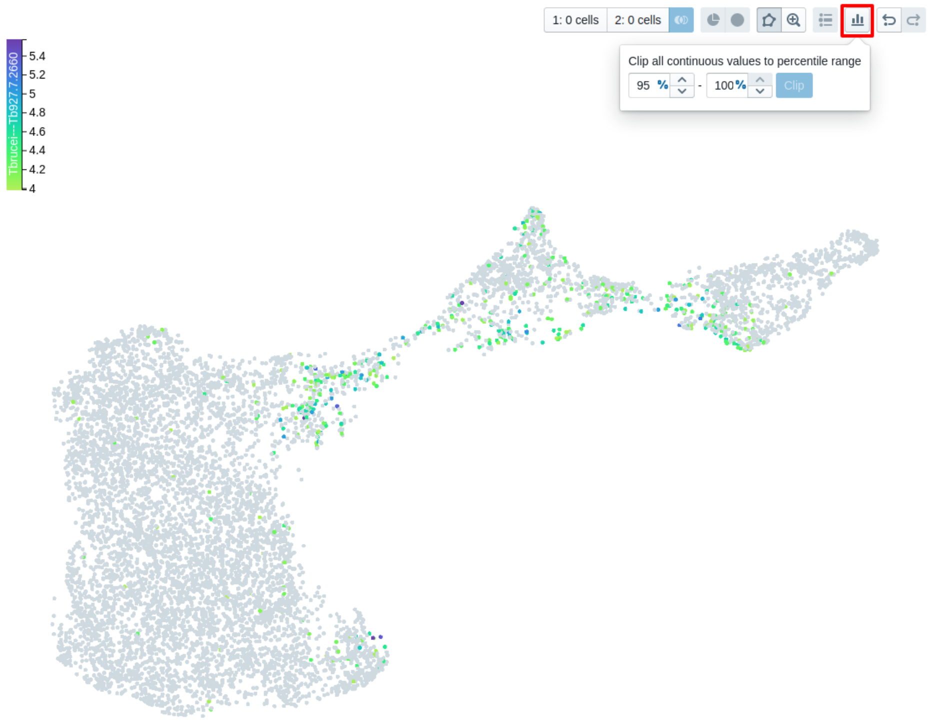
- Copy one of the gene IDs and explore it in TriTrypDB. Can you come up with a rational reason why your selected gene might be important in stumpy development? Note that copying gene IDs from cellxgene is frustrating. If you click on the expand icon for the individual gene, it becomes easier to copy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

k. 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. If the gene is found, you can select it to explore it further. 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



l. 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.

- The gene id is Tb927.7.2660. Find this gene using the quick search, and color the UMAP with expression values for this gene.
- Which cells are expressing this gene at the highest levels? Is it easy to see a pattern just by coloring for this gene?
- 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
- 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.

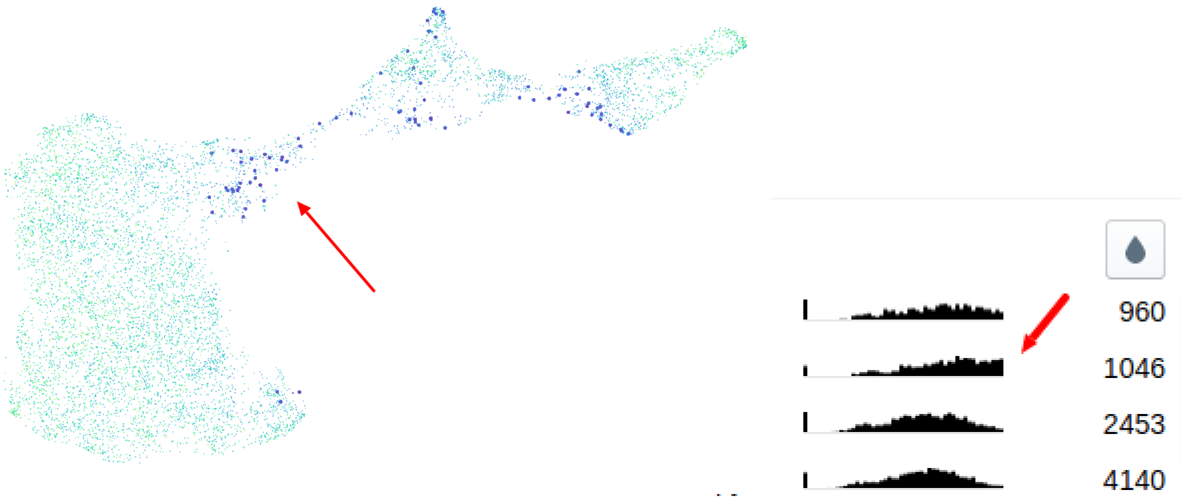

Clip all continuous values to percentile range

95 % - 100 %

- What happens to the UMAP and the histogram? Is it easier to find the cells with the highest expression levels for this gene now?
- Looking at the expression levels, why do you think the authors chose this transcriptional regulator for further study?

4. 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.

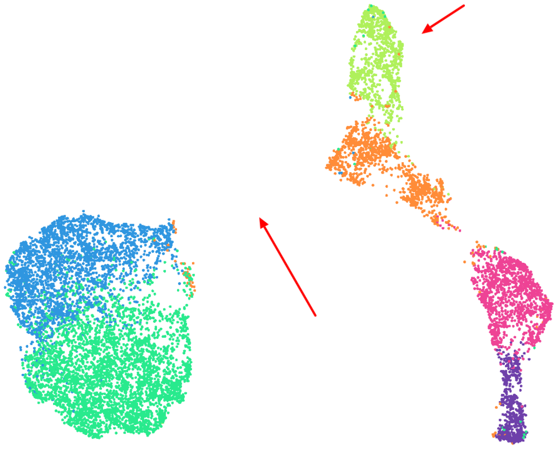
Note that this gene is most highly expressed in the slender B population, and the cells that are transitioning between slender to stumpy. This can be observed in the histograms in the left pane, or by using the gene expression histogram in the right pane to select the cells with the highest expression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

a. 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?

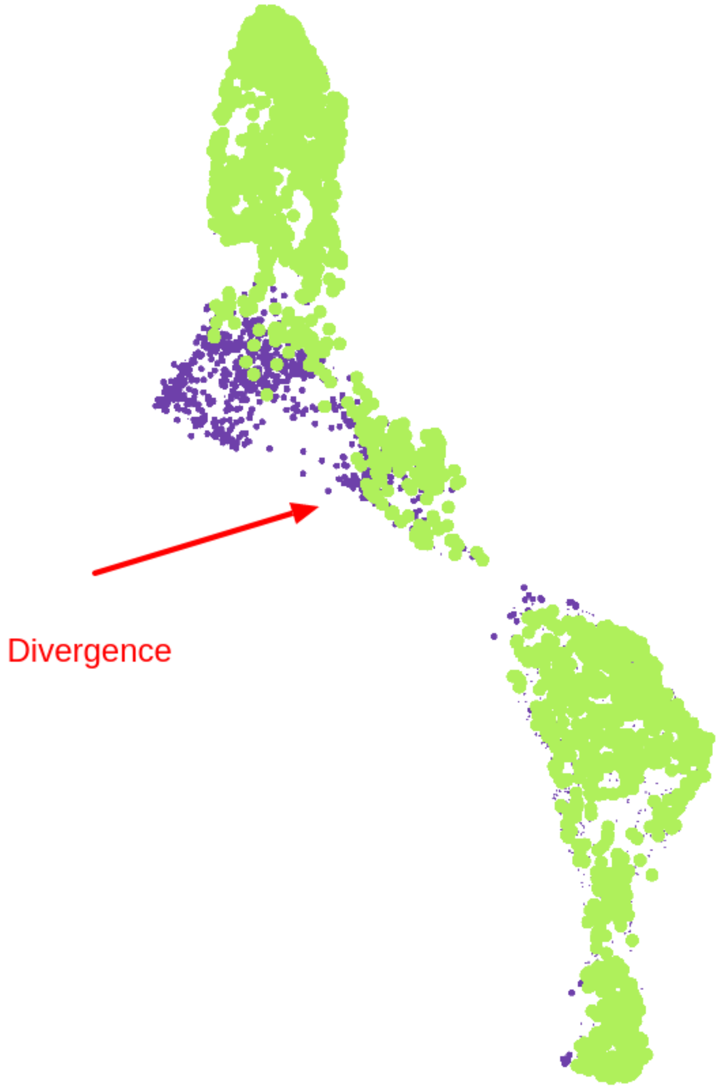


b. Color the UMAP for the expression of the gene that was knocked out (Tb927.7.2660). Use the tools you've already learnt about to explore this.
    - Does this look as you expect?
    - Are there any cell populations where this gene is not expressed?
    - Do you think the knockout was successful?

c. 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.

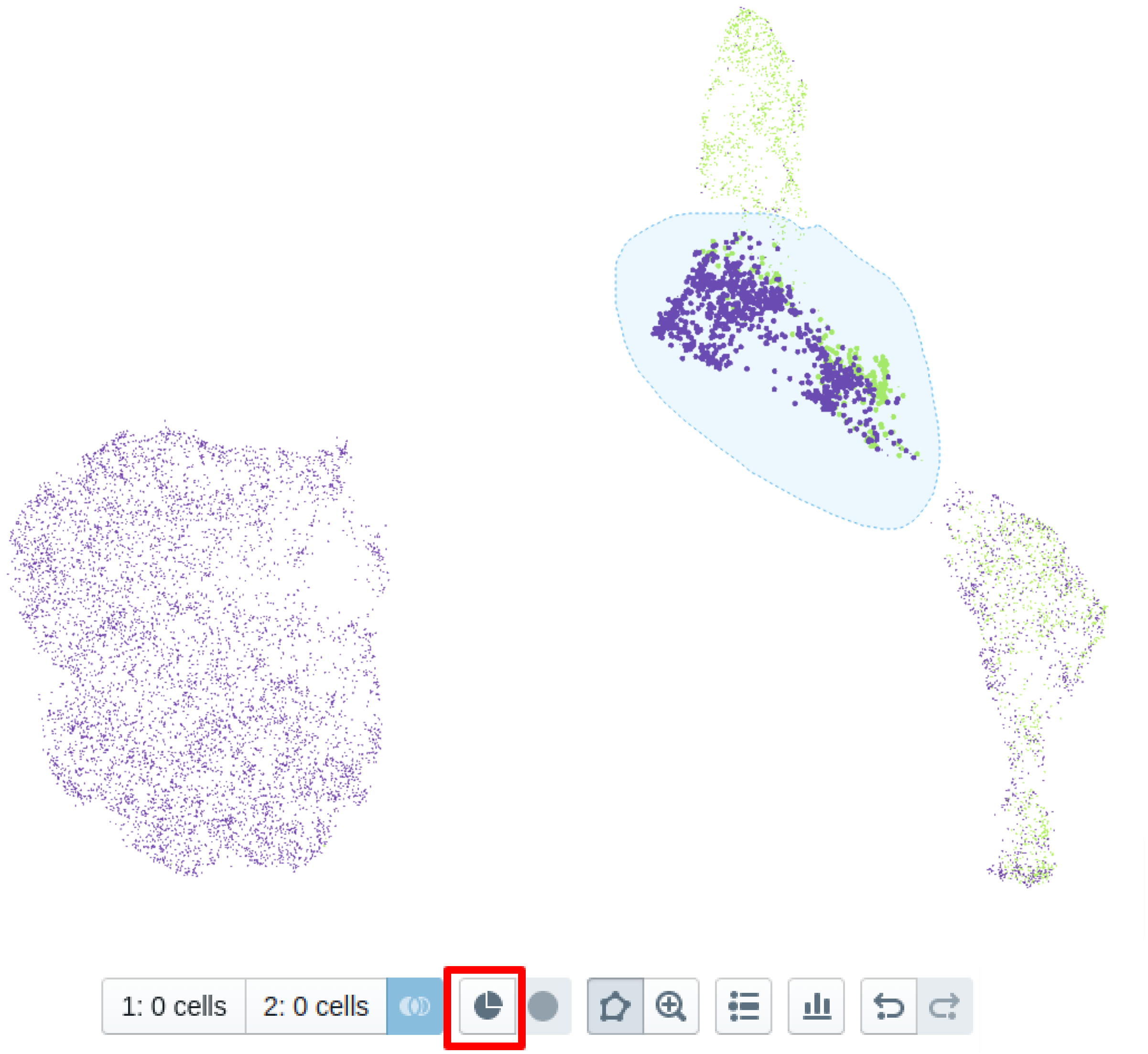
d. It appears that knock-out cells cannot differentiate from slender to stumpy. Instead, they form a novel cluster labeled “Long Slender B.2”. Let’s see what genes are highly expressed in this cluster:
   - Uncheck all the clusters except “Long Slender B.2”. Add this to Population 1
   - Inverse this 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?

e. 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 programme here



Divergence

- 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?


1: 0 cells
2: 0 cells